PsyQA: A Chinese Dataset for Generating Long Counseling Text for Mental Health Support

Hao Sun[1*], **Zhenru Lin**[2*], **Chujie Zheng**[1], **Siyang Liu**[3], **Minlie Huang**[1†]

[1]The CoAI group, DCST, Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems,
[1]Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
[2]Dept. of Electronic Engineering, Tsinghua University, Beijing 100084, China
[3]Tsinghua-Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School,
[3]Tsinghua University, Shenzhen, China
{h-sun20,linzr18}@mails.tsinghua.edu.cn, aihuang@tsinghua.edu.cn

Abstract

Great research interests have been attracted to devise AI services that are able to provide mental health support. However, the lack of corpora is a main obstacle to this research, particularly in Chinese language. In this paper, we propose PsyQA, a Chinese dataset of psychological health support in the form of question and answer pair. PsyQA is crawled from a Chinese mental health service platform, and contains 22K questions and 56K long and well-structured answers. Based on the psychological counseling theories, we annotate a portion of answer texts with typical strategies for providing support, and further present in-depth analysis of both lexical features and strategy patterns in the counseling answers. We also evaluate the performance of generating counseling answers with the generative pretrained models. Results show that utilizing strategies enhances the fluency and helpfulness of generated answers, but there is still a large space for future research.

1 Introduction

The burden of mental disorders continues to grow with significant impacts on human health and social development (Organization et al., 2011; The World Health Organization, 2020). As an effective therapy for mental disorders (Reynolds Jr et al., 2013), online mental health counseling, which mostly refers to communicating anonymously, has become popular in recent years (Fu et al., 2020).

Great research interests have been endeavored to devise AI services that are able to provide mental health support (Bucci et al., 2019; Liu et al., 2021). Based on the online-text psychotherapy corpora, previous works have utilized text mining techniques to detect empathy (Sharma et al., 2020; Zheng et al., 2021), linguistic development of counselors (Zhang et al., 2019), and self-injurious thoughts and behaviors (Franz et al., 2020). However, the research of text-based mental health counseling is still largely limited due to the lack of relevant corpora, particularly in Chinese language.

To this end, we collect **PsyQA** in this work, a Chinese dataset of **Psy**chological health support in the form of **Q**uestion-**A**nswer pair. An example data of PsyQA is shown in Figure 1. In each example, the ***question*** along with a detailed ***description*** and several ***keyword*** tags is posted by an anonymous help-seeker, where the description generally contains dense persona and emotion information about the help-seeker. The ***answer*** is usually quite long (524 words on average). The answers are replied asynchronously from the well-trained volunteers or professional counselors, and contain both the detailed analysis of the seeker's problem and the guidance for the seeker. Moreover, a portion of the answers are also additionally annotated by professional workers with typical support ***strategies***, which are based on the psychological counseling theories (Hill, 2009).

Our collected PsyQA has three distinct characteristics. Firstly, the corpus covers abundant mental health topics from 9 categories including emotion, relationships, and so on (refer to Appendix for topic statistics). Secondly, the answers in PsyQA are mostly provided by experienced and well-trained volunteers or professional counselors. Thirdly, we provide support strategy annotations for a portion of answers, which can greatly facilitate future research on our corpus. As will be shown later, there are not only lexical features in the texts of different support strategies (Section 4.2), but also explicit patterns of strategy organization and utilization in the answers (Section 4.3).

To validate whether existing models can generate long counseling answers to mental health questions, we conduct experiments on both strategy identifi-

[*]Equal contribution.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1489–1503
August 1–6, 2021.

Question (Post Title)

为什么有些事情越想心越闷?
The more I think about some things, the more upset I feel. Why?

Description (Post Content)

比如家里人说我和别人相亲的事，我是不太主动比较被动的人……又怪自己，又恨别人到处说，搞得心里很难受很郁闷，这该怎么办?
For example, my family asked me to go on a blind date with others. I am not an active but passive person. ... I blame myself and blame others for speaking ill of me everywhere, making me very uncomfortable and depressed. What should I do?

Keywords

情绪 表达情绪 情绪调节 情绪智力
Emotion, Emotion Expression, Emotion Regulation, EQ

Answer (Response)

你好呀～事情越想越闷可能是陷入了反刍思维中。反刍式思考是指 …… 反刍思维作为一种认知，对情绪也有重要的影响。在这种情况下，你首先要冷静下来 ……比如自己闷在家里没出去相亲，家人就说自己是不是想打光棍儿。其实你仔细看这两件事情并没有因果关系。……但这样的逻辑也是不太合情理的。当然，在这种情况下，你也可以使用转移注意力的方式，让自己的情绪稍微平复下来。比如做一次冥想练习，或者出去做运动。

Hi ~ The more you think about it, the more depressed you feel. This is maybe because you are trapped in ruminant thinking. Ruminant thinking means that ..... Ruminant thinking, as a form of cognition, also has an important effect on emotion. In this case, you need to calm down first… For example, you stayed at home and didn't go out for a blind date, and your family said that you just wanted to be single. When you look at it carefully, there is no causal relationship between the two events. ..... But this logic doesn't work. Of course, in this case, you can also distract your attention to calm yourself down a bit. Take a meditation practice, or go outside to exercise.

Figure 1: An example from PsyQA. $(Question, Description, Keyword)$ triples are posted by help-seekers while $Answer$ is provided by help-supporters. Different strategies in the answer are colored differently. Strategies Information, Interpretation, Restatement, and Direct Guidance are used in this answer. Note that a question may have multiple answers.

cation (Section 5) and answer generation (Section 6). We find that the contextual information greatly benefits the performance of support strategy identification. Experimental results also demonstrate that utilizing support strategies improves the answers generated by the models in terms of their language fluency, coherence, and the ability to be on-topic and helpful. However, there is still much room for further research compared to the answers written by well-trained volunteers or professional counselors.

Our contributions are summarized as follows:

- We collect PsyQA, a high-quality Chinese dataset of psychological health support in the form of QA pair. The answers in PsyQA are usually long, which are provided by well-trained volunteers or professional counselors.
- We annotate a portion of answer texts with a set of strategies for mental health support based on psychological counseling theories. Our analysis reveals that there are not only typical lexical features in the texts of different strategies, but also explicit patterns of strategy organization and utilization in the answers.
- We conduct experiments of both strategy identification and answer text generation on PsyQA. Results demonstrate the importance of using support strategies, meanwhile indicating a large space for future research.

