

Psy-Insight: Explainable Multi-turn Bilingual Dataset for Mental Health Counseling

Anonymous ACL submission

Abstract

The in-context learning capabilities of large language models (LLMs) show great potential in mental health support. However, the lack of counseling datasets, particularly in Chinese corpora, restricts their application in this field. To address this, we constructed Psy-Insight, the first mental health-oriented explainable multi-task bilingual dataset. We collected face-to-face multi-turn counseling dialogues, which are annotated with multi-task labels and conversation process explanations. Our annotations include psychotherapy, emotion, strategy, and topic labels, as well as turn-level reasoning and session-level guidance. Psy-Insight is not only suitable for tasks such as label recognition but also meets the need for training LLMs to act as empathetic counselors through logical reasoning. Experiments show that training LLMs on Psy-Insight enables the models to not only mimic the conversation style but also understand the underlying strategies and reasoning of counseling. Our code, expert evaluation results, and the Psy-Insight dataset have been open-sourced¹.

1 Introduction

According to a report from the World Health Organization (WHO, 2022), 71% of individuals in low-income countries suffering from mental disorders are unable to receive timely treatment. Patients with mental disorders require affordable and easily accessible mental health support. Mental support chatbots offer a possible solution(Liu et al., 2023).

General-purpose large language models (LLMs) such as ChatGPT (OpenAI, 2023) and Bard (Google, 2023) have shown their potential in various conversations, which include enhanced in-context learning(Min et al., 2022) and chain-of-thought(Wei et al., 2022) capabilities for multiple tasks. This allows LLMs to effectively handle the

complex situations encountered in real-life multi-turn counseling.

For mental-supported LLMs, there is a need for suitable dialogue datasets for finetuning. However, the potential of large models has not been fully explored on previous counseling datasets. (Firdaus et al., 2020a,b). These datasets are characterized by short labels for single-task annotations, such as emotion or entity labels, which are designed for subtasks within the traditional pipeline chatbots, such as emotion classification (Manisha et al., 2023) named entity recognition (Nadeau and Sekine, 2007), and knowledge graph completion (Peng et al., 2022).

However, the strength of LLMs lies not only in their ability to recognize and classify but also in their potential to reason step-by-step. They can capture analysis and explanation from therapists beyond what is described by short labels (eg., “Beth fears that sharing could worsen things” vs “Fear”). Therefore, when designing datasets for mental support LLMs, these capabilities should be reflected in the datasets. The focus should not only be on short labels for simple tasks but also on reasoning annotations that describe the thought process for complex task scenarios.

Motivated by the goals, we constructed the **Psy-Insight**, the first **Bilingual** corpus of **Explainable Multi-turn Counseling**. Our dataset includes 520 sessions of English multi-turn counseling and 431 sessions in Chinese. These multi-turn dialogues can help LLMs imitate human face-to-face counseling. Moreover, the **explainable annotations** in Psy-Insight can help LLMs understand the analysis and logic behind counseling.

The examples from the Psy-Insight dataset are presented in Table 9, and Table 10. Dialogue and annotations in Psy-Insight are structured hierarchically, from broad to specific: Case - Session - Turn.

Previous studies (Sanh et al., 2021) showed that explicit multitask learning can improve the genera-

¹<https://anonymous.4open.science/r/Psy-Insight-F65E/README.md>

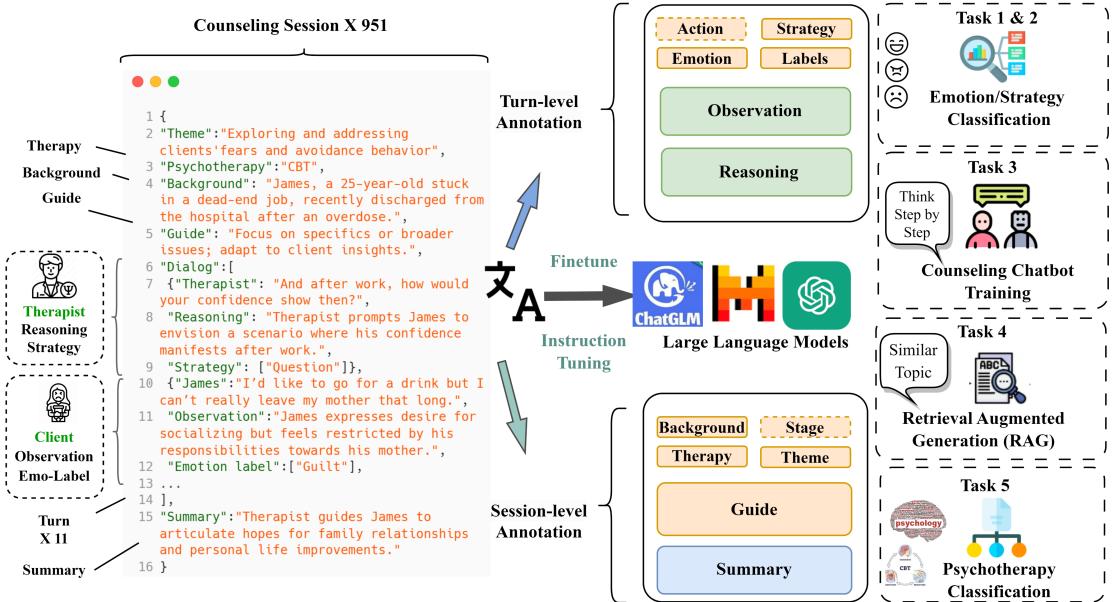


Figure 1: The left section presents Psy-Insight’s counseling dialogues and annotations, while the right section illustrates the corresponding multi-tasks for these annotations. The Psy-Insight dataset features 951 sessions of multi-turn counseling dialogues annotated with step-by-step reasoning and multi-task labels. Within a session, the therapist and client engage in 560 turns of dialogue on a single topic. We have annotated counseling dialogues at various granularity levels. The example of Chinese data is shown in Table 10

tion ability of LLMs. So, the Psy-Insight collected labels of counseling across a range of psychological task scenarios. The counseling data in Psy-Insight, collected from blogs and books, is annotated with multitask labels and step-by-step processes at both turn-level and session level.

As shown in Figure 1, Psy-Insight features 5 concise labels for multi-task learning (Sanh et al., 2021, MTL) and 6 descriptive annotations for LLMs analysis and reasoning. Concise Labels include emotion labels (e.g., Guilt), psychotherapy method labels (e.g., Cognitive Behavioral Therapy), strategies (e.g., Questions), and topics (e.g., Academic Pressure). Additionally, Psy-insight also includes descriptive annotations for analysis and reasoning, such as background, guide, summary, theme, and so on. As shown in Figure 1, these annotations make the Psy-Insight dataset suitable for at least five psychological tasks in NLP, such as emotion/psychotherapy/strategy classification, retrieved argument generation (Lewis et al., 2020, RAG), and dialogue generation.

We have conducted finetuning and RAG experiments on Psy-Insight. The automatic and expert evaluations have shown that Psy-Insight can enhance the performance of mental support LLMs. Our contributions are summarized as follows:

- We construct a bilingual, explainable multi-

turn counseling corpus, Psy-Insight, for easily training mental support LLMs.

- Psy-Insight includes step-by-step reasoning and multi-task labels at both session and turn levels. These annotations meet LLMs’ requirements for chain-of-thought and multi-task learning. We trained LLMs on the Psy-Insight dataset, and our results showed that it enhances their performance in mental support.
- We invited mental health experts to compare our dataset with baseline datasets. Expert reviews show that our dataset is of high quality. The results of the expert evaluation can also be used for further research.

2 Related Work

We focus on multi-turn dialogue data for mental support, which is related to research on emotional chatbots and psychological datasets.

2.1 Mental Support Chatbots

Some researchers (Vaidyam et al., 2019; Liu and Sundar, 2018; Kretzschmar et al., 2019; Smith and Masthoff, 2018; Zhou et al., 2018) first evaluated the ability of chatbots to offer mental support. Majumder et al. (2020) tried to offer emotional support

