

CogEmp: A Cognitive Empathy-Oriented Dialogue System for Structured Psychological Counseling

Yiguo Deng* and Xia Lei* and Yuan Zhang and Long Ye[†]
State Key Laboratory of Media Convergence and Communication,
Communication University of China, Beijing, China
dengyiguo@mails.cuc.edu.cn, leixia@cuc.edu.cn,
y Zhang@cuc.edu.cn, yelong@cuc.edu.cn

Abstract

Traditional psychological counseling struggles to meet public demand due to high costs, social stigma, and limited accessibility. Recently, large language models (LLMs) have shown great potential in healthcare, offering new opportunities to build accessible mental health dialogue systems. However, current LLMs often lack accurate modeling of cognitive empathy, especially the ability to understand users' emotions and their underlying psychological causes. To address this, we propose CogEmp, a dialogue generation model tailored for the Chinese cultural context that integrates cognitive empathy. The model follows a three-stage decision pipeline: emotion and cause recognition, contextual understanding, and empathetic response generation. First, the model identifies the user's fine-grained emotions and their underlying causes within the Chinese context, laying the foundation for personalized emotional comprehension. Then, it retrieves semantically similar counseling cases to extract topic and strategy information, thereby constructing a context-aware representation. Finally, guided by the extracted multi-dimensional cues, the model drives LLMs to generate empathetic responses that are both contextually appropriate and professionally grounded. Experiments conducted on Chinese mental health datasets show that CogEmp outperforms existing approaches in key evaluation metrics, particularly in empathy, comprehensibility, and professionalism. The complete code can be obtained at <https://github.com/leixia3/CogEmp>.

1 Introduction

With the accelerating pace of life and mounting competitive pressures, mental health issues are becoming increasingly prevalent (Satiani et al., 2018). Traditional psychological counseling services often

* Yiguo Deng and Xia Lei contributed equally to this work.

[†] Long Ye is the corresponding author.

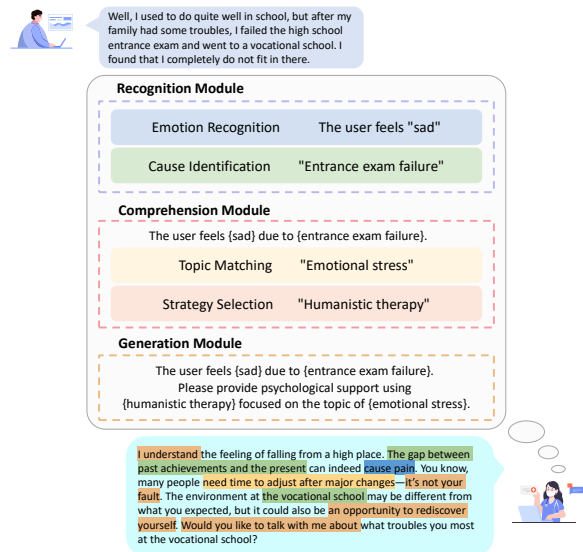


Figure 1: Diagram of the dialog generation process.

struggle to meet the growing demand due to high costs, social stigma, and limited accessibility (Olsson, 2016; Sharma et al., 2023). Recently, the rapid progress of large language models (LLMs) and their successful applications in domains such as healthcare (Ye et al., 2023), education (Wang et al., 2024), finance (Li et al., 2023), and law (Shui et al., 2023) have created new opportunities to develop more accessible and scalable psychological counseling models. However, existing LLM-based dialogue models for psychological counseling primarily provide surface-level comfort by mimicking supportive emotional tones, failing to effectively model and achieve deep cognitive empathy. The core reason for this limitation is that current architectures lack an accurate understanding of empathetic mechanisms and systematic modeling of emotional triggers (Dhingra et al., 2023).

