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

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 and 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 that prompts ChatGPT to rewrite public single-turn dialogues into multi-turn ones. Our work begins with the analysis of language transformation, validating the feasibility of the proposed method when compared with other baseline methods. We then conduct a study on dialogue diversity, including lexical features, semantic features, and dialogue topics, demonstrating the effectiveness of our proposed method. Furthermore, we implement an expert evaluation and the results demonstrate that the dialogues generated with our proposed method are of higher quality than those generated with other baseline methods. Thus, we employ our method to generate a large-scale, diverse, and high-quality dialogue dataset named SMILECHAT, comprising 55,165 dialogues in total with an average of 10.4 turns per dialogue. Finally, we utilize the collected corpus to develop a mental health chatbot, MECHAT. To better assess the overall quality of SMILECHAT, we collect a real-life chat dataset comprising 82 counseling dialogues for model evaluation. Both automatic and human evaluations demonstrate that our trained dialogue system exhibits significant improvements, showcasing that SMILECHAT is high-quality and practical.

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.

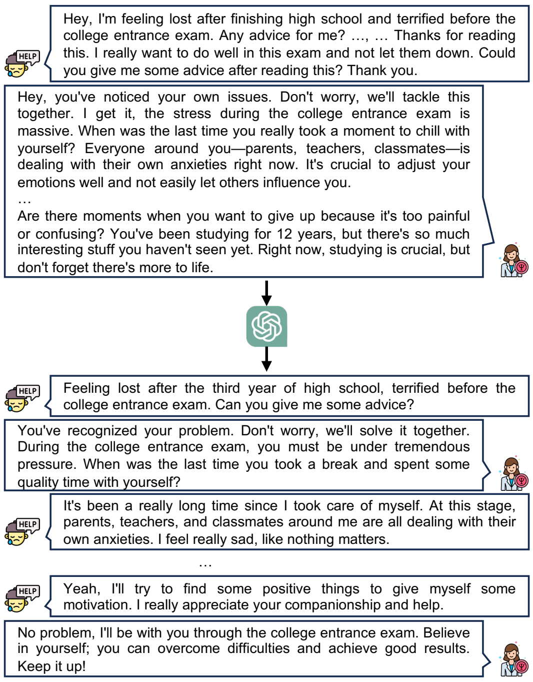

Figure 1: Example of instructing ChatGPT to rewrite a single-turn dialogue into a multi-turn one.

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 a dialogue system provides 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 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. Specifically, we instruct ChatGPT to transform publicly available question-answer pairs (public QA), which can also be considered as single-turn dialogues from the real world, into multi-turn conversations, as illustrated in Figure 1. 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, dialogue generation, and dialogue retention, and mainly elaborate on the SMILE method and baseline methods.
- We demonstrate the feasibility of the SMILE method through language transformation (§4).
- We demonstrate the effectiveness of the SMILE method by utilizing three diversity indicators (§5): lexical features, semantic features, and dialogue topics.
- We implement an expert evaluation (§6) to demonstrate that the dialogues generated with our proposed method are of higher quality than those generated with baseline methods.
- Following the validation of feasibility and effectiveness, and expert evaluation, we leverage the SMILE method to generate a large-scale, and diverse multi-turn chat dataset, SMILECHAT (§7), for mental health support.
- Finally, we propose training a dialogue system to explore the quality of conversation (§8) and collecting a set of 82 real-life counseling dialogues to construct an authentic 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 provides a novel perspective for alleviating the scarcity of multi-turn conversations in mental health.
- Through the analysis of language transformation and dialogue diversity, 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 82 counseling dialogues to build an authentic test set, PSYTEST, which contains 4045 utterances spoken by supporters. Both automatic and human evaluations demonstrate that SMILECHAT improves the performance of the dialogue system in mental health.
- We release SMILECHAT, which comprises 55,165 Chinese multi-turn dialogues with an

average of 10.4 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., 2023).

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.

3 Method

PsyQA[2], a high-quality Chinese dialogue dataset related to mental health support in the form of one question mapping to multiple answers, has gone through a data anonymization process. Our dataset creation pipeline based on PsyQA comprises three main stages: (1) data preprocessing, (2) dialogue generation, and (3) dialogue retention.

3.1 Data Preprocessing

Considering the distinction between QAs in PsyQA and multi-turn dialogues and the context window limitation of 4096 tokens in ChatGPT, we recommend performing data preprocessing for PsyQA. This process involves wording cleaning and length truncation.

Wording Cleaning This work aims 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 help-seeker and a supporter, 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.

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

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

This process includes both automatic and manual cleaning procedures. For detailed content, see Appendix A.

Length Truncation After conducting a statistical analysis of the PsyQA dataset released by Sun et al. (2021), we find that 757 QAs had a total length exceeding 1800 characters. Additionally, we identify 9 QAs in which the total discourse length exceeds 4000 characters. Furthermore, the model `gpt-3.5-turbo`[3] has a maximum context length of 4096 tokens. To ensure reliable and smooth rewrites, we limit the length of the QA pairs, maximizing the number of rewritten dialogue turns. Specifically, we truncate the length of the QA pairs at 1800 characters and truncate any excess text. This control ensures that the generated text is limited to approximately 2,000 tokens. It is worth noting that the PsyQA data used in this study undergo a data preprocessing process.

3.2 Dialogue Generation

3.2.1 Task Formulation

Considering the remarkable capabilities of ChatGPT and its cost-effective, plug-and-play characteristics, we consider ChatGPT to be a suitable choice for this task compared to other alternative models. Our study is designed to prompt ChatGPT to rewrite a single-turn dialogue into a multi-turn one. Generally, text generation can be formulated as follows:

$$P_{\mathcal{M}_\theta}(Y|x) = \prod_{t=1}^{L} P_{\mathcal{M}_\theta}(y_t|Y_{<t}, x) \tag{1}$$

where $\mathcal{M}$ denotes a large language model with parameters θ. Y represents the output of the model generation, while L signifies the variable length of the output. The symbol y_t represents the t-th token generated by the model, and x denotes the conditional model input.

3.2.2 Prompt Design

To establish some conventions regarding specific terms for prompt design, we define the "help-seeker" and "supporter" as "求助者" and "支持者" in Chinese, respectively. A single-turn dialogue is defined as "求助者：u^H 支持者：u^S", where u^H and u^S represent the utterances of the help-seeker and supporter, respectively, and H and S refer to the help-seeker and supporter.

[3]The model we use is `gpt-3.5-turbo-0613`.

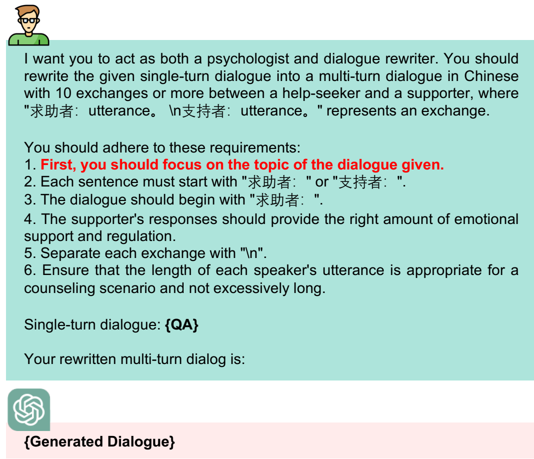

Figure 2: The SMILE method used to generate dialogues for mental health support.

In this section, we mainly focus on detailing our methodology. Furthermore, we are building two baseline methods for comparison.

Standard Prompt As its name suggests, the standard prompt does not contain any single-turn dialogues and instead uses only the simplest prompt to generate multi-turn dialogues. The specifics of the standard prompt are illustrated in Figure 8 in Appendix C. The conditional model input in Equation 1 is $x := (\mathcal{I})$, where $\mathcal{I}$ represents the standard prompt. We simplify the method name as `standard` and consider this method as our baseline.