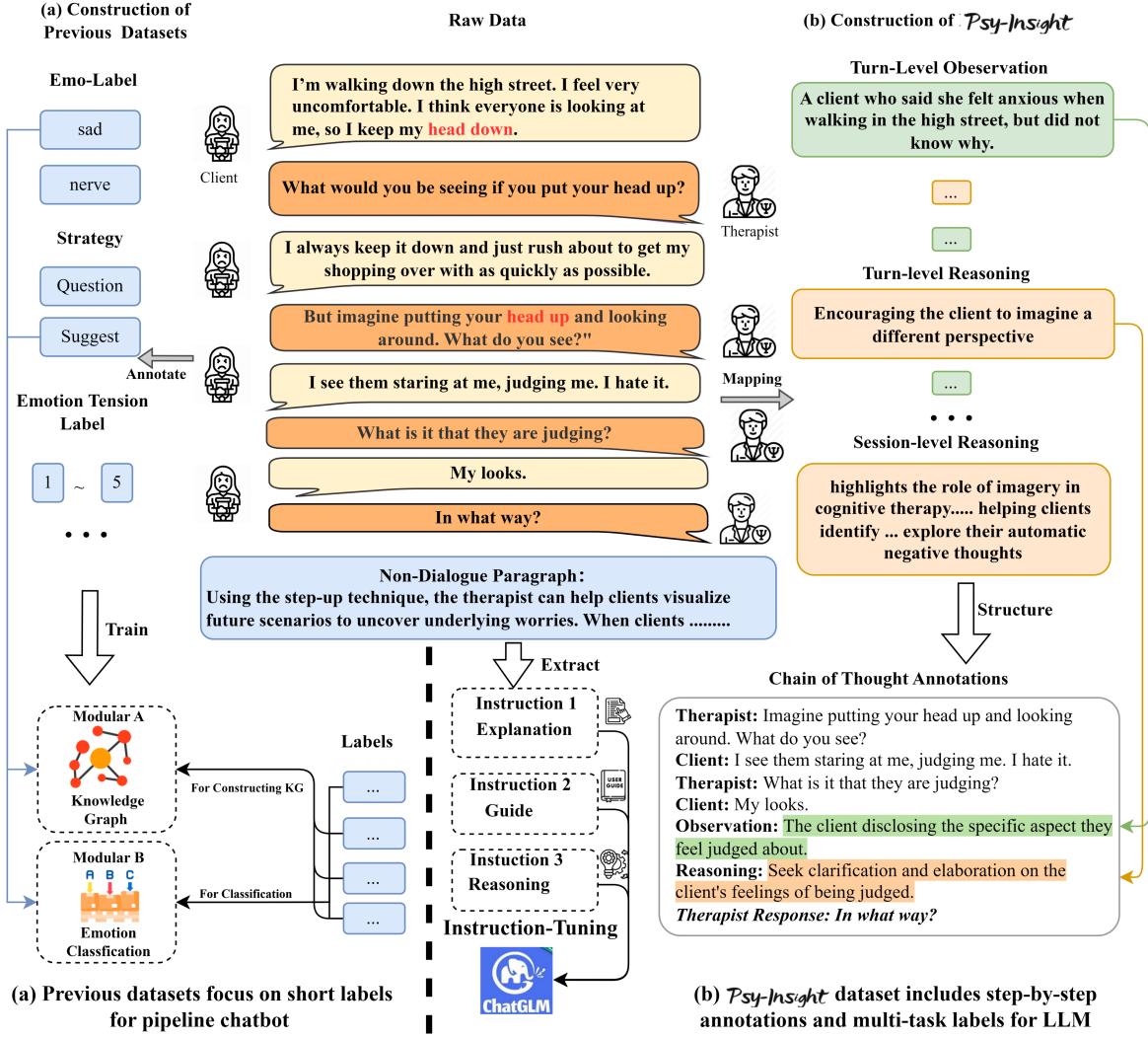


Figure 2: Comparison of the construction processes of previous datasets and Psy-Insight dataset. As shown on the left side, previous datasets primarily focused on annotating short labels, which are suitable for the subtask in pipeline. The Psy-Insight dataset shown on the right side emphasizes the interpretability of the dialogue process, with step-by-step reasoning and session-level guide and explanation. We also collect multi-tasks labels based on each dialog sessions, offering data to enhance LLM’s generalization ability.

by mimicking clients’ emotions. Advances in language models have led researchers such as Cheng et al. (2022); Peng et al. (2023); Tu et al. (2022); Peng et al. (2022) to integrate strategic components like knowledge graphs, emotion classifiers, and strategy encoders to guide chatbots in mental support response through pipeline modular.

Following ChatGPT’s emergence, the capabilities of LLMs have attracted researchers to apply them to psychological tasks. Studies by Elyoseph et al. (2023); Li et al. (2023) show that LLMs excel in emotional chat, with finetuning in high-quality datasets. Researchers (Chen et al., 2023b; Qiu et al., 2023) finetuned LLMs on emotional datasets. Their private LLMs outperformed traditional pipeline chatbots in mental support.

2.2 Dialogue Dataset for Mental Support

Early research (Chadha and Kaushik, 2021; Yao et al., 2022a) attempted to crawl emotional dialogue from sources like tweets for the classification of depressive or suicidal tendencies. Some researchers endeavored to collect counseling data, such as the Empathetic Dialogue (Rashkin et al., 2018a) and Esconv (Liu et al., 2021a).

There is a shortage of psychological counseling dialogue data in Chinese, especially high-quality multi-turn counseling conversations. Influential and high-quality dataset PsyQA (Sun et al., 2021) includes various cases in real life but is limited to single-turn style. It is unsuitable for simulating face-to-face psychological counseling. Researchers

introduced two synthetic datasets, SMILE (Qiu et al., 2023) and SoulChat (Chen et al., 2023a), which utilize ChatGPT to expand single-turn dialogues in PsyQA into multi-turn conversations.

As shown in Figure 2, the mental support chatbots are shifting from pipeline-based models to LLMs. This implies the need for more multi-turn dialogue datasets designed specifically for LLMs. Synthetic dataset fills gaps in the volume of multi-turn counseling data for pre-training, but they still have significant differences compared to human counseling in quality.

As detailed in Appendix F, we invited psychological experts and students to compare and score ChatGPT-synthetic and human counseling sessions (Table 13). The experts’ comments indicate that AI-synthesized counseling texts exhibit shallow empathy, less emotional interaction, and insufficient emotional attention. In contrast to human therapists, LLMs tend to overemphasize problem-solving and fail to fully engage the seeker’s intrinsic help-seeking motivation. This reveals the weaknesses of LLMs in establishing deep emotional connections and exploring underlying motivations.

3 Psy-Insight Construction

The goal of the Psy-Insight dataset is to collect **non-synthetic multi-turn bilingual counseling** for training LLMs and synthesize **step-by-step explanations** for multi-tasks with the help of original explanatory texts. We use synthetic labels to improve the interpretability of datasets while maintaining the quality of dialogues based on real counseling data. Our method aims to fully unlock the potential of real counseling and narrative context in raw data, maximizing their utility.

3.1 Workflow

We collect dialogues from blogs, books, and common crawled websites. The data source records and copyright information are shown in Appendix 6. Figure 3 shows the whole workflow of corpus construction.

3.2 Psychotherapy

Table 1 shows the psychotherapy labels in Psy-Insight. Following suggestions for professional therapists, we utilize these psychotherapy’s keywords as anchors (eg. Keyword: Solution-focused Brief Therapy) for crawling and locating counseling cases in raw data. Psychotherapies such as

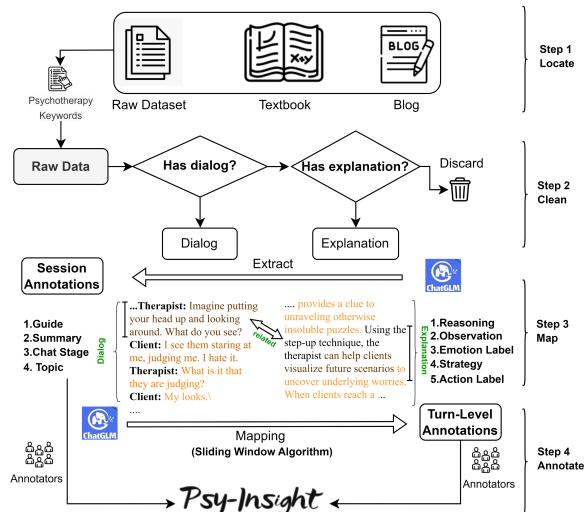


Figure 3: The construction workflow of Psy-Insight Dataset. Our workflow involves 4 steps: (1) Locating dialogues and explanation with psychotherapy keywords; (2) Data cleaning; (3) Mapping dialogues and explanation with sliding window algorithm, and computing similarity with embedding models and LLMs; (4) Checking annotations by human annotators.

SFBT and CBT focus on case studies, with their dialogue data frequently found in blogs and books. This makes them take a significant part of the whole dataset. These therapy labels are beneficial for psychological research.

3.3 Text Segmentation

We utilized psychotherapy labels and regex methods to locate the conversation structure and then crawled the text fragment from raw data. In these text fragments, we found numerous descriptive paragraphs that are related to dialogues. These non-dialogue sections contain rich psycho-explanations, including background, therapist reflections, explanations so on. We retained these text for label mapping.

3.4 Labels Mapping in Sliding Window

We map descriptive paragraphs into session-level labels and turn-level explanations for related dialogue. Specifically, we use a sliding window algorithm to trace related descriptive paragraphs and dialogue in the order they appear. By adjusting the window size, we can obtain dialog-description key-value pairs at different granularity levels. For example, a descriptive paragraph may explain multiple turns of dialogue within a session. By setting the sliding window, we map the content of a specific dialogue turn with a descriptive text. We then

Psychotherapy	Session	Ratio
CBT(1987)	191	24.29%
REBT(1996)	13	1.37%
SFBT(1997)	398	41.85%
Adlerian Counseling(1953)	20	2.1%
Client-centered Therapy(1946)	28	2.94%
Family Therapy(1984)	43	4.52%
Gestalt Therapy(1993)	12	1.26%
Multicultural Therapy(1991)	14	1.47%
Postmodern Therapy(1996)	16	1.68%
Psychoanalytic Therapy(1980)	46	4.84%
Psychodynamic Model(1994)	23	2.42%
Reality Therapy(2010)	10	1.05%
Unknown	10	13.15%
Overall	786	100%

Table 1: The distribution for psychotherapy labels in Psy-Insight. Most of the counseling cases are from case-centered modern psychotherapies, such as Solution-Focused Brief Therapy, SFBT(1997) and Cognitive Behavioral Therapy, CBT(1987), Rational emotive behavior therapy, REBT(1996). We only found a few cases in traditional psychotherapies in crawled data (eg., Psychodynamic Therapy (1998)).

use sentence-transformer(Reimers and Gurevych, 2019) and GLM4-9B (Du et al., 2022) to determine if the dialogue turn and its descriptive text are relevant.

These related dialogue-description text pairs include a dialogue window and its corresponding descriptive information. The dialogue window acts as the key, and the descriptive text as the value. If the descriptive text covers multiple dialogue turns, we break it into session-level labels, such as background introductions before first counseling. If the dialogue window size is less than three turns, we break the descriptive text into turn-level labels, such as observations of the client’s expressions or emotions. By analyzing the dialogue window length and descriptive content, we convert unstructured paragraphs into labels of varying granularity.

At the end of the workflow, annotators will review and verify the labels for each session. Not all dialogues have related explanatory texts, so some labels are marked as Unknown.

4 Psy-Insight Analysis

We collected 189 cases of face-to-face dialogues, with 75 cases in the Chinese dataset. Each case includes several sessions in the counseling cycle, and each session involves several turns of dialogue.

