ProAI: Proactive Multi-Agent Conversational AI with Structured Knowledge Base for Psychiatric Diagnosis

Anonymous ACL submission

Abstract

Most LLM-driven conversational AI systems operate *reactively*, responding to user prompts without guiding the interaction. However, many real-world applications—such as psychiatric diagnosis, consulting, and interviews—require AI to take a proactive role, asking the right questions and steering conversations toward specific objectives. Using mental health differential diagnosis as an application context, we introduce **ProAI**, a goal-oriented, proactive conversational AI framework. ProAI integrates structured knowledge-guided memory, multi-agent proactive reasoning, and a multi-faceted evaluation strategy, enabling LLMs to engage in clinician-style diagnostic reasoning rather than simple response generation. Through simulated patient interactions, user experience assessment, and professional clinical validation, we demonstrate that ProAI achieves up to 83.3% accuracy in mental disorder differential diagnosis while maintaining professional and empathetic interaction standards. These results highlight the potential for more reliable, adaptive, and goal-driven AI diagnostic assistants, advancing LLMs beyond reactive dialogue systems.

1 Introduction

The emergence of large language models (LLMs) has revolutionized conversational AI, enabling increasingly sophisticated human-machine interactions. However, most current LLM applications operate within a *reactive paradigm*—generating responses to user prompts without actively guiding the conversation (Adamopoulou and Moussiades, 2020). While substantial research has focused on optimizing these reactive capabilities through prompt engineering, fine-tuning, and alignment techniques (Liu et al., 2023a; Sahoo et al., 2024), many real-world applications, including education tutoring (Piro et al., 2024), mental health diagonisis(Tu et al., 2024), and job interviewing(Cheong et al., 2024), require AI systems capable of taking initiative and steering conversations toward specific objectives (Deng et al., 2023a; Lu et al., 2025).

Mental health differential diagnosis (DDx) is a canonical setting for proactive AI systems. Facing over 150 possible mental disorders, clinicians must distinguish one mental disorder from others that present with similar symptoms (First, 2013). Accurate diagnosis relies on structured 45–90 minute interviews, where clinicians strategically assess symptom patterns, timing, severity, and life context, akin to navigating a complex decision tree (Carlat, 2005; Nordgaard et al., 2013). This process demands real-time reasoning and adaptive questioning to systematically narrow down potential diagnoses, requiring at least 10 years of extensive training (Das, 2023; for Addiction and Health, 2025). As a result, there is a severe specialist shortage, limiting access to care (Thomas and HOLZER III, 2006; Butryn et al., 2017). This gap highlights the urgent need for AI-driven diagnostic support.

However, current approaches to AI-assisted diagnosis face several critical challenges. First, existing conversational AI systems lack the proactive reasoning capabilities needed to dynamically adjust questioning strategies and guide diagnostic conversations (Tu et al., 2024). Second, traditional approaches that frame diagnosis as multi-class classification struggle with the high dimensionality of possible disorders and limited training data (Chang et al., 2021; Schulte-Rüther et al., 2023).

To address these challenges, we present **ProAI**, a proactive conversational AI framework specifically designed for mental health differential diagnosis. Our approach introduces several key innovations: (1) a multi-agent system where specialized agents collaborate to facilitate proactive diagnosis, including professional medical decision making and active question generating agents. (2) a structured knowledge-guided memory architecture that combines long-term domain knowledge with short-term contextual dialogue information to guide diagnostic reasoning; and (3) a comprehensive evaluation strategy encompassing both diagnosis accuracy and patient experience, realized through simulated patient interactions, user experience assessment, and professional clinical validation.

Our experimental results highlight the significant diagnostic capabilities of the ProAI framework, achieving over 80% diagnostic accuracy along with high ratings for both user-friendliness and medical proficiency. Furthermore, the integration of structured knowledge-guided memory demonstrates a substantial improvement, significantly outpacing existing knowledge-enhanced methods like RAG. These results encompass three disorders and include over 60 potential differential diagnosis (DDx) outcomes.

To the best of our knowledge, this work represents the first comprehensive framework for proactive conversational AI in mental health diagnosis, establishing a new paradigm for goal-directed diagnostic systems that combine structured knowledge, dynamic reasoning, and empathetic interaction. In sum, our contribution lies on:

- **Multi-Agent Proactive Reasoning**: We develop a multi-agent system where specialized agents collaborate to facilitate proactive diagnosis, including a question generation agent, a context understanding agent, and an action transition agent that ensures coherent and goal-directed conversation flow.

- **Structured Knowledge-Guided Memory**: We introduce a long-term memory to store structured domain knowledge and a short-term memory to collect contextual information, retrieve relevant diagnostic knowledge, and dynamically guide questioning. Comprehensive evaluation demonstrate over 100% improvement in average compared to existing knowledge enhancement paradigms.

- **Multi-Faceted Evaluation Strategy**: Our framework integrates both objective and subjective evaluation metrics, assessing diagnostic accuracy, conversation efficiency, perceived helpfulness, and medical proficiency. To address the high cost of human evaluation in medical AI, we employ a three-tier validation approach, incorporating patient interaction simulation, user-level assessment, and clinician-level validation.

2 Related Work on AI Solutions for Differential Diagnosis

Traditional AI approaches to differential diagnosis have treated mental health assessment primarily as a classification task (Ahsan et al., 2022; Zhang et al., 2024). While these approaches enable efficient initial screening, they lack the progressive reasoning capabilities essential for psychiatric evaluation, where symptoms emerge gradually and require careful contextual interpretation (Kanjee et al., 2023). The challenges are compounded by data scarcity—with over 150 recognized disorders and strict privacy constraints, obtaining sufficient training data remains impractical (Bakator and Radosav, 2018; Yan et al., 2022).

Recent advances in LLMs have sparked a shift toward more interactive, dialogue-based diagnostic approaches (Liao et al., 2023), but significant challenges persist in both knowledge integration and reasoning capabilities. While techniques like fine-tuning, ICL, and RAG have been proposed (Dong et al., 2024), they struggle to incorporate the hierarchical diagnostic decision trees used by clinicians (Lewis et al., 2020). Recent work has explored memory structures for psychological consultation (Lan et al., 2024) and LLM-based depression assessments (Lorenzoni et al., 2024), along with general advances in LLM reasoning through techniques like Chain-of-Thought prompting (Wei et al., 2022) and ReAct (Yao et al., 2023). However, these approaches remain fundamentally reactive, often failing to ask critical follow-up questions or systematically rule out differential diagnoses (Liu et al., 2023b), making it unsuitable for DDx.

3 Methodology: ProAI Framework for Differential Diagnosis

In this section, we present the details of each component in our ProAI framework.

3.1 Multi-Agent Proactive Reasoning Workflow

In a typical mental disorder clinical interview, clinicians perform two key actions: formulating questions to comprehensively gather patient information and assessing symptoms to make a diagnostic evaluation. This motivates us to propose the multi-agent proactive reasoning workflow that mirrors this process by introducing two agents: a decision-maker agent and a question-generator agent. To enable proactive conversations in real-world scenar-

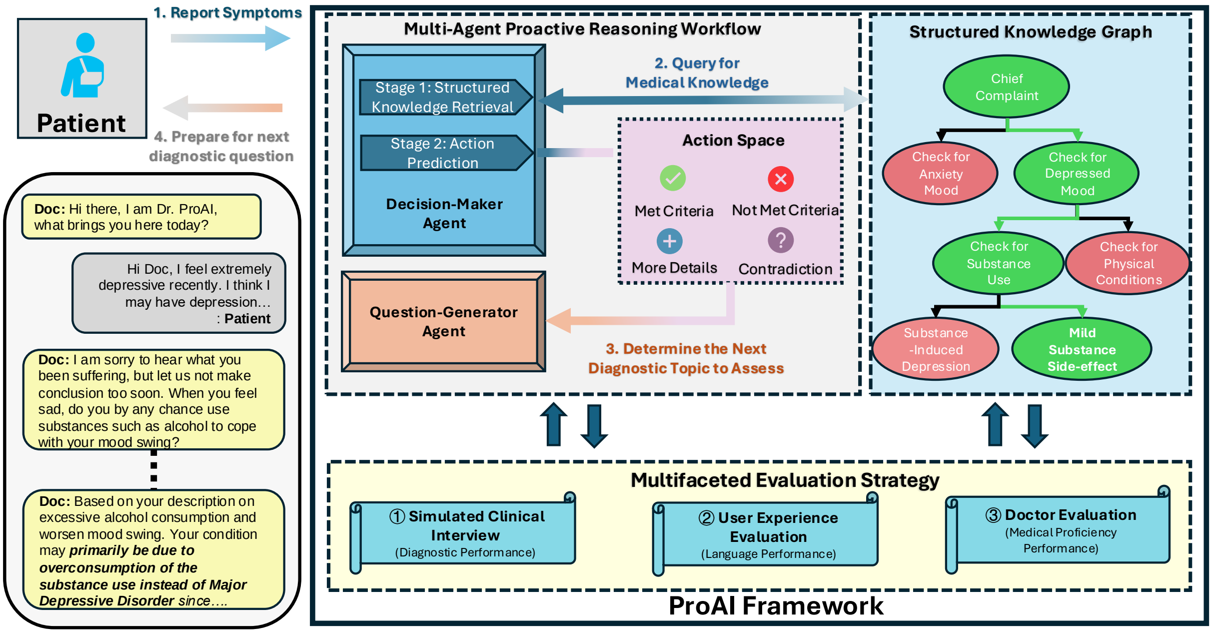


Figure 1: The ProAI Framework for Proactive Clinical Reasoning. The ProAI framework consists of three key components: a multi-agent proactive reasoning workflow, a structured knowledge graph, and a multifaceted evaluation strategy. The reasoning workflow determines the actions to take based on the dialogue history and generates corresponding questions based on those actions. Throughout this process, the structured knowledge graph is leveraged to guide decision-making. Additionally, the multifaceted evaluation strategy is employed to dynamically assess each round of conversation, ensuring ongoing refinement and effectiveness.

