

PsyDial: A Large-scale Long-term Conversational Dataset for Mental Health Support

Huachuan Qiu^{1,2} Zhenzhong Lan^{2*}

¹ Zhejiang University ² School of Engineering, Westlake University
{qiuhuachuan, lanzhenzhong}@westlake.edu.cn

Abstract

Dialogue systems for mental health counseling aim to alleviate client distress and assist individuals in navigating personal challenges. Developing effective conversational agents for psychotherapy requires access to high-quality, real-world, long-term client-counselor interaction data, which is difficult to obtain due to privacy concerns. Although removing personally identifiable information is feasible, this process is labor-intensive. To address these challenges, we propose a novel privacy-preserving data reconstruction method that reconstructs real-world client-counselor dialogues while mitigating privacy concerns. We apply the **RMRR (Retrieve, Mask, Reconstruct, Refine)** method, which facilitates the creation of the privacy-preserving PsyDial dataset, with an average of 37.8 turns per dialogue. Extensive analysis demonstrates that PsyDial effectively reduces privacy risks while maintaining dialogue diversity and conversational exchange. To fairly and reliably evaluate the performance of models fine-tuned on our dataset, we manually collect 101 dialogues from professional counseling books. Experimental results show that models fine-tuned on PsyDial achieve improved psychological counseling performance, outperforming various baseline models. A user study involving counseling experts further reveals that our LLM-based counselor provides higher-quality responses. Code, data, and models are available at <https://github.com/qiuhuachuan/PsyDial>, serving as valuable resources for future advancements in AI psychotherapy.

1 Introduction

According to statistics from the World Health Organization, nearly 1 billion people worldwide suffer from mental health issues, and this number is steadily rising (Wu et al., 2023). Despite the availability of effective treatments, many lack access

to care due to a shortage of qualified therapists (Phillips and Tucker, 2023). Stigma, discrimination, and violations of human rights further prevent individuals from seeking help. The development of dialogue systems (Li et al., 2023b; Zheng et al., 2024a; Qiu et al., 2024a) in mental health counseling shows promise in alleviating distress and helping individuals navigate personal challenges by providing mental health support and fostering self-awareness and healing (Sharma et al., 2023).

A critical factor in developing effective conversational agents for psychotherapy is access to high-quality, real-world, long-term client-counselor interaction datasets (Li et al., 2023a). Extensive efforts to create dialogue datasets, including crowdsourcing (Liu et al., 2021), model augmentation (Zheng et al., 2023a), text transformation (Qiu et al., 2024a; Chen et al., 2023; Zhang et al., 2024), and LLM synthesis (Liu et al., 2023), indicate significant progress has been made. We present several typical datasets used for training dialogue support systems in Table 1. However, these datasets have critical limitations: (1) they are typically sourced from simulations or synthetic data, which do not reflect real-world interactions between clients and counselors (Sohail et al., 2023); (2) they exhibit significant differences in utterance length between clients and counselors, because clients tend to produce longer utterances than counselors in real-world scenarios (Li et al., 2023a); and (3) they often consist of very short dialogues (fewer than 15 turns), which is less than half the length of turns in real-world counseling dialogues (Liu et al., 2021; Qiu et al., 2024a).

However, collecting a large-scale set of real-world client-counselor interaction dialogues with user permission for scientific purposes is labor-intensive. Furthermore, manually anonymizing both explicit and implicit privacy-related content embedded in the dialogues presents a significant challenge (Mandal et al., 2025). Our work is mo-

* Corresponding author.

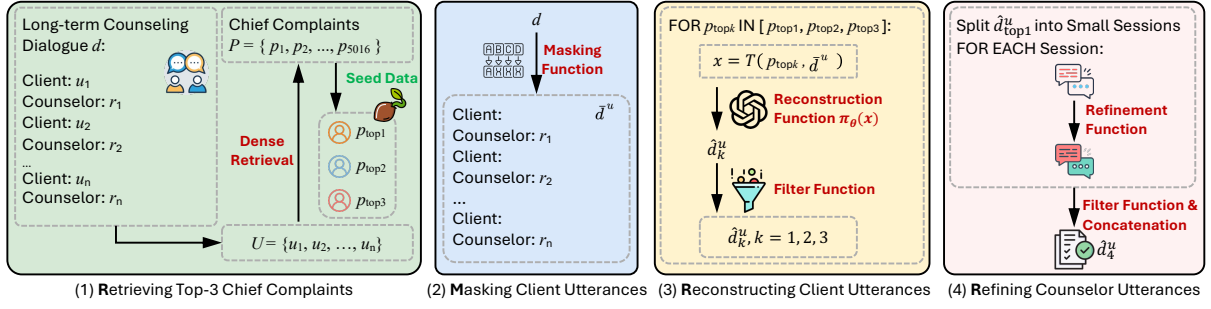


Figure 1: Our method, RMRR, reconstructs counseling dialogue data without leaking user privacy. The following is a description of the terms: $d \in \mathcal{D}_0$, $\hat{d}_1^u \in \mathcal{D}_1$, $\hat{d}_2^u \in \mathcal{D}_2$, $\hat{d}_3^u \in \mathcal{D}_3$, and $\hat{d}_4^u \in \mathcal{D}_4$. For more details, refer to Section 4. For details of the filter function, see Figures 16 and 17 in Appendix D. The former filter function is used for reconstruction, while the latter is used for refinement.

tivated by the need to bridge this data gap and enable the creation of a large-scale, long-term, privacy-preserving dataset that accurately reflects real-world client-counselor dynamics while safeguarding sensitive information.

In this paper, we introduce a novel privacy-preserving data reconstruction method, called **RMRR (Retrieve, Mask, Reconstruct, Refine)**, which facilitates the creation of the PsyDial dataset, comprising 2,382 high-quality, long-term counseling dialogues. This method involves four key steps: (1) retrieving relevant client chief complaints from an open-source dataset; (2) masking client utterances to protect user privacy; (3) reconstructing masked client utterances using a large language model (GPT-4o) guided by the retrieved chief complaints; and (4) refining counselor utterances to maintain dialogue fluency and contextual relevance. Through data analysis, our method successfully balances privacy concerns with the therapeutic value of dialogues, ensuring dialogue diversity. Extensive evaluations demonstrate that models fine-tuned on the PsyDial dataset outperform various baseline models.

This paper makes five main contributions:

1. We release a dataset, \mathcal{D}_{0m} , consisting of large-scale, real-world, long-term counseling dialogues, wherein only the counselor’s part is included, with client utterances masked. The released dataset does not contain any user privacy data, as it has undergone rigorous auditing by expert counselors.
2. We propose a novel method called RMRR (Retrieve, Mask, Reconstruct, Refine, as shown in Figure 1), which facilitates to build a privacy-preserving dataset, PsyDial, with an average of 37.8 turns per dialogue. Extensive analysis demonstrates that PsyDial effectively reduces

privacy risks while maintaining dialogue diversity and conversational exchanges.

3. To evaluate the performance of models fine-tuned on our dataset, we manually collected 101 dialogues from professional counseling books as a test set, \mathcal{D}_{101} .
4. Models fine-tuned on PsyDial show improved psychological counseling performance, outperforming several baseline models on the test set. A user study involving counseling experts further reveals that our LLM-based counselor provides higher-quality responses.
5. We will release the datasets (\mathcal{D}_1 , \mathcal{D}_2 , \mathcal{D}_3 , and \mathcal{D}_4), code, and models to the research community, providing valuable resources for future advancements in AI-driven psychotherapy.

2 Related Work

Recent research in mental health support has increasingly relied on publicly available datasets (Liu et al., 2021; Sun et al., 2021; Chen et al., 2023; Qiu et al., 2024a). Large-scale conversational datasets have enabled researchers to explore various facets of mental health, including predicting suicidal ideation (Qiu et al., 2024c), diagnosing mental health conditions (Srivastava et al., 2022; Yao et al., 2022), understanding client reactions (Li et al., 2023a), forecasting support strategies (Liu et al., 2021; Qiu et al., 2024b; Li et al., 2023a), providing personalized interventions (Balaskas et al., 2024), and assessing response safety within dialogue histories (Qiu et al., 2023). In this paper, we present several typical datasets used for training dialogue support systems, considering both English and Chinese perspectives, which are summarized in Table 1.

English. To directly improve the performance of counseling models, various approaches have been

Dataset	Source	Open-Sourced	Size	Avg. Turns	Language	Avg. Len. of Client	Avg. Len. of Counselor
ESConv (Liu et al., 2021)	Crowdsourcing	✓	1053	14.1	English	15.7	20.2
AugESC (Zheng et al., 2023a)	Synthetic	✓	65k	13.4	English	17.4	19.8
ExTES (Zheng et al., 2024b)	Synthetic	✓	11177	7.5	English	22.5	34.9
Psych8k (Liu et al., 2023)	Synthetic	✓	8187	1	English	52.4	72.4
PsyQA (Sun et al., 2021)	Real-world	✓	56k	1	Chinese	190.5	524.6
SmileChat (Qiu et al., 2024a)	Synthetic	✓	55165	5.7	Chinese	56.9	106.3
SoulChat (Chen et al., 2023)	Synthetic	✓	258353	5.9	Chinese	41.4	90.0
CPsyCounD (Zhang et al., 2024)	Synthetic	✓	3084	8.1	Chinese	32.9	52.6
Xinling (part) (Li et al., 2023a)	Real-world	✗	550	40	Chinese	34.5	26.1
Xinling (full) (Li et al., 2023a)	Real-world	✗	2382	38.7	Chinese	32.7	24.3
PsyDial (Ours, e.g., \mathcal{D}_4)	Semi-real	✓	2382	37.8	Chinese	33.7	31.1

Table 1: Comparisons of existing datasets, including ours, in the realm of mental health.

explored to obtain high-quality dialogues. In the English-speaking domain, a typical example is ESConv (Liu et al., 2021), a crowdsourced emotional support dataset comprising 1,053 dialogues that offer a range of support strategies. However, this data collection method is costly and time-consuming, resulting in a small-scale dataset with limited diversity. To address this challenge, new approaches have been proposed to augment the dataset size using dialogue posts (Zheng et al., 2023a) and self-chats (Zheng et al., 2024b). The Psych8k (Liu et al., 2023) dataset contains about 8,187 single-turn query-answer pairs derived from 260 real-world counseling recordings via the GPT-4 API.

The first four datasets, although limited to English, represent significant progress in the field. *Our research aims to bridge the data gap and advance the development of LLM-based counselors in Chinese, thereby contributing to the growth and flourishing of the research community.*

Chinese. Furthermore, developing automated dialogue systems for psychological counseling in Chinese is of significant importance, as Chinese is the second most widely spoken language in the world. To advance research in this field, the PsyQA dataset (Sun et al., 2021) was created by crawling QA posts from a public mental health support platform. However, single-turn conversations often fall short in addressing mental health issues, which typically require multiple exchanges in real-world interactions. Recent research, such as SoulChat (Chen et al., 2023), SMILE (Qiu et al., 2024a), and CPsyCoun (Zhang et al., 2024), has used LLMs and data transformation techniques to generate multi-turn dialogues from single-turn exchanges or counseling reports. While these methods can produce large-scale dialogues, they still tend to have a limited number of turns. The Xinling dataset (Li et al., 2023a), which includes real-world client-counselor interactions with an average of 38.7 turns, has not been fully open-sourced to the research community

due to user privacy concerns, although a fine-tuned model (Qiu et al., 2024b) is available.

Unlike previous studies, our research aims to pioneer the release of real-world, long-term counseling dialogues using an innovative, privacy-preserving method for data reconstruction.

3 Real-world Counseling Data

This work mainly focuses on creating a privacy-preserving, long-term counseling dataset based on the blueprint of real-world client–counselor dialogues. To achieve this purpose, we use the dataset Xinling, collected by Li et al. (2023a). The Xinling dataset contains a total of 2,382 conversation dialogues, of which 550 are annotated for client reaction analysis in the existing work (Li et al., 2023a). To the best of our knowledge, this represents the most extensive real-world counseling conversation corpus in Chinese. A detailed statistical summary of the dataset is provided in Table 6 in the Appendix. A significant characteristic is that these conversational turns are much longer than those collected through crowdsourcing (38.7 turns compared to 14.1 turns in ESConv (Liu et al., 2021)), demonstrating that in real-world scenarios, conversations between human counselors and clients typically involve more interactions. Furthermore, clients tend to speak in longer utterances than counselors (an average of 32.7 characters compared to 24.3 characters; see Table 1). This dataset emphasizes a client-centered therapeutic approach, prioritizing active listening and empathic understanding rather than directive problem-solving or frequent questioning. This aligns with therapeutic best practices (Hill and O’Brien, 1999; Beck, 2011; Rogers, 1946) that encourage clients to explore their thoughts and emotions, enhance their autonomy, and effectively facilitate changes in mind and behavior.

4 Method

In this section, we will elaborate on the details of our approach, which aims to reconstruct an entirely new dialogue dataset that addresses user privacy concerns.

Consider the dialogue dataset $\mathcal{D}_0 = \{d_i\}_{i=1}^j$, which contains 2,382 dialogues. A sampled, full multi-turn dialogue $d \sim \mathcal{D}_0$ is represented as $d = \{(u_1, r_1), \dots, (u_i, r_i), \dots, (u_n, r_n)\}$, where u_i represents the client’s utterance and r_i represents the counselor’s utterance at the i -th turn.

We observe that user privacy in dialogues is primarily conveyed through client utterances rather than counselor utterances. Inspired by the cloze test, our objective is to fill in the blanks of client utterances, which are masked prior to the reconstruction process. To generate high-quality dialogues, we introduce an innovative framework for constructing a privacy-preserving, long-term dialogue dataset. This framework, as illustrated in Figure 1, comprises four steps: (1) retrieving the top- k open-source client chief complaints based on the entirety of the client’s utterances, (2) masking all client utterances, (3) reconstructing the client utterances using an LLM and the retrieved chief complaints, and (4) refining the counselor’s utterances to maintain contextual relevance, ensure fluency, and uphold high quality. Before delving into the framework, we first introduce some preliminaries.

4.1 Preliminaries

We denote the LLM as π_θ , which generates an output sequence (e.g., a response) y for a given input sequence (e.g., a prompt) x , i.e., $y \sim \pi_\theta(\cdot|x)$. Our objective is to enable π_θ to act as an agent capable of delivering desired responses to various input prompts.

Masking Function. The motivation for fully masking, instead of using other redaction techniques, is that, to the best of our knowledge, there is no Named Entity Recognition (NER) model suitable for counseling data. If the current NER model cannot fully identify sensitive entities, privacy may be compromised. Considering the complexity of privacy issues within the Scope of Confidential Information in Appendix I, fully masking is preferable to redaction prior to data reconstruction. To completely protect user privacy in the original dialogue, masking all user utterances is more optimal than redaction. Therefore, we define a masking function $M : u_i \rightarrow \bar{u}_i$ as follows:

$$M(u_i) = \begin{cases} \bar{u}_i, & \text{if complete masking is applied to } u_i, \\ u_i, & \text{if no masking is applied.} \end{cases} \quad (1)$$

where u_i represents the original utterance spoken by the client at the i -th turn, and \bar{u}_i represents an empty placeholder created by complete masking.

