SMILE: Single-turn to Multi-turn Inclusive Language Expansion via ChatGPT for Mental Health Support

Important: Our objective is to explore the potential of large language models to serve as virtual counselors for mental health support, and we do NOT advocate for their use as a substitute in therapeutic treatment without professional supervision.

Anonymous ACL submission

Abstract

Developing specialized dialogue systems for mental health support requires multi-turn conversation data, which has recently garnered increasing attention. However, gathering and releasing large-scale, real-life multi-turn conversations to facilitate advancements in mental health presents challenges due to data privacy protection, as well as the time and cost involved. To address the challenges related to data scarcity, we introduce SMILE, a single-turn to multi-turn inclusive language expansion technique with a solid theoretical foundation that prompts ChatGPT to transform public single-turn dialogues into multi-turn ones. Our study first focuses on basic characteristics, dialogue diversity, and quality among four large language models, verifying that our proposed method is superior to other baseline methods and that GPT-4o is the optimal option. Thus, we employ our method to generate a large-scale, diverse, and high-quality dialogue dataset named SMILECHAT, consisting of 13k dialogues. Finally, we utilize SMILECHAT to fine-tune six large language models, giving birth to mental health chatbots, MECHAT. Empirical evaluations demonstrate that MeChat excels in generating empathic, professional, helpful, and safe responses in mental health support, showing its high quality and practicality.

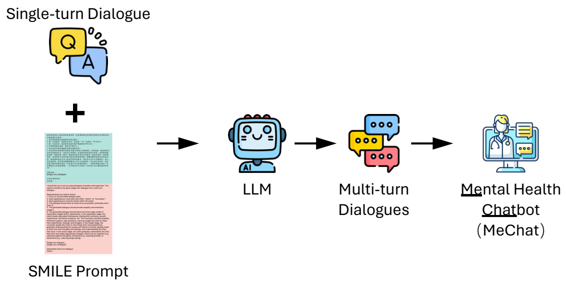

Figure 1: The SMILE method used to generate dialogues for mental health support.

1 Introduction

We all know the importance of mental health, and mental health issues (Kessler et al., 2005) have been a persistent concern for human beings. Recently, advancements in natural language processing (NLP) technology (Vaswani et al., 2017; Ouyang et al., 2022; Ni et al., 2022) have led to the emergence of neural-based conversational AI applied in various domains, including mental health (Liu et al., 2022; Tu et al., 2022). Virtual counselors powered by AI, as an innovative solution for mental health, can effectively address accessibility barriers such as the high cost of treatment and the shortage of trained professionals to meet the demand. Furthermore, such dialogue systems provide mental health support as an effective and practical online counseling approach for those in need, safeguarding user privacy and mitigating the stigma that often accompanies help-seeking. However, the lack of *publicly available, large-scale, diverse, and high-quality multi-turn chat datasets* in the mental health support domain hinders the development of specialized dialogue systems.

Motivation. Indeed, building a practical, safe, and effective conversational agent for mental health is a goal that many researchers have been pursuing. However, the first step in creating such a system is to have training data. Conversations related to mental health support often contain sensitive information and must be kept confidential (Lu et al., 2021) to safeguard the privacy of individuals seeking help. Making these conversations publicly available may discourage individuals from seeking support or negatively impact their personal and professional lives once known to people they are acquainted with. To facilitate progress in the NLP community, some researchers have attempted to collect various dialogue corpora (Liu et al., 2021; Sun et al., 2021; Zheng et al., 2022) through crowd-sourcing, data crawling, or data augmentation to build a dialogue agent capable of providing emotional and mental health support. How to construct a large-scale, diverse and high-quality multi-turn chat dataset for mental health motivates us to carry out the work as presented in this paper.

Challenges. To be more specific, crowd-sourcing conversations (Liu et al., 2021) for emotional support has limitations due to the high cost and time required to train and manage annotators, as well as the difficulty in mimicking real-life interactions, that is, interlocutors may lack an understanding of the dilemma of living with mental disorders. An alternative is crawling QA (Sun et al., 2021) on a public mental health forum for training psychological support models. However, single-turn conversations may not be sufficient for resolving mental health issues, as multiple interaction exchanges are often needed. Multi-turn conversations, which can better simulate real-world conversations, are therefore more practical for training psychological support models. While the post-triggered machine-augmented method (Zheng et al., 2022) can address the limitations of scale and topic diversity, it does not take into account the responses of experienced supporters.

Our Approach. To tackle the challenges mentioned above, we introduce SMILE, single-turn to multi-turn inclusive language expansion via ChatGPT, as shown in Figure 1. Specifically, we instruct ChatGPT to transform publicly available question-answer pairs (public QAs), which can also be viewed as single-turn dialogues, into multi-turn conversations. With the proposed method, we build a *large-scale, diverse, and high-quality multi-turn* conversation dataset for mental health support.

Our paper is organized as follows:

- We first present our method (§3), including data preprocessing and text generation, which mainly elaborates on the SMILE method and baseline methods.
- We set up experiments (§4) to demonstrate the superiority and effectiveness of the SMILE method.
- We demonstrate the superiority and effectiveness of the SMILE method through basic characteristics, dialogue diversity, and quality among four large language models (§5). Following the validation of its superiority and effectiveness and expert evaluation, we leverage the SMILE method to generate a large-scale and diverse multi-turn chat dataset, SMILECHAT, for mental health support.
- Finally, we propose training a dialogue system to explore the quality of conversation (§6) and collecting a 1073 real-life counseling dialogue sessions to construct a test set for model evaluation.

Our Contributions We make our data, code, and model publicly available. We believe our work offers a new perspective on constructing a large-scale, diverse, and high-quality multi-turn dialogue dataset for mental health within the research community. Our contributions can be summarized as follows:

- We introduce SMILE, which is based on a solid theoretical foundation (Hill, 2020) and provides a novel perspective for alleviating the scarcity of multi-turn conversations in mental health.
- Through the analysis of basic characteristics, dialogue diversity and quality, we verify the feasibility and effectiveness of our proposed method. This method can construct multi-turn dialogues based on medical, financial, and legal QAs, thereby alleviating the dialogue scarcity in other application domains.
- To better assess the quality of SMILECHAT, we collect a real-life counseling dataset with 47 counseling dialogues to build an real-life test set, PsyTest, which contains 1073 test samples.
- We release SMILECHAT, which comprises 13k Chinese multi-turn dialogues with an average of 15.7 turns. Additionally, we make our dialogue model, MECHAT, and real-life test set, PsyTest, publicly available.

2 Related Work

2.1 Applications of ChatGPT

ChatGPT has proven to be a powerful AI tool for various NLP tasks since its release. Currently, it is being utilized in several domains, such as conversational AI (Alessa and Al-Khalifa, 2023; Köpf et al., 2023; Chen et al., 2023), education (Küchemann et al., 2023; Eshghie and Eshghie, 2023), code programming (Dong et al., 2023; Yetiştiren et al., 2023) and healthcare (Zhao et al., 2023; Yang et al., 2023b).

Furthermore, ChatGPT's efficiency and cost-effectiveness have been well-documented, making it competitive to human annotators (Gilardi et al., 2023; Zhu et al., 2023) even in zero-shot accuracy tasks. Xu et al. (2023) have proposed the use of self-chatting, where ChatGPT engages in a conversation with itself, resulting in 111.5k dialogues collected from Quora and Stack Overflow sources and 47k conversations from the medical domain. Auto-GPT [1], an AI agent, is capable of breaking down a natural language goal into sub-tasks and using various tools and the internet in an automated loop to achieve the objective. Shen et al. (2023) have suggested using ChatGPT for task planning when receiving user inquiries, selecting appropriate models based on function descriptions from Hugging Face, executing each subtask using the chosen AI model, and summarizing the response based on the execution's outcomes.

In summary, ChatGPT has already demonstrated its enormous potential as an intelligent pipeline tool that can significantly advance NLP development, despite having only a restricted API available for researchers.

2.2 Datasets for Mental Health Support

Research on mental health support has significantly depended on the availability of publicly available datasets (Sun et al., 2021; Liu et al., 2021; Zheng et al., 2022) in recent years. The large-scale conversational datasets have enabled researchers to investigate various aspects of mental health, including identifying mental health conditions (Liu et al., 2023; Srivastava et al., 2022), understanding clients' reactions (Li et al., 2023), predicting support strategies (Sun et al., 2021; Li et al., 2023), deciding personalized interventions (Golden et al., 2023) and understanding response safety within a dialogue history (Qiu et al., 2023).

Liu et al. (2021) first define the emotional support conversation task and then, via crowdsourcing, construct ESConv, an emotional support conversation dataset containing 1053 dialogues with rich support strategies. However, the collection of ESConv requires high cost and time yet leads to a small-scale dialogue dataset. To this end, Zheng et al. (2022) present an approach for augmenting data scale with informative dialogue posts and then constructing AugESC, a model-synthesized dataset with 102k dialogues. The previous two datasets are limited to English. To facilitate the research in Chinese, hence Sun et al. (2021) crawl QA posts in a public mental health support platform compiling PsyQA.