2 Related Work

Our work primarily concerns linguistic behavior for counseling, NLP for mental health detection and therapy, and text-based mental health-related datasets.

2.1 Linguistic Behaviors in Counseling

Hill's model (Hill, 2009) consists of three stages: exploration, insight, and action in which helpers guide clients in exploring their thoughts and feelings, discovering the origins and consequences of maladaptive thoughts and behaviors, and acting on those discoveries to create positive long-term change. We draw on Hill's model and apply it to formulate the answer in the PsyQA dataset.

Some previous work explored how mental health support is sought and provided. For example, some studies measure how the language of comments in Reddit mental health communities influences risk

to suicidal ideation in the future (De Choudhury and Kiciman, 2017), and seek to understand how counselors' behaviors develop over time (Zhang et al., 2019). While these previous studies model implicit linguistic behaviors of counselors, we focus on linguistic strategy development in a long psychological response, considering the strategies as a skeleton to generate the whole response.

2.2 NLP for Mental Health Detection and Therapy

Some prior work analyzed the posts and blogs of users with the rise of social networking sites (SNS), attempting to employ NLP techniques to detect depression (Tadesse et al., 2019; Yates et al., 2017), suicidal ideation (Zirikly et al., 2019; Cao et al., 2019), and other general mental health problems (Xu et al., 2020). In another line of work, some researchers endeavored to construct "therapybots" (Fitzpatrick et al., 2017; Inkster et al., 2018), and focused on therapy and attempted to create dialogue agents to provide therapeutic benefit, where the effectiveness of web-based cognitive-behavioral therapeutic (CBT) apps or mobile mental well-being apps are explored. Adopting a more straightforward method, we make the machine generate answers to a detailed question, mimicking a mental health counselor. Though the ultimate goal is to develop systems for real-world treatment, there is still a long way to go in this direction and our corpus can be the first step towards building intelligent systems for this purpose, and offers the opportunity for studying the effectiveness of using explicit strategies in the systems.

2.3 Text-based Mental Health-Related Datasets

There are some datasets for mental health detection and therapy. However, most of them are collected from general social networking sites such as Twitter, Reddit, and Weibo (Harrigian et al., 2020). General social networking sites contain irrelevant posts or unprofessional responses, which might put NLP systems trained on these corpora at huge risk. Thus, some previous work focused on the counseling part in the online mental health communities (forums), such as TeenHelp (Franz et al., 2020), TalkLife (Sharma et al., 2020). In Chinese domain, Wang et al. (2020) collected a public counseling conversation dataset by crawling. However, most responses in this dataset are short and general without any suggestion. Crisis Text Line (Althoff et al., 2016) presents the best mental health counseling dataset up to now. It contains a large-scale multi-turn counseling conversation by experienced volunteer counselors[1]. Different from Crisis Text Line, PsyQA focuses on Chinese long-text response in a single-turn asynchronous counseling conversation.

From the perspective of the mental health domains, most of the prior work is focusing on single-domain like depression, suicidal ideation, and eating disorders (Harrigian et al., 2020). Instead, PsyQA contains all sorts of general mental health disorders, concerning nine topics labeled by help-seekers including self-growth, emotion, love problem, relationships, behaviors, family, treatment, marriage, and career.

3 Data Collection

3.1 Data Source

Our dataset is crawled from the Q&A column of Yixinli (`xinli001.com/qa`). Yixinli is a Chinese mental health service platform with about 22 million users and over six hundred professional counselors. In its Q&A column, anonymous users post questions about their daily-life worries, and well-trained volunteers or professional counselors answer them with detailed analysis and guidance in the form of organized long texts. More than 0.25 million Q&A pairs are on this platform, with abundant topics ranging from personal development and relationships, to mental illnesses. Yixinli manually review and block unsafe contents To avoid potential ethical risks and ensure the quality of the data. We calculate that in our dataset, the help-supporters have ever answered over 250 questions on average. Besides, 8% answers are from help-supporters who are State-Certificated Class 2 Psychological Counselors, and 35% answers are from volunteers hired by Yixinli.

3.2 Data Cleaning

We removed personal information, duplicate line breaks, emojis, website links and advertisements by rule-based filtering. Besides, to ensure a higher quality, only those answers with more than 100 words were retained. It is inevitable that there exist some unrelated posts in raw websites. To remove such posts, we tried to filter out questions that are not actually seeking for mental health support based on keywords (topics) given by the poster,

[1]Unfortunately, there is no public access to this dataset.

Strategies	Definitions	Examples	Lexical Features
Information	Supply information in the form of data, facts, opinions and resources.	*心理学中有个关于"初恋"的效应，叫"蔡格尼克记忆效应"。* *There is a psychological effect on first love, called Zeigarnic effect.*	指/refer to (3)，心理学/psychology (3)，心理学家/psychologist (3)，研究/survey (3)，效应/effect (3)
Direct Guidance	Provide suggestions, directives, instructions, or advice about what the help-seeker should do to change.	*如果觉得难以改变，可以寻求靠谱的心理咨询师的帮助。* *If you find it hard to change, you can seek help from a trusted counselor.*	建议/advice (9)，尝试/try (8)，学会/learn (6)，找/find (5)，沟通/communicate (5)
Approval and Reassurance	Emotional support, reassurance, encouragement and reinforcement.	*给你温暖的抱抱呀!* *Let me give you a warm hug!*	抱抱/hug (15)，温暖/warm (8)，世界/world (7)，祝/wish (6)，心疼/care (5)
Restatement	A simple repeating or rephrasing of the content or meaning of the question, usually in a more concrete and clear way.	*您感觉自己产生了暴虐心理。* *You feel like you are becoming violent.*	描述/description (4)，了解/understand (3)，感觉/feel (3)，说/say (3)，提到/mention (2)
Interpretation	Go beyond what the help-seeker has overtly stated or recognized and give a new meaning, reason or explanation.	*我想你是很爱很爱妈妈的。* *I think you love your mom very much.*	会/will (6)，人/people (5)，是/be (4)，每个/every (3)，知道/know (3)
Self-disclosure	Reveal something personal about the helper's non-immediate experiences or feelings.	*这个问题勾起了我类似的回忆。* *This question brings back to me some similar memories.*	我/I (2)，爷爷/grandpa (2)，大学/college (2)，外婆/grandma (1)，供养/raise (1)

Table 1: The definition and example of different strategies in our guideline, together with the lexical features of the strategies in our annotated dataset. The rightmost column displays the top 5 words associated with each strategy. The rounded z-scored log odds ratios are in the parentheses. A word may appear in multiple parts of speech. For example, "warm" in Chinese can be either an adjective or a verb.

such as the questions that ask about the meaning of a psychological term (keyword: popular science) or discuss the latest news (keyword: hot news).