Reconstruction Function. For each dialogue after the masking process, we reconstruct the masked utterances using the reconstruction function: $\hat{u}_i = \pi_\theta(\cdot|x, \hat{u}_{i-1})$, where \hat{u}_i is the i -th element of the utterance sequence after reconstruction, and x represents the available context for reconstruction. The variable $i \geq 1$ denotes the position of the element in the sequence, and \hat{u}_0 is an empty string.

4.2 Reconstructing Counseling Data

In this section, we describe our method, as shown in Figure 1, for reconstructing psychological counseling dialogues based on a blueprint derived from real-world counseling data with masked client utterances. To ensure that user privacy is not compromised in the counselor’s responses, we require two expert counselors to review and revise the counselor’s parts, transforming names and ages into those commonly found on the internet while maintaining the fluency and completeness of the dialogue. Refer to the guidelines in Appendix I.

4.2.1 Chief Complaints Retrieval

Privacy information is often embedded in what the client says during a counseling dialogue. Therefore, the two most critical steps for privacy mitigation are masking the client utterances and subsequently reconstructing them using an LLM (e.g., GPT-4o). However, effectively and appropriately reconstructing the masked client utterances is a non-trivial task. To address this challenge, we propose to reconstruct the client utterances through a cloze test with a retrieved relevant client chief complaint.

The chief complaint is the main reason a client seeks therapy, encompassing their symptoms, difficulties, or challenges. We use the open-source PsyQA dataset, which contains client chief complaints in the form of user-posted questions. To ensure comprehensive data, we retain only those complaints longer than 300 Chinese characters, resulting in 5,016 chief complaints, denoted as $P = \{p_1, p_2, \dots, p_{5016}\}$. We then extract client utterances, excluding role tokens, into a single string $U_c = \{u_1, u_2, \dots, u_n\}$ and use dense retrieval (gte-Qwen2-7B-instruct (Li et al., 2023c)) to

identify the top- k candidate complaints for dialogue reconstruction, where $k = 3$. We hypothesize that the more similar the retrieved complaints are to the client’s original utterance, the better the model performance (Rows 8–10 in Table 3). Thus, the top-3 complaints are $p_{\text{top}1}, p_{\text{top}2}, p_{\text{top}3} = R(U_c, P)$, where R is the retrieval function.

4.2.2 Masking Client Utterances

To fully protect user privacy, we mask all client utterances. The masked dialogue, \bar{d}^u , is then generated by applying M to each u_i using a complete masking strategy:

$$\begin{aligned} \bar{d}^u &= \{(M(u_1), r_1), (M(u_2), r_2), \dots, (M(u_n), r_n))\} \\ &= \{(\bar{u}_1, r_1), (\bar{u}_2, r_2), \dots, (\bar{u}_n, r_n)\} \end{aligned} \quad (2)$$

4.2.3 Reconstructing Client Utterances

To reconstruct counseling dialogues, we use GPT-4o to reconstruct client utterances in a cloze test setting. In collaboration with two expert counselors in psychological counseling, we develop a prompt for this purpose, as shown in Figures 6 and 7 in Appendix A. For each dialogue session with masked client utterances, we apply the reconstruction function π_θ to \bar{d}^u , obtaining \hat{d}_k^u , where $k \in 1, 2, 3$. Specifically, each reconstructed client utterance is generated by $\hat{u}_i = \pi_\theta(x, \hat{u}_{i-1})$. Here, we define $x = T(p_{\text{top}k}, \bar{d}^u)$ as the combination of T , $p_{\text{top}k}$, and \bar{d}^u , where T denotes the prompt template for reconstructing client utterances and $p_{\text{top}k}$ represents the retrieved client chief complaint. The reconstructed dialogue $\hat{d}_k^u = \{(\hat{u}_1, r_1), (\hat{u}_2, r_2), \dots, (\hat{u}_n, r_n)\}$ is then generated by applying the LLM π_θ to each masked client utterance:

$$\begin{aligned} \hat{d}_k^u &= \{(\pi_\theta(x), r_1), (\pi_\theta(x, \hat{u}_1), r_2), \\ &\quad \dots, (\pi_\theta(x, \hat{u}_1, \dots, \hat{u}_{n-1}), r_n)\} \end{aligned} \quad (3)$$

Considering $k = 3$ (see Section 4.2.1), we obtain three datasets, denoted as \mathcal{D}_1 , \mathcal{D}_2 , and \mathcal{D}_3 .

4.2.4 Refining Counselor Utterances

We observe a context-fluency gap in the \mathcal{D}_1 , \mathcal{D}_2 , and \mathcal{D}_3 datasets. To address this issue, we refine the counselor’s utterances in these three datasets. Considering that we proposed the aforementioned hypothesis and conducted a preliminary manual check on the three datasets (\mathcal{D}_1 , \mathcal{D}_2 , and \mathcal{D}_3), we found that the dialogues in these datasets are comparable in terms of quality. To balance cost and efficiency,

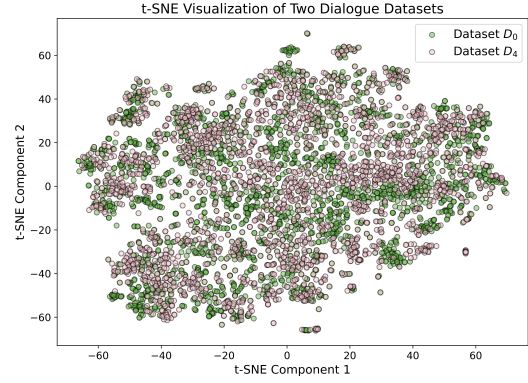


Figure 2: t-SNE visualization of two dialogue datasets, \mathcal{D}_0 and \mathcal{D}_4 .

we used only \mathcal{D}_1 to refine the counselor’s utterances. Consequently, we obtained a new dataset named \mathcal{D}_4 .

5 Data Analysis

5.1 Data Statistics

In this section, we present the data statistics for five datasets, as shown in Table 6 in Appendix N. The datasets \mathcal{D}_0 , \mathcal{D}_1 , \mathcal{D}_2 , and \mathcal{D}_3 are regarded as baseline datasets, while dataset \mathcal{D}_4 is considered the optimal dataset obtained through our method.

5.2 Privacy Elimination

Evaluating user privacy in counseling dialogues is challenging, as privacy is often implicit in client utterances and requires careful reasoning to identify. In this section, we focus on privacy elimination through comparative analysis. We collaborated with experts to verify that the dataset \mathcal{D}_4 contains no user privacy after reconstruction, compared to the dataset \mathcal{D}_0 . We randomly sampled 500 dialogue pairs from both datasets and had experts check for any privacy content from \mathcal{D}_0 in \mathcal{D}_4 , following our guidelines (see Appendix I). Human validation reports confirmed that none of the sampled dialogues in \mathcal{D}_4 contained user privacy from \mathcal{D}_0 , demonstrating the effectiveness of our method. The t-SNE visualizations (van der Maaten and Hinton, 2008) are presented in Figures 28 and 29 in Appendix R.

5.3 Dialogue Diversity

To demonstrate the diversity of the reconstructed dataset, we employ the state-of-the-art embedding extraction model gte-Qwen2-7B-instruct for feature extraction. The t-SNE visualization of the two dialogue datasets, \mathcal{D}_0 and \mathcal{D}_4 , is shown in Figure 2. As can be seen, the data points from both datasets are evenly distributed, indicating that the

data generated using our method retains its diversity. For additional data visualizations, please refer to Appendix S.

Further, to comprehensively evaluate vocabulary usage in a dialogue corpus, we propose a composite metric termed **Lexical Diversity Density (LDD)**. This metric is defined as the product of **Proportion of Unique Words** and **Unique Words Per Dialogue**. Formally,

$$\begin{aligned} \text{LDD} &= \left(\frac{\text{Total Unique Words}}{\text{Total Words}} \right) \times \left(\frac{\text{Total Unique Words}}{\text{Number of Dialogues}} \right) \\ &= \frac{(\text{Total Unique Words})^2}{\text{Total Words} \times \text{Number of Dialogues}} \end{aligned} \quad (4)$$

This formulation captures not only the overall richness of the vocabulary used (i.e., how many distinct words appear relative to total word count), but also how those words are distributed across individual dialogues. Relying solely on **Proportion of Unique Words** may lead to misleading conclusions, as this metric can be artificially inflated in short texts or dominated by outliers with high repetition. Conversely, **Unique Words Per Dialogue** overlooks intra-dialogue repetition and may be skewed by varying dialogue lengths. By combining these two dimensions, **LDD** offers a more robust and discriminative measure of lexical diversity, penalizing both excessive repetition and overly narrow usage of vocabulary. This metric is well-suited for assessing whether clients or counselors in psychological counseling dialogues use a diverse and widely distributed vocabulary, helping to avoid mechanical or repetitive language. Therefore, we present the statistics of LDD (see Table 2) from the perspectives of the client and counselor, respectively, to provide a quantitative measure of the dataset’s diversity beyond visual inspection. We find that our dataset significantly outperforms existing open-source counseling datasets in terms of the LDD, indicating that our dataset is of high quality. The word segmentation tool is Jieba¹.

5.4 Dialogue Topic Distribution

We use the automatic topic annotation method proposed in previous work (Qiu et al., 2024a), which presents 60 types of dialogue topics and their corresponding topic annotation prompts. To protect data privacy, we use the LLM "Qwen/Qwen2.5-72B-Instruct" (Team, 2024) to annotate dialogue topics on private GPUs locally.

For brevity, we present the distribution of dialogue topics within the dialogues in Table 4 in Appendix H. The total number of topics in dataset \mathcal{D}_0 is 18,693, while the total in dataset \mathcal{D}_4 is 17,950, representing a decrease of only 743 topics. We find that dataset \mathcal{D}_4 contains comparably rich dialogue topics to dataset \mathcal{D}_0 , highlighting that \mathcal{D}_4 maintains topic diversity.

6 Experiments

6.1 Experimental Setups

6.1.1 Backbone Models

To fully understand the benefits of our proposed method and highlight our contributions, we train two open-source LLMs: Qwen2.5-7B-Instruct (Yang et al., 2024; Team, 2024) and deepseek-llm-7b-chat (DeepSeek-AI et al., 2024). In the main body of this paper, we present the results obtained by fine-tuning Qwen2.5-7B-Instruct, while the results obtained by fine-tuning deepseek-llm-7b-chat are placed in Appendix Q.

6.1.2 Model Training

To validate the effectiveness of our collected dataset, we directly adopt the supervised fine-tuning (SFT) paradigm to train LLM-based counselors, using the loss function shown in Equation 5. This stage involves training the model to generate the counselor’s responses based on the dialogue history. The implementation details are provided in Appendix F.

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\sum_{t=1}^m \log \pi_{\theta}(y_t | y_{<t}, x) \right] \quad (5)$$

6.1.3 Baselines

We evaluate our approach against three types of baseline models: (1) **Golden Response**, which is considered a highly persuasive baseline in this paper; (2) **Counseling Models**, including SoulChat, MeChat, PsyChat, and CPsyCounX. These models are Chinese LLM-based counselors designed for mental health support, and we use the inference code available on Hugging Face; and (3) **Fine-tuned Models**, which include the models fine-tuned on the datasets \mathcal{D}_0 , \mathcal{D}_1 , \mathcal{D}_2 , \mathcal{D}_3 , and \mathcal{D}_4 , corresponding to the models π_0 , π_1 , π_2 , π_3 , and π_4 .

¹<https://github.com/fxsjy/jieba>

Dataset	Client					Counselor				
	Total Unique Words	Total Words	Proportion of Unique Words (%)	Unique Words Per Dialogue	LDD	Total Unique Words	Total Words	Proportion of Unique Words (%)	Unique Words Per Dialogue	LDD
SmileChat	50537	11734921	0.43	0.92	0.3956	71297	21022978	0.34	1.29	0.4386
SoulChat	89572	41396709	0.22	0.35	0.0770	95152	86625643	0.11	0.37	0.0407
CpsyCounD	10418	541736	1.92	3.38	6.4896	9246	826879	1.12	3.00	3.3600
\mathcal{D}_0	41231	2001910	2.06	17.31	35.6586	25100	1505603	1.67	10.54	17.6018
\mathcal{D}_1	21771	1991034	1.09	9.14	9.9626	24744	1514125	1.63	10.39	16.9357
\mathcal{D}_2	21890	1982515	1.10	9.19	10.1090	24675	1514374	1.63	10.36	16.8868
\mathcal{D}_3	21836	1983623	1.10	9.17	10.0870	24714	1513405	1.63	10.37	16.9031
\mathcal{D}_4 (PsyDial)	21723	1997853	<u>1.09</u>	<u>9.12</u>	<u>9.9408</u>	16421	1868782	<u>0.88</u>	<u>6.89</u>	<u>6.0632</u>

Table 2: Statistics of lexical diversity density. The metric LDD denotes lexical diversity density, which is defined in Equation 4. The best results are highlighted in bold, and the results of PsyDial are underlined.

6.1.4 Evaluation Set

To objectively and fairly evaluate our SFT model in comparison to other baseline models, we use dialogues collected from professional psychological counseling books (see Appendix M). We manually collected 101 dialogues from nine well-known psychological counseling books. To ensure the accuracy of the collected dialogues, we employed a two-stage process, including manual transcription and verification. Finally, these collected dialogues serve as our test set for evaluating our models. For detailed data statistics, refer to Appendix O.

6.1.5 Evaluations

We evaluate the mental health counseling capabilities of various models across two settings: (1) counseling skills assessment and (2) pairwise response selection.

Counseling Skills Assessment. This evaluation focuses on nine key aspects: (1) Fine-grained Skills: We assess a model’s proficiency in eight fine-grained components, informed by prior research (Zhang et al., 2024; Zheng et al., 2024b; Qiu et al., 2024a; Zhao et al., 2024). These components include Empathy, Active Listening, Issue Clarification, Open-ended Questioning, Encouraging Self-Exploration, Cognitive Restructuring, Guided Questioning, and Non-judgmental and Accepting Attitude. (2) Overall Score: This score provides a high-level measure of counseling ability by combining the eight components.

Each criterion is evaluated on a five-point scale using GPT-4o. Detailed descriptions of the metrics are provided in Appendix J.