Standard Prompt with a Dialogue Topic Intuitively, feeding a single, fixed prompt into a large language model often results in the generation of low diversity. To address this issue, 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 56 distinct types, each accompanied by its corresponding explanation. For more details, please refer to Appendix K. The conditional model input in Equation 1 is $x := (\mathcal{I}, \mathcal{T})$, where $\mathcal{I}$ represents the dialogue topic chosen in uniform sampling. We simplify the method name to `stand.`$_{\text{DT}}$ and adopt this method as our baseline, as illustrated in Figure 8 in Appendix C.

SMILE Method Our proposed method, referred to as the SMILE method, instructs the ChatGPT to rewrite single-turn dialogues into multi-turn ones. Figure 2 depicts the specific prompt template. The conditional model input in Equation 1 is $x := (\mathcal{I}, \mathcal{T}, \mathcal{D})$, where $\mathcal{T}$ and $\mathcal{D}$ represent the dialogue topics hidden in the QA and single-turn dialogue, respectively.

3.3 Dialogue Retention

ChatGPT may exhibit potential instability during the dialogue generation process, but this occurs very infrequently. Therefore, we employ an automatic filtering mechanism to exclude dialogues that fail to meet our specified requirements, which encompass dialogue format and dialogue turns as detailed in Appendix B.

4 Language Transformation

4.1 Experimental Setup

QA Sampling To ensure a fair comparison and prevent duplicate instances of the same question with different answers, we first randomly select 500 non-duplicate questions from the first 5,000 QAs in PsyQA. We then randomly choose one answer to serve as the corresponding question's response. The data samples obtained are employed as seed dialogues, which are subsequently restructured into multi-turn conversations via ChatGPT. We name this 500 sampled data as **PsyQA*** to distinguish it from the whole dataset.

Hyperparameters On the whole, we present three prompt methods in this paper. For each prompt method, we instruct ChatGPT to generate 500 dialogues to study language transformation and dialogue diversity, thereby validating the feasibility and effectiveness of our proposed SMILE method, respectively. To enhance the diversity of the generated dialogues, we set ChatGPT's hyperparameters during text generation to the officially recommended default values, with τ (temperature) = 1.0 and p (top_p) = 1.0.

Dialogue Filtering When a dialogue falls short of the criteria in §3.3, we prompt ChatGPT to generate the dialogue anew until it aligns with our specified requirements only in the initial 500 dialogues. In 500 conversations, instability in generation is observed only in a small number of cases, and the issue is resolved by the third attempt, addressing situations that do not meet the conversational requirements.

Text Representation Text representation is used for analyzing language transformation and semantic diversity. A dialogue between a help-seeker and a supporter is represented as $d = \{u_0^H, u_1^S, u_2^H, ..., u_t^X, ..., u_{n-1}^H, u_n^S\}$, where u_t^X represents the utterance of the t-th turn spoken by either the help-seeker or supporter. A string of a dialogue can be denoted as $d_s = [u_0^H; u_1^S; u_2^H; ...; u_{n-1}^H; u_n^S]$, where $[;]$ denotes the operation of textual concatenation.

To obtain the text embedding of a dialogue, we employ OpenAI's model *text-embedding-ada-002*[4], which accepts a maximum context length of 8191. Each dialogue is first preprocessed into a single string without any speaker tokens and is then mapped to a 1536-dimensional vector. For example, to compute the cosine similarity between two different dialogues, we can obtain

$$\cos(d_i, d_j) = \frac{E_i \cdot E_j}{\|E_i\| \, \|E_j\|} \qquad (2)$$

where E_i and E_j denote the text embeddings from two distinct dialogues.

4.2 Analysis

To assess the transformation feasibility of our SMILE method, we employ cosine similarity to calculate the transformation ratio. Specifically, a single-turn dialogue d is transformed into a multi-turn dialogue $\hat{d}$ using the SMILE method. A multi-turn dialogue generated by `standard` or `stand.`$_{\mathrm{DT}}$ is denoted as $\bar{d}$. We can calculate $\cos(d, \hat{d})$ for the first hypothesis and $\cos(d, \bar{d})$ for the second hypothesis. We utilize the text embeddings obtained in §4.1. The mechanism of language transformation is presented in Figure 9 in Appendix D.

Figure 3 presents the distribution of dialogue transformation among three methods. According to our analysis, we conclude that single-turn dialogues can be successfully rewritten into multi-turn dialogues.

5 Dialogue Diversity

To demonstrate the effectiveness of the SMILE method, we mainly focus on three diversity aspects: lexical features, semantic features, and dialogue topics. For brevity, we only present semantic features in the main body of our paper. For details of another two diversity aspects, see §F in Appendix.

5.1 Semantic Features

To measure the semantic diversity of a dialogue dataset, we propose computing the cosine similarity

[4] https://platform.openai.com/docs/guides/embeddings

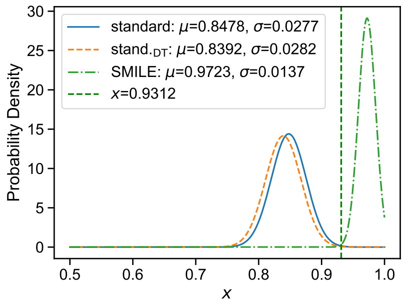

Figure 3: Distribution of dialogue transformation among three methods. The line $x = 0.9312$ represents the boundary of $\mu - 3\sigma$ in the SMILE method.

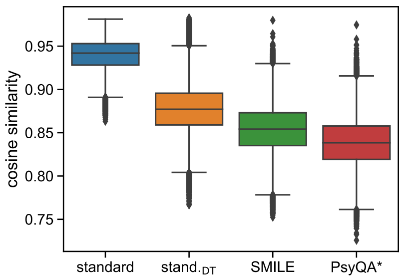

Figure 4: Pairwise dialogue cosine similarity among four settings: our proposed three methods and a reference point using sampled data from PsyQA.

Category	Total	Help-seeker	Supporter
# Dialogues	55165	-	-
# Utterances	1833856	693756	1140100
# Avg. turns per dialogue	10.4	-	-
Avg. utterances per dialogue	33.2	12.6	20.7
Avg. length per utterance	27.9	26.1	28.9

Table 1: Statistics of the dialogue dataset, SMILECHAT.

between every two different dialogues. We utilize the text embeddings obtained in §4.1.

To calculate the pairwise cosine similarity between every two distinct dialogues, this produces $\binom{500}{2}$ pairs of dialogues and their corresponding cosine values in each prompt method, including the reference point of sampled data from PsyQA.

We present the results in Figure 4, 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, aligning closely with the sampled dialogue of PsyQA. However, it's worth noting that the median of the SMILE method is larger than that of PsyQA*. The reason behind this may be the introduction of token distribution from ChatGPT.

6 Expert Evaluation

We elaborate on data quality analysis from two main professional perspectives: (1) the evaluation of dialogue quality by three professional counselors, and (2) alignment with the supporters' strategy flow patterns in real-life counseling conversations (For brevity, we include this section in Appendix G.).

Method	Q1	Q2	Q3
`standard`	1.95	2.12	1.21
`stand.`$_{DT}$	1.25	1.2	1.82
SMILE	**2.8**	**2.68**	**2.97**

Table 2: Evaluation of dialogue quality by three professional counselors.

To further validate that the dialogue quality generated using the SMILE method is superior to that produced by the two baseline methods, we propose to recruit three professional psychological counselors and randomly sample 20 dialogues for each method. Each set comprises three dialogues generated from different methods. Three counselors assign scores (3, 2, and 1, where 3 is the best, and 1 is the worst) based on three metrics: assessing (Q1) the overall quality of the conversation, (Q2) the authenticity of the user's stated issues, and (Q3) the professionalism of the counselor's responses.

