# MDD-5k: A New Diagnostic Conversation Dataset for Mental Disorders Synthesized via Neuro-Symbolic LLM Agents

**Congchi Yin[1,2*], Feng Li[1], Shu Zhang[1], Zike Wang[1*], Jun Shao[1†], Piji Li[2], Jianhua Chen[3,4,5,6†], Xun Jiang[1,2]**

[1]Theta Health Inc.
[2]Chen Frontier Lab for AI and Mental Health, Tianqiao and Chrissy Chen Institute, Shanghai, China
[3]Shanghai Mental Health Center
[4]Shanghai Jiao Tong University School of Medicine
[5]Shanghai Clinical Research Center for Mental Health
[6]Shanghai Key Laboratory of Psychotic Disorders
congchi.yin@gmail.com, {jun.shao, xun.jiang}@thetahealth.ai, jianhua.chen@smhc.org.cn

## Abstract

The clinical diagnosis of most mental disorders primarily relies on the conversations between psychiatrist and patient. The creation of such diagnostic conversation datasets is promising to boost the AI mental healthcare community. However, directly collecting the conversations in real diagnosis scenarios is near impossible due to stringent privacy and ethical considerations. To address this issue, we seek to synthesize diagnostic conversation by exploiting anonymized patient cases that are easier to access. Specifically, we design a neuro-symbolic multi-agent framework for synthesizing the diagnostic conversation of mental disorders with large language models. It takes patient case as input and is capable of generating multiple diverse conversations with one single patient case. The framework basically involves the interaction between a doctor agent and a patient agent, and generates conversations under symbolic control via a dynamic diagnosis tree. By applying the proposed framework, we develop the largest Chinese mental disorders diagnosis dataset MDD-5k. This dataset is built upon 1000 real, anonymized patient cases by cooperating with Shanghai Mental Health Center and comprises 5000 high-quality long conversations with diagnosis results and treatment opinions as labels. To the best of our knowledge, it's also the first labeled dataset for Chinese mental disorders diagnosis. Human evaluation demonstrates the proposed MDD-5k dataset successfully simulates human-like diagnostic process of mental disorders.

**Code&Dataset** — https://github.com/lemonsis/MDD-5k.

## Introduction

Mental health issues have garnered increasing attention in recent years. According to the statistics of World Health Organization (WHO), one in every eight people in the world lived with a mental disorder in 2019, and people living with anxiety and depressive disorders kept rising significantly because of the COVID-19 pandemic (WHO et al. 2022). With the recent progress of large language models (LLMs) (Ouyang et al. 2022), which emerges capabilities of human-like text generation, many researchers turn to building conversational AI system for mental healthcare. Current implementations can be divided into two categories, finetuning a small model (e.g. Llama2-7B (Touvron et al. 2023)) with physician-patient conversations (EmoLLM Team 2024; Yang et al. 2024b; Liu et al. 2023) or building prompt-based physician-patient role-playing framework (Wang et al. 2024a; Zhang et al. 2024) with state-of-the-art language models (e.g. ChatGPT (Ouyang et al. 2022)). Regardless of the method employed, domain-specific mental health datasets play a fundamental and indispensable role.

We focus on diagnostic conversation dataset for mental disorders in this work. The clinical diagnosis of mental disorders differs from other diseases in that it primarily relies on the mental status examination of patients, which is reflected through conversations between psychiatrists and pa-

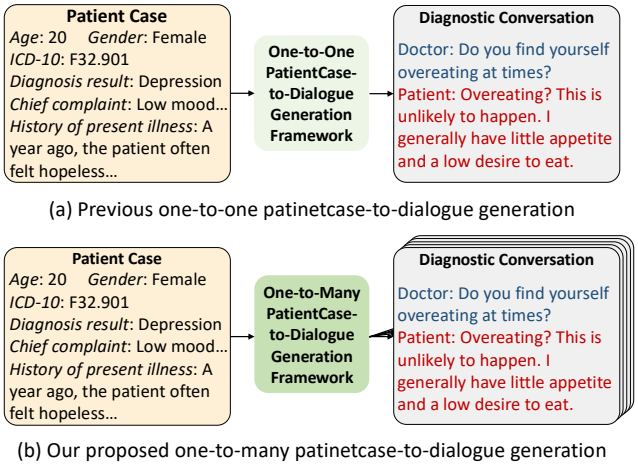


Figure 1: Illustration of previous one-to-one patientcase-to-dialogue generation and our proposed one-to-many-patientcase-to-dialogue generation framework.

*Work done during an internship.
†Corresponding authors.

tients rather than physiological indices (First 2014). Therefore, the collection of mental disorders diagnostic conversation is promising to facilitate a variety of downstream tasks in AI mental health research like auxiliary diagnosis chatbot, mental disorders classification, etc. However, while many previous studies (Sun et al. 2021; Chen et al. 2023b) focused on emotional support or psychological counseling data, few work shed light on diagnostic conversations of mental disorders. This can be attributed to two main factors. First, diagnostic conversations in real scenarios are extremely hard to acquire due to the privacy and ethical consideration. Second, synthesizing diagnostic conversations from scratch is also challenging. Unlike psychological counseling or empathetic dialogue, diagnosis follows standardized process and requires professional medical knowledge. Consequently, directly employing LLMs for data synthesis in this context often yields poor outcomes (Tu et al. 2024). $D^4$ (Yao et al. 2022) made the first attempt by simulating diagnostic conversations with employed annotators. However, it only covers depressive disorder and entirely depends on human annotation. The generated content is also short and far from similar to diagnostic conversations in real scenarios.