Pairwise Response Selection. Drawing on prior research, we use a chatbot arena (Zheng et al., 2023b), where, given a dialogue history and two shuffled candidate responses, experts determine which response is more appropriate based on pre-defined selection guidelines (see Appendix K). For all cases of pairwise response selection, we first

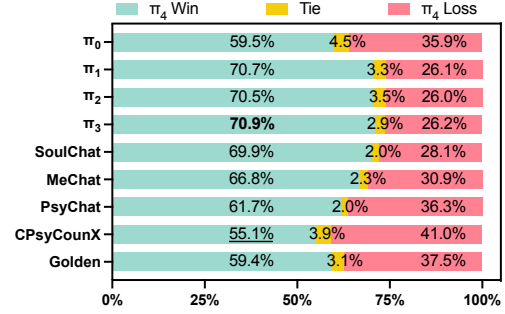


Figure 3: Results of pairwise response evaluation. Our model π_4 outperforms various baselines. The best and worst win rates are highlighted in bold and underlined.

use majority voting to resolve label disagreements. If all three annotators assign different labels, they must discuss the instance to adjudicate the disagreement and determine the final label. We then visualize the model’s win rate relative to baseline models.

6.2 Main Results

6.2.1 Counseling Skills Analysis

The performance comparisons of counseling skills between our model and the baseline models are presented in Table 3. To validate the reliability and effectiveness of our GPT-based evaluation of counseling skills, we calculate the Spearman rank correlation coefficient (Sedgwick, 2014) with ratings from licensed counselors. All metrics show positive correlations (ranging from 0.412 to 0.603), most of which are statistically significant. Further details are provided in Appendix L.

Our SFT model, π_4 (Row 11), outperforms both existing counseling models and fine-tuned baseline models across most metrics of counseling skills, such as Empathy, Active Listening, Open-ended Questioning, Encouraging Self-exploration, Non-judgmental and Accepting Attitude, and Overall Score. These improvements highlight the practicality of our model π_4 and underscore the value of the PsyDial dataset created by our method.

Our model demonstrates significantly stronger

Category	Models	Empathy \uparrow	Active Listening \uparrow	Issue Clarification \uparrow	Open-ended Questioning \uparrow	Encouraging Self-Exploration \uparrow	Cognitive Restructuring \uparrow	Guided Questioning \uparrow	Non-judgmental and Accepting Attitude \uparrow	OA \uparrow
Golden	Counselor Response	3.18	3.80	3.49	3.29	3.54	2.75	3.79	4.48**	3.42*
Counseling Models	SoulChat	3.16*	3.64**	3.24**	3.20**	3.31*	2.64	3.41**	4.37**	3.32**
	MeChat	3.22	3.68**	3.27**	3.34	3.37	2.68	3.56	4.43**	3.37**
	PsyChat	3.33	3.78*	3.28**	3.37	3.47	2.70	3.61	4.50**	3.41*
	CPsyCounX	3.26	3.84	3.54	3.44	3.48	2.78	3.79	4.49**	3.49
Fine-tuned Models	π_0	3.21	3.75*	3.45	3.41	3.46	2.59	3.63	4.58**	3.38**
	π_1	3.10**	3.57**	3.22**	3.32	3.33*	2.51	3.44*	4.55**	3.22**
	π_2	3.06**	3.52**	3.16**	3.23**	3.32*	2.48*	3.36**	4.52**	3.22**
	π_3	3.08**	3.47**	3.11**	3.16**	3.30*	2.50	3.31**	4.45**	3.20**
	π_4	3.27	3.95	3.49	3.45	3.48	2.64	3.63	4.80	3.55

Table 3: Comparison results among different baselines compared to our SFT model π_4 (backbone model: Qwen2.5-7B-Instruct). Statistically significant differences compared to our model π_4 are marked with ** (p -value < 0.05), * (p -value < 0.1), as determined by the Mann-Whitney U test. The best and second results are highlighted in bold and underlined.

<p>f'''(Ours)</p> <p>Client: On Valentine's Day, I was busy writing my thesis, and my girlfriend complained that I wasn't paying attention to her. We ended up arguing.</p> <p>Counselor: It seems like you were very busy at the time, and your girlfriend may have felt neglected. Is that right?</p> <p>Client: Yes. It's not easy for me either, you know? I wake up early and work late every day. I'm really exhausted.</p> <p>Counselor: I understand how tough that must be. That time must have been especially hard for you.</p> <p>Client: The pressure is huge for PhD students. I have to focus on publishing papers. It's very competitive now, and I often can't sleep at night</p> <p>↪ because I keep thinking about my thesis.</p> <p>Counselor: It sounds like you're under a lot of stress, and the lack of sleep makes it even harder to recharge.</p> <p>Client: You really understand me. With the pressure from my family and supervisor, I still need to find time to spend with my girlfriend. There's just</p> <p>↪ never enough time.</p> <p>Counselor: It seems like you're juggling a lot of pressure from different areas, and time feels really tight.</p> <p>Client: Exactly, that's why I didn't care about Valentine's Day. I'm too busy trying to finish my thesis, so there's no room for expectations.</p> <p>Counselor: I can understand your feelings. The pressure from your thesis really makes it hard to enjoy any holidays.</p> <p>Client: I just want to get my thesis done and submitted. Do you think that's the right approach?</p> <p>Counselor: That is important; completing your thesis is your primary task.</p> <p>Client: But the conflict with my girlfriend is also confusing me. I don't know how to balance it-writing my thesis and also caring for her feelings.</p> <p>↪ It feels like an impossible task.</p> <p>Counselor: It sounds like you're really struggling between these two priorities. Is that right?</p> <p>Client: Yes, I feel lost and torn. Sometimes when I feel really down, I go to the gym and run like crazy to clear my mind. The only thing I focus on is</p> <p>↪ running until I'm exhausted, just to forget about the emotional struggle.</p> <p>Counselor: Using running to relax seems like a good strategy. Do you find that it helps you feel better?</p> <p>(CPsyCounX)</p> <p>Client: On Valentine's Day, I was busy writing my thesis, and my girlfriend complained that I wasn't showing her enough attention, and we had a fight.</p> <p>Counselor: It sounds like your busyness led you to overlook your girlfriend, which made her feel neglected. Can you share how that made you feel?</p> <p>Client: It's not easy trying to graduate. I work early in the morning and late at night. I'm really tired.</p> <p>Counselor: I understand how difficult that is. But it's also important to acknowledge how your girlfriend feels. How do you think you could balance</p> <p>↪ your studies with your relationship?</p> <p>Client: I hope we can find a solution where I can meet my thesis deadlines and still spend time with her on Valentine's Day.</p> <p>Counselor: It sounds like you care a lot about your relationship. Let's explore how we might make that work. What kind of activities would you ideally</p> <p>↪ like for Valentine's Day?</p> <p>Client: I have no expectations for Valentine's Day. My thesis is unfinished-why would I expect anything?</p> <p>Counselor: I can feel the pressure you're under, but perhaps we can shift your mindset a bit. Even amidst your workload, try to carve out some time for</p> <p>↪ both yourself and your girlfriend, so you can enjoy a pleasant Valentine's Day together.</p> <p>Client: Alright, I'll try to adjust my mindset and make an effort not to work on my thesis during important holidays. I'll try to spend more time with</p> <p>↪ her.</p> <p>Counselor: That's a great attitude. Now, let's work together to come up with a plan that helps you balance your academic commitments while also making</p> <p>↪ time for your relationship. How does that sound?'''</p>
--

Figure 4: Comparative example dialogues. For the Chinese version, see Figure 23 in Appendix Q.

active listening abilities compared to other baseline models, surpassing the golden baseline by 0.15. The results indicate that responses within the original dialogue history already exhibit a high level of active listening, and our model surpasses the performance of the gold responses. This suggests that in real-world psychological counseling, therapists frequently engage in active listening by using phrases such as "Mm-hmm" or "I see, go on," which involves fully concentrating, understanding, responding to, and remembering what the client says.

Regarding issue clarification, cognitive restructuring, and guided questioning, our model also consistently outperforms other baseline models, though it falls slightly short of CPsyCounX. This discrepancy may be attributed to the fact that CPsyCounX was trained on dialogues derived from psy-

chotherapy notes, which could enhance its ability to structure and clarify issues. Furthermore, the average number of sentences in CPsyCounX's responses is 3.5 (compared to an average of 2.0 sentences in golden responses), and it consistently employs questioning techniques in its replies. However, the difference in performance between our model and CPsyCounX is not statistically significant.

With respect to cognitive restructuring, all models score below 3, indicating that techniques for challenging a client's thought patterns are less frequently employed in psychotherapy. Comparing the scores of the gold responses and our model, we can conclude that a score of 3 represents an optimal level of cognitive restructuring. Excessive use of cognitive restructuring may weaken the therapeutic relationship, potentially hindering the development

of a strong therapeutic alliance (Wilmots et al., 2020).

Our model fosters a client-friendly therapeutic environment that is non-judgmental and accepting. It creates a safe and supportive space where clients feel comfortable sharing their thoughts and emotions without fear of criticism or negative judgment. The counselor maintains a neutral and respectful stance, accepting the client’s experiences and emotional expressions without imposing personal values or opinions, especially for AI counselors that provide suggestions.

6.2.2 Win Rate Analysis

We present the results of the pairwise response evaluation in Figure 3. To validate the reliability and effectiveness of our GPT-based evaluation of win rate (see Appendix L), we assess its agreement with judgments from licensed counselors. Cohen’s κ (Cohen, 1960; Artstein and Poesio, 2008) is used to measure inter-rater agreement, yielding a value of 0.67, which falls within the range of substantial agreement (0.60–0.80).

Our SFT model, π_4 , consistently outperforms all baseline models, highlighting the superiority of our dataset. First, π_4 achieves a win rate exceeding 60% against three counseling models: SoulChat, MeChat, and PsyChat. Second, it surpasses three baseline models fine-tuned with initial reconstruction datasets, \mathcal{D}_1 , \mathcal{D}_2 , and \mathcal{D}_3 . Third, the π_4 model achieves nearly equal win rates of 59.5% and 59.4% when competing against the π_0 and the golden response, respectively. This also confirms that our initial data is of high quality. Lastly, while our model achieves the lowest win rate (55.1%) against CPsyCounX, it remains competitive overall.

6.2.3 Expert Evaluation

In this study, a licensed counselor role-plays as a client to interact with two different models independently. To this end, we first require two licensed counselors to portray 60 distinct client cases (Qiu et al., 2024a), allowing for a qualitative analysis of our model and several baseline models. *We conclude that our model is more suited for long-term counseling.*

For example, the primary concern expressed by the client is the difficulty of balancing the demands of his doctoral dissertation work with the need to attend to his girlfriend’s emotional needs, which has led to feelings of helplessness and distress. We provide two comparative example dialogues, as shown

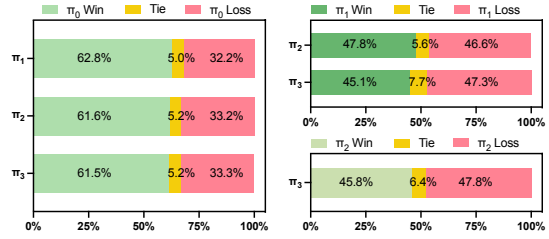


Figure 5: Results of pairwise response evaluation among models π_1 , π_2 , and π_3 .

in Figure 4. Our observations suggest that our model demonstrates superior active listening skills and offers more encouraging responses, whereas CPsyCounX tends to rely on mechanical empathy and frequent question-asking. These results highlight that our model is superior to existing counseling models because active listening better enables clients to express their innermost feelings from the perspective of client-centered therapy.

6.2.4 Ablation Study for Data Refinement

We present the ablation study for data refinement from two perspectives: counseling skills and win rate.

Counseling Skills. π_0 outperforms models π_1 , π_2 , and π_3 in all eight counseling skill components, as well as in the overall score. In contrast, while our model, π_4 , exceeds these four models in both the individual counseling skill components and the overall score, its performance further underscores the critical role of data refinement.

Win Rate. π_0 clearly outperforms models π_1 , π_2 , and π_3 in terms of win rate, as shown in Figure 5, with pairwise comparisons among π_1 , π_2 , and π_3 revealing comparable performance, indicating that these three models are essentially neck and neck. In stark contrast, π_4 stands out by significantly surpassing all four models, further emphasizing the critical impact of data refinement, as illustrated in Figure 3.

7 Conclusions

In conclusion, this paper presents a novel, privacy-preserving data reconstruction method, contributing to the development of PsyDial, a large-scale, semi-real client-counselor dialogue dataset. Extensive experiments demonstrate that the dataset is both privacy-preserving and of high quality. A user study conducted with counseling experts underscores its effectiveness in long-term counseling contexts. This work lays a strong foundation for future research in AI psychological counseling.

Limitations

In this work, we focus on leveraging LLMs to reconstruct text-based counseling conversations to protect user privacy in dialogues. Text-based counseling is a widely used and accessible form of therapy, and it is particularly well-suited for LLMs, which can precisely generate counselor responses conditioned with a dialogue history. In the future, a promising direction for online counseling across other formats involves integrating multi-modal models that combine video counseling dialogues.

Our data is sourced from a single counseling platform in China. While the demographic diversity of both clients and counselors in these conversations contributes to a degree of diversity, future research could benefit from expanding the data collection to multiple counseling platforms across different regions. This would allow for a more comprehensive data collection.

One potential direction for future work is to use the LLM judge mentioned in the paper to construct large-scale, high-quality preference data. This could be achieved by generating candidate responses with two candidate models, and then using GPT-4o as a preference selector. We believe this direction can significantly enhance the model performance.

Despite the promising potential of LLM judges, our findings indicate that they tend to favor longer responses and assign higher scores, which introduces a bias that must be addressed to ensure their scores are consistent with those of human experts. *How to address biased scores for long responses using LLM judges is an extremely important direction for future research.*

Our conversation system can be used for peer counselor training, where two novice counselors take on different roles, one acting as the client and the other as the counselor. The counselor can learn professional response suggestions from the system, which can help enhance their skills. *In future human-in-the-loop experiments, this setup not only improves the novice counselors' expertise but also provides human feedback to the system, allowing for further refinement of the virtual counseling system.*

Ethical Considerations

Data Privacy. Our research is reviewed and approved by the Westlake University Institutional

Ethics Committee (20211013LZZ001). Following the informed consent, we rigorously anonymized the counselors' dialogues to ensure the highest level of privacy protection for both clients and counselors. Furthermore, only dialogues with masked client utterances were sent to the proprietary API (GPT-4o) to ensure compliance with our platform's informed consent guidelines.

Data Release. To promote interdisciplinary research at the intersection of Natural Language Processing (NLP) and psychological counseling, we plan to release these datasets to the research community upon acceptance of this article. Additionally, the dataset \mathcal{D}_{0m} , in which all client utterances are fully masked, is also released for researchers interested in this field.

Application. This study proposes an automatic approach for reconstructing long-term counseling dialogues while safeguarding user privacy. Additionally, we fine-tune several LLM-based counselors to validate the effectiveness of our dataset. We advocate for the integration of LLM-based counselors as supplementary tools to assist human counselors in providing text-based counseling to their clients.