We present the results of an expert evaluation of dialogue quality in 20 sets of dialogues in Table 2. Each value in the table represents the average score from three professional counselors, and results show that the dialogue quality generated with the SMILE method is of higher quality than that generated by the two baseline methods. The intraclass correlation coefficients (James et al., 1984; Koo and Li, 2016), including nine coefficients for the three metrics, are all greater than 0.8, indicating almost perfect agreement.

7 SMILECHAT Dataset

Through the analysis of language transformation, 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 ChatGPT in generating all multi-turn conversations based on PsyQA one round, leading to a **large-scale** dialogue dataset.

7.1 Data Statistics

To ensure data quality, we impose stricter requirements on dialogue turns, retaining only dialogues with at least 5 turns. Thus, we compile a collection of 55,165 conversations, SMILECHAT. Table 1 presents the statistics of the collected corpus.

7.2 Dialogue Exemplars

Multi-turn dialogue examples generated with the `standard` and `stand.`$_{\text{DT}}$ methods are illustrated in Figure 13 and 15. Further, a dialogue generated by the SMILE method is shown in Figure 17.

8 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.

8.1 Task Formulation

Our collected dataset can be represented as $D = \{d_1, d_2, ..., d_n\}$, where each d_j represents a single multi-turn dialogue.

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:

$$d_t = \{u_0^H, u_1^S, u_2^H, ..., u_{t-1}^H, u_t^S\} \sim D \quad (3)$$

We build a dialogue model that can predict the supporter's utterance u_t^S based on the dialogue history $h_t = \{u_0^H, u_1^S, u_2^H, ..., u_{t-1}^H\}$. Our objective is to maximize the likelihood probability as follows:

$$\underset{\theta}{\text{maximize}}\ \mathcal{F}(\theta; h_t, u_t^S) := \mathbb{E}_{d_t \sim D} \prod_{i=1}^{L} \mathbb{P}(u_t^S|\theta, h_t) \quad (4)$$

where L is the sequence length of u_t^S.

8.2 Experimental Setup

Baseline Model To validate the dialogue quality of our collected dataset, we conduct a fine-tuning experiment on ChatGLM2-6B (Zeng et al., 2023). Moreover, we also add two prompt engineering baselines (Prompt A and B, for more details, see Appendix I).

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 ChatGLM2-6B model for efficient fine-tuning.

Category	Total	Help-seeker	Supporter
# Dialogues	82	-	-
# Utterances	7672	3627	4045
# Avg. turns per dialogue	74.8	-	-
Avg. utterances per dialogue	93.6	44.2	49.3
Avg. length per utterance	24.6	29.7	20.1

Table 3: Statistics of test set, PSYTEST.

Hyperparameters We present the hyperparameters for constructing a dialogue model for mental health in Table 10. For additional details on the training and generation processes, please refer to Appendix H.

8.3 Evaluation

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 50 minutes each time. This duration is widely recognized as a standard time setting in psychological counseling. Therefore, we compile a collection of 82 high-quality real-life counseling dialogues with data desensitization for model evaluation. We name this test set PSYTEST, which contains 4045 utterances spoken by supporters, as shown in Table 3.

8.3.1 Automatic Evaluation

Metrics To conduct automatic evaluation, the evaluation metrics we use consist of Perplexity (PPL) (Jelinek et al., 1977), METEOR (Banerjee and Lavie, 2005), BLEU-1/2/3 (Papineni et al., 2002), Rouge-L (Lin, 2004), and Distinct-1/2/3 (D-1/2/3) (Li et al., 2016).

Results The results of the automatic evaluation, including 9 metrics, are presented in Table 4. Notably, the evaluated dialogues are based on real-world counseling data rather than generated dialogues, which excludes the influence stemming from ChatGPT. All automatic evaluation metrics we use indicate improved performance. Our results show that the model trained with SMILECHAT is effective and practical. Consequently, the automatic evaluation demonstrates that our collected dataset is of high quality.

8.3.2 Human Evaluation

Metrics We conduct a pairwise human evaluation to study the model performance trained with

Methods	PPL (⇓)	METEOR (⇑)	BLEU-1 (⇑)	BLEU-2 (⇑)	BLEU-3 (⇑)	Rouge-L (⇑)	D-1 (⇑)	D-2 (⇑)	D-3 (⇑)
Baseline	3.38	7.58†/14.46‡	5.77†/8.55‡	1.71†/3.75‡	0.95†/1.79‡	8.15†/11.79‡	75.00†/63.22‡	93.18†/86.69‡	96.51†/93.62‡
Baseline + Prompt A	3.43	6.69†/12.56‡	3.79†/5.62‡	1.05†/2.47‡	0.57†/1.16‡	6.21†/9.63‡	64.15†/51.55‡	89.00†/80.11‡	94.87†/90.02‡
Baseline + Prompt B	3.99	7.90†/14.45‡	5.18†/7.48‡	1.48†/3.32‡	0.78†/1.58‡	7.55†/11.18‡	71.79†/57.43‡	93.05†/84.58‡	96.89†/92.94‡
Fine-tuned	**1.50**	**13.31†/17.40‡**	**12.07†/14.60‡**	**5.08†/7.00‡**	**3.07†/3.70‡**	**14.81†/16.45‡**	**87.36†/80.55‡**	**98.43†/95.70‡**	**99.47†/98.38‡**

Table 4: Results of automatic evaluation in our compiled real-life test set. † denotes the tokenizer we use is `THUDM/chatglm2-6b`, while ‡ denotes the tokenizer we use is `hfl/chinese-roberta-wwm-ext-large`.

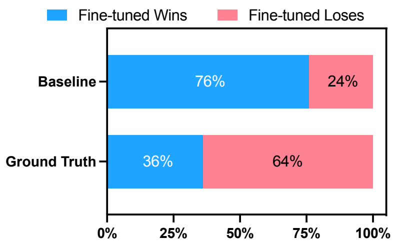

Figure 5: Pairwise human evaluation results, comparing the fine-tuned model to the baseline and ground truth, are reported. We present the win and lose rates of each compared pair in 100 randomly sampled real-life dialogue sessions. Fleiss' kappa (Fleiss et al., 1981) is used to measure the inter-rater agreement, and all values fall within moderate agreement with $0.5 \leq \kappa \leq 0.6$.

Help-seeker: I just went through a breakup, and it hurts so much.
Supporter:

✓ **Response 1 [Fine-tuned]:** Breakups are very painful experiences, and I can understand your feelings. *(Affirmation and Reassurance)* Can you tell me more about the situation surrounding your breakup? *(Inquiring Subjective Information)*
Response 2 [Baseline]: I'm really sorry to hear that. Breakups can be incredibly painful. *(Affirmation and Reassurance)* I understand that it can bring about feelings of disappointment, sadness, loneliness, and loss. *(Reflection of Feelings)* The end of a relationship may make us feel like we've invested a lot, only to get an ending in return, which can leave us feeling even more lost and helpless. *(Interpretation)*

If you're willing, we can discuss the impact of the end of this relationship on you and explore your current emotions. We can also talk about how to cope with these emotions and find some constructive solutions. *(Invite to Explore New Actions)*

Figure 6: Case study. Counseling strategies used in the two responses are presented in parentheses. Strategies in green are supporting, while those in red are challenging.

our proposed dialogue corpus. Initially, we randomly sample 100 multi-turn dialogue sessions, each comprising a minimum of 9 turns selected from PSYTEST. Each dialogue session concludes with the utterance spoken by the counselor (also referred to as the supporter). 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 such as naturalness, human-like qualities, correctness, and appropriateness. The evaluation is conducted based on the ethical principles of psychologists and the code of conduct (Association et al., 2016).

Results We employ majority voting to reach final decisions among three professional counselors. As depicted in Figure 5, 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 36% of all dialogue sessions.

