$\mathbb{S}^4$: Operationalizing Speech Act Theory for Strategic Semi-Structured Psychiatric Interview

Guanqun Bi[1], **Zhoufu Liu**[1,2], **Zhuang Chen**[3,†], **Dazhen Wan**[4],
Xiyao Xiao[4], **Minlie Huang**[1]

[1]CoAI Group, DCST, IAI, BNRIST, Tsinghua University, [2]Beijing Institute of Technology,
[3]Central South University, [4]Lingxin AI,
biguanqun@mail.tsinghua.edu.cn zhchen18@foxmail.com

Abstract

Psychiatric interviewing is a strategic, goal-oriented interaction that requires proactively steering the conversation to elicit latent information. However, existing methods often degenerate into rigid interrogation or aimless chitchat due to a lack of strategic planning. In this work, we introduce $\mathbb{S}^4$, a comprehensive framework grounded in Speech Act Theory, modeling the interview as a unified process of internal strategy (Illocution and Perlocution) and external realization (Locution). We synthesize a large-scale dataset with fine-grained psychiatric speech act annotations. Trained on this data, $\mathbb{S}^4$ employs reinforcement learning driven by long-term therapeutic effects to optimize the strategic chaining of atomic acts, aiming to maximally elicit information and maintain patient engagement. Experiments demonstrate that $\mathbb{S}^4$ significantly outperforms baselines, validating the effectiveness of our effect-driven strategic modeling.

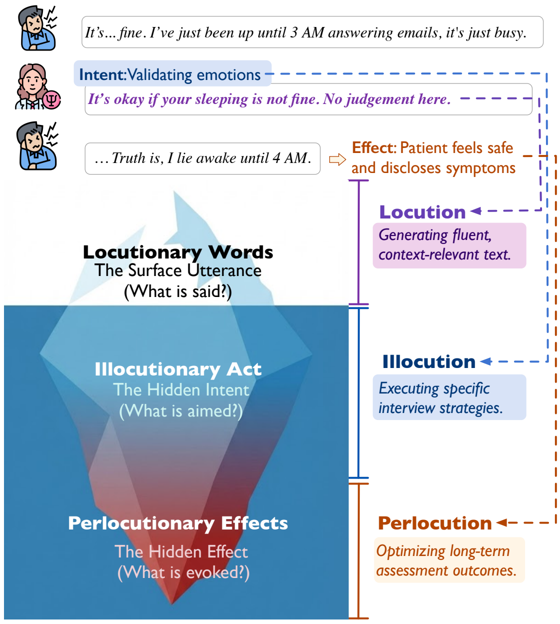


Figure 1: The iceberg of psychiatric interviewing: while existing work focus on surface utterance generation (Locution), effective interviewing requires explicit modeling of hidden strategic intent (Illocution) to achieve long-term assessment outcomes (Perlocution).

1 Introduction

Psychiatric assessment is foundational to clinical treatment. While expert interviews are considered the gold standard, they are labor-intensive, time-consuming, and expensive (Gratch et al., 2014). To scale access, traditional approaches rely on self-report questionnaires (*e.g.*, PHQ-9 (Kroenke et al., 2001)). Although standardized, these tools often trigger patient resistance due to their interrogative rigidity (Lucas et al., 2014; Miner et al., 2016; Daley et al., 2020).

The advent of Large Language Models (LLMs) offers a promising alternative to enable more natural, conversational interactions (OpenAI, 2023a; Anthropic, 2024). However, effectively leveraging this capability for psychiatric interviewing remains an unresolved challenge. One line of work integrates LLMs into questionnaire pipelines (Kebe et al., 2025). Yet, these systems typically inherit the rigidity of traditional forms, prioritizing checklist completion over the rapport required to elicit truthful answers. Another line of work adopts *task-oriented chat* frameworks (Yao et al., 2022; Lan et al., 2024b; Ren et al., 2024) to blend symptom inquiry with empathetic dialogue.

Despite these efforts, a fundamental limitation persists across current paradigms: whether constrained by rigid templates or free-form generation, they rely on surface-level imitation rather than strategic intent planning. They fail to recognize that a professional interview is a semi-structured *strategic loop*: the interviewer must maintain interview rapport as a *prerequisite* to encourage natural disclosure and ultimately achieve precise symptom probing. Lacking this strategic

† Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 38881–38899

grounding, existing models cannot navigate the dependency between interview intent and patient reaction, leaving them unable to sustain the rapport-driven interview flow. Consequently, they often degenerate into either rigid interrogation or aimless chitchat.

To rigorously quantify this deficiency, we first establish a set of evaluation metrics derived from interview standards. Our analysis reveals that state-of-the-art models struggle to coordinate these objectives. Addressing this requires looking beyond surface text to the functional nature of communication. We draw upon **Speech Act Theory (SAT)** (Austin, 1962; Searle, 1969), which posits that speaking is not merely describing the world, but *performing actions* upon it. Based on this, we propose $\mathbb{S}^4$, a framework that operationalizes the interview process through the lens of SAT, as illustrated in Figure 1: (1) **Locution (What is said):** The surface linguistic realization, corresponding to the generation of fluent, context-relevant text. (2) **Illocution (What is aimed):** The hidden strategic intent (e.g., *Validating Emotions*), representing the execution of specific interview strategies. (3) **Perlocution (What is evoked):** The consequential interview effect (e.g., *Patient discloses symptoms* and *feels safe*), describing the dual impact of information utility and rapport building that defines the success of the interview. Guided by the grounding, we first construct a high-quality dataset of simulated interviews where every utterance (locution) is explicitly annotated with interview actions (illocution) and their interview effects (perlocution). By defining a discrete action space within this framework, we transform the interview into a decision process and employ Reinforcement Learning (RL) to optimize long-horizon policies that effectively navigate the full interview flow.

Our contributions are as follows: (1) We propose a SAT-based framework $\mathbb{S}^4$ to redefine psychiatric interviewing, transforming the problem from sequence modeling to goal-oriented strategy planning. (2) We construct a psychiatric interview dataset explicitly annotated with SAT-based action labels and flow structures. (3) We propose an RL-driven framework that optimizes long-horizon interview strategies. (4) Experiments demonstrate that $\mathbb{S}^4$ achieves superior performance over state-of-the-art baselines, validating the necessity of explicit strategic intent modeling grounded in SAT for effective psychiatric interviewing.

2 Related Work

Early automated mental health assessment relied on passive social media mining (Choudhury et al., 2013; Coppersmith et al., 2014; Hassan et al., 2024), which is scalable but lacks the interactivity to probe specific symptoms. To address this, interactive diagnostic systems emerged. Initial approaches adopted rigid slot-filling methods (Kebe et al., 2025; Roy et al., 2025; Zheng et al., 2025), administering questionnaires like a checklist; recent works extend this to clinical settings (Tang et al., 2025; Tu et al., 2024), yet such interrogation-style methods ensure clinical rigor at the cost of rapport. With the advent of LLMs, various control strategies have been explored: generic LLMs (Lorenzoni et al., 2024) offer flexibility but lack protocol adherence; prompting methods (Zhang et al., 2024; Cao et al., 2025) improve controllability via knowledge injection but without global optimization; schema-guided methods (Gu et al., 2025; Wan et al., 2025; Lan et al., 2024b,a) introduce psychological schemas yet mimic surface interactions without capturing strategic intent. Hybrid methods (Yao et al., 2022; Yin et al., 2024; Liu et al., 2025) attempt to blend diagnosis with empathy, but treat them as a binary dichotomy rather than integrated clinical acts. Recently, Ren et al. (2024) applied RL to align with empathy preference. In contrast, $\mathbb{S}^4$ redefines the task as a unified strategic process grounded in SAT, learning to chain therapeutic acts for long-term outcomes.

3 $\mathbb{S}^4$ Framework

We present $\mathbb{S}^4$, a comprehensive framework that operationalizes SAT (Austin, 1962; Searle, 1969) for clinical psychological interviews, as shown in Figure 2.

Speech Act Theory. Our framework builds upon Speech Act Theory (SAT), a foundational theory in pragmatics introduced by Austin (1962) and further developed by Searle (1969). SAT posits that utterances are not merely descriptions of the world but constitute *actions* that speakers perform through language. The theory distinguishes three dimensions of a speech act: (1) *Locution*: the literal content of what is said; (2) *Illocution*: the intended communicative function (*e.g.*, requesting, asserting, or comforting); and (3) *Perlocution*: the actual effect achieved on the listener

Framework	Theoretical Basis	Intent Resolution	Control Paradigm	Optimization Goal
Generic LLMs (*e.g.* OpenAI (2023a))	○*(None)*	○*(Latent/Black-box)*	○*(Probabilistic)*	○*(Text Similarity)*
Slot-filling Methods (*e.g.* Kebe et al. (2025))	◐*(Clinical Scales)*	◐*(Fixed Questions)*	○*(Sequential Slots)*	◐*(Scale Completion)*
Prompting Methods (*e.g.* Cao et al. (2025))	○*(Strategies)*	○*(Implicit/Prompted)*	○*(Instruction-based)*	○*(Task Success)*
Hybrid Methods (*e.g.* Yao et al. (2022))	◐*(Ad-hoc Rules)*	◐*(Coarse/Binary)*	○*(Rigid Pipeline)*	◐*(Static Heuristics)*
Schema-Guided Methods (*e.g.* Lan et al. (2024b))	◐*(Ad-hoc Ontology)*	◐*(Coarse Intents)*	◐*(Supervised)*	◐*(Label Accuracy)*
WundtGPT (*e.g.* Ren et al. (2024))	○*(None)*	○*(Implicit/Prompted)*	◐*(Single-turn)*	◐*(Empathy Preference)*
$\mathbb{S}^4$ **(Ours)**	● **(SAT)**	● **(Atomic Acts)**	● **(Global Policy)**	● **(Therapeutic Effect)**

Table 1: Comparison of modeling paradigms. Symbols denote the level of strategic sophistication: ○ Implicit/Low, ◐ Coarse/Local, ● Explicit/Global. Unlike baselines that rely on surface mimicry or ad-hoc rules, $\mathbb{S}^4$ explicitly models the interview intent (Illocution) and optimizes for long-term therapeutic outcomes (Perlocution).