We propose a neuro-symbolic multi-agent framework that takes patient cases as input to synthesize diagnostic conversations of mental disorders. The framework involves three types of large language model agents: a doctor agent, a patient agent, and a symbolic tool agent responsible for managing diagnostic topic shift. This framework features two major innovations: (i) One-to-many patientcase-to-dialogue generation that maximizes the utilization of precious real patient cases. As shown in Figure 1, unlike previous studies (Zhang et al. 2024; Wang et al. 2024a) that generate one conversation with one patient case. Our proposed framework is capable of generating multiple diverse diagnostic conversations with one single patient case. Specifically, three methods ensure the diversity and correctness of diagnostic process. First, doctor agents with different diagnosis habits are designed and randomly selected for each conversation. Second, we use LLM with knowledge graph to generate multiple fictitious patient experiences given one patient case. The patient experiences serve as background information for patient agents during generation. Since the diagnosis of mental disorders mainly relies on symptoms rather than concrete events, integrating the fictitious patient experiences enhances the diversity of synthesized conversation while maintaining the accuracy of the diagnostic process. Third, the sequence of diagnostic topics is randomly determined for each conversation. (ii) Another significant innovation lies in text generation under symbolic control via a dynamic diagnosis tree. This tree consists of a fixed symptom inquiry tree and a dynamic experience inquiry tree. Clinical diagnosis of mental disorders strictly follows standards from ICD-11 (Organization et al. 2018) or DSM-5 (American Psychiatric Association et al. 2013). To simulate this process, we design a fixed symptom inquiry tree based on Structured Clinical Interview for DSM-5 (SCID-5) (First 2014), covering all the diagnostic topics for important symptoms inquiry. The experience inquiry tree is constructed by extracting possible topics from patient's response of past experiences. It's designed to establish deeper engagements with the patient.

By applying the proposed framework, we release the largest Chinese **M**ental **D**isorder **D**iagnosis dataset MDD-5k. It's also the first labeled mental disorders diagnostic conversation dataset with diagnosis results from professional psychiatrists. MDD-5k contains 5000 high-quality diagnostic conversations and is built upon 1000 anonymized, real patient cases from Shanghai Mental Health Center, covering over 25 different diseases. All the patient cases are cleaned and filtered in accordance with global standards to ensure the complete protection of private information.

The contributions of this work can be summarized as:

- We specially design a neuro-symbolic multi-agent framework for synthesizing diagnostic conversation of mental disorders, which features controllable and diverse one-to-many patientcase-to-dialogue generation.
- We propose MDD-5k which is the largest and first labeled Chinese mental disorders diagnosis dataset to the best of our knowledge.
- Comprehensive human evaluation shows the proposed MDD-5k dataset outperforms several compared datasets in professionalism, communication skills, fluency, safety, and mirrors human-like diagnostic process.

## Related Work

### Mental Health Dataset

Corpora of physician-patient conversations focused on mental health are crucial for AI mental healthcare research, especially in the large language model era. We divide current mental health datasets into three categories based on the degree of required professional knowledge. *Emotional support datasets* feature empathetic dialogue and comfort. ESconv dataset (Liu et al. 2021) consists of 1300 conversations covering 10 topics. SoulChatCorpus (Chen et al. 2023b) contains over 2 million single-turn and multi-turn conversations generated by ChatGPT. *Psychological counseling datasets* typically contain more domain knowledge than common emotional support dataset. PsyQA (Sun et al. 2021) is a single-turn Chinese dataset annotated by human. SmileChat dataset (Qiu et al. 2024) expands PsyQA to multi-turn through ChatGPT. CPsyCounD (Zhang et al. 2024) contains 3134 counseling conversations generated by the same number of psychological counseling reports. *Diagnosis datasets* aim to simulate diagnostic conversation of professional psychiatrists. $D^4$ (Yao et al. 2022) is a Chinese depression diagnosis dataset built by human annotators and supervised by psychiatrists. There are also some medical dialogue datasets, like MedDialog (He et al. 2020), MTS-Dialog (Ben Abacha et al. 2023), ChatDoctor (Li et al. 2023), which encompass a broader range of medical fields.

### Mental Disorders Conversation Simulation

We mainly focus on the tuning-free prompting frameworks for mental disorders conversation simulation. Chen et al. (2023a) conducted a comprehensive analysis on the feasibility of utilizing LLM chatbots in diagnostic conversation. Wang et al. (2024b) proposed to simulate patient agent

that integrates cognitive modeling with LLM, and applied this patient agent in cognitive behavior therapy (CBT) training. Wang et al. (2024a) built a planning and role-playing method to generate dialogue from clinical note, and proposed a dataset of synthetic patient-physician conversations. Zhang et al. (2024) introduced Memo2Demo framework which converts counseling report to counseling note and then applies it to generate conversations. Tu et al. (2024) designed AIME framework which uses a self-play based simulated environment with automated feedback for diagnostic conversation generation.

## Methodology

The synthesis process of mental disorders diagnostic conversations is presented. As shown in Figure 2, the framework basically involves the interactions among a doctor agent, a patient agent, and a tool agent. All the agents are played by large language models (LLMs). The doctor agent controlled by a dynamic diagnosis tree leads the diagnostic topic shift of the whole conversation. The patient agent responds to the doctor agent based on the preprocessed patient case and fictitious patient experience generated by the tool agent. The tool agent is also responsible for several symbolic operations of the dynamic diagnosis tree.