Case Study We present a case study, as shown in Figure 6. First, during counseling conversations, the text should not be too long for the help-seeker to understand. Second, at the beginning of a conversation, a supporter generally will not challenge the help-seeker; instead, they will provide support. Third, response 1 is more human-like and natural. Therefore, response 1 is better than response 2 overall.

9 Conclusion

This paper introduces SMILE, a simple yet effective solution for addressing the scarcity of multi-turn conversations in mental health. Through language transformation and diversity analysis, we confirm the feasibility and effectiveness of our approach, demonstrating that our method is transferable. We also conduct expert evaluations, demonstrating that the SMILE method exhibits high superiority. Thus, the proposed method enables the automatic creation of a large-scale, diverse, and high-quality dialogue corpus, SMILECHAT, consisting of 55,165 dialogues with an average of 10.4 turns. Both automatic and human evaluations using the PSYTEST dataset, consisting of 82 real-life counseling dialogues, demonstrate that SMILECHAT significantly improves dialogue system performance in mental health. With the release of multi-turn dialogues (SMILECHAT), our dialogue model (MECHAT), and an authentic test set (PSYTEST), we contribute 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 ChatGPT. Consequently, the dataset unavoidably incorporates ChatGPT's model knowledge. Furthermore, we discuss how ChatGPT does not fully utilize the rich vocabulary and content of single-turn conversations, as reflected in the distinct-n metric.

Additionally, there is inevitably a gap in counseling practicality between models trained on this synthetic dataset and those trained on 50-minute counseling conversations, which will be our future work. While dialogues generated by the three methods still exhibit significant gaps when compared to the Xingling (Oracle) dataset with respect to the flow patterns of supporters' strategies, there exists a large space for us to improve in our future work.

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

Ethical Considerations

We present a detailed illustration for the section on ethical considerations, which includes **Data Usage Agreement**, **Usage Policy**, and **Suggestions for Real-life Deployment**. Furthermore, the section on suggestions for real-life deployment highlights four main aspects: Anonymity and Consent, Handling Sensitive Topics, Ensuring Client Confidentiality, and Preventing Misinformation.

Data Usage Agreement

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 purposes only**. 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.

Usage Policy

Our chatbot is trained with the SmileChat dataset end-to-end. As illustrated in our paper, SmileChat still exhibits a significant gap in the flow patterns of counseling strategies compared to real-life counseling dialogues. Access to SmileChat and Mechat **should be strictly limited to research purposes and should NOT be utilized for real-world deployment, commercial purposes, or any other use beyond academic research**. Researchers using SmileChat and MeChat should be aware of the dataset and model limitations and make efforts to acknowledge and/or mitigate them to the best of their ability.

Suggestions for Real-life Deployment

When deploying our chatbot on the server for research purposes, we recommend implementing safeguards to protect help-seekers and, to a large extent, mitigate potential physical risks.

Anonymity and Consent We strongly recommend that researchers allow users to interact with the chatbot anonymously, especially when discussing sensitive issues. Additionally, researchers should clearly communicate the level of anonymity and obtain informed consent from users before collecting any personal information.

Handling Sensitive Topics In real-life human-machine interactions, we strongly recommend constructing a classifier to identify user intent. When users discuss sensitive topics, if detected by the intent classifier, a "human-in-the-loop" system allows a human moderator or healthcare professional to intervene in complex or sensitive situations.

Furthermore, we recommend ensuring clear escalation paths for users who require immediate human assistance, such as in cases of suicide and self-injury.

Ensuring Client Confidentiality Researchers should clearly disclose to users that they are interacting with a chatbot, not a human, setting appropriate expectations for the system's capabilities and limitations.

Researchers should provide clear information on how data is handled, stored, and for what purposes.

Misinformation Prevention Researchers should implement fact-checking mechanisms to verify the accuracy of information provided by the chatbot.

Researchers should clearly communicate the limitations of the chatbot's knowledge and advise users to consult a human professional for critical decisions.

References

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

American Psychological Association et al. 2016. Ethical principles of psychologists and code of conduct. *American Psychological Association*.

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.

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.

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*.

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*.

Lawrence R James, Robert G Demaree, and Gerrit Wolf. 1984. Estimating within-group interrater reliability with and without response bias. *Journal of applied psychology*, 69(1):85.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

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.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

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*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023. 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.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations (ICLR)*.

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*.

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 7 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 7: 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.

B Requirements for Dialogue Retention

Here are two main requirements for dialogue filtering: data format and dialogue turns.

B.1 Data Format

We provide the requirements for data format as follows:

1. The generated conversations do not start with "求助者：" or "支持者：".

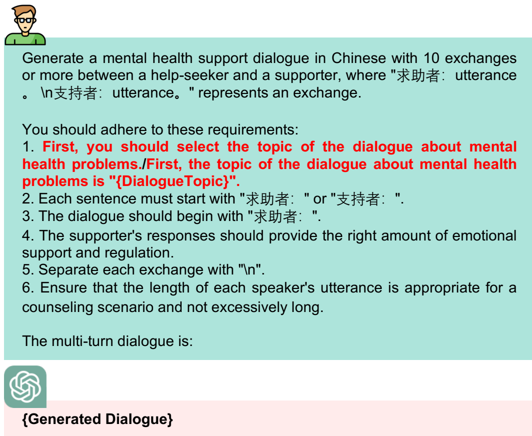

Figure 8: The standard prompt and the standard prompt with a dialogue topic. The two are separated by using the mark of "/" to differentiate between them. The key difference lies in whether a dialogue topic is introduced.

2. The generated dialogue does not contain any "\n", which is used for splitting the utterance from the help-seeker or supporter.
3. Each utterance in generated conversations does not start with "求助者：", "求助者:", "支持者：" or "支持者:".
4. The last utterance in generated conversations contains an English sentence.

B.2 Dialogue Turns

Conversations comprising fewer than 5 turns will be discarded.

C Method

For brevity, we present the standard prompt and the standard prompt with a dialogue topic in Figure 8.

D Mechanism of Language Transformation

We propose two hypotheses: (1) *When a single-turn dialogue is rewritten into a multi-turn dialogue, the similarity between the two is high (Attract).* (2) *When ChatGPT generates a multi-turn dialogue without introducing a single-turn dialogue, the similarity between the generated dialogue and a randomly selected single-turn dialogue is low (Repel).* For better understanding, we present the mechanism of language transformation in Figure 9.

E Dialogue Topics Annotation

In this paper, to label the dialogue topics of generated dialogues, the hyperparameters of ChatGPT

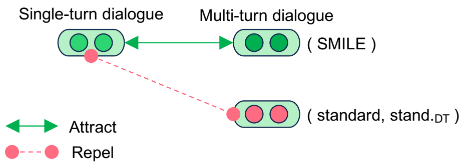

Figure 9: Mechanism for language transformation.

API during generation we used are set to the officially recommended default values, where temperature $\tau = 1.0$ and $p = 1.0$. Figure 10 shows the pipeline of dialogue topics annotation.

The dialogue topics include **{the list of dialogue topics}**. Please select dialogue topics that are as relevant as possible to the given dialogue. Your output format should be exactly as follows: "Dialogue topics are: {topic 1}, {topic 2}, {topic 3}, ..., {topic n}. "

Given dialogue: **{dialogue}**

Dialogue topics are:

Figure 10: Pipeline of Dialogue Topics Annotation, where the content **in bold** is a placeholder.

F Dialogue Diversity

F.1 Lexical Features

For lexical analysis, we utilize the ChatGLM2-6B tokenizer[5] 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, with **PsyQA*** serving as the reference point, detailed in Table 5.