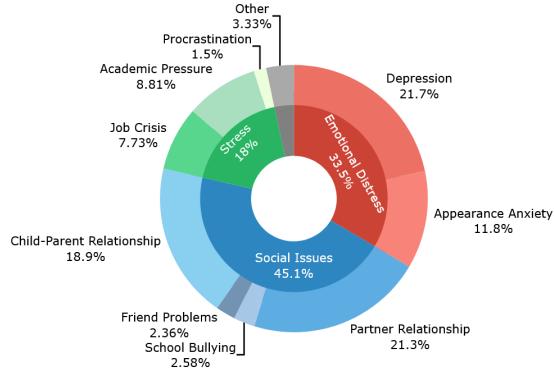


Figure 4: Statistics of topics in counseling of Psy-Insight. Top-3 topics are Depression (21.7%), Partner Relationship (21.3%), Child-Parent Relationship (18.9%).

4.1 Conversation Statistics in Psy-Insight

Table 2-3 and Figure 5 shows statistical data of counseling in Psy-Insight.

Face-to-Face Counseling Compared to GPT-synthetic (Qiu et al., 2023; Chen et al., 2023b) datasets, the conversations in Psy-Insight have longer average dialogue turns (55 vs 24 turn/case). As shown in Table 8, human therapists tend to respond with shorter turns to help and guide their patients in a multi-turn interaction (1.7 sentence/response). However, ChatGPT often gives long suggestions straight away, compared to shorter conversations (avg. 10 turn/response in the Smile dataset).

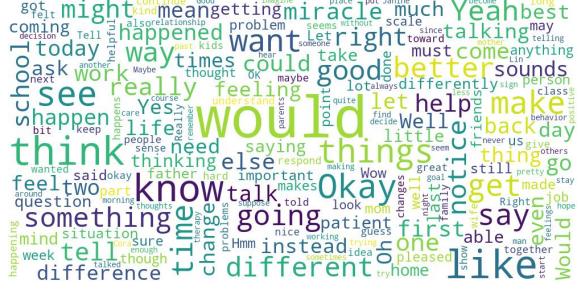
Topics Figure 4 shows the topic distribution in Psy-insight, over 50% topics belong to top-3 topics

High-frequency Words We also analyze the high-frequency word distribution in Psy-Insight’s dialog. Figure 5a displays the top 100 English consulting words, while Figure 5b does the same for Chinese. Appendix D details high-frequency words in the annotations.

Bilingual Corpus Our statistics reveal that high-frequency words and topics show notable similarities in both Chinese and English conversations, with the same words accounting for as much as 42%. Our bilingual dataset holds potential value for cross-cultural psychological research.

4.2 Label Analysis in Psy-Insight

As shown in Table 1,4, we also analyze the multi-task labels and other explainable labels. Concise labels are suitable for multi-task learning (eg. Classification). Session-level descriptive annotations such as topics and guidance are suitable as cases’



(a) Top-100 common words in English counseling.



(b) Top-100 frequent words in Chinese counseling

Figure 5: Word cloud figures of Chinese and English counseling.

Category	Total	Therapist	Client
Cases	114	-	-
Sessions	520	-	-
Turn	6,208	3,147	3,151
Avg. Utts. / Case	54.46	27.15	27.31
Avg. Utts. / Session	11.94	5.95	5.99
Reasoning	3,061	3,061	-
Observation	3,090	-	3,090

Table 2: Statistics of English Counseling in Psy-Insight. Dialogue in Psy-Insight is organized into different levels, ranging from large to small: Case - Session - Turn. During the whole counseling cycle in a specific case, the therapist always engages in a session for a topic. Each session consists of dozens of dialogue turns.

Category	Total	Therapist	Client
Cases	75	-	-
Sessions	431	-	-
Turn	5,776	2,911	2,895
Avg. Utts. / Case	77.01	38.81	38.2
Avg. Utts. / Session	13.4	6.75	6.65
Reasoning	2,831	2,831	-
Observation	2,797	-	2,797

Table 3: Statistics of Chinese Counseling in Psy-Insight.

Therapist's Strategy	Num	Proportion
Reassurance	168	2.77%
Information	139	2.29%
Unknown	551	10.73%
Providing Suggestions	122	2.01%
Question	3,896	64.14%
Reflection of Feelings	868	14.29%
Restatement	121	1.99%
Role-play	89	1.47%
Self-disclosure	19	0.31%
Overall	5,973	100%

Client's Emotion	Num	Proportion
Anger	622	9.03%
Anxiety	912	13.23%
Depression	425	6.17%
Fear	340	4.93%
Guilty	248	3.6%
Happiness	867	12.58%
Neutral	1,777	25.79%
Unknown	319	15.85%
Sadness	493	7.15%
Shame	115	1.67%
Overall	6,118	100%

Table 4: Label distribution for emotional labels and therapist's strategies in Psy-Insight dataset.

300 keys for LLM generation (eg. retrieved argument
301 generation).

302 **Therapist's Strategy** As shown in Table 4, we
303 annotated therapists' strategies based on their utterances.
304 We have also discovered that the question
305 is the most likely strategy employed by a therapist,
306 especially in the early stages of counseling.

307 **Client's Emotion** The Table 4 also shows statistics
308 of clients' emotion labels. The distribution of user
309 sentiments is relatively even. Approximately
310 50% of user sentiments are various types of nega-
311 tive emotions (e.g., Anger, Anxiety, Fear).

4.3 Additional Annotation Statistics

In addition to these labels, we also annotated labels
313 such as session-level guidance, dialog summary,
314 turn-level step-by-step reasoning labels, and so on.
315 Examples of these annotations can be found in
316 Appendix E.

5 Evaluation

To explore the potential of the Psy-Insight dataset,
319 we have researched the impact of explainable anno-
320 tations on the results for multiple tasks. We use the
321 dialogue context and the annotations for finetuning

Model & Task	English-Dataset					
	BertScore-P	Bleu-1	Bleu-3	Meteor	RougeL	Distinct-2
Mistral-7B _{Base}	0.824	0.101	0.023	0.098	0.097	0.007
Mistral-7B _{SFT dialog}	0.903	0.270	0.119	0.379	0.290	0.007
Mistral-7B _{SFT reasoning+dialog}	0.912	0.266	0.107	0.397	0.278	0.001
Mistral-7B _{SFT observation+reasoning}	0.909	0.256	0.095	0.383	0.270	0.001
Mistral-7B _{SFT mix-instructions}	0.877	0.250	0.09	0.285	0.247	0.005
ChatGPT _{Base}	0.861	0.163	0.035	0.193	0.148	0.001
ChatGPT _{SFT dialog}	0.888	0.219	0.071	0.315	0.216	0.001
ChatGPT _{SFT reasoning+dialog}	0.876	0.204	0.057	0.287	0.199	0.001
ChatGPT _{SFT observation+reasoning}	0.876	0.199	0.065	0.282	0.202	0.004
Model & Task	Chinese-Dataset					
	BertScore-P	Blue-1	Blue-3	Meteor	RougeL	Distinct-2
GLM4-9B _{RAG dialog}	0.905	0.111	0.025	0.291	0.298	-
GLM4-9B _{RAG dialog+Explanation}	0.912	0.123	0.029	0.281	0.295	-

Table 5: Result for finetuning LLMs with different combinations of annotations. The input formats of different models can be seen in Figure 4. The mix-instruction chatbot is trained by performing instruction-tuning on multiple tasks simultaneously.

(Task 3 in Table 1) and RAG (Task 4 in Table 1).

5.1 Multi-stage Generation

Each counseling session includes a guide Ins and background B_i and dialogue. In the i th session, the conversation consists of utterances from the therapist t_i and the client c_i . For j th turns in i th session, we annotated observation o_j for client utterance c_j , and annotated reasoning r_j for therapist utterance t_j .

We conduct ablation experiments and treat annotation as the variable. Inspired by the chain-of-thought (Wei et al., 2022), we insert turn-level annotation into dialog history, to guide the model first to generate objective observations, then engage in subjective reasoning, and finally produce a response.

$$Input_{i,j} = Ins + b_i + \sum_{k=0}^{j-1} (o_{i,k} + r_{i,k} + t_{i,k} + c_{i,k}) \quad (1)$$

To assess the impact of step-by-step annotations on generation results, we employed a method similar to chain-of-thought (Wei et al., 2022). LLMs treat explanatory texts o_x and r_x as targets for the first stage of generation, and generate dialogue responses t_i in the second stage. So, the target of the training is $Target_{i,j}$

$$Target_{i,j} = o_{i,j} + r_{i,j} + t_{i,j} \quad (2)$$

During the evaluation stage, the model first generates the annotation part and then the response. For evaluation, we only compute similarity metric

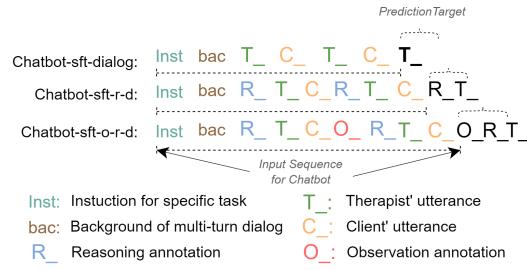


Figure 6: LLMs accomplish tasks by predicting the next token. For the training dataset of the Mistral7B-SFT-Reasoning model, incorporating reasoning annotations into the dialogue history and subsequently treating both reasoning and response as prediction targets allows LLMs to learn to reason first and then generate a response.

between generated response $t_{i,j}$ with the ground truth response.

5.2 Implementation

We use GhatGPT-3.5-turbo and Mistral-7b-instruc-v0.2 for finetuning on the English corpus for evaluation. These models are pre-trained in the instruction dataset and suitable for learning multi-turn dialogue. And we also test ChatGLM4-9B on Chinese dataset with retrieve argument generation.

Previous studies (Huerta-Enochian, 2024) showed that Prompt Loss Weight (PLW) is suitable for learning in long multi-turn dialog. In the training stage, we applied PLW masks technique to mask the instruction, background, and the client's utterances. They are not involved in loss computation.