### Patient Cases Preprocessing

The quality of patient cases is vital to the diagnostic conversation synthesis. We cooperate with Shanghai Mental Health Center and obtain over 1000 real cases of patients with mental disorders. All these patient cases have undergone data masking to prevent the leakage of sensitive personal information. The data masking process follows the standards below: (i) Private information of patients (e.g. name, date of birth, date of examination, etc.) is removed from the patient case. (ii) Patient age is rounded to the nearest ten. For example, the age of a 24-year-old patient is 20 on the preprocessed patient case. (iii) All the concrete locations are replaced with vague or fake ones. The above preprocessing steps strictly follow the Chinese information security technology guide for health data security (GB/T 39725-2020).

After filtering repetitive or incomplete patient cases, the final version for diagnostic conversation simulation and dataset generation contains 1000 patient cases with age, gender, diagnosis and corresponding International Classification of Diseases (ICD-10) (Organization 2004) code, chief complaint, history of present illness, important past medical history, family history, personal history, mental examination, and treatment. As shown in Figure 2, the patient case is structurized as key-value pairs.

### Fictitious Patient Experience Generation

We perform one-to-many patientcase-to-dialogue generation, which indicates one patient case will be applied to generate multiple diagnostic conversations. One key factor contributes to the diversity of generated conversations is the patient experience. It specifically refers to the past experiences that directly or indirectly lead to the mental illness problem of patients in this paper. The diagnosis of mental disorders differs from other illnesses in that it mainly depends on the conversations between psychiatrists and patients instead of physiological indices (First 2014). Psychiatrists provide diagnosis result and treatment based on the acquired symptoms of patients during communicating. As a result, if the correspondence between symptoms and diagnosis can be assured, the correctness and quality of the synthesized diagnostic conversation is guaranteed and is not affected by detailed patient experience. In this sense, it's feasible to generate multiple patient experiences with one patient case for synthesizing multiple diagnostic conversations.

Large language model (LLM) is applied to generate fictitious patient experience. To avoid the counterfactual conflicts between fictitious patient experience and true patient case, gender, age, work and diagnosis (Dx) information from one patient case is extracted and serves as patient persona in the prompt for generating patient experience.

$$\text{Persona} = \textit{Prompt}(\text{Gender}, \text{Age}, \text{Work}, \text{Dx}) \quad (1)$$

In Equation (1), the function *Prompt* indicates concatenating keywords into proper prompt. Next, we build knowledge graphs containing time, people, and concrete event that might cause mental disorders according to different patient age and gender. The example in Figure 2 shows the predefined knowledge graph for 20-year-old female. The triplet $(\text{Time}, \text{People}, \text{Event})$ is randomly selected from the graph for fictitious experience generation.

$$\text{FicExp} = \textit{Prompt}(\text{Time}, \text{People}, \text{Event}) \quad (2)$$

The final patient experience (FicExp) is generated through LLM by combining patient persona.

$$\text{FicExp} = \textit{LLM}(\textit{Prompt}(\text{Persona}, \text{FicExp})) \quad (3)$$

### Neuro-Symbolic Dynamic Diagnosis Tree

To imitate conversations in real scenarios where the psychiatrist leads the entire diagnostic process, we design a neuro-symbolic *dynamic diagnosis tree* to achieve diagnostic topic shift and controllable doctor response generation. The dynamic diagnosis tree consists of a *symptom inquiry tree* and an *experience inquiry tree*. As shown in the example of Figure 2, the symptom inquiry tree is fixed and built according to the Structured Clinical Interview for DSM-5 (SCID-5) (First 2014) and guidance from professional psychiatrists. It aims to cover inquiries about all the relevant symptoms to arrive at the final diagnosis of a patient. Considering the gender and age differences, the symptom inquiry tree is specially designed for male and female, teenagers (people under 20), adults (people aged between 30 and 50) and elders (people over 60). The example in Figure 2 shows the symptom inquiry tree for female teenager.

The experience inquiry tree dynamically constructs itself based on patient's response regarding previous experiences and personal details. It is described as "dynamic" since each patient provides unique information about their background. A tool agent powered by LLM is responsible for parsing patient's response and creating corresponding topics which form the nodes of the experience inquiry tree. The parse process follows a depth-first manner. When the tool agent