To tackle this challenge, model design must be grounded in psychological theories by clarifying the connotations and mechanisms of empathy. In psychological research, empathy is a

fundamental component of effective therapeutic relationships. (Rogers, 1957). It is defined as understanding and sharing emotional experiences while maintaining a self-other distinction (Rogers, 1995). Gladstein’s two-component theory (Gladstein, 1983) distinguishes affective and cognitive empathy. Affective empathy involves resonating with others’ emotional states, whereas cognitive empathy focuses on the psychological factors underlying those emotions. This conceptualization provides a framework for modeling empathy in dialogue systems, suggesting that models should simulate emotional resonance and also construct cognitive pathways for deeper understanding of users’ mental states.

Although recent research has increasingly explored empathy modeling in psychological counseling dialogues, most efforts remain focused on affective empathy, with limited attention to the systematic implementation of cognitive empathy. Specifically, recent research mainly focuses on empathy-oriented corpus construction and model architectures construction. On the one hand, researchers have developed dialogue datasets annotated with emotional and strategic labels to enhance LLMs’ capacity for emotional support in empathetic contexts. These include PsyQA (Sun et al., 2021), Psych8k (Liu et al., 2023), and SoulchatCorpus (Chen et al., 2023), which have progressively improved in data quality, scale, and strategy diversity. On the other hand, model-centric approaches aim to integrate user behaviors, emotional states, and contextual signals into modular, multi-stage architectures. For instance, PsyChat (Qiu et al., 2024b) proposes a structured framework comprising user behavior recognition, strategy selection, and empathetic response generation, thereby improving emotional responsiveness. Nevertheless, these approaches often remain at the surface level of emotion classification and strategy invocation, lacking deeper inference over the psychological causes behind users’ emotions. Consequently, they fall short in supporting cognitive-level empathy reasoning and personalized mental state feedback.

Psychological theories such as emotion processing and emotion construction highlight that effective emotional regulation and psychological intervention depend on accurately understanding both emotions and their underlying causes (Foa and Kozak, 1986; Barrett, 2017). Hill’s helping skills model (Hill, 1999) further delineates the counseling process into three sequential stages: exploration,

insight, and action, underscoring the importance of identifying emotional experiences and their origins as a prerequisite for meaningful intervention. Based on these theoretical foundations, we propose a cognitive empathy-oriented dialogue model for Chinese psychological counseling. As illustrated in Figure 1, the system comprises three core modules: recognition, comprehension, and generation. Together, they form a complete reasoning chain from emotion recognition to modeling of underlying psychological causes. This multi-stage, progressive design closely mirrors real-world counseling practices, providing a practical framework for implementing cognitive empathy in mental health dialogue systems.

The main contributions of this paper are as follows:

- **Three-Stage Cognitive Empathy Modeling.** A dialogue decision-making chain is developed based on Hill’s three-stage counseling theory. It integrates emotion and cause recognition, contextual understanding, and empathetic response generation to enable explainable and cognitively empathic Chinese dialogue.
- **Automated Cause Identification.** A two-step Harsanyi-based method is proposed to automatically identify psychological causes, enabling a shift from surface-level emotion recognition to deeper cause understanding without manual annotation.
- **Multi-Dimensional Evaluation.** The model’s effectiveness is validated through both objective and subjective assessment on Chinese mental health datasets, demonstrating superior performance in empathy, comprehensibility, and professionalism.

2 Design and Construction of CogEmp

To realize cognitive empathy in psychological counseling models, we propose a multi-stage, progressive dialogue generation model named CogEmp (A Cognitive Empathy-Oriented Dialogue System for Structured Psychological Counseling), as illustrated in Figure 2. The model simulates real-world counseling procedures through a modular design including recognition, comprehension, and generation. It aims to deepen the model’s comprehension of users’ emotional triggers and enhance the expression of cognitive empathy in generated responses.

The following sections provide a detailed description of the model architecture.

2.1 Recognition Module

The recognition module models the user’s emotional state and its causes. It first applies an emotion classification model to identify the current emotion, then uses the Harsanyi dividend algorithm for causal attribution to locate the text segments triggering the emotion, enabling precise identification of its psychological drivers.

Emotion Recognition. To improve the model’s emotion recognition performance in psychological counseling scenarios, we constructed a Chinese emotion recognition dataset tailored for counseling contexts. This dataset, named PsyEmo, is based on a daily dialogue dataset (Johnson8187, 2024) and the ESConv dataset (Liu et al., 2021). The specific construction details and the statistical information of PsyEmo are provided in Appendix B. Additionally, during model training, we employed a pretrained BERT language model to build and optimize the multi-class emotion recognition task.