Action ($\mathcal{A}$)	Sample Locution ($\mathcal{L}$)	Definition & Intended Perlocution ($\mathcal{P}$)
I. Information Seeking *(Directives: Eliciting Disclosure)*		
`Explore`	"*How have you been sleeping?*"	**Solicit Narrative**: Ask open-ended questions to elicit detailed disclosure and **expand symptom scope**.
`Probe`	"*Could you tell me more about that?*"	**Deepen Inquiry**: Follow up on ambiguity to **clarify details** and deepen focus.
`Confirm`	"*Do you feel this way every day?*"	**Pinpoint Fact**: Ask closed-ended questions to **verify diagnostic criteria**.
`Clarify`	"*By 'fatigue', I mean tiredness.*"	**Resolve Confusion**: Provide explanations to **align cognition** and correct misunderstandings.
II. Affective Regulation *(Expressives: Modifying State)*		
`Validate`	"*That sounds incredibly hard.*"	**Affirm Emotion**: Acknowledge patient distress to **lower defensiveness** and build trust.
`Support`	"*I understand. Please go on.*"	**Maintain Flow**: Use back-channeling to demonstrate active listening and **boost efficacy**.
`Ease`	"*Do you have any hobbies?*"	**Reduce Tension**: Engage in non-clinical conversation to **de-escalate anxiety** and humanize the agent.
III. Interview Management *(Representatives: Setting Frame)*		
`Initiate`	"*Hi, I'm your AI counselor.*"	**Set Frame**: Establish professional boundaries and the **purpose of the session**.
`Conclude`	"*Thanks for sharing. Take care.*"	**Ensure Closure**: Formally end the session to provide a **safe exit** and consolidation.

Table 2: The $\mathbb{S}^4$ Action Space Definition. Our framework models the interview as a top-down decision process: the agent first selects an Illocutionary Action ($\mathcal{A}$), which constrains the surface Locution ($\mathcal{L}$) to achieve a targeted Perlocutionary Effect ($\mathcal{P}$).

(*e.g.*, persuading, calming, or informing). This tripartite distinction provides a principled lens for analyzing strategic communication, making it particularly suitable for modeling psychiatric interviews where the interviewer's intent and the patient's response are both critical to assessment success.

3.1 Task Formalization via SAT

Drawing from SAT, we formalize semi-structure psychiatric interview as a strategic dialogue generation process. Unlike traditional SAT analysis which often focuses on post-hoc classification, our framework is constructive: First, at the *illocutionary level*, the policy π_θ selects a strategic action $a_t \sim \pi_\theta(a|\mathcal{H}_t)$ (*e.g.*, `Probe`, `Validate`) from a predefined clinical action space $\mathcal{A}$ based on dialogue history $\mathcal{H}_t$. Second, at the *locutionary level*, the generator produces the utterance $s_t \sim \pi_\theta(s|a_t, \mathcal{H}_t)$ explicitly conditioned on the selected intent. Finally, the system is optimized to maximize *perlocutionary* goals: information utility and rapport building. See Appendix A for the detailed mathematical formulation.

3.2 Action Space Design

To explicitly control the illocutionary dimension, we define a fine-grained action space $\mathcal{A}$ comprising 9 atomic acts. We structure these acts into three functional groups corresponding to the fundamental dimensions of psychiatric interviewing: information gathering, affective regulation, and session management. Each action type corresponds to a specific illocutionary function, originally extracted from DAIC-WOZ (Gratch et al., 2014) and subsequently reviewed and refined by clinical psychology experts to ensure alignment with professional interview practice. As illustrated in Table 2, the action space is organized into three functional groups: **Information Seeking** for eliciting clinical disclosures, **Affective Regulation** for managing emotional states and interview alliance, and **Interview Management** for maintaining conversational structure.

3.3 Dialogue Synthesis

We construct $\mathbb{S}^4\mathbb{DATA}$, a synthesized interview dataset with explicit action annotations.

Dialogue Synthesis Pipeline. We construct $\mathbb{S}^4\mathbb{DATA}$ through a rigorous process designed to

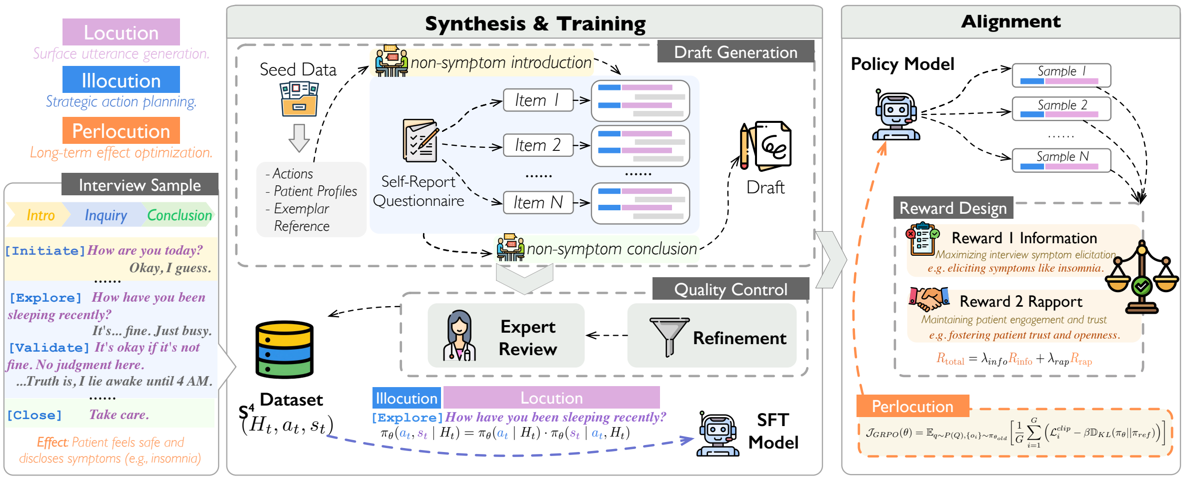
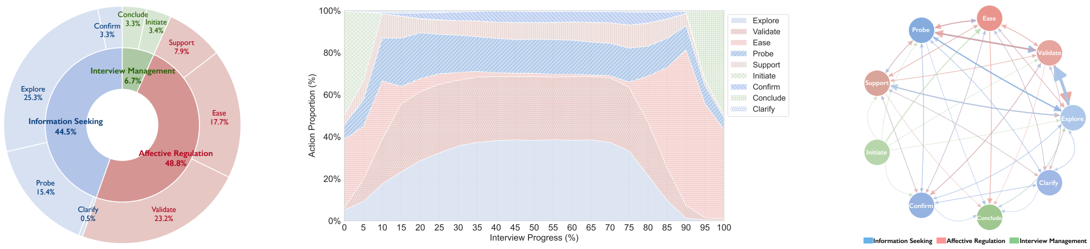


Figure 2: Overview of $\mathbb{S}^4$, a framework aligned with SAT. Left: synthesizing $\mathbb{S}^4$DATA with explicit action annotations and expert validation. Right: a two-stage paradigm where SFT teaches action-conditioned generation (Locution + Illocution), and RL optimizes long-term outcomes (Perlocution) through dual rewards for information elicitation and rapport.

ensure symptom coverage and realism. We utilize a simulated patient dataset where patient personas are derived from the DAIC-WOZ dataset (Gratch et al., 2014) and augmented by LLMs. The patient profiles consist of two components: *Basic Information* and *Scale Information*. The *Basic Information* includes key attributes such as gender, age, occupation, and mental state, with profile keys adapted from professional clinical intake standards. The *Scale Information* is derived from validated PHQ-8 scales, where each item is scored from 0 (not at all) to 3 (nearly every day), covering eight symptom dimensions: anhedonia, depressed mood, sleep disturbance, fatigue, appetite changes, feelings of worthlessness, concentration difficulties, and psychomotor changes. The patient roles are enacted by GPT-4o (OpenAI, 2024), selected for its proven capabilities in role-playing and emotional expressiveness (Engel, 1977; Macneil et al., 2012; Beck et al., 1979; Barrows, 1993). A detailed example of a simulated patient profile is provided in Appendix B.1.

Based on these profiles, dialogues are synthesized via a two-phase pipeline. In the *draft generation* phase, we first generate an initial interview focused on rapport building and general inquiry conditioned on the patient profile, then inject diagnostic segments for symptoms not adequately covered at optimal insertion points to maintain context coherence, and finally append casual turns to ensure a safe ending. In the *quality control* phase, we apply three refinement strategies to bridge the gap between synthetic and real clinical data: (i) *Action Filtering* uses heuristic rules to correct inconsistent action labels; (ii) *Response Diversification* retrieves and adapts diverse empathetic responses from a real-world counseling corpus to prevent generic outputs; (iii) *Style Transfer* rewrites patient utterances to match the linguistic style (*e.g.*, hesitation markers, sentence length) of real DAIC-WOZ transcripts. Additionally, expert clinical psychologists review a random subset of synthesized dialogues to correct labels and responses, validating clinical fidelity.