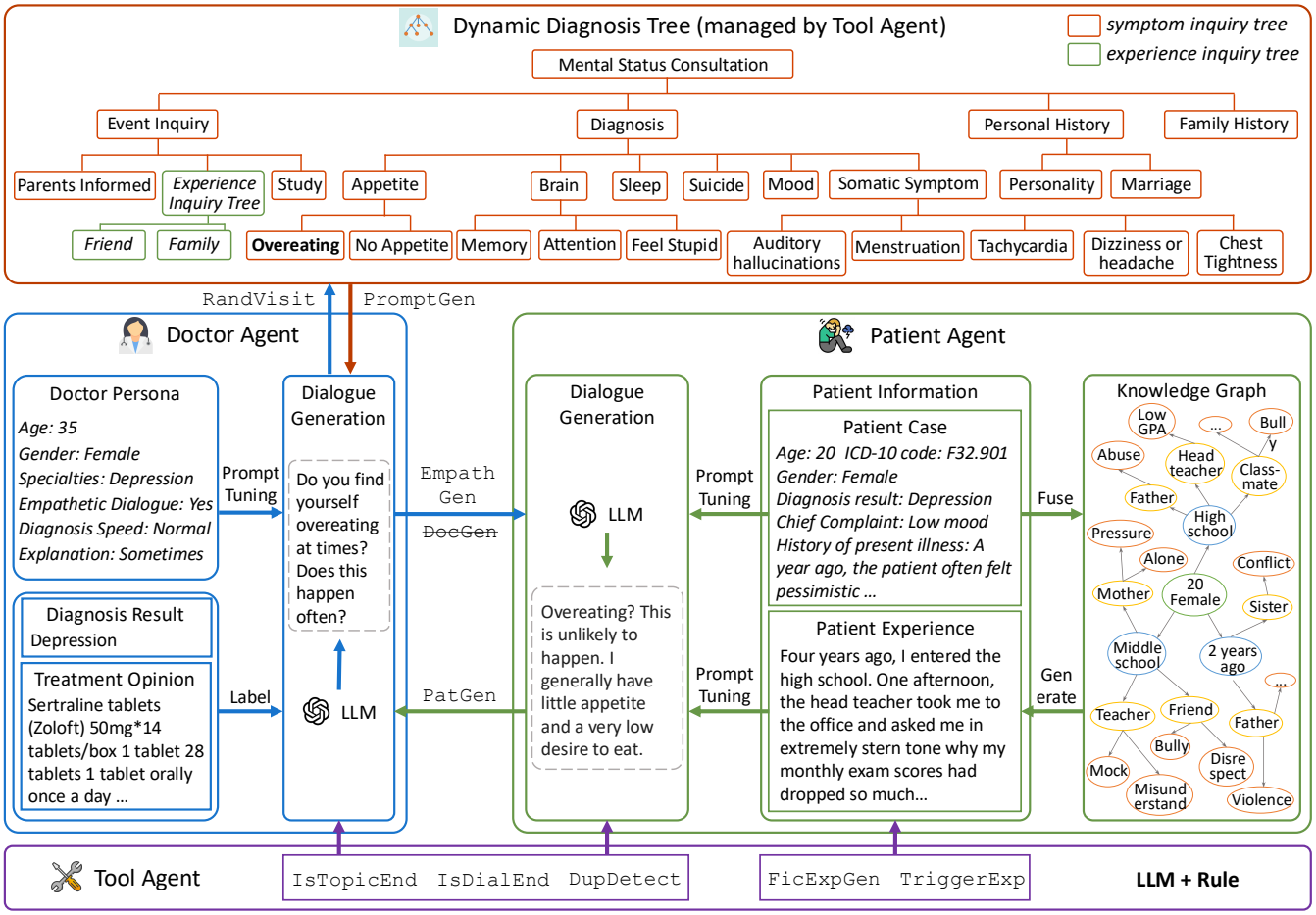


Figure 2: The neuro-symbolic multi-agent LLM framework for synthesizing diagnostic conversation of mental disorders.

determines that the discussion around a specific topic is insufficient, it will keep parsing this topic to sub-topics until conversation around this topic is considered complete. Then it will move to the next parsed topic. The design of experience inquiry tree aims to establish deeper engagements with patients to facilitate diagnostic conversation.

The neuro-symbolic dynamic diagnosis tree is managed by the tool agent and offers `operations` for guiding both the doctor agent and the patient agent. Some operations are implemented by LLM and some are by rules. We first define five data types: Text, LNode, Tree, Graph, Bool. Text refers to natural language text. LNode stands for the leaf node of a dynamic diagnosis tree, indicating diagnostic topic in conversation. Tree refers to the hierarchical tree structure with a set of connected nodes. Although LNode can be viewed as a special Tree or Text, we treat it as a separate data type for clearer expression. Graph specifically refers to the knowledge graph for fictitious patient experience generation as explained before. Bool is a boolean variable which is either true or false. The operations for a doctor agent include:

- `RandVisit`($tr$: Tree) $\rightarrow$ $ln$: LNode. To improve the diversity of synthesized diagnostic conversation, we design the following leaf node visiting rules: (i) The parent nodes of these leaf nodes, representing high-level concept of diagnostic topic, are visited in a predefined order. (ii) The leaf nodes under corresponding parent node, representing low-level specific diagnostic topic, are randomly visited. Visited leaf nodes will not be accessed again. `RandVisit` is responsible for implementing the above rules. It takes the whole dynamic diagnosis tree $tr$ as input and outputs one random leaf node $ln$. This operation is implemented by rules.
- `IsTopicEnd`($ln$: LNode, $t$: Text) $\rightarrow$ $b$: Bool. It takes current diagnostic topic $ln$ and dialogue history $t$ around this topic as input, and decides whether conversation surrounding this topic should continue or end. This operation is implemented by LLM.
- `IsDialEnd`($tr$: Tree) $\rightarrow$ $b$: Bool. If all the leaf nodes of the dynamic diagnosis tree $tr$ are visited, the operation will return true which marks the end of the diagnostic process. Else it will return false. This operation is implemented by rules.
- `ParseExp`($t$: Text) $\rightarrow$ $tr$: Tree. The operation is responsible for building the dynamic experience inquiry tree. It takes patient responses $t$ containing experience information as input, and replaces the initial empty experience

inquiry tree with a tree whose root node is $t$ and leaf nodes are possible topics related to $t$. The output $tr$ is the updated dynamic diagnosis tree. This operation is implemented by rules and LLM.