We randomly split the Psy-Insight into a training set (85%) and a test set (15%). For reproduction,

Model	Int.	Help.	Con.	Exp.
Human	3.44	3.58	3.78	3.43
GLM4-9B	2.90	2.76	2.79	3.21
GLM4-9B _{reasoning}	3.12	2.88	2.85	3.25
GLM4-9 _{obs+reasoning}	3.00	2.62	2.56	2.78

Table 6: Human interactive evaluation focuses on interactivity, helpfulness, comfort, and explainability.

we utilized open-source models (GLM, Mistral) for the experiment. Our code and expert evaluation results are available as open-source on our Github.

5.3 Automatic Evaluation

Metrics The automatic evaluation metrics including BertScore-P (Zhang et al., 2020), BLEU (Papineni et al., 2002), Meteor (Banerjee and Lavie, 2005), RougeL (Lin, 2004), and Distinct-2 (Li et al., 2016).

Result Table 5 shows the automated metrics results of the instruction-tuning. Compared with the baseline, all finetuned models show improvement in automatic metrics. This suggests that Psy-insight is helpful for the model to better generate responses for mental support. Compared to pure dialog finetuning, incorporating step-by-step annotations results in increased metrics such as BertScore-P, Meteor, and Distinct-2 scores.

5.4 Human Evaluation

We designed two types of human evaluation experiments.

Human Interactive Evaluation As shown in Table 6, we conducted an ablation experiment with the Psy-Insight dataset to investigate the generation quality of ChatGLM4 with dialog and annotation. We used different labels from Psy-Insight to fine-tune ChatGLM4-9B models and compared their generated results with the original human counseling dialogue.

Similar Cases Evaluation As shown in Table 7, to compare the counseling quality across datasets, we randomly selected English cases from the Psy-Insight and compared them with similar cases from the Esconv (Liu et al., 2021a) English dataset. In the same way, we compared Psy-Insight’s Chinese cases with similar cases from the synthetic Smile (Qiu et al., 2023) dataset.

Metrics We recruited 10 student volunteers and 5 experts for human evaluation on 60 random selected responses. They were asked to compare the responses based on the following metrics: **Inter-**

Psy-Insight vs.	Esconv(2021a)		Smile(2023)	
	Win	Lose	Win	Lose
Interactivity	55	21	57	17
Helpfulness	44	28	52	24
Comforting	56	27	46	34
Explainability	45	29	42	30
Overall	50	32	51	25

Table 7: Result of the expert A/B test evaluation among Psy-Insight and other baseline dataset. Experts not only provided scores but also offered comparative assessments for similar cases, as detailed in Appendix F.

activity: Does the therapist have the intention to continue the conversation? **Helpfulness:** Whether the suggestion suitable for client? **Comforting:** Whether the response is useful for comforting client? **Explainability:** Is the self-explanation by the model reasonable in counseling? We standardized the evaluation results to ensure consistency among raters. In case comparison, we ensured randomness and similarity by clustering cases’s topics with sentence-Bert (Reimers and Gurevych, 2019), selecting the top 5 detailed cases per major category, and assessing content/theme similarity with ChatGLM4-9b. We then chose 20 pairs for expert scoring (Table 7) and 10 for comparative assessments (Appendix F).

Result As shown in Table 6, LLMs trained with reasoning annotation generates high-quality counseling conversations. However, there is still a remarkable gap between the mental support LLMs and professional therapists. Table 7 shows that Psy-Insight achieves better quality than Esconv in English dialogue data, and its Chinese dialogue quality surpasses that of the synthetic dataset Smile.

6 Conclusion and Future Work

We provide Psy-Insight, a high-quality bilingual psychological counseling corpus, and annotated it with step-by-step reasoning and multi-task labels. Rich annotations make the Psy-Insight dataset suitable for multi-task learning in mental support. Additionally, with step-by-step explanations, LLMs can understand the reasoning behind counseling.

In future work, we plan to utilize expert assessments in Appendix F. These comments from professional counselors are high-quality human feedback. We hope to utilize them with techniques such as reinforcement learning or DPO (Rafailov et al., 2024), enabling LLMs to counsel based on human preferences.

448 Limitations

449 Although the Psy-Insight dataset offers rich psycho-
450 logical annotation for at least 5 NLP tasks, we only
451 evaluate the instruction tuning in English dialogue
452 due to the limited resources. We hope the statistical
453 data we provide can help future research.

454 Copyright

455 We collect datasets from crawled blogs and books
456 and also extract conversation from raw common
457 crawl datasets including the book3 and Massive
458 Never-ending BT Vast Chinese corpus project. All
459 the copyright information and data sources are
460 recorded in our GitHub repository.

461 References

462 Franz Alexander. 1980. *Psychoanalytic therapy: Prin-
463 ciples and application*, volume 1946. U of Nebraska
464 Press.

465 Satanjeev Banerjee and Alon Lavie. 2005. METEOR:
466 An automatic metric for MT evaluation with im-
467 proved correlation with human judgments. In *Pro-
468 ceedings of the ACL Workshop on Intrinsic and Ex-
469 trinsic Evaluation Measures for Machine Transla-
470 tion and/or Summarization*, pages 65–72, Ann Arbor,
471 Michigan. Association for Computational Linguis-
472 tics.

473 Akshma Chadha and Baijnath Kaushik. 2021. Machine
474 learning based dataset for finding suicidal ideation on
475 twitter. In *2021 Third International Conference on
476 Intelligent Communication Technologies and Virtual
477 Mobile Networks (ICICV)*, pages 823–828. IEEE.

478 Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng,
479 Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023a.
480 **SoulChat: Improving LLMs' empathy, listening, and**
481 **comfort abilities through fine-tuning with multi-turn**
482 **empathy conversations.** In *Findings of the Associa-
483 tion for Computational Linguistics: EMNLP 2023*,
484 pages 1170–1183, Singapore. Association for Com-
485 putational Linguistics.

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

493 Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui
494 Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng.
495 2022. Improving multi-turn emotional support dia-
496 logue generation with lookahead strategy planning.
497 *arXiv preprint arXiv:2210.04242*.

498 Steve De Shazer and Insoo Kim Berg. 1997. ‘what
499 works?’remarks on research aspects of solution-
500 focused brief therapy. *Journal of Family therapy*,
501 19(2):121–124.

502 Raymond DiGiuseppe. 1996. The nature of irrational
503 and rational beliefs: Progress in rational emotive
504 behavior theory. *Journal of Rational-Emotive and
505 Cognitive-Behavior Therapy*, 14(1):5–28.

506 Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding,
507 Jiezhang Qiu, Zhilin Yang, and Jie Tang. 2022. Glm:
508 General language model pretraining with autoregres-
509 sive blank infilling. In *Proceedings of the 60th An-
510 nual Meeting of the Association for Computational
511 Linguistics (Volume 1: Long Papers)*, pages 320–335.

512 Zahor Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and
513 Maya Lvovsky. 2023. Chatgpt outperforms humans
514 in emotional awareness evaluations. *Frontiers in
515 Psychology*, 14:1199058.

516 Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and
517 Pushpak Bhattacharyya. 2020a. MEISD: A multi-
518 modal multi-label emotion, intensity and sentiment
519 dialogue dataset for emotion recognition and senti-
520 ment analysis in conversations. In *Proceedings of
521 the 28th International Conference on Computational
522 Linguistics*, pages 4441–4453, Barcelona, Spain (On-
523 line). International Committee on Computational Lin-
524 guistics.

525 Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and
526 Pushpak Bhattacharyya. 2020b. MEISD: A multi-
527 modal multi-label emotion, intensity and sentiment
528 dialogue dataset for emotion recognition and senti-
529 ment analysis in conversations. In *Proceedings of
530 the 28th International Conference on Computational
531 Linguistics*, pages 4441–4453, Barcelona, Spain (On-
532 line). International Committee on Computational Lin-
533 guistics.

534 Google. 2023. Bard. <https://bard.google.com>. Ac-
535 cessed: 2023-03-31.

536 Ralph W Heine. 1953. A comparison of patients' re-
537 ports on psychotherapeutic experience with psycho-
538 analytic, nondirective and adlerian therapists. *Ameri-
539 can Journal of Psychotherapy*, 7(1):16–23.

540 William P Henry, Hans H Strupp, Thomas E Schacht,
541 and Louise Gaston. 1994. Psychodynamic ap-
542 proaches.

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

547 Mathew Huerta-Enochian. 2024. Instruction fine-
548 tuning: Does prompt loss matter? *Preprint*,
549 arXiv:2401.13586.

550 Kira Kretzschmar, Holly Tyroll, Gabriela Pavarini, Ari-
551 anna Manzini, Ilina Singh, and NeurOx Young Peo-
552 ple's Advisory Group. 2019. Can your phone be

553	your therapist? young people’s ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. <i>Biomedical informatics insights</i> , 11:1178222619829083.	608
554		609
555		610
556		
557	Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. <i>Advances in Neural Information Processing Systems</i> , 33:9459–9474.	611
558		612
559		613
560		614
561		615
562		
563	Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. <i>arXiv preprint arXiv:2307.11760</i> .	616
564		617
565		618
566		
567		
568	Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. <i>Preprint</i> , arXiv:1510.03055.	619
569		620
570		621
571		
572	Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In <i>Text Summarization Branches Out</i> , pages 74–81, Barcelona, Spain. Association for Computational Linguistics.	622
573		623
574		
575		
576	Marsha M Linehan. 1987. Dialectical behavioral therapy: A cognitive behavioral approach to parasuicide. <i>Journal of Personality disorders</i> , 1(4):328–333.	624
577		625
578		626
579	Bingjie Liu and S Shyam Sundar. 2018. Should machines express sympathy and empathy? experiments with a health advice chatbot. <i>Cyberpsychology, Behavior, and Social Networking</i> , 21(10):625–636.	627
580		628
581		629
582		630
583	June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language model for mental health support. <i>arXiv preprint arXiv:2309.15461</i> .	631
584		632
585		633
586		634
587	Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021a. Towards emotional support dialog systems. <i>arXiv preprint arXiv:2106.01144</i> .	635
588		
589		
590		
591	Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021b. Towards emotional support dialog systems. <i>arXiv preprint arXiv:2106.01144</i> .	636
592		637
593		638
594		639
595	Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 8968–8979, Online. Association for Computational Linguistics.	640
596		641
597		642
598		643
599		644
600		
601		
602		
603	Manisha, William Clifford, Eugene McLaughlin, and Paul Stynes. 2023. A deep learning emotion classification framework for low resource languages. In <i>International Conference on Big Data Analytics</i> , pages 113–121. Springer.	645
604		646
605		647
606		648
607		649
553	Sheila McNamee. 1996. Therapy and identity construction in a postmodern world. <i>Constructing the self in a mediated world</i> , pages 141–155.	650
554		651
555		652
556		653
557		
558		
559		
560		
561		
562		
563		
564		
565		
566		
567		
568		
569		
570		
571		
572		
573		
574		
575		
576		
577		
578		
579		
580		
581		
582		
583		
584		
585		
586		
587		
588		
589		
590		
591		
592		
593		
594		
595		
596		
597		
598		
599		
600		
601		
602		
603		
604		
605		
606		
607		

