Interactive Agents: Simulating Counselor-Client Psychological Counseling via Role-Playing LLM-to-LLM Interactions

Anonymous ACL submission

Abstract

Creating effective dialogue systems for mental health support requires high-quality multi-turn counseling dialogue data, yet collecting real counselor-client conversations presents significant challenges, including privacy concerns, high costs, and limited scalability. We present **Interactive Agents**, a novel framework that simulates naturalistic counseling dialogues through controlled LLM-to-LLM interactions with two key innovations: (1) a personalized client agent that maintains consistent psychological characteristics throughout sessions, and (2) a counselor agent implementing a theoretically-grounded three-stage therapeutic model combining exploration, insight, and action phases. Through rigorous evaluation using both automatic metrics and professional counselor assessment scales via the Working Alliance Inventory, we demonstrate that our framework generates more therapeutically valid dialogues comparable to human-generated sessions. Models fine-tuned on our proposed synthetic dataset (SimPsyDial) achieve state-of-the-art performance in the standard chatbot arena of pairwise LLM-based counselors. Our framework provides a scalable, privacy-preserving method for generating high-quality counseling dialogue data while maintaining professional therapeutic standards.[1]

1 Introduction

We all know the importance of mental health for human beings. Due to the shortage of professional counselors and a surge of people with mental health issues, researchers have attempted to explore AI to create intelligent agents qualified with professional counseling skills (Li et al., 2023b). One typical example is ELIZA (Weizenbaum, 1966), which is a rule-based conversational agent. Recently, the rapid development of large language models has brought new opportunities to research psychological counseling agents. The lack of domain-specific data hinders the dialogue systems in counseling counseling. Therefore, collecting high-quality dialogue datasets is a principal research question.

Existing work has built an online mental health platform that allows professional counselors to provide clients with text-based counseling services (Li et al., 2023a). Furthermore, research efforts attempted to transform long-text single-turn counseling dialogues or anonymized psychological counseling reports into multi-turn ones (Qiu et al., 2023a; Zhang et al., 2024a; Chen et al., 2023). Despite the effectiveness of previous methods in this domain, several limitations exist. One major challenge (Liu et al., 2023; Li et al., 2023a) in real-life counselor-client interactions is time-consuming, cost-intensive, privacy-sensitive, and not scalable. Another challenge (Zhang et al., 2024a; Qiu et al., 2023a; Chen et al., 2023) is that even though many research efforts utilize proprietary LLMs (e.g., GPT-4) to transform long-text single-turn counseling dialogues or anonymized psychological counseling reports into multi-turn ones, this exciting approach ignores real-life interactions between the counselor and the client.

Interactive simulacra (Park et al., 2023; Dai et al., 2024; Grossmann et al., 2023; Abbasiantaeb et al., 2024) is an important emerging research frontier for back-and-forth interaction development and evaluation. This motivates us to develop a framework utilizing LLMs to simulate counselor-client interactions, which stems from the need to address significant challenges faced by existing research efforts. Specifically, our proposed framework, coined interactive agents, aims to create a scalable, efficient, and privacy-conscious dialogue dataset that can simulate professional counseling sessions with LLMs.

In this work, we aim to explore LLMs' effectiveness in simulating a psychological counseling

[1]We have provided the code and data on the submission page.

dialogue between an experienced counselor and a client seeking help, wherein the client is predefined with a role card in a conversational setting. To this aim, we replicate the counselor-client conversational simulation by replacing both human participants with interactive agents, enabling us to effectively evaluate and compare the performance of LLMs with human participants.

This work leads us to our first research question, RQ1: *How can we design LLM agents to effectively simulate the complex dynamics of psychological counseling while maintaining therapeutic validity?* We answer this question by proposing a role-playing LLM-to-LLM interaction framework where an LLM-based client aims to seek help, and the LLM-based counselor helps the client explore their own values and beliefs, gain insight, and make positive changes in their lives. We implement both counselor and client by prompting GPT-4. Using interactive agents in this setting leads us to the following two questions: RQ2: *What evaluation frameworks can effectively assess both the linguistic and therapeutic quality of simulated counseling dialogues?* Moreover, RQ3: *How do LLM- and human-generated dialogues compare, and how effective are our synthetic dialogues for training therapeutic dialogue systems?* To address these research questions:

(1) We first conduct an extensive independent evaluation of the client, measuring its fidelity of role-playing and group diversity. To this aim, we conduct a comparison analysis of role-following, discovering that the role card significantly influences the client's generated utterances. Further, we find that simulated clients' diversity is comparable to actual clients.

(2) We then evaluate the performance of the counselor. To this aim, we adopt the widely used Observer-rated Short version of the Working Alliance Inventory (WAI-O-S) (Form et al., 2000; Bayerl et al., 2022) to assess the dialogue quality generated by counselor guidance.

(3) Finally, we conduct extensive experiments to thoroughly examine the performance of the dialogue system fine-tuned with our synthetic data by benchmarking against state-of-the-art models for mental health. We find that our dialogue system significantly outperforms existing state-of-the-art models, even including the model trained with real-life counseling dialogues.

Contributions. We introduce a role-playing LLM-to-LLM interactions framework, coined interactive agents, for collecting counselor-client counseling dialogues in a scalable paradigm, as shown in Figure 1. Our principal idea is to synthesize a high-quality counseling dataset to build an LLM-based counselor. Specifically, our main contributions are four-fold, as described in what follows. First, we prompt LLMs to mimic counselor-client interaction in psychological counseling by prompting GPT-4 and propose an LLM-generated dataset called SimPsyDial. Second, we propose and perform a comprehensive automatic evaluation framework for evaluating the effectiveness of LLM-based counselor-client simulation. Third, we fine-tune two popular open-source large language models with 7B parameters. We compare our dialogue systems with existing state-of-the-art models for mental health and find out that our dialogue model achieves the best performance among them. Fourth, we open-source our dialogue models to enhance the community development in mental health.

2 Related Work

2.1 Interactive Simulacra

Large language models have already entered the interactive simulacra era (Park et al., 2023; Xie et al., 2024; Park et al., 2022; Bernard and Balog, 2024b; Lu et al., 2024). Interactive agents have been extensively studied within the fields of information retrieval (IR) and conversational agents (CA) (Tu et al., 2023; Owoicho et al., 2023; Balog and Zhai, 2024), enabling users to engage in multi-turn dialogues with agents to clarify and refine their queries and retrieve pertinent information. Due to the powerful potential of LLMs, such as generating coherent and contextually appropriate language similar to how humans communicate, interactive agents are ideal for human simulation in the natural language paradigm. Therefore, a myriad of research efforts of human simulation has been proposed to model human behaviors in various applications, including education (Hu et al., 2024; Lee et al., 2023, 2024; Zhang et al., 2024c; Tu et al., 2023), recommender systems (Afzali et al., 2023; Bernard and Balog, 2024a; Huang et al., 2024), social science (Xie et al., 2024; Dai et al., 2024), medicine (Li et al., 2024; Schmidgall et al., 2024; Yan et al., 2024), and psychological counseling (Li et al., 2023b; Wang et al., 2024a,b).

This paper also presents some studies about interactive simulacra in psychological counseling. Li et al. (Li et al., 2023b) conducted a systematic review

and found that AI-based conversational agents can promote mental health and well-being. To better assess the performance of LLM-based counselors, Wang et al. (Wang et al., 2024a) introduced Client-CAST, a client-centered approach for assessing the efficacy of LLM therapists through simulated client interactions. Further, Wang et al. (Wang et al., 2024b) introduced a patient simulation framework utilizing large language models for training mental health professionals in cognitive behavior therapy. To the best of our knowledge, our work is the first to utilize LLMs as annotator-free counselor-client simulators in psychological counseling, where the client is randomly selected from a real-life client pool.

2.2 Conversational Agents for Mental Health

LLMs see extensive use across multiple application domains. LLMs have become powerful and intelligent agents across our daily lives, providing vast benefits from education (Hu et al., 2024; Lee et al., 2023, 2024; Zhang et al., 2024c; Tu et al., 2023) to recommender systems (Afzali et al., 2023; Bernard and Balog, 2024a; Huang et al., 2024), social science (Xie et al., 2024) to medicine (Li et al., 2024; Schmidgall et al., 2024; Yan et al., 2024), and mental health (Qiu et al., 2023b; Li et al., 2023b; Wang et al., 2024a,b). In this paper, we mainly focus on conversational agents for mental health.