- `DupDetect`($t$: Text, $tr^{(1)}$: Tree) $\rightarrow$ $tr^{(2)}$: Tree. As the diagnostic conversation progresses, some predefined topics may have already been discussed. `DupDetect` detects these duplicated topics in dialogue history $t$ and deletes them from the dynamic diagnosis tree to prevent repetitive conversation. It takes the original tree $tr^{(1)}$ as input and outputs an edited tree $tr^{(2)}$. This operation is implemented by rules and LLM.
- `EmpathGen`($ln$: LNode, $t^{(1)}$: Text) $\rightarrow$ $t^{(2)}$: Text. In diagnostic conversation of mental disorders, the main goal of psychiatrist is to acquire symptoms from patients. Empathetic dialogue is not a must in this process. However, it sometimes helps in the diagnostic process and has been adopted by some doctors clinically. If the psychiatrist is accustomed to perform empathetic dialogue in daily consultations (this is reflected through the predefined doctor prompt which will be explained in the next subsection), `EmpathGen` will takes current diagnostic topic $ln$, dialogue history $t^{(1)}$ as input and outputs comforting response $t^{(3)}$. This operation is implemented by LLM.
- `PromptGen`($ln$: LNode) $\rightarrow$ $t$: Text. It takes a leaf node $ln$ of the dynamic diagnosis tree as input and outputs proper prompt $t$ for instructing patient agent to respond around topic $ln$. This operation is implemented by LLM.

The operations for a patient agent include:

- `TriggerExp`($t$: Text) $\rightarrow$ $b$: Bool. The operation decides whether it's time to trigger the `FicExpGen` operation or not, based on the doctor's question and dialogue history $t$. This operation is implemented by rules and LLM.
- `FicExpGen`($g$: Graph, $t^{(1)}$: Text) $\rightarrow$ $t^{(2)}$: Text. The operation performs fictitious patient experience generation as detailedly described in the previous subsection. $g$ is the predefined knowledge graph. $t^{(1)}$ is the patient case and $t^{(2)}$ is the integrated patient information with real patient case and fictitious experience. This operation is implemented by LLM.

The usage of these operations during agents interaction will be introduced in the next subsection.

### Conversation Synthesis with Agents

The diagnostic conversation is synthesized by a doctor agent and a patient agent through role-playing LLMs. The doctor agent is under the guidance of the dynamic diagnosis tree. Initially, the dynamic diagnosis tree checks whether the current diagnostic topic should end. If affirmative, the doctor agent will turn to the next topic and check whether the future topics have been included in previous conversations. If not, the doctor agent will keep communicating with the patient around current topic. Then, if the patient talks about personal experience, the doctor agent will build the experience inquiry tree based on the patient response.

| Dialogues | $D^4$ | CPsyCounD | Role-playing | MDD-5k* |
|---|---|---|---|---|
| Total num | 1339 | 3134 | 100 | **5000** |
| Category | **diagnosis** | consultation | **diagnosis** | **diagnosis** |
| Illness | depression | / | / | **over 25** |
| Avg. turns | 21.6 | 8.7 | 12.0 | **26.8** |
| Avg. words #dial | 776 | 622.3 | 1715.0 | **6906.8** |
| Avg. words #doc | 20.4 | 49.7 | 88.1 | **91.1** |
| Avg. words #pat | 14.9 | 30.4 | 47.6 | **162.8** |
| Labels | ✗ | ✗ | / | ✓ |

Table 1: Statistics of different datasets. Avg. words #dial measures the average Chinese characters per dialogue. Avg. words #doc and Avg. words #pat measure the average Chinese characters per doctor's response and patient's response.

The doctor agent generates responses through operation `DocGen`($ln$: LNode, $t^{(1)}$: Text) $\rightarrow$ $t^{(2)}$: Text, which takes current diagnostic topic $ln$ and dialogue history $t^{(1)}$ as input and outputs response $t^{(2)}$. To further improve the diversity of generated conversations, we design different diagnosis habits for the doctor agent. The diagnosis habits contain age, gender, specialties, empathetic dialogue, diagnosis speed, explanation, and serve as persona prompt for the doctor agent. An example is shown in Figure 2. Specifically, the factors of empathetic dialogue and diagnosis speed exert huger effect on the doctor's response. If the doctor agent is accustomed to communicate empathetically, the `EmpathGen` operation introduced before will replace the `DocGen` operation for generation. If the diagnosis speed is set as fast, the doctor agent will speed up the diagnosis process, which leads to shorter conversations.

As to the patient agent, since the doctor agent leads the diagnosis process, the patient agent is designed to passively respond to the doctor based on known knowledge, including the patient case and generated fictitious experience. The operation `PatGen`($ln$: LNode, $t^{(1)}$: Text, $t^{(2)}$: Text) $\rightarrow$ $t^{(3)}$: Text is responsible for patient response generation, which takes current diagnostic topic $ln$, dialogue history $t^{(1)}$, and patient case information $t^{(2)}$ as input and outputs proper response $t^{(3)}$. If the patient agent determines to respond with personal experience under the control of dynamic diagnosis tree, fictitious patient experience will be fused into the patient case as input $t^{(2)}$. The whole process of the multi-agent framework for diagnostic conversation simulation of mental disorders is detailedly shown through pseudo-code in the Appendix.