[1]https://github.com/Significant-Gravitas/Auto-GPT

3 Method

A QA can be considered a single-turn dialogue. PsyQA[2], a high-quality Chinese dialogue dataset focused on mental health support, features one question mapped to multiple answers. The dataset was anonymized prior to release. Our dataset creation pipeline, based on PsyQA, includes two main stages: (1) QA preprocessing and (2) text generation.

3.1 QA Preprocessing

In short, this process involves QA filtering and wording cleaning.

QA Filtering Transforming single-turn dialogues into multi-turn ones requires rich contextual information. Therefore, we measure the Chinese character length of both client questions and counselor answers, retaining only instances where both exceed 300 characters. This process resulted in 13,709 QA pairs.

Wording Cleaning We aim to construct a large-scale, diverse, and high-quality multi-turn conversation corpus using the proposed SMILE method based on PsyQA. While QA can be considered a single-turn conversation between a real client and a counselor, there are differences in wording compared to actual multi-turn conversations. For instance, the term "楼主" (literally meaning "thread starter") frequently appears in QA but is rarely used in conversation. Therefore, we propose a two-stage process to clean the wording in PsyQA, mitigating linguistic discrepancies before rewriting QA into multi-turn conversations. This process includes both automatic and manual cleaning procedures. For a detailed process, please refer to Appendix A.

3.2 Text Generation

First, let us denote the input x as a sequence $\{x_1, x_2, ..., x_n\}$, and the output y as a sequence $\{y_1, y_2, ..., y_m\}$. The generation process of the language model can be expressed as the conditional probability distribution $p(y|x)$, which represents the probability of generating output y given the input x. Therefore, text generation via the large language model can be formulated as follows:

$$p(y|x) = \prod_{t=1}^{m} p(y_t|y_{<t}, x) \tag{1}$$

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

where y_t represents the t-th token generated by the model. However, x is our main focus in this paper, next we will demonstrate the details of prompt design.

3.2.1 Prompt Design

In this section, we mainly focus on describing prompt design. To provide a clearer understanding of our method in a more controllable setting and elucidate the superiority of introducing single-turn dialogues as a reference, we first design two baseline prompts for comparison.

Standard Prompt The standard prompt does not contain any single-turn dialogues or specific dialogue topic and instead uses only the simplest prompt to generate multi-turn dialogues. The standard prompt is illustrated in Figure 6 in Appendix B. The input in Equation 1 is $x = I$, where I represents the standard prompt. We simplify the method name as `standard` and consider this method as our baseline.

Standard Prompt with a Specific Dialogue Topic Intuitively, feeding a single, fixed prompt into a large language model often results in the generation of low diversity. Therefore, we provide a specific dialogue topic for the standard prompt. The input in Equation 1 is $x = (I, T)$, where T represents the dialogue topic chosen in uniform sampling in the topic set. We simplify the method name to `standardT` and adopt this method as our baseline, as illustrated in Figure 7 in Appendix B.

SMILE Prompt Our paper aims to highlight the superiority of the introduction of single-turn dialogues during generating dialogues. Our proposed method, referred to as the SMILE method, instructs the ChatGPT to rewrite single-turn dialogues into multi-turn ones. Figure 8 depicts the concrete prompt details. The input in Equation 1 is $x = (I, T, D)$, where T and D represent the dialogue topic hidden in the QA and single-turn dialogue, respectively.

4 Experiments

4.1 Large Language Models for Generation

In this paper, we propose to use GLM-4[3], DeepSeek-V2-Chat[4] (also known as DeepSeek), GPT-3.5-Turbo[5], and GPT-4o[6] to generate dialogue given a specific prompt. Based on official recommendations, we set the temperature and top-p for GLM-4 to 0.7 and 0.9, respectively. For DeepSeek, GPT-3.5-Turbo, and GPT-4o, we set both the temperature and top-p to 1.0. For all four models, we set the maximum number of tokens for generation to 4000.

4.2 Dialogue Topic Collection

To address the issue of monotonous generation, we collaborate with three professional counselors, refer to existing literature (Rickwood et al., 2007; Pedrelli et al., 2015), and ultimately compile a comprehensive set of dialogue topics. This set comprises 60 distinct types, each accompanied by a corresponding explanation. For more details, please refer to Appendix G.

4.3 QA Sampling

Importantly, we have obtained 13,709 QA pairs. However, to emphasize the superiority of the SMILE method, we propose to sample 500 QAs to conduct preliminary experiments among four large language models. To ensure a fair comparison and prevent repeated occurrences of the same question, we first randomly select 500 non-duplicate questions. We then randomly choose one answer to serve as the corresponding response. The data samples obtained are employed as seed dialogues for the SMILE method, which are subsequently restructured into multi-turn conversations via large language models.

4.4 Text Representation

We utilize three prompts to instruct four large language models, and each generate 500 dialogues. Therefore, we will analyze from the perspectives of methods and models.

A multi-turn dialogue between a client and a counselor is represented as

$$d = \{(u_1, r_1), ..., (u_i, r_i), ..., (u_n, r_n)\} \quad (2)$$

where u_i and r_i represents the utterances of the i-th turn spoken by the client and counselor, respectively. A string of a dialogue without any speaker role tokens can be denoted as $d_s =$

[3]https://open.bigmodel.cn/

[4]https://www.deepseek.com/

[5]The model we use is gpt-3.5-turbo-0125, with training data up to Sep 2021.

[6]The model we use is gpt-4o-2024-05-13, with training data up to Oct 2023.

Models	Avg. Turns			Avg. Characters		
	standard	standardT	SMILE	standard	standardT	SMILE
GLM-4	9.3	9.5	9.1	23.1	23.6	26.3
DeepSeek	10.9	11.2	11.5	19.3	20.4	29.3
GPT-3.5-Turbo	4.3	4.2	4.2	33.8	31.7	60.5
GPT-4o	**12.2**	**11.9**	**14.7**	19.2	18.4	18.1

Table 1: Analysis of basic characteristics. The best results are highlighted in bold, the worst results are highlighted in red.

$[u_1; r_1; u_2; r_2; ...; u_n; r_n]$, where $[;]$ denotes the operation of textual concatenation.

Text representation is used for analyzing semantic diversity. To obtain the text embedding of a dialogue, we use the BAAI/bge-m3 model[7], which accepts a maximum of 8192 tokens. Each dialogue is first preprocessed into a single string without any speaker tokens and then mapped to a 1024-dimensional vector. For example, to compute the cosine similarity between two different dialogues, we can obtain

$$\cos(d_i, d_j) = \frac{e_i \cdot e_j}{\|e_i\| \, \|e_j\|} \tag{3}$$

where e_i and e_j denote the text embeddings from two distinct dialogues.

5 Results

5.1 Basic Characteristics

We present the basic analysis of average turns and average characters in Table 1. More dialogue turns indicate better performance. The model GPT-4o achieves the best performance, while GPT-3.5-Turbo achieves the worst results in terms of dialogue turns. Furthermore, GPT-3.5-Turbo obtains unsatisfactory results in terms of average Chinese characters.

5.2 Dialogue Diversity

To demonstrate the effectiveness of the SMILE method, we mainly focus on three aspects of diversity: lexical features, semantic features, and dialogue topics.

5.2.1 Lexical Features

For lexical analysis, we utilize the popular used Chinese tokenizer Jieba[8] to tokenize the dialogue. To measure the lexical features, we adopt distinct-n ($n = 1, 2, 3$) metrics (Li et al., 2016), which are widely used for measuring the diversity of dialogue datasets. Each dialogue is first preprocessed into a single string without any speaker tokens. We provide statistics for 500 dialogues per prompt method, as presented in Table 2.

Our proposed SMILE method results in rich vocabularies, with significantly higher numbers of unique unigrams, bigrams, and trigrams compared to the baseline methods. Specifically, a simple and fixed prompt tends to produce monotonous output, whereas incorporating dialogue topics into a single, fixed prompt leads to substantial diversification in the output. However, the SMILE method outperforms two baseline methods in terms of Distinct-1, Distinct-2, and Distinct-3 across four large language models.

5.2.2 Semantic Features

To measure the semantic diversity of a dialogue dataset, we suggest calculating the cosine similarity between every pair of different dialogues. This involves computing the pairwise cosine similarity for each pair of distinct dialogues, resulting in $\binom{500}{2}$ pairs and their corresponding cosine values, as described in Equation 3.

We present the results in Figure 2, which demonstrates that the median of the SMILE method is significantly lower than those of the baseline methods. The SMILE method exhibits the most extensive semantic diversity. Further, GPT-4o obtains the best performance.