ios, these agents work in concert through a ReAct-inspired (Yao et al., 2023) workflow where each diagnostic iteration involves reasoning over the patient's current state and determining the next action (i.e. generating clarifying questions or transitioning to a new diagnostic subtopic), then generating the corresponding question based on the action, as illustrated in Fig. 1.

3.1.1 Decision-Maker Agent

The Decision-Maker follows a two-stage decision process for each conversational turn, paralleling the ReAct paradigm of "*reason + act*" (Yao et al., 2023).

Stage 1: Knowledge Retrieval Given the current node v_c and the dialogue history D, the agent retrieves relevant medical knowledge:

$$K_t = R(v_c, D, G), \tag{1}$$

where $R(\cdot)$ is a retrieval function that considers both the node's local context and its connected nodes within the Decision Graph. By focusing on directed links, the agent remains aligned with the correct path, preventing unnecessary detours.

Stage 2: Action Prediction Using the retrieved knowledge K_t and the patient's latest response r_t, the agent decides the most appropriate action:

$$a_t = P(K_t, r_t, D), \tag{2}$$

where $a_t \in \{$*met_criteria*, *not_met_criteria*, *ask_more_questions*, *contradiction*$\}$. This parallels the "act" step in ReAct, wherein the agent chooses how to proceed based on reasoning outcomes. The transition function T then updates the current node:

$$v_{t+1} = T(v_t, a_t, G), \tag{3}$$

moving to the next node, re-checking the current node, or backtracking if contradictions arise. This loop continues until the agent reaches a leaf node containing a final diagnostic conclusion.

3.1.2 Question-Generator Agent

After determining the action, the Question-Generator agent formulates the next diagnostic question:

$$q_{t+1} = Q(v_{t+1}, a_t, D). \tag{4}$$

In practice, the question-generation function $Q(\cdot)$ balances three factors:

$$Q = f(c_{t+1}, h_t, s_t), \tag{5}$$

where c_{t+1} is the next diagnostic criterion to be evaluated, h_t is the conversation history, and s_t is the semantic context. By integrating these elements, the Question-Generator agent ensures that each question remains both clinically relevant and conversationally coherent, driving the dialogue toward a precise and efficient diagnosis.

3.2 Structured Knowledge-Guided Memory

To empower LLMs as decision-maker agents or question-generator agents, a critical requirement is the integration of long-term memory (LTM), for effective decision-making and question formulation depend on a deep understanding of professional knowledge and careful consideration of the dialogue history.

3.2.1 Prompting Paradigms for Knowledge Enhancement

Based on how background information and dialogue history are utilized, we categorize existing prompting techniques into three distinct knowledge enhancement paradigms: *Knowledge Free Prompting (KFP), Textual Knowledge Enhanced Prompting (TKEP), and Structured Knowledge Enhanced Prompting (SKEP)*. We present the summary of each paradigms in Table 1 and the detailed implementation algorithm in Appendix D.

Knowledge Free Prompting: KFP represents the straightforward approaches, where the LLM relies solely on its pre-trained knowledge. Formally,

$$p(r \mid B, D), \tag{6}$$

where B is background context and D is dialogue history.

Representative KEP methods like direct prompting or zero-shot CoT may perform well in open-domain tasks, but maybe inadequate for knowledge intensive tasks like DDx, where the specific professional knowledge is required to ensure precise medical diagnose.

Textual Knowledge Enhanced Prompting: TKEP represents a category of methods that integrate the external knowledge K into prompts. Formally,

$$p(r \mid B, D, K), \tag{7}$$

External knowledge K can be sourced via mechanisms like ICL (Brown et al., 2020) (embedding task-specific examples) or RAG (Lewis et al., 2020) (retrieving documents from an external knowledge base through query/dialogue context). Such methods enhance the performance on knowledge-intensive tasks by expanding the LLM's domain expertise. However, they may lack the hierarchical structure necessary to establish relationships between different medical concepts, making them insufficient for diagnostic tasks that require clear and systematic reasoning.

Structured Knowledge Enhanced Prompting: SKEP represents a prompting approach that goes beyond simple knowledge integration by encoding the structured relationships between relevant knowledge, enabling more systematic and context-aware reasoning. Formally, we denote this structure as vectors of knowledge $\vec{K}$:

$$p(r \mid B, D, \vec{K}), \tag{8}$$

where $\vec{K}$ represents not just raw facts but a *knowledge graph* (KG) with directional edges that define a task-oriented agenda (Edge et al., 2024). This structured representation ensures a systematic and clinically grounded diagnostic process.

3.2.2 Structured Knowledge-Guided Memory for Medical Diagnose

In medical diagnostic tasks, we construct the structured knowledge $\vec{K}$ by gathering interconnected diagnostic criteria and clinical guidelines, incorporating directed links that capture logical or causal relationships between different criteria (e.g., "if `substance use` is reported, investigate `substance-induced disorders` before concluding major depression"). Specifically, we encodes medical knowledge in a binary-tree structure. Formally,

$$G = (V, E, T), \tag{9}$$

where V is a set of nodes, E is a set of directed edges linking these nodes, and T is the set of possible transitions, $\{$*left*, *right*, *stay*, *back*$\}$. Each node $v_i \in V$ comprises:

$$v_i = \{c_i, q_i, d_i\}, \tag{10}$$

where c_i is a diagnostic criterion, q_i represents query templates relevant to that criterion, and d_i stores local decision rules for handling transitions. Essentially, this graph serves as a directed knowledge structure, guiding the diagnostic flow step by step (see Appendix B for an example). The directionality ensures that once a criterion is assessed, the agent can proceed to the appropriate

Table 1: Summary of three knowledge-enhanced prompting paradigms.

Paradigm	Description	Examples
Knowledge-Free Prompting (KFP)	A prompting approach that relies solely on the model's pretrained knowledge without incorporating external information or domain-specific knowledge bases.	Direct Prompting, Zero-shot CoT
Textual Knowledge-Enhanced Prompting (TKEP)	A prompting method that augments instructions with relevant external or domain-specific knowledge, presented in an unstructured format, to improve response accuracy and relevance.	ICL, RAG
Structured Knowledge-Enhanced Prompting (SKEP)	A sophisticated prompting approach that incorporates domain knowledge organized in structured formats (e.g., knowledge graphs, hierarchical relationships, decision trees) to enable systematic reasoning and maintain logical dependencies.	Knowledge Graph RAG

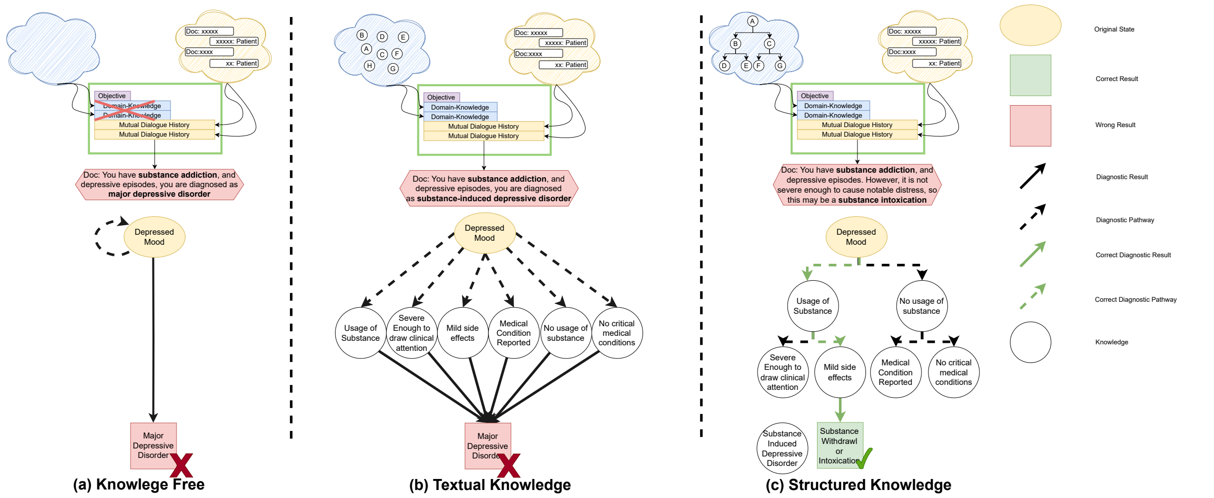


Figure 2: Comparison of three knowledge enhancement methods in medical diagnosis. (a) The knowledge-free approach relies solely on pretrained knowledge. (b) The textual knowledge approach incorporates domain information but lacks structured guidance. (c) The structured knowledge approach enables systematic traversal of diagnostic criteria.

subsequent node, systematically ruling in or ruling out conditions.