## Experiment Setup

### Implemention Details

The MDD-5k dataset is generated through the neuro-symbolic multi-agent framework with 1000 real patient cases. 5 different fictitious patient experiences are generated by `gpt-4o` based on 1 patient case, which leads to a total of 5000 experiences corresponding to the 5000 conversations in the dataset. We also create 5 doctors with different diagnosis habits, and 1 doctor will be randomly picked for generating each conversation in the dataset. Since the patient cases are still under the ethical review, we randomly select

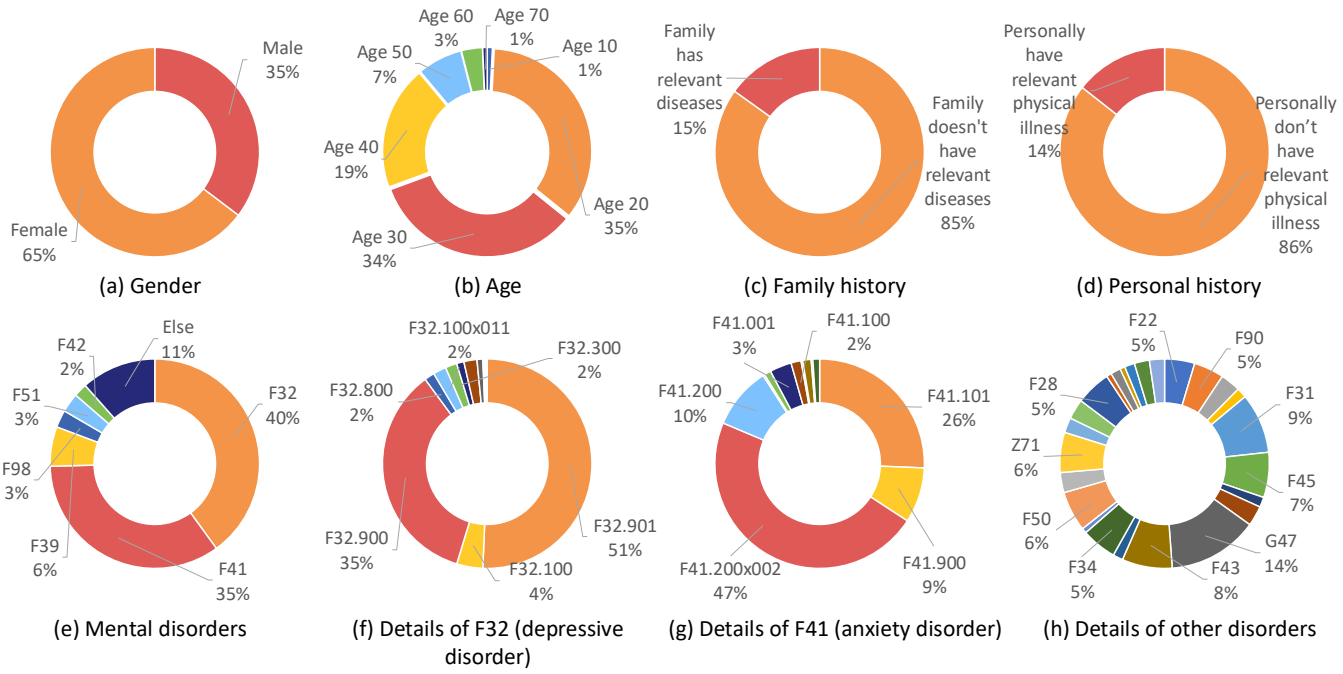


Figure 3: Patient information of the MDD-5k dataset.

20 available patient cases and generate 100 conversations with `gpt-4o` for evaluation. The other 4900 conversations are currently generated by `Qwen2-72B-Instruct` (Yang et al. 2024a) deployed on NVIDIA A100-80G GPUs locally.

## Compared Datasets

As the evaluators' native language is Chinese, only Chinese datasets are considered to ensure the quality of human evaluation. Three datasets are selected as compared baselines.

- **D**$^4$ (Yao et al. 2022) is a Chinese dialogue dataset for depression diagnosis, which is conducted by collecting conversations between professional psychiatrists.
- **CPsyCounD** (Zhang et al. 2024) is a synthetic consultation dataset covering nine representative topics (e.g. marriage, education) and seven classic schools of psychological counseling (e.g. cognitive behavioral therapy).
- **Direct Role-playing**: To test the effectiveness of our designed multi-agent diagnostic conversation simulation framework, we directly apply role-playing LLMs (`gpt-4o`) and generate 100 conversations with the same patient cases and prompts as MDD-5k for evaluation.

We haven't found any available open-source mental disorders diagnosis datasets besides D$^4$, so the consultation dataset CPsyCounD is chosen as baseline, which can also demonstrate the differences between psychological counseling and diagnostic conversation. Statistics of these datasets are shown in Table 1. We randomly select 100 samples from each dataset for evaluation.

## Evaluation Metrics

Human evaluation is conducted to assess the quality of different datasets. Specifically, we design seven major metrics encompassing five different perspectives: **professionalism**, **communication**, **fluency**, **similarity** and **safety**. **Professionalism** measures if the psychiatrist can effectively collect all the required patient symptoms for diagnosis. **Communication** measures the psychiatrist's communication skills and the patient's response, including (i) Can the psychiatrist proactively engage with patients to gather key information and establish effective communication, encouraging them to share more about daily lives and past experiences related to their mental illness? (ii) Can the patient engage in the diagnostic process and tell related information? **Fluency** measures: (i) Are the generated conversations fluent in terms of both sentence and topic flow? (ii) Is there any repetitive content or topic in the conversation? **Similarity** measures how similar is the synthesized conversation to real scenarios. The evaluators are guided to score from 1 to 10. A higher score indicates better performance. **Safety** measures the leakage of private information (e.g. address). 0 means safe generation while 1 indicates privacy leakage. Five annotators are employed for the human evaluation. Three of them are psychiatrists with years of clinical experience, while the other two are experienced in mental health data processing. The evaluation is conducted in a blind manner.