Dataset Statistics. We processed 189 distinct role cards across 6 severity levels, resulting in 938 synthesized sessions with a total of 26,607 dialogue turns. The dataset covers diverse symptom severities and patient backgrounds. As shown in Figure 3, the action distributions and transitions align with professional protocols: *Affective Regulation* actions constitute the largest proportion (48.8%), underscoring the critical role of rapport-building, while *Information Seeking* (44.5%) dominates diagnostic phases and *Interview Management* (6.7%) provides structural framing. The temporal dynamics further reveal a natural phase progression—from rapport-building (`Initiate`, `Ease`) in the opening, through intensive information gathering (`Explore`, `Probe`) in the diagnostic phase, to emotional consolidation (`Conclude`) at closure. The transition network shows clinically meaningful patterns: `Explore` leading into `Probe` and then `Confirm` forms recurring diagnostic loops, empathetic pivots to



(a) Action distribution (b) Strategy evolution (c) Action transitions

Figure 3: Analysis of action space in $\mathbb{S}^4\mathbb{DATA}$: (a) distribution across functional categories, (b) temporal evolution across interview phases, and (c) transition network revealing strategic flow patterns.

`Validate` follow distress disclosure, and `Clarify` serves as a repair pathway. Extended analysis is in Appendix B.2.

3.4 Training Pipeline of $\mathbb{S}^4\mathbb{DIAL}$

We adopt a two-stage training pipeline to instill interview capabilities into $\mathbb{S}^4$. The first stage (SFT) focuses on the illocutionary and locutionary levels, teaching the model *how* to speak; the second stage (RL) focuses on the perlocutionary level, teaching the model *why* to speak.

Stage I: Supervised Fine-Tuning (SFT) To instill the basic clinical capabilities defined in the illocutionary and locutionary layers, we first fine-tune the base model π_θ on the synthesized dataset $\mathbb{S}^4\mathbb{DATA}$. This stage focuses on teaching the model to validly select actions from $\mathcal{A}$ and generate fluent, coherent utterances conditioned on these actions. We optimize the standard negative log-likelihood loss:

$$\mathcal{L}_{SFT} = -\mathbb{E}_{(\mathcal{H}_t, a_t, s_t)\sim \mathbb{S}^4\mathbb{DATA}}[\log \pi_\theta(a_t, s_t|\mathcal{H}_t)] \tag{1}$$

where $\mathcal{H}_t$ is the dialogue history, a_t is the strategic action, and s_t is the utterance. While SFT ensures the model follows interview norms and action protocols, it primarily mimics the surface form of the expert demonstrations and lacks the long-horizon planning required for strategic interviewing.

Stage II: Reinforcement Learning (RL) To bridge the gap between "saying the right thing" (locution/illocution) and "achieving the right effect" (perlocution), we employ Reinforcement Learning to optimize a long-horizon policy driven by the ultimate goals of interview practice. We formulate a dual-reward function R_{total}, capturing the two fundamental dimensions of interview effectiveness:

$$R_{\text{total}} = \lambda_{info} R_{\text{info}} + \lambda_{rap} R_{\text{rap}} \tag{2}$$

The first component, **Information Utility** ($R_{\textbf{info}}$), targets the *instrumental* perlocutionary goal. It rewards the effective elicitation of symptom information by incentivizing the use of informative language and precise skills to efficiently clarify symptom ambiguity, while penalizing redundant or irrelevant inquiries. The second component, **Rapport Building** ($R_{\textbf{rap}}$), targets the *relational* perlocutionary goal. It evaluates whether the interviewer's response effectively maintains the interviewer-patient alliance, thereby reducing patient resistance and fostering a safe environment for truthful disclosure.

By optimizing these complementary signals via Group Relative Policy Optimization (GRPO) (Shao et al., 2024), $\mathbb{S}^4$ learns to jointly enhance both information gathering and rapport building, where effective emotional support facilitates more open disclosure. The objective is:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}\sim\pi_{\theta_{old}}}\left[\frac{1}{G}\sum_{i=1}^{G}\Big(\mathcal{L}_i^{clip} - \beta\mathbb{D}_{KL}(\pi_\theta||\pi_{ref})\Big)\right] \tag{3}$$

where A_i is the advantage derived from reward R. This formulation enables stable credit assignment across the full dialogue trajectory.

4 Experiments and Analysis

We design experiments to answer the following Research Questions (RQs):

RQ1 (Overall Effectiveness): Does SAT-based framework $\mathbb{S}^4$ outperform state-of-the-art baselines in both automatic and human evaluations?

RQ2 (Illocutionary): Can $\mathbb{S}^4$ successfully model and execute strategic interview intents?

RQ3 (Locutionary): Does explicit action guidance produce efficient and normative language?

RQ4 (Perlocutionary): Do strategic interventions achieve superior diagnostic outcomes and patient rapport?

4.1 Experimental Setup

Dataset. We additionally constructed 200 simulated patients homologous to those in $\mathbb{S}^4\mathbb{DATA}$ (see Section 3.3) for evaluation. In the experiments, the models play the role of the Interviewer to conduct psychiatric interviews with these simulated patients.

Implementation Details. We performed SFT on Qwen3-8B using 8 Nvidia H20 GPUs (BS=32, LR=5e-5, 3 epochs). RL training employed GRPO on 1 Nvidia H20 GPU with LoRA (rank=32, alpha=16), setting LR=5.0e-6 and generating 2 sequences per prompt for preference optimization. Evaluation used GPT-4o (temp=0) as the judge. Full details are in Appendix C.

Baselines. We compare our method against several state-of-the-art general-purpose LLMs, including Qwen3-8B/32B (Yang et al., 2025), GPT-4o (OpenAI, 2024), GPT-OSS-120B, DeepSeek-V3 (DeepSeek-AI, 2024), and DeepSeek-R1 (DeepSeek-AI, 2025). Additionally, we include task-specific baselines fine-tuned on psychiatric interview dialogue datasets, specifically DAIC-WOZ (Gratch et al., 2014) and MDD5k (Yin et al., 2024), which are fine-tuned on the same Qwen3-8B backbone as our model for fair comparison. We also include WundtGPT, which is trained on the D4 dataset using KTO (Ren et al., 2024). All baseline models are prompted with carefully designed instructions specifying the role of a psychological interviewer conducting PHQ-8 assessment interviews, using few-shot prompting with 2-3 demonstration examples. Detailed descriptions of each baseline model are provided in Appendix C.2.

Evaluation Metrics. We establish a multi-dimensional evaluation framework derived from clinical expert consensus and professional interviewing standards (Gratch et al., 2014; Engel, 1977). We assess eight response quality dimensions: Information Density (Dens.) measures how efficiently utterances convey information; Linguistic Accuracy (Acc.) evaluates naturalness of symptom description; Skill Diversity (Div.) evaluates the variety of distinct action types used; Skill Balance (Bal.) assesses whether action proportions align with clinical norms; Skill Precision (Prec.) measures contextual appropriateness of each action; Cognitive Ease (Cog.) evaluates how readily patients can process utterances; Relationship Building (Rel.) tracks therapeutic alliance; and Emotional State (Emo.) monitors patient affect trajectory. Additionally, we report three diagnostic outcome metrics: Symptom Recall Rate (Symp.R.), Binary Classification F1 (Bin.F1), and Five-class Severity F1 (5cls.F1). All eight response quality metrics are evaluated using GPT-4o as an independent judge (temperature $= 0$, seed $= 998244353$); detailed prompts and criteria are provided in Appendix D.

4.2 RQ1: Overall Effectiveness

Automatic Evaluation. Table 3 presents a comprehensive comparison between $\mathbb{S}^4$ and state-of-the-art baselines. We analyze the results across two key dimensions:

Response Quality. Compared to general LLMs like GPT-4o and DeepSeek-V3, $\mathbb{S}^4$ demonstrates superior performance in most response quality metrics, particularly in skill precision (9.19) and cognitive ease (8.64). This indicates that our hierarchical framework effectively translates interview intent into precise, concise, and low-burden language, avoiding the "verbosity bias" and aimless chitchat common in baseline models. Notably, $\mathbb{S}^4$ achieves a near-perfect emotional state score (9.17), significantly outperforming DeepSeek-R1 (5.55), proving that our model maintains high emotional support rather than just mechanically extracting information.

Diagnostic Metrics. $\mathbb{S}^4$ also surpasses baselines as a highly effective diagnostic tool. It dominates in symptom recall rate (97.76%) and binary F1 (89.67%), significantly outperforming both general-purpose models (e.g., GPT-4o's 57.06% F1) and task-specific baselines (e.g., MDD5k's 37.50% F1). This confirms that $\mathbb{S}^4$ successfully balances clinical efficiency with human-centric care, achieving the highest weighted average score (8.01) among all systems.