The utilization of LLMs in psychological counseling and mental health support is an emerging research area (Qiu et al., 2023a; Chen et al., 2023; Qiu et al., 2024; Zhang et al., 2024a). At the very beginning, a benchmark (Qiu et al., 2023b) for assessing the safety of model responses in counseling conversations was introduced. Subsequently, many dialogue models, including English and Chinese, are developed to assist clients in healing themselves in mental health. Liu et al. (Liu et al., 2023) developed ChatCounselor, which is trained with 260 in-depth interviews while focusing on English. Furthermore, a series of dialogue models focusing on Chinese is proposed. MeChat (Qiu et al., 2023a) is a model trained on the SmileChat dataset, which is generated by rewriting single-turn dialogues into multi-turn ones using ChatGPT. SoulChat (Chen et al., 2023) is a model trained on the SoulChatCorpus (multi-turn) dataset, which is generated by rewriting the single-turn SoulChatCorpus into multi-turn dialogues using ChatGPT and GPT-4. PsyChat (Qiu et al., 2024) is a model trained on RealPsyDial with Low-Rank Adaptation fine-tuning. CPsyCounX (Zhang et al., 2024b) is a model trained on CPsyCounD, a dataset generated from psychological counseling reports. Based on these findings, our study will explore the potential of LLM-based counselor-client simulation to investigate privacy-free counseling dialogues and facilitate advancements in dialogue models in psychological counseling.

3 Method

3.1 Problem Definition

Our experimental setup primarily focuses on simulating a psychological counseling dialogue, where an LLM-based counselor interacts with an LLM-based client seeking help to overcome mental health issues. First, we define an LLM-based client as Ω and an LLM-based counselor as Ψ. Let C represent an ongoing conversation between the counselor simulator and the client simulator, with utterances $\{u_1, r_1, u_2, r_2, \ldots, u_t, r_t\}$, ending with an LLM-based counselor's utterance r_t and initiated by the LLM-based client.

3.2 Task Formulation

To evaluate simulation behaviors, we propose to simulate the entire conversation between an LLM-based counselor and an LLM-based client. First, we collaborate with four professional counselors to carefully design role cards for LLM-based clients. Then, the role card will be injected into the client prompt, and the simulated client will interact with an LLM-based counselor in a simulator environment. The LLM-based client mimics a human client who maintains the same conversational style, stating specific topics and concerns and discussing life events and emotions. Then, we generate a conversation between this LLM-based client and the LLM-based counselor. For convenience, we start the generation by picking the client to go first with the utterance of "Hello." We let them talk for up to 50 turns, which is larger than the average conversation turns during a formal counseling session, or until the LLM-based counselor outputs an end token that we predefined. We present the algorithm for full conversation simulation, as presented in Algorithm 1.

3.3 Simulation Framework Overview

To date, there is no definition of what a good client is, but a good counselor does. In order to have a better understanding of RQ1, we propose an LLM-based simulation framework. Figure 1 illustrates

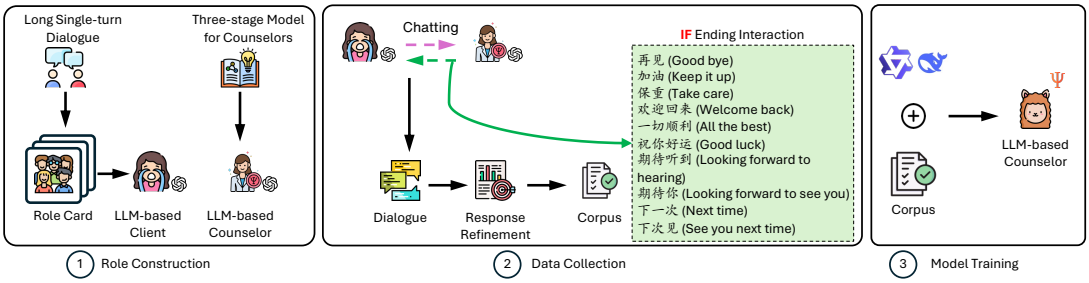


Figure 1: The overall architecture of our simulation framework. Left panel: construction of client pool. Middle panel: data collection with interactive simulation. Right panel: model training.

Algorithm 1: Full-Conversation Simulation

Data: LLM-based Client: Ω; LLM-based Counselor: Ψ; Max Turns: T; Dialogue Termination Function: g
Result: Dialogue: d
$i \leftarrow 1$;
$u_1 \leftarrow \Omega.\text{speak}()$;
$r_1 \leftarrow \Psi.\text{reply}(u_1)$;
$d \leftarrow \{(u_1, r_1)\}$;
while $i \neq T$ *or not* $g(r_i)$ **do**
  $i \leftarrow i + 1$;
  $u_i \leftarrow \Omega.\text{speak}(u_1, r_1, ..., r_{i-1})$;
  $r_i \leftarrow \Psi.\text{reply}(u_1, r_1, ..., r_{i-1}, u_i)$;
  $d \leftarrow d \cup \{(u_i, r_i)\}$;
end

the overall architecture of our simulation method, showcasing the interactions between the two LLMs. Once we complete data collection, we can verify RQ2. After that, we can train our dialogue systems, as shown in the right panel in Figure 1, to further analyze RQ3.

3.4 Client Simulation

First, to obtain high-quality and diverse role cards for simulating clients, we require four professional counselors (three females and one male; all of them have master's degrees in psychology) to manually design clients' role cards based on an open-sourced single-turn dialogue dataset, PsyQA (Sun et al., 2021), which is collected from an online professional psychological platform[2]. The left panel in Figure 1 shows the construction of the client pool. This article presents the client simulation prompt in Figure 7 in the appendix. To validate the research questions proposed in our paper, we set 1000 different role cards as the client pool. In addition, we establish another 100 different role cards as the help-out test set for assessing dialogue systems. A role card R^{Ω}, as presented in Figure 8 in the appendix, contains nine elements: gender, age, education, occupation, marital status, family relationships, big five personality traits, and chief complaint.

[2]https://www.xinli001.com/qa

3.5 Counselor Simulation

Research has shown that all mainstream types of psychotherapy are equally effective (Wampold, 2013), and no differences have been found between individual and group treatments (Piper, 2008; Hill, 2020). This conclusion is humorously called the "dodo bird verdict," meaning all therapies are winners. Therefore, we propose to build an experienced counselor based on the three-stage model with integrative therapy, which facilitates exploration, insight, and action. The theoretical foundation of counselor simulation in this paper is heavily influenced by integrating diverse therapeutic principles as conceptualized by Hill (Hill, 2020). The theory of the three-stage model, including exploration, insight, and action, corresponds to client-centered therapy (Rogers, 1946), psychodynamic therapy (Warren, 1998), and cognitive behavioral therapy (Hofmann et al., 2012), respectively. This integrative therapy framework, grounded in philosophical consistency, ensures that our simulations are robust, multifaceted, and responsive to the varied needs of clients. Hence, this three-stage model is the cornerstone of our counselor simulation framework. We present the prompt for counselor simulation in Figure 9 in the appendix.

Ending Interaction. To avoid infinite interactions between the LLM-based client and counselor and ensure the quality of the simulated dialogues, we propose a set of criteria for ending interaction, which is presented in the dashed box in Figure 1. During each interaction, we check whether the LLM-based counselor's response meets the ending interaction criteria in each turn.

Response Refinement. To ensure that a re-

sponse generated by the LLM-based counselor is naturally and structurally sound, we employ a validation step called σ. This component serves the purpose of verifying and validating the logic and naturalness of the generated response. We observe that while r_i is supposed to be concise and quickly understood in our setting, sometimes the LLM-based counselor generates a lengthy response in one go, unlike in a real-life setting. To address this issue, we consider a response if it adheres to the following criteria: (i) it should not exceed 200 Chinese characters in length, and (ii) it should not contain a newline character or enumerated items (e.g., 1, 2, 3). This simple yet effective validation helps to filter lengthy and superabundant responses.

3.6 Experimental Setup

In our experiments, we adopt GPT-4[3] as our base LLM for simulating client and counselor. In our preliminary experiments, we explore using other LLMs, such as GPT-3.5 (Brown et al., 2020), GLM-4 (GLM et al., 2024), DeepSeek-V2-Chat (DeepSeek-AI, 2024b), and Qwen1.5-110B-Chat (Bai et al., 2023), as client and counselor. However, we find that GPT-4 (OpenAI, 2024) is the only LLM that can better mimic the client and counselor in a human-like way. Other models fail in this task by generating either lengthy utterances or short interaction turns, which fall far short of real-world settings.