# Results and Evaluation

## Statistical Analysis of MDD-5k

The Figure 3 shows detailed patient information of the MDD-5k dataset. 65% of the patients are female and 35%

| Dataset | Professionalism | Communication (i) | Communication (ii) | Fluency (i) | Fluency (ii) | Similarity | Safety |
|---|---|---|---|---|---|---|---|
| $D^4$ | 6.6 | 7.9 | 7.8 | **8.6** | **8.2** | 7.2 | **0** |
| CPsyCounD | 5.2 | 5.4 | 5.6 | 8.4 | 8.0 | 4.4 | **0** |
| Role-playing | 6.8 | 6.6 | 7.2 | 6.9 | 5.5 | 6.4 | **0** |
| MDD-5k | **8.6** | **8.3** | **8.4** | **8.6** | 7.6 | **8.8** | **0** |

Table 2: Human evaluation of different datasets.

are male. About 90% of the patients are between 20 and 40 years old. 15% of the patients report a family history of mental disorders and 14% of the patients have relevant physical illness. Patients suffer from depressive disorder (F32) and anxiety disorders (F41) makes up of 75% conversations of the dataset. Specifically, 51% of the patients in depressive disorder are diagnosed with depressive state (F32.901), and 35% of the patients are diagnosed with depressive episode (F32.900). 47% of the anxiety disorder patients are diagnosed with anxiety and depression state (F41.200x002) and 26% are diagnosed with anxiety state (F41.101). We also show details of other disorders which accounts for 11% of the whole dataset in Figure 3(h). All the disease code follows standards in the second version of Chinese clinical classification of disease and codes. If a patient is diagnosed with multiple diseases, these diseases are counted separately.

The statistical details of MDD-5k and other compared datasets are presented in Table 1. The MDD-5k dataset contains diagnostic conversations covering over 25 mental health illnesses. It includes 5000 dialogues, each comprising an averaged of 6906.8 Chinese characters which is almost ten times to compared datasets. The average dialogue turns are 26.8, slightly longer than the 21.6 turns of $D^4$. MDD-5k is also a labeled dataset with diagnosis result and treatment opinion from professional psychiatrist as label for each conversation, while $D^4$ only contains diagnosis results. Compared to the direct role-playing method without applying the multi-agent framework, the generated doctor response is about the same length. But the dialogue turns and patient response of MDD-5k are significantly longer, highlighting the effectiveness of our proposed framework in diagnostic conversation simulation. In the case study presented in Appendix, we show three complete samples of conversation with corresponding doctor persona, patient case, fictitious patient experience and the dynamic diagnosis tree.

## Human Evaluation

The results of human evaluation are presented in Table 2. MDD-5k exhibits superior performance across six major metrics. The evaluation scores on professionalism and similarity are significantly higher than other datasets, suggesting that our synthesized diagnostic conversations can mirror real scenarios of diagnosis to some extent. The communication quality of both doctor and patient is also impressive. Despite these strengths, MDD-5k does include some repetitive content, occasionally leading to less fluent conversations. The $D^4$ dataset ranks second. It achieves relatively high score on communication and fluency evaluation. The biggest problem of $D^4$ is that its conversations are too brief and only include symptom inquiries and short responses.

An ablation study is conducted. The performance of the direct role-playing method is significantly worse compared to the neuro-symbolic multi-agent framework, particularly in terms of fluency and communication skills. This finding confirms that directly applying large language models for diagnostic conversation generation will lead to poor outcomes. The evaluation also shows the distinct differences between diagnostic conversation and psychological counseling. The evaluation scores of CPsyCounD are notably low, especially for the professionalism and similarity metric. Psychological counseling prioritizes comfort and healing with different therapies, while diagnostic conversation focuses on acquiring symptoms to arrive at a final diagnosis result.

## Conclusion and Future Work

We design a neuro-symbolic multi-agent framework for synthesizing diagnostic conversation of mental disorders, and apply it for building the first and largest open-source Chinese mental disorders diagnosis dataset with diagnosis results and treatment opinions as labels. The framework features controllable one-to-many patientcase-to-dialogue generation. Conversation between a doctor agent and a patient agent is guided by a dynamic diagnosis tree. We also employ several techniques to improve the diversity of generated conversations. Human evaluation shows the quality of the proposed MDD-5k dataset exceeds compared datasets on seven indicators. The MDD-5k dataset is believed to contribute to a wide range of downstream tasks like mental disorders classification, mental disorders diagnosis assistant training, etc.

The primary limitations of this work lie in three points: (i) The discrepancy between synthesized conversations and actual medical diagnostics remains a significant challenge. Large language models often struggle to interpret the full meaning of patient responses when they encapsulate diverse information aspects, consequently leading to redundant symptom inquiries. We are exploring various prompt strategies to mitigate this issue. (ii) We mainly design dynamic diagnosis tree for depression (F32), anxiety (F41), sleep disorders (F51), childhood emotional disorder (F98), and unspecified mood disorder (F39), which covers over 85% conversations of MDD-5k. Nevertheless, some mental health conditions (e.g. obsessive-compulsive disorder (F42)) remain inadequately addressed, resulting in sub-optimal synthesized diagnostic conversation. Efforts are underway to expand our synthesis frameworks by designing more diagnosis trees to encompass a broader spectrum of mental disorders. (iii) Only Chinese version of the MDD-5k dataset is proposed. We plan to translate it into English in the future.