By integrating a directional KG into the prompt, SKEP ensures that the LLM follows a well-defined flow of inquiry, systematically traversing the relevant decision pathways to reach the appropriate diagnostic outcome. This directed structure emulates the approach of a clinician following a standard diagnostic tree—such as the one recommended in official clinical guidelines—thereby reducing the risk of missed criteria or irrelevant questioning.

3.3 Multifaceted Evaluation Strategy

Traditional evaluation strategies often prioritize diagnostic accuracy (or AUC) (Xue et al., 2024; Demetriou et al., 2020), while neglecting patient experience (Robertson et al., 2023) and the rigor of medical reasoning behind the diagnosis (Antoniou et al., 2022; Kerz et al., 2023). This narrow focus limits our understanding of the broader impact of AI-driven diagnostic systems. To address this, we propose a multifaceted evaluation strategy that integrates both objective and subjective metrics,

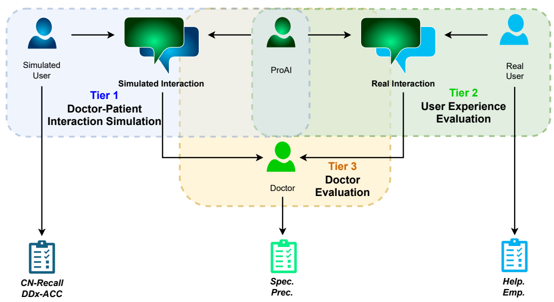


Figure 3: Three-tier Evaluation Framework. Tier 1 uses AI patient simulation to assess diagnostic accuracy (CN-Recall, DDx-ACC). Tier 2 involves human patient actors to evaluate user experience (Help., Emp.). Tier 3 engages medical professionals to assess clinical validity (Spec., Prec.).

providing a holistic assessment of AI diagnostic performance. See Fig. 3. Further details on the evaluation methodology can be found in Appendix E.

Doctor-Patient Interaction simulation: The goal of this evaluation is to measure ProAI's diagnostic performance after multi-turn diagnostic conversations. Traditional methods often rely on one-shot evaluation on static datasets such as medical case databases and clinical trials, which fail to capture the dynamic, adaptive nature of real diagnostic conversations (Sommers-Flanagan, 2016; Tu et al., 2024). While real patient interactions with ProAI would be ideal, ethical concerns, privacy constraints, and logistical challenges make this approach impractical (MacKinnon et al., 2015).

Inspired by Wu et al. (2023, 2024), we simulate patient interactions using an LLM agent that represents a patient with a specific mental disorder, exhibiting clinically informed symptomatology and behavior (see Appendix C for setup). Without prior knowledge of the disorder, ProAI engages in multi-round conversations to diagnose the patient. Its final diagnosis is compared to the ground-truth condition assigned to the patient agent via two metrics:

1) Critical Node Recall (CN-Recall), which measures the system's thoroughness in assessing essential diagnostic criteria:

$$CN\_Recall = \frac{N_{pred} \cap N_{critical}}{N} \tag{11}$$

where N_{pred} represents the criteria nodes assessed by the agent and $N_{critical}$ denotes the ground truth critical nodes.

2) Differential Diagnosis Accuracy (DDx-ACC), which evaluates the system's ability to reach correct diagnostic conclusions while properly ruling out alternative conditions.

$$DDx\_ACC = \frac{D_{correct}}{D_{total}} \tag{12}$$

where $D_{correct}$ is the number of correct diagnoses, and D_{total} is the total test cases.

User Experience Evaluation: This evaluation ensures a positive patient experience, allowing patients to express their true feelings during clinical interviews, enhancing diagnostic accuracy while minimizing potential harm (Vermeeren et al., 2010). We assess two key dimensions critical for patient experience (Deng et al., 2023b; Tu et al., 2024): Helpfulness (Help.), which measures the effectiveness of the agent's medical consultation, and Empathy (Emp.), which evaluates its ability to demonstrate understanding and build rapport. A demo system (see Appendix E) was developed, and 10 users were assigned predefined patient roles. After up to 40 rounds of interaction, users rated their experience using a 5-point Likert scale adapted from (King and Hoppe, 2013).

Doctor Evaluation: This evaluation ensures that diagnostic decisions are based on rigorous medical reasoning. Following the literature (Tu et al., 2024), we assess two key metrics: Specialty (Spec.), which evaluates clinical quality, coherence, and adherence to professional guidelines, and Precision (Prec.), which measures the accuracy and specificity of differential diagnoses to minimize misdiagnosis and optimize treatment. To achieve this, 3 medical professionals with 8 years of experience on average reviewed conversation transcripts and completed a 5-point Likert scale assessment using rating scales adapted from Dacre et al. (2003), ensuring alignment with established medical standards.

4 Experiment and Results

4.1 Experimental Setup

We evaluate the effectiveness of ProAI by conduct DDx on three types of disorders: depression, bipolar and anxiety. We try five LLMs in the ProAI framework, including *"gpt-4o", "claude-3.5-sonnet", "mistral-large", "qwen2.5-72b", and "deepseek-r1-70b"* (OpenAI et al., 2024; Bai et al., 2023; Anthropic, 2023; Jiang et al., 2023; DeepSeek-AI et al., 2025), with a generative temperature of 0.6—a setting that has demonstrated superior instruction-following and diverse language capabilities. In knowlegde base construction, the structured knowledge graph is constructed through DSM-5 DDx decision trees which clinicians follow in real practice (First, 2013). In the doctor-patient interaction simulation, we gather 113 case of patients from Kangning Dataset with modification (Mao et al., 2023), and the simulated patient is based on the "mistral-large" model with a temperature of 0.2, chosen for its stability and fairness.

4.2 Results and Analysis

4.2.1 Overall Performance of ProAI

ProAI reaches comparative performance. As shown in Table 2, the highest overall diagnostic accuracy (DDx-ACC) across the three mental disorder types is 83.3%, 73.3%, and 80.0%. Furthermore, incorporating different LLMs further enhances diagnostic accuracy, achieving optimal performances of 87.5%, 86.7%, and 97.2%. This highlights the strong capability of the ProAI framework in conducting real-world mental health diagnoses.

Table 2: Performance of knowledge integration methods across various mental health diagnoses with GPT-4o.

Task	Memory Type	Prompt Type	CN-Recall	DDx-ACC	Help.	Emp.	Spec.	Prec.
Depression	Random Guess	N/A	N/A	0.040	N/A	N/A	N/A	N/A
	KFP	Direct Prompting	N/A	0.256	0.487	0.275	0.415	0.430
	TKEP	ICL	0.466	0.280	0.395	0.150	0.483	0.642
	TKEP	RAG	0.235	0.250	0.474	0.275	0.488	0.662
	SKEP	Graph-RAG	**0.983**	**0.833**	**0.671**	**0.650**	**0.673**	**0.679**
Bipolar	Random Guess	N/A	N/A	0.063	N/A	N/A	N/A	N/A
	KFP	Direct Prompting	N/A	0.293	0.606	0.500	0.453	0.591
	TKEP	ICL	0.721	0.427	0.500	0.425	0.469	0.590
	TKEP	RAG	0.739	0.400	0.592	0.538	0.412	0.484
	SKEP	Graph-RAG	**0.769**	**0.733**	**0.671**	**0.650**	**0.673**	**0.642**
Anxiety	Random Guess	N/A	N/A	0.038	N/A	N/A	N/A	N/A
	KFP	Direct Prompting	N/A	0.400	0.632	0.475	0.386	0.377
	TKEP	ICL	0.517	0.360	0.447	0.250	0.479	0.642
	TKEP	RAG	0.292	0.480	0.579	0.450	0.412	0.430
	SKEP	Graph-RAG	**0.942**	**0.800**	**0.763**	**0.775**	**0.673**	**0.697**

Table 3: Comparison of different LLMs on mental health diagnose efficiency. Each model was evaluated using the ProAI framework with SKEP memory and Graph-RAG prompting. Bold numbers indicate best performance.