Annotator Salary. We compensated two licensed counselors with a competitive salary, ensuring that the pay scale aligns with industry standards in our city.

LLM Usage. Throughout this paper, we use LLM for grammar checks.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. This work was supported by the Scientific Research Project of Westlake University (Grant No. WU2024B003).

References

- Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.
- Andreas Balaskas, Stephen M. Schueller, Kevin Doherty, Anna L. Cox, and Gavin Doherty. 2024. [Designing personalized mental health interventions for anxiety: Cbt therapists' perspective](#). *International Journal of Human-Computer Studies*, 190:103319.
- Judith S Beck. 2011. Cognitive-behavioral therapy. *Clinical textbook of addictive disorders*, 491:474–501.

- 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, Singapore. Association for Computational Linguistics.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiusi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. 2024. [Deepseek llm: Scaling open-source language models with longtermism](#). Preprint, arXiv:2401.02954.
- Clara E Hill and Karen M O’Brien. 1999. Helping skills. *Washington, DC: American Psychological Association*.
- 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.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023c. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. [Chatcounselor: A large language models for mental health support](#). Preprint, arXiv:2309.15461.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). In *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)*, pages 3469–3483, Online. Association for Computational Linguistics.
- Aishik Mandal, Tanmoy Chakraborty, and Iryna Gurevych. 2025. Towards privacy-aware mental health ai models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2502.00451*.
- Lindsey Phillips and Suzanne Tucker. 2023. A closer look at the mental health provider shortage. *Counseling Today*, 65(11):25–39.
- Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024a. [SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636, Miami, Florida, USA. Association for Computational Linguistics.
- Huachuan Qiu, Anqi Li, Lizhi Ma, and Zhenzhong Lan. 2024b. [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, Lizhi Ma, and Zhenzhong Lan. 2024c. [PsyGUARD: An automated system for suicide detection and risk assessment in psychological counseling](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4581–4607, Miami, Florida, USA. Association for Computational Linguistics.
- Huachuan Qiu, Tong Zhao, Anqi Li, Shuai Zhang, Hongliang He, and Zhenzhong Lan. 2023. 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.
- Philip Sedgwick. 2014. Spearman’s rank correlation coefficient. *Bmj*, 349.
- Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.
- Shahab Saquib Sohail, Faiza Farhat, Yassine Himeur, Mohammad Nadeem, Dag Øivind Madsen, Yashbir Singh, Shadi Atalla, and Wathiq Mansoor. 2023.

- Decoding chatgpt: a taxonomy of existing research, current challenges, and possible future directions. *Journal of King Saud University-Computer and Information Sciences*, page 101675.
- Aseem Srivastava, Tharun Suresh, Sarah P Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Counseling summarization using mental health knowledge guided utterance filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3920–3930.
- 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](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503, Online. Association for Computational Linguistics.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Laurens van der Maaten and Geoffrey Hinton. 2008. [Visualizing data using t-sne](#). *Journal of Machine Learning Research*, 9(86):2579–2605.
- Eva Wilmots, Nick Midgley, Lisa Thackeray, Shirley Reynolds, and Maria Loades. 2020. The therapeutic relationship in cognitive behaviour therapy with depressed adolescents: A qualitative study of good-outcome cases. *Psychology and Psychotherapy: Theory, Research and Practice*, 93(2):276–291.
- Yang Wu, Lu Wang, Mengjun Tao, Huiru Cao, Hui Yuan, Mingquan Ye, Xingui Chen, Kai Wang, and Chunyan Zhu. 2023. Changing trends in the global burden of mental disorders from 1990 to 2019 and predicted levels in 25 years. *Epidemiology and psychiatric sciences*, 32:e63.
- 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, 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, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. 2022. [D4: a Chinese dialogue dataset for depression-diagnosis-oriented chat](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2438–2459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. [CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13947–13966, Bangkok, Thailand. Association for Computational Linguistics.
- Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. 2024. [ESC-eval: Evaluating emotion support conversations in large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810, Miami, Florida, USA. Association for Computational Linguistics.
- Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023a. [AugESC: Dialogue augmentation with large language models for emotional support conversation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.
- 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. 2023b. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Yanxin Zheng, Wensheng Gan, Zefeng Chen, Zhenlian Qi, Qian Liang, and Philip S Yu. 2024a. Large language models for medicine: a survey. *International Journal of Machine Learning and Cybernetics*, pages 1–26.
- Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024b. [Self-chats from large language models make small emotional support chatbot better](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11325–11345, Bangkok, Thailand. Association for Computational Linguistics.

A Prompt for Reconstructing Client Utterances

We provide the prompt for reconstructing client utterances in Figures 6 and 7, which include the English and Chinese versions.

```
f'''You will play the role of a client undergoing a psychological
↪ counseling session. Your personal background information is as
↪ follows: {INSERT client chief complaint here}

# Dialogue between counselor and client:
{INSERT the masked dialogue here}

# Task instructions:
- (Important!) Based on the provided background information and
↪ context, complete the dialogue by filling in what the "Client"
↪ says in response.
- (Important!) Do not modify the counselor's words; they must be
↪ included as they are while completing the dialogue.
- Make sure to respond from the client's perspective, reflecting the
↪ emotions and reactions appropriate to the situation, keeping the
↪ language and tone consistent.
- Focus on maintaining logical flow and emotional coherence, ensuring
↪ that the interaction feels real and credible.
- If appropriate, you may show the client's inner conflict, confusion,
↪ or desire for help.
- Ensure that each part of the conversation pushes the narrative
↪ forward and sets the stage for future exchanges.

# Task objective:
- By completing the client's responses, explore their psychological
↪ state and the interaction with the counselor in depth, in order to
↪ better understand the client's emotional needs and psychological
↪ motivations.'''
```

Figure 6: Prompt for reconstructing client utterances. For the Chinese version, see Figure 7.

```
f'''你将扮演一位正在接受心理咨询的来访者。你的个人背景信息如
↪ 下: {INSERT client chief complaint here}

# 咨询师与来访者之间的对话:
{INSERT the masked dialogue here}

# 任务说明:
- (重要!) 根据提供的背景信息和上下文内容, 将“来访者:”说的内容补充
↪ 完整, 完成你与咨询师之间的对话。
- (重要!) 不改动咨询师说的话, 但需要在补充过程中将咨询师的话原封不
↪ 动地放入补充过程中。
- 请务必以来访者的身份回应, 展现出你对所处情境的真实反应与情感, 保持
↪ 语言和态度的一致性。
- 在对话中, 注意逻辑性和情感连贯性, 使互动真实可信。
- 如果适合, 可以展示来访者的内在冲突、困惑或希望获得帮助的愿望。
- 确保每一段对话能够推动情节发展, 并为后续对话提供铺垫。

# 任务目标:
- 通过完成来访者的话, 深入探索来访者的心理状态及其与咨询师的互动方
↪ 式, 以便更好地理解来访者的情感需求和心理动机。'''
```

Figure 7: Prompt for reconstructing client utterances. For the English version, see Figure 6.

B Prompt for Refining Counselor Utterances

We provide the prompt for refining counselor utterances in Figures 8 and 9, which include the English and Chinese versions, respectively. Meanwhile, we also use a filter function (with its corresponding criteria) to ensure high-quality refinement. The corresponding algorithm is presented in Figure 17.

```
f'''# Dialogue between Counselor and Client
{ctx}

You are a professional psychological counselor. Based on the provided
↪ dialogue and task instructions, please complete the specified
↪ tasks.

# Task Instructions:
- Do not modify the client's statements or continue the dialogue.
- Analyze the provided dialogue and identify any abrupt or
↪ inappropriate responses from the counselor.
- Suggest suitable replacements for the counselor's statements.
- Output both your analysis and the revised complete dialogue.

# Your Analysis:
# Revised Complete Dialogue:'''
```

Figure 8: Prompt for refining counselor utterances. For the Chinese version, see Figure 9.

```
f'''# 咨询师与来访者之间的对话
{ctx}

你是一位专业的心理咨询师, 根据所提供的对话和任务说明完成指定的任务。

# 任务说明:
- 不改动来访者说的话, 且不续写对话。
- 分析这段对话, 指出咨询师说话比较突兀的地方。
- 对咨询师说的话进行相应的替换。
- 输出分析和替换后的完整对话。

# 你的分析:
# 替换后的完整对话:'''
```

Figure 9: Prompt for refining counselor utterances. For the English version, see Figure 8.

C Case Study for RMRR

We present three examples to illustrate the incremental improvements at each stage of the framework. First, we provide an example segment (see Figures 10 and 11) of a masked dialogue for reconstruction. Second, we generate a new dialogue (see Figures 12 and 13) through the reconstruction process. Third, we obtain a refined dialogue (see Figures 14 and 15) using the refinement function.

```
'''
1. 来访者:
1. 咨询师: 嗯嗯你说
2. 来访者:
2. 咨询师: 你是怎么觉察到这部分
3. 来访者:
3. 咨询师: 嗯嗯, 哪怕你都知道这其中的不合理的地方, 还是无法拒绝。是
  ↳ 什么让你哪怕会付出更多, 更心累还是做不到拒绝
4. 来访者:
4. 咨询师: 我记得上次你谈到过你有能力去察觉身边人的情绪变化, 完全能感
  ↳ 同身受给予人心灵上的安慰和分担, 感觉是一种天赋, 给予人很珍贵的情感
  ↳ 价值, 正常情况下会是身边的人担心你厌倦了失去你。可是好像在你这里是
  ↳ 相反, 你付出更多, 但是怕融不进去
5. 来访者:
5. 咨询师: 我想问你哦, 你是一开始就会很自然的付出很多吗? 超过你和对方
  ↳ 当时的情谊
6. 来访者:
6. 咨询师: 当你如此付出的时候, 你回过头来看当时的心态里有对对方回应的
  ↳ 期待吗
7. 来访者:
7. 咨询师: 嗯, 好像“给别人帮助”是你的天赋, 很自然会去行使。而且我在想
  ↳ 你很强的共情能力是否会加强这一点。在对方还没有意识到需要的时候就提
  ↳ 前付出了
8. 来访者:
8. 咨询师: 嗯嗯, 这都是你现在的反思。你后来有调整吗。好像恐惧失去, 恐
  ↳ 惧融入不了本身加上你对人对待好坏的界限把握度的不确定感会让你有这些
  ↳ 忧虑和自我怀疑
9. 来访者:
9. 咨询师: 嗯嗯, 好像现在有一种无解的感觉。不知道该怎么办“我会不自主的
  ↳ 认为出现在我生命里的每一个人都是必要的”这个是怎么来的
10. 来访者:
10. 咨询师: 你是说是亲人给你的信息, 你慢慢内化了的
'''
```

Figure 10: A segment in a masked dialogue. For the English version, see Figure 11.

```
'''
1. Client:
1. Counselor: Mm-hmm, go ahead.
2. Client:
2. Counselor: How did you come to notice this?
3. Client:
3. Counselor: Mm, even though you clearly recognize the irrational
  ↳ aspects, you still can't refuse. What makes it so difficult to say
  ↳ no, despite knowing you'll end up more exhausted and paying a
  ↳ heavier emotional price?
4. Client:
4. Counselor: I recall last time you mentioned having the ability to
  ↳ notice changes in other people's emotions and empathize deeply with
  ↳ them, offering genuine comfort and emotional support. It seems to
  ↳ be your gift-providing valuable emotional connection. Normally, it
  ↳ would be others who worry about losing you due to fear of tiring
  ↳ you out, but it seems the opposite happens with you; you give more
  ↳ but fear not being able to fully integrate or be accepted.
5. Client:
5. Counselor: Can I ask, do you usually find yourself investing deeply
  ↳ in relationships right from the start, perhaps even more than the
  ↳ relationship itself warrants at the time?
6. Client:
6. Counselor: When you give so much, do you notice yourself expecting
  ↳ something in return?
7. Client:
7. Counselor: Mm, it seems that helping others is your gift, something
  ↳ you naturally and instinctively do. I wonder if your strong empathy
  ↳ actually reinforces this behavior, prompting you to offer help even
  ↳ before the other person realizes they need it.
8. Client:
8. Counselor: Mm-hmm, that's a reflection you're having now. Have you
  ↳ tried making adjustments since then? It seems your fear of losing
  ↳ connections, anxiety about not fitting in, combined with
  ↳ uncertainty around setting healthy emotional boundaries,
  ↳ contributes significantly to these feelings of worry and
  ↳ self-doubt.
9. Client:
9. Counselor: Mm-hmm. It seems there's a feeling of hopelessness now.
  ↳ When you say, "I automatically assume everyone who appears in my
  ↳ life is necessary," where does this belief originate from?
10. Client:
10. Counselor: You're suggesting it's a message you received from your
  ↳ family and gradually internalized?
'''
```

Figure 11: A segment in a masked dialogue. For the Chinese version, see Figure 10.