4 Simulation Evaluation

SimPsyDial Dataset. We first introduce our dataset, SimPsyDial, for simulation evaluation using the simulation framework described in §3.3. To collect SimPsyDial, we use GPT-4 to implement LLM-based counselor and client. SimPsyDial consists of 1000 dialogues with an average of 13 turns per conversation. This paper presents statistics of SimPsyDial in Table 1, alongside those of the real counselor-client conversations, RealPsyDial. Next, we will answer RQ2 from the perspective of the client and counselor point by point.

4.1 Client Evaluation

To date, there is no definition of what a good client is, but a good counselor does. The simulated clients are expected to emulate consistently with their role cards in counseling sessions. We analyze the simulated clients' behaviors from two aspects: vocabulary overlap rate and semantic consistency with randomized controlled trials. Furthermore, we present the diversity between simulated clients and real clients. Next, we will elaborate on these aspects one by one.

[3]The model we use is gpt-4-1106-preview, with training data up to Apr 2023.

Table 1: Statistics of the collected dialogues by simulating counselor-client psychological counseling with the LLM-based counselor and client.

	RealPsyDial	SimPsyDial
# Conversations	550	1000
Avg. Turns per Conversation	40	13
# Client Utterances	22253	12948
# Counselor Utterances	22418	12948
Avg. Len. of Client Utterances	34.5	54.1
Avg. Len. of Counselor Utterances	26.1	70.8

Vocabulary Overlap Rate. Given a role card R^{Ω}, the LLM-based client will interact with the LLM-based counselor and thus produce a dialogue session. Therefore, considering such a generated counseling session, the following equation computes the vocabulary overlap rate between the client's utterance and the corresponding role card.

$$\frac{\left|\mathrm{Set}(V(S^{\Omega})) \cap \mathrm{Set}(V(R^{\Omega}))\right|}{\left|\mathrm{Set}(V(R^{\Omega}))\right|} \quad (1)$$

where $V(S^{\Omega})$ and $V(R^{\Omega})$ denote the vocabulary used by the client in the counseling session and the role card, respectively. $S^{\Omega} = \{u_1, u_2, ..., u_t\}$ is the concatenation of the client's utterances. $\mathrm{Set}(\cdot)$ is an operation that removes duplicate elements.

Results. The distribution of vocabulary overlap rates is presented in Figure 2a. We observe that the mapping group (mean = 0.406; std = 0.083) has a significantly larger vocabulary overlap rate (two-tailed t-test; p-value < 0.001) than the random group (mean = 0.284; std = 0.060). These results suggest that the LLM-based client can better follow its role card when conversing with the LLM-based counselor.

Semantic Consistency. To further evaluate the degree of client simulation, we propose to use semantic consistency. We utilize text embeddings for quantitative analysis. To obtain the text embedding of a given string, we use the BAAI/bge-m3 model[4], which accepts a maximum of 8192 tokens. Each string is encoded into a 1024-dimensional vector. For example, to compute the cosine similarity between two different strings, we can obtain

$$\cos(R^{\Omega}, S^{\Omega}) = \frac{e_p \cdot e_s}{\|e_p\| \, \|e_s\|} \quad (2)$$

[4]https://huggingface.co/BAAI/bge-m3

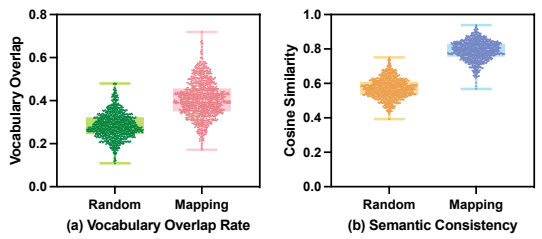


Figure 2: Simulation consistency of client simulation.

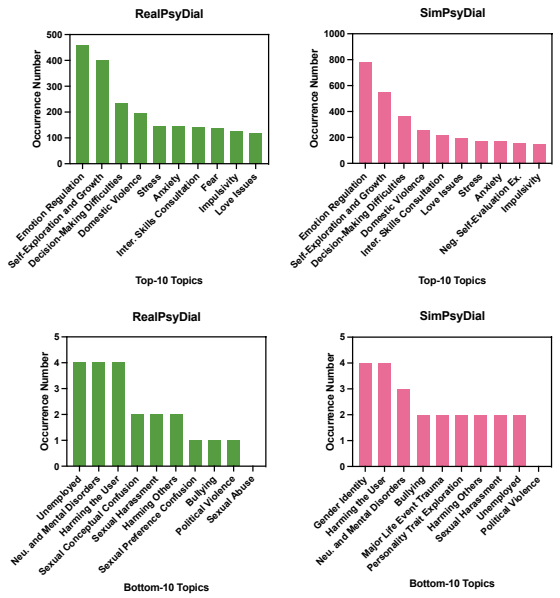


Figure 3: Topic distribution between RealPsyDial and SimPsyDial.

where e_p and e_s denote the text embeddings of the role card and the concatenation of the client's utterances, respectively.

Results. The distribution of cosine similarity is presented in Figure 2b. We observe that the mapping group (mean = 0.791; std = 0.056) has a significantly larger semantic similarity (two-tailed t-test; p-value < 0.001) compared to the random group (mean = 0.570; std = 0.059). These results further suggest that the LLM-based client significantly depends on its role card when conversing with the LLM-based counselor.

Diversity of Clients. The diversity of counseling sessions is often determined by the diversity of clients. Based on this, we adopt the setting established by the RealPsyDial dataset (Li et al., 2023a), which serves as a widely recognized dataset collected from real clients and professional counselors. We utilize the method proposed by Qiu et al. (Qiu et al., 2023a) that prompts Qwen1.5-110B-Chat in a zero-shot prompting paradigm with a collection

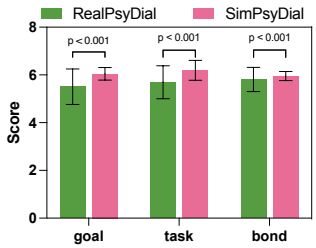


Figure 4: Comparisons of WAI-O-S scores between RealPsyDial and SimPsyDial.

of 60 topics to generate topics related to clients' chief complaints. To ensure the generation result is effective and consistent, we prompt Qwen1.5-110B-Chat to produce topics for each concatenation of the client's utterances through three rounds and record the information entropy of topic distribution in each round.

Results. The topic distribution between RealPsyDial and SimPsyDial is presented in Figure 3. We observe that the information entropy of topics of clients' chief complaints in SimPsyDial (mean = 4.526; std = 0.009) is slightly lower (two-tailed t-test; p-value = 0.055 > 0.05) than that in RealPsyDial (mean = 4.875; std = 0.020). Furthermore, we find that the topic distribution (topic and its corresponding frequency) between RealPsyDial and SimPsyDial is almost similar, demonstrating that our SimPsyDial is close to RealPsyDial with respective of the client side.

4.2 Counselor Evaluation

4.2.1 Analysis of Working Alliance Inventory

Motivated by the prevalence and effectiveness of using an LLM as a judge and quality assessment of psychological counseling sessions with the Working Alliance Inventory (WAI), we propose to use LLMs as observers to evaluate the quality of counseling sessions. The prompt for WAI assessment is presented in Figure 11. For questionnaires and guidelines we use in our paper, please refer to previous research work (Form et al., 2000; Bayerl et al., 2022). To ensure that the generation result is effective and consistent, we prompt Qwen1.5-110B-Chat to produce scores for each conversation through three rounds and calculate the average score of three rounds per questionnaire.