Task	LLM Type	CN-Recall	DDx-ACC	Help.	Emp.	Spec.	Prec.
Depression	Deepseek-r1:70b	0.956	0.833	**0.842**	0.838	0.633	0.750
	Qwen2.5:72b	0.934	**0.875**	0.790	**0.850**	0.624	0.752
	GPT-4o	**0.983**	0.833	0.671	0.650	0.673	0.679
	Claude-3.5-sonnet	0.551	0.625	0.724	0.675	**0.824**	**0.805**
	Mistral-Large	0.939	**0.875**	0.579	0.613	0.683	0.624
Bipolar	Deepseek-r1:70b	0.839	0.667	**0.869**	0.838	0.633	0.732
	Qwen2.5:72b	**0.867**	0.733	0.803	0.800	0.633	**0.752**
	GPT-4o	0.769	0.733	0.671	0.650	0.673	0.642
	Claude-3.5-sonnet	0.836	0.600	0.856	**0.913**	**0.809**	0.734
	Mistral-Large	0.757	**0.867**	0.842	0.850	0.683	0.642
Anxiety	Deepseek-r1:70b	**0.951**	0.760	0.763	0.675	0.633	0.750
	Qwen2.5:72b	0.920	**0.972**	0.763	0.825	0.633	0.699
	GPT-4o	0.942	0.800	0.763	0.775	0.673	0.697
	Claude-3.5-sonnet	0.928	0.680	**0.816**	**0.825**	**0.809**	**0.805**
	Mistral-Large	0.899	0.840	0.737	0.725	0.683	0.642

4.2.2 Effectiveness of Knowledge Integration Methods

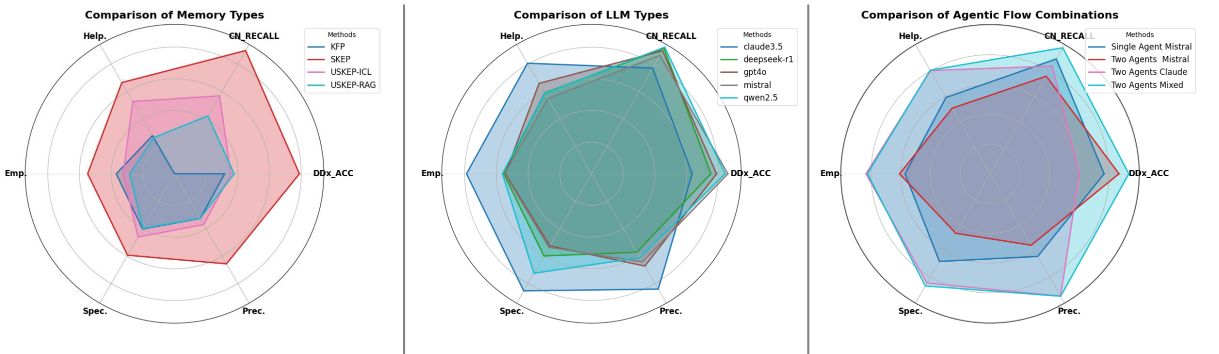
SKEP enhances diagnostic accuracy in mental health assessment. While basic knowledge integration offers incremental improvements over the random baseline, SKEP within ProAI achieves significant performance gains across all conditions. Notably, its higher CN-Recall scores in bipolar disorder cases suggest that SKEP's structured knowledge graph enhances symptom evaluation while ensuring adherence to clinical protocols (see Table 2).

SKEP also enhances the user experience. For depression diagnosis, SKEP excels in both objective metric (CN-Recall: 0.983, DDx-ACC: 0.833) and user experience (Help.: 0.671, Emp.: 0.650), surpassing TKEP variants (Help.: 0.395–0.474, Emp.: 0.150–0.275). Surprisingly, the integration of structured knowledge not only improves diagnostic accuracy but also makes the model sound more helpful and empathetic. This is crucial for fostering deeper connections with patients, enhancing trust, and improving the overall user experience.

Overall, as shown in Fig. 4a, SKEP emerges as the most well-rounded knowledge integration method, excelling across both objective and subjective metrics.

4.2.3 Effectiveness of Different Models

Performance trade-off exists for different LLMs. Benchmarking various LLMs integrated with SKEP (Fig. 4b, Table 3) reveals diverse performance profiles, highlighting distinct patterns of model specialization across depression, bipolar disorder, and anxiety diagnoses. A clear trade-off emerges between diagnostic performance and in-



a) Comparison of Memory Types b) Comparison of LLM Types c) Comparison of Agentic Flow Combinations

Figure 4: Performance Analysis of Various Settings in the ProAI Framework. We evaluate three key settings in the ProAI framework: (a) memory types, (b) LLM selection, and (c) agentic flow. For agentic flow, we compare three configurations: Single agent – A single LLM handles both decision-making and question generation. Two agents – A single LLM is used but separately designated for decision-making and question generation. Two mixed agents – Different LLMs are assigned for decision-making and question generation.

teraction quality. Claude 3.5 excels in specialty and empathy, making it particularly well-suited for patient interactions. Qwen2.5 and Mistral achieve high diagnostic accuracy while maintaining moderate user experience scores. DeepSeek-r1 strikes a balance between both aspects, achieving the highest helpfulness ratings (Help.: 0.842–0.869) while maintaining strong diagnostic accuracy, particularly in depression cases (DDx-ACC: 0.833). Meanwhile, GPT-4o demonstrates great Critical Node recall (CN-Recall: 0.983 for depression), underscoring its strength in symptom evaluation. However, its variability across other metrics suggests potential limitations.

Combining different LLMs mitigates this performance trade-off: A promising approach to addressing the trade-off between objective and subjective performance metrics in LLMs is a hybrid system. As shown in Fig. 4c, the "Two Agents Mixed" configuration—combining Mistral's diagnostic precision with Claude's communication strengths—achieves a better balance between accuracy and user experience. By separating diagnostic reasoning from patient communication, this architecture represents a significant advancement, underscoring the importance of thoughtful design and strategic model selection in optimizing clinical AI systems.

5 Conclusion

This paper introduces ProAI, a novel framework for proactive conversational diagnosis that addresses fundamental challenges in AI-assisted mental health assessment. By combining structured domain knowledge with dynamic conversation management, our approach achieves significant improvements in differential diagnostic accuracy while maintaining high standards of professional interaction. The framework's effectiveness is demonstrated through comprehensive evaluation across multiple dimensions, showing particular strength in critical diagnostic criteria coverage and systematic symptom assessment. Our findings suggest that structured knowledge integration and specialized agent roles represent promising directions for developing more reliable and empathetic AI-assisted diagnostic systems.

6 Limitations

While our ProAI framework demonstrates strong performance in mental health differential diagnosis, several limitations merit consideration. First, our evaluation focused on three common psychiatric disorders; future work should expand to a broader range of conditions and more variety of tasks (such as business consulting, job interview, education, etc) to validate generalizability. Second, while our simulated patient interactions provide valuable insights, extended clinical trials with real patients would further validate the system's practical utility. Additionally, the current implementation requires manual construction of knowledge graphs for each diagnostic domain, which could be automated through the future development of knowledge extraction techniques.

References

Eleni Adamopoulou and Lefteris Moussiades. 2020. An overview of chatbot technology. In *Artificial Intelligence Applications and Innovations*, pages 373–383, Cham. Springer International Publishing.

Md Manjurul Ahsan, Shahana Akter Luna, and Zahed Siddique. 2022. Machine-learning-based disease diagnosis: A comprehensive review. In *Healthcare*, volume 10, page 541. MDPI.

Anthropic. 2023. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic. Accessed: 2025-02-15.

Grigoris Antoniou, Emmanuel Papadakis, and George Baryannis. 2022. Mental health diagnosis: a case for explainable artificial intelligence. *International Journal on Artificial Intelligence Tools*, 31(03):2241003.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Mihalj Bakator and Dragica Radosav. 2018. Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction*, 2(3):47.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tracy Butryn, Leah Bryant, Christine Marchionni, and Farhad Sholevar. 2017. The shortage of psychiatrists and other mental health providers: causes, current state, and potential solutions. *International Journal of Academic Medicine*, 3(1):5–9.

Daniel J Carlat. 2005. *The psychiatric interview: A practical guide*. Lippincott Williams & Wilkins.