...
1. 来访者：你好，真的不知道该怎么办了
1. 咨询师：嗯嗯你说。（“嗯嗯你说。”显得有些敷衍，尽管可能是为了让来访者继续说下去，但可以更具建设性。）
2. 来访者：关于人际关系已经困扰我好多年了，虽然一次次开导自己要去看开，但到事情发生时还是会陷进去，我是一个感性的人，把什么想的都单纯，我特别缺乏自信和
↳ 依赖别人，觉得彼此聊的开，相处的融洽就是好朋友了，但其实在她们眼里我也不过是一时的过客而已，对我来讲和别人发生不愉快的事情再见面就会很尴尬，但我
↳ 想不明白为什么有些人发生矛盾甚至吵架了还能笑嘻嘻的话，然后在分别时删了对方面，不累吗？我认识把一个当朋友就要不离不弃，但其实她的交际圈中你只是其
↳ 中一个，有时把自己放的太卑微，听着她们说自己不喜欢不认同的话也不敢去反驳，怕她们生气，人际关系和我想的不一樣，感觉很虚伪，觉得自己很懦弱，很悲
↳ 哀，没有一个真正的朋友，好像都藏着事，彼此防着，很累，很消极，我特别羡慕那些独立的人，觉得她们真棒，拿的起放的下。感觉自己把自己弄的一团糟，滥好
↳ 人一个也两面得不到好。不知道该怎么办。
2. 咨询师：你是怎么觉察到这部分？（“你是怎么觉察到这部分？”虽然是开放性问题，但没有具体指向，让来访者不一定知道具体要回答什么。）
3. 来访者：我一直觉得关系应该是真诚的，但总是看到大家在矛盾后还能迅速和好，这让我很困惑。我往往在矛盾后会感到不自在，不知道如何面对对方，还有很多其他
↳ 的事情困扰着我。
3. 咨询师：嗯嗯，哪怕你都知道这其中的不合理的地方，还是无法拒绝。是什么让你哪怕会付出更多，更心累还是做不到决绝？（“哪怕你都知道这其中的不合理的地
↳ 方，还是无法拒绝。是什么让你哪怕会付出更多，更心累还是做不到决绝？”有些复杂，且可能让来访者感到困惑和不安。）
4. 来访者：可能是因为我太害怕孤独了吧？我害怕一旦拒绝或是表示自己不满，就会被大家排挤。即使知道这种想法不合理，我还是无法摆脱这种恐惧。
4. 咨询师：我记得上次你谈到过你有能力去察觉身边人的情绪变化，完全能感同身受给予人心灵上的安慰和分担，感觉是一种天赋，给予人很珍贵的情感价值，正常情况
↳ 下会是身边的人担心你厌倦了失去你。可是好像在你这里是相反，你付出更多，但是怕融不进去。
5. 来访者：对，我总是能够察觉到别人需要什么，然后不由自主地去满足他们。但问题是，我觉得自己总是在付出，却得不到应有的回报或是理解。很多时候，我觉得自
↳ 己的存在感很薄弱，很难真正融入到任何一个群体中。
5. 咨询师：我想问你哦，你是一开始就会很自然的付出很多吗？超过你和对方当时的情谊。（咨询师的疑问“我想问你哦，你是一开始就会很自然的付出很多吗？超过你和
↳ 对方当时的情谊。”显得有些突兀，且有点评判意味，不太符合咨询师的中立态度。）
6. 来访者：是的，一开始我就会全身心地投入，总想通过多付出来赢得友谊和认可。但是后来我发现，这样做不仅没有让关系更紧密，反而让我感觉更加孤独和疲惫。
6. 咨询师：当你如此付出的时候，你回过头来看当时的心态里有对对方回应的期待吗？
7. 来访者：有的，我总是希望对方能同样真心待我，但往往事与愿违。我感到很失望，有时候甚至怀疑自己是不是做错了什么。
7. 咨询师：嗯，好像“给别人帮助”是你的天赋，很自然会去行使。而且我在想你很强的共情能力是否会加强这一点。在对方还没有意识到需要的时候就提前付出了。
8. 来访者：是的，那是我很自然的反应。我总是担心别人会遇到麻烦，所以总是尽可能去帮助。但这样一来，自己也变得无所适从，有时候甚至感觉对方并不真正需要我
↳ 的帮助。
8. 咨询师：嗯嗯，这都是你现在的反思。你后来有调整吗？好像恐惧失去，恐惧融入不了本身加上你对人对待好坏的界限把握度的不确定感会让你有这些忧虑和自我怀
↳ 疑。（“好像恐惧失去，恐惧融入不了本身加上你对人对待好坏的界限把握度的不确定感会让你有这些忧虑和自我怀疑”这句话可能过于复杂，对来访者来说不容易被理
↳ 解。）
9. 来访者：我试着去调整，但效果并不明显。我总是无法摆脱对于失去朋友的恐惧，这让我在很多时候犹豫不决，害怕去真正表达自己的意见或情感。
9. 咨询师：嗯嗯，好像现在有一种无解的感觉。不知道该怎么办“我会不自主的认为出现在我生命里的每一个人都是必要的”这个是怎么来的？（“好像现在有一种无解的
↳ 感觉”这句话显得较为绝对且负面，可能更好地引导来访者探索自己的感受，而不是直接给出评价。）
10. 来访者：我想可能是小时候家人告诉我，每一个出现在我们生命中的人都是有意义的，所以我特别重视每一段关系。但后来发现，这种想法给我带来了许多困扰。
10. 咨询师：你是说是亲人给你的信息，你慢慢内化了的？（这句话在表达上不够简单明了。）
...

Figure 12: A dialogue segment after reconstruction. Refinement feedback is highlighted in parentheses. For the English version, see Figure 13.

```

'''
1. Counselor: Mm-hmm, go ahead. (The original response, "Mm-hmm, go ahead," felt somewhat superficial. While intended to encourage the client to
↳ continue, it could have been more engaging.)
2. Client: Interpersonal relationships have troubled me for many years. I keep trying to persuade myself to let go, but whenever issues arise, I still
↳ get stuck. I'm sensitive and tend to see things simply. I often lack confidence and become dependent on others. When I feel comfortable and
↳ connected with someone, I quickly consider them a good friend. But in their eyes, I'm probably just a temporary acquaintance. For me, encountering
↳ someone again after a conflict is awkward, yet I can't understand how others can still smile and chat after arguments, only to delete each other
↳ afterward. Isn't that exhausting? For me, friendship means loyalty and staying together no matter what, but for others, I seem to be just another
↳ casual acquaintance. Sometimes I put myself in such a humble position that even when they say things I dislike or disagree with, I dare not
↳ express my true feelings, fearing they'll get upset. Relationships are so different from what I imagined. They feel hypocritical, and I feel weak
↳ and pathetic, like I have no real friends. It seems everyone has secrets and guards against each other. I'm exhausted and negative, and I envy
↳ those independent people who can easily move on. I feel like I've made a mess of things, trying to be nice to everyone but getting nothing in
↳ return. I just don't know what to do.
2. Counselor: How did you come to notice this? (While open-ended, the original question "How did you become aware of this?" lacks specific direction,
↳ potentially leaving the client uncertain about how to respond.)
3. Client: I've always believed relationships should be genuine, but seeing how quickly others reconcile after conflicts confuses me. After conflicts,
↳ I feel uneasy and unsure how to face the other person. There are also many other things bothering me.
3. Counselor: Mm, even though you clearly recognize the irrational aspects, you still can't refuse. What makes it so difficult to say no, despite
↳ knowing you'll end up more exhausted and paying a heavier emotional price? (The original phrasing was overly complex and potentially confusing for
↳ the client.)
4. Client: Maybe because I'm too afraid of loneliness? I'm scared that if I refuse or express dissatisfaction, I'll be excluded. Even though I realize
↳ this fear isn't rational, I can't shake it off.
4. Counselor: I recall last time you mentioned having the ability to notice changes in other people's emotions and empathize deeply with them,
↳ offering genuine comfort and emotional support. It seems to be your gift—providing valuable emotional connection. Normally, it would be others who
↳ worry about losing you due to fear of tiring you out, but it seems the opposite happens with you; you give more but fear not being able to fully
↳ integrate or be accepted.
5. Client: Exactly. I'm always aware of others' needs and instinctively try to meet them. But the issue is, I feel I'm always giving without receiving
↳ the acknowledgment or understanding I deserve. Often, I feel invisible, unable to genuinely integrate into any group.
5. Counselor: Can I ask, do you usually find yourself investing deeply in relationships right from the start, perhaps even more than the relationship
↳ itself warrants at the time? (The original phrasing felt abrupt and somewhat judgmental, not neutral enough for a counselor.)
6. Client: Yes, from the beginning, I fully invest myself, hoping to win friendship and acceptance through extra effort. But eventually, I've found
↳ that doing this doesn't make relationships closer. Instead, it leaves me feeling more isolated and drained.
6. Counselor: When you give so much, do you notice yourself expecting something in return?
7. Client: Yes, I always hope they'll treat me with the same sincerity, but often reality disappoints me. I feel let down, sometimes wondering if I've
↳ done something wrong.
7. Counselor: Mm, it seems that helping others is your gift, something you naturally and instinctively do. I wonder if your strong empathy actually
↳ reinforces this behavior, prompting you to offer help even before the other person realizes they need it.
8. Client: Exactly, that's my instinctive reaction. I'm always worried others might face problems, so I help proactively. But then, I feel confused
↳ and uncertain, sometimes realizing the other person may not actually need my help.
8. Counselor: Mm-hmm, that's a reflection you're having now. Have you tried making adjustments since then? It seems your fear of losing connections,
↳ anxiety about not fitting in, combined with uncertainty around setting healthy emotional boundaries, contributes significantly to these feelings
↳ of worry and self-doubt. (The original sentence was overly complex and difficult to understand clearly.)
9. Client: I've tried making adjustments, but I haven't seen much improvement. I still struggle with the fear of losing friends, making me hesitant
↳ and unable to genuinely express my thoughts or feelings.
9. Counselor: Mm-hmm. It seems there's a feeling of hopelessness now. When you say, "I automatically assume everyone who appears in my life is
↳ necessary," where does this belief originate from? (The original statement "It seems there's a feeling of hopelessness now" appeared overly
↳ negative and conclusive. Reframing it as empathetic and exploratory would better encourage the client to reflect.)
10. Client: Maybe it's because my family used to tell me that everyone we encounter in life has significance, so I've always valued every relationship
↳ deeply. But I've realized this belief has caused me a lot of distress.
10. Counselor: You're suggesting it's a message you received from your family and gradually internalized? (Simplified from the original statement to
↳ be clearer.)
'''

```

Figure 13: A dialogue segment after reconstruction. Refinement feedback is highlighted in parentheses. For the Chinese version, see Figure 12.

```

'''
1. 来访者：你好，真的不知道该怎么办了
1. 咨询师：我能理解你的感受，请告诉我具体发生了什么？
2. 来访者：关于人际关系已经困扰我好多年了，虽然一次次开导自己要看开，但到事情发生时还是会陷进去，我是一个感性的人，把什么想的都单纯，我特别缺乏自信 and
↳ 依赖别人，觉得彼此聊的开，相处的融洽就是好朋友了，但其实在她们眼里我也不过是一时的过客而已，对我来讲和别人发生不愉快的事情再见面就会很尴尬，但我
↳ 想不明白为什么有些人发生矛盾甚至吵架了还能笑嘻嘻的说话，然后在分别时删了对方，不累吗？我认识把一个当朋友就要不离不弃，但其实她的交际圈中你只是其
↳ 中一个，有时把自己放的太卑微，听着她们说自己不喜欢不认同的话也不敢去反驳，怕她们生气，人际关系和我想的不一樣，感觉很虚伪，觉得自己很懦弱，很悲
↳ 哀，没有一个真正的朋友，好像都藏着事，彼此防着，很累，很消极，我特别羡慕那些独立的人，觉得她们真棒，拿的起放的下。感觉自己把自己弄的一团糟，滥好
↳ 人一个也两面得不到好。不知道该怎么办。
2. 咨询师：你觉得你的感受始终很真实，但为什么别人处理矛盾的方式让你困扰呢？
3. 来访者：我一直觉得关系应该是真诚的，但总是看到大家在矛盾后还能迅速和好，这让我很困惑。我往往在矛盾后会感到不自在，不知道如何面对对方，还有很多其他
↳ 的事情困扰着我。
3. 咨询师：我听到了你的困惑。不过，是什么让你即使很不舒服，还是害怕表达自己的不满呢？
4. 来访者：可能是因为我太害怕孤独了吧？我害怕一旦拒绝或是表示自己不满，就会被大家排挤。即使知道这种想法不合理，我还是无法摆脱这种恐惧。
4. 咨询师：你的天赋是能细腻地察觉到他人情绪，能提供安慰和支持。可是在你看来，尽管你付出很多，还是担心无法融入，是这样吗？
5. 来访者：对，我总是能够察觉到别人需要什么，然后不由自主地去满足他们。但问题是，我觉得自己总是在付出，却得不到应有的回报或是理解。很多时候，我觉得自
↳ 己的存在感很薄弱，很难真正融入到任何一个群体中。
5. 咨询师：你有没有意识到，在关系刚开始时，你是否会投入很多精力去维护彼此的关系，这可能超过你和对方的实际情感联系？
6. 来访者：是的，一开始我就会全身心地投入，总想通过多付出来赢得友谊和认可。但是后来我发现，这样做不仅没有让关系更紧密，反而让我感觉更加孤独和疲惫。
6. 咨询师：当你如此付出的时候，你回过头来看当时的心态里有对对方回应的期待吗？
7. 来访者：有的，我总是希望对方能同样真心待我，但往往事与愿违。我感到很失望，有时候甚至怀疑自己是不是做错了什么。
7. 咨询师：嗯，好像“给别人帮助”是你的天赋，很自然会去行使。而且我在想你很强的共情能力是否会加强这一点。在对方还没有意识到需要的时候就提前付出了。
8. 来访者：是的，那是我很自然的反应。我总是担心别人会遇到麻烦，所以总是尽可能去帮助。但这样一来，自己也变得无所适从，有时候甚至感觉对方并不真正需要我
↳ 的帮助。
8. 咨询师：能够理解。你有时会对对方的回应有较高的期望，这种期望带来了失望和疲惫。那么后来你有做出一些调整吗？
9. 来访者：我试着去调整，但效果并不明显。我总是无法摆脱对于失去朋友的恐惧，这让我在很多时候犹豫不决，害怕去真正表达自己的意见或情感。
9. 咨询师：这种感受很自然，特别是在你害怕失去关系的情况下。你觉得这种恐惧和犹豫来自哪里呢？
10. 来访者：我想可能是小时候家人告诉我，每一个出现在我们生命中的人都是有意义的，所以我特别重视每一段关系。但后来发现，这种想法给我带来了许多困扰。
10. 咨询师：我明白了，这种观念是你的亲人传递给你的，对吗？你慢慢地内化了这种观念，使得你更加重视每一个关系。
'''