- 663 Carl R Rogers. 1946. Significant aspects of client- 719
 664 centered therapy. *American psychologist*, 1(10):415- 720
 665 422.
 666 Victor Sanh, Albert Webson, Colin Raffel, Stephen H 721
 667 Bach, Lintang Sutawika, Zaid Alyafeai, Antoine 722
 668 Chaffin, Arnaud Stiegler, Teven Le Scao, Arun 723
 669 Raja, et al. 2021. Multitask prompted training 724
 670 enables zero-shot task generalization. *arXiv preprint 725*
 671 arXiv:2110.08207.
 672 Kirsten A Smith and Judith Masthoff. 2018. Can a 726
 673 virtual agent provide good emotional support? In 727
 674 *Proceedings of the 32nd International BCS Human 728*
 675 Computer Interaction Conference. BCS Learning & Development.
 677 Suzette L Speight, Linda J Myers, Chikako I Cox, and 731
 678 Pamela S Highlen. 1991. A redefinition of multicultur- 732
 679 tural counseling. *Journal of counseling & development*, 733
 680 70(1):29–36.
 681 Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and 734
 682 Minlie Huang. 2021. Psyqa: A chinese dataset for 735
 683 generating long counseling text for mental health 736
 684 support. *arXiv preprint arXiv:2106.01702*.
 685 Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong 737
 686 Wen, and Rui Yan. 2022. Misc: a mixed strategy- 738
 687 aware model integrating comet for emotional support 739
 688 conversation. *arXiv preprint arXiv:2203.13560*.
 689 Aditya Nrusimha Vaidyam, Hannah Wisniewski, 740
 690 John David Halama, Matcheri S Kashavan, and 741
 691 John Blake Torous. 2019. Chatbots and conversational 742
 692 agents in mental health: a review of the psychiatric 743
 693 landscape. *The Canadian Journal of Psychiatry*, 744
 694 64(7):456–464.
 695 C Seth Warren. 1998. Models of brief psychodynamic 745
 696 therapy: A comparative approach.
 697 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten 746
 698 Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, 747
 699 et al. 2022. Chain-of-thought prompting elicits reasoning 748
 700 in large language models. *Advances in neural 749*
 701 information processing systems, 35:24824–24837.
 702 WHO. 2022. Who highlights urgent need to transform 750
 703 mental health and mental health care. Accessed: 751
 704 2023-03-31.
 705 Thomas Wolf, Lysandre Debut, Victor Sanh, Julien 752
 706 Chaumond, Clement Delangue, Anthony Moi, Pier- 753
 707 ric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, 754
 708 Joe Davison, Sam Shleifer, Patrick von Platen, 755
 709 Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, 756
 710 Teven Le Scao, Sylvain Gugger, Mariama Drame, 757
 711 Quentin Lhoest, and Alexander Rush. 2020. *Trans- 758*
 712 *formers: State-of-the-art natural language processing*. 759
 713 In *Proceedings of the 2020 Conference on Empirical 760*
 714 *Methods in Natural Language Processing: System 761*
 715 *Demonstrations*, pages 38–45, Online. Association 762
 716 for Computational Linguistics.
 717 Robert E Wubbolding. 2010. Reality therapy. *The 763*
 718 *Corsini Encyclopedia of Psychology*, pages 1–3.

747 A Ethical Evaluation

748 Privacy

749 Considering the potential ethical risks. We filter
750 the privacy information in Psy-Insight to ensure
751 that our data is only related to the cases and not to
752 personal identities. The raw crawled data contains
753 sensitive data like names, URL links, contact information
754 phone numbers so on. We put a lot of effort
755 in the anonymization process.

756 For example, as the title of both sides in the
757 consultation, real names are frequently collected
758 in counseling data. In our experiment, we observe
759 that masking the real name directly will cause a per-
760 formance decrease in model finetuning. As a result,
761 We take the rule-based method to replace the title
762 in the dialogue content with pronouns and annotate
763 common names based on gender and family role di-
764 alogues, such as "therapist", "client", and "Smith"
765 for adult men, and "Mary" for adult women, and
766 so on. Similarly, we use roles-based tools to fil-
767 ter sensitive personal information in raw data. We
768 guarantee that our Psy-Insight does not include any
769 personal privacy information. All names in Psy-
770 Insight are aliases.

771 Ethical Risk

772 We safeguard the anonymity of the data. In our
773 repository, users must agree to a statement that
774 they cannot trace or de-anonymize the content of
775 the consultation data. Our dataset is intended solely
776 for academic research.

777 Beyond privacy concerns, we are also aware of
778 the high risks associated with training LLMs to
779 provide psychological counseling services. To pre-
780 vent models trained on Psy-Insight from generat-
781 ing misleading responses, we have cleaned and
782 risk-assessed the counseling dialogues in the Psy-
783 Insight dataset. First, we filtered out some counse-
784 lings lacking psychotherapy labels from the crawled
785 dataset. We cannot ensure the controllability and
786 scientific nature of dialogues with unclear data
787 sources. Next, we separated counseling dialogues
788 with negative emotional (eg., anxiety) and behav-
789 ior keywords based on emotional tags, which
790 may include actions such as suicide and depression.
791 Finally, we had three professional psychological
792 therapists assess the dialogue content and filter the
793 useless responses. All of the psycho-experts in our
794 work have at least national second-level psycho-
795 logical certificates. Our effort aims to ensure that
796 the counseling dialogues in the Psy-Insight dataset

797 align with psychological standards and ethics.

798 B Reproduction

799 In this section, we provide an overview of our ex-
800 perimental setting in Evaluation. All experiments
801 were conducted on 2 NVIDIA RTX 3090 GPUs.
802 For finetuning, we apply (Hu et al., 2021) Lora for
803 instruction-tuning LLM. We set alpha at 128 and
804 rank at 256.

805 We finetune the English part of Psy-Insight with
806 a batch size of 8 and a learning rate of 0.00002.
807 For equality in evaluation, all models are trained
808 for 220 epochs. We take the transformers-4.36.2
809 (Wolf et al., 2020) and peft-0.9.1 frameworks for
810 training.

C Construction Workflow Example

811

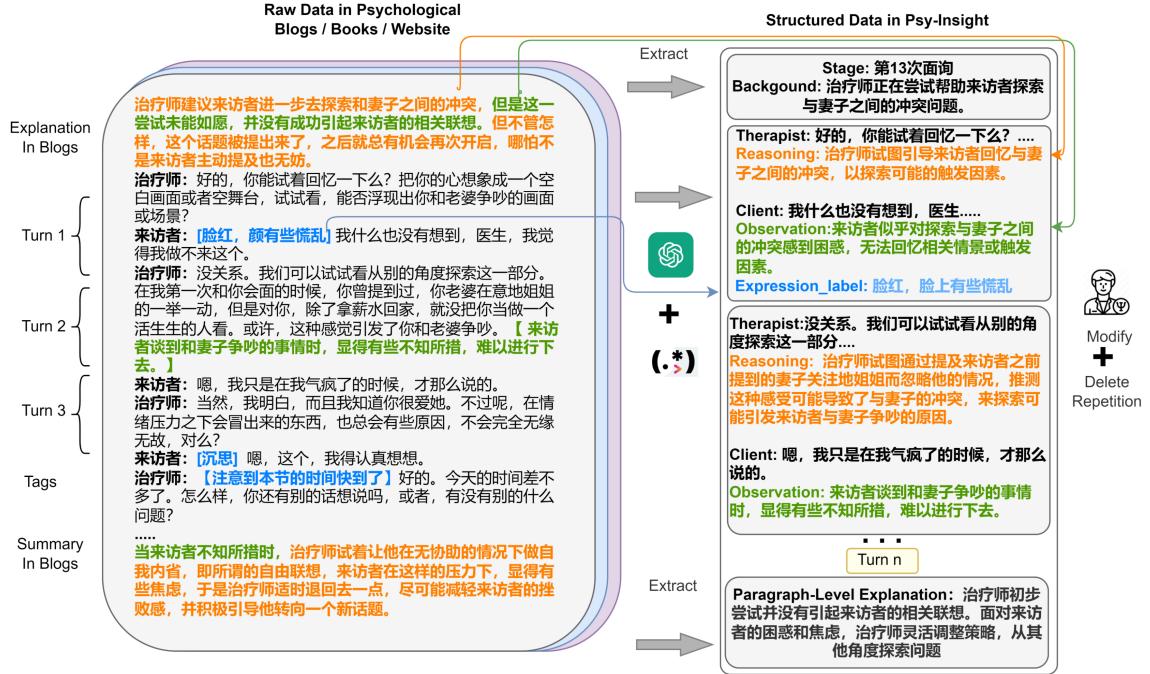


Figure 7: The data collection process in Psy-Insight Dataset. We use psychological labels as anchors to locate dialogue-like structures from raw web data. After positioning, we structure the dialogue data and descriptive text using regular expressions and LLMs. In workflow of dialog collection, GLM4-9B and Mistral7B are used for formatting rather than synthesizing them from scratch. Non-dialog labels are extracted from explanatory paragraph in raw textbooks. Furthermore, a long-context LLM agent is used to map these structured labels into related dialog turn.