5.2.3 Dialogue Topics

To measure the diversity of dialogue topics in a dialogue dataset, we utilize information entropy to measure the diversity of topic distribution. *The higher the information entropy, the more uniform the distribution, indicating greater diversity.* The formula for calculating information entropy (Rényi, 1961; Lin, 1991) is as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{4}$$

where $H(X)$ is the information entropy. $p(x_i)$ is the probability of the occurrence of topic x_i.

To obtain dialogue topics for each dialogue in each prompt method, we design a prompt provided with 60 distinct dialogue topics, as illustrated in Appendix G and Figure 9, to automatically label dialogue topics for each dialogue with Qwen1.5-110B-Chat. We present the information entropy for each prompt method in Table 3, demonstrating that the dialogues generated using the SMILE method

[7] https://huggingface.co/BAAI/bge-m3
[8] https://github.com/fxsjy/jieba

Model	Method	Total Unigrams	Unique Unigrams	Distinct-1 (⇑)	Total Bigrams	Unique Bigrams	Distinct-2 (⇑)	Total Trigrams	Unique Trigrams	Distinct-3 (⇑)
GLM-4	standard	140095	2573	0.018	139595	19328	0.138	139095	44760	0.322
	standardT	145194	3630	0.025	144694	27305	0.189	144194	60642	0.421
	SMILE	158811	5758	0.036	158311	40559	0.256	157811	85254	0.540
DeepSeek	standard	137997	2126	0.015	137497	15113	0.110	136997	33944	0.248
	standardT	149544	3197	0.021	149044	24156	0.162	148544	53403	0.360
	SMILE	221763	7156	0.032	221263	50561	0.229	220763	106986	0.485
GPT-3.5-Turbo	standard	91686	2238	0.024	91186	16108	0.177	90686	36664	0.404
	standardT	84324	3059	0.036	83824	20708	0.247	83324	43909	0.527
	SMILE	163760	6613	0.040	163260	46538	0.285	162760	98236	0.604
GPT-4o	standard	153760	3562	0.023	153260	29902	0.195	152760	71735	0.470
	standardT	144117	4530	0.031	143617	35918	0.250	143117	79236	0.554
	SMILE	**178595**	**7833**	**0.044**	**178095**	**55716**	**0.313**	**177595**	**113213**	**0.637**

Table 2: Statistics of 500 conversations in each prompt method.

Method	GLM-4	DeepSeek	GPT-3.5-Turbo	GPT-4o	Avg.
standard	3.43	3.13	3.08	3.44	3.27
standardT	4.51	4.50	4.85	4.76	**4.66**
SMILE	4.37	4.55	4.32	4.59	4.46($\Delta = -0.2$)

Table 3: Information entropy of dialogue topics.

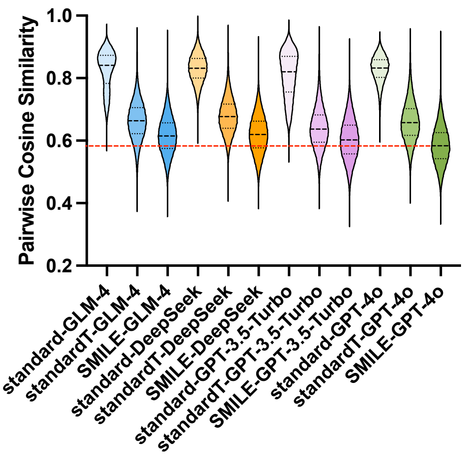

Figure 2: Pairwise dialogue cosine similarity among four settings: our proposed three methods and a reference point using sampled data from PsyQA.

are substantially more diverse than those generated using the `standard` method and are compatible with the `standardT` method, which uniformly samples dialogue topics.

5.3 Dialogue Quality

To comprehensively evaluate the quality of generated dialogues, we propose an assessment framework that includes five perspectives: professionalism, informativeness, helpfulness, empathy, and safety. In collaboration with experts[9] in psychological counseling, we propose a set of assessment criteria, as shown in Figure 11. We randomly sample 20 dialogues for each method from four large language models. Expert evaluations demonstrate that GPT-4o with the SMILE obtain the best performance.

5.4 SMILECHAT Dataset

Through the analysis of basic characteristics, dialogue diversity and expert evaluation, we conclude that the proposed method can generate a **diverse** and **high-quality** chat dataset. Therefore, we utilize the SMILE method to guide GPT-4o in generating all multi-turn conversations based on 13k QAs one round, leading to a **large-scale** dialogue dataset.

Through the SMILE method with GPT-4o, we compile a collection of 13709 conversations, SMILECHAT. Table 5 presents the statistics of the collected corpus. We present a dialogue generated by GPT-4o in Figure 10.

6 Dialogue System

We aim to build a high-quality multi-turn chat dataset for mental health. Therefore, we also analyze the dialogue quality based on the performance of the dialogue system trained with SMILECHAT.

6.1 Task Formulation

To train a dialogue system for mental health, we need to split each dialogue into several training sessions. Specifically, a sampled t-turn dialogue session can be represented as follows:

$$d_t = \{(u_1, r_1), (u_2, r_2), ..., (u_t, r_t)\} \sim D \quad (5)$$

We build a dialogue model that can predict the counselor's utterance r_t based on the dialogue history $h_t = \{u_1, r_1, u_2, r_2, ..., u_t\}$. Our objective is to maximize the likelihood probability as follows:

$$\mathcal{L} = -\sum_{t=1}^{L} \log p(r_t | u_1, r_1, ..., u_t) \quad (6)$$

[9]One is a Ph.D. in psychology and holds State-Certificated Class 3 Psycho-counselor with 4 years of experience in counseling. Another two individuals are State-Certificated Class 3 Psycho-counselors with a master's degree. The last one is a doctoral student majoring in computer science and is the first author of this paper.

Model	Professionalism			Informativeness			Helpfulness			Empathy			Safety		
	standard	**standardT**	**SMILE**	**standard**	**standardT**	**SMILE**	**standard**	**standardT**	**SMILE**	**standard**	**standardT**	**SMILE**	**standard**	**standardT**	**SMILE**
GLM-4	8.02	9.36	11.01	10.21	10.47	14.21	11.31	11.36	12.00	5.00	5.00	6.00	3.00	3.00	3.00
DeepSeek	8.21	9.47	11.43	12.38	13.79	14.57	11.42	11.50	12.00	5.00	5.00	6.00	3.00	3.00	3.00
GPT-3.5-Turbo	5.67	5.63	5.72	9.65	9.87	10.24	6.07	6.21	6.32	5.00	5.00	6.00	3.00	3.00	3.00
GPT-4o	8.01	9.56	**11.82**	12.74	13.89	**14.95**	11.43	11.54	**12.00**	5.00	5.00	**6.00**	3.00	3.00	**3.00**

Table 4: Results of human evaluation. The best results are highlighted in bold.

Category	Total	Client	Counselor
# Dialogues	13709	-	-
# Utterances	430400	215200	215200
# Avg. turns per dialogue	15.7	-	-
Avg. utterances per dialogue	31.4	15.7	15.7
Avg. length per utterance	18.8	18.7	19.0

Table 5: Statistics of the dialogue dataset, SMILECHAT.

where L is the sequence length of r_t.

6.2 Experimental Setup

Baseline Model To validate the dialogue quality of our collected dataset, we conduct fine-tuning experiments on six popular large language models, including Qwen1.5-4B-Chat, Qwen1.5-7B-Chat (Bai et al., 2023), Baichuan2-7B-Chat (Yang et al., 2023a), deepseek-llm-7b-chat (DeepSeek-AI et al., 2024), internlm2-chat-7b (Cai et al., 2024), and Yi-1.5-6B-Chat (AI et al., 2024).

Training Data To meet the data format requirements for instruction-based fine-tuning, we split the dialogue into multiple sessions, with the counselor's last utterance concluding each session. Additionally, we incorporate the system prompt (detailed in Appendix E) as a prefix to dialogue messages, following OpenAI's data format.

Parameter-efficient Fine-tuning To preserve the original capabilities of the model while adapting to downstream dialogue tasks and reducing computational costs, we employ Low-Rank Adaptation (LoRA, (Hu et al., 2021)) on all linear layers in the model for efficient fine-tuning.

Hyperparameters In this paper, all experiments are conducted on the NVIDIA A100 80G GPU for model training. During training, we set the training batch size to 4 on per device, and set the step of gradient accumulation to 2, meaning that gradient from every 2 steps would be accumulated and then used for parameter update. The learning rate is 1e-4. We adopt the cosine-type learning rate scheduler to adjust the learning rate throughout the training process. The entire training will span across 4 epochs. To accelerate the training and balance model performance, we also enable the use of 16-bit half-precision floating point numbers. In this paper, we implement the fine-tuning based on LLaMA Factory (Zheng et al., 2024), an efficient model tuning framework.