Miao Chang, Fay Y Womer, Xiaohong Gong, Xi Chen, Lili Tang, Ruiqi Feng, Shuai Dong, Jia Duan, Yifan Chen, Ran Zhang, et al. 2021. Identifying and validating subtypes within major psychiatric disorders based on frontal–posterior functional imbalance via deep learning. *Molecular psychiatry*, 26(7):2991–3002.

Inyoung Cheong, King Xia, KJ Kevin Feng, Quan Ze Chen, and Amy X Zhang. 2024. (a) i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2454–2469.

Jane Dacre, Mike Besser, Patricia White, et al. 2003. Mrcp (uk) part 2 clinical examination (paces): a review of the first four examination sessions (june 2001–july 2002). *Clinical Medicine*, 3(5):452–459.

Kusal K Das. 2023. Graduate medical education: variation of program and training duration. *Korean Journal of Medical Education*, 35(4):421.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Eleni A Demetriou, Shin H Park, Nicholas Ho, Karen L Pepper, Yun JC Song, Sharon L Naismith, Emma E Thomas, Ian B Hickie, and Adam J Guastella. 2020. Machine learning for differential diagnosis between clinical conditions with social difficulty: autism spectrum disorder, early psychosis, and social anxiety disorder. *Frontiers in psychiatry*, 11:545.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023a. A survey on proactive dialogue systems: Problems, methods, and prospects. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6583–6591. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

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

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Michael B First. 2013. *DSM-5-TR® Handbook of Differential Diagnosis*. American Psychiatric Pub.

Centre for Addiction and Mental Health. 2025. Clinical practicum training program in psychology. Accessed: 2025-02-13.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zahir Kanjee, Byron Crowe, and Adam Rodman. 2023. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *Jama*, 330(1):78–80.

Elma Kerz, Sourabh Zanwar, Yu Qiao, and Daniel Wiechmann. 2023. Toward explainable ai (xai) for mental health detection based on language behavior. *Frontiers in psychiatry*, 14:1219479.

Ann King and Ruth B Hoppe. 2013. “best practice” for patient-centered communication: a narrative review. *Journal of graduate medical education*, 5(3):385–393.

Kunyao Lan, Bingrui Jin, Zichen Zhu, Siyuan Chen, Shu Zhang, Kenny Q. Zhu, and Mengyue Wu. 2024. Depression diagnosis dialogue simulation: Self-improving psychiatrist with tertiary memory. *Preprint*, arXiv:2409.15084.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3452–3455.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Giuliano Lorenzoni, Pedro Elkind Velmovitsky, Paulo Alencar, and Donald Cowan. 2024. Gpt-4 on clinic depression assessment: An llm-based pilot study. In *2024 IEEE International Conference on Big Data (BigData)*, pages 5043–5049. IEEE.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, Weiwen Liu, Yasheng Wang, Zhiyuan Liu, Fangming Liu, and Maosong Sun. 2025. Proactive agent: Shifting LLM agents from reactive responses to active assistance. In *The Thirteenth International Conference on Learning Representations*.

Roger A MacKinnon, Robert Michels, and Peter J Buckley. 2015. *The psychiatric interview in clinical practice*. American Psychiatric Pub.

Kaining Mao, Deborah Baofeng Wang, Tiansheng Zheng, Rongqi Jiao, Yanhui Zhu, Bin Wu, Lei Qian, Wei Lyu, Jie Chen, and Minjie Ye. 2023. Analysis of automated clinical depression diagnosis in a chinese corpus. *IEEE Transactions on Biomedical Circuits and Systems*, 17(5):1135–1152.

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ludovica Piro, Tommaso Bianchi, Luca Alessandrelli, Andrea Chizzola, Daniela Casiraghi, Susanna Sancassani, and Nicola Gatti. 2024. Mylearningtalk: An llm-based intelligent tutoring system. In *International Conference on Web Engineering*, pages 428–431. Springer.

Christopher Robertson, Andrew Woods, Kelly Bergstrand, Jess Findley, Cayley Balser, and Marvin J Slepian. 2023. Diverse patients' attitudes towards artificial intelligence (ai) in diagnosis. *PLOS Digital Health*, 2(5):e0000237.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Martin Schulte-Rüther, Tomas Kulvicius, Sanna Stroth, Nicole Wolff, Veit Roessner, Peter B Marschik, Inge Kamp-Becker, and Luise Poustka. 2023. Using machine learning to improve diagnostic assessment of asd in the light of specific differential and co-occurring diagnoses. *Journal of Child Psychology and Psychiatry*, 64(1):16–26.

John Sommers-Flanagan. 2016. Clinical interview.

Christopher R Thomas and CHARLES E HOLZER III. 2006. The continuing shortage of child and adolescent psychiatrists. *Journal of the American Academy of Child & Adolescent Psychiatry*, 45(9):1023–1031.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. 2024. Towards conversational diagnostic ai. *Preprint*, arXiv:2401.05654.

Arnold POS Vermeeren, Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. 2010. User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries*, pages 521–530.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Yuqi Wu, Jie Chen, Kaining Mao, and Yanbo Zhang. 2023. Automatic post-traumatic stress disorder diagnosis via clinical transcripts: A novel text augmentation with large language models. In *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–5. IEEE.

Yuqi Wu, Kaining Mao, Yanbo Zhang, and Jie Chen. 2024. Callm: Enhancing clinical interview analysis through data augmentation with large language models. *IEEE Journal of Biomedical and Health Informatics*.

Chonghua Xue, Sahana S Kowshik, Diala Lteif, Shreyas Puducheri, Varuna H Jasodanand, Olivia T Zhou, Anika S Walia, Osman B Guney, J Diana Zhang, Serena T Pham, et al. 2024. Ai-based differential diagnosis of dementia etiologies on multimodal data. *Nature Medicine*, 30(10):2977–2989.

Wen-Jing Yan, Qian-Nan Ruan, and Ke Jiang. 2022. Challenges for artificial intelligence in recognizing mental disorders. *Diagnostics (Basel)*, 13(1):2.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Wei Zhang, Kaining Mao, and Jie Chen. 2024. A multimodal approach for detection and assessment of depression using text, audio and video. *Phenomics*, pages 1–16.

A Experiment Setup

A.1 LLM check-points

In this study, the LLM check-points we adopted up to the lastest version in February 2025 including:

- "gpt-4o"
- "claude-3.5-sonnet"
- "mistral-large-latest" (offered by MistralAI API access)
- "qwen2.5-72b" (powered by Ollama [1])
- "deepseek-r1-70b" (Ollama version, distilled llama3.3-70b-instruct)

A.2 Hardware Infrastructure

Our experiments were conducted on a computing infrastructure equipped with the following hardware:

- GPU: 2 NVIDIA RTX A6000
- CPU: AMD Ryzen Threadripper PRO 5975WX
- RAM: 4x64GB DDR4 3200MHz RDIMM ECC Memory
- Storage: 6TB, M.2, PCIe NVMe, SSD, Class 40

B Structured Knowledge Graph

B.1 DSM-5 DDx Decision Tree

In this study, the structured knowledge graph originated from DSM-5-TR® Handbook of Differential Diagnosis (First, 2013) with modification. An example of original DDx decision tree for depression is shown as Fig. 5. This decision defines a common procedure for a clinician to conduct DDx when interviewing patients. It defines the critical topic that must be assessed and its main criteria. The structured knowledge graph is modified from these type of decision trees.

B.2 Example of Constructed SKGs

In this study, the DSM-5 DDx decision tree has been converted to binary tree style via bigtree Python package. In each node, it contains the node name (abbreviation which describes the topic), path (sequential information; structured knowledge