```

Figure 14: A dialogue segment after refinement. For the English version, see Figure 15.


```

'''
1. Client: Hi, I really don't know what to do anymore.
1. Counselor: I can understand your feelings. Could you tell me exactly what's been going on?
2. Client: Interpersonal relationships have troubled me for many years. I keep trying to persuade myself to let go, but whenever issues arise, I still
↪ get stuck. I'm sensitive and tend to see things simply. I often lack confidence and become dependent on others. When I feel comfortable and
↪ connected with someone, I quickly consider them a good friend. But in their eyes, I'm probably just a temporary acquaintance. For me, encountering
↪ someone again after a conflict is awkward, yet I can't understand how others can still smile and chat after arguments, only to delete each other
↪ afterward. Isn't that exhausting? For me, friendship means loyalty and staying together no matter what, but for others, I seem to be just another
↪ casual acquaintance. Sometimes I put myself in such a humble position that even when they say things I dislike or disagree with, I dare not
↪ express my true feelings, fearing they'll get upset. Relationships are so different from what I imagined. They feel hypocritical, and I feel weak
↪ and pathetic, like I have no real friends. It seems everyone has secrets and guards against each other. I'm exhausted and negative, and I envy
↪ those independent people who can easily move on. I feel like I've made a mess of things, trying to be nice to everyone but getting nothing in
↪ return. I just don't know what to do.
2. Counselor: Your feelings seem very genuine. Why do you think other people's ways of handling conflicts trouble you?
3. Client: I've always believed relationships should be genuine, but seeing how quickly others reconcile after conflicts confuses me. After conflicts,
↪ I feel uneasy and unsure how to face the other person. There are also many other things bothering me.
3. Counselor: I hear your confusion. But what makes you afraid to express your dissatisfaction, even when you feel uncomfortable?
4. Client: Maybe because I'm too afraid of loneliness? I'm scared that if I refuse or express dissatisfaction, I'll be excluded. Even though I realize
↪ this fear isn't rational, I can't shake it off.
4. Counselor: You have a gift for sensitively noticing others' emotions, and you're good at offering comfort and support. Yet despite giving so much,
↪ you still fear not fitting in. Is that right?
5. Client: Exactly. I'm always aware of others' needs and instinctively try to meet them. But the issue is, I feel I'm always giving without receiving
↪ the acknowledgment or understanding I deserve. Often, I feel invisible, unable to genuinely integrate into any group.
5. Counselor: Have you noticed that at the beginning of relationships, you tend to invest a lot of energy into maintaining connections, possibly more
↪ than your actual emotional closeness would justify?
6. Client: Yes, from the beginning, I fully invest myself, hoping to win friendship and acceptance through extra effort. But eventually, I've found
↪ that doing this doesn't make relationships closer. Instead, it leaves me feeling more isolated and drained.
6. Counselor: When you invest so much, do you recognize having expectations for how others should respond to your efforts?
7. Client: Yes, I always hope they'll treat me with the same sincerity, but often reality disappoints me. I feel let down, sometimes wondering if I've
↪ done something wrong.
7. Counselor: It sounds like helping others comes naturally to you, a real strength of yours. I'm wondering if your strong sense of empathy further
↪ reinforces this, prompting you to provide help even before others realize they need it.
8. Client: Exactly, that's my instinctive reaction. I'm always worried others might face problems, so I help proactively. But then, I feel confused
↪ and uncertain, sometimes realizing the other person may not actually need my help.
8. Counselor: I can understand that. Sometimes your high expectations about their responses leave you disappointed and worn out. Have you tried making
↪ any adjustments?
9. Client: I've tried making adjustments, but I haven't seen much improvement. I still struggle with the fear of losing friends, making me hesitant
↪ and unable to genuinely express my thoughts or feelings.
9. Counselor: Such feelings are natural, especially when you fear losing relationships. Where do you think this fear and hesitation come from?
10. Client: Maybe it's because my family used to tell me that everyone we encounter in life has significance, so I've always valued every relationship
↪ deeply. But I've realized this belief has caused me a lot of distress.
10. Counselor: I understand. This belief was conveyed by your family, correct? Over time, you've internalized this idea, leading you to value each
↪ relationship deeply.
'''

```

Figure 15: A dialogue segment after refinement. For the Chinese version, see Figure 14.

D Filter Function for Dialogue Reconstruction

We present the filter function for dialogue reconstruction in Figure 16 and set the maximum number of iterations to 8. If all 8 dialogues generated through dialogue reconstruction fail the filter function, we will select the dialogue with the highest score among them.

```
from difflib import SequenceMatcher
def filter_function(original_dialogue, new_dialogue):
    original_counselor_utters = get_counselor_utters(original_dialogue)
    new_counselor_utters = get_counselor_utters(new_dialogue)
    s = SequenceMatcher(None, original_counselor_utters,
    ↪ new_counselor_utters)
    score = round(s.ratio(), 3)
    if score < 0.85:
        # repeat dialogue_reconstruction function
```

Figure 16: Filter function for dialogue reconstruction.

E Filter Function for Dialogue Refinement

We present the filter function for dialogue refinement in Figure 17 and set the maximum number of iterations to 8. If all 8 dialogues generated through dialogue refinement fail the filter function, we will select the dialogue with the highest score among them.

```
from difflib import SequenceMatcher
def filter_function(original_dialogue, new_dialogue):
    original_client_utters = get_client_utters(original_dialogue)
    new_client_utters = get_client_utters(new_dialogue)
    s = SequenceMatcher(None, original_client_utters,
    ↪ new_client_utters)
    score = round(s.ratio(), 3)
    if score < 0.85:
        # repeat dialogue_refinement function
```

Figure 17: Filter function for refining dialogue.

F Implementation Details

The data used in this paper follows OpenAI’s data format. All experiments are conducted on NVIDIA A100 80G GPUs for model training. During fine-tuning, we utilize four GPUs, set the training batch size to 1 per device, and configure the gradient accumulation step to 2, meaning that gradients from every two steps are accumulated before updating the model parameters. The learning rate is set to $1e-5$, with a cosine learning rate scheduler dynamically adjusting it throughout the training process. The training spans two epochs. To enhance efficiency and balance model performance, we enable 16-bit half-precision floating-point computation.

G Dialogue Example

We provide a complete dialogue with 27 turns from dataset \mathcal{D}_4 , as shown in Figure 18. For brevity, we omit the intermediate turns of this dialogue.

```

'''
Client: Hello, I really don't know what to do anymore. . .
Counselor: I can understand your feelings. Could you tell me exactly what's been going on?
Client: Interpersonal relationships have troubled me for years. Even though I repeatedly try to persuade myself to let go, when something happens, I
↪ still get caught up in it. I'm an emotional person who tends to see things simply. I lack self-confidence and tend to rely on others. If I get
↪ along well with someone, I easily consider them a good friend, but I later realize that to them, I'm just a temporary acquaintance. When conflicts
↪ happen, I feel incredibly awkward when meeting again, yet I can't understand why others seem fine talking and laughing even after arguments, then
↪ casually delete each other afterward—doesn't it exhaust them? For me, friendship means loyalty and persistence, but in their social circles, I'm
↪ just one of many. Sometimes I belittle myself too much; I don't dare express disagreement even when they say things I dislike, fearing their anger.
↪ Relationships seem very superficial, leaving me feeling weak, pathetic, and without a genuine friend. Everyone seems guarded, hiding their true
↪ selves. It's exhausting and demoralizing. I really envy those independent individuals who seem strong and able to move on easily. I feel I've made
↪ a mess of things, trying to please everyone but ultimately satisfying no one, including myself. I just don't know what to do.
Counselor: Your feelings seem very genuine. Why do you think other people's ways of handling conflicts trouble you?

...

Client: Exactly. Here, I feel safe and can express my true thoughts without much worry.
Counselor: I notice you do a lot of reflecting. Are you used to managing problems alone?
Client: Yes, I'm accustomed to handling things myself without relying on others. Yet deep down, I genuinely long for a friend whom I can trust and
↪ depend on.
Counselor: Do you find it easier to express yourself through writing?
Client: Yes, I prefer writing my feelings down. It feels safer. Speaking face-to-face always worries me—I fear people's reactions, misunderstanding,
↪ or rejection.
Counselor: Are you afraid that your vulnerability and genuine feelings might make others perceive you as troublesome?
Client: Exactly. I'm worried my vulnerability and honesty might burden others or make them uncomfortable.
Counselor: If you gave others a chance and they responded positively to your feelings, do you think the outcome would be different?
Client: I've considered that possibility but always lacked the courage to try. I'm scared of failure and even more afraid of losing what few
↪ connections I already have.
Counselor: Perhaps you could pay attention to your feelings in such situations and notice any changes within yourself.
Client: I'll try, though overcoming this psychological barrier is difficult. Still, I'll work to confront my fears and attempt to trust others more.
Counselor: That's understandable. Do you sometimes feel you lack something that makes people interested in deeper interactions?
Client: Yes, I sometimes think I'm ordinary and have nothing particularly attractive about me. I fear others might lose interest.
Counselor: Everyone has uniqueness, even if we consider ourselves ordinary. But maybe at times, we feel more mundane compared to others. Is this how
↪ you're feeling?
Client: Yes, I always feel that I'm not particularly special, especially when I see independent, confident people. It deepens my sense of inferiority.
Counselor: I can sense your confusion. Perhaps we can explore the reasons behind this internal emptiness and whether specific situations intensify
↪ these feelings.
Client: Yes, sometimes I genuinely feel hollow inside, like there's nothing compelling enough for others to stay.
Counselor: Is that how you feel?
Client: Exactly. Sometimes it seems all my efforts are pointless, neither making myself nor others truly happy.
Counselor: I understand. From your description, I sense you're experiencing considerable stress and possibly feelings of helplessness. Our time is up
↪ for today, but let's continue this conversation next week, alright?
Client: Alright, thank you for listening and being here today.
Counselor: You're welcome. Lastly, please fill out this feedback form to help us better serve you.
'''

```

Figure 18: An example dialogue in \mathcal{D}_4 .

H Dialogue Topic Distribution

We present a comparison of the dialogue topic distributions between \mathcal{D}_0 and \mathcal{D}_4 , as shown in Table 4.

Dataset	Mate Selection	Love Issues	Post-Love Issues	Marital Issues	Sexual Conceptual Confusion	Sexual Preference Confusion	Gender Identity	Sexual Orientation	Family Conflict	Child Education
\mathcal{D}_0	201	566	124	237	42	19	97	94	788	119
\mathcal{D}_4	111	500	125	187	19	5	25	42	859	70
Dataset	Domestic Violence	Sexual Harassment	Sexual Abuse	Bullying	Loss	Setback	Political Violence	Secondary Trauma	Major Life Event Trauma	Psychological Counseling Trauma
\mathcal{D}_0	79	17	21	29	131	119	12	52	71	259
\mathcal{D}_4	78	8	10	15	90	81	10	36	62	272
Dataset	Health Issues	Psychosomatic Symptoms	School Environment Adaptation	Workplace Environment Adaptation	Role Transition Adaptation	Cultural Adaptation	Self-Exploration and Growth	Personality Trait Exploration	Negative Self-Evaluation Exploration	Exploration of Life Meaning
\mathcal{D}_0	212	495	94	62	432	109	1187	86	1377	627
\mathcal{D}_4	98	436	106	49	357	79	1061	28	1397	613
Dataset	Time Management	Emotion Regulation	Depression	Anxiety	Stress	Obsessive-Compulsive Disorder	Fear	Decision-Making Difficulties	Impulsivity	Interpersonal Skills Counseling
\mathcal{D}_0	425	1609	631	629	407	65	205	775	196	688
\mathcal{D}_4	366	1784	791	941	645	81	186	799	159	525
Dataset	Interpersonal Conflicts	Social Anxiety	Learning Efficiency	Work Efficiency	Job Dissatisfaction	Learning Dissatisfaction	Occupational Burnout	Learning Burnout	Job Challenges	Learning Pressure
\mathcal{D}_0	1315	459	146	59	317	46	466	184	325	759
\mathcal{D}_4	1431	474	123	44	242	27	444	150	235	776
Dataset	School Dropout	Unemployed	Unemployment	Career Planning	Suspected Neurosis and Mental Disorders	Neurosis and Mental Disorders	Self-Harm Tendency	Suicidal Tendency	Harming Others	Harming the User
\mathcal{D}_0	22	40	90	447	143	53	156	133	43	102
\mathcal{D}_4	20	27	72	333	98	26	132	116	21	53

Table 4: A comparison of the dialogue topic distribution between \mathcal{D}_0 and \mathcal{D}_4 .

I Privacy Review Guidelines

I.1 Purpose of Privacy Review

The purpose of this privacy review is to ensure the confidentiality and protection of all personal and sensitive information shared by clients during the counseling process. It is essential to maintain trust and provide a safe environment for clients.

I.2 Scope of Confidential Information

Personal Information: This includes the client's name, age, gender, contact details (phone number, email address, home address), and any other identifying information.

Mental Health Information: Any details regarding the client's emotional state, psychological symptoms, diagnoses, treatment history, and progress in counseling.

Health History: Information about the client's medical history, including past and current physical or mental health conditions, and family medical history.

Counseling Content: All discussions and disclosures made by the client during counseling sessions, including personal issues, family matters, and any other sensitive topics.

Assessment and Test Results: Any psychological assessments or tests conducted during counseling and their results.

Other Sensitive Information: Any other information that may be considered private or confidential, such as suicidal or self-harm thoughts, violence-related concerns, or experiences of abuse or neglect.

J Evaluation Metrics for Counseling Skills

The following is a comprehensive set of definitions and corresponding rating scales for the skills that are aligned with the evaluation principles that a competent counselor should master. Each skill is defined with a detailed definition, and a 5-point Likert scale is provided to assess how well these skills are applied in a counseling context. To summarize, rating criteria for eight dimensions are defined as follows: **5: Excellent** – Strong demonstration of the skill. **4: Good** – Competent demonstration with room for minor improvement. **3: Average** – Some strengths but notable gaps or inconsistencies. **2: Poor** – Weak demonstration with several critical issues. **1: Very Poor** – Significant issues with the skill, hindering the therapeutic process.

```
def get_rating_prompt(ctx: str, response: str, metric: str):
    system_prompt = f'''You are a professional psychological counseling
    ↳ supervisor. You should evaluate how well the counselor applies
    ↳ their skills in a counseling context.'''
    rating_prompt = f'''The following is a counseling context.
    Dialogue history: {ctx}
    Counselor's response: {response}

    {metric}

    Provide a brief reasoning for your rating based on these criteria, and
    ↳ then assign a numerical rating. Provide your answer in the
    ↳ following format.
    - Reasoning: (Your explanation here)
    - Rating: (Ranging from 1 to 5)'''

    return system_prompt, rating_prompt
```

Figure 19: Prompt for evaluating counseling skills.

We provide a detailed prompt template for evaluating counseling skills, as shown in Figure 19. Here, `ctx` refers to the dialogue history, and `response` denotes the content generated by an LLM-based counselor. The metrics used in the prompt template are described in Sections J.1 to J.9.

J.1 Empathy

Definition. Empathy refers to the counselor's ability to understand, resonate with, and validate the client's emotions and experiences. It involves not only recognizing the client's feelings but also communicating a deep sense of emotional understanding and support.

Rating Criteria.

- 5: The counselor demonstrates deep empathy, consistently validating and responding to the client's emotions and experiences in a way that fosters connection.
- 4: The counselor shows empathy, but it may occasionally lack depth or clarity in certain moments.
- 3: The counselor shows basic empathy, but the emotional understanding feels somewhat distant or incomplete.
- 2: The counselor struggles to demonstrate empathy, and emotional understanding feels superficial or lacking.
- 1: The counselor does not show empathy or seems indifferent to the client's emotional experiences.

J.2 Active Listening

Definition. Active listening ensures a thorough understanding of the client's problems and emotions. The counselor attentively listens to both verbal and non-verbal cues, confirming the primary concerns and emotional state of the client. This

helps build rapport and trust, while also making the client feel fully heard.

Rating Criteria.

- 5: The counselor listens attentively without interruption, demonstrates full understanding, and accurately reflects the client's feelings and concerns.
- 4: The counselor listens well but may occasionally miss small details or interrupt slightly.
- 3: The counselor listens but struggles to pick up key details or misinterprets some aspects of the client's communication.
- 2: The counselor listens partially, often missing important cues or failing to grasp the main concerns.
- 1: The counselor does not listen actively, often interrupting or showing little engagement with the client's message.

J.3 Issue Clarification

Definition. Clarification involves seeking additional details or further explanation when the client's communication is unclear. The counselor asks specific questions to gain a better understanding of the client's situation, ensuring that all critical aspects of the problem are comprehended fully.

Rating Criteria.