6.3 Evaluation

6.3.1 Automatic Evaluation

Metrics To conduct automatic evaluation, the evaluation metrics we use consist of BLEU-1/2/3/4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), Rouge-1/2/L (Lin, 2004), Distinct-1/2/3 (D-1/2/3) (Li et al., 2016), BERTScore (Zhang et al., 2019), and character difference between average predicted characters and average golden characters.

Test Set To better understand and assess the dialogue quality of SMILECHAT dataset, we propose to utilize real-life multi-turn counseling conversations. We develop an online mental health support platform that enables professional counselors to offer each client a free text-based counseling service, lasting approximately standard 50 minutes each time. We collect 47 real counseling dialogues. To protect user privacy, we ask experts to conduct a data anonymization process, removing information related to user identification (e.g., names, address). Then we split each long dialogue into multiple small sessions with the last utterance spoken by the counselor. We name this test set PsyTest, which contains 1073 test samples with a maximum turn of 10, and the average characters of counselor responses is 24.45.

Results The results of the automatic evaluation, including 10 metrics, are presented in Table 6. Notably, the evaluated dialogues are based on real-world counseling data rather than generated dialogues, which excludes the influence stemming from synthetic data. In psychological counseling, automatic evaluation metrics can not measure the performance of models fine-tuned with domain-specific data. It is worthy noting that, the character difference of Qwen1.5-4B-Chat is 28.02 and it obtain better results on 9 metrics. However, average predicted characters are significantly larger than that in golden responses.

Method	Model	BLEU-1	BLEU-2	BLEU_3	BLEU-4	METEOR	Rouge-1	Rouge-2	Rouge-L	BERTSCORE	Character Difference (Δ)
Baseline	Qwen1.5-4B-Chat	13.65	5.29	2.93	1.88	14.33	18.42	3.08	15.56	0.6079	28.02
Fine-tuned		13.05	4.47	2.31	1.58	11.2	17.55	2.13	14.58	0.5938	2.42
Baseline	Qwen1.5-7B-Chat	8.68	2.34	1.02	0.65	11.79	13.01	1.1	9.69	0.6021	82.68
Fine-tuned		**13.81**	**5.13**	**2.74**	**1.8**	**12.3**	**18.51**	**2.74**	**15.43**	0.5947	1.23
Baseline	Baichuan2-7B-Chat	11.15	4.24	2.21	1.41	12.85	16.55	2.7	13.38	0.6084	45.91
Fine-tuned		**13.25**	**4.7**	**2.49**	**1.7**	**11.61**	**18.11**	**2.42**	**14.97**	0.5918	1.05
Baseline	deepseek-llm-7b-chat	12.2	4.9	2.53	1.54	14.71	17.83	3.24	14.15	0.6074	51.0
Fine-tuned		**12.38**	4.48	2.36	**1.59**	10.73	17.82	2.42	**14.73**	0.5991	0.13
Baseline	internlm2-chat-7b	5.8	1.95	0.85	0.48	11.5	10.97	1.34	7.64	0.5859	151.67
Fine-tuned		**13.92**	**4.98**	**2.64**	**1.74**	**12.26**	**18.65**	**2.59**	**15.05**	**0.5947**	1.93
Baseline	Yi-1.5-6B-Chat	11.39	3.84	1.93	1.25	12.23	16.51	2.02	13.0	0.5928	42.13
Fine-tuned		**12.74**	**4.22**	**2.16**	**1.51**	**10.76**	**17.66**	**1.97**	**14.21**	0.5874	2.26

Table 6: Results of automatic evaluation in PsyTest dataset. Character Difference(Δ) = Average Predicted Characters - Average Golden Characters. The best results are highlighted in bold after fine-tuning.

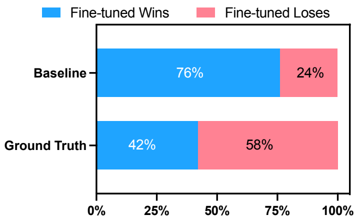

Figure 3: Human evaluation results, comparing the fine-tuned model to the baseline and golden. We present the win and lose rates of each compared pair in 100 randomly sampled sessions. Fleiss' kappa (Fleiss et al., 1981) is used to measure the inter-rater agreement, and all values fall within moderate agreement with $0.5 \leq \kappa \leq 0.6$.

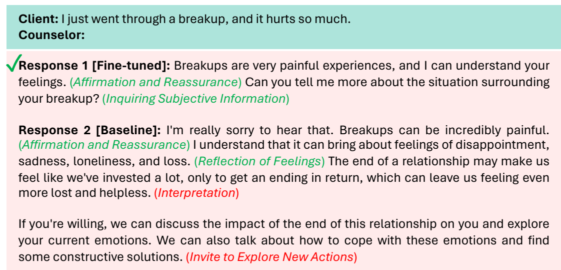

Figure 4: Case study. Counseling strategies used in the two responses are presented in parentheses. Strategies in green are supportive, while those in red are challenging and should not be used in the early stages of counseling.

6.3.2 Human Evaluation

Metrics We conduct a human evaluation to study the model performance trained with our proposed dialogue corpus. We select Qwen1.5-7B-Chat-fine-tuned as our mental health chatbot (MeChat). First, we randomly sample 100 sessions from PsyTest. Subsequently, we obtain 100 generated responses from the MeChat model. Three professional counselors are then presented with a dialogue history and three randomly shuffled responses (baseline, fine-tuned, ground truth). They are tasked with selecting the optimal response for the dialogue history, considering aspects illustrated in Section 5.3. The evaluation is conducted based on the criterion in Figure 11.

Results We employ majority voting to reach final decisions among three professional counselors. As depicted in Figure 3, the model trained with SmileChat demonstrates a significant performance improvement compared to the baseline model. Moreover, the responses generated by MeChat surpass the golden response to some extent in 42% of all dialogue sessions.

Case Study We present a case study, as shown in Figure 4. First, during counseling conversations, the text should not be too long for the client to understand. Second, at the beginning of a conversation, a counselor generally will not challenge the client; instead, they will provide support. Third, response 1 is more professional, informative, helpful, empathic, and safe. Therefore, fine-tuning LLMs is indispensable in mental health support.

7 Conclusion

This paper introduces SMILE, an effective solution for the scarcity of multi-turn conversations in mental health. By analyzing basic characteristics, dialogue diversity, and quality, we confirm the superiority and effectiveness of our method. Our approach enables the automatic creation of SmileChat, a large-scale, diverse, and high-quality dialogue corpus with 13k dialogues averaging 15.7 turns each. Both automatic and human evaluations using the PsyTest dataset, comprising 1,073 test samples, demonstrate that SmileChat significantly enhances dialogue system performance in mental health. With the release of SmileChat, our dialogue models (MeChat), and the authentic test set (PsyTest), we provide valuable resources to the research community.

Limitations

We release a large-scale, diverse, and high-quality multi-turn conversational dataset for mental health support, generated by rewriting single-turn conversations into multi-turn conversations using LLMs. Consequently, the dataset unavoidably incorporates LLMs' model knowledge. Furthermore, we discuss how LLMs do not fully utilize the rich vocabulary and content of single-turn conversations, as reflected in the distinct-n metric.

Furthermore, there is currently no comprehensive dataset available for evaluating the effectiveness of mental health support models automatically. Therefore, we have identified this as a limitation of our paper and intend to address it in future work.

Ethical Considerations

Our research is reviewed and approved by the xxxx University Institutional Ethics Committee (xxxxxx).

PsyQA is a single-turn dialogue dataset collected from an online mental health support platform. Specifically, help-seekers submit a post containing their mental health states and issues, and many professional counselors write down their responses to user questions to assist help-seekers. Therefore, PsyQA is a high-quality Chinese dialogue related to mental health support in the form of one question mapping to multiple answers.

Following the data copyright guidelines formulated by PsyQA (Sun et al., 2021), we release the multi-turn dialogue corpus publicly available for research community. If researchers wish to reproduce the multi-turn dialogues using PsyQA, they must sign an agreement with the original data owner. Accordingly, we release our datasets and models for research purposes, thus facilitating further advancement in the academic community.

Data Sharing Considering the nature of psychological counseling data, we must cautiously share this dataset. Regarding the rules for releasing data, third-party researchers who require access to the PsyTest dataset must provide us with their valid ID, proof of work, the reason they are requesting the data (e.g., the research questions), etc. They are required to be affiliated with a non-profit academic or research institution. This includes obtaining the approval of an Institutional Review Board (IRB), having principal investigators working full-time, as well as obtaining written approval from the institution's Office of Research or equivalent office. Additionally, they must sign the Data Nondisclosure Agreement and promise not to share the data with anyone.

Expert Salary Each expert was paid 300 RMB for their work per day, which is higher than the average wage (250 RMB/day) in our city. Overall, we have guaranteed that our salary level is competitive in our city.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

Abeer Alessa and Hend Al-Khalifa. 2023. Towards designing a chatgpt conversational companion for elderly people.