Our proposed SMILE method results in rich vocabularies, with a significantly higher number of unique unigrams, bigrams and trigrams compared to the baseline methods. Specifically, a simple and fixed prompt tends to produce monotonous output. When dialogue topics are incorporated into a single, fixed prompt, the model's output demonstrates substantial diversification. Furthermore, the SMILE method outperforms the baseline methods in terms of Distinct-1, Distinct-2 and Distinct-3.

F.2 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:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \quad (5)$$

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 56 distinct dialogue topics, as illustrated in Appendix K and Figure 10, to automatically label dialogue topics for each dialogue with GPT-4. The model we use is `gpt-4-0613`. We present the information entropy for each prompt method in Table 6, demonstrating that the dialogues generated using the SMILE method are substantially more diverse than those generated using the `standard` method and are compatible with the `stand.`$_{\text{DT}}$ method, which uniformly samples dialogue topics.

G Expert Evaluation

To illustrate the patterns of supporters' strategies, we compile statistics on the flow patterns of these strategies, including "Supporting", "Challenging" and "Other".

G.1 Experimental Setup

PsyQA (Sun et al., 2021) contains 4012 QAs with supporters' strategies in the original dataset. Xinling (Li et al., 2023) is a real-life multi-turn counseling dataset containing 550 dialogues with supporters' strategies in the original dataset.

Three methods (`standard`, `stand.`$_{\text{DT}}$, and SMILE) each contain 500 dialogues, respectively. We use Xinling to train a classifier to predict the supporters' strategies in dialogues generated with these three methods. The performance of the trained classifier is comparable to the results in the original paper.

G.2 Results

Supporters have three main strategies: "Supporting", "Challenging", and "Other", and we present the results in Table 7, 8 and 9. We highlight the gap greater the 0.2 in bold and the corresponding gap is in parentheses.

[5]https://huggingface.co/THUDM/chatglm2-6b

Metrics	# Unique Unigrams	# Total Unigrams	Distinct-1 (⇑)	# Unique Bigrams	# Total Bigrams	Distinct-2 (⇑)	# Unique Trigrams	# Total Trigrams	Distinct-3 (⇑)
PsyQA*	11785	203306	0.058	120049	202806	0.592	182942	202306	0.904
standard	4174	153536	0.027	32281	153036	0.211	72340	152536	0.474
$\texttt{stand.}_{\text{DT}}$	6032	175319	0.034	52141	174819	0.298	105921	174319	0.608
SMILE	**10447**	**254585**	**0.041**	**111662**	**254085**	**0.439**	**196367**	**253585**	**0.774**

Table 5: Statistics of 500 conversations in each prompt method, including PsyQA*.

Setting	standard	$\texttt{stand.}_{\text{DT}}$	**SMILE**
Limited Topics	8.11	**14.28**	14.07
Unlimited Topics	8.40	14.76	**15.02**
Average	8.26	14.52	**14.55**

Table 6: Information entropy of dialogue topics. The setting of limited topics indicates that the dialogue topics in the statistics are confined to the 56 dialogue topics we provide. The setting of unlimited topics denotes that the dialogue topics not only belong to our provided 56 distinct topics but also include topics generated spontaneously by GPT-4.

Strategy	**Supporting**				
Methods	standard	stand.DT	SMILE	PsyQA	**Xinling (Oracle)**
Stage 1	0.854(+0.101)	0.847(+0.094)	0.615(-0.138)	**0.474(-0.279)**	0.753
Stage 2	**0.547(-0.278)**	**0.520(-0.305)**	**0.432(-0.393)**	**0.312(-0.513)**	0.825
Stage 3	**0.419(-0.367)**	**0.365(-0.421)**	**0.360(-0.426)**	**0.179(-0.607)**	0.786
Stage 4	**0.468(-0.258)**	**0.387(-0.339)**	**0.390(-0.336)**	**0.107(-0.619)**	0.726
Stage 5	**0.808(+0.382)**	**0.774(+0.348)**	**0.662(+0.236)**	**0.154(-0.272)**	0.426

Table 7: Pattern flow of "Supporting" strategy.

G.3 Conclusion

Dialogues generated by the three methods exhibit similar flow patterns of supporters' strategies. Instructing ChatGPT to rewrite QAs into multi-turn dialogues will inevitably impact the strategy pattern in the generated dialogues. Dialogues generated by the three methods still exhibit significant gaps when compared to the Xingling (Oracle) dataset with respect to the flow patterns of supporters' strategies.

H Implementation Details

Model Training. We implement ChatGLM2-6B with the Transformers library. Table 10 shows the parameters of parameter-efficient fine-tuning.

Dialogue Generation We use ChatGLM2-6B as the baseline model, and all hyperparameters during generation are set to their default values from the official repository[6]. After tuning ChatGLM2-6B with LoRA, we set the maximum generation length to 2040, temperature to 0.8, and adopt nucleus sampling with $p = 0.8$.

[6]https://huggingface.co/THUDM/chatglm2-6b

Strategy	**Challenging**				
Methods	standard	stand.DT	SMILE	PsyQA	**Xinling(Oracle)**
Stage 1	0.140(+0.084)	0.150(+0.094)	**0.372(+0.316)**	**0.348(+0.292)**	0.056
Stage 2	**0.452(+0.305)**	**0.478(+0.331)**	**0.565(+0.418)**	**0.598(+0.451)**	0.147
Stage 3	**0.576(+0.384)**	**0.632(+0.440)**	**0.636(+0.444)**	**0.752(+0.560)**	0.192
Stage 4	**0.516(+0.272)**	**0.603(+0.359)**	**0.586(+0.342)**	**0.847(+0.603)**	0.244
Stage 5	0.112(-0.127)	0.157(-0.082)	0.231(-0.008)	**0.794(+0.555)**	0.239

Table 8: Pattern flow of "Challenging" strategy.

Strategy	**Other**				
Methods	standard	stand.DT	SMILE	PsyQA	**Xinling(Oracle)**
Stage 1	0.006(-0.185)	0.003(-0.188)	0.013(-0.178)	0.178(-0.013)	0.191
Stage 2	0.001(-0.027)	0.002(-0.026)	0.003(-0.025)	0.090(+0.062)	0.028
Stage 3	0.005(-0.017)	0.003(-0.019)	0.004(-0.018)	0.069(+0.047)	0.022
Stage 4	0.016(-0.014)	0.010(-0.020)	0.024(-0.006)	0.046(+0.016)	0.03
Stage 5	**0.080(-0.255)**	**0.069(-0.266)**	**0.107(-0.228)**	**0.052(-0.283)**	0.335

Table 9: Pattern flow of "Other" strategy.

I Prompt Details

Prompt A is directly utilized from an online GitHub repository, while prompt B is designed by our professional counselors.

Prompt A[7] 我想让你担任心理医生。我将为您提供一个寻求指导和建议的人，以管理他们的情绪、压力、焦虑和其他心理健康问题。您应该利用您的认知行为疗法、冥想技巧、正念练习和其他治疗方法的知识来制定个人可以实施的策略，以改善他们的整体健康状况。(English translation: I want you to take on the role of a psychologist. I will provide you with someone seeking guidance and advice to manage their emotions, stress, anxiety, and other mental health issues. You should use your knowledge of cognitive-behavioral therapy, meditation techniques, mindfulness practices, and other therapeutic methods to formulate strategies that individuals can implement to improve their overall health.)

Prompt B 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。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

[7]https://github.com/PlexPt/awesome-chatgpt-prompts-zh

Epoch	Learning Rate	Batch Size	LoRA Rank	LoRA Dropout	LoRA α	Seed
2	1e-4	1	16	0.1	64	1234

Table 10: Parameters of parameter-efficient fine-tuning.

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.)

J 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 11.

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.

Pairwise human evaluation results, comparing the baseline model to ground truth, are reported, as illustrated in Figure 12. Results show that the baseline model, without training with SMILECHAT, lags significantly behind compared to the ground truth.