Ablation Study. To validate the contribution of each component, we compare $\mathbb{S}^4$ with its variants. Removing the RL stage (w/o RL) leads to a drastic drop in binary F1 (89.67% $\rightarrow$ 63.44%) and emotional state (9.17 $\rightarrow$ 8.08), although skill precision remains relatively high (8.89). This suggests that while SFT teaches the model *how* to speak (Illocution/Locution), it is the perlocutionary-driven RL that teaches it *what* strategy achieves the best diagnostic outcomes. Conversely, removing the SFT stage (w/o SFT) results in a significant de-

Model	Response Quality Metrics (0-10 Scale)								Diagnostic Metrics (%)			Avg.
	Dens.	Acc.	Div.	Bal.	Prec.	Cog.	Rel.	Emo.	Symp.R.	Bin.F1	5cls.F1	
General-Purpose LLMs (Baseline)												
Qwen3-8B	5.87	6.90	6.80	7.17	4.60	7.06	5.99	5.17	94.41	52.00	24.10	5.94
Qwen3-32B	6.69	7.87	6.84	5.83	6.22	6.37	6.46	5.01	96.45	56.57	34.66	6.33
GPT-4o	6.82	7.62	6.77	6.70	7.60	7.36	7.29	5.29	92.59	57.06	24.07	6.36
GPT-OSS-120B	7.24	**8.17**	6.50	4.13	6.84	7.09	7.45	5.23	95.01	56.57	31.27	6.34
DeepSeek-V3	7.78	7.24	7.24	5.83	7.77	7.20	7.43	5.36	95.24	76.66	30.87	6.87
DeepSeek-R1	7.32	7.74	**7.70**	6.03	6.78	7.28	7.57	5.55	95.45	75.47	45.33	7.10
Task-Specific Baselines												
DAIC-WOZ	7.80	0.93	4.96	3.40	6.56	7.97	4.66	4.42	85.06	59.89	28.07	5.43
MDD-5k	**8.19**	5.32	6.31	3.33	8.01	5.95	4.62	0.72	96.75	37.50	10.12	5.06
WundtGPT	6.53	2.18	3.52	3.33	5.17	7.66	3.76	2.50	94.27	47.25	22.00	4.89
Our Models												
$\mathbb{S}^4$	6.90	6.65	6.19	7.80	**9.19**	**8.64**	7.65	**9.17**	**97.76**	89.67	59.72	**8.01**
w/o SFT	5.50	6.82	6.66	5.47	6.36	7.37	5.97	6.14	91.46	**91.79**	**70.12**	7.37
w/o RL	7.04	7.24	6.34	**8.10**	8.89	8.52	**7.67**	8.08	92.90	63.44	34.50	7.05

Table 3: Main results comparing baseline LLMs, task-specific baselines, and our models. **Bold** indicates best performance, underline indicates second-best.

cline in skill precision (9.19 $\rightarrow$ 6.36) and skill balance (7.80 $\rightarrow$ 5.47). This confirms that the SFT phase is crucial for grounding the model in the clinical action space and ensuring professional adherence. The full $\mathbb{S}^4$ model effectively integrates the strengths of both stages.

Human Evaluation. To complement the automatic metrics, we conduct human evaluation with three licensed psychologists (see Appendix E for details). Evaluators rated 50 randomly sampled dialogues (presented in randomized order with model identities concealed) on a 1-5 Likert scale across four dimensions: Symptom Elicitation (SE, whether the interviewer effectively gathers symptom information), Empathic Authenticity (EA, genuineness of emotional responses), Facilitation of Disclosure (FD, ability to encourage patient openness), and Cognitive Appropriateness (CA, logical coherence of responses). Inter-rater agreement was substantial (Cohen's κ = 0.71). Additionally, experts conducted a role-play task simulating patients for 10 sessions per model, assessing Diagnostic Success (DS)—whether systems could complete PHQ-8 assessment with accurate conclusions. As shown in Table 4, $\mathbb{S}^4$ achieves the highest scores across all dimensions (avg. 4.59) and 93.3% DS rate, outperforming GPT-4o (80.0%) and DeepSeek-R1 (73.3%).

Model	SE	EA	FD	CA	Avg.	DS (%)
GPT-4o	4.12	3.45	3.68	3.10	3.59	80.0
DeepSeek-R1	4.25	3.30	3.75	3.25	3.64	73.3
MDD-5k	2.80	1.95	2.10	2.55	2.35	36.7
$\mathbb{S}^4$	**4.65**	**4.48**	**4.52**	**4.70**	**4.59**	**93.3**

Table 4: Human evaluation results. SE/EA/FD/CA are rated on 1-5 Likert scale; DS (Diagnostic Success) is the percentage of successful PHQ-8 assessments in expert role-play sessions.

4.3 RQ2: Illocutionary Level

We examine whether $\mathbb{S}^4$ can effectively select appropriate Illocutionary Acts to drive the conversation. Table 3 shows $\mathbb{S}^4$ achieves the highest skill precision and skill balance among all models, significantly outperforming both general-purpose LLMs and task-specific baselines. Figure 4 visualizes the distribution of system actions conditioned on patient emotional states, revealing distinct adaptive strategies across three conditions. When patients exhibit `High Distress`, $\mathbb{S}^4$ strongly prioritizes `Validate` and substantially reduces information-seeking actions like `Explore`, reflecting an emotion-first strategy that stabilizes the patient before proceeding with assessment. Under `Normal` states, the action distribution shifts markedly: the proportion of `Explore` increases while `Validate` recedes, indicating that $\mathbb{S}^4$ leverages stable emotional moments to advance the clinical assessment more actively. Notably, the `Resistance` state reveals a third distinct pattern. Here the model maintains a substantial level of `Validate` alongside continued `Explore`, while the use of `Ease` rises compared to the other two states. This combination suggests a de-escalation strategy

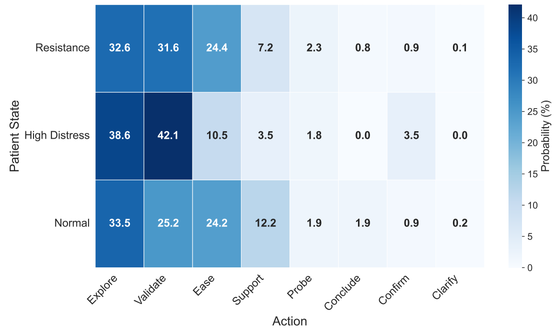


Figure 4: Analysis of illocutionary strategic patterns: the heatmap visualizes action selection probabilities across three patient emotional states. `Initiate` is excluded as it always occurs at the dialogue start.

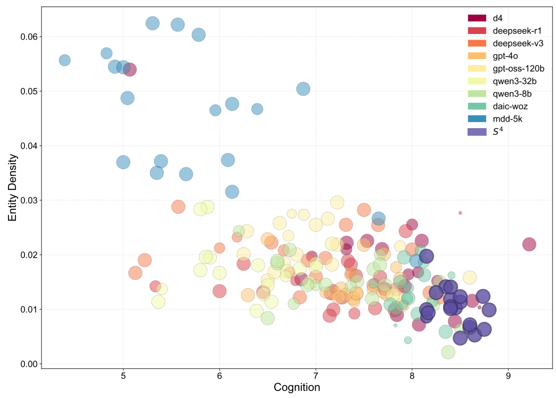


Figure 5: Analysis of locutionary linguistic efficiency: the bubble chart visualizes cognitive ease (x-axis), entity density (y-axis), and symptom recall rate (bubble size).

where the model simultaneously acknowledges patient reluctance through validation, softens the interaction through rapport-building, and gently persists with clinical inquiry rather than abandoning the assessment entirely. In contrast, baseline models without explicit action modeling lack this granular adaptability. General-purpose LLMs tend to apply a uniform conversational style regardless of patient emotional shifts, and task-specific baselines trained on fixed-format interviews cannot dynamically rebalance their strategies in response to patient states.

4.4 RQ3: Locutionary Level

We evaluate whether explicit action guidance produces clinically efficient language. Table 3 shows $\mathbb{S}^4$ achieves the highest cognitive ease and information density among all models. As shown in Figure 5, baseline models exhibit a scattered distribution that reveals a fundamental tension between clinical thoroughness and patient accessibility. MDD-5k occupies the upper-left region with high entity density and low cognitive ease, indicating that its responses are clinically dense but linguistically burdensome for patients. Qwen3-8b sits in the high-cognitive-ease region but with a notably small bubble, reflecting language that is accessible yet diagnostically shallow. GPT-4o falls in the middle of both dimensions without a clear advantage in either. In contrast, $\mathbb{S}^4$ is positioned in the bottom-right quadrant with the largest bubble, simultaneously achieving high cognitive ease, low entity density, and superior symptom recall rate. This demonstrates that the explicit illocutionary-to-locutionary mapping enables the model to generate language that is both clinically precise and cognitively light, extracting more diagnostic information per utterance without burdening the patient with dense or complex phrasing. Figure 6 presents two qualitative case studies that illustrate this principle. In Case 1, $\mathbb{S}^4$ demonstrates clinical foresight by linking concentration difficulties to psychomotor symptoms across PHQ items 7 and 8, while GPT-4o offers generic advice unrelated to the diagnostic framework. In Case 2, $\mathbb{S}^4$ prioritizes the `Validate` action to establish therapeutic alliance when the patient exhibits self-deprecation, whereas GPT-4o mechanically pivots to the next symptom domain without acknowledging the emotional disclosure.

4.5 RQ4: Perlocutionary Level

Finally, we evaluate the perlocutionary effect of $\mathbb{S}^4$ on two critical outcomes: symptom collection completeness and patient rapport maintenance. Table 3 shows $\mathbb{S}^4$ achieves the highest symptom recall rate and emotional state score among all models, while also leading in binary diagnostic F1.