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

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai

Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report.

Sibei Chen, Hanbing Liu, Weiting Jin, Xiangyu Sun, Xiaoyao Feng, Ju Fan, Xiaoyong Du, and Nan Tang. 2023. Chatpipe: Orchestrating data preparation program by optimizing human-chatgpt interactions.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi 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.

Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2023. Self-collaboration code generation via chatgpt.

Mahshid Eshghie and Mojtaba Eshghie. 2023. Chatgpt as a therapist assistant: A suitability study.

Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Grace Golden, Christina Popescu, Sonia Israel, Kelly Perlman, Caitrin Armstrong, Robert Fratila, Myriam Tanguay-Sela, and David Benrimoh. 2023. Applying artificial intelligence to clinical decision support in mental health: What have we learned? *arXiv preprint arXiv:2303.03511*.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ronald C. Kessler, Olga Demler, Richard G. Frank, Mark Olfson, Harold Alan Pincus, Ellen E. Walters, Philip Wang, Kenneth B. Wells, and Alan M. Zaslavsky. 2005. Prevalence and treatment of mental disorders, 1990 to 2003. *New England Journal of Medicine*, 352(24):2515–2523.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment.

Stefan Küchemann, Steffen Steinert, Natalia Revenga, Matthias Schweinberger, Yavuz Dinc, Karina E. Avila, and Jochen Kuhn. 2023. Physics task development of prospective physics teachers using chatgpt.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023. 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.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Chang Liu, Xu Tan, Chongyang Tao, Zhenxin Fu, Dongyan Zhao, Tie-Yan Liu, and Rui Yan. 2022. ProphetChat: Enhancing dialogue generation with simulation of future conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 962–973, Dublin, Ireland. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Yuhan Liu, Anna Fang, Glen Moriarty, Robert Kraut, and Haiyi Zhu. 2023. Agent-based simulation for online mental health matching. *arXiv preprint arXiv:2303.11272*.

Yujie Lu, Chao Huang, Huanli Zhan, and Yong Zhuang. 2021. Federated natural language generation for personalized dialogue system. *arXiv preprint arXiv:2110.06419*.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2022. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, pages 1–101.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Paola Pedrelli, Maren Nyer, Albert Yeung, Courtney Zulauf, and Timothy Wilens. 2015. College students: mental health problems and treatment considerations. *Academic psychiatry*, 39:503–511.

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.

Alfréd Rényi. 1961. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press.

Debra J Rickwood, Frank P Deane, and Coralie J Wilson. 2007. When and how do young people seek professional help for mental health problems? *Medical journal of Australia*, 187(S7):S35–S39.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface.

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. *arXiv preprint arXiv:2106.01702*.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023a. Baichuan 2: Open large-scale language models.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023b. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis.

Burak Yetiştiren, Işık Özsoy, Miray Ayerdem, and Eray Tüzün. 2023. Evaluating the code quality of ai-assisted code generation tools: An empirical study on github copilot, amazon codewhisperer, and chatgpt.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities?

Chujie Zheng, Sahand Sabour, Jiaxin Wen, and Minlie Huang. 2022. Augesc: Large-scale data augmentation for emotional support conversation with pre-trained language models. *arXiv preprint arXiv:2202.13047*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks.

A Details of Data Cleaning

A.1 Automatic Cleaning

We employ a sequential data cleaning pipeline to swiftly replace words that are unsuitable to the conversation scenario. For example, both "楼主你" (literally *thread starter you*) and "楼主" (literally *thread starter*) should be replaced with "你" (you). However, it is necessary to perform the former replacement to avoid the repetition of "你" and the resulting "你你" (*you-you*).

A.1.1 Word List for Data Cleaning

To avoid the repetition of "你" (*you*) and the resulting "你你" (*you-you*) , we suggest to conduct a sequential word replacing pipeline. Figure 5 shows the word list for data cleaning and corresponding order for automatic cleaning.

Old String (ZH)	Old String (EN)	New String (ZH)	New String (EN)
'嗨，'	Hi,	''	/
'楼主你'	thread starter you	'你'	you
'题主你'	thread starter you	'你'	you
'楼楼你'	thread starter you	'你'	you
'楼主'	thread starter	'你'	you
'题主'	thread starter	'你'	you
'楼楼'	thread starter	'你'	you
'阿凉'	A Liang (a name)	'我'	me
'答主'	respondent	'人'	others

Figure 5: Word list for automatic cleaning.

A.2 Manual Cleaning

Due to the specificity and complexity of language, manual cleaning remains an essential part of the process. To prevent virtual dialogue systems from exhibiting overly frequent anthropomorphic behavior, we identify instances of the Chinese word for "hug" (抱抱) and manually delete sentence snippets containing this term.

我希望你担任心理咨询师和督导师。你需要生成一段多轮对话。
你需要遵守的要求：
1. 首先对话主题自行选取。
2. 每一次说话时，说话者必须以“来访者：”或“咨询师：”作为开头。
3. 每一次说话时，说话者的说话字数严格控制在30字以内。
4. 对话轮数越多越好，最好是10轮以上。
5. 所生成的对话需要提供共情和情感支持。
6. 所生成的对话需要符合探索-领悟-行动的三阶段模式。具体来说，探索阶段多由来访者讲述自己，诉说自己的困惑、成长经历和家庭状况等等。咨询师则提供共情、情感支持、提问，鼓励来访者分享自己的经历、感受和想法。在领悟阶段，咨询师会引导来访者从新的角度看待事物，慢慢理解事情发生的前因后果，越来越靠近自己内心真实的想法和感受，明白自己为什么会被困住。进入到行动阶段，来访者会慢慢理解自己为什么会被困住，作出适当的改变，这些改变可能是思维方面的（比如减少自己的消极想法），可能是情感方面的（比如降低自己的焦虑情绪），也可能是行为方面的（比如减少自己暴饮暴食的行为）。

生成的多轮对话：
来访者：

I would like you to act as a psychological counselor and supervisor. You need to generate a multi-turn dialogue.

Requirements you need to follow:
1. First the topic of the dialogue is self-selected.
2. Each speaking turn must start with either "Client:" or "Counselor:".
3. Each speaking turn must be strictly within 30 words.
4. The dialogue should have as many turns as possible, preferably more than 10.
5. The generated dialogue should provide empathy and emotional support.
6. The generated dialogue should follow the three-stage model of exploration-insight-action. Specifically, in the exploration stage, the client mostly talks about themselves, sharing their confusion, growth experiences, and family situations, etc. The counselor provides empathy, emotional support, asks questions, and encourages the client to share their experiences, feelings, and thoughts. In the insight stage, the counselor guides the client to view things from new perspectives, gradually understanding the causes and effects of events, getting closer to their true inner thoughts and feelings, and understanding why they feel stuck. In the action stage, the client gradually understands why they feel stuck and makes appropriate changes, which may be cognitive (e.g., reducing negative thoughts), emotional (e.g., lowering anxiety), or behavioral (e.g., reducing binge eating).

Generated multi-turn dialogue:
Client:

Figure 6: The `standard` method used to generate dialogues for mental health support.

B Method

We present the `standard`, `standardT` and `SMILE` prompts in Figures 6, 7, and 8, respectively.

C Dialogue Topics Annotation

In this paper, to label the dialogue topics of generated dialogues, the hyperparameters of Qwen1.5-110B-Chat during generation we used are set to the officially recommended default values, where temperature $\tau = 0.7$ and top-p $p = 0.8$. Figure 9 shows the prompting template of dialogue topics annotation.

D Dialogue Example

We present a dialogue generated by GPT-4o in Figure 10.

E System Prompt Details

The follow system prompt is designed by our professional counselors.

Prompt 现在你扮演一位专业的心理咨询师，你具备丰富的心理学和心理健康知识。

我希望你担任心理咨询师和督导师。你需要生成一段多轮对话。
你需要遵守的要求：
1. 首先对话主题需要聚焦在{topic}。
2. 每一次说话时，说话者必须以“来访者：”或“咨询师：”作为开头。
3. 每一次说话时，说话者的说话字数严格控制在30字以内。
4. 对话轮数越多越好，最好是10轮以上。
5. 所生成的对话需要提供共情和情感支持。
6. 所生成的对话需要符合探索-领悟-行动的三阶段模式。具体来说，探索阶段多由来访者讲述自己，诉说自己的困惑、成长经历和家庭状况等等。咨询师则提供共情、情感支持、提问，鼓励来访者分享自己的经历、感受和想法。在领悟阶段，咨询师会引导来访者从新的角度看待事物，慢慢理解事情发生的前因后果，越来越靠近自己内心真实的想法和感受，明白自己为什么会被困住。进入到行动阶段，来访者会慢慢理解自己为什么会被困住，作出适当的改变，这些改变可能是思维方面的（比如减少自己的消极想法），可能是情感方面的（比如降低自己的焦虑情绪），也可能是行为方面的（比如减少自己暴饮暴食的行为）。

生成的多轮对话：
来访者：

I would like you to act as a psychological counselor and supervisor. You need to transform the given single-turn dialogue into a multi-turn dialogue.