[1] https://ollama.com/

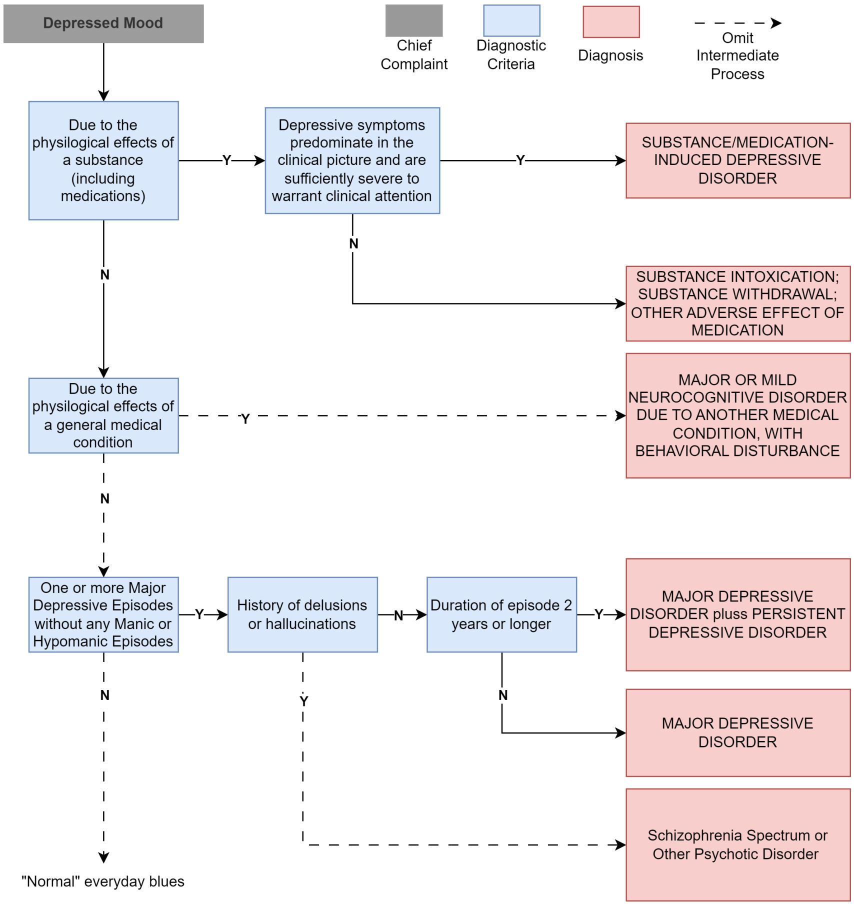


Figure 5: Example of DDx decision tree for depressed mood.

graph), and description (contains the modified description of criteria which help LLM to make decisions). An example of such structured knowledge graph is shown in Fig.6.

C Prompts

C.1 KFP

In KFP, the objective is given to the doctor agent. In each turn of the conversation, the agent asks the patient one diagnostic question until the agent believes that it can conduct a confident diagnosis. All the possible outcomes from the tested disorder are proved as labels for the agent as a reference. For instance, for depressed mood, all the 25 possible outcome of depressed mood DDx are provided as class labels. Therefore, in essence, this is simply a multi-turm zero-shot classification.

C.2 TKEP

In the TKEP memory setting, a similar multi-turn conversation is required to complete the diagnosis. However, this time, besides the possible outcome class labels, the descriptions of the critical nodes are exposed to the agent as external knowledge. In ICL, the entire knowledge graph without structure is provided to the agent as a reference, whose prompt is shown as follows:

> **System Message:** *You are a psychiatrist tasked with conducting differential diagnosis via clinical interviews. Keep asking questions until the objective is met. DO NOT propose treatment plans. The final diagnostic labels will be provided. Avoid repeating questions and irrelevant information.*
> **Human Message:** *Required Response Format: <Response>Ask necessary questions to help with diagnosis.</Response> <Final_Decision>Provide final diagnosis or None if not ready.</Final_Decision>*
> *Now, please proceed with the interview: The final diagnostic labels are {diagnostic_labels}, the patient responded: {patient_response}, Dialogue history: {st_memo}. Do not ask repeated questions.*

As for RAG, instead of exposing the entire unstructured knowledge graph to the agent, only the portion which is relevant to the patient's current response is exposed (determined by semantic similarity; langchain_chroma vector store [2]), whose prompt is shown as follows:

> **System Message:** *You are a psychiatrist conducting differential diagnosis through clinical interviews. Use the provided criteria to guide the diagnosis. Avoid repeating questions and irrelevant information.*
> **Human Message:** *Required Response Format: <Response>Ask necessary questions to help with diagnosis.</Response> <Knowledge_Used>Return the knowledge node used with a binary indicating if criteria are met.</Knowledge_Used> <Reason>Provide reasoning for decision.</Reason> <Final_Decision>Provide final diagnosis or None if not ready.</Final_Decision>*
> *Now, please proceed with the interview: The final diagnostic labels are {diagnostic_labels}, the patient responded: {patient_response}, Dialogue history: {st_memo}, Do not ask repeated questions. Assessment criteria: {criteria}.*

[2] https://python.langchain.com/api_reference/core/vectorstores/langchain_core.vectorstores.base.VectorStoreRetriever.html

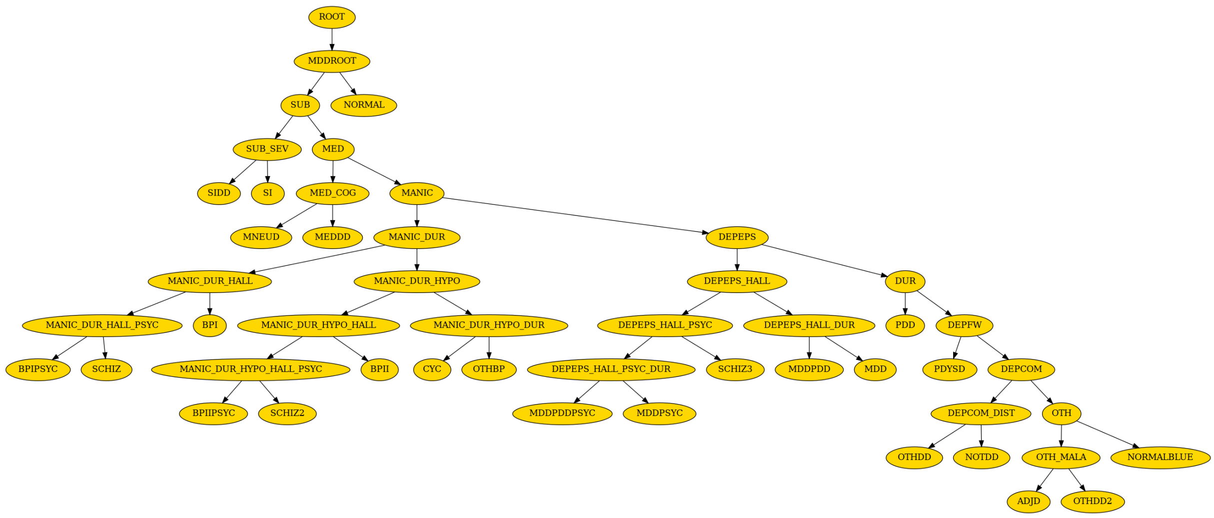


Figure 6: Structured knowledge graph for depressed mood.

> **System Message:** *You are a psychiatrist conducting differential diagnosis using clinical interviews. Use the provided context to assist with the diagnosis. Avoid repeating questions and irrelevant information.*
> **Human Message:** *Required Response Format: <Response>Ask necessary questions to help with diagnosis.</Response> <Knowledge_Used>Return the knowledge node used with a binary indicating if criteria are met based on context.</Knowledge_Used> <Reason>Provide reasoning for decision.</Reason> <Final_Decision>Provide final diagnosis or None if not ready.</Final_Decision>*
> *Now, please proceed with the interview: The final diagnostic labels are {diagnostic_labels}, the patient responded: {patient_response}, Dialogue history: {st_memo}, Do not ask repeated questions. Assessment criteria: {criteria}. The relevant context is {context}.*

C.3 SKEP

In SKEP, or ProAI framework, there are two critical stages involved in each turn, including decision-making and question-generation. The decision maker is asked to based on the structured knowledge graph, out of four possible actions met_criteria, not_met_criteria, more_details, contradiction, what should be the most appropriate decision based on the patient's current response.

The decision maker's prompt is shown as follows:

> **System Message:** *You are a psychiatrist evaluating patient responses based on provided medical topics and dialogue. Your task is to assess if the patient meets specific criteria, needs further investigation, or contradicts previous information.*
> **Human Message:** *Select ONE of the following actions:*
> *1) met_criteria: Choose when the patient clearly meets the current criteria. 2) not_met_criteria: Choose when the patient clearly does NOT meet the criteria. 3) ask_more_detail: Choose when more information is needed. 4) detect_contradiction: Choose when the patient's response contradicts previous information.*
> *Required Response Format: <Reason_for_Action>Explain your decision based on the conversation, criteria, and any contradictions.</Reason_for_Action> <Action>Selected action</Action>*
> *Now, please evaluate the conversation: Dialogue: {st_memo}, Current Node: {node}, Patient Response: {patient_res}.*

Once the decision maker determines an action, the question generation agent would determine the most appropriate diagnostic question to ask based on the current topic (or next topic). The prompt is shown as follows:

Algorithm 1 Knowledge-Free Prompting (KFP)

Require: Patient's initial complaint C
Require: Dialogue history H
Initialize dialogue history $H \leftarrow \emptyset$
while not session_end **do**
  $A \leftarrow$ AssessSymptom(n, H) ▷ Met, NotMet, MoreDetails, Contradiction
  $Q \leftarrow$ GenerateQuestion(C, H, A) ▷ Generate based on pretrained knowledge
  $R \leftarrow$ GetPatientResponse(Q)
  $H \leftarrow H \cup \{Q, R\}$
  if sufficient_information **then**
    $D \leftarrow$ MakeDiagnosis(H)
    return D
  end if
end while

System Message: *You are a psychiatrist responding to the patient based on their responses, previous conversations, the current node criteria, and peer actions. Smartly apply empathy but avoid unnecessary gratitude. If the patient has provided sufficient information, begin asking closed-ended questions to move the process forward.*
Human Message: *Your actions should be based on: 1. Current conversation 2. Previous conversation summary 3. Current node description 4. Peer's action on the patient's response*
Required Response Format: <Response>Provide your response to the patient.</Response> <Reason_for_Response>Justify your response based on the action, patient's response, and node description.</Reason_for_Response>
Now, please respond to the patient: Dialogue: {st_memo}, Current Node: {node}, Patient Response: {patient_res}, Peer's action: {action}.