As shown in Figure 7, baseline models exhibit diverse and often problematic trajectory patterns. DeepSeek-R1, despite its strong reasoning capabilities, shows a characteristic pattern where aggressive early symptom elicitation corresponds to a steady decline in patient emotional state over the course of the interview, suggesting that its interrogation-like approach damages therapeutic rapport as the conversation progresses. GPT-4o maintains a more moderate emotional trajectory but advances slowly on PHQ-8 coverage, reflecting a tendency toward cautious, open-ended conversation that prioritizes comfort over diagnostic efficiency. $\mathbb{S}^4$ exhibits a notably superior pattern that resolves this apparent tension. The PHQ

Case Studies: $\mathbb{S}^4$ vs. Baselines

Case 1: Symptom Association | *Detecting Psychomotor issues from Attention deficits (PHQ Item 7 → 8)*

User: "When reading, I might lose focus after a while... TV is a bit better but sometimes I get distracted there too."

$\mathbb{S}^4$**:** **[Action: Probe]** "You mentioned losing focus or finding it hard to concentrate. I'd like to understand more: have you noticed if your **movement or speaking speed has changed significantly?** For example, becoming notably slower, or conversely, feeling restless and unable to sit still?"

GPT-4o (Failure: Generic Advice): "I understand that focusing can be challenging. Sometimes when we're dealing with difficult emotions, it helps to organize our thoughts. Have you tried making a playlist of songs that make you feel better? Also, can you summarize for me how you've been feeling overall in the past two weeks?"

MDD-5K (Failure: Topic Drift): "That's interesting. Would you describe yourself as more of an introvert or extrovert? I'm asking because sometimes personality traits can affect how we experience these symptoms."

Case 2: Emotional Validation | *Prioritizing Alliance over Efficiency (PHQ Item 6)*

User: "I feel pretty useless. I can't even start small tasks. Others must think I'm just like this."

$\mathbb{S}^4$**:** **[Action: Validate]** "It sounds like you actually have **high standards for yourself**, and when things don't go as expected, that disappointment is truly hard to bear. I can understand why you'd feel that way. Thank you for sharing this with me."

GPT-4o (Failure: Mechanical): "Thank you for sharing that with me. I can hear that this has been really difficult for you. Let's move on to another area—how has your sleep been recently? Have you noticed any changes in your sleep patterns over the past two weeks?"

DeepSeek-R1 (Failure: Ignore Emotion): "I understand. Can you tell me more about your energy levels? How would you rate your fatigue on a scale from 1 to 10? And have you noticed any changes in your appetite?"

Figure 6: Comparative analysis of interview strategies. Our model $\mathbb{S}^4$ demonstrates superior strategic intent (Illocution) compared to baselines, which suffer from generic advice (Case 1) or mechanical slot-filling (Case 2).

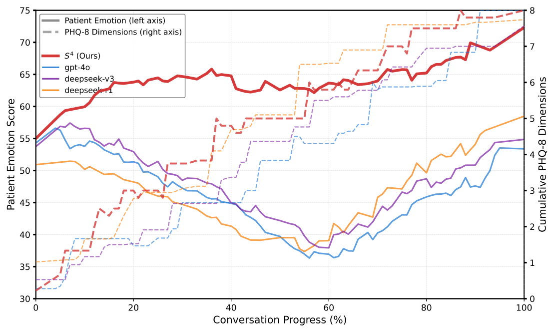


Figure 7: Analysis of perlocutionary effects: the dual-axis chart visualizes Emotional State (solid line) and PHQ-8 Information Collection (dashed line) across different models.

coverage line ascends steadily and comprehensively across all eight symptom domains, while the emotional state line remains consistently elevated throughout the dialogue, significantly outpacing all baselines at every stage. This dual achievement confirms that the perlocutionary-driven RL optimization successfully teaches the model to treat diagnostic completeness and patient well-being not as competing objectives but as mutually reinforcing goals, where maintaining rapport facilitates deeper disclosure and thus more efficient information gathering.

5 Conclusion

We presented $\mathbb{S}^4$, a framework that reimagines psychiatric interviewing as a strategic process grounded in Speech Act Theory. By unifying internal strategy (Illocution) with external realization (Locution) and optimizing via long-horizon reinforcement learning, $\mathbb{S}^4$ effectively balances structured symptom inquiry with rapport. Experiments on our expert-annotated dataset confirm that $\mathbb{S}^4$ significantly outperforms baselines in both diagnostic precision and patient engagement. These results validate the efficacy of explicit intent modeling for high-stakes psychiatric interactions. Future work will extend this approach to multimodal and longitudinal care scenarios.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Key Research and Development Program of China (No. 2024YFC3606800), the National Science Foundation for Distinguished Young Scholars (No. 62125604), the National Natural Science Foundation of China (No. 62441614), and the Beijing Natural Science Foundation (No. L252009).

Limitations

Despite the promising results, our work has several limitations that future research should address.

Simulated Data Validity. While $\mathbb{S}^4\mathbb{DATA}$ is constructed based on real clinical personas (derived from DAIC-WOZ), undergoes rigorous quality control with multi-stage filtering and consistency checks, and is reviewed by clinical experts to ensure authenticity and diagnostic validity, the dialogues are ultimately synthesized by LLMs. Real patients often exhibit more subtle linguistic cues, non-linear narratives, and complex non-cooperative behaviors (e.g., denial, cognitive distortion) that may not be fully captured by current simulations. Direct validation with real patients in a controlled clinical setting is a necessary next step.

Modality Constraints. Our current framework operates solely on textual data. In actual psychiatric practice, paralinguistic features (e.g., speech rate, prosody, latency) and non-verbal cues (e.g., facial expressions, body language) provide critical diagnostic evidence. Ignoring these modalities limits the system's ability to detect subtle affective states, such as flat affect or psychomotor retardation.

Scope of Disorder. We primarily focused on depression assessment using the PHQ-8 standard. While the SAT-based framework is theoretically transferable, distinct psychiatric conditions (e.g., anxiety, PTSD, bipolar disorder) require different interviewing strategies and safety protocols. The generalization capability of $\mathbb{S}^4$ to these broader domains remains to be verified.

Ethical Considerations

Clinical Disclaimer and Human Oversight. $\mathbb{S}^4$ is designed as a research tool for studying dialogue strategies in mental health conversations, not a replacement for human clinicians. It aims to assist professionals in the intake and screening process by gathering information and building initial rapport. Final diagnostic decisions, risk assessments, and treatment planning must strictly remain under the supervision of qualified human psychiatrists or psychologists.

Data Privacy and Safety. Our experiments utilize $\mathbb{S}^4\mathbb{DATA}$, which consists of synthesized dialogues based on anonymized public datasets (DAIC-WOZ), effectively mitigating direct privacy risks associated with real patient data. However, any future deployment involving real subjects must comply with strict data protection regulations (e.g., HIPAA, GDPR). We have ensured that the synthetic data generation process does not produce personally identifiable information (PII).

Risk Management and Crisis Intervention. Mental health interactions carry inherent risks, particularly regarding self-harm or suicide. Although the PHQ-8 scale used in this study excludes the suicide ideation item (Item 9) to minimize risk in automated settings, real-world deployment requires robust safety mechanisms. A production system must include a "human-in-the-loop" protocol to detect trigger phrases or high-risk signals immediately, triggering a fallback mechanism that transfers control to a human professional or provides emergency resources.

Bias and Fairness. We acknowledge that the underlying LLMs and the seed personas (derived from DAIC-WOZ) may contain demographic or cultural biases. There is a risk that the model might perform unevenly across different gender, age, or ethnic groups, potentially leading to underdiagnosis or misinterpretation of symptoms for minority populations. Continuous fairness auditing and diverse value alignment are essential before any real-world application.

Human Evaluation and IRB Considerations. Our human evaluation involved three licensed clinical psychologists who evaluated AI-generated dialogue quality (see Appendix E for complete protocol). This study was determined exempt from formal IRB review for the following reasons: (1) the evaluation task involved professional assessment of synthetic AI outputs rather than research on human subjects; (2) no sensitive personal data was collected from the evaluators beyond their professional ratings; (3) the evaluators were compensated professionals providing expert consultation rather than research participants; and (4) no patient data or real clinical interactions were involved—all dialogues were synthetically generated. Evaluators provided informed consent acknowledging the study objectives, voluntary participation, and confidential handling of their ratings.

References

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2024-06-20.

John Langshaw Austin. 1962. *How to do things with words*. Oxford University Press, Oxford, UK.

Howard S Barrows. 1993. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. *Academic Medicine*, 68(6):443–451.

Aaron T Beck, A John Rush, Brian F Shaw, and Gary Emery. 1979. *Cognitive Therapy of Depression*. Guilford Press, New York.

Jieming Cao, Chen Huang, Yanan Zhang, Ruibo Deng, Jincheng Zhang, and Wenqiang Lei. 2025. Breaking the Stigma! Unobtrusively Probe Symptoms in Depression Disorder Diagnosis Dialogue. *Preprint*, arXiv:2501.15260.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013*. The AAAI Press.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych@ACL 2014, Baltimore, Maryland, USA, June 27, 2004*, pages 51–60. Association for Computational Linguistics.

Kate Daley, Ines Hungerbuehler, Kate Cavanagh, Heloísa Garcia Claro, Paul Alan Swinton, and Michael Kapps. 2020. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. *Frontiers Digit. Health*, 2:576361.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

George L Engel. 1977. The need for a new medical model: a challenge for biomedicine. *Science*, 196(4286):129–136.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, and 1 others. 2014. The distress analysis interview corpus of human and computer interviews. In *Lrec*, volume 14, pages 3123–3128. Reykjavik.

Yiyang Gu, Yougen Zhou, Qin Chen, Ningning Zhou, Jie Zhou, Aimin Zhou, and Liang He. 2025. Enhancing Depression-Diagnosis-Oriented Chat with Psychological State Tracking. *Preprint*, arXiv:2403.09717.