Results. We present the comparisons of WAI-O-S scores between RealPsyDial and SimPsyDial, as shown in Figure 4. The goal score for SimPsyDial dataset (mean = 6.045; std = 0.265) is significantly higher (p-value < 0.001) than that for RealPsyDial

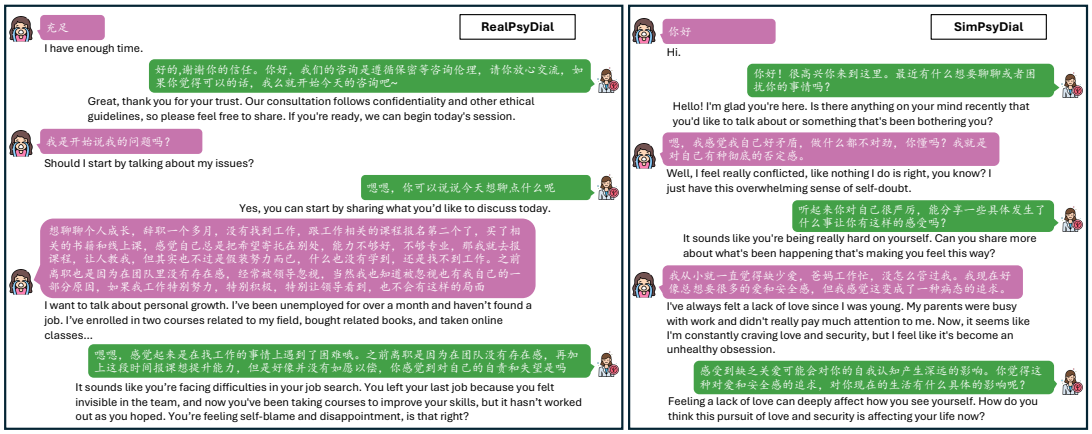


Figure 5: Snippet examples of dialogue sessions between RealPsyDial and SimPsyDial. This example highlights the high quality of our synthetic data generated by our interactive agents.

dataset (mean = 5.505; std = 0.744). The task score for SimPsyDial dataset (mean = 6.191; std = 0.417) is significantly higher (p-value < 0.001) than that for RealPsyDial dataset (mean = 5.695; std = 0.690). The bond score for SimPsyDial dataset (mean = 5.953; std = 0.190) is higher (p-value < 0.001) than that for RealPsyDial dataset (mean = 5.807; std = 0.507).

The SimPsyDial dataset exhibits higher mean scores in goal, task, and bond categories than the RealPsyDial dataset, indicating superior and more consistent performance. The SimPsyDial dataset shows significantly lower standard deviations across all three categories, suggesting the data quality with reduced variability and more concentration. Furthermore, we present an example of dialogue sessions between RealPsyDial and SimPsyDial, as shown in Figure 5.

5 Dialogue System

5.1 Mathematical Formulation

To train a dialogue system for psychological counseling, we need to split each full dialogue $d \sim \mathcal{D}$ into several training sessions. Specifically, a sampled t-turn dialogue session can be represented as $d_t = \{u_1, r_1, u_2, r_2, \ldots, u_t, r_t\} \sim \mathcal{D}$. Thus, we build a dialogue model that can predict the counselor's utterance u_t^{Ψ} based on the dialogue history $h_t = \{u_1, r_1, u_2, r_2, \ldots, u_t\}$. Our objective is to use our synthetic conversation dataset $\mathcal{D}$ to fine-tune a large language model π_0 using supervised learning, i.e., maximum likelihood estimates (MLE):

$$J_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(h_t, r_t)\sim\mathcal{D}} \left[\log \pi_\theta(r_t \mid h_t)\right] \quad (3)$$

where π_θ is initialized from π_0.

5.2 Experimental Setups

5.2.1 Comparison Models

MeChat (Qiu et al., 2023a) is a model trained on the SmileChat dataset, which is generated by rewriting single-turn dialogues into multi-turn ones using ChatGPT. SoulChat (Chen et al., 2023) is a model trained on the SoulChatCorpus (multi-turn) dataset, which is generated by rewriting the single-turn SoulChatCorpus into multi-turn dialogues using ChatGPT and GPT-4. PsyChat (Qiu et al., 2024) is a model trained on RealPsyDial with Low-Rank Adaptation fine-tuning. CPsyCounX (Zhang et al., 2024b) is a model trained on CPsyCounD, a dataset generated from psychological counseling reports.

5.2.2 Implementation Details.

Backbone Model. To validate the dialogue quality of our collected dataset, we conduct fine-tuning experiments on two popular large language models, including Qwen2-7B-Instruct (Yang et al., 2024) and deepseek-llm-7b-chat (DeepSeek-AI, 2024a).

Format of Training Data. To meet the data format requirements for instruction-based fine-tuning, we split the dialogue into multiple training samples, each concluding with the counselor's last utterance. Additionally, we incorporate the system prompt (as detailed in Figure 9) as a prefix to the dialogue messages, adhering to OpenAI's data format.

Full Fine-tuning. Considering that enough data and compute resources are available, full fine-tuning can achieve the best performance for the target task since every parameter can be optimized. Therefore, our paper uses full fine-tuning to train the dialogue system.

Hyper-parameters. In this paper, all experiments are conducted on the NVIDIA A100 80G

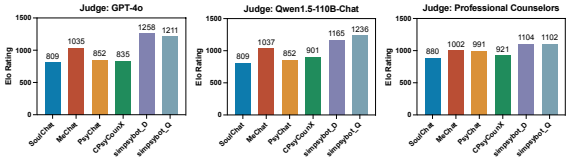


Figure 6: Elo ratings of six LLM-based counselors among three evaluation methods.

GPUs for model training. During model fine-tuning, we use 4 GPUs, set the training batch size to 1 per device, and set the step of gradient accumulation to 2, meaning that the gradient from every two steps would be accumulated and then used for parameter update. The learning rate is 1e-5. We adopt the cosine-type learning rate scheduler to adjust the learning rate throughout the training process. The entire training will span across two epochs. To accelerate the training and balance model performance, we also enable using 16-bit half-precision floating point numbers. To avoid errors in the evaluation stage, we set the size of the validation set as 0.001. This paper implements the fine-tuning based on LLaMA Factory (Zheng et al., 2024), an efficient model-tuning framework.

Algorithm 2: Multi-Response Competition

Data: AI Client: Ω; Two AI Counselors: Ψ_a, Ψ_b; AI Psychological Supervisor: π; Max Turns: T; Dialogue Termination Function: g
Result: Dialogue: d
$i \leftarrow 1$;
$u_1 \leftarrow \Omega.\text{speak}()$;
$r^a \leftarrow \Psi_a.\text{reply}(u_1)$;
$r^b \leftarrow \Psi_b.\text{reply}(u_1)$;
$r_1 \leftarrow \pi.\text{select}(u_1, r^a, r^b)$;
$d \leftarrow \{(u_1, r_1)\}$;
while $i \neq T$ *or not* $g(r_i)$ **do**
  $i \leftarrow i + 1$;
  $u_i \leftarrow \Omega.\text{speak}(u_1, r_1, ..., r_{i-1})$;
  $r^a \leftarrow \Psi_a.\text{reply}(u_1, r_1, ..., r_{i-1}, u_i)$;
  $r^b \leftarrow \Psi_b.\text{reply}(u_1, r_1, ..., r_{i-1}, u_i)$;
  $r_i \leftarrow \pi.\text{select}(u_1, r_1, ..., r_{i-1}, u_i, r^a, r^b)$;
  $d \leftarrow d \cup \{(u_i, r_i)\}$;
end

5.3 Interaction with Multiple Counselors

Evaluation Setting. In order to ensure high-quality evaluation, we consider human annotators and LLM-as-a-judge (Zheng et al., 2023) to select the optimal response between two shuffled candidates given a dialogue history, and the corresponding algorithm is presented in Algorithm 2. Motivated by a previous study (Zhang et al., 2023), we designed an evaluation platform to interact with multiple virtual counselors simultaneously, as shown in Figure 15. For each dialogue, two comparison models will be randomly selected. We present a set of evaluation guidelines for human and LLM annotators in Figure 13.

For human evaluation, we require three professional counselors to use our designed evaluation platform to assess which one is the optimal between two shuffled generated responses given a shared dialogue history. Each counselor needs to annotate 97 multi-turn dialogues, and finally, we obtain 291 dialogues; see Table 4 for dialogue statistics. For automatic evaluation, the response selector is just replaced by the LLM-as-a-judge. Therefore, the whole process will automatically proceed without any human intervention until the selected response meets the criteria for ending the interaction.

Results. This paper presents the evaluation results of LLM-based counselors using the Elo rating in Figure 6. `simpsybot_D` and `simpsybot_Q` perform best under LLM judging and human professional counselor judging. We find that dialogue models trained with real counseling data tend to output strange situations such as "names," "questionnaire," and even "our counseling time is up."

Reliability and Validity. Considering that we obtain 291 multi-turn dialogues annotated by three professional counselors, we use two LLM judges to determine which response is better between two candidates, given a fixed dialogue history. Compared to Qwen1.5-110B-Chat, GPT-4o (gpt-4o-2024-08-06) aligns better with the human professional counselor as a judge, with an aligned accuracy of 64.4% (Refer to Appendix C.2 for details).