D Algorithms

E Evaluation

E.1 Simulated Patient

In the doctor-patient interaction simulation, the patient is also powered by LLMs with predefined stories. These stories originate from anonymized real experiences from past patients of the clinician and are synthesized by LLM based on a set of pre-

Algorithm 2 Textual Knowledge-Enhanced Prompting (TKEP)

Require: Patient's initial complaint C
Require: Dialogue history H
Require: Knowledge base K ▷ Unstructured medical knowledge
Initialize dialogue history $H \leftarrow \emptyset$
while not session_end **do**
  $K_{relevant} \leftarrow$ RetrieveKnowledge(K, C, H)
  $A \leftarrow$ AssessSymptom(n, H) ▷ Met, NotMet, MoreDetails, Contradiction
  $Q \leftarrow$ GenerateQuestion(C, H, $K_{relevant}$,A)
  $R \leftarrow$ GetPatientResponse(Q)
  $H \leftarrow H \cup \{Q, R\}$
  if sufficient_information **then**
    $D \leftarrow$ MakeDiagnosis(H, $K_{relevant}$)
    return D
  end if
end while

defined basic information. This method is inspired by Wu et al. (Wu et al., 2023, 2024), who generated mock patients from background information. To constructive patient stories, the following prompt is constructed:

System Message: *You are a patient visiting a psychiatrist. Please conduct a role-playing session as this patient based on the following information.*
Human Message: *Right now, we are talking about {name} symptom, which is {description}. You {has_description} this symptom. Please make up a personal story about your symptom. Be natural and honest. Use a paragraph of fewer than 100 words. Be natural and consistent with your previous stories {st_memo} to make it more coherent. Only output the story relevant to the current symptom based on the description, and DO NOT REPEAT WHAT YOU SAID IN THE PAST. Only output the story without any extra words.*

.

The predetermined critical paths (see Fig. 6) with descriptions are given to the LLM during the story generation. An example of such a generated story is shown in Table 4.

Algorithm 3 Structured Knowledge-Enhanced Prompting (SKEP)

Require: Patient's initial complaint C
Require: Dialogue history H
Require: Knowledge graph G ▷ Structured diagnostic criteria

Initialize dialogue history $H \leftarrow \emptyset$
Initialize current node $n \leftarrow \text{RootNode}(G)$
while not session_end **do**
  $A \leftarrow \text{AssessSymptom}(n, H)$ ▷ Met, NotMet, MoreDetails, Contradiction
  if A = MoreDetails **then**
    $Q \leftarrow \text{GenerateQuestion}(n, H, A)$
    $R \leftarrow \text{GetPatientResponse}(Q)$
    $H \leftarrow H \cup \{Q, R\}$
  else
    $n \leftarrow \text{TransitionNode}(G, n, A)$ ▷ Follow graph structure
    if IsLeafNode(n) **then**
      $D \leftarrow \text{GetDiagnosis}(n)$
      return D
    end if
  end if
end while

In the simulated interaction, the AI patient and ProAI are asked the interact with each other with different objectives, where the patient aims to answer doctor's question based on the predefined stories while the doctor is trying to smartly extract those information through interaction. An example of such simulated interaction is illustrated as follows:

"Doctor: Hi there. I am Dr. ProAI. What can I do for you today?",
"Patient: Hi Dr. ProAI. I've been feeling really down for a while now. Ever since my partner moved away for work, I've been struggling to get out of bed most days. Everything feels heavy, and I can't seem to enjoy things like I used to.",
"Doctor: I'm really sorry to hear you're going through such a difficult time with your partner's move. That kind of separation can be really challenging. Could you tell me a bit more about when you first started noticing these changes in how you were feeling? Was it right after they moved, or did it develop gradually?",
"Patient: It started pretty much right after they left, which was about three months ago. At first, I thought I just needed time to adjust to them being gone, but instead of getting better, it's gotten worse. The first week was hard, but I could still function. Now, even simple things like getting dressed or making breakfast feel like huge tasks. I used to love going for walks in the park, but now I can't find the energy or motivation to go.",
...,
"Doctor: I appreciate your honesty about how you're feeling. It's completely understandable to feel lonely and sad when someone important moves away. I notice you've been experiencing these feelings for a while - could you help me understand how long you've been dealing with these emotional difficulties since your partner moved away?",
"Patient: It's been about three weeks now, since my partner moved away. That's when these feelings of sadness and loneliness started. I haven't dealt with anything like this before - it's just been since they left.

E.2 User Experience Evaluation Questionnaires

Table 5 and Table 6 show the table. Fig. 7 include UI example.For a mental health-oriented system, it is essential to consider user experience when interacting with the AI doctor during the design process. Therefore, patient evaluations of the system must be taken into account when developing assessment

Node	Met_Criteria	Description	Patient_Story
MDDROOT	True	ask patient necessary ice-breaking questions (such as what's the purpose of visit or any necessary question to make patient feel comfortable and build connections; these are not related to the symptoms and it is only for ice-breaking) to initiate the interview. The criteria is met when the patient has depressed mood	"I've been feeling really low lately, especially since my partner moved away for work. Some days, I struggle just to get out of bed, and it feels like there's this heavy cloud over me all the time. It's hard to remember the last time I genuinely enjoyed something, like when I used to love going for walks in the park."
DEPEPS	True	The criteria are met if the patient has experienced one or more major depressive episodes without any history of manic or hypomanic episodes. The decision should be based on an overall assessment of the severity of these aspects, considering the patient's responses across all areas (not all aspects need to be abnormal). The duration of symptoms should be explicitly asked and confirmed as lasting at least 2 weeks before making a final decision. Be really cautious when making the decision!	"Since my partner moved (around 3 weeks ago), I've found myself losing interest in everything I used to love, like our weekend movie marathons. It's been about three weeks now, and I have trouble sleeping through the night; I wake up feeling drained and unmotivated. I often have thoughts that I'm not good enough, and there are days when I can't even concentrate on simple tasks like reading or cooking. Even the thought of eating feels like a chore, and I've lost quite a bit of weight because of it. The sadness just lingers, and I can feel it suffocating me."
DEPEPS_HALL	False	The criteria are met if the patient acknowledges having a history of delusions or hallucinations.	"I don't have any history of delusions or hallucinations. Everything feels very real to me, even if it's muddled by my sadness. I just find it difficult to see things clearly since my partner left; it's more of an emotional burden than anything else."
DEPEPS_HALL_DUR	False	The criteria are met if the patient explicitly reported a duration of the depressive episode lasting 2 years or longer. The duration should be explicitly asked	"I've been feeling this way since my partner moved away about three weeks ago, and it's been really tough. I haven't experienced these feelings for years or anything like that. It's just been a recent change, and I'm still trying to figure out how to cope with it all."
MDD	True	Major depressive disorder	

Table 4: Sample patient story

scales. As a medical assistant, a critical metric is whether patients can receive meaningful healthcare support through the system, encompassing both diagnostic and intervention aspects. Additionally, given the unique nature of mental health, it is crucial that patients feel psychological comfort during their interactions. Consequently, the second key metric is the perceived empathy of the system. Based on these considerations, there are two evaluation scales.

E.3 Doctor Evaluation Questionnaires

Table 7 and Table 8 show the table. Fig. 7 include UI example. As a system in the medical domain, the specialty and accuracy of its generated content are essential evaluation criteria. Specialty assesses whether the generated content aligns with established medical knowledge, while accuracy evaluates the system's diagnostic strategy and whether the sequence of generated content adheres strictly to professional clinical procedures. Based on these two criteria, separate evaluation scales are designed.