Abdelrahaman A. Hassan, Radwa J. Hanafy, and Mohammed E. Fouda. 2024. Automated Multi-Label Annotation for Mental Health Illnesses Using Large Language Models. *Preprint*, arXiv:2412.03796.

Gaoussou Youssouf Kebe, Jeffrey M. Girard, Einat Liebenthal, Justin Baker, Fernando De la Torre, and Louis-Philippe Morency. 2025. LlaMADRS: Prompting Large Language Models for Interview-Based Depression Assessment. *Preprint*, arXiv:2501.03624.

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):606–613.

Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B W Williams, Joyce T. Berry, and Ali H. Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114 1-3:163–173.

Kunyao Lan, Bingrui Jin, Zichen Zhu, Siyuan Chen, Shu Zhang, Kenny Q. Zhu, and Mengyue Wu. 2024a. Depression Diagnosis Dialogue Simulation: Self-improving Psychiatrist with Tertiary Memory. *Preprint*, arXiv:2409.15084.

Kunyao Lan, Cong Ming, Binwei Yao, Lu Chen, and Mengyue Wu. 2024b. Towards Reliable and Empathetic Depression-Diagnosis-Oriented Chats. *Preprint*, arXiv:2404.05012.

June M. Liu, Mengxia Gao, Sahand Sabour, Zhuang Chen, Minlie Huang, and Tatia M. C. Lee. 2025. Enhanced Large Language Models for Effective Screening of Depression and Anxiety. *Preprint*, arXiv:2501.08769.

Giuliano Lorenzoni, Pedro Elkind Velmovitsky, Paulo Alencar, and Donald Cowan. 2024. GPT-4 on Clinic Depression Assessment: An LLM-Based Pilot Study. *Preprint*, arXiv:2501.00199.

Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Comput. Hum. Behav.*, 37:94–100.

Craig A Macneil, Melissa K Hasty, Philippe Conus, and Michael Berk. 2012. Is diagnosis enough to guide interventions in mental health? Using case formulation in clinical practice. *BMC Medicine*, 10:111.

Adam S. Miner, Amanda Chow, Sarah Adler, Ilia Zaitsev, Paul Tero, Alison Darcy, and Andreas Paepcke. 2016. Conversational agents and mental health: Theory-informed assessment of language and affect. In *Proceedings of the Fourth International Conference on Human Agent Interaction, HAI 2016, Biopolis, Singapore, October 4-7, 2016*, pages 123–130. ACM.

OpenAI. 2023a. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2023b. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-05-13.

Chenyu Ren, Yazhou Zhang, Daihai He, and Jing Qin. 2024. WundtGPT: Shaping Large Language Models To Be An Empathetic, Proactive Psychologist. *Preprint*, arXiv:2406.15474.

Kaushik Roy, Harshul Surana, Darssan Eswaramoorthi, Yuxin Zi, Vedant Palit, Ritvik Garimella, and Amit Sheth. 2025. Large Language Models for Mental Health Diagnostic Assessments: Exploring

The Potential of Large Language Models for Assisting with Mental Health Diagnostic Assessments – The Depression and Anxiety Case. *Preprint*, arXiv:2501.01305.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge University Press, Cambridge, UK.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Jinwen Tang, Qiming Guo, Wenbo Sun, and Yi Shang. 2025. A Layered Multi-Expert Framework for Long-Context Mental Health Assessments. *Preprint*, arXiv:2501.13951.

Sichang Tu, Abigail Powers, Natalie Merrill, Negar Fani, Sierra Carter, Stephen Doogan, and Jinho D. Choi. 2024. Automating PTSD Diagnostics in Clinical Interviews: Leveraging Large Language Models for Trauma Assessments. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 644–663.

Tianxi Wan, Jiaming Luo, Siyuan Chen, Kunyao Lan, Jianhua Chen, Haiyang Geng, and Mengyue Wu. 2025. From Medical Records to Diagnostic Dialogues: A Clinical-Grounded Approach and Dataset for Psychiatric Comorbidity. *Preprint*, arXiv:2510.25232.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. 2022. D4: A Chinese Dialogue Dataset for Depression-Diagnosis-Oriented Chat. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2438–2459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Congchi Yin, Feng Li, Shu Zhang, Zike Wang, Jun Shao, Piji Li, Jianhua Chen, and Xun Jiang. 2024. MDD-5k: A New Diagnostic Conversation Dataset for Mental Disorders Synthesized via Neuro-Symbolic LLM Agents. *arXiv preprint arXiv:2408.14142*.

Xiaoheng Zhang, Weigang Cui, Junjie Wang, and Yang Li. 2024. Chat, Summary and Diagnosis: A LLM - Enhanced Conversational Agent for Interactive Depression Detection. In *2024 4th International Conference on Industrial Automation, Robotics and Control Engineering (IARCE)*, pages 343–348, Chengdu, China. IEEE.

Wenjie Zheng, Qiming Xie, Zengzhi Wang, Jianfei Yu, and Rui Xia. 2025. Towards Explainable Multimodal Depression Recognition for Clinical Interviews. *Preprint*, arXiv:2501.16106.

A Detailed Task Formalization via SAT

This section provides the mathematical formalization of the $\mathbb{S}^4$ framework based on Speech Act Theory (SAT), supplementing Section 3.1.

Let $\mathcal{D} = \{(u_1, s_1), \ldots, (u_t, s_t), \ldots\}$ denote a dialogue session, where u_t is the patient's utterance and s_t is the system's response at turn t. The system operates through a hierarchical decision-making process:

Illocutionary Level (Intent Selection). The system first determines a strategic action a_t from a predefined action space $\mathcal{A}$ based on the dialogue history $\mathcal{H}_t$:

$$a_t \sim \pi_\theta(a|\mathcal{H}_t) \tag{4}$$

where π_θ is the policy model. The action space $\mathcal{A}$ consists of 9 atomic acts organized into three functional groups: *Information Seeking*, *Affective Regulation*, and *Interview Management* (see Table 2).

Locutionary Level (Utterance Realization). Conditioned on the selected illocutionary intent a_t and history $\mathcal{H}_t$, the system generates the natural language response s_t:

$$s_t \sim \pi_\theta(s|a_t, \mathcal{H}_t) \tag{5}$$

This hierarchical dependency ($a_t \rightarrow s_t$) ensures that the generated text strictly adheres to the strategic purpose defined by the illocutionary act.

Perlocutionary Level (Goal Achievement). The system's objective is to maximize a dual reward R_{total} reflecting two fundamental dimensions of interview effectiveness:

- **Information Utility (R_{info}):** The effectiveness of eliciting accurate clinical information.
- **Rapport Building (R_{rap}):** The effectiveness of maintaining patient engagement and trust.

The optimization objective is:

$$\max_{\pi_\theta} \mathbb{E}\left[\sum_t \left(R_{\text{info}}(s_t, u_{t+1}) + \lambda \cdot R_{\text{rap}}(s_t, u_{t+1})\right)\right] \tag{6}$$

B Dataset Details

This section supplements Section 3.3 with construction details and analysis.

B.1 Construction Pipeline

Patient Profile Generation. We do not generate profiles from scratch but extract and adapt them from real consultation transcripts in DAIC-WOZ (Gratch et al., 2014). We first extract comprehensive "Role Cards" capturing demographics, psychological states, and life events. To ensure clinical diversity, we systematically inject symptoms corresponding to 6 conditions (combinations of Depression severity and Anxiety presence). Finally, the model (acting as the patient) completes a PHQ-8 questionnaire (Kroenke et al., 2009) to establish ground truth scores. Figure 8 presents a detailed example of a simulated patient profile.

Dialogue Synthesis. The synthesis pipeline ensures naturalness and clinical meaning. Conditioned on the Role Card, the model generates an initial conversation dominated by `Ease` to establish baseline rapport. We then inject targeted diagnostic segments for PHQ-8 symptoms using open-ended questioning (1-3 turns per symptom). Heuristic rules correct action labels, restricting `Explore` and `Confirm` tags to clinical inquiries. To improve quality, we diversify empathetic responses using a real-world corpus and align the patient's linguistic style with random DAIC-WOZ transcripts.

B.2 Action Space Analysis

The action space dynamics in $\mathbb{S}^4\mathbb{DATA}$ exhibit clinically meaningful patterns (see Figure 3 in the main text). **Distribution**: Affective Regulation actions constitute the largest share at 48.8%, underscoring the critical role of rapport-building, with `Validate` (23.2%) and `Ease` (17.7%) as the most prominent. Information Seeking accounts for 44.5%, led by `Explore` (25.3%) and `Probe` (15.4%). Interview Management (6.7%) provides structural framing. **Temporal dynamics**: In the opening phase, `Initiate` and `Ease/Support` dominate for rapport establishment. The diagnostic phase shifts to `Probe` and `Explore`, interspersed with `Validate` for emotional safety. The closing phase returns to `Conclude` for consolidation. **Transition patterns**: Diagnostic loops (`Explore` $\rightarrow$ `Probe` $\rightarrow$ `Confirm`) form the backbone of information gathering. Empathetic pivots to `Validate` after distress disclosure interleave diagnostic work with emotional support. Repair pathways via `Clarify` enable misunderstanding correction before resuming inquiry. These patterns capture the