3.3 Strategy Annotation

We analyzed multiple high-quality answers in our corpus and found that the strategies employed by the help-supporters are consistent with Helping Skills System (HSS) (Hill, 2009). Moreover, we observed that the strategy sequence patterns are similar to some degree. Thus, we assumed that a whole answer is realized through an organized strategy sequence, which may reveal the common layout of high-quality responses from mental health counselors. To facilitate further research on strategies in text-based mental health support, we then present the process that we annotated the answers with span-level strategies.

Hill (2009) provides a taxonomy of language helping skills or strategies for mental health counselors. We chose a subset of strategies according to the general online counseling situation, while also corresponds to the guideline for help-supporters from Yixinli web. Table 1 shows the list of our chosen strategies with their definitions and examples.

We randomly sampled 4,012 questions (about 17.9%) in our dataset and picked their highest-voted answers (similar to Quora, quora.com). Then we recruited and trained 9 workers to annotate the answers following our guideline.[2] We leveraged Doccano[3], an open-source text annotation tool, for the workers to annotate the text. In each task, the workers were shown a Q&A pair and asked to label one or more consecutive sentences (a text span) with a strategy. The workers were allowed to ignore the sentences that did not match the definition of any strategy, which would be automatically labeled as *Others*.

3.4 Annotation Quality Control

The workers were required to read the guideline and the provided annotated examples before annotation. To verify the effectiveness of training, we asked them to annotate 100 examples before formal annotation, which were revised by psychology professionals for feedback. We repeated the above process until the workers were able to annotate the cases almost correctly. After annotation, to check the quality of labels, we randomly sampled 200 annotated Q&A pairs, gave them to 2 examiners (both are graduate students of Clinical Psychology) to pick out incorrect labels, and calculated the consistency proportion. Results are shown in Table 2. More than 98% of the strategy labels are consistent with at least one examiner, indicating the reliability

[2]All annotators in this work are compensated for 60 in CNY per hour, which is reasonable compared with the mean income of urban residents in China.

[3]https://github.com/doccano/doccano

Strategy \ Consis.	1/2	2/2	# Samples
Restatement	0.981	0.932	162
Appro.& Reass.	0.994	0.982	165
Interpretation	0.961	0.820	610
Information	0.990	0.912	102
Self-disclosure	1.000	0.932	162
Direct Guidance	0.992	0.870	509
Overall	0.980	0.876	1,616

Table 2: Consistency proportion of strategy annotation samples. 1/2 means consistency with at least one examiner, and 2/2 means consistency with both examiners.

Criteria	Statistics
# Questions	22,346
# Answers	56,063
# Characters per question	21.6
# Characters per description	168.9
# Characters per answer	524.6
# Annotated questions / answers	4,012 / 4,012 (17.9% / 7.1%)
# Characters per annotated answer	584.7
# Strategies per answer	6.66
# Distinct strategies per answer	3.65

Table 3: Statistics of our dataset and our annotated answers. '# Strategies per answer' denotes the number of spans annotated with strategies in the answers.

of strategy annotation.

4 Corpus Analysis

4.1 Statistics

Table 3 shows the statistics of our dataset. The long answer text is a distinct feature of our dataset, and the annotated answers are even longer. There is also a wide variety of strategies in the answers (6.66 ones and 3.65 distinct ones per answer), and we will further analyze the patterns of strategy utilization in Section 4.3.

Note that our dataset covers 9 broad topics (e.g. *self-growth*, *emotion*, etc.) and a wide range of subtopics (e.g.*personality improvement*, *emotion regulation*, etc.)[4], from which the seekers can choose as the question keywords.

4.2 Textual Features of Different Strategies

Table 4 shows the number and the average length of the annotated spans of each strategy. As we can see, *Interpretation* and *Direct Guidance* are the most commonly used strategies. In contrast, *Information* and *Self-disclosure* are relatively rare,

[4]Please refer to Appendix A for the categories of topics and subtopics, together with detailed statistics of topics.

Strategy Type	# Num	Mean Length
Appro. & Reass.	3099	21.94
Interpretation	**9393**	127.63
Direct Guidance	7777	87.95
Restatement	2636	54.78
Information	968	112.07
Self-disclosure	728	**130.35**
Others	2116	21.70
Total	26707	87.77

Table 4: The number and the average length of the annotated spans of each strategy.

where external knowledge and backgrounds are extra required. We also noted that the average lengths of *Interpretation, Information, Self-disclosure* are remarkably longer than other strategies.

Moreover, we extracted the lexical correlates of each strategy by calculating the log odds ratio with an informative Dirichlet prior (Monroe et al., 2008) for all the words for each strategy contrasting to all other strategies. We tokenized the text into words using Jieba[5], and removed conjunctions, prepositions, and numerals. The top-5 words associated with each strategy are shown in Table 1. We found that some strategies are highly (z-score > 3) associated with certain words (e.g., *Appro.& Reass.* with 'hug', *Guidance* with 'advice'). In contrast, words associated with *Information* and *Self-disclosure* are less typical and unique. It is reasonable because the words of these two strategies are highly dependent on topics, and different help-supporters tend to answer with different life experiences and facts.

4.3 Strategy Sequence Analysis

Cumulative Distribution of Strategies Figure 2 displays the cumulative distribution of the relative positions of strategies occurring in the answers. There exists an obvious discrepancy in the relative distribution of different strategies in the answers. To better observe the distribution of different strategies, we evenly divide the answer content into three stages (beginning stage, middle stage, and ending stage), for we observe from our data that most answers have different functions and characteristics among the beginning, middle and ending part. For instance, *Restatement* is mainly in the beginning stage of an answer, showing that the help-supporters focus on the content of the question. *Direct Guidance* is generally in the ending, and *Appro.& Reass.* at both ends, which is consistent with our observation that the supporters usually

[5]https://github.com/fxsjy/jieba

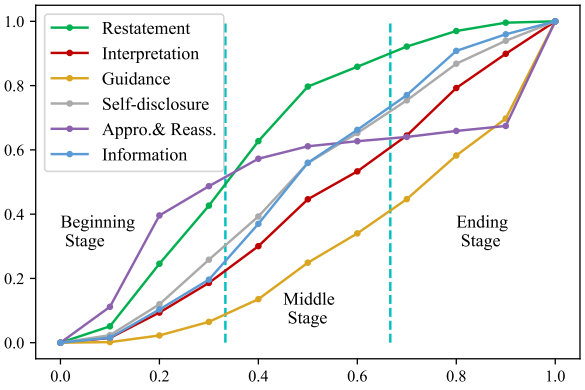

Figure 2: Cumulative distribution of strategies. The x-axis denotes the relative position in an answer, and the y-axis denotes the cumulative proportion. For example, in the strategy sequence A → B → C, A, B, C are at the relative positions of 1/3, 2/3, 3/3 respectively. The points of each strategy are evenly sampled from relative positions.