D Detailed Statistics

812

Dataset	Domain	Dialog Source	Avg. Turn	Total Turn	Language	Annotations
Empathetic Dialogue (2018b)	Emotional Support	forum	4.31	24,850	en	Emotion
ESConv (2021b)	Emotional Support	crowd workers	29.8	1,053	en	Background,Strategy
MedDialog (2020)	Medical Dialogue	hospital	3.19	1,145,231	en	Emotion,Intensity
D4 (2022b)	Depression Diagnosis	hospital	21.6	257,454	zh	Depression
PsyQA (2021)	Mental Health	forum	2.51	1,339	zh	-
Smile (2023)	Mental Health	ChatGPT	10.4	22,346	zh	Background,Strategy
SoulChat (2023b)	Mental Health	ChatGPT	20	55,165	zh	Emotion,Intensity
Psy-Insight	Mental Health	book blog	46	2,300,248	zh	-
			77	6,208	en	Explainable Annotations
				5,776	zh	Multi-task Labels

Table 8: Comparison of the Psy-Insight dataset with baseline datasets. Compared to previous datasets, Psy-Insight includes face-to-face counseling from real life, featuring longer average dialogue turns (46 in English, 77 in Chinese). Unlike previous datasets that primarily focus on short labels and single subtasks, Psy-Insight annotates dialogues at various granularities with step-by-step explanatory and rich multi-task labels.



(a) Word clouds of English explainable annotations in Psy-Insight.



(b) Word clouds of Chinese explainable annotations in Psy-Insight.

Figure 8: Common words in annotations.



(a) The translation of previous Chinese explainable annotation dialog.



(b) The translation of Chinese reasoning picture.

Figure 9: Translation for previous Chinese word picture.

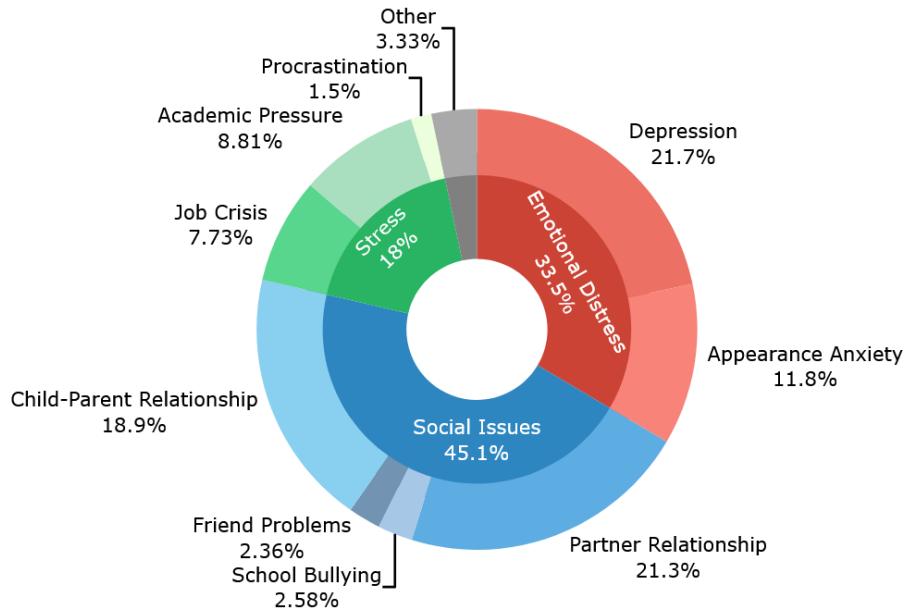


Figure 10: Statistics of topics in counseling of Psy-Insight.

E Data Examples of Psy-Insight

813

Topic	Session ID	Psychotherapy
Addressing academic challenges in students.	32	Solution Focused Brief Therapy (SFBT)
Background	Guide	
A student struggling with attendance and academic performance, as discussed in the previous session.	Acknowledge efforts, ask open-ended, reflect feelings, respect boundaries, and encourage sharing.	
Dialog	Strategy / Emo-label	Reasoning / Observation
Therapist: Yes, you've been able to get to school sometimes but not as much as you may want. I'm impressed, though, with how you are at least trying to make it to school these last few weeks since we talked. There must be a lot going on that makes it harder for you to come to school as much as you would like. I'm curious; is there anything that would be helpful to talk about that might help even a little bit?	Question	Therapist acknowledges the student's efforts and explores potential topics for discussion to offer support.
Beth: Maybe ... I don't know if I can ... not sure ... it is very hard and I don't know what will happen if I talk about it.	Anxiety, Fear	Beth expresses uncertainty and difficulty in opening up.
Therapist: Well, in what way do you think it might help if you did share it with me or someone else? Do you think it would make things better for you ... like feeling like coming to school?	Question	Therapist encourages Beth to consider the potential benefits of sharing her concerns.
Beth: I don't know. It might make things even worse.	Fear	Beth fears that sharing could worsen things
Therapist: It must be very important to you if it might make things worse. Even though it sounds like if you were to get some help with this issue, it might make your life easier in doing the things you seem to want to do, like school. That is a tough place to be. What would be the most helpful for us to do that might help move you to a better place and make you feel better?	Question, Reflection	Therapist acknowledges the significance of Beth's concerns and explores ways to support her in improving her situation.
Summary		
The student begins to share her experience, possibly due to feeling accepted and supported by the therapist, indicating a growing trust in the therapeutic relationship.		

Table 9: An example session in the English Dataset of Psy-Insight.

The data in the Psy-Insight dataset is collected and annotated in session units. Table 9 and Table 10 show the data units for English and Chinese data, respectively. Each session contains 5 to 50 rounds of dialogue between a counselor and a client on a specific topic. Our annotation labels include Session-Level and Turn-Level labels. Session-Level labels include background, guidance, session topic, and counseling summary. Turn-Level labels are annotations for each turn of dialogue in

multi-turn conversations. We labeled different aspects of the counselor's and client's dialogue content. For the counselor's dialogue, we annotated counseling strategies and subjective reasoning. For the client's dialogue, we annotated emotional classification results and observations of dialogue facts from the counselor's perspective. Due to space limitations, more Chinese dialogue examples and English dialogue cases can be found on our github website³

824

825

826

827

828

829

830

831

832

833

Stage	Session ID	Psychotherapy
The 4th Session	23	Postmodern Therapy
Background	Guide	
来访者是一个年轻女性，面临与家庭和亲密关系相关的挑战，希望在情感认知和复杂性方面得到支持和理解。可能正在寻求解决家庭和婚姻中的困境，希望更清晰地了解自己的情感和选择。		治疗师需要通过逐步询问，帮助她分离问题，减轻负面情绪并澄清问题影响，以提升她的投入感和生活应对能力。
Dialog	Strategy / Emo-label	Reasoning / Observation
治疗师: 你愿意挑战自己的恐惧并更加开放地表达自己，但是你还是希望一点一点慢慢来，是吗？	Question	治疗师通过逐步询问引导来访者澄清问题。
来访者: 当然。	Neutral	None
治疗师: 来访者，你从哪里学到如何做一名女性、妻子、母亲的？	Question	通过提问探索来访者的性别角色认知。
来访者: 我不知道。我从来没有想过这个问题。	Neutral	来访者对这个主题缺乏反思。
治疗师: 嗯，我想知道你从哪里学到要照顾他人，将自己的需要摆在他人需要的后面，有时甚至应该牺牲小我成全他人的？	Question	进一步询问来访者自我牺牲和他人优先的观念来源。
来访者: （现在依然如此）我想我是从父母那里学来的，还有我父亲对待我母亲的方式。	Others	来访者认识到她的行为模式和观念可能来自父母的影响。
治疗师: 你不认为你现在的生活方式就是在学习你母亲的做法吗——关于如何做一名女性、妻子和母亲？	Question	让来访者反思她是否在重复母亲的生活模式。
来访者: 我不确定，我从没想过我是否在重复我母亲的生活方式。我知道她从未想过要出去工作或去读书。从这个角度来看，我和她有很大的不同。	Neutral	来访者表示不确定，也指出她们之间的显著差异。
治疗师: 是的，这看起来的确是个真正的差异。我想知道你能否找到你和她一致的地方？	Question	治疗师承认差异，并鼓励来访者寻找和母亲生活方式相似的部分。
来访者: 嗯，我认为我母亲十分传统。在她看来，男人是整个家庭的领导，是家庭收入的主要来源，如果你希望的话，他也可以成为家庭的保护者。而女性的工作就是养育孩子、照料整个家庭，我猜还包括照顾丈夫。	Neutral	来访者描述她母亲的传统观念和性别角色分工。
治疗师: 在你母亲教授给你的诸多观念中，你接纳了其中的多少？	Question	治疗师询问来访者对母亲观念的内化程度。
来访者: 嗯，我猜在我结婚的前15年中，我一直在步母亲的后尘。实质上我在我的婚姻生活中一直在做她那样的女性。现在，这成为了我的问题。我觉得我不希望再这样下去了。我不希望以牺牲我的利益为代价去担负起让来访者幸福的责任。我不希望让我的生活继续围着来访者转。但是我又痛恨自己的这些想法。我为自己想要追求自己的职业而感到内疚。	Guilty	来访者反思自己在婚姻中的行为模式，并表达出对改变的渴望和内心的矛盾与内疚感。
Summary		
作为咨询师，我应当引导来访者质疑、反思并逐步替换这些不健康的文化信念，建立更积极的自我认知。		

Table 10: A data unit in the Chinese part of Psy-Insight. Due to the page limitation, we only show half part of dialog in this 19-turn dialog. Table 15 shows its English translation.