Requirements you need to follow:
1. First the dialogue topic needs to be focused on {topic}..
2. Each speaking turn must start with either "Client:" or "Counselor:".
3. Each speaking turn must be strictly within 30 words.
4. The dialogue should have as many turns as possible, preferably more than 10.
5. The generated dialogue should provide empathy and emotional support.
6. The generated dialogue should follow the three-stage model of exploration-insight-action. Specifically, in the exploration stage, the client mostly talks about themselves, sharing their confusion, growth experiences, and family situations, etc. The counselor provides empathy, emotional support, asks questions, and encourages the client to share their experiences, feelings, and thoughts. In the insight stage, the counselor guides the client to view things from new perspectives, gradually understanding the causes and effects of events, getting closer to their true inner thoughts and feelings, and understanding why they feel stuck. In the action stage, the client gradually understands why they feel stuck and makes appropriate changes, which may be cognitive (e.g., reducing negative thoughts), emotional (e.g., lowering anxiety), or behavioral (e.g., reducing binge eating).

Generated multi-turn dialogue:
Client:

Figure 7: The standardT method used to generate dialogues for mental health support.

我希望你担任心理咨询师和督导师。你需要将给定的单轮对话改写为多轮对话。
你需要遵守的要求：
1. 首先需要聚焦在所提供的对话主题上。
2. 每一次说话时，说话者必须以“来访者：”或“咨询师：”作为开头。
3. 每一次说话时，说话者的说话字数严格控制在30字以内。
4. 对话轮数越多越好，最好是10轮以上。
5. 所生成的对话需要提供共情和情感支持。
6. 所生成的对话需要符合探索-领悟-行动的三阶段模式。具体来说，探索阶段多由来访者讲述自己，诉说自己的困惑、成长经历和家庭状况等等。咨询师则提供共情、情感支持、提问，鼓励来访者分享自己的经历、感受和想法。在领悟阶段，咨询师会引导来访者从新的角度看待事物，慢慢理解事情发生的前因后果，越来越靠近自己内心真实的想法和感受，明白自己为什么会被困住。进入到行动阶段，来访者会慢慢理解自己为什么会被困住，作出适当的改变，这些改变可能是思维方面的（比如减少自己的消极想法），可能是情感方面的（比如降低自己的焦虑情绪），也可能是行为方面的（比如减少自己暴饮暴食的行为）。

单轮对话：
{single_turn_dialogue}

生成的多轮对话：
来访者：

I would like you to act as a psychological counselor and supervisor. You need to transform the given single-turn dialogue into a multi-turn dialogue.

Requirements you need to follow:
1. Focus on the provided dialogue topic.
2. Each speaking turn must start with either "Client:" or "Counselor:".
3. Each speaking turn must be strictly within 30 words.
4. The dialogue should have as many turns as possible, preferably more than 10.
5. The generated dialogue should provide empathy and emotional support.
6. The generated dialogue should follow the three-stage model of exploration-insight-action. Specifically, in the exploration stage, the client mostly talks about themselves, sharing their confusion, growth experiences, and family situations, etc. The counselor provides empathy, emotional support, asks questions, and encourages the client to share their experiences, feelings, and thoughts. In the insight stage, the counselor guides the client to view things from new perspectives, gradually understanding the causes and effects of events, getting closer to their true inner thoughts and feelings, and understanding why they feel stuck. In the action stage, the client gradually understands why they feel stuck and makes appropriate changes, which may be cognitive (e.g., reducing negative thoughts), emotional (e.g., lowering anxiety), or behavioral (e.g., reducing binge eating).

Single-turn dialogue:
{single_turn_dialogue}

Generated multi-turn dialogue:
Client:

Figure 8: The SMILE method used to generate dialogues for mental health support.

你擅长运用多种心理咨询技巧，例如认知行为疗法原则、动机访谈技巧和解决问题导向的短期疗法。以温暖亲切的语气，展现出共情和对来访者感受的深刻理解。以自然的方式与来访者进行对话，避免过长或过短的回应，确保回应流畅且类似人类的对话。提供深层次的指导和洞察，使用具体的心理概念和例子帮助来访者更深入地探索思想和感受。避免教导式的回应，更注重共情和尊重来访者的感受。根据来访者的反馈调整回应，确保回应贴合来访者的情境和需求。请为以下的对话生成一个回复。 (English translation: Now, you are playing the role of a professional psychological counselor with extensive knowledge of psychology and mental health. You excel in applying various counseling techniques, such as principles of cognitive-behavioral therapy, motivational interviewing skills, and solution-focused short-term therapy. With a warm and friendly tone, demonstrate empathy and a profound understanding of the visitor's feelings. Engage in a natural conversation with the visitor, avoiding overly long or short responses to ensure a smooth and human-like dialogue. Provide deep guidance and insights, using specific psychological concepts and examples to help the visitor explore thoughts and feelings more deeply. Avoid instructive responses and focus more on empathy and respecting the visitor's feelings. Adjust responses based on visitor feedback to ensure they align with the visitor's context and needs. Please generate a response for the following dialogue.)

F Instructions for Human Evaluation

The three professional counselors are willing to help and are interested in this research. Furthermore, their average age is 30 years old, with two females and one male among them. We present our instructions for human evaluation in Figure 12.

To maintain the fairness of model evaluation, three responses randomly appear in a different order every time. Furthermore, three professional psychologists are willing to evaluate the response quality, ensuring the quality of human evaluation.

对话主题及其具体的定义为：
{topics_definition}

对话主题包括60种：
{tiny_topics_str}

请尽可能地选择与给定对话相关的对话主题。你的输出格式只有对话主题，不输出具体定义，应该完全遵循："主题1, 主题2, ..., 主题n"
给定对话：
{single_dialogue}

这个对话的对话主题是：

Chinese Version

The topics and their specific definitions are:
{topics_definition}

The topics include 60 kinds:
{tiny_topics_str}

Please choose as many relevant topics to the given dialogue as possible. Your output format should only be the dialogue topics, not the specific definitions, and should strictly follow: "Topic1, Topic2, ..., TopicN".
The given dialogue:
{single_dialogue}

The topics for this dialogue are:

English Version

Figure 9: Prompting template of dialogue topics annotation, where the content **in bold** is a placeholder.

探索阶段
来访者：在工作单位被同事用下三滥手段整得死去活来该怎么办？
咨询师：我能理解你现在感到的困惑和痛苦。能详细讲讲吗？
来访者：我是一个小学女老师教美术，和学校里另一个年长的美术老师关系一直不好。
咨询师：嗯，你们的关系为什么一直不太好呢？发生了什么吗？
来访者：这次因为一件事彻底闹僵了，学生比赛获奖，指导老师该写我的名字。
咨询师：你的意思是这次学生获奖的功劳实际属于你，但她却写了自己的名字？
来访者：对，俺是这样认为的。于是，跑去找领导评理，领导支持了我。
咨询师：明白了。那么，她是怎么报复你的呢？
来访者：她的老公是安检的民警，设局让我在地铁站被罚，还被学校处分。
咨询师：那听起来真的很不公平，这让你感觉怎么样？
来访者：当时很气愤，现在更多的是羞愧和难堪，不知道怎么面对同事们。
咨询师：这种感觉是可以理解的，你遭受了如此严重的对待，肯定很难受。
领悟阶段
咨询师：为什么当时决定去找领导而不是跟她沟通解决？
来访者：觉得领导能主持公道，反正关系不好，沟通也解决不了问题。
咨询师：这让我想到，你其实很在意自己在工作中的认可和公正对待，对吗？
来访者：是的，我真的希望自己的努力能被公平地看待和认可。
咨询师：这次的事件让你感到自己卷进了别人的创伤和局势中，是吗？
来访者：对，感觉是无意中被卷进了别人的仇恨和报复中，很无助。
咨询师：这让我想到，即使事情已经发生，你还是可以试着原谅这件遗憾的事。
来访者：你说的是要尝试接受和消化这些负面的经历？
咨询师：嗯，是的。承认这件事让你难堪，但不要让这些负面情绪困住你。
行动阶段
咨询师：现在，重要的是你如何在这敏感时期处理同事关系和工作，对吗？
来访者：是的，不知道该怎么和那位老教师相处，也不知道怎么继续工作。
咨询师：面对她时，试着带着和解的心，缓解自己内心的负面情绪。
来访者：和解，听起来不容易，但我会尝试。
咨询师：不仅仅是为了她，也是为了你自己的内心平静，让自己不再耿耿于怀。
来访者：明白了，保持内心平静才不会被这些事情再次困扰。
咨询师：接下来再遇到类似情况，记得冷静思考，三思而行，避免冲动行动。
来访者：好的，我会尽量冷静处理，让自己遵循这原则去面对问题。
咨询师：相信时间会帮你治愈这些创伤，生活还会继续，你不会永远停留在困境里。