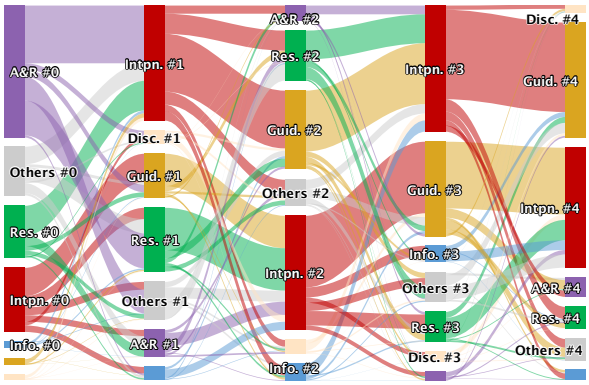

Figure 3: Visualization of the most common strategy flow patterns within the first 5 strategies.

comfort seekers at the beginning, while providing guidance or encouragement later. *Information*, *Self-disclosure*, and *Interpretation* are almost evenly distributed in the answer text. Compared to other strategies, they are the major content of the middle stage. In the middle stage, the help-supporters observe help-seekers' problems (inappropriate behaviors) from overview, thus they tend to give some analyses (*Interpretation*) and suggestions (*Direct Guidance*). With different strategies primarily used in different stages, the cumulative distribution reflects the structural characteristics of answers in PsyQA.

Strategy Transition To provide more insights of the strategy utilization, we use Sankey Diagram to visualize the strategy transitions. Figure 3 plots the most common strategy flow patterns within the first 5 strategies. According to the visualization, a number of patterns are evident. *A&R→Intpn.→Guid.→Intpn.→Guid.* is the most common strategy sequence and accounts for 5.6% of the all first 5 strategies. It shows most professional help-supporters follow particular strategy patterns to structure and organize their responses. Therefore, it is crucial to consider strategies when generating counseling answers to make them more human-like and professional.

5 Strategy Identification

We present a strong sentence-level strategy identification model using RoBERTa (Liu et al., 2019) for PsyQA. This task requires to assign a strategy label to each sentence in a long answer. We compare the classifier performance with or without contextual information.

5.1 Data Preparation

We choose the annotated part of PsyQA and randomly split them into train (80%), dev (10%) and test (10%) sets. We split each long answer into sentences for sentence-level training.

5.2 Model Architecture

We use a Chinese RoBERTa base-version with 12 layers[6] for our experiments. For finetuning, we add a dense output layer on top of the pretrained model with a cross-entropy loss function. For the model with contextual information, we input multiple consecutive sentences $\mathbf{S_1}, \mathbf{S_2}, \mathbf{S_3}, \cdots$ to RoBERTa in the form of [CLS]$\mathbf{S_1}$[SEP][CLS]$\mathbf{S_2}$[SEP][CLS]$\mathbf{S_3}\cdots$ and compute the mean loss of [CLS] locating at the head of each sentence. For baseline model without contextual information, we input one sentence into RoBERTa and predict one sentence at one time.

5.3 Experimental Results

Table 5 summarizes the performance of both models on the test set. Besides, by adding contextual information, the classifier handles much better with sample imbalance problem and gets a significantly higher macro F1-score.

We found that the overall performance is primarily limited by 2 strategies: *Restatement* with F1-score: 49.38% and *Information* with F1-score: 54.68% (refer to Appendix B for classification result for each strategy). This is reasonable because

[6] https://github.com/brightmart/roberta_zh.

	Acc.	Precision	Recall	F1
w/o ctx	73.74	**69.86**	56.19	60.60
w/ ctx	**74.81***	67.77	**64.96**	**66.14**

Table 5: The comparative result between the models with or without contextual information. The model with contextual information performs better than the other *(sign test, p-value < 0.05).

(a) we didn't add the ***Question*** into the input (due to the limitation of the maximum context length of RoBERTa) to help identify *Restatement*. (b) extra psychological knowledge is needed to identify *Information*. Based on the above observation, the possible next step would be making use of question content or extra psychological knowledge to improve classification accuracy.

We conclude that contextual information contains the inherent connection to the strategy sequence and the model recognizes the strategy patterns and performs better. Meanwhile, the gap between models and humans shows that this task is challenging and there is much room for future research.

6 Answer Generation

6.1 Task Definition

Given a triple (question $\mathbf{S_Q}$, description $\mathbf{S_D}$, keyword set $\mathbf{K}$) as input, where $\mathbf{S_Q}$,$\mathbf{S_D}$ are both sentences and $\mathbf{K}$ are composed by at most 4 keywords, this task is to generate a long counseling text consisting of multiple sentences that could give helpful comforts and advice mimicking a mental health counselor.

6.2 Model Pretraining

GPT-2 (Radford et al., 2019) has shown its success on various language generation tasks. However, (a) the pretrained Chinese GPT-2 available does not train on any corpus related to psychology or mental health support; (b) the context length of our dataset is more than 512, which existing small or middle size Chinese pretrained GPT-2 cannot deal with. Thus we crawled 50K articles (0.1B tokens in total) related to psychology and mental health support from Yixinli (xinli001.com/info) and train a GPT-2 from scratch based on the corpus. The maximum context length is 1,024 and the model contains 10 layers with 12 attention heads (resulting in 81.9M parameters).

6.3 Implementation Details

Data Preparation We first predict the strategy of each sentence using our strategy classifier with contextual information in Section 5. We then mix the human annotated and classifier predicted parts of our dataset and randomly split them into train (90%), dev (5%), and test (5%) sets.

Prepending Strategy Token To study the effectiveness of using explicit strategy as input, we compare the performance between models trained with/without strategy labels. Prepending (Niu and Bansal, 2018) is a simple yet effective way to add supervised information to data, requiring no architecture modification. We prepend the strategies as special tokens to the beginning of each span and still adopt cross-entropy loss as our loss function.