F Expert Assessments of Psy-Insight and Baseline Datasets

834

English Counseling Case	
Psy-Insight	Esconv (Liu et al., 2021a)
Source	
Textbook& Psychological blogs	Crowdworkers & Volunteers
<u>seeker</u> : Well, I certainly would be a lot more cheerful.	<u>seeker</u> : Hello there, would you be able to give me some advice?
<u>supporter</u> : Mmm-hmm. Would that ... do you think that would have some effect on this getting better business?	<u>supporter</u> : I would be glad to. What is troubling you?
<u>seeker</u> : Probably.	<u>seeker</u> : The first lockdown was bad enough but the second has really thrown me into despair. I simply cannot cope without being able to see friends and family for much longer.
<u>supporter</u> : Yeah. Oh. Okay. That was the only question to pop into my head while I was ... Do you have any? All right. I think we should take a break and we'll be back in 10 minutes. Okay. Just relax.	<u>supporter</u> : I completely understand and had similar experiences. What were some ways you tried to interact with people last shut down?
<u>seeker</u> : Yes, we do.	<u>seeker</u> : I use facebook quite a lot and it helps but it cannot compensate for face to face contact. I really need to be able to see people.
<u>supporter</u> : Okay. You need more practice in just relaxing. Just ...	<u>supporter</u> : Digital contact does have its limitations. Could you meet with people outside at a distance?
<u>supporter</u> : Well, I'm glad you both came today and I thought I saw some new talent that I hadn't heard about before today.	<u>seeker</u> : Well, part of the problem with that is that I am not sure what the rules are any more! We have been told so many different things that I am out of the loop.
<u>supporter</u> : I wrote down some of the stuff the team said about you.	<u>supporter</u> : I understand. Try sticking to the basics to make it less stressful. Wear a mask outside at a distance.
<u>seeker</u> : Yes.	<u>seeker</u> : I always do wear a mask.. I think we legally have to :). I really miss little things like going out for coffee or to the gym though. It makes me feel like will never be the same again.
<u>supporter</u> : And they are very impressed by how concerned you are and how caring you are and your perseverance. And you—at the same time being smart, taking one day at a time. But also, I mean you have been seeking help for four years.	<u>supporter</u> : Have to talked about this with friends? They may be feeling the same way. Just talking this out with the may be reassuring.
<u>seeker</u> : Yes.	<u>seeker</u> : I have a little bit but I don't want to burden them. The other aspect is that it is making me anxious about many others things too. I keep having morbid thoughts about what will happen if the pandemic never ends.
<u>supporter</u> : And you never gave up. And it must—	<u>supporter</u> : That's understandable. Maybe talking to a professional who you don't have to worry about burdening with your concerns will allow you to talk about your concerns.
<u>seeker</u> : I couldn't.	<u>seeker</u> : That's why I'm here ;). You are being very helpful, thank you! It's just nice to be able to tell people what I am worried about and not have them think that I am just being silly.
<u>supporter</u> : Oh, sure, but it also must take some knowing about what is possible to sort of give you the power to continue to search knowing that you could.	<u>supporter</u> : Many people are struggling with the current state of affairs. Don't be afraid to talk to people they may also be afraid to talk to people and you could miss out on supporting each other.
<u>seeker</u> : Yes.	<u>seeker</u> : That does sound like really good advice. The other thing I haven't mentioned yet is that I have been having financial pressures because I am on furlough. I am worried that if this lasts much longer I will end up deep in debt.
<u>supporter</u> : Not believing that it's impossible. There's force in that. And at the same time being cautious at this point is a sign of wisdom. Okay. And someone pointed to the fact that you're very, very good at sharing your happiness about him being better and showing him that.	
<u>seeker</u> : Yes, but I think that's ... that's one of ... that you have pointed out and you have tried ... to tell me to do that sometimes when I just feel: "What shall I do, what can I do, how can I do?" And I think of what you are simply saying during these meetings we have had in $1\frac{1}{2}$ years. And you have said, "Try to see the good things that were happening just now. And, my older sons, I think we are very close, and we are talking, and they are seeing when I am not feeling so well about everything and they try to support me there.	
... turn X 10	

Table 11: The case comparison of Psy-insight and Esconv on the similar topic: Distress.

Psy-Insight	Esconv(Liu et al., 2021a)
Dialog ID	
1	2
Experts' Assessments	
<p><u>专业咨询师A</u>: 1更像真实的咨询，双方有情感流动，有非言语信息。2比较像机器人，一直在回答来访者的问题，情感互动较弱，对求助者的情绪关注不够。</p> <p><u>专业咨询师B</u>: 1非常尊重求助者的决定，给出建议让求助者自己做决定；2依然更关注问题解决。咨询师不应替求助者做决定。</p> <p><u>心理系学生C</u>: 对话1更好，有共情和尝试让来访者提出自己曾经尝试过有效方法。</p> <p><u>心理系学生D</u>: 对话2更好，提出了具体的解决办法；对话1的处理方式可能并不太合适。</p> <p><u>心理系学生E</u>: 1有一种逼着seeker做决定的感觉，这让我觉得不太好。因此尽管1提出的解决方案可能有效，也未必真的就起到作用，还可能产生二次伤害。和其他的2相比，感觉这个2在安慰有效性上比较欠缺，也是说了一些“正确的废话”。</p> <p><u>心理系学生F</u>: 对话1更好，在敏感话题给予了隐私维护。</p> <p><u>心理系学生G</u>: 对话1更好，对话1中清楚询问了情况并给出了具体的建议，并安抚了求助者的情绪。对话2则更像机器人的回答。对话1更自然。</p>	
Translation of Assessments	
<p><u>Professional Counselor A</u>: Dialogue 1 feels more like real counseling, with emotional flow and non-verbal communication. Dialogue 2 seems more robotic, constantly answering the seeker's questions, with weaker emotional interaction and insufficient attention to the seeker's emotions.</p> <p><u>Professional Counselor B</u>: Dialogue 1 respects the seeker's decisions, offering suggestions for the seeker to make their own decisions; Dialogue 2 still focuses more on problem-solving. Counselors should not make decisions for the seeker.</p> <p><u>Psychology Student C</u>: Dialogue 1 is better, with empathy and attempts to get the seeker to suggest effective methods they have tried before.</p> <p><u>Psychology Student D</u>: Dialogue 2 is better, as it proposes specific solutions; the approach in Dialogue 1 may not be entirely appropriate.</p> <p><u>Psychology Student E</u>: Dialogue 1 gives a sense of forcing the seeker to make a decision, which I find uncomfortable. Therefore, even if the solutions proposed in Dialogue 1 are potentially effective, they may not actually work and could cause secondary harm. Compared to the other Dialogue 2, this Dialogue 2 seems to lack effectiveness in comfort, also saying some "correct but useless" words.</p> <p><u>Psychology Student F</u>: Dialogue 1 is better, providing privacy protection on sensitive topics.</p> <p><u>Psychology Student G</u>: Dialogue 1 is better, as it clearly inquires about the situation, offers specific advice, and soothes the seeker's emotions. Dialogue 2 feels more like a robotic response. Dialogue 1 is more natural.</p>	

Table 12: The expert assessments of Table 13. All assessments of the case comparisons are open-source on our GitHub website.

Chinese Counseling Case	
Psy-Insight	Smile(Qiu et al., 2023)
Source	
Textbook& Psychological blogs	ChatGPT-Synthesis from PsyQA
Topic	
Depression	
来访者: 说实话, 因为我自己无法解决自己的问题, 所以我需要得到专业的帮助, 这一点让我觉得很沮丧。	来访者: 我最近总感觉心情很低落, 没有办法集中精力。
治疗师: 如果你寻求外界的帮助却不觉得沮丧的话, 那我才要吃惊了。这是一种健康的感受。但是事实上, 你已经迈出了重要的一步, 这是值得庆贺的进步。这一定需要极大的勇气。	治疗师: 听起来你可能是焦虑或者抑郁了。你有考虑过找专业人士帮助你么?
来访者: 嗯, 我从来没有想过这就算是一个大进步。	
治疗师: 现在你坐在这里, 你是否相信治疗能够给你带来帮助呢? 你是否相信治疗能在很短的时间内改善你的生活呢?	
来访者: 我来到这里就是因为我觉得这能帮助我。	
治疗师: 我了解了你的过去, 我们也已经进行了一段时间的交谈, 我相信你会感觉好些的。我不会做出什么保证, 但是我想, 改善你的生活并非不可能。我这样想并非基于一个无意义的愿望, 而是基于你给出的四个关键信息: 你已经迈出了重要一步—你来到了这里, 你在一定程度上相信你的生活会得到改善。此外, 你已经为自己建立了一个目标: 得到新的生活。最后, 你对自己的痛苦能保持一种开放性的态度。换句话说, 你可以描述出你的痛苦来。	
来访者: 那意思就是我还是有希望的。	
治疗师: 是的, 的确有希望。我还相信你很快就能感觉好起来, 只需要一个条件。	
来访者: 什么条件?	
治疗师: 那就是你愿意付出努力, 哪怕面对十分困难的工作时也会势力不屈。这项困难的工作听起来会很奇怪, 那就是你要尽力不再那么费力去和痛苦斗争。	
来访者: 我愿意尝试一下。	
来访者: 你说的有道理, 我会试着去做的。可是我总是觉得如果我失败了, 再次陷入低谷, 那我就真的一无所有了。	
治疗师: 不要害怕失败, 每个人都有过不如意的经历。但是你可以从失败中寻找经验, 继续努力前行。并且不要忘记, 你永远不是孤身一人, 你身边有很多支持你的人。	
来访者: 谢谢你, 现在觉得好多了。我会尽快去做的。	
治疗师: 没关系的, 把握好自己的节奏, 慢慢来就好。需要的时候, 我一直在你的身边。	

Table 13: The Chinese case comparison of Psy-insight and Smile dataset on the similar topic: Distress. Table 16 shows its English translation.