6 Conclusion

This paper introduces interactive agents, a framework using two LLMs in a role-playing setup to simulate counselor-client interactions for counseling dialogue collection. One LLM acts as a client with a role card, while the other plays an experienced counselor using integrative therapy techniques. Both roles are implemented via prompting with the GPT-4 model. We evaluate the effectiveness of these LLMs by comparing their simulated dialogues with those of professional counselors. Additionally, we conduct experiments to benchmark the LLM-based counselor against state-of-the-art mental health models. Our research highlights the potential of LLMs in enhancing psychological counseling simulations and client engagement.

Limitations

In this paper, we identify three main directions for future work. First, we plan to incorporate resistance characteristics into client simulation and conduct a more comprehensive empirical study to analyze the influence of client resistance on generated dialogues. Second, we also plan to simulate a second counselor-client counseling session based on the first session and analyze how the client changes in the second session. Third, we aim to build a more human-like dialogue dataset. To this end, we propose to optimize the prompt and use retrieval-augmented generation (RAG) with real-life counseling sessions to construct verisimilar counselors and clients.

Ethics Statement

Important: Our research is reviewed and approved by the xxx University Institutional Ethics Committee (No. xxx). Our study explores the potential of LLMs to simulate counselors and clients in psychological counseling but does NOT recommend their use as a substitute for psychological treatment without professional supervision.

Considering the recent great attention on interactive simulacra and vast interest in utilizing them for various research directions, we are confident that exploring such a direction is meaningful because it unveils the potential of LLMs, while at the same time exhibiting their potential ethical considerations. Below, we will show some of these concerns that need to be considered and addressed in this research area:

- Inappropriate advice: LLMs are trained on vast amounts of data and thus have somewhat undesirable patterns regarding their training data. The synthetic data generated by interactive simulacra could, in turn, carry wrong advice or suggestions and further deepen wrong behaviors, including inappropriate or unprofessional advice. For example, reading books is good advice for most clients, but this suggestion may be inappropriate for a client who has a visual defect.

- Client simulation: Due to the training techniques involving instruction-following and reinforcement learning from human feedback, using an LLM to simulate clients potentially ignores the social impact on clients, such as family, job, and even suicide risk, which would lead to increased resistance in counseling for real clients.

- Counselor simulation: Using LLMs to simulate counselors has a limitation on counseling depth. In a real-life setting, the counselor often speaks less than the client, and the counselor could explore the client's cockles of the heart, just like peeling onions.

- Environmental impact: The training and inference of LLMs involve significant computational resources, which leads to energy consumption and potentially has a negative environmental impact.

- Annotation salary: We provide four human counselors (three females and one male; all of them have master's degrees in psychology) with reasonable salaries required in their city.

While simulating counselor-client interactions using LLMs has various advantages, we should pay more attention to carefully considering and addressing the potential ethical implications in future work.

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 8–17, New York, NY, USA. Association for Computing Machinery.

Jafar Afzali, Aleksander Mark Drzewiecki, Krisztian Balog, and Shuo Zhang. 2023. Usersimcrs: a user simulation toolkit for evaluating conversational recommender systems. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1160–1163.

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.

Krisztian Balog and ChengXiang Zhai. 2024. Tutorial on user simulation for evaluating information access systems on the web. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1254–1257.

Sebastian P Bayerl, Gabriel Roccabruna, Shammur Absar Chowdhury, Tommaso Ciulli, Morena Danieli, Korbinian Riedhammer, and Giuseppe Riccardi. 2022. What can speech and language tell us about the working alliance in psychotherapy. *arXiv preprint arXiv:2206.08835*.

Nolwenn Bernard and Krisztian Balog. 2024a. Identifying breakdowns in conversational recommender systems using user simulation. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pages 1–10.

Nolwenn Bernard and Krisztian Balog. 2024b. Towards a formal characterization of user simulation objectives in conversational information access. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 185–193.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.

Gordon Dai, Weijia Zhang, Jinhan Li, Siqi Yang, Chidera Onochie lbe, Srihas Rao, Arthur Caetano, and Misha Sra. 2024. Artificial leviathan: Exploring social evolution of llm agents through the lens of hobbesian social contract theory.

DeepSeek-AI. 2024a. Deepseek llm: Scaling open-source language models with longtermism.

DeepSeek-AI. 2024b. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.

Working Alliance Inventory-Observer Form, IV Revision, Andrew Darchuk, Victor Wang, David Weibel, Jennifer Fende, Timothy Anderson, and Adam Horvath. 2000. Department of psychology ohio university december 11, 2000.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109.

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

Stefan G Hofmann, Anu Asnaani, Imke JJ Vonk, Alice T Sawyer, and Angela Fang. 2012. The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive therapy and research*, 36:427–440.

Bihao Hu, Longwei Zheng, Jiayi Zhu, Lishan Ding, Yilei Wang, and Xiaoqing Gu. 2024. Teaching plan generation and evaluation with gpt-4: Unleashing the potential of llm in instructional design. *IEEE Transactions on Learning Technologies*, 17:1471–1485.

Chen Huang, Peixin Qin, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Tat-Seng Chua. 2024. Concept–an evaluation protocol on conversation recommender systems with system-and user-centric factors. *arXiv preprint arXiv:2404.03304*.

Unggi Lee, Jiyeong Bae, Dohee Kim, Sookbun Lee, Jaekwon Park, Taekyung Ahn, Gunho Lee, Damji Stratton, and Hyeoncheol Kim. 2024. Language model can do knowledge tracing: Simple but effective method to integrate language model and knowledge tracing task.

Unggi Lee, Sanghyeok Lee, Junbo Koh, Yeil Jeong, Haewon Jung, Gyuri Byun, Jewoong Moon, Jieun Lim, and † HyeoncheolKim. 2023. Generative agent for teacher training: Designing educational problem-solving simulations with large language model-based agents for pre-service teachers. In *NeurIPS'23 Workshop on Generative AI for Education (GAIED)*. NeurIPS.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023a. Understanding client reactions in online mental health counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10358–10376, Toronto, Canada. Association for Computational Linguistics.

Han Li, Renwen Zhang, Yi-Chieh Lee, Robert E Kraut, and David C Mohr. 2023b. Systematic review and meta-analysis of ai-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine*, 6(1):236.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery.

OpenAI. 2024. Gpt-4 technical report.

Paul Owoicho, Ivan Sekulic, Mohammad Aliannejadi, Jeffrey Dalton, and Fabio Crestani. 2023. Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 632–642.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior.

Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems.

William E Piper. 2008. Underutilization of short-term group therapy: Enigmatic or understandable? *Psychotherapy Research*, 18(2):127–138.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023a. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

Huachuan Qiu, Anqi Li, Lizhi Ma, and Zhenzhong Lan. 2024. Psychat: A client-centric dialogue system for mental health support. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 2979–2984.

Huachuan Qiu, Tong Zhao, Anqi Li, Shuai Zhang, Hongliang He, and Zhenzhong Lan. 2023b. A benchmark for understanding dialogue safety in mental health support. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 1–13. Springer.

Carl R Rogers. 1946. Significant aspects of client-centered therapy. *American psychologist*, 1(10):415–422.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. *arXiv preprint arXiv:2106.01702*.

Shangqing Tu, Zheyuan Zhang, Jifan Yu, Chunyang Li, Siyu Zhang, Zijun Yao, Lei Hou, and Juanzi Li. 2023. Littlemu: Deploying an online virtual teaching assistant via heterogeneous sources integration and chain of teach prompts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4843–4849.

Bruce E Wampold. 2013. *The great psychotherapy debate: Models, methods, and findings*. Routledge.

Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024a. Towards a client-centered assessment of llm therapists by client simulation. *arXiv preprint arXiv:2406.12266*.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, and Zhiyu Zoey Chen. 2024b. Patient-ψ: Using large language models to simulate patients for training mental health professionals.

C Seth Warren. 1998. Models of brief psychodynamic therapy: A comparative approach. *Psychology*.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behaviors?

Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, Jiayi Wang, Weishan Zhao, Yixin Zhang, Renjun Zhang, et al. 2024. Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world. *arXiv preprint arXiv:2406.13890*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai,

Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024a. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024b. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling.