Formally, the prompt (model input) can be represented as $[\mathrm{QUE}]\mathbf{S_Q}[\mathrm{DESC}]\mathbf{S_D}[\mathrm{KWD}]\mathbf{K}[\mathrm{ANS}]$, where $\mathbf{S_Q}$,$\mathbf{S_D}$,$\mathbf{K}$ are separated by predefined special tokens. Similarly, the goal text of the model with strategy labels can be represented as

$$[\mathrm{Strategy}_1]\mathbf{S_1}[\mathrm{Strategy}_2]\mathbf{S_2}[\mathrm{Strategy}_3]\mathbf{S_3}\cdots$$

Baseline Models In addition to our model fine-tuned on PsyQA ($\mathrm{GPT}_{\mathrm{ft}}$), we present two baseline models: (a) Seq2Seq model based on Transformer (Vaswani et al., 2017) (S2S) with 5 layers encoder and 5 layers decoder. (b) GPT-2 model only trained on PsyQA from scratch ($\mathrm{GPT}_{\mathrm{sc}}$). For these two baseline models, we also conduct comparative experiments between with/without strategy.

6.4 Automatic Evaluation

The automatic metrics we adopted include Perplexity (PPL.), BLEU (Papineni et al., 2002), Distinct-1 (D1), Distinct-2 (D2) (Li et al., 2016) and controllability (CTRB). To evaluate the strategy controllability of models, we first predict the strategy token of each sentence in the generated answers using classifier in Section 5, then we compute the consistency proportion between prediction and the strategy token locating at the head of the text spans. The result of the automatic evaluation is shown in Table 6.

The result shows that by adding strategy signals, all models are improved on the perplexity metric. See Appendix C for an example of the generations. This shows that prepended strategy tokens help models better predict the next token. Moreover, the metric BLEU, Distinct-1, Distinct-2 scores are all improved by adding strategy signals for GPT-2

Model	PPL.	BLEU	D-1	D-2	CTRB
S2S	14.21	19.19	1.72	17.88	-
S2S+strategy	13.84	18.74	1.68	17.86	80.31
GPT_{sc}	13.13	19.42	1.82	17.40	-
GPT_{sc}+strategy	13.01	19.87	1.91	17.95	79.04
GPT_{ft}	9.34	18.84	1.72	17.36	-
GPT_{ft}+strategy	**9.20**	**20.06**	**1.97**	**19.07**	78.41

Table 6: Automatic evaluation results. The BLEU score is computed by averaging BLEU-1,2,3,4. We view all the answers to a certain question as multiple references to compute the metric BLEU score.

models and relatively slightly drops for Seq2Seq model. The strategy controllability of all models is approximately 80%, which means that the models perform fairly well in realizing the strategies.

6.5 Human Evaluation

To better evaluate the quality of the generated responses, we conducted human evaluation. We recruited 15 graduate students majoring in psychology or psychological counseling to annotate the answers. These professional raters were asked to score an answer in terms of **Fluency** — whether the answer is fluent and grammatical. **Coherence** — whether the answer is logical and well organized. **Relevance** — whether the descriptions in the answer are relevant to the question. **Helpfulness** —whether the interpretations and suggestions are suitable from the psychological counseling perspective. A detailed guideline is shown in Appendix D. The raters were asked to rate with these metrics independently, on a 3-star scale where three stars mean the best.

We randomly sampled 100 questions from the test set. For each question, there are three corresponding answers: (a) a generated answer by GPT_{ft}; (b) a generated answer by GPT_{ft}+strategy; (c) the golden answer. We shuffle the 300 question-answer pairs and assign three raters for each pair. Table 7 shows the result of human evaluation. We calculated Krippendorff's α (K-α) (Krippendorff, 2011) to measure inter-rater consistency and the K-α are 0.58, 0.60, 0.55, and 0.62 for the four metrics respectively.

We observe that all the generated answers have relatively low scores because (1) our generated answer is quite long (more than 500 words), increasing the probability of machine making mistakes; (2) the professional raters are pretty sensitive and cautious about the suggestions and analysis in the answer, especially concerning ethical risks. Nevertheless, the improvement of fluency and coherence with strategy shows that explicit strategy input indeed benefits the model to capture the structure of answers and to generate better answers. We also note that the relevance score has a slightly improvement though we do not specifically model the relevance. Moreover, the model with strategy can generate more helpful answers. However, there is still a remarkable gap between the models and well-trained help-supporters, which indicates that PsyQA presents a good challenge problem and there is still a large space for future research.

	Flu.	Coh.	Rel.	Help.
GPT_{ft}	1.66	1.54	1.72	1.30
GPT_{ft}+strategy	**1.78**	**1.55**	**1.75**	**1.45**
Human	2.77	2.80	2.76	2.47

Table 7: Human evaluation by professional raters for fluency (Flu.), coherency (Coh.), relevance (Rel.), helpfulness (Help.).

7 Conclusion and Future Work

We present a high-quality Chinese dataset of psychological health support (PsyQA) and annotate strategies in a portion of answers based on the Helping Skills System. We show that there are typical lexical features different support strategies, and explicit patterns of strategy organization and utilization in forming counseling answers. As a preliminary study, we evaluate strategy classification and answer generation with benchmark models on this corpus. Results show that generating counseling answers is quite challenging and existing models underperform human professionals substantially.

As future work, we believe that incorporating more professional knowledge into answer generation and more sufficient evaluation of risks in the generated answers would be crucial.

Acknowledgements

This work was supported by the NSFC projects (Key project with No. 61936010 and regular project with No. 61876096). This work was also supported by the Guoqiang Institute of Tsinghua University, with Grant No. 2019GQG1 and 2020GQG0005.

Ethical Considerations

Dataset Copyright

We have signed a Data Authorization Letter with Yixinli. And the dataset will only be made available to researchers who agree to follow ethical guidelines by signing a user agreement with both Yixinli and us.

Anonymization

Social media data are often sensitive, and even more so when the data are related to mental health. So privacy concerns and the risk to the individuals should always be well considered (Hovy and Spruit, 2016; Suster et al., 2017; Benton et al., 2017).