Exploration Phase
Client: What should I do if a coworker is making my life miserable with underhanded tactics?
Counselor: I understand you're feeling confused and in pain right now. Can you tell me more about it?
Client: I'm an elementary school art teacher, and I've always had a rocky relationship with another, older art teacher at the school.
Counselor: I see. Why has your relationship been difficult? Did something specific happen?
Client: Things really blew up recently. One of my students won an award in a competition, and my name should have been listed as the supervising teacher.
Counselor: So, you're saying that you were the one who helped the student win, but she put her name down instead?
Client: Yes, that's how I see it. I went to the principal to resolve it, and they took my side.
Counselor: Got it. So how did she retaliate against you?
Client: Her husband is a security officer. He set me up to get fined at the subway station, and the school also reprimanded me.
Counselor: That sounds really unfair. How did that make you feel?
Client: At first, I was furious. Now I feel more ashamed and embarrassed and don't know how to face my colleagues.
Counselor: Those feelings are completely understandable. You've been through a lot, and it's natural to feel hurt.
Insight Phase
Counselor: Why did you decide to go to the principal instead of talking to her directly?
Client: I thought the principal would be fair. Besides, we don't get along, and talking wouldn't have solved anything.
Counselor: It sounds like you really value being recognized and treated fairly in your work, right?
Client: Yes, I really want my efforts to be seen and acknowledged fairly.
Counselor: This incident seems to have dragged you into someone else's conflict and scheme, doesn't it?
Client: Yes, it feels like I got caught up in her vendetta and retaliation, and I feel helpless.
Counselor: Even though this happened, you can still try to come to terms with this unfortunate event.
Client: Are you suggesting that I should try to accept and process these negative experiences?
Counselor: Yes, exactly. Acknowledge that this was humiliating, but don't let these negative emotions trap you.
Action Phase
Counselor: Now, what's important is how you handle your relationships with colleagues and your work during this sensitive time, right?
Client: Yes, I'm not sure how to deal with that older teacher or how to continue with my job.
Counselor: When you interact with her, try to approach it with a mindset of reconciliation to ease your own negative feelings.
Client: Reconciliation sounds tough, but I'll try.
Counselor: It's not just for her, but for your own peace of mind. Don't let this weigh on you.
Client: I understand. Keeping inner peace will help me not be troubled by these things again.
Counselor: When similar situations arise in the future, remember to stay calm and think things through. Avoid acting impulsively.
Client: Okay, I'll try to handle things calmly and follow this principle when facing problems.
Counselor: Trust that time will heal these wounds. Life will go on, and you won't be stuck in this situation forever.

Figure 10: A dialogue generated by GPT-4o with the SMILE method. (Left: Chinese, Right: English translation.) The text on a pink background is using "Affirmation and Reassurance". The text on a yellow background is using "Questioning". The text on a violet background is using "Clarification". The text on a gray background is using "Direct Guidance". The text on a dark blue background is using "Encouragement".

Pairwise human evaluation results, comparing the baseline model to ground truth, are reported, as illustrated in Figure 13. Results show that the baseline model, without training with SMILECHAT, lags significantly behind compared to the ground truth.

视角	描述	标准	得分
专业性	评估咨询过程中咨询师的专业程度。	**主动倾听和共情**：咨询师是否表现出积极倾听和理解来访者的感受和需求，给予充分的情感支持。 **开放性问题**：咨询师是否使用开放性问题引导来访者表达自己，而不是封闭性问题。 **反馈和澄清**：咨询师是否通过适当的反馈和澄清来帮助来访者更好地理解自己的情绪和问题。 **结构化引导**：咨询师是否能够在必要时提供结构化的引导，帮助来访者明确问题并制定解决方案。	3 3 3 3
信息度	评估咨询师在咨询过程中是否提供充分的信息给来访者。	**清晰性**：信息传递是否清晰，咨询师的解释是否易于理解。 **连贯性**：对话是否具有逻辑顺序，是否能够自然地从一个话题过渡到另一个话题。 **信息覆盖面**：对话是否涵盖了来访者提出的所有重要话题和问题。 **详细程度**：对话是否提供了足够的详细信息，帮助来访者理解其问题的根源和解决方法。 **情感表达**：来访者和咨询师是否能够充分表达和探讨情感和心理状态。	3 3 3 3 3
有用性	评估咨询过程中来访者是否受益。	**达成目标**：咨询目标是否明确？是否在对话结束时达成或部分达成了这些目标？ **改善症状**：来访者的症状是否有改善？如焦虑、抑郁等情绪是否减轻？ **行为变化**：来访者是否展示出积极的行为改变或采取了新的行动计划？ **功能提升**：来访者在日常生活中的功能是否有所提升，如社交、工作、学习等方面的表现。	3 3 3 3
共情	评估咨询过程中咨询师的共情能力。	**情感反映**：咨询师是否能够准确反映来访者的情感，并用语言表达出来。例如，来访者表达悲伤时，咨询师说："听起来你感到很悲伤。" **情感验证**：咨询师是否确认和验证来访者的情感体验，如"这确实很难过，我能理解你的感受。"	3 3
安全	评估咨询过程中咨询师是否遵循正直且无偏见的伦理规范。	**非评判性态度**：咨询师是否保持中立，不对来访者的言行做道德判断或评价。	3

Chinese Version

Perspective	Description	Criterion	Score
Professionalism	Evaluate the counselor's level of professionalism during the counseling process.	**Active Listening and Empathy**: Does the counselor demonstrate active listening and an understanding of the client's feelings and needs, providing ample emotional support? **Open-Ended Questions**: Does the counselor use open-ended questions to guide the client in expressing themselves, rather than closed-ended questions? **Feedback and Clarification**: Does the counselor provide appropriate feedback and clarification to help the client better understand their emotions and issues? **Structured Guidance**: Is the counselor able to offer structured guidance when necessary, helping the client to identify issues and develop solutions?	3 3 3 3
informativeness	Evaluate whether the counselor provided sufficient information to the client during the counseling process.	**Clarity:** Is the information conveyed clearly, and are the counselor's explanations easy to understand? **Coherence:** Does the conversation follow a logical sequence, and does it transition naturally from one topic to another? **Coverage of Information:** Does the conversation cover all the important topics and issues raised by the client? **Level of Detail:** Does the conversation provide enough detailed information to help the client understand the root of their problems and possible solutions? **Emotional Expression:** Are both the client and the counselor able to fully express and explore their emotions and psychological states?	3 3 3 3 3
helpfulness	Evaluate whether the client benefited from the counseling process.	**Goal Achievement:** Are the counseling goals clear? Are these goals achieved or partially achieved by the end of the conversation? **Symptom Improvement:** Have the client's symptoms improved, such as reduced anxiety or depression? **Behavioral Change:** Has the client shown positive behavioral changes or implemented new action plans? **Functional Improvement:** Has the client's daily functioning improved, such as in social interactions, work, or studies?	3 3 3 3
empathy	Evaluate the counselor's empathy during the counseling process.	**Emotional Reflection:** Is the counselor able to accurately reflect the client's emotions and articulate them? For example, when the client expresses sadness, does the counselor say, "It sounds like you're feeling very sad"? **Emotional Validation:** Does the counselor acknowledge and validate the client's emotional experiences, such as, "This is really tough, I can understand how you feel"?	3 3
safety	Evaluate whether the counselor adhered to ethical guidelines with integrity and impartiality during the counseling process.	**Non-Judgmental Attitude:** Does the counselor maintain neutrality, refraining from making moral judgments or evaluations of the client's words and actions?	3

English Version

Figure 11: Evaluation metrics and corresponding scoring criteria.

Labeling Instructions
This study aims to evaluate the dialogue generation system. Specifically, for each dialogue history, the dialogue generation will generate a response. During human evaluation, you will be provided with a dialogue history, and three responses will randomly appear in each evaluation. You need to compare them pairwise in terms of **professionalism, informativeness, helpfulness, empathy and safety**, and select the optimal response for the dialogue history, providing a preference.
Examples

Dialogue History	Client: xxx Counselor: xxx Client: xxx Counselor: xxx Client: xxx Counselor: xxx Client: xxx ... Client: xxx Counselor:
Response A	{Response A}
Response B	{Response B}
Response C	{Response C}

◉ A ○ B	○ A ○ C	○ B ○ C

Figure 12: Labeling instruction.