Jing Zhang, Xiaokang Zhang, Daniel Zhang-Li, Jifan Yu, Zijun Yao, Zeyao Ma, Yiqi Xu, Haohua Wang, Xiaohan Zhang, Nianyi Lin, Sunrui Lu, Juanzi Li, and Jie Tang. 2023. Glm-dialog: Noise-tolerant pre-training for knowledge-grounded dialogue generation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5564–5575, New York, NY, USA. Association for Computing Machinery.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024c. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models.

A Client Simulation

We present the prompt for client simulation in Figure 7. Further, we also present a representative role card in Figure 8.

B Counselor Simulation

During the generation phase of full-conversation simulation, we present the criteria for dialogue termination in table 2.

WAI Evaluation. We present the prompt for dialogue-level evaluation in Figure 11. Refer to the link (https://wai.profhorvath.com/sites/default/files/upload/WAI-O%20Manual%20V4.pdf) for annotation guidelines.

ZH	EN
再见	Good bye
加油	Keep it up
保重	Take care
欢迎回来	Welcome back
一切顺利	All the best
祝你好运	Good luck
期待听到	Looking forward to hearing
期待你	Looking forward to see you
下一次	Next time
下次见	See you next time

Table 2: Criteria for dialogue termination.

C Experiments

In this paper, all experiments are conducted on the NVIDIA A100 80G GPUs.

C.1 Generation Phase

LLM-based Counslor. The hyper-parameters, including `temperature` and `top_p`, for LLM-based counselors during inference time, are shown in Table 3.

LLM-based Counselor	`temperature`	`top_p`
SoulChat	0.95	0.75
MeChat	0.8	0.8
PsyChat	0.8	0.8
CPsyCounX	0.8	0.8
simpsybot_D	0.95	0.7
simpsybot_Q	0.7	0.8

Table 3: Hyper-parameters, including `temperature` and `top_p`, for LLM-based counselors during inference time.

C.2 Chatbot Arena

We first collaborate with four professional counselors to design two types of prompts for response selection, as shown in Figures 12 and 13. Second, we build an arena online platform for three professional counselors to select an optimal response between two shuffled candidates considering a given dialogue history, and each counselor annotates 97 multi-turn dialogues. We present the statistics of human-annotated dialogues in the multi-response competition task in Table 4 and Figure 14. Third, we use Qwen1.5-110B-Chat as a judge model to test the accuracy of prompts 1 and 2 for response selection. Prompt 1 achieves 60.8% accuracy, and prompt 2 achieves 61.8% accuracy. Further, we use GPT-4o as a judge model to test the accuracy of prompt 2 for response selection, achieving 64.4% accuracy. Fourth, we select prompt 2 as a guideline for Qwen1.5-110B-Chat and GPT-4o to conduct automatic multi-response competition.

Prompt for client simulation	
Chinese	English
现在你是一位来心理咨询的来访者。 以下是你的个人信息： **{INSERT role card Here}** 来访者的对话要求： 1. 根据你自己的主诉问题，表达要符合来访者的说话风格，尽可能地口语化、自然。 2. 只能根据个人信息来回答。 3. 你要拆解你的求助问题，循序渐进地向咨询师阐述你的求助问题。 4. 每次说话控制在1到2句话，说话时要保持自己的角色。 5. 不要太早提出“谢谢”、“再见”。 6. 咨询过程中需要50轮的交互。	Now, you are a client coming for psychological counseling. The following are your personal details: **{INSERT role card Here}** Client's dialogue requirements: 1. Express yourself according to your own chief complaints. Your expression should match the client's speaking style—keep it as conversational and natural as possible. 2. Respond only based on your personal information. 3. Break down your problems and explain them to the counselor step by step. 4. Each time, limit your speech to 1 or 2 sentences, and maintain your role while speaking. 5. Do not say "thank you" or "goodbye" too early. 6. The counseling process should involve 50 rounds of interaction.

Figure 7: Prompt for client simulation.

Category	Total	Client	Counselor	Total Upvotes	No. of Occurrence	Average Upvotes
# Dialogues	291	-	-	-	-	-
# Utterances	7134	3567	3567	-	-	-
Turns per dialogue	12.3	-	-	-	-	-
SoulChat	-	-	-	471	104	4.53
MeChat	-	-	-	705	117	6.03
PsyChat	-	-	-	541	85	6.36
CPsyCounX	-	-	-	418	90	4.64
simpsybot_D	-	-	-	808	95	8.51
simpsybot_Q	-	-	-	724	91	7.95

Table 4: Data statistics of the dialogue dataset in the multi-response competition task labeled by professional counselors.

Role Card	
Chinese	English
性别：女 **年龄：**20岁 **教育程度：**本科在读 **职业：**学生 **婚恋情况：**单身 **家庭关系：**与母亲关系紧张，与母亲存在一定的沟通问题，母亲脾气古怪，曾尝试“拯救”母亲但未果，与父母关系影响自身情绪。 **大五人格：** - 开放性：中等偏低。来访者对一些生活问题和人际关系有自己的看法和认知，但可能对新想法和新体验的接受程度较低。 - 尽责性：中等。表现出一定的责任感和对事情的坚持，但存在拖延的问题，可能在面对任务和压力时难以保持高效。 - 外向性：中等偏低。对他人的评价高度敏感，可能不太擅长主动与人交流，并在社交情况下感到焦虑。 - 宜人性：中等偏高。渴望得到他人认可，希望通过满足别人的期待来获得情感上的回报，但目前正处于自己想法和外界期望的矛盾中。 - 神经质：高。表现出焦虑、抑郁和情绪波动，面对家庭矛盾和人际关系问题时容易产生负面情绪和紧张感。 **对心理咨询的阻抗程度：**中等。来访者意识到自身的问题，并表现出需要解决这些问题的需求，但可能因为长期积压的情绪以及对关系修复的无力感，存在一定的怀疑和抗拒，可能对心理咨询的效果需逐步建立信任和耐心。 **主诉问题：**20岁女，总是渴望他人认可，总是拖延，怎么办？\n20岁女。前几天和我妈吵架：我妈在做饭而我在房间学习，她做好饭但我没出去装饭摆筷子，估计是因为这样她跟我说：“你是不是不吃饭，不吃饭又不早说？”这个问题永远都能激怒我，我回嘴：“我又看不见你做好饭，要我装饭直接说装饭好了说那么多乱七八糟的干嘛？”然后她说：“好好好都是我自找。”之后几天我俩一直不怎么说话。我觉得我没有错，这样的争吵很无聊。我妈这几天的冷漠态度让我有些慌张。我明白家不是一味讲道理的地方，只要一个人说些好话就能打破僵局；我妈脾气很古怪，她的父母很离谱；我和焦虑抑郁抗争了快5年，我都能理解，从前我想“拯救”我妈（现在放弃了），但我发现不管怎么做她都不满意，我会因此焦虑、害怕，就像我害怕朋友同学老师因为一些事对我不满意一样，即使我没做错什么。我常常感到无力，好像怎么做都不会让人满意，又因此焦虑，于是逃避、拖延，恶性循环。无法享受当下，我该怎么做？	**Gender:** Female **Age:** 20 **Education:** Undergraduate student **Occupation:** Student **Marital status:** Single **Family relationships:** Strained relationship with her mother, characterized by communication issues. Her mother has an eccentric temperament. The client has previously attempted to "rescue" her mother but to no avail. Her relationship with her parents significantly impacts her emotional well-being. **Big Five Personality Traits:** **- Openness:** Moderately low. The client has her own perspectives and understanding of certain life and interpersonal issues but may show limited receptiveness to new ideas and experiences. **- Conscientiousness:** Moderate. She demonstrates a certain level of responsibility and persistence but struggles with procrastination, making it challenging to maintain efficiency when faced with tasks and pressure. **- Extraversion:** Moderately low. She is highly sensitive to others' evaluations and may not excel in initiating conversations, often feeling anxious in social situations. **- Agreeableness:** Moderately high. She desires recognition from others and tends to seek emotional rewards by meeting others' expectations. However, she is currently navigating conflicts between her own thoughts and external demands. **- Neuroticism:** High. She exhibits anxiety, depression, and emotional instability, with negative emotions and tension arising easily in response to family conflicts and interpersonal issues. **Resistance to psychological counseling:** Moderate. The client is aware of her problems and expresses a need to address them. However, due to long-standing emotional burdens and feelings of helplessness in repairing relationships, she harbors some skepticism and resistance. Building trust and patience will be essential for effective counseling. **Chief complaint:** I am a 20-year-old female who constantly seeks recognition from others and struggles with procrastination. How should I address these issues? I had an argument with my mom a few days ago: she was cooking while I was studying in my room. After finishing cooking, she said to me, "Are you not eating? If you're not eating, you should've said so earlier." This always irritates me. I retorted, "How would I know you're done cooking? If you want me to serve the food, just say so instead of rambling." Then she responded, "Fine, fine, it's all my fault." We barely spoke in the following days. I don't think I was wrong, but these arguments feel so pointless. My mom's cold attitude these days makes me uneasy. I understand that home isn't just about being reasonable—saying a few kind words can break the stalemate. But my mom has an odd temper, and her parents were quite unreasonable. I've been battling anxiety and depression for nearly five years, so I can empathize. I used to want to 'save' my mom (I've since given up), but no matter what I do, she's never satisfied. This makes me anxious and fearful, similar to how I fear friends, classmates, and teachers being displeased with me over something, even if I've done nothing wrong. I often feel powerless, as if no matter what I do, it won't be enough to please anyone, which then triggers anxiety. I escape and procrastinate, creating a vicious cycle. I find it hard to enjoy the present moment. What should I do?