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 naturalness, human-like qualities, correctness, and appropriateness, and select the optimal response for the dialogue history, providing a preference.	
Examples	
Dialogue History	Help-seeker: xxx Supporter: xxx Help-seeker: xxx Supporter: xxx Help-seeker: xxx Supporter: xxx Help-seeker: xxx … Help-seeker: xxx Supporter:
Response A	{Response A}
Response B	{Response B}
Response C	{Response C}

A B | A C | B C

Figure 11: Labeling instruction.

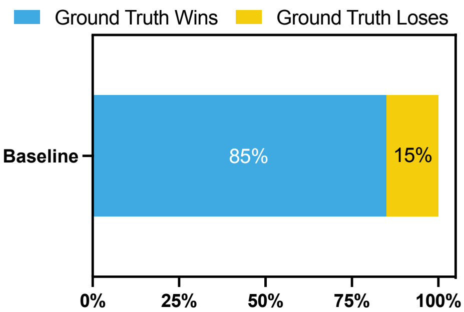

Figure 12: Pairwise human evaluation results, comparing the baseline model to ground truth, are reported. We present the win and lose rates of each compared pair in 100 randomly sampled real-life dialogue sessions. Fleiss' kappa (Fleiss et al., 1981) is used to measure the inter-rater agreement, and the value falls within moderate agreement with $0.5 \leq \kappa \leq 0.6$.

K Definitions of Dialogue Topics

1. Mate Selection: Do not know how to enter into an intimate relationship, have fear or aversion towards intimate relationships, are unwilling to date, or encounter some difficulties in choosing a spouse.

2. Love Issues: Problems encountered by individuals in a romantic state during the process of love, including both long-distance and non-long-distance relationships.

3. Post-love Issues: Interpersonal relationship management with ex-partners after the end of an intimate relationship.

4. Marriage Issues: Limit to relationship problems between spouses. Marriage issues refer to a series of problems that arise between a husband and wife after marriage. They generally include extramarital affairs, emotional infidelity, domestic violence, personality defects, personality disorders, mother-in-law and daughter-in-law relationships, dull marriage, interpersonal communication, cultural differences, sexual life, and problems in marriages caused by practices like polygamy or polyandry.

5. Concepts of Sex Distress: Sexual concept distress refers to the distress present in understanding and views regarding sexual physiology, psychology, behavior, morality, and civilization.

6. Sexual Preference Distress: Sexual preference distress refers to the psychological distress caused by paraphilias such as fetishism, transvestic fetishism, exhibitionism, friction fetishism, voyeurism, zoophilia, pedophilia, sadomasochism, and necrophilia.

7. Gender Identity/Cognition: It refers to the deeply felt gender within a person, based on personal experience, which may align with the gender assigned at birth (i.e., cisgender) or differ from it (i.e., transgender).

8. Sexual Orientation: Refers to the gender(s) a person is attracted to (including the opposite sex, same-sex, no specific sex, both sexes).

9. Family Conflict: Conflicts, disputes, and communication issues among family members, including those originating from both the nuclear family and the extended family.

10. Child Education: Literally meaning the education of children.

11. Domestic Violence: Physically, mentally, or otherwise harmful actions carried out between family members, including acts such as physical assault, binding, harming, and restricting personal freedom, as well as consistent verbal abuse, intimidation, and similar behaviors.

12. Sexual Harassment: Sexual harassment refers to using sexually suggestive language or actions towards the target of harassment, coercing the victim to comply and making them uncomfortable.

13. Sexual Assault: Sexual assault involves various forms of unwanted sexual contact and coerced sexual activities, including rape, forced kissing, sexual harassment, sexual abuse, exhibitionism, voyeurism, and such actions might also be considered sexual assault in legal precedents.

14. Bullying: Typically refers to the harassment and oppression between individuals with unequal power, which has long existed in society. This includes physical or verbal attacks, resistance and exclusion in interpersonal interactions. It can also involve talking about someone in a sexual harassment-like manner or mocking and making comments about their body parts, or it could be insults and sarcasm driven by personal reasons like jealousy.

15. Loss: The loss of significant others or pets.

16. Setbacks: Inevitable emotional reactions resulting from the obstruction of purposeful actions by individuals. It can cause substantial harm and manifest as disappointment, pain, distress, and unease.

17. Political Violence: Rooted in class domination and relying on the state as its entity, political coercion is the force exerted by the state, relying on various means of coercion such as the military, police, courts, and prisons.

18. Secondary Trauma: After witnessing a large number of cruel and destructive scenes, the degree of psychological and emotional distress exceeds the tolerance limit of some individuals, leading to various abnormal psychological phenomena. These phenomena are often a result of sympathy and empathy for survivors and their trauma, causing severe mental and emotional distress and even psychological breakdown. The main symptoms of secondary trauma include loss of appetite, easy fatigue, decreased physical energy, sleep disturbances (difficulty falling asleep, easy awakening), nightmares, irritability or anger, easy startle response, lack of concentration, feeling numb, fearful, and hopeless about one's own experiences, accompanied by trauma reactions and interpersonal conflicts.

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

20. Psychological Counseling Trauma: Harm

caused by therapists during psychological counseling.

21. Health Issues: Life and psychological distress caused by diseases such as heart disease, thyroid nodules, polycystic ovary syndrome, et al.

22. Psychosomatic Symptoms: Including palpitations (a subjective feeling of discomfort with heartbeats), sleep problems (such as insomnia/nightmares), which, if caused by sleep issues leading to other problems, are categorized as sleep problems; eating issues (such as loss of appetite/emotional eating); memory problems; stomach pain; dizziness; fainting; difficulty breathing; physical weakness/tiredness without apparent reason; unexplained bodily pains et al.

23. School/Workplace Adaptation: Literal meaning.

24. Role Transition Adaptation: For example, new mothers/fathers, new wives/husbands, retirees.

25. Cultural Adaptation: Adaptive issues encountered during the ongoing direct contact between individuals from different cultural groups in the process of changes occurring in one or both parties due to their original cultural backgrounds.

26. Self-exploration and Growth: Self-exploration and growth in the course of life development (adolescence, early adulthood, middle age, late adulthood).

27. Personality Trait Exploration: Traits, formation, origins, influences, et al., of personality and character.

28. Exploration of Negative Self-evaluation: Not knowing how to love oneself, feelings of inferiority, low self-esteem, self-denial, self-doubt, internal conflicts, self-contradictions, fearing being different, sensitivity, lack of security, feeling inadequate, et al.

29. Exploration of Life Meaning: Sense of meaninglessness, existential emptiness.

30. Emotion Regulation/Control Methods Consultation: Literal meaning.

31. Depression: An emotion characterized by "low mood, slow thinking, reduced speech and movement."

32. Anxiety: A restless emotion arising from excessive worry about the safety of loved ones, one's own life, future, destiny, et al.

33. Stress: Stress is a cognitive and behavioral experiential process formed by stressors and stress responses, namely psychological stress.

34. Obsession: Marked by an excessive pursuit of perfection, precision, and a tendency to rationalize conflicts. It involves strong self-control tendencies and self-regulatory behavior to the point of entanglement and nitpicking. Behaviorally, it adheres excessively to rules, regulations, and orders, to the extent that even life details are sought to be procedural and ritualistic, demanding step-by-step compliance.

35. Fear: A strong and repressed emotional state felt deeply by an individual or group in the face of real or imagined danger. Manifestations include heightened nervousness, overwhelming inner fear, inability to concentrate, mental blankness, impaired judgment or self-control, and increased impulsiveness.

36. Dilemma/Decision-Making Difficulty: Faced with numerous choices and concerns, dealing with dilemmas or even multiple predicaments, struggling to make decisions.

37. Impulsiveness/Loss of Control: Often refers to acting recklessly without considering consequences. It involves intense emotions and weak rational control.