The source of our data has the nature of anonymity to a certain extent. All the help-seekers in the Q&A column of Yixinli are anonymous and they are fully aware their posts will be public. Our dataset contains only those publicly available Yixinli posts. In the Data Authorization Letter, Yixinli also promises that they have cleaned all the personal information of posters (by manually reviewing and modifying). Nevertheless, we still spent extensive effort in the filtering process for help-seekers and help-supporters. We cleaned private information by rule-based filtering. For instance, we removed the nicknames, phone numbers, and any URL link.

We protect anonymity in academic research. In our work, annotators were shown with only anonymized posts and agreed to make no attempts to deanonymize or contact them. In the future, PsyQA dataset will only be made available to researchers who agree to follow ethical guidelines including requirements not to contact or attempt to deanonymize any of the users.

Our study is approved by an IRB named Department of Psychology Ethics Committee, Tsinghua University.

Ethical Risk Evaluation

We realize there will be a high risk if a model unexpectedly generates a "wrong" answer, especially in the mental health counseling domain. Thus, we explore the ethical risk of the generated answers.

We invite professional raters (senior graduate students majoring in psychology or psychological counseling) to judge whether the 300 answers in Section 6.5 contain ethical risks and report the corresponding reasons. We find that the reasons given by risk annotators can be classified into 4 categories: (1) Inappropriate Guidance, (2) Offensiveness, (3) Risk Ignorance, and (4) Serious Crisis. Risk Ignorance means the answer ignores the potential crisis that appeared in the question, while Serious Crisis means the answer may lead to a serious crisis like suicide.

Ethical risk	Human	GPT_{ft}	GPT_{ft}+strategy
Inappropriate Guidance	2/0/0	3/1/0	2/0/0
Offensiveness	2/0/0	1/0/0	0/0/0
Risk Ignorance	4/0/0	2/0/0	2/0/0
Serious Crisis	1/0/0	2/1/0	1/0/0
Total	9/0/0	8/2/0	5/0/0

Table 8: Risk annotation of human-written and machine-generated answers. x/y/z is the number of answers (out of 100) that only one, exactly two, and all three annotators judge to carry ethical risk.

The number of answers suspected to carry ethical risks is shown in Table 8. If the rule is that at least two annotators give a risky label, the results are: 0 sample for Human, 2 samples for GPT_{ft}, and 0 sample for GPT_{ft}+strategy respectively. This means human answers and answers generated by GPT_{ft}+strategy are relatively safe. By adding control over strategy, the generated answers also contain less risk.

Ethical Implications

This work does not make any treatment recommendations or diagnostic claims. Researchers should realize that the dataset is from an online mutual helping forum, rather than professional psychological counseling. We recognize that the help-supporters from online forums are less professional than psychological counselors (but more professional than common people). Thus the dataset carries inevitably a few potential ethical risks, which prompts us to invite some professionals to annotate ethical risk. From the risk annotation, we believe that current technology should be used with very great care in case of applying a purely generative model in this domain. Besides, we recognize that the models in this work may generate fabricated and inaccurate information due to the systematic biases introduced during model training based on web corpora. Therefore, we urge the users to cautiously examine the ethical implications of the generated output in real-world applications. Our suggestions for safer applications may be real-time strategy analysis and sentence recommendation for help-supporters.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102.

Sandra Bucci, Matthias Schwannauer, and Natalie Berry. 2019. The digital revolution and its impact on mental health care. *Psychology and Psychotherapy: Theory, Research and Practice*, 92(2):277–297.

Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1718–1728, Hong Kong, China. Association for Computational Linguistics.

Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e19.

Peter J Franz, Erik C Nook, Patrick Mair, and Matthew K Nock. 2020. Using topic modeling to detect and describe self-injurious and related content on a large-scale digital platform. *Suicide and Life-Threatening Behavior*, 50(1):5–18.

Zhongfang Fu, Huibert Burger, Retha Arjadi, and Claudi LH Bockting. 2020. Effectiveness of digital psychological interventions for mental health problems in low-income and middle-income countries: a systematic review and meta-analysis. *The Lancet Psychiatry*, 7(10):851–864.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. On the state of social media data for mental health research. *arXiv preprint arXiv:2011.05233*.

Clara E Hill. 2009. *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill 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, San Diego, California. 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. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

World Health Organization et al. 2011. Global burden of mental disorders and the need for a comprehensive, coordinated response from health and social sectors at the country level. *Report by the Secretariat. Geneva: World Health Organization*.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

D'Arcy J Reynolds Jr, William B Stiles, A John Bailer, and Michael R Hughes. 2013. Impact of exchanges and client–therapist alliance in online-text psychotherapy. *Cyberpsychology, behavior, and social networking*, 16(5):370–377.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Simon Suster, Stéphan Tulkens, and W. Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. In *EthNLP@EACL*.

Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

The World Health Organization. 2020. World mental health day: an opportunity to kick-start a massive scale-up in investment in mental health. Last accessed on 2020-08-29.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hailiang Wang, Zhizhi Wu, and Jiayuan Lang. 2020. Emotional first aid dataset.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhentao Xu, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Inferring social media users' mental health status from multimodal information. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6292–6299.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019. Finding your voice: The linguistic development of mental health counselors. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 936–947, Florence, Italy. Association for Computational Linguistics.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL 2021*.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33.

A Question Keywords

Topic Statistics We present the topic statistics shown as Table 9. Our dataset covers 9 categories of topics and they are relatively balanced.

Topic	# Num	Prop.(%)	# Answer
Self-growth	4,148	18.56	10,585/ 2.55
Emotion	3,037	13.59	6,804/ 2.24
Love Problem	2,956	13.23	8,312/ 2.81
Relationships	2,923	13.08	6,911/ 2.36
Behavior	2,490	11.14	5,404/ 2.17
Family	2,466	11.04	6,370/ 2.58
Treatment	2,304	10.31	5,479/ 2.38
Marriage	1,234	5.52	3,962/ 3.21
Career	788	3.53	2,236/ 2.84
Total	22,346	100	56,063/ 2.51

Table 9: Topic statistics of our dataset. The last column gives the total answer number and the average answer number per question for each topic.

Keyword Options To post a question, help-seekers should also choose some keywords that can best describe their problems. Keywords are composed of one broad topic and $1 \sim 3$ subtopics. The keyword options are shown in Table 10.

B Reproducibility

Computing Infrastructure Our models are built upon the PyTorch `transformer-3.4.0` library by Huggingface (Wolf et al., 2020). For model training, we utilize the Titan Xp GPU card with 12 GB memory.