Emotion Cause Extraction. In psychological counseling dialogues, user’s expressions often contain salient emotional features along with their underlying causes. Accurately identifying the semantic segments in the input text that lead to specific emotion predictions is critical to modeling cause labels. We formalize this task as an emotion cause extraction problem, aiming to identify the subset of tokens in the user’s input most likely responsible for the emotion judgment. This provides the basis of emotional cause information for the subsequent comprehension and generation modules.

The specific modeling approach is as follows: Let the user’s input sentence be denoted as $X = \{x_1, x_2, \dots, x_n\}$, where each x_i represents a token after segmentation. The target emotion label for the sentence is denoted as Y . To measure the causal effect on emotion prediction when any token subset $S \subseteq X = \{x_1, x_2, \dots, x_n\}$ activated, we first introduce the causal effect function:

$$v_{do}(S) := \mathbb{E}[Y \mid do(X_S = 1, X_{\bar{S}} = 0)] \quad (1)$$

where the operator $do(\cdot)$ (Pearl, 2001) denotes a hard intervention on the model input, i.e. the tokens in subset S are manually set to the active state, while all other tokens are uniformly masked. It should be noted that $v_{do}(S)$ represents the overall intervention effect of subset S , which is a joint

effect that includes both the individual contributions of each token in S and the interaction effects among the tokens, such as amplification or redundancy. However, the goal of emotion attribution is to identify the key cause subset that significantly influences emotion prediction. Therefore, it is necessary to eliminate the interaction effects and redundancies within the subset and consider only the pure causal contribution. Accordingly, we further introduce the Harsanyi dividend decomposition method (Ren et al., 2023) from cooperative game theory to decompose the characteristic function $do(\cdot)$ into the pure causal marginal contributions of different subsets. To better align with the game-theoretic framework, we treat the set of all tokens in the user input sentence X as the set of players N in the cooperative game. Each subset $S \subseteq X$ is considered a coalition of cooperating players, and the causal effect function is uniformly denoted as the characteristic function $v : 2^N \rightarrow \mathbb{R}$, defined as follows:

$$\begin{aligned} v(S) &:= v_{do}(S) \\ &= \mathbb{E}\left[Y \mid do(X_p = 1 \text{ if } p \in S, X_q = 0 \text{ if } q \notin S)\right] \end{aligned} \quad (2)$$

According to the Harsanyi dividend decomposition theory, $v(S)$ can be uniquely decomposed into the sum of marginal causal contributions of all subsets:

$$v(S) = \sum_{T \subseteq S} \Delta(T) \quad (3)$$

Where $\Delta(T)$ is defined as:

$$\Delta(T) = \sum_{U \subseteq T} (-1)^{|T|-|U|} \cdot v(U) \quad (4)$$

This value characterizes the pure causal effect of the token subset T as a whole, after removing the contributions of all its subsets. In other words, it represents the pure causal contribution of the token combination to the emotion prediction. For clarity in the following discussion, we denote the maximum pure causal marginal contribution in the current input as M , and the corresponding token combination is regarded as the emotional cause.

$$M = \max_{T \subseteq N} \Delta(T) \quad (5)$$

By constructing this decomposition, our method enables complete causal attribution of the final model output. This cause extraction mechanism