- 5: The counselor actively seeks clarification whenever necessary, asking precise questions that help unpack the client's issues clearly and comprehensively.
- 4: The counselor seeks clarification in most situations, though some questions may be slightly general or unclear.
- 3: The counselor asks some clarifying questions, but they may miss key aspects or fail to probe deeply enough.
- 2: The counselor rarely asks for clarification, leaving gaps in understanding that could hinder progress.
- 1: The counselor does not seek clarification, resulting in a poor understanding of the client's issues.

J.4 Open-ended Questioning

Definition. Open-ended questions are designed to encourage the client to explore their thoughts, feelings, and experiences in greater depth. These questions usually start with "how," "what," or "can

you tell me more about," and allow the client to provide expansive, reflective answers.

Rating Criteria.

- 5: The counselor consistently uses open-ended questions that encourage deep exploration and self-reflection, promoting rich, meaningful dialogue.
- 4: The counselor asks open-ended questions but may sometimes rely on closed or leading questions.
- 3: The counselor occasionally uses open-ended questions but often defaults to yes/no questions or questions with limited scope.
- 2: The counselor uses very few open-ended questions, limiting the client's ability to explore their own thoughts and feelings.
- 1: The counselor avoids open-ended questions entirely, only asking yes/no or directive questions.

J.5 Encouraging Self-Exploration

Definition. Encouraging self-exploration means asking questions and providing prompts that help the client reflect on their own emotions, thoughts, behaviors, and decision-making. This promotes greater self-awareness and empowers the client to make their own insights and choices.

Rating Criteria.

- 5: The counselor frequently encourages the client to explore their own thoughts and feelings, fostering significant self-awareness and insight.
- 4: The counselor encourages self-exploration but may not consistently prompt the client to explore deeper layers of their experiences.
- 3: The counselor occasionally encourages self-exploration, but it may lack depth or clarity, limiting the client's reflection.
- 2: The counselor provides limited opportunities for self-exploration, directing the conversation more than encouraging self-reflection.
- 1: The counselor does not encourage self-exploration, instead providing solutions or interpretations without engaging the client's own thoughts.

J.6 Cognitive Restructuring

Definition. Cognitive restructuring involves helping the client identify and challenge distorted or unrealistic thought patterns. The counselor assists the client in reframing negative or maladaptive

thoughts, fostering more realistic and helpful cognitive patterns that promote emotional well-being.

Rating Criteria.

- 5: The counselor skillfully helps the client identify distorted thoughts and gently guides them to more balanced, realistic perspectives.
- 4: The counselor helps challenge distorted thinking, but may not consistently provide clear alternatives or insight.
- 3: The counselor offers some cognitive restructuring, but the process feels incomplete or lacks sufficient exploration of thought patterns.
- 2: The counselor rarely engages in cognitive restructuring, providing minimal guidance for challenging negative thoughts.
- 1: The counselor does not address cognitive distortions or fails to help the client change unhelpful thinking patterns.

J.7 Guided Questioning

Definition. Guided questioning refers to the use of focused questions to help the client narrow down their concerns or focus on specific goals. This approach helps the client clarify their thoughts and find solutions to specific problems, often by prompting deeper reflection on particular aspects of their experience.

Rating Criteria.

- 5: The counselor uses guided questions effectively, helping the client focus on specific issues or goals in a way that enhances clarity and progress.
- 4: The counselor uses guided questions but may not always focus them as effectively on the client's immediate needs or goals.
- 3: The counselor uses some guiding questions, but they may be overly broad or fail to narrow in on the client's main concerns.
- 2: The counselor rarely uses guiding questions, or their questions lack focus, making it difficult for the client to concentrate on specific issues.
- 1: The counselor does not use guiding questions, and the session lacks focus or clarity on specific goals.

J.8 Non-judgmental and Accepting Attitude

Definition. A non-judgmental and accepting attitude means creating a safe and supportive environment where the client feels free to share their

thoughts and feelings without fear of criticism or negative judgment. The counselor maintains a neutral, respectful approach, accepting the client's experiences and emotional expressions without imposing their own values or opinions.

Rating Criteria.

- 5: The counselor consistently maintains a non-judgmental, accepting stance, allowing the client to share openly without fear of judgment.
- 4: The counselor is generally non-judgmental, with occasional lapses in maintaining complete neutrality or acceptance.
- 3: The counselor maintains a neutral attitude but may unintentionally come across as judgmental in some instances.
- 2: The counselor's attitude is occasionally critical or dismissive, potentially creating discomfort for the client.
- 1: The counselor is overtly judgmental or dismissive, making the client feel unsafe or unsupported.

J.9 Overall Score

Definition. This overall score combines the key evaluation principles into one holistic rating scale. The counselor's performance will be evaluated based on how effectively they apply each principle in their practice. Each principle contributes equally to the final score, providing a comprehensive assessment of the counselor's abilities in understanding and responding to the client's needs. The overall score will reflect the counselor's skill in fostering a supportive, non-judgmental, and effective therapeutic environment.

Evaluation Principles:

- **Active Listening:** Ensure a complete understanding of the client's issues and emotions, while confirming the client's main concerns and feelings.
- **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 sad."
- **Issue Clarification:** If the client's communication is unclear, the counselor can ask specific questions to ensure a full understanding of their situation.
- **Open-ended Questions:** The counselor can use open-ended questions to encourage the

client to provide more information and elaborate on their thoughts and feelings.

- Encouraging Self-Exploration: The counselor can ask questions that encourage the client to explore their feelings, thoughts, and behaviors in order to promote self-reflection.
- Cognitive Restructuring: Help the client identify and challenge unrealistic or distorted thought patterns, guiding them toward more balanced thinking.
- Guided Questioning: The counselor can use guiding questions to help the client focus on specific issues or goals, clarifying their thoughts and moving toward resolution.
- Non-judgmental and Accepting Attitude: Avoid making judgments about the client's experiences or emotions. Use neutral language and respect the client's perspectives and choices.

Rating Criteria.

- 5: Excellent – The counselor consistently demonstrates mastery of all principles. Their approach is empathetic, insightful, and highly effective in addressing the client's needs. The counselor maintains a strong, non-judgmental rapport while promoting self-awareness and growth.
- 4: Good – The counselor effectively applies the principles in most areas. There may be occasional lapses in one or two principles, but overall the counselor's approach is competent, and the client feels supported and understood.
- 3: Average – The counselor demonstrates adequate skills in applying the principles. However, there are noticeable gaps or inconsistencies in their practice. Some areas may need improvement to enhance the therapeutic process.
- 2: Below Average – The counselor struggles to effectively apply several principles. There are significant gaps in understanding or responding to the client's needs, resulting in limited therapeutic progress. The approach may sometimes feel disconnected or judgmental.
- 1: Poor – The counselor demonstrates minimal or no proficiency in applying the principles. Their responses are ineffective, and they may create an unhelpful or even detrimental therapeutic environment. The client is likely to feel unsupported or misunderstood.

K Evaluation Guidelines for Pairwise Response Selection

We present the evaluation guidelines for pairwise response selection in Figures 20 and 21.

```
f''' 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。如果两个
回复一样好，那么最后的选择为tie。

评估原则如下：
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：{ctx}
回复选项：{choices}

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。如果两个回复一样好，那么最后的选择为tie。
其中x属于[A, B, tie]，你的输出格式为如下：
- 选择的依据：先输出选择的依据
- 最后的选择：x'''
```

Figure 20: Prompt template for pairwise response selection. For the English version, refer to Figure 21.

```
f'''You're a professional psychotherapy supervisor. I will provide you
with a conversation history between a client and a therapist, and
you need to select the most appropriate response based on the
following principles. If both responses are equally good, choose
"tie" as the final answer.

The evaluation principles are as follows:
1. Active Listening: Ensure full understanding of the client's issues
   and feelings, and confirm the client's main concerns and emotions.
2. Empathy: Show understanding and care for the client's feelings,
   using empathetic language like "I understand how you feel" or "It
   sounds like you're really upset."
3. Issue Clarification: If the client is unclear, the therapist can ask
   for specific details to ensure full understanding of their
   situation.
4. Open-ended Questioning: The therapist can use open-ended questions
   to encourage the client to share more information.
5. Encouraging Self-Exploration: The therapist can ask questions that
   encourage the client to explore their feelings, thoughts, and
   behaviors to promote self-reflection.
6. Cognitive Restructuring: Help the client identify and challenge
   unrealistic or distorted thought patterns.
7. Guided Questioning: The therapist can use guiding questions to help
   the client focus on specific issues or goals.
8. Non-judgmental and Accepting Attitude: Avoid making judgments about
   the client's experiences or emotions. Use neutral language that
   respects the client's viewpoints and choices.

Conversation History: {ctx}
Response Options: {choices}

Based on these principles, first, consider which principle is most
important for the current conversation history, and then select the
response that fits best. If both responses are equally appropriate,
choose "tie" as the final answer.

Your output format should be as follows:
- Reasoning for choice: First, state the reasoning behind your choice.
- Final Choice: x'''
```

Figure 21: Prompt template for pairwise response selection. For the Chinese version, refer to Figure 20.

L Validation of Automatic Evaluation

L.1 Automatic Evaluation of Counseling Skills

To assess the reliability of GPT-4o in evaluating counseling skills, we analyzed the correlation between expert evaluation scores and GPT-generated scores. For this analysis, we randomly selected 500 test cases (4500 in total) from our models and baselines for each dimension. Spearman’s rank correlation was applied to compare the individual test case scores generated by GPT-4o with the expert evaluations. As shown in Table 5, significant correlations were observed across all metrics. These findings suggest that automatic evaluation can reliably approximate human assessments of counseling skills, thereby supporting the validity of our experimental results.

Metric	Correlation
Empathy	0.531**
Active Listening	0.547**
Issue Clarification	0.421**
Open-ended Questioning	0.422**
Encouraging Self-Exploration	0.514**
Cognitive Restructuring	0.412**
Guided Questioning	0.562**
Non-judgmental and Accepting Attitude	0.603**
OA	0.513**

Table 5: Spearman’s rank correlation between expert and GPT-generated scores. Significant correlations are marked with ** (p -value < 0.05).

L.2 Automatic Evaluation of Pairwise Response Selection

To assess the reliability of GPT-4o in evaluating the win rate between two anonymous candidate responses, we examined the correlation between expert evaluations and GPT-generated judgments. For this analysis, we randomly selected 3,000 test cases from our models and baselines. Cohen’s κ was applied to compare the judgments generated by GPT-4o with those of the expert evaluators, quantifying inter-rater agreement. The resulting κ value of 0.67 indicates substantial agreement (0.60–0.80). These findings suggest that automatic evaluation can reliably approximate human assessments of win rate, thereby reinforcing the validity of our experimental results.

M Psychological Counseling Books

We present all the professional counseling books used to compile the test set for model evaluation, as shown in Figure 22.

N Data Statistics

We present the data statistics for five datasets: \mathcal{D}_0 , \mathcal{D}_1 , \mathcal{D}_2 , \mathcal{D}_3 , and \mathcal{D}_4 , as shown in Table 6.

O Test Set

We present the data statistics for the test set in Table 7.

P Open-source Models

We present all the open-source models, including both backbone and counseling models, used in this paper, along with their corresponding official repositories on Hugging Face.

- Qwen2.5-7B-Instruct: <https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>
- deepseek-llm-7b-chat: <https://huggingface.co/deepseek-ai/deepseek-llm-7b-chat>
- gte-Qwen2-7B-instruct: <https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct>
- SoulChat: <https://huggingface.co/scutcyr/SoulChat>
- MeChat: <https://huggingface.co/qiuhuachuan/MeChat>
- PsyChat: <https://huggingface.co/qiuhuachuan/PsyChat>
- CPsyCounX: <https://huggingface.co/CAS-SIAT-XinHai/CPsyCounX>



- (1) <https://books.google.com/books/about/%E5%BF%83%E7%90%86%E5%92%A8%E8%AF%A2%E4%B8%8E%E6%B2%BB%E7%96%97%E7%9A%84%E6%A1%88%E4%BE%8B%E8%AF%84%E4%BC%B0%E5%92%8C.html?id=8aElxQEACAAJ>
- (2) <https://www.google.com/books/edition/%E8%AE%A4%E7%9F%A5%E8%A1%8C%E4%B8%BA%E7%96%97%E6%B3%95%E5%85%A5%E9%97%A8/4IT20AEACAAJ?hl=en>
- (3) https://books.google.com/books/about/%E8%AE%A4%E7%9F%A5%E8%A1%8C%E4%B8%BA%E6%B2%BB%E7%96%97.html?id=_mJd0AEACAAJ
- (4) <https://books.google.com/books/about/%E5%BF%83%E7%90%86%E5%92%A8%E8%AF%A2%E7%9A%84%E6%8A%80%E5%B7%A7%E5%92%8C%E7%AD%96%E7%95%A5.html?id=fELFswEACAAJ>
- (5) <https://books.google.com/books/about/%E5%BF%83%E7%90%86%E5%92%A8%E8%AF%A2%E4%B8%8E%E6%B2%BB%E7%96%97%E7%9A%84%E7%90%86%E8%AE%BA%E5%8F%8A%E5%AE%9E%E8%B7%B5.html?id=DsvMYgEACAAJ>
- (6) <https://books.google.com/books/about/%E8%AE%A4%E7%9F%A5%E7%96%97%E6%B3%95.html?id=F4ehAAAAACAAJ>
- (7) <https://books.google.com/books/about/ACT%E5%B0%B1%E8%BF%99%E4%B9%88%E7%AE%80%E5%8D%95.html?id=C-FUAEACAAJ>
- (8) <https://www.amazon.com.mx/%E5%8A%A8%E6%9C%BA%E5%BC%8F%E8%AE%BF%E8%B0%88%E6%89%8B%E5%86%8C-%E5%BF%83%E7%90%86%E5%AD%A6%E4%B9%A6%E7%B1%8D%E8%B0%88%E8%AF%9D%E7%96%97%E6%B3%95%E5%BF%83%E7%90%86%E5%92%A8%E8%AF%A2%E6%B2%BB%E7%96%97%E5%B8%88%E5%8F%82%E8%80%83-%E7%90%86%E5%92%A8%E8%AF%A2%E4%B8%8E%E6%B2%BB%E7%96%97%E7%9A%84%E6%8A%80%E5%B7%A7%E5%92%8C%E7%AD%96%E7%95%A5.html?id=Dh1NzwEACAAJ>
- (9) <https://books.google.com/books/about/%E5%BF%83%E7%90%86%E5%92%A8%E8%AF%A2%E4%B8%8E%E6%B2%BB%E7%96%97%E7%9A%84%E6%8A%80%E5%B7%A7%E5%92%8C%E7%AD%96%E7%95%A5.html?id=Dh1NzwEACAAJ>

Figure 22: Professional counseling books used to collect the test set for model evaluation.