Simulated Patient Profile Example

Basic Information
- **Gender**: Female
- **Age**: 32
- **Location**: Beijing, China
- **Education**: Majored in Marine Biology but feels ineffective in academic research, feeling her efforts have no real impact; previously studied Graphic Design but was dissatisfied, traveled around China, then chose Marine Biology, yet still hasn't found a true sense of belonging.
- **Family Relations**: Only child; relationship with parents is superficial with plain daily communication; few conflicts, but feels a lack of deep care, finding parental companionship distant and hollow.
- **Occupation**: Fresh graduate with no clear direction. Interested in marine ecological management but sees limited prospects; feels helpless in job seeking; questions her own experience and ability, frequently engaging in self-denial and feeling incompetent for meaningful work.
- **Emotional State**: Long-term low mood, lack of motivation; difficulty rising in the morning and sleeping at night; anhedonia (unable to feel pleasure) even in previously liked activities like travel; constant feeling of suppression, rarely feeling relaxed.
- **Cognitive Mode & Worldview**: Previously believed herself to be smart and responsible, but this belief is fading; feels incapable of achieving goals; losing interest in the world; questions the meaning of life; dissatisfied with injustice but feels powerless to change anything.
- **Behavioral Control**: Difficulty controlling emotions, especially tears in solitude or anxiety; occasional self-harm to temporarily relieve inner pain; severe dissatisfaction with sleep quality, sometimes abusing medication to force sleep.
- **Social Function**: Reduced social activities; estranged from old friends; feels like a burden to her supportive boyfriend; believes others cannot understand her state, refusing to seek help and often isolating herself.
- **Psychological State**: Reports life feels meaningless; feels deep powerlessness and fear regarding work and future; admits psychotherapy was ineffective and isolating; sits alone staring blankly when low, passive and exhausted.
- **Others**: Mentions interest in travel/adventure but admits they no longer attract her; tried meditation and exercise without success; often gives up action, trapped in a loop of negative thinking and emotion.

Scale Results (PHQ-8)
- Little interest or pleasure in doing things: **2**
- Feeling down, depressed, or hopeless: **3**
- Trouble falling or staying asleep, or sleeping too much: **2**
- Feeling tired or having little energy: **3**
- Poor appetite or overeating: **2**
- Feeling bad about yourself (or that you are a failure): **3**
- Trouble concentrating on things: **2**
- Moving or speaking so slowly that other people could have noticed (or being so fidgety/restless): **2**

Figure 8: A detailed example of a simulated patient profile used in $\mathbb{S}^4\mathbb{DATA}$.

strategic grammar of clinical dialogue—the implicit rules governing how interviewers chain actions.

C Implementation Details and Baselines

This section supplements Section 4.1 with full implementation details and baseline descriptions.

C.1 Implementation Details

Data Synthesis. We utilized GPT-4o with a temperature of 1.0 for all data synthesis tasks. The process involved three stages:

1. **Patient Profile Extraction**: We extracted basic information from DAIC-WOZ transcripts and rewrote them according to specified depression severities (None, Mild, Moderate, Moderately Severe, Severe). The model then filled out the PHQ-8 questionnaire based on these profiles to generate ground truth scores.

2. **Dialogue Generation**: An initial dialogue was generated based on the patient profile and clinical guidelines. We then scanned the PHQ-8 results to identify missing symptom discussions. For each missing symptom, we identified the optimal insertion point, generated relevant turns, and rewrote the context to ensure coherence.

3. **Refinement**: We appended 5 chit-chat turns at the end of each session to simulate natural closure.

The final dataset contains 938 sessions with a total of 26,607 turns.

Supervised Fine-Tuning (SFT). We fine-tuned the Qwen3-8B model using the LLaMA-Factory framework on 8 Nvidia H20 GPUs with DeepSpeed acceleration. The training configuration was as follows:

- Batch size per GPU: 4
- Gradient accumulation steps: 4
- Initial learning rate: 5e-5 (Cosine decay)

- Precision: bfloat16 mixed precision
- Epochs: 3

Reinforcement Learning (RL). We employed Group Relative Policy Optimization (GRPO) to fine-tune the SFT model using LoRA on a single Nvidia H20 GPU. The reward model used the same architecture as the SFT model but was trained on a subset of 1,083 annotated turns for 5 epochs. The GRPO hyperparameters were:

- LoRA: rank=32, alpha=16
- Batch size: 2
- Generations per prompt: 2 (for preference optimization)
- Gradient accumulation steps: 4
- Learning rate: 5.0e-6
- Epochs: 1

Evaluation Setup. For evaluation, the counselor model generated responses with temperature 0.7 and top-p 0.9 (max tokens=512). The simulated patient (GPT-4o) operated at temperature 1.0. The evaluator (GPT-4o) used temperature 0 and a fixed random seed (998244353) to ensure reproducibility. The evaluation process involved generating a full session first, followed by a post-hoc analysis by the evaluator model.

C.2 Baseline Models

Qwen3-8B and Qwen3-32B. Qwen3-8B and Qwen3-32B (Yang et al., 2024) are open-source large language models from the Qwen2 series developed by Alibaba Cloud. These models are pretrained on massive multilingual corpora covering over 30 languages and demonstrate strong performance across various natural language understanding and generation tasks. We evaluate both versions to assess the impact of model scale on therapeutic dialogue quality.

GPT-4o. GPT-4o (OpenAI, 2024) is a multimodal model from OpenAI, building upon the GPT-4 (OpenAI, 2023b) architecture with enhanced efficiency and multilingual capabilities. It serves as a strong upper baseline for our task.

GPT-OSS-120B. GPT-OSS-120B refers to a large-scale open-source GPT-style model with approximately 120 billion parameters. This baseline evaluates whether raw model scale alone can achieve competitive performance on specialized therapeutic dialogue tasks.

DeepSeek-V3. DeepSeek-V3 (DeepSeek-AI, 2024) employs mixture-of-experts (MoE) architecture to achieve efficient scaling, activating only a subset of parameters for each input. This baseline evaluates whether general-purpose reasoning capabilities transfer to therapeutic dialogue scenarios.

DeepSeek-R1. DeepSeek-R1 (DeepSeek-AI, 2025) is a reasoning-focused model trained with reinforcement learning to enhance multi-step reasoning capabilities, providing a strong baseline for evaluating whether enhanced reasoning transfers to therapeutic dialogue.

Prompting Strategy. For all baseline models, we use carefully designed prompts specifying the role as a professional psychological counselor conducting PHQ-8 depression screening, with 2-3 high-quality demonstration dialogues covering different patient profiles and symptom presentations.

D Evaluation Framework: Detailed Metrics

This appendix provides detailed specifications for all evaluation metrics introduced in Section 4.1. Table 5 gives a complete overview; the subsections below detail the scoring methodology for each metric, organized by SAT layer.

D.1 Summary of Metrics

Dimension	Metric	Abbr.	Scale
Locutionary	Information Density	Dens.	0-10
	Linguistic Accuracy	Acc.	0-10
Illocutionary	Skill Diversity	Div.	1-3
	Skill Balance	Bal.	1-3
	Skill Precision	Prec.	0-3
Perlocutionary	Cognitive Ease	Cog.	0-10
	Relationship Building	Rel.	0-10
	Emotional State	Emo.	0-10
Diagnostic	Symptom Recall	Symp.R.	0-100%
	Binary Class. F1	Bin.F1	0-100%
	5-class F1	5cls.F1	0-100%

Table 5: Summary of evaluation metrics with abbreviations and scales.

D.2 Dimension 1: Locutionary Quality

D.2.1 Information Density (Dens.)

Objective. Evaluate how effectively each counselor utterance conveys information relative to its length, mimicking human counselor communication patterns.

Scoring Methodology.

1. For each counselor utterance, we use LLM-based scoring to assess information density.
2. Each utterance receives a score from 0-10 based on:
   - Amount of meaningful information conveyed
   - Conciseness and efficiency
   - Absence of redundancy or filler content
3. The final score is the mean across all utterances in the dialogue.

LLM Prompt Template. We provide the LLM with the counselor utterance and ask: *"Rate the information density of this counselor statement on a scale of 0-10, where 10 indicates maximum information conveyed with minimal words, and 0 indicates verbose or uninformative content."*

D.2.2 Linguistic Accuracy (Acc.)

Objective. Assess whether the counselor accurately describes PHQ-8 symptoms using flexible, natural language rather than verbatim repetition of the original criteria.

Scoring Methodology.

1. For each of the 8 PHQ symptoms, locate the corresponding inquiry in the dialogue.
2. Use LLM-based evaluation with the following three-tier scoring scheme:
   - **High (8-10)**: Accurately conveys the PHQ-8 symptom with flexible, natural language that paraphrases rather than copies the criterion text. The description maintains semantic accuracy while using varied vocabulary and sentence structure.
   - **Medium (5-7)**: Relatively accurate but uses rigid language or contains exact textual matches with the criterion (identical word sequences, not just shared vocabulary). The meaning is preserved but the language is less natural.
   - **Low (1-4)**: Significant deviation from the PHQ-8 symptom description. The inquiry may be tangentially related but fails to capture the core symptom being assessed.
   - **Zero (0)**: No corresponding dialogue found for this symptom.
3. Additional Principles:
   - If the utterance contains text identical to the PHQ-8 criterion (not just shared words), the score cannot exceed 7.
   - Shared vocabulary is acceptable; verbatim phrase repetition is penalized.
4. The final score is the mean across all 8 PHQ symptoms.

PHQ-8 Symptoms. The 8 symptoms assessed are: (1) Little interest or pleasure, (2) Feeling down/depressed/hopeless, (3) Sleep problems, (4) Feeling tired, (5) Appetite problems, (6) Feeling bad about self, (7) Concentration problems, (8) Moving/speaking slowly or being fidgety.

D.3 Dimension 2: Illocutionary Competence

D.3.1 Skill Diversity (Div.)