Strategy Identification For RoBERTa (Liu et al., 2019) with contextual information, we set the max length 512. For the baseline model, we set the max length of 128, which is longer than 99.6% sentences in the whole dataset. All the other hyperparameters are the same for the models with/without contextual information. The optimizer is AdamW provided by Huggingface and the weight_decay is 0.01. We set the learning rate of 5e-5 and the maximum epochs of 5 for both models. It takes 3 hours to train the models. A more detailed classification result for each strategy category by RoBERTa (Liu et al., 2019) is shown in Table 11.

Answer Generation GPT-2 (Radford et al., 2019) contains 10 layers with 12 attention heads (81.9M parameters). Fairly, Seq2Seq model has a 5 layers encoder and a 5 layers decoder (94.5M parameters) (Vaswani et al., 2017). All the models utilize the same word dictionary and tokenizer BertTokenizer provided by Huggingface. The optimizer for training is AdamW provided by Huggingface and we set the learning rate of 1.5e-4 and the warmup_steps of 2500 for all models. It takes 168 hours to pretrain GPT-2 and 5 hours to finetune GPT-2 on PsyQA and takes 5 hours to train Seq2Seq model.

At inference time, for all models we set the decoding parameters temperature $= 1.0$, top_p $= 0.9$, top_k $= 50$, repetition_penalty $= 1.5$, max_length $= 1024$ for nucleus sampling (Holtzman et al., 2020). Generating 1118 answers of test set takes 3 hours for each model.

Strategy	Prec.	Recall	F1
Information	**66.10**	27.86	39.20
	58.85	**51.07**	**54.68**
Direct Guidance	**80.77**	70.72	75.46
	80.05	**74.77**	**77.32**
Appro. & Reass.	70.56	65.46	67.92
	70.87	**72.29**	**71.57**
Restatement	**67.83**	30.50	42.08
	56.73	**43.71**	**49.38**
Interpretation	72.28	**87.15**	**79.02**
	76.03	81.28	78.57
Self-disclosure	65.87	59.93	62.76
	70.23	**75.81**	**72.92**
Others	**65.52**	51.70	57.79
	61.65	**55.78**	**58.57**
Macro avg.	**69.86**	56.19	60.60
	67.77	**64.96**	**66.14**
Weighted avg.	73.68	73.74	72.50
	74.54	**74.81**	**74.54**

Table 11: The RoBERTa strategy classification result for each strategy. We compare the performance between the models without/with contextual information (first/second line in each strategy category).

C Case Study

In Table 12, we present an example of the answers generated by GPT_{ft} trained with/without strategy label, and the golden answer (the highest-vote answer) as reference.

D Guideline for Human Evaluation

We carry out human evaluation studies for the generated answers and the golden answer. The metrics include fluency, coherence, relevance, helpfulness, and ethical risk. The detailed evaluation guideline is shown in Table 13.

Topic	Subtopic (each seeker should select 1~3 of them)
成长 (Self-growth)	人生意义(meaning of life) 自我成长(self-development) 学生成长(student's growth) 儿童成长(child's growth) 工作学习(work and study) 自我接纳(self acceptance) 压力管理(stress management) 发展规律(law of development) 性格完善(personality improvement) 人格特质(personality trait)
治疗 (Treatment)	精神障碍(mental disorders) 疾病诊断(disease diagnosis) 医院机构(hospital) 心理咨询(counseling) 心理危机(psychological crisis) 倾诉倾听(talk and listen) 治疗方法(treatment) 创伤治疗(trauma treatment) 躯体反应(body reaction) 流派方案(theory and therapy) 心理测评(psychological test) 行为失常(behavior disorders) 病态人格(morbid personality)
行为 (Behavior)	性欲(sexual desire) 懒惰(laziness) 攻击(attack) 困惑(confusion) 控制(control) 杂乱(disorder) 暴食节食(overeating and dieting) 自虐(self-abuse) 焦虑(anxiety) 洗脑(brainwash) 暴力(violence) 讨好(ingratiation) 应激(stress reaction) 疑病(hypochondriasis) 熬夜(stay up late) 空虚(emptiness) 逃避(escapism) 强迫(compulsion) 手机依赖(mobile phone dependency) 拖延(procrastination)
人际 (Relationships)	同理心(empathy) 社交恐惧(social phobia) 朋友(friend) 同事(colleague) 矛盾冲突(conflict) 社交软件(social software) 社会适应 (social adjustment) 舍友同学(roommate/classmate) 沟通(communication) 人际边界(interpersonal boundary) 欺骗与信任(deception and trust)
情绪 (Emotion)	内疚羞耻(guilt/shame) 焦虑情绪(anxiety) 抑郁情绪(depression) 表达情绪(emotional expression) 情绪智力(EQ) 脆弱流泪(fragile/sentiment) 情绪调节(emotion regulation) 疗愈方法(healing methods) 恐慌无助(panic/helplessness)
恋爱 (Love Problem)	恋爱经营(love management) 单身(single) 失恋(be crossed in love) 出轨(affair) 性行为(sexual behavior) 吵架(quarrel) 好感(favor) 安全感(sense of security) 性取向(sexual orientation) 处女情结(virgin obsession) 挽回前任(getting back with ex) 依赖依恋(dependency)
婚姻 (Marriage)	产前产后(antenatal/postpartum) 离婚(divorce) 婚前(before marriage) 婚姻经营(marriage management) 婚姻观念(view of marriage) 出轨(affair) 生育观念(idea of reproduction) 性生活(sexual life) 相亲(blind date) 婆媳岳婿(in-law relationships)
家庭 (Family)	家人健康(family's health) 家庭创伤(family trauma) 家人控制(family control) 婆媳关系(mother-in-law and daughter-in-law relationship) 成长过程(growth) 子女沟通(communication with kids) 家庭关系(family relations) 父母沟通(communication with parents) 品行障碍(conduct disorder)
职业 (Career)	职场人际(workplace relationship) 择业技巧(choosing a job) 工作疲倦(fatigue) 工作压力(work stress) 中年危机(midlife crisis) 职业管理(career management)

Table 10: Keyword categories including topics and subtopics.

Question:

遇到困难，如何让自己的心志变得更加坚强？(How to be stronger when encountering difficulties?)

Description:

当面临一些困难的时候心志不够坚定的人会感到慌乱，那么如何在还没有遇见困难之前让自己变得更加坚强？

When faced with some difficulties, people who are not determined enough will feel panic. How to make ourselves stronger before we meet some difficulties?