Dataset	# Tokens	# Dialogues	Max Turns	Min Turns	Avg. Turns	Num. of client utterances	Num. of counselor utterances	Avg. Len. of client	Avg. Len. of counselor
\mathcal{D}_0	6461213	2382	148	9	38.7	92286	92286	32.7	24.3
\mathcal{D}_1	6435130	2382	118	9	37.6	89455	89455	33.9	25.0
\mathcal{D}_2	6422349	2382	126	9	37.6	89671	89671	33.6	25.0
\mathcal{D}_3	6420533	2382	117	8	37.6	89473	89473	33.7	25.0
\mathcal{D}_4	7017401	2382	120	9	37.8	90033	90033	33.7	31.1

Table 6: Data statistics for five datasets, including \mathcal{D}_0 , \mathcal{D}_1 , \mathcal{D}_2 , \mathcal{D}_3 , and \mathcal{D}_4 .

Avg. Turns	Avg. Len. of client	Avg. Len. of counselor
12.7	25.5	42.6

Table 7: Data statistics for the test set.

Q Results

Q.1 Expert Evaluation

For example, the primary concern expressed by the client is the difficulty of balancing the demands of his doctoral dissertation work with the need to attend to his girlfriend’s emotional needs, which has led to feelings of helplessness and distress. We provide two comparative example dialogues, as shown in Figure 23. Our observations suggest that our model demonstrates superior active listening skills and offers more encouraging responses, whereas CPsyCounX tends to rely on mechanical empathy and frequent question-asking. These results highlight that our model is superior to existing counseling models because active listening better enables clients to express their innermost feelings from the perspective of client-centered therapy.

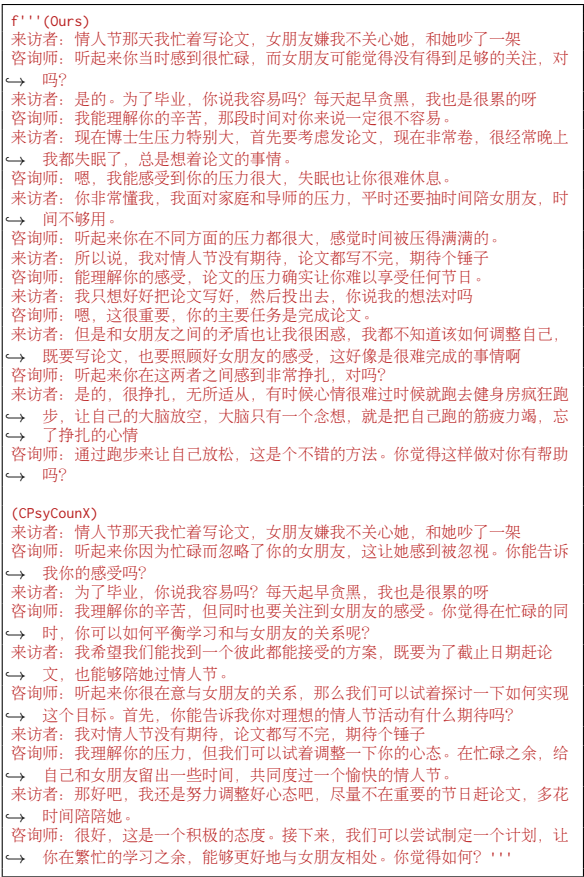


Figure 23: Comparative example dialogues. For the English version, see Figure 4.

Q.2 Counseling Skills Analysis

A comparison of counseling skills performance between our model (backbone: deepseek-llm-7b-chat) and the baseline models is presented in Table 8.

Q.3 Quantitative Analysis

An example with quantitative scores is presented in Figures 24 and 25. We can see that the overall score assigned by both the experts and GPT is identical.

Category	Models	Empathy ↑	Active Listening ↑	Issue Clarification ↑	Open-ended Questioning ↑	Encouraging Self-Exploration ↑	Cognitive Restructuring ↑	Guided Questioning ↑	Non-judgmental and Accepting Attitude ↑	OA ↑
Golden	Counselor Response	3.18	3.80	3.49	3.29*	3.54	2.75	3.79	4.48**	3.42*
Counseling Models	SoulChat	3.16	3.64**	3.24**	3.20**	3.31*	2.64	3.41**	4.37**	3.32**
	MeChat	3.22	3.68**	3.27**	3.34	3.37	2.68	3.56	4.43**	3.37**
	PsyChat	3.33*	3.78*	3.28**	3.37	3.47	2.70	3.61	4.50**	3.41*
	CPsyCounX	3.26	3.84	3.54	3.44	3.48	2.78	3.79	4.49**	3.49
Fine-tuned Models	π_0	3.16	3.69**	3.38	3.37	3.35	2.55	3.55	4.49**	3.36**
	π_1	3.02**	3.46**	3.15**	3.19**	3.18**	2.45	3.45**	4.46**	3.22**
	π_2	3.05**	3.50**	3.19**	3.23**	3.29*	2.55	3.43**	4.35**	3.16**
	π_3	3.04**	3.52**	3.19**	3.18**	3.27*	2.52	3.45**	4.43**	3.20**
	π_4	3.24	3.91	3.42	3.46	3.48	2.64	3.67	4.72	3.54

Table 8: Comparison results among different baselines compared to our SFT model π_4 (backbone model: deepseek-llm-7b-chat). Statistically significant differences compared to our model π_4 are marked with ** (p -value < 0.05), * (p -value < 0.1), as determined by the Mann-Whitney U test.

```

messages = [
  {
    "role": "user",
    "content": "你好。"
  },
  {
    "role": "assistant",
    "content": "你好，你今天想要讨论哪方面的话题呢？"
  },
  {
    "role": "user",
    "content": "我想跟你谈一谈我和我男朋友感情的问题。"
  },
  {
    "role": "assistant",
    "content": "嗯嗯。"
  },
  {
    "role": "user",
    "content": "就是我现在有点犹豫，要不要跟他分手。我跟男朋友在之间有一个学姐，他们关系很好。"
  },
  {
    "role": "assistant",
    "content": "是男朋友的学姐？"
  },
  {
    "role": "user",
    "content": "男朋友比我高一级，刚开始我和他都是学委，在工作中有过接触。因为他比我高一级，在很多事情上都会指导我，所以有一次我就用一个陌生的QQ试
    ↪ 探他到底有没有女朋友，他说没有，后来他感觉到是我，就跟我确认了。然后我就问他，“我做你女朋友好不好？”他说好，我们两个就在一起了。”
  },
  {
    "role": "assistant",
    "content": "嗯嗯，那刚刚你说到的学姐是一个怎样的情况呢？"
  },
  {
    "role": "user",
    "content": "我男朋友有一个小团体，里面是三男两女，这个学姐是团体中的一个，我男朋友之前有一个谈了6年的女朋友，从高中开始的。”
  },
  {
    "role": "assistant",
    "content": "哦，以前的女朋友。”
  },
  {
    "role": "user",
    "content": "对，后来不知道为什么那个前女友说了些狠话就出国了。那段时间我男朋友特别痛苦，这个学姐就跟他关系特别好，跟另一个男生一起就陪着他度过
    ↪ 这段时间。”
  },
  {
    "golden_response": {
      "role": "assistant",
      "content": "所以是另一个男生和这个学姐陪你的男朋友度过了之前分手后最艰难的那段时间。”
    }
  },
  {
    "Qwen7B_D0": "嗯嗯，好像学姐和男朋友的关系很亲近",
    "Qwen7B_D4": "听起来那段时间对你男朋友来说确实很难熬。”
  }
]

GPT_OA_scores = {
  "Qwen7B_D0": 2
  "Qwen7B_D4": 3
}

Expert_OA_scores = {
  "Qwen7B_D0": 2
  "Qwen7B_D4": 3
}

```

Figure 24: An example with quantitative scores. For the English version, see Figure 25.


```

messages = [
  {
    "role": "user",
    "content": "Hello."
  },
  {
    "role": "assistant",
    "content": "Hi there. What would you like to talk about today?"
  },
  {
    "role": "user",
    "content": "I'd like to talk to you about my relationship with my boyfriend."
  },
  {
    "role": "assistant",
    "content": "Okay."
  },
  {
    "role": "user",
    "content": "I'm currently unsure whether I should break up with him. There's a senior female student who's very close to him."
  },
  {
    "role": "assistant",
    "content": "She's his senior?"
  },
  {
    "role": "user",
    "content": "Yes, my boyfriend is one year above me. We first met because we were both class representatives and worked together. Since he's a
    ↳ year ahead, he often guided me in various matters. At one point, I wanted to know if he had a girlfriend, so I messaged him using an
    ↳ anonymous QQ account to test him. He said he didn't. Later, he realized it was me and asked me directly. I then asked him, 'Would you be my
    ↳ boyfriend?' He said yes, and that's how we got together."
  },
  {
    "role": "assistant",
    "content": "I see. So, what's the situation with this senior you mentioned?"
  },
  {
    "role": "user",
    "content": "My boyfriend is part of a small group—three guys and two girls. This senior is one of them. Before me, he had a girlfriend he'd been
    ↳ with for six years, since high school."
  },
  {
    "role": "assistant",
    "content": "His ex-girlfriend."
  },
  {
    "role": "user",
    "content": "Yes. For some reason, she said some harsh things and then went abroad. That breakup was really painful for him. During that time,
    ↳ this senior and another guy from the group were the ones who stayed by his side and helped him get through it."
  }
],
"golden_response": {
  "role": "assistant",
  "content": "So it was this senior and another male friend who supported him during the hardest time after the breakup."
},
"Qwen7B_D0": "Yes. It seems like this senior and my boyfriend have a very close relationship.",
"Qwen7B_D4": "It sounds like that period was especially difficult for your boyfriend."

GPT_OA_scores = {
  "Qwen7B_D0": 2
  "Qwen7B_D4": 3
}

Expert_OA_scores = {
  "Qwen7B_D0": 2
  "Qwen7B_D4": 3
}

```

Figure 25: An example with quantitative scores. For the Chinese version, see Figure 24.

Q.4 Win Rate Analysis

We present the results of the pairwise response evaluation in Figure 26.

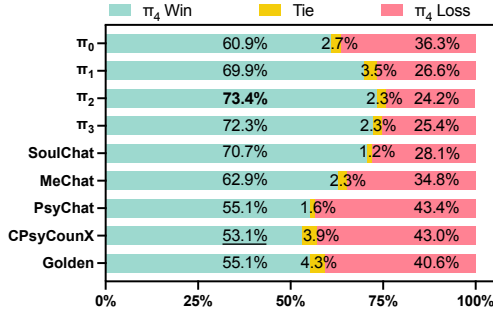


Figure 26: Results of pairwise response evaluation. Our model π_4 (backbone model: deepseek-llm-7b-chat) outperforms various baselines. The best and worst win rates are highlighted in bold and underlined.

Q.5 Ablation Study

We present the ablation study for data refinement with the backbone model deepseek-llm-7b-chat from two perspectives: counseling skills and win rate. The corresponding results are presented in Table 8 and Figure 27.

Counseling Skills. π_0 outperforms models π_1 , π_2 , and π_3 in all eight counseling skill components, as well as in the overall score. In contrast, while our model, π_4 , exceeds these four models in both the individual counseling skill components and the overall score, its performance further underscores the critical role of data refinement.

Win Rate. π_0 clearly outperforms models π_1 , π_2 , and π_3 in terms of win rate, as shown in Figure 27, with pairwise comparisons among π_1 , π_2 , and π_3 revealing comparable performance, indicating that these three models are essentially neck and neck. In stark contrast, π_4 stands out by significantly surpassing all four models, further emphasizing the critical impact of data refinement, as illustrated in Figure 26.

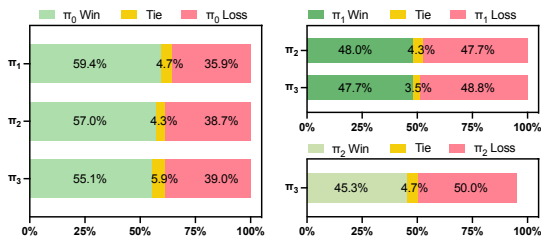


Figure 27: Results of pairwise response evaluation among models π_1 , π_2 , and π_3 (backbone model: deepseek-llm-7b-chat).

R Privacy Elimination

We present a t-SNE visualization of two dialogue datasets, \mathcal{D}_0 and \mathcal{D}_4 , from the client's perspective, as shown in Figures 28 and 29. These two figures demonstrate that the overlap of data points between \mathcal{D}_0 and \mathcal{D}_4 is significantly reduced, indicating that our method can effectively achieve privacy elimination.

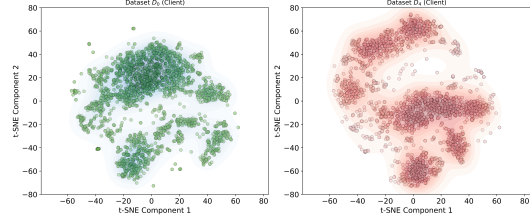


Figure 28: t-SNE visualization of two dialogue datasets, \mathcal{D}_0 and \mathcal{D}_4 , from the perspective of client.

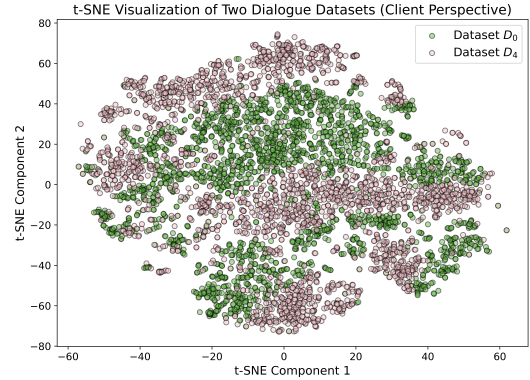


Figure 29: t-SNE visualization of two dialogue datasets, \mathcal{D}_0 and \mathcal{D}_4 , from the perspective of client.

S Dialogue Diversity

We present a t-SNE visualization of two dialogue datasets, \mathcal{D}_0 and \mathcal{D}_4 , from the counselor’s perspective, as shown in Figure 30, demonstrating that our dataset is comparably diverse to the real-world dataset.

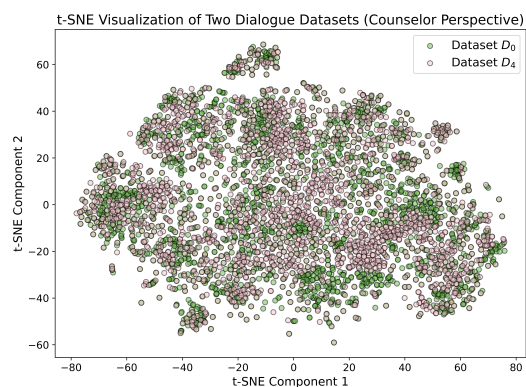


Figure 30: t-SNE visualization of two dialogue datasets, \mathcal{D}_0 and \mathcal{D}_4 .