38. Interpersonal Communication Skills/Methods Consultation: Inquire about handling interpersonal relationships and more.

39. Interpersonal Conflicts/Disputes: Conflicts, disagreements, dissatisfaction, or communication issues arise during interpersonal interactions. If there are other issues caused by interpersonal relationships, they are also categorized as interpersonal interactions.

40. Social Difficulties/Fears: Characterized by involuntarily feeling nervous and frightened when interacting with others (especially in public settings), leading to confusion, incoherent speech, and even severe fear of social encounters.

41. Learning Efficiency/Methods: Literal meaning.

42. Work Efficiency/Methods: Literal meaning.

43. Job Dissatisfaction: Includes dissatisfaction with salary, benefits, environment, systems, personnel, et al. in a work setting.

44. Learning Dissatisfaction: Includes dissatisfaction with interpersonal relationships, environment, systems, et al. in an educational setting.

45. Occupational Burnout: Refers to the state of physical and mental fatigue and exhaustion experienced by individuals under heavy work pressure.

46. Learning Fatigue: Refers to a phenomenon where students hold a negative attitude towards school courses and academics, accompanied by the

following behavioral manifestations: loss of enthusiasm for academic tasks and school activities, displaying a negative state, and showing indifference and estrangement towards classmates and friends.

47. Occupational Stress/Challenges: Various pressures/challenges that arise or form in the workplace, including stress/challenges caused by factors such as heavy workload, difficulties in interpersonal communication, and the impact of changes in the work environment.

48. Academic Pressure: Refers to the mental burden that individuals bear during learning activities. This includes various tense stimuli from the environment experienced during the learning process, as well as measurable and assessable abnormal reactions in the physiological, psychological, and social behavior of students.

49. Dropping Out: Stopping attending school midway or losing the opportunity to attend school.

50. Underemployment/Unemployment: Underemployment refers to the behavior of not finding a job and waiting for job opportunities. Unemployment refers to the situation of a person within a certain age range who is willing and capable of working for remuneration but has yet to find a job.

51. Career Planning: Refers to the ongoing systematic process of planning for one's career and even life. It includes three elements: career orientation, goal setting, and pathway design.

52. Professional/Career/Occupational Choices/Changes: Literal meaning.

53. Suspected Neuroses and Mental Disorders: Highly suspected of having neuroses and mental disorders, recommended type of consultation.

54. Neuroses and Mental Disorders: It is a general term for a group of mental disorders, including neurasthenia, obsessive-compulsive disorder, anxiety disorder, phobia, somatic symptom disorder, et al. Patients experience significant distress and impairment in psychological or social functioning, but there is no confirmed organic pathological basis. The course of the illness is mostly prolonged or episodic. Mid- to long-term consultations are needed. Mental disorders refer to a range of conditions that meet the diagnostic criteria of DSM-5 (substance addiction or dependence also included) and require hospital consultation.

55. Self-harm/Suicidal Tendencies: Describing incidents of suicide/self-harm/self-mutilation/contemplating suicide. Triggers crisis alerts and referrals.

56. Harming Others/Killing: Describing incidents of harming others/killing/thinking of killing.

L Dialogue Examplars

Multi-turn dialogue examples generated with the `standard` and `stand.`$_{\text{DT}}$ methods are illustrated in Figure 13 and 15. Further, a dialogue generated by the SMILE method is shown in Figure 17.

求助者：我最近总是感到焦虑和压力很大，无法入睡。我不知道怎么办才好。
支持者：听起来你正面临着一些困扰。首先，你要知道你并不孤单，很多人都经历过类似的情况。要不要先尝试一些放松的方法，比如深呼吸或者听一些舒缓的音乐来帮助你缓解压力？
求助者：可能我太过紧张，我试过做深呼吸和听音乐，但只能短暂缓解，然后焦虑又回来了。
支持者：那可能是因为你需要找到更有效的应对方式。你可以试试通过身体活动来释放压力，比如散步、跑步或者瑜伽。运动不仅可以疏导你的负面情绪，还能帮助你保持身心健康。
求助者：或许我确实需要开始运动了。除了运动，你有没有其他的建议来帮我减轻焦虑？
支持者：当然，除了身体活动，均衡的饮食和良好的睡眠也非常重要。你可以尝试每天保持规律的作息时间，避免摄入刺激性食物和饮料，如咖啡和巧克力。同时，也可以试试冥想和深层放松练习，它们对缓解焦虑有帮助。
求助者：听起来这些方法都很有效。我会试试看的。但有时候我觉得自己控制不住焦虑情绪，它总是在困扰我。
支持者：明白你的感受。焦虑情绪确实让人不安。但我们要记住，我们可以通过一些技巧来调节自己的情绪。比如，当你感到焦虑时，可以尝试用目光聚焦来分散注意力，或者快速数数来制止负面思绪。这样可以帮助你重新获得情绪的平衡。
求助者：这主意不错，下次我会试试。谢谢你的建议和支持。
支持者：不客气，我很高兴能够为你提供帮助。除了这些方法，如果你觉得焦虑情绪依然无法控制，不妨考虑寻求专业的心理辅导，专家会给你更专业的意见和指导。
求助者：如果情况再次恶化，我会考虑去找专业帮助。谢谢你的鼓励和支持，我觉得好多了。
支持者：很高兴听到你感觉好多了。记住，无论何时，我都在这里支持你。不要犹豫，随时都可以来找我倾诉。你不是一个人在战斗，我们一起度过这段困难时期。

Figure 13: A dialogue example generated with the `standard` method. (Chinese Version)

Help-seeker: I've been feeling very anxious and stressed lately, and I can't sleep. I don't know what to do.
Supporter: It sounds like you're facing some challenges. First, you need to know that you're not alone; many people have experienced similar situations. Would you like to try some relaxation techniques, such as deep breathing or listening to soothing music, to help alleviate your stress?
Help-seeker: Maybe I'm too tense. I've tried deep breathing and listening to music, but it only provides temporary relief, and then the anxiety comes back.
Supporter: That might be because you need to find more effective coping mechanisms. You could try releasing stress through physical activity, such as walking, running, or yoga. Exercise not only helps channel negative emotions but also contributes to maintaining your overall well-being.
Help-seeker: Perhaps I do need to start exercising. Besides physical activity, do you have any other suggestions to help reduce my anxiety?
Supporter: Certainly. In addition to physical activity, a balanced diet and good sleep are also crucial. Try maintaining a regular sleep schedule, avoiding stimulants like coffee and chocolate. Additionally, consider trying meditation and deep relaxation exercises; they can be helpful in relieving anxiety.
Help-seeker: It sounds like these methods are effective. I'll give them a try. But sometimes I feel like I can't control my anxiety; it keeps bothering me.
Supporter: I understand your feelings. Anxiety can indeed be unsettling. But remember, we can use techniques to regulate our emotions. For instance, when you feel anxious, try using focused gaze to distract your attention, or count quickly to interrupt negative thoughts. This can help you regain emotional balance.
Help-seeker: That's a good idea; I'll try it next time. Thank you for your advice and support.
Supporter: You're welcome. I'm glad I could help. In addition to these methods, if you find that your anxiety remains uncontrollable, consider seeking professional psychological counseling. Experts can provide more specialized advice and guidance.
Help-seeker: If the situation worsens, I'll consider seeking professional help. Thank you for your encouragement and support; I feel much better now.
Supporter: I'm glad to hear that you're feeling better. Remember, I'm here to support you anytime. Don't hesitate to reach out. You're not alone in this battle; we'll get through this challenging period together.