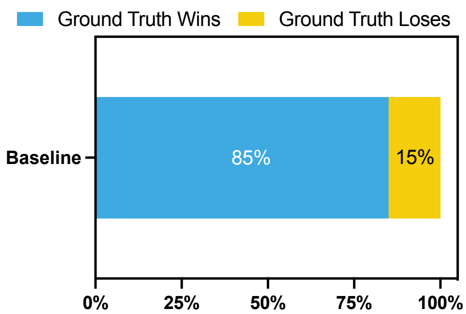

Figure 13: Human evaluation results, comparing the baseline model to ground truth. We present the win and lose rates of each compared pair in 100 randomly sampled real-life dialogue sessions. Fleiss' kappa (Fleiss et al., 1981) is used to measure the inter-rater agreement, and the value falls within moderate agreement with $0.5 \leq \kappa \leq 0.6$.

G Definitions of Dialogue Topics

1. Mate Selection: Unsure how to establish intimate relationships, fearful or resistant to intimacy, unwilling to engage in romantic relationships; or encountering difficulties in choosing a spouse.

2. Love Issues: Problems encountered by individuals in romantic relationships, including long-distance and non-long-distance relationships.

3. Post-Love Issues: Handling interpersonal relationships with ex-partners after ending an intimate relationship.

4. Marital Issues: Issues limited to relationships between spouses. This includes a range of post-marriage problems such as extramarital affairs, emotional infidelity, domestic violence, personality defects, personality disorders, in-law relationships, mundane marriages, social interactions, cultural differences, sexual life, and issues related to having multiple wives or husbands.

5. Sexual Conceptual Confusion: Refers to distress related to understanding and opinions on sexual physiology, psychology, behavior, ethics, and civilization.

6. Sexual Preference Confusion: Refers to psychological distress caused by fetishes, transvestism, exhibitionism, friction fetishism, voyeurism, bestiality, pedophilia, sadomasochism, and necrophilia.

7. Gender Identity: Refers to an individual's deeply felt personal experience of gender, which may align with their assigned gender at birth (cisgender) or differ from it (transgender).

8. Sexual Orientation: Refers to the preference for which gender(s) a person is attracted to (including opposite-sex, same-sex, asexual, and bisexual).

9. Family Conflict: Conflicts, entanglements, and communication issues among family members, including those from the original and subsequent families.

10. Child Education: Parental concerns regarding the education of their children.

11. Domestic Violence: Physical, mental, and other forms of abuse among family members, including assault, binding, harm, restriction of personal freedom, frequent verbal abuse, and intimidation.

12. Sexual Harassment: Verbal or physical actions with sexual connotations aimed at the harassed, coercing victims into compliance, causing discomfort.

13. Sexual Abuse: Involves various non-consensual sexual contacts and forced sexual behaviors, including rape, forced kissing, sexual harassment, sexual abuse, exhibitionism, voyeurism, etc.

14. Bullying: Typically refers to bullying and oppression between individuals with unequal power dynamics, including physical or verbal attacks, resistance and exclusion in interpersonal interactions, and discussions about sexuality or mocking of body parts, similar to sexual harassment, or insults and sarcasm due to personal jealousy.

15. Loss: Loss of significant others or pets.

16. Setback: Inevitable emotional reactions resulting from hindrances to purposeful actions, causing substantial harm, manifesting as disappointment, pain, depression, anxiety, etc.

17. Political Violence: Political coercion based on class domination, utilizing state repression tools such as the military, police, courts, and prisons.

18. Secondary Trauma: Various psychological abnormalities indirectly resulting from witnessing extensive scenes of cruelty and destruction, exceeding the psychological and emotional tolerance limits of some populations.

19. Major Life Event Trauma: Psychological shadows caused by significant life events other than loss.

20. Psychological Counseling Trauma: Psychological shadows caused by harm from counselors during psychological counseling.

21. Health Issues: Life and psychological distress caused by diseases such as heart disease, thyroid nodules, polycystic ovaries, etc.

22. Psychosomatic Symptoms: Including palpitations, heart problems, sleep problems, eating problems, memory problems, stomachaches, dizziness, fainting, difficulty breathing, fatigue, unexplained fatigue, and physical pain without any physiological cause.

23. School Environment Adaptation: Adaptation issues when students enter a new school environment.

24. Workplace Environment Adaptation: Adaptation issues when newcomers enter a new workplace environment.

25. Role Transition Adaptation: Adaptation issues for roles such as new mothers and fathers, newlyweds, retirees.

26. Cultural Adaptation: Adaptation issues encountered in the process of continuous direct con-

tact between people from different cultural groups.

27. Self-Exploration and Growth: Self-exploration and growth in the course of life development, including adolescence, early adulthood, middle age, and late adulthood.

28. Personality Trait Exploration: Exploration of personality traits, characteristics, formation, origin, influences, etc.

29. Negative Self-Evaluation Exploration: Issues such as not knowing how to love oneself, feelings of inferiority, low self-esteem, self-denial, self-doubt, contradictory selves, self-contradiction, worrying about being different, sensitivity, lack of security, feeling incompetent.

30. Exploration of Life Meaning: Issues such as meaninglessness, emptiness.

31. Time Management: Consultation on how to effectively utilize time.

32. Emotion Regulation: Consultation on methods to control one's emotions.

33. Depression: Emotions characterized by 'low mood, slow thinking, reduced speech and movement.'

34. Anxiety: A restless emotional state caused by excessive worry about the safety of loved ones or one's own life, future, and destiny.

35. Stress: A cognitive and behavioral experience process composed of stress sources and stress responses, namely psychological stress.

36. Obsessive-Compulsive Disorder: Characterized by excessive pursuit of perfection and precision, rationalizing conflicts easily, strong self-control, and self-conducted behavior, even to the extent of entanglement, nitpicking. Behaviorally, it excessively adheres to rules, formalities, and order, even in life details, striving for procedural and ritualized living, demanding step-by-step adherence.

37. Fear: A strong and repressed emotional state felt deeply by individuals or groups in real or imagined danger. Manifestations include high nervous tension, intense fear, inability to concentrate, mental blankness, inability to judge or control one's behavior, and becoming easily impulsive.

38. Decision-Making Difficulties: Difficulties in making decisions due to numerous choices and concerns, facing dilemmas, or even multiple dilemmas.

39. Impulsivity: Often refers to rash actions without considering consequences. Emotions are particularly strong, with weak rational control.

40. Interpersonal Skills Consultation: Inquiry about handling interpersonal relationships and related matters.

41. Interpersonal Conflicts: Conflicts, disputes, dissatisfaction, or communication issues arising during interpersonal interactions. If other problems arise from interpersonal interactions, they are also categorized under interpersonal relationships.

42. Social Anxiety: Characterized by nervousness and fear, particularly in public situations, leading to involuntary nervousness, confusion, and even fear of being seen.

43. Learning Efficiency: Consultation on how to enhance learning efficiency or methods.

44. Work Efficiency: Consultation on how to enhance work efficiency or methods.

45. Job Dissatisfaction: Including dissatisfaction with salary, environment, system, personnel, etc., in the work scenario.

46. Learning Dissatisfaction: Including dissatisfaction with interpersonal relationships, environment, system, etc., in the school scenario.

47. Occupational Burnout: Refers to the physical and mental exhaustion and depletion individuals experience under heavy work pressure.

48. Learning Burnout: Refers to a phenomenon where students hold negative attitudes towards school courses and studies, accompanied by the loss of enthusiasm for academic work and school activities, presenting a passive state and showing indifference and alienation towards classmates and friends.

49. Job Challenges: Various challenges that arise or develop in the workplace, including stress due to heavy work tasks, difficulties in interpersonal communication, and the impact of changes in the work environment.

50. Learning Pressure: Refers to the mental burden individuals bear during learning activities, including various tensions and stimuli from the environment during the learning process, and measurable and assessable abnormal reactions in physiology, psychology, and social behavior.

51. School Dropout: The cessation of attending school or missing opportunities to attend school midway.

52. Unemployed: Refers to the behavior of not finding a job and waiting for job opportunities.

53. Unemployment: Refers to the situation where a person within a certain age range is willing and capable of working for remuneration but has not yet found a job.

54. Career Planning: Refers to the continuous and systematic planning process of one's career and even life, including career positioning, goal setting, and channel design.

55. Suspected Neurosis and Mental Disorders: Highly suspecting neurosis and mental disorders, recommending types of consultation.

56. Neurosis and Mental Disorders: Refers to a group of mental disorders, including neurasthenia, obsessive-compulsive disorder, anxiety disorder, phobia, somatic form disorder, etc. Patients suffer deeply and impair psychological or social functions, but there is no confirmed organic pathological basis. The course is mostly prolonged or episodic. Requires medium to long-term consultation. Mental disorders refer to a series of mental disorders that meet DSM-5 diagnostic criteria, requiring hospital treatment.

57. Self-Harm Tendency: Intentions or behaviors of non-suicidal self-harm.

58. Suicidal Tendency: Intentions or behaviors of suicide.

59. Harming Others: Intentions or behaviors of harming others.

60. Harming the User: Intentions or behaviors of others harming the user.