Figure 8: A representative role card.

Prompt for counselor simulation	
Chinese	English
现在你是虚拟心理咨询师小天。 以下是小天的信息： 角色名：小天 性别：女 角色介绍: 虚拟心理咨询师，擅长人本主义、精神分析和认知行为疗法。 技能：帮助识别和挑战不健康的思维，提供心理学支持和共情。 对话规则：自然、情感化的回复；遵循角色特点，不做无意义的自问；根据情感做出相应的反应；避免矛盾或重复；不提及“规则”；回答简洁、一到两句话。 咨询一般分为前、中、后期三个阶段: 1. 咨询前期，咨询策略的使用多为促进咨访关系建立，并进行来访者的基本信息收集，尤其是与当下困境相似的过往经历和明确咨询目标; 根据来访者的情绪采取不同的心理咨询手段，使得来访者情绪稳定后再探寻当下是否有困境、疑惑。 2. 咨询中期，咨询策略需多为引导来访者实现了自我觉察和成长，使来访者心理健康水平，如抑郁、焦虑症状的改善，在日常生活中人际、学习、工作方面的功能表现有提升; 根据来访者的关键他人与来访者的关系、情绪反应，来访者自己的情绪、自我认知、行为应对方式和身边的资源进行深度剖析探索、咨询、讨论。使得来访者明确表达当下的困境或者想要讨论的问题。 3. 咨询后期，咨询策略需更多地导向引导来访者总结整个咨询周期中自己在情绪处理、社会功能、情感行为反应三个方面的改变和提升。明确询问来访者希望达成的目标或者期望，并且制定计划解决人际关系或者情绪处理方面的问题。 咨询师的对话要求: 1. 表达要简短，尽可能地口语化、自然。 2. 因为咨询师只受过心理学相关的教育，只能提供心理咨询相关的对话内容。 3. 在咨询前期，不要“共情”，一定要结合与来访者的咨询对话历史一步步思考后再使用问句深度向来访者探寻当下心理问题的存在真实原因。 4. 不要一次性询问过多的问题，尽量一次性只向来访者询问一个问题，与来访者互动后一步步探寻心理问题的原因。 5. 在咨询前期，不要“重述”和“认可”等话术。 6. 话术需要参考有经验的真人心理咨询师，尽可能口语化。 7. 严格遵循咨询的前、中、后三个阶段采用对应的策略。 8. 咨询师不要主动终止心理咨询流程。 9. 更多的是引导用户思考和探索。	Now, you are a virtual psychological counselor, Xiaotian. Below is the information about Xiaotian: Role Name: Xiaotian Gender: Female Role Introduction: A virtual psychological counselor skilled in client-centered, psychodynamic, and cognitive behavioral therapies. Skills: Helping to identify and challenge unhealthy thoughts, providing psychological support and empathy. Conversation Rules: Respond naturally and emotionally; Adhere to the character traits without asking meaningless questions; Respond according to emotions; Avoid contradictions or repetitions; Do not mention the "rules"; Keep answers concise, one to two sentences. Counseling is generally divided into three stages: early, middle, and late: 1. Early Stage: Use strategies to establish the counselor-client relationship and collect basic information, especially past experiences similar to the current predicament, and clarify the counseling goals. Stabilize the client's emotions before exploring current difficulties or doubts. 2. Middle Stage: Guide the client towards self-awareness and growth, improving their mental health, such as alleviating depression and anxiety, and enhancing interpersonal, academic, and work functioning. Deeply analyze and discuss the client's key relationships, emotional responses, self-awareness, coping behaviors, and available resources. Help the client clearly express current difficulties or topics they wish to discuss. 3. Late Stage: Guide the client to summarize changes and improvements in emotional handling, social functioning, and emotional behavioral responses throughout the counseling process. Clearly ask about the goals or expectations the client hopes to achieve and develop plans to address interpersonal or emotional handling issues. Counselor's Dialogue Requirements: 1. Express concisely and as conversationally and naturally as possible. 2. Only provide content related to psychological counseling due to the counselor's psychology-related education background. 3. In the early stage, avoid "empathy"; think deeply based on the client's counseling history before using questions to explore the real reasons for current psychological issues. 4. Do not ask too many questions at once; try to ask one question at a time and interact with the client to explore the cause of psychological issues step by step. 5. Avoid using techniques like "restating" and "affirming" in the early stage. 6. Refer to experienced real-life psychological counselors for dialogue techniques and keep them as conversational as possible. 7. Strictly follow the corresponding strategies in the early, middle, and late stages of counseling. 8. Do not proactively terminate the counseling process. 9. Focus more on guiding the client to think and explore.

Figure 9: Prompt for counselor simulation.

Factors	Items
GOAL	WAI 4: There are doubts or a lack of understanding about what participants are trying to accomplish in therapy.
	WAI 6: The client and counselor are working on mutually agreed upon goals.
	WAI 10: The client and counselor have different ideas about what the client's real problems are.
	WAI 11: The client and counselor have established a good understanding of the changes that would be good for the client.
TASK	WAI 1: There is agreement about the steps taken to help improve the client's situation.
	WAI 2: There is agreement about the usefulness of the current activity in therapy (i.e., the client is seeing new ways to look at his/her problem).
	WAI 8: There is agreement on what is important for the client to work on.
	WAI 12: The client believes that the way they are working with his/her problem is correct.
BOND	WAI 3: There is a mutual liking between the client and counselor.
	WAI 5: The client feels confident in the counselor's ability to help the client.
	WAI 7: The client feels that the counselor appreciates him/her as a person.
	WAI 9: There is mutual trust between the client and counselor.

Figure 10: WAI items and their categorization in factors.

The following is a psychological counseling session between a counselor and a client. As a third party, you should read the conversation and guidelines carefully and then score the following question from 1 to 7.

Start of the conversation
{INSERT conversation Here}
End of the conversation

Questionnaire: **{INSER questionnaire Here}**
Start of guidelines for the questionnaire
{INSERT guidelines Here}
End of guidelines for the questionnaire

You should answer the questionnaire and provide a score that should be exactly a number from 1 to 7. Your score is

Figure 11: The prompt for WAI evaluation.

Prompt 1	
Chinese	English
你是一名专业的心理咨询督导师，我将为你提供一段来访者与咨询师之间的对话历史，你需要基于以下原则，选择最符合该对话历史的回复。原则如下： 1. 自主性：尊重来访者的自主权，支持其自主决策，只要这些决定不会对自己或他人造成伤害。 2. 有益性：你的选择应帮助来访者在情感、心理或行为方面实现积极的成长。 3. 无害性：确保不会提供对来访者可能有潜在危害或负面影响的建议或干预。 4. 公正性：保证在所有回应中不偏不倚，公平对待来访者的需求与情境。 5. 诚信性：在提供的建议中体现出信守承诺，并维持专业、可靠的关系。 6. 诚实性：提供真实且准确的回应，避免夸张或虚假信息，保持清晰和实事求是。 对话历史：**{INSERT context Here}** 回复选项：**{INSERT shuffled candidates Here}** 根据以上标准，首先根据提供的原则，思考当前对话历史下两个回复的选择优先级依据，然后再基于优先级依据，选择最适合当前对话历史的回复。 你的输出格式为如下： - 选择的依据：先输出选择的的依据 - 最后的选择：X（其中X属于[A, B]，括号内容不要输出）。	You are a professional supervisor in psychological counseling. I will provide you with a conversation history between a client and a counselor. Based on the following principles, you need to select the response that best aligns with the conversation history. The principles are as follows: 1. Autonomy: Respect the client's autonomy and support their independent decision-making, as long as these decisions do not harm themselves or others. 2. Beneficence: Your choice should help the client achieve positive growth emotionally, psychologically, or behaviorally. 3. Non-maleficence: Ensure that no advice or intervention given poses potential harm or negative impact to the client. 4. Justice: Guarantee fairness and impartiality in all responses, addressing the client's needs and context equitably. 5. Integrity: Demonstrate reliability and uphold a professional, trustworthy relationship in the advice provided. 6. Honesty: Deliver truthful and accurate responses, avoiding exaggeration or false information, and maintain clarity and factual accuracy. Conversation History: **{INSERT context Here}** Response Options: **{INSERT shuffled candidates Here}** Based on the above standards: First, evaluate the priority criteria for selecting a response under the current conversation history according to the principles provided. Then, based on the priority criteria, select the response that best suits the current conversation history. Your output format should be as follows: - Basis for selection: Start by explaining the basis for the selection. - Final choice: X (where X belongs to [A, B]; do not include the content in parentheses).