Objective. Evaluate whether the counselor employs a variety of skills rather than relying on repetitive strategies.

Skill Repertoire. We define 9 skill types: `Initiate`, `Ease`, `Explore`, `Confirm`, `Probe`, `Support`, `Validate`, `Clarify`, `Conclude`.

Scoring Methodology.

1. For each counselor utterance, identify which skills are employed.
2. Count the total usage of each skill type throughout the dialogue.
3. Calculate the number of distinct skill types used.
4. Assign rating based on threshold:
   - **Good (3 points)**: ≥ 6 skill types used
   - **Okay (2 points)**: ≥ 4 skill types used
   - **Bad (1 point)**: < 4 skill types used

D.3.2 Skill Balance (Bal.)

Objective. Evaluate whether the counselor uses skills in appropriate proportions that align with clinical best practices.

Scoring Methodology.

1. Building on skill diversity analysis, calculate the usage proportion of each skill type.
2. Use LLM-based evaluation to assess whether the distribution aligns with clinical guidelines.
3. Provide the LLM with:
   - Skill distribution statistics (percentage for each skill)
   - Full dialogue transcript
4. Clinical best practice guidelines:
   - `Explore` questions should dominate (30-50%)
   - `Probe` should be moderate and used when patients exhibit symptoms (not excessive)
   - `Validate` should be appropriately responsive to emotions (30-40%)
   - `Confirm` (closed questions) should be used sparingly for key confirmations (10-20%)
   - `Support` should be interspersed throughout
5. Rating scheme: Good (3 points) / Okay (2 points) / Bad (1 point).

D.3.3 Skill Precision (Prec.)

Objective. Evaluate whether the counselor uses skills at appropriate moments based on patient responses.

Scoring Methodology.

1. For each counselor utterance, extract:
   - The immediately preceding patient utterance
   - The current counselor utterance
   - The skill(s) employed
2. Use LLM-based evaluation to assess contextual appropriateness:
   - Is the chosen skill appropriate given the patient's response?
   - Does it advance the therapeutic goal?
   - Is the timing contextually suitable?
3. Each utterance receives a binary rating: Good (appropriate) or Bad (inappropriate).
4. The final score is computed as: Score $= 3 \times \frac{\text{\# Good}}{\text{\# Total}}$
5. We also report macro-F1 across skill categories to assess precision-recall balance.

D.4 Dimension 3: Perlocutionary Effect

D.4.1 Cognitive Ease (Cog.)

Objective. Assess how easily patients can process the interviewer's utterances.

Factors Evaluated.

- **Utterance Length**: Are counselor statements appropriately concise?
- **Complexity**: Is the language complexity suitable for the patient?
- **Question Pacing**: Does the counselor allow sufficient processing time between questions?
- **Language Abstractness**: Does the counselor use concrete, accessible language?
- **Vocabulary Difficulty**: Are the words used appropriate for lay understanding?

Scoring Methodology.

1. For each counselor utterance, use LLM-based evaluation to assess cognitive ease.
2. Each utterance receives a score from 0-10:
   - **High (8-10)**: High cognitive ease; language is clear, concise, and accessible.
   - **Medium (4-7)**: Moderate ease; some complexity or verbosity.
   - **Low (0-3)**: Low ease; overly complex, lengthy, or abstract.
3. The final score is the mean across all utterances.

D.4.2 Relationship Building (Rel.)

Objective. Evaluate how well the counselor builds therapeutic rapport and stabilizes patient emotions.

Key Indicators.

- Timely empathy or encouragement when patient expresses distress
- Natural expression of understanding
- Emotional connection and warmth
- Appropriate use of validation
- Respectful boundary maintenance

Scoring Methodology.

1. For each counselor utterance, assess its contribution to relationship building.
2. Each utterance receives a score from 0-10:
   - **High (8-10)**: Strong rapport-building; empathetic, warm, and well-timed.
   - **Medium (4-7)**: Adequate but could be more empathetic or better timed.
   - **Low (0-3)**: Weak rapport-building; cold, dismissive, or poorly timed.
3. The final score is the mean across all utterances.

D.4.3 Emotional State Tracking (Emo.)

Objective. Simulate and track how the counselor's language influences the patient's emotional trajectory.

Scoring Methodology.

1. Initialize patient emotional state with a score from 0-100:
   - 0-30: Highly distressed
   - 31-60: Moderately distressed
   - 61-80: Mildly distressed
   - 81-100: Calm/stable
2. For each counselor utterance:
   (a) Use an LLM to simulate the patient's emotional response.
   (b) Update the emotion score based on whether the utterance:
      - Provides comfort and validation (increase score)
      - Shows understanding and empathy (increase score)
      - Is dismissive or invalidating (decrease score)
      - Rushes or pressures the patient (decrease score)
3. Track the emotional trajectory throughout the dialogue.
4. Report the final emotion score as the primary metric.
5. Secondary metrics include: emotional stability (variance), recovery rate (positive slope), and lowest point reached.

LLM Simulation. We provide the LLM with:

- Patient's current emotional state (score)
- Counselor's utterance
- Recent dialogue context (last 2-3 turns)

The LLM evaluates the emotional impact and returns an updated emotion score with justification.

D.5 Dimension 4: Diagnostic Outcome Metrics

D.5.1 Symptom Recall Rate (Symp.R.)

Objective. Measure the percentage of PHQ-8 symptom items for which the model successfully elicited sufficient information to form a clinical judgment.

Methodology. The elicited information is compared against the ground-truth patient profile. An item is considered "recalled" if the conversation contains explicit confirmation or denial of the symptom matching the profile.

D.5.2 Binary Classification F1 (Bin.F1)

Objective. Evaluate the model's accuracy in identifying the presence of depression.

Methodology. Based on the elicited symptoms, the total PHQ-8 score is calculated. A score ≥ 10 indicates depression. We calculate the Macro-F1 score for the binary classification task (Depressed vs. Non-Depressed) against the ground-truth labels.

D.5.3 Five-class Classification F1 (5cls.F1)

Objective. Evaluate the fine-grained diagnostic accuracy across depression severity levels.

Methodology. The total PHQ-8 score is mapped to five severity categories: None (0-4), Mild (5-9), Moderate (10-14), Moderately Severe (15-19), and Severe (20-24). We report the Macro-F1 score for this multi-class classification task.

E Human Evaluation Protocol

This section provides complete details of our human evaluation procedure, including evaluator recruitment, compensation, and the full instructions provided to participants.

E.1 Evaluator Recruitment and Compensation

Three licensed clinical psychologists were recruited through professional networks at collaborating academic institutions. All evaluators held valid clinical licenses and had at least 3 years of experience in mental health assessment.

Compensation. Evaluators were compensated at a rate of 150 CNY per hour (approximately 21 USD), which exceeds the average hourly wage for professional psychological consultation services in the local region (typically 100-120 CNY per hour for similar expert annotation tasks). The total evaluation task required approximately 8-10 hours per evaluator, and all evaluators confirmed that the compensation was fair and adequate for the work involved.

E.2 Evaluation Instructions

The following instructions were provided to all human evaluators in their native language (Chinese). Evaluators were asked to rate 50 psychiatric interview dialogues (presented in randomized order with model identities hidden) on a 1-5 Likert scale. Table 6 presents the complete scoring rubric.

Dimension	Scoring Criteria (1-5 Scale)
SE	*Symptom Elicitation: How effectively does the counselor gather PHQ-8 symptom information?* 5: Covers all 8 symptoms with appropriate depth 3: Covers some symptoms but misses important ones 1: Fails to elicit meaningful symptom information
EA	*Empathic Authenticity: How genuine and well-timed are the empathic responses?* 5: Genuine, well-timed, therapeutically effective 3: Adequate but sometimes feels mechanical 1: Absent, inappropriate, or counterproductive
FD	*Facilitation of Disclosure: How well does the counselor create a safe space for self-disclosure?* 5: Excellent rapport; patient appears comfortable 3: Adequate facilitation but some awkward moments 1: Fails to create safe space; feels interrogative
CA	*Cognitive Appropriateness: Is the language clear, concise, and clinically appropriate?* 5: Highly professional, clear, and concise 3: Adequate but sometimes unclear or verbose 1: Inappropriate language that hinders interview

Table 6: Human evaluation scoring rubric for subjective dimensions (SE, EA, FD, CA). Evaluators were instructed to focus only on counselor utterances and consider what would be appropriate in a real clinical screening context.

E.3 Diagnostic Success (DS) Protocol

In addition to the subjective evaluation, we designed an objective role-play diagnostic task to assess real-world diagnostic capability. Each of the three experts simulated patients with depressive tendencies and interacted with each system for 10 sessions (30 sessions per model).

Role-Play Design. Before each session, experts were assigned a patient profile with predetermined PHQ-8 item scores (ground truth), covering diverse severity levels: None (0-4), Mild (5-9), Moderate (10-14), Moderately Severe (15-19), and Severe (20-24). Experts role-played according to their assigned profiles while interacting naturally with the system.

Success Criteria. A diagnostic session was considered successful if the system: (1) successfully elicited information for at least 6 out of 8 PHQ-8 symptoms (symptom recall $\geq 75\%$), and (2) correctly classified the patient's depression status (binary classification: depressed if PHQ-8 total $\geq$ 10).

Metric Calculation. The Diagnostic Success (DS) rate is computed as:

$$\text{DS} = \frac{\text{\# Successful Sessions}}{\text{\# Total Sessions}} \times 100\% \quad (7)$$

This objective metric complements the subjective ratings by directly measuring whether systems can complete clinically valid PHQ-8 assessments in realistic interaction scenarios.