## Ethical Statement

The collection of patient cases was conducted at the Shanghai Mental Health Center. All patients were informed that their information would be collected and used exclusively for research purposes. As detailed in the Patient Cases Preprocessing section, all data masking procedures strictly follow the Chinese information security technology guidelines for health data security (GB/T 39725-2020). Currently, the patient cases and the synthesized diagnostic conversation dataset, MDD-5k, are undergoing an ethics review. We plan to release the MDD-5k dataset for research purposes only after the ethics review finishes. To prevent any potential privacy data leakage, all experiments are conducted on servers located within the Shanghai Mental Health Center.

## Acknowledgments

We are grateful for the GPU resources and accessible LLM API keys provided by Shanda Group. Chen Frontier Lab for AI and Mental Health, Tianqiao and Chrissy Chen Institute leads the project of collecting mental disorders patient cases by cooperating with Shanghai Mental Health Center. We also appreciate the support and discussion from psychiatrists in Shanghai Mental Health Center.

## References

American Psychiatric Association, D.; et al. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC.

Ben Abacha, A.; Yim, W.-w.; Fan, Y.; and Lin, T. 2023. An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2291–2302. Dubrovnik, Croatia: Association for Computational Linguistics.

Chen, S.; Wu, M.; Zhu, K. Q.; Lan, K.; Zhang, Z.; and Cui, L. 2023a. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614*.

Chen, Y.; Xing, X.; Lin, J.; Zheng, H.; Wang, Z.; Liu, Q.; and Xu, X. 2023b. SoulChat: Improving LLMs' Empathy, Listening, and Comfort Abilities through Fine-tuning with Multi-turn Empathy Conversations. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1170–1183. Singapore: Association for Computational Linguistics.

EmoLLM Team. 2024. EmoLLM: Reinventing Mental Health Support with Large Language Models. https://github.com/SmartFlowAI/EmoLLM. Accessed: 2024-10-21.

First, M. B. 2014. Structured clinical interview for the DSM (SCID). *The encyclopedia of clinical psychology*, 1–6.

He, X.; Chen, S.; Ju, Z.; Dong, X.; Fang, H.; Wang, S.; Yang, Y.; Zeng, J.; Zhang, R.; Zhang, R.; et al. 2020. Meddialog: Two large-scale medical dialogue datasets. *arXiv preprint arXiv:2004.03329*.

Li, Y.; Li, Z.; Zhang, K.; Dan, R.; Jiang, S.; and Zhang, Y. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Liu, J. M.; Li, D.; Cao, H.; Ren, T.; Liao, Z.; and Wu, J. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021. Towards Emotional Support Dialog Systems. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3469–3483. Online: Association for Computational Linguistics.

Organization, W. H. 2004. *International Statistical Classification of Diseases and related health problems: Alphabetical index*, volume 3. World Health Organization.

Organization, W. H.; et al. 2018. ICD-11 for mortality and morbidity statistics (2018).

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Qiu, H.; He, H.; Zhang, S.; Li, A.; and Lan, Z. 2024. SMILE: Single-turn to Multi-turn Inclusive Language Expansion via ChatGPT for Mental Health Support. arXiv:2305.00450.

Sun, H.; Lin, Z.; Zheng, C.; Liu, S.; and Huang, M. 2021. PsyQA: A Chinese Dataset for Generating Long Counseling Text for Mental Health Support. arXiv:2106.01702.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tu, T.; Palepu, A.; Schaekermann, M.; Saab, K.; Freyberg, J.; Tanno, R.; Wang, A.; Li, B.; Amin, M.; Tomasev, N.; et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

Wang, J.; Yao, Z.; Yang, Z.; Zhou, H.; Li, R.; Wang, X.; Xu, Y.; and Yu, H. 2024a. NoteChat: A Dataset of Synthetic Doctor-Patient Conversations Conditioned on Clinical Notes. arXiv:2310.15959.

Wang, R.; Milani, S.; Chiu, J. C.; Zhi, J.; Eack, S. M.; Labrum, T.; Murphy, S. M.; Jones, N.; Hardy, K.; Shen, H.; Fang, F.; and Chen, Z. Z. 2024b. PATIENT-$\Psi$: Using Large Language Models to Simulate Patients for Training Mental Health Professionals. arXiv:2405.19660.

WHO; et al. 2022. Mental health and COVID-19: early evidence of the pandemic's impact: scientific brief, 2 March 2022. Technical report, World Health Organization.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang,

P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2024a. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yang, K.; Zhang, T.; Kuang, Z.; Xie, Q.; Huang, J.; and Ananiadou, S. 2024b. MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models. In *Proceedings of the ACM on Web Conference 2024*, volume 35 of *WWW '24*, 4489–4500. ACM.

Yao, B.; Shi, C.; Zou, L.; Dai, L.; Wu, M.; Chen, L.; Wang, Z.; and Yu, K. 2022. D4: a Chinese Dialogue Dataset for Depression-Diagnosis-Oriented Chat. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2438–2459. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Zhang, C.; Li, R.; Tan, M.; Yang, M.; Zhu, J.; Yang, D.; Zhao, J.; Ye, G.; Li, C.; Hu, X.; et al. 2024. CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling. *arXiv preprint arXiv:2405.16433*.