Figure 14: A dialogue example generated with the `standard` method. (English Version)

求助者：最近我有点困扰。我不知道怎样进入一段亲密的关系，对亲密关系有些恐惧和厌恶，不愿意约会，或者在选择伴侣方面遇到了一些困难。
支持者：听起来你对进入一段亲密的关系有些困惑和困扰。可以告诉我你具体是如何感受的吗？
求助者：我感到很害怕和不舒服，一想到要和别人建立深入的感情联系就感到无法忍受。
支持者：这种感受可以是非常困扰的。或许你可以试着了解一下自己内心的原因，或者是过去的经历对你产生了什么影响。
求助者：我曾经经历过一段情感上的创伤，可能是导致我对亲密关系和约会感到害怕的原因。
支持者：了解，经历过情感创伤可以对我们产生深远的影响。你曾经尝试过寻求专业帮助吗？
求助者：没有，我一直都觉得这是我自己的问题，不知道该找谁帮助。
支持者：你不必独自承受这个问题。专业心理咨询师可以给予你更具体的帮助和指导，帮助你应对这个挑战。我可以为你提供一些相关的资源和建议。
求助者：谢谢你的帮助，我会考虑尝试寻找专业帮助的。
支持者：非常好，这是一个积极的决定。无论你在这个过程中遇到什么困难，都记得你不是一个人在面对这个问题。
求助者：我会记住的。但我现在还是感到有些害怕去谈论和处理这个问题。
支持者：害怕是正常的，但要相信自己的内在力量和能力。专业咨询师会帮助你逐渐面对和处理这个问题，你也可以通过与亲密的朋友和家人分享来减轻压力。
求助者：嗯，我会试试这些方法。谢谢你的鼓励和支持。
支持者：我相信你能够克服这个困难，你并不孤单。记得，你可以随时来找我聊天，我会一直支持你的。

Figure 15: A dialogue example generated with the `stand.`$_{\mathrm{DT}}$ method. (Chinese Version)

Help-seeker: Recently, I've been feeling a bit troubled. I don't know how to enter into an intimate relationship, and I have some fear and aversion towards intimacy. I'm unwilling to date, and I'm facing difficulties in choosing a partner.
Supporter: It sounds like you're confused and troubled about entering an intimate relationship. Can you tell me more about how you specifically feel?
Help-seeker: I feel very scared and uncomfortable. The thought of establishing deep emotional connections with others feels unbearable to me.
Supporter: This feeling can be very distressing. Perhaps you can try to understand the reasons within yourself or explore how past experiences may have influenced you.
Help-seeker: I have experienced emotional trauma before, which may be the reason I feel scared about intimate relationships and dating.
Supporter: I see. Experiencing emotional trauma can have a profound impact on us. Have you ever tried seeking professional help?
Help-seeker: No, I've always felt it's my own problem and didn't know who to turn to for help.
Supporter: You don't have to face this problem alone. Professional psychologists can provide you with more specific assistance and guidance to help you cope with this challenge. I can offer you some relevant resources and suggestions.
Help-seeker: Thank you for your help. I will consider trying to seek professional help.
Supporter: That's very positive. Whatever difficulties you encounter in this process, remember that you are not alone in facing this issue.
Help-seeker: I will remember that. But right now, I still feel a bit afraid to talk about and deal with this problem.
Supporter: Fear is normal, but trust in your inner strength and abilities. Professional counselors can help you gradually face and address this issue, and you can also alleviate the pressure by sharing with close friends and family.
Help-seeker: Yes, I will try these methods. Thank you for your encouragement and support.
Supporter: I believe you can overcome this challenge. You're not alone. Remember, you can always come to talk to me, and I'll be here to support you.

Figure 16: A dialogue example generated with the `stand.`$_{\rm DT}$ method. (English Version)

求助者：高三后的迷茫，高考前的恐惧，能给我一些建议么？
支持者：你看到你的问题了。别担心，我们一起慢慢解决。高考期间的你应该正在面临巨大的压力，有多久没有静下心来好好和自己相处了。
求助者：确实很长时间没有好好照顾自己了。在这个阶段的父母，老师，身边的同学都在承担着各自的焦虑。我真的感觉好难过，好像一切都不需要我。
支持者：我理解你的感受。在这个年纪，我们都有时候会觉得自己被否认，被忽视，甚至觉得自己多余。但是请相信，你对这个世界和他人都是有价值的。
求助者：是的，我需要找回对自己的价值感。我现在总是过着非常累的生活，几乎没有朋友，只能一个人独处。我不知道该怎么面对高考。
支持者：高考的确是个很大的挑战，但我们不能让它完全定义我们的价值。我们要学会调整好自己的情绪，不被别人轻易影响，不要盲目去证明自己。我们需要冷静下来，好好思考怎样更好地改善自己。
求助者：你说的对，高考只是一个阶段。但是我感觉我不知道该怎么学习，每次考试都给我打击。我真的很迷茫。
支持者：学习是个具体的技巧，我们可以一起学习如何安排时间、休息，甚至在休息的时候做些什么。而且，我们也可以通过和同学、老师交流来不断提高自己的学习方法。只要我们保持积极的心态，相信会有进步的。
求助者：谢谢你的鼓励和建议。我觉得我需要更多的支持和理解，不再觉得孤独和无助。你有什么其他建议吗？
支持者：当你感到难过和迷茫的时候，可以寻找一些积极向上的事物来激励自己，比如读一些励志书籍，观看一些鼓舞人心的视频。还有，记得不要忘记和朋友一起交流，他们的支持和鼓励也会帮助你度过难关。
求助者：嗯，我会尝试去找一些正能量的事情来给自己一些动力。我现在真的很感激你的陪伴和帮助。
支持者：没关系，我会一直陪你度过高考的。相信自己，你一定能够克服困难，取得好成绩。加油！
求助者：谢谢你的信任和鼓励。我会努力的！

Figure 17: An example of multi-turn dialogue generated by SMILE method. (Chinese Version)

Help-seeker: I'm feeling lost after finishing high school, and I'm scared about the college entrance exam. Can you give me some advice?
Supporter: You've identified your concerns. Don't worry; we'll work through this together. You're likely under immense pressure during the college entrance exam period. When was the last time you took a moment to be with yourself and relax?
Help-seeker: It's been a long time since I took care of myself properly. At this stage, parents, teachers, and classmates around me are all dealing with their anxieties. I feel really sad, like nothing matters.
Supporter: I understand how you feel. At this age, we sometimes feel denied, ignored, or even redundant. But please believe that you have value to the world and others.
Help-seeker: Yes, I need to rediscover my sense of self-worth. I'm living a very tiring life now, almost no friends, always alone. I don't know how to face the college entrance exam.
Supporter: The college entrance exam is indeed a significant challenge, but we shouldn't let it completely define our value. We need to learn to adjust our emotions, not be easily influenced by others, and not blindly prove ourselves. Let's calm down and think about how to improve ourselves.
Help-seeker: What you're saying is right; the college entrance exam is just a phase. But I feel lost; I don't know how to study, and every exam is a setback for me.
Supporter: Learning is a specific skill, and we can learn together how to manage time, rest, and even what to do during breaks. Moreover, we can continuously improve our study methods through communication with classmates and teachers. As long as we maintain a positive mindset, progress is possible.
Help-seeker: Thank you for your encouragement and advice. I feel like I need more support and understanding, not to feel lonely and helpless anymore. Do you have any other suggestions?
Supporter: When you feel sad and lost, look for positive things to motivate yourself, like reading inspirational books or watching uplifting videos. Also, don't forget to communicate with friends; their support and encouragement will help you get through tough times.
Help-seeker: Yeah, I'll try to find some positive things to give myself motivation. I'm really grateful for your companionship and help.
Supporter: It's okay; I'll be with you throughout the college entrance exams. Believe in yourself; you can overcome difficulties and achieve good results. Keep going!
Help-seeker: Thank you for your trust and encouragement. I'll do my best!

Figure 18: An example of multi-turn dialogue generated by SMILE method. (English Version)