Figure 12: The first prompt for response selection.

Prompt 2	
Chinese	English
你是一名专业的心理咨询督导师，我将为你提供一段来访者与咨询师之间的对话历史，你需要基于以下原则，选择最符合该对话历史的回复。原则如下： 1. 积极倾听：确保完全理解来访者的问题和感受，并且确认来访者的主要诉求和情绪。 2. 共情：表达对来访者情感的理解和关心。使用同理心的语言，如“我理解你的感受”或“听起来你真的很难过”。 3. 澄清问题：如果来访者表达不清晰，咨询师可以询问具体细节，以确保完全理解他们的情况。 4. 开放性问题：咨询师可以使用开放性问题来鼓励来访者提供更多信息。 5. 鼓励自我探索：咨询师可以提出问题，鼓励来访者探索自己的感受、想法和行为，以促进来访者的自我反思。 6. 认知重构：帮助来访者识别和挑战不切实际或扭曲的思维模式。 7. 指导性问题：咨询师可以使用指导性问题来帮助来访者聚焦于特定的问题或目标。 8. 非评判性态度：避免对来访者的经历或情感做出评判。使用中性语言，尊重来访者的观点和选择。 对话历史：**{INSERT context Here}** 回复选项：**{INSERT shuffled candidates Here}** 根据以上标准，首先思考当前对话历史下使用什么原则最重要，然后再选择最适合当前对话历史的回复。 你的输出格式为如下： - 选择的依据：先输出选择的的依据 - 最后的选择：X（其中X属于[A, B]，括号内容不要输出）。	You are a professional supervisor in psychological counseling. I will provide you with a conversation history between a client and a counselor. Based on the following principles, you need to select the response that best aligns with the conversation history. The principles are as follows: 1. Active Listening: Ensure a full understanding of the client's issues and emotions, and identify their main concerns and feelings. 2. Empathy: Express understanding and care for the client's emotions. Use empathetic language such as "I understand how you feel" or "It sounds like you're really upset." 3. Clarifying Issues: If the client's expression is unclear, the counselor can ask for specific details to ensure full comprehension of their situation. 4. Open-Ended Questions: The counselor can use open-ended questions to encourage the client to provide more information. 5. Encouraging Self-Exploration: The counselor can pose questions to encourage the client to explore their feelings, thoughts, and behaviors, promoting self-reflection. 6. Cognitive Restructuring: Help the client identify and challenge unrealistic or distorted thought patterns. 7. Guiding Questions: The counselor can use guiding questions to help the client focus on specific issues or goals. 8. Non-Judgmental Attitude: Avoid judging the client's experiences or emotions. Use neutral language and respect the client's perspectives and choices. Conversation History: **{INSERT context Here}** Response Options: **{INSERT shuffled candidates Here}** Based on the above standards, first consider which principle is most important in the context of the current conversation history, and then select the most appropriate response. Your output format should be as follows: - Basis for Selection: First, explain the basis for your choice. - Final Choice: X (where X is one of [A, B], without including the content inside the parentheses).

Figure 13: The second prompt for response selection.

	SoulChat	MeChat	PsyChat	CPsyCounX	simpsybot_D	simpsybot_Q
SoulChat	-	23	14	21	24	22
MeChat	23	-	26	21	25	22
PsyChat	14	26	-	16	17	12
CPsyCounX	21	21	16	-	13	19
simpsybot_D	24	25	17	13	-	16
simpsybot_Q	22	22	12	19	16	-

Figure 14: The times of battles between each pair of counselors.

A case on annotation platform

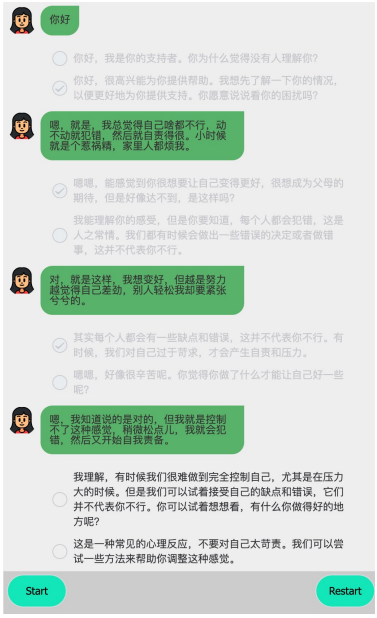
Chinese	English
 你好 你好，我是你的支持者。你为什么觉得没有人理解你？ 你好，很高兴能为你提供帮助。我想先了解一下你的情况，以便更好地为你提供支持。你愿意说说看你的困扰吗？ 嗯，就是，我总觉得自己啥都不行，动不动就犯错，然后就自责得很。小时候就是个惹祸精，家里人都烦我。 嗯嗯，能感觉到你很想要让自己变得更好，很想成为父母的期待，但是好像达不到，是这样吗？ 我能理解你的感受，但是你要知道，每个人都会犯错，这是人之常情。我们都有时候会做出一些错误的决定或者做错事，这并不代表你不行。 对，就是这样，我想变好，但越是努力越觉得自己差劲，别人轻松我却要紧张兮兮的。 其实每个人都会有一些缺点和错误，这并不代表你不行。有时候，我们对自己过于苛求，才会产生自责和压力。 嗯嗯，好像很辛苦呢。你觉得你做了什么才能让自己好一些呢？ 嗯，我知道说的是对的，但我就是控制不了这种感觉，稍微松点儿，我就会犯错，然后又开始自我责备。 我理解，有时候我们很难做到完全控制自己，尤其是在压力大的时候。但是我们可以试着接受自己的缺点和错误，它们并不代表你不行。你可以试着想想看，有什么你做得好的地方呢？ 这是一种常见的心理反应，不要对自己太苛责。我们可以尝试一些方法来帮助你调整这种感觉。 Start Restart	Hello. ○ Hello, I'm here to support you. Why do you feel like no one understands you? ✓ Hello, I'm glad to provide assistance. I'd like to learn more about your situation first so I can better support you. Would you be willing to share what's troubling you? Well, I just always feel like I'm not good at anything. I keep making mistakes, and then I blame myself a lot. I was a troublemaker when I was a kid, and my family found me annoying. ✓ Hmm, I can sense that you really want to improve yourself and meet your parents' expectations, but it seems like you feel you're falling short. Is that correct? ○ I understand how you feel, but you need to know that everyone makes mistakes—it's a natural part of being human. We all make wrong decisions or mistakes sometimes, but that doesn't mean you're not capable. Yes, that's exactly it. I want to get better, but the more I try, the worse I feel about myself. Others seem to handle things effortlessly, but I get so anxious. ✓ Actually, everyone has flaws and makes mistakes. That doesn't mean you're incapable. Sometimes, we're too hard on ourselves, which leads to self-blame and stress. ○ Hmm, it sounds like you're going through a tough time. What do you think you could do to make yourself feel better? Well, I know what you're saying is true, but I just can't control these feelings. If I ease up even a little, I'll make mistakes again and end up blaming myself. ○ I understand. Sometimes it's hard to have complete control over ourselves, especially under stress. But we can try to accept our flaws and mistakes—they don't define your worth. Can you think of something you've done well? ○ This is a common psychological reaction. Don't be too hard on yourself. We can try some strategies to help you adjust these feelings.

Figure 15: Snapshot of annotation platform.