Psy-Insight	Smile(Qiu et al., 2023)
Dialog ID	
1	2
Experts' Assessments	
<u>专业咨询师A</u> : 感觉1更能调动求助者的求助动机。另外，咨询师只能评估来访者的状态，不会直接说你可能焦虑或抑郁了。但2后面说“你永远不是孤身一人”“需要的时候，我一直在你的身边”比较温暖。	
<u>专业咨询师B</u> : 对话1更好，对话2首次对话中仅由“心情低落，无法集中精力”就回应求助者“可能是焦虑或者抑郁”，有点过于草率，以及“真正勇敢的人会主动接受或寻求帮助”这句话有点把不敢正视心理问题归因于“不够勇敢”，不如对话1中承认一些负面情绪是自然的这样的表述好。	
<u>心理系学生C</u> : 对话1更好，对话2更加自然，因为对话一探讨问题的角度更加深入，更像心理咨询，而对话2更像是日常放松状态下人们的对话。	
<u>心理系学生D</u> : 1更好，1有共情，并给出了具体建议，不过1的建议有些模式化口号化，感觉有效性有限。2虽然也有建议，但是ai味较重，感觉不是特别真诚。1的对话比较自然。但是对于一个求助者来说，1可能也不足够亲近。	
<u>心理系学生E</u> : 1更好。虽然两个都有一点说教的生硬，但是1的安慰相对更多，注意肯定了来访的行为，而不是像2的祈使句命令提建议。	
<u>心理系学生F</u> : 2更好，1太夸张了。	
<u>心理系学生G</u> : 对话1更好，支持者对求助者的话会给出总结，并且很明显地在引导对话，对话2里说求助者可能是焦虑或抑郁有些过于直接和冰冷了。对话1更自然。	
Translation of Assessments	
<u>Professional Counselor A</u> : Dialogue 1 feels more capable of motivating the seeker's help-seeking motivation. Additionally, counselors can only assess the state of the visitor and will not directly say that you may be anxious or depressed. However, Dialogue 2 later says "You are never alone" and "I am always by your side when you need me," which feels warmer.	
<u>Professional Counselor B</u> : Dialogue 1 is better. In the first dialogue of Dialogue 2, the response to the seeker's statement "feeling low and unable to concentrate" is "you may be anxious or depressed," which seems a bit hasty. Also, the phrase "truly brave people will actively accept or seek help" somewhat attributes the avoidance of mental issues to "not being brave enough." This is not as good as the acknowledgment in Dialogue 1 that some negative emotions are natural.	
<u>Psychology Student C</u> : Dialogue 1 is better. Dialogue 2 feels more natural because Dialogue 1 explores the problem more deeply, resembling psychological counseling, while Dialogue 2 feels more like casual conversations in a relaxed state.	
<u>Psychology Student D</u> : Dialogue 1 is better. It shows empathy and provides specific advice, although some of the advice in Dialogue 1 is somewhat formulaic and slogan-like, feeling limited in effectiveness. Dialogue 2, while also offering advice, has a stronger AI flavor, feeling less sincere. Dialogue 1's conversation is more natural. However, for a seeker, Dialogue 1 may not be close enough.	
<u>Psychology Student E</u> : Dialogue 1 is better. Although both have a bit of didactic stiffness, Dialogue 1 offers more comfort, paying attention to and affirming the visitor's behavior, unlike Dialogue 2's imperative sentences commanding suggestions.	
<u>Psychology Student F</u> : Dialogue 2 is better. Dialogue 1 is too exaggerated.	
<u>Psychology Student G</u> : Dialogue 1 is better. The supporter in Dialogue 1 provides summaries of the seeker's words and clearly guides the conversation. Saying in Dialogue 2 that the seeker may be anxious or depressed is too direct and cold. Dialogue 1 is more natural.	

Table 14: The expert assessments of human counseling and AI-synthesized counseling. AI-generated counseling dialogues often lack empathy, interaction, and focus too much on problem-solving.

G Translation of Previous Cases

835

Chat Stage	Case ID	Psychotherapy Guide
The 4th Session	23	Postmodern Therapy
Background	Strategy / Emo-label	Reasoning / Observation
The visitor is a young woman facing challenges related to family and intimate relationships, seeking support and understanding in emotional cognition and complexity. She may be seeking solutions to dilemmas in her family and marriage, hoping for clearer insight into her emotions and choices.	Question	The therapist needs to help her separate issues through gradual questioning, alleviate negative emotions, and clarify the impact of these issues to enhance her engagement and coping abilities in life.
Dialogue In Psy-Insight	Strategy / Emo-label	Reasoning / Observation
Therapist: You're willing to challenge your fears and be more open in expressing yourself, but you still prefer to take it step by step, right?	Neutral Question	The therapist guides the visitor to clarify issues through gradual questioning.
Client: Of course.	None	
Therapist: Where did you learn how to be a woman, wife, and mother?	Exploring the visitor's gender role perceptions through questioning.	
Client: I don't know. I've never thought about that question.	Neutral	The visitor lacks reflection on this topic.
Therapist: Well, I'm curious where you learned to care for others, putting your own needs behind theirs, sometimes even sacrificing your own self to fulfill others?	Question	Further probing into the origins of the visitor's self-sacrificial and others-first concepts.
Client: (Still thinking) I guess I learned it from my parents, especially from how my father treated my mother.	Others	The visitor recognizes that her behavioral patterns and beliefs may stem from parental influence.
Therapist: Don't you think your current way of life is learning from your mother—about how to be a woman, wife, and mother?	Question	Prompting the visitor to reflect on whether she is replicating her mother's life patterns.
Client: I'm not sure. I've never thought about whether I'm repeating my mother's way of life. I know she never thought about going out to work or studying. From this perspective, there are significant differences between us.	Neutral	The visitor expresses uncertainty and points out significant differences between herself and her mother.
Therapist: Yes, that does seem like a real difference. I wonder if you can find any similarities between you and her?	Question	Acknowledging the differences, the therapist encourages the visitor to identify similarities with her mother's way of life.
Client: Well, I think my mother is very traditional. In her view, men are the leaders of the family, the primary breadwinners, and if you want, they can also be the protectors of the family. The role of women is to raise children, take care of the entire family, and I guess, including taking care of their husbands.	Neutral	The visitor describes her mother's traditional views and gender role division.
Therapist: Among the many beliefs your mother taught you, how many have you accepted?	Question	The therapist asks about the degree to which the visitor has internalized her mother's beliefs.
Summary		
As a counselor, I should guide the visitor to question, reflect, and gradually replace these unhealthy cultural beliefs, establishing a more positive self-awareness.		

Table 15: The English translation of Table 10.

Chinese Counseling Case	
Psy-Insight	Smile(Qiu et al., 2023)
Source	
Textbook& Psychological blogs	ChatGPT-Synthesis from PsyQA
Topic	
Depression	
<u>Client:</u> To be honest, I feel frustrated because I can't solve my own problems, so I need professional help.	<u>Client:</u> I've been feeling very low lately and unable to concentrate.
<u>Counselor:</u> If you didn't feel frustrated seeking outside help, I would be surprised. This is a healthy feeling. But in fact, you've taken an important step, which is a commendable progress. It must have taken great courage.	<u>Counselor:</u> It sounds like you might be anxious or depressed. Have you considered seeking help from a professional?
<u>Client:</u> Well, I never thought of that as a big step.	
<u>Counselor:</u> Now that you're here, do you believe that counseling can help you? Do you believe that counseling can improve your life in a short time?	<u>Client:</u> I'm actually afraid of seeing a psychologist, always felt it's a place for crazy people.
<u>Client:</u> I came here because I believe it can help me.	<u>Counselor:</u> Truly brave people will actively accept and seek help, this applies to mental health as well. It's nothing to be ashamed of.
<u>Counselor:</u> I understand your past, and we've had a period of conversation, I believe you'll feel better. I won't make any promises, but I think improving your life is not impossible. I think this way not based on a meaningless wish, but based on the four key information you've given: you've taken an important step - you're here, you somewhat believe your life will improve. In addition, you've set a goal for yourself: to get a new life. Finally, you can maintain an open attitude towards your pain. In other words, you can describe your pain.	<u>Client:</u> But I really don't know what to do, I feel like I'm stuck in a rut.
<u>Client:</u> So there's still hope for me.	<u>Counselor:</u> Don't worry, we'll figure out a way together. You can try some exercise or learn new knowledge to distract yourself. We can also seek help from professionals.
<u>Counselor:</u> Yes, there is indeed hope. I also believe you'll feel better soon, with just one condition.	<u>Client:</u> You're right, I'll try to do that. But I always feel that if I fail and fall into a low again, I'll truly have nothing left.
<u>Client:</u> What condition?	<u>Counselor:</u> Don't be afraid of failure, everyone has had unsatisfactory experiences. But you can find lessons from failure and continue to strive forward. And don't forget, you're never alone, there are many people around you who support you.
<u>Counselor:</u> That is, you're willing to exert effort, even when facing very difficult tasks. This difficult task sounds strange, that is, you should try not to struggle so hard with your pain.	<u>Client:</u> Thank you, I feel much better now. I'll do it as soon as possible.
<u>Client:</u> I'm willing to try.	<u>Counselor:</u> It's okay, take your own pace, take it slow. I'm always here when you need me.

Table 16: The English translation of Table 14.

Guideline For human evaluation

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, and 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.

Interaction – As the patient, are you willing to continue the consultation? ...

e.g.

(Low interaction) The response is long and wordy or general.

(High interaction) The AI asks questions that facilitate the conversation/gives targeted suggestions.

Helpfulness – If you were the patient, would this response solve your problem?

e.g.

(High helpfulness) Responses to patients' questions that provide targeted suggestions.

Comforts – If you put yourself in the patient's shoes, would this response make you feel comforted?

e.g.

(High comfort) The reply expressed sympathy

(Low comfort) The reply asked a stupid question

Explainability – Put yourself in the shoes of a consultant and pay attention to judging from the consultant's perspective whether the model's explanation of the response is understandable? Can you understand the intention behind the model's generated response?

e.g.

(Interpretability) Combined with the explanation information generated by the partial model and the conversation context, can you understand how this response contributes to the consultation conversation?

Table 17: Evaluation guideline for rater.