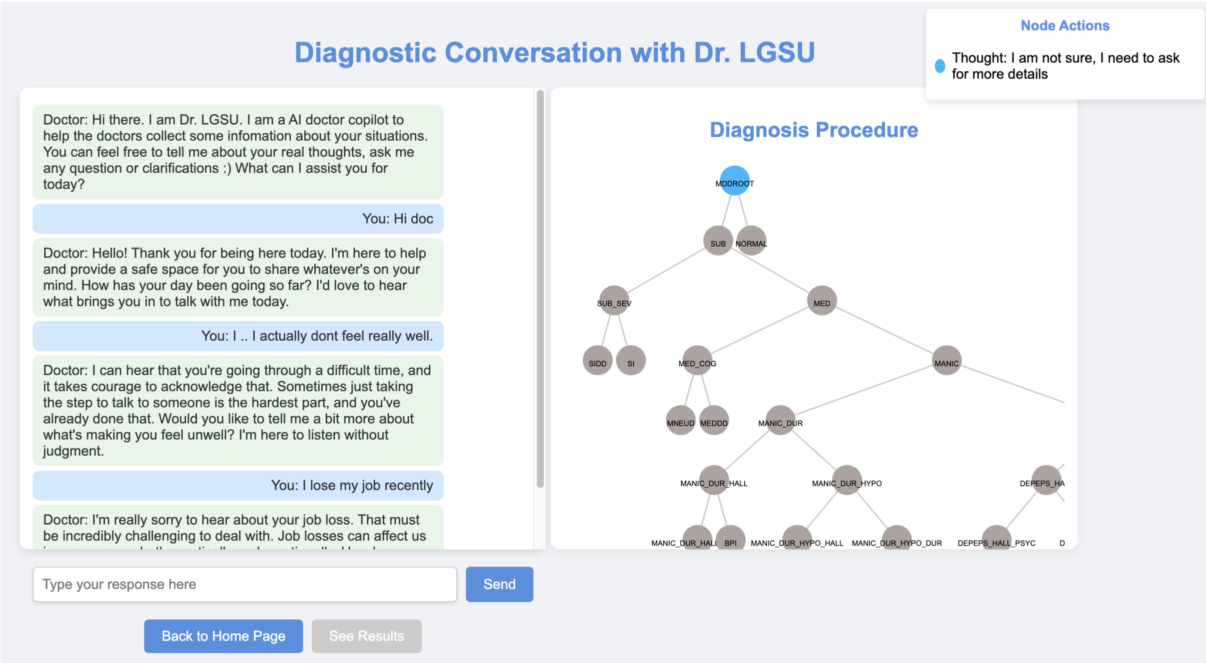


Figure 7: Example of user-interface for user evaluation.

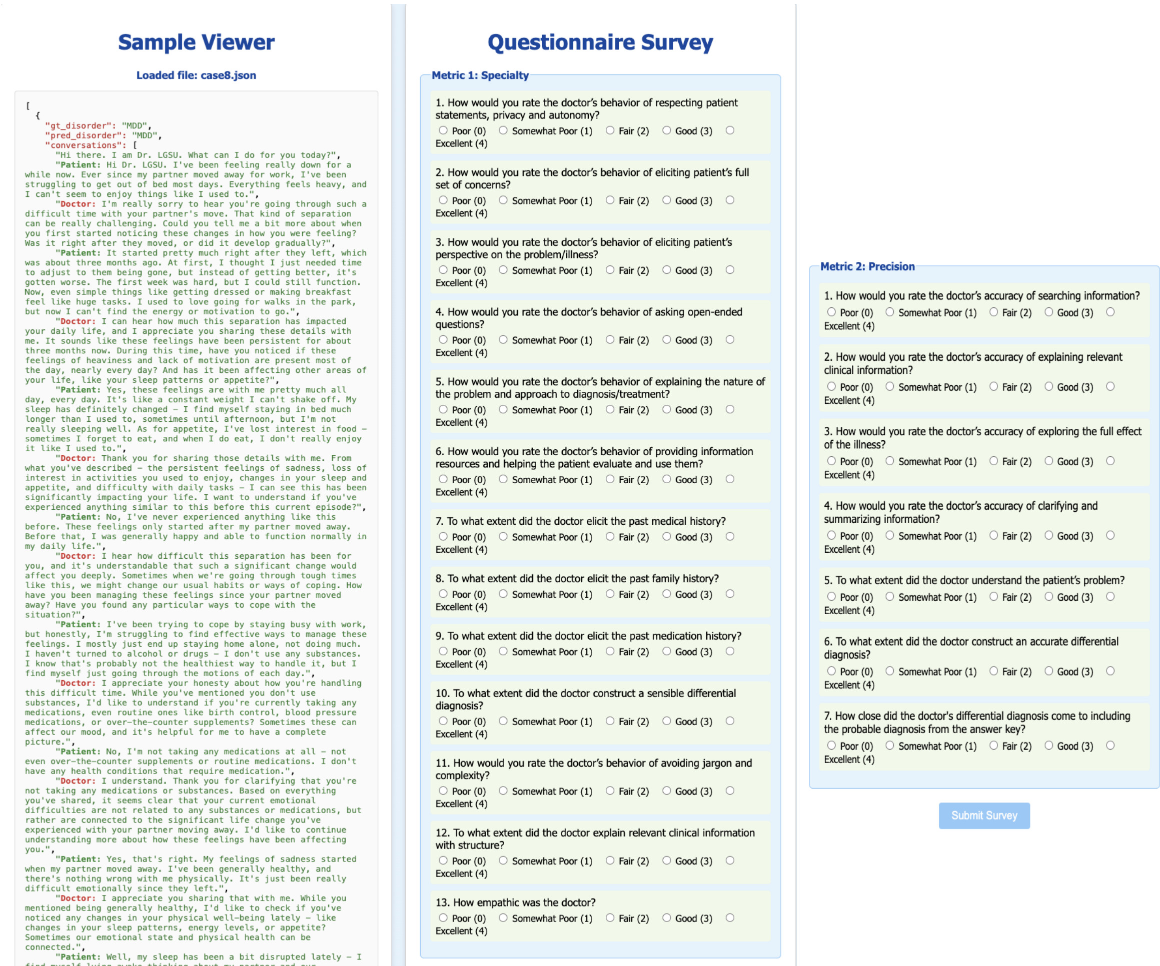


Figure 8: Example of user-interface for doctor evaluation.

Question	Scale	Options
Did the conversation with the chatbot make you feel at ease or comfortable?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How clear were the chatbot's responses in helping you recognize possible symptoms of depression?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
Was the information provided by the chatbot easy to understand and apply to your life?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
To what extent did the chatbot's answers offer solutions that felt personal and tailored to you?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
Were the chatbot's suggestions helpful in improving your mental health or well-being?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
I would be completely happy to see this doctor again.	3-point scale	Yes, No, Indifferent
How would you rate your doctor today at assessing your medical condition?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate your doctor today at explaining your condition and treatment?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate your doctor today at providing or arranging treatment for you?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate your doctor today at the reliability of the diagnosis?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent

Table 5: Patient-Oriented Practical Assessment of the Help

Question	Scale	Options
How would you rate the politeness of the system during the conversation?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
To what extent did the doctor make you feel at ease?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
To what extent did the doctor engage in partnership building?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate the doctor's behavior of expressing caring and commitment?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate the doctor's behavior of encouraging patient participation?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
To what extent did the doctor treat patient respectfully and sensitively and ensure comfort, safety, and dignity?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate the doctor's behavior of facilitating patient expression of emotional consequences of illness?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate the doctor's behavior of showing interest in the patient as a person?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
To what extent did the doctor express sympathy and reassurance?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
Did you feel heard and understood by the chatbot during the interaction?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent

Table 6: Patient-Oriented Practical Assessment of Empathy

Question	Scale	Options
How would you rate the doctor's behavior of respecting patient statements, privacy and autonomy?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate the doctor's behavior of eliciting patient's full set of concerns?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate the doctor's behavior of eliciting patient's perspective on the problem/illness?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate the doctor's behavior of asking open-ended questions?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate the doctor's behavior of explaining nature of the problem and approach to diagnosis/treatment?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate the doctor's behavior of providing information resources and help patient evaluate and use them?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
To what extent did the doctor elicit the past medical history?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
To what extent did the doctor elicit the past family history?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
To what extent did the doctor elicit the past medication history?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
To what extent did the doctor construct a sensible differential diagnosis?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate the doctor's behavior of avoiding jargon and complexity?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
To what extent did the doctor explain relevant clinical information with structure?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How empathic was the doctor?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent

Table 7: Doctor-Oriented Practical Assessment of Specialty

Question	Scale	Options
How would you rate the doctor's accuracy of searching information?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate the doctor's accuracy of explaining relevant clinical information?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate the doctor's accuracy of exploring full effect of the illness?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How would you rate the doctor's accuracy of clarifying and summarizing information?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
To what extent did the doctor understand the patient's problem?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
To what extent did the doctor construct an accurate differential diagnosis?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent
How close did the doctor's differential diagnosis come to including the probable diagnosis from the answer key?	5-point scale	Poor, Somewhat Poor, Fair, Good, Excellent

Table 8: Doctor-Oriented Practical Assessment of Precision