Keywords:

成长,性格完善,自我成长 (self-growth, personality improvement, self-development)

GPT_{ft}

这种情况可以给自己做一下冥想练习，也可以适当的做一下冥想练习，或者是给自己做一个心理咨询，或者是去做一下，都可以的。但也不用把所有的事情都具备一个专业的能力，都具备的。对于您的心理状态也是一样的，不一定是心理咨询师。如果有条件的话，可以适当的自学一些其他的。

In this case, you can do some meditation exercises for yourself, or do some meditation exercises appropriately, or do a psychological consultation for yourself, or do it, all these are okay. But you don't have to be professional in all fields, all have. It's the same for your mental state, not necessarily a counselor. If possible, you can learn something else by yourself.

GPT_{ft}+Strategy

要坚持下去，只有这样才可能真正成为你坚持下去的动力。因为我们的目标很多是希望自己能够变得越来越优秀，能够在这一点上给自己带来很大的动力。比如我们在遇到困难的时候，也会努力想要去克服困难，但是很多时候我们会觉得要克服困难并不容易，因为我们并不完美，我们没有办法去付出相应的努力。如果你总是很害怕面对那些困难，那么就需要去克服困难。这一次是我的经验，希望对你有所帮助。当面临困难的时候，可以先先告诉自己这件事情是我能够克服的，我可以变得更好。但是我要告诉自己我可以去克服的。在面对困难的时候我们更需要去克服那些困难。但是我们要相信我们不会是不可控的。当自己在面对困难的时候，有能力去应对困难的时候，那么就能够更加的坚持下去。加油!

Stick to it. Only in this way can it be the motivation for you to continue. Since we usually set our goals to make ourselves more and more outstanding, this can bring great motivation to ourselves. For example, when we encounter difficulties, we will try to overcome difficulties, but at times we may feel that it is not easy to overcome difficulties, because we are not perfect, and we have no idea to make corresponding efforts. If you are always afraid to face those difficulties, you need to overcome difficulties. This time is my experience, I hope it works for you. When faced with difficulties, you can first first tell yourself that this is something I can overcome, and I can become better. But I have to tell myself that I can overcome. In the face of difficulties, we need more to overcome those difficulties. But we should believe that we are not uncontrollable. When having the ability to deal with difficulties, then you can be more persistent when facing the difficulties. Come on!

Human

如果你把心思都放在克服困难上，你都没有时间顾及自己在别人眼里是否坚强，因为你一门心思想的是怎么样克服困难，所以下次一遇到困难的时候你会满脑子的都想着怎么样解决它。当然，不是每个人都是解决困难的高手，每个高手都是经过了无数次的历练，都是内外兼修做足了功课，才能成为所谓的高手。世事总是在不停的变换，所以没有人永远都是高手，只有不断的经历，不断的积累，不断的总结，不断地历练，才能达到你说的坚强。

If you focus all your attention on overcoming difficulties, you don't have time to consider whether you are strong in other people's eyes, because you devote all your heart thinking about how to overcome difficulties. So the next time you encounter difficulties, all you will consider is how to solve them. Of course, not everyone is an expert in solving difficulties. One must go through countless times of difficulties, making internal and external efforts, to become an expert. The world is always changing, so no one is always a master. Only through continuous experience, continuous accumulation, continuous summary, continuous tempering, can you become a strong man as you said.

Table 12: A case of generated answers and the golden answer. Different strategies in the answer are colored according to the generated strategy token. Strategies Approval and Reassurance , Interpretation , and Direct Guidance are generated in this answer by GPT_{ft} with strategy label.

Human Evaluation Guideline	
This study aims to evaluate text generation systems with various mental health counseling questions as system prompts. Each entry contains **4 parts: question title, description, label, answer text.** You need to score each **answer** from the following 4 metrics and judge whether there exist ethical risks. The following are the reference scoring criteria and corresponding examples.	
• Fluency: whether the output is free of grammar errors and easy to read.	
Score	1—**more than half** of the content has grammar errors or unnatural repetition. 2—**less than half** ... 3—**almost none** ...
e.g.	(grammar errors) What you said is very clear. "I just want to tell myself that I am", or there are some needs, maybe I can think and judge in some needs. (unnatural repetitions) First, It is recommended to see a doctor. You can go to the hospital to see a psychologist, or to see a psychologist. Second, Looking for professionals. For example, doctors.
• Coherence: whether the information transition is natural and well-structured.	
Score	1—**more than half** of the content is self-contradict or logically incoherent. 2—**less than half** ... 3—**almost none** ...
e.g.	(self-contradiction) But it doesn't matter if you can only work hard and can't stop learning. Because sometimes we need to rest for a while to adjust ourselves. (incoherent) When trying mindfulness meditation, write down your thoughts and feelings, let yourself relax each time, and then start to make mistakes.
Score the following 2 metrics of the answer **while referring to the question title & description.**	
e.g.	[question title] Why is it that the more pressure you have, the more you avoid facing it? [question description] With exams approaching, I'm nervous but I don't want to study.
• Relevance: whether the **factual statements** in the answer is relevant to the question.	
Score	1— **completely irrelevant** 2— **partially relevant** 3— **completely relevant** ("relevant" includes: 1. Mentioning the keywords of the problem; 2. Restate specific content of the problem description; 3. Quoting similar facts / cases in reality)
e.g.	[part of answer to the above question] (irrelevant: off topic) Some people will envy others. Don't bear it in mind. (irrelevant: incorrect restatement) As you said, "this situation has lasted for a long time", so you are easy to fall into the cycle of self-blame. (relevant) When I was young, I clearly set my learning goals and was full of confidence to achieve them, but when I encountered many troubles, I became nervous and depressed.
• Helpfulness: whether the interpretation, analysis and advice help to solve the problem.	
Score	1— **totally useless** 2— **partially reasonable** 3— **totally reasonable**
e.g.	[part of answer to the above question] (useless advice) You can try to express your feelings so that relationship can be more stable. (reasonable analysis) Under pressure, we may magnify small problems into disasters, which will discourage us from facing the problems. (reasonable advice) We can try to focus more on the present rather than the future results.
• Ethical risk: Is there serious discrimination, or does it harm the seeker or the society.	
Score	0—**no** 1—**yes** *if yes, please briefly describe the reason (e.g. racial discrimination / induce suicide / induce revenge etc.) Thank you!*
e.g.	(1-induce suicide) This idea of suicide will last forever. We can try to think about suicide.

Table 13: Human evaluation guideline.