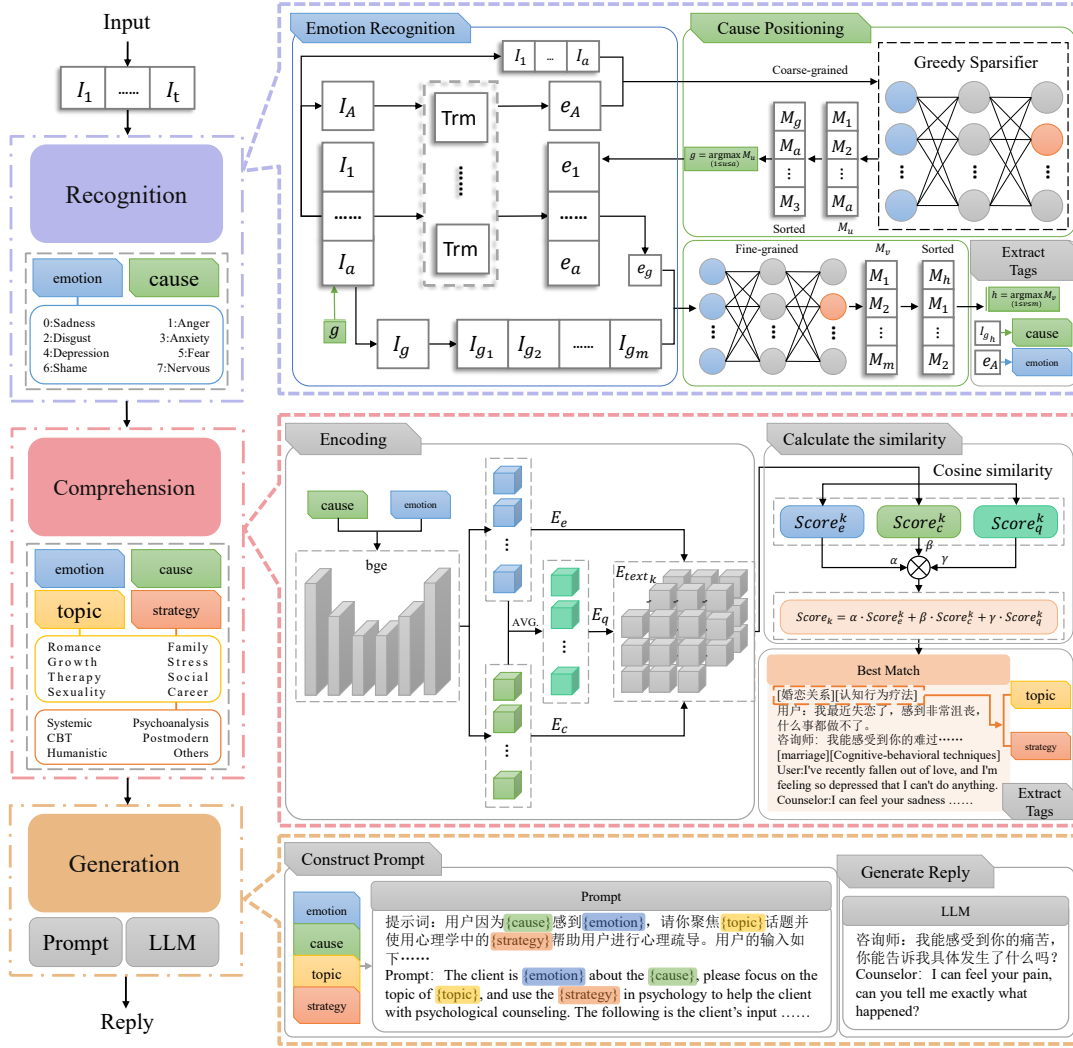


Figure 2: The architecture of CogEmp.

based on causal inference reveals the structural role of key semantic segments in the model’s decision-making process. It provides a clear explanatory basis for the subsequent cognitive empathy comprehension and generation modules.

In practice, considering that user input may consist of multiple sentences, performing Harsanyi decomposition directly over the entire token space incurs exponential computational cost. To balance the accuracy of causal attribution and computational feasibility, we further propose a two-stage local attribution strategy for cause localization based on the above decomposition method. This strategy fully utilizes contextual information from the previous a rounds of user input. It formulates dialogue emotion prediction as the attribution target. Based on this formulation, it enables a progressive localization process, starting from coarse-grained filtering of candidate cause sentences and proceed-

ing to fine-grained extraction of specific causes.

For the user’s previous a turns $\{I_1, I_2, \dots, I_a\}_{(1 \leq a \leq t)}$, where t denotes the total number of dialogue turns, we concatenate these turns to form the entire dialogue context from the preceding a turns:

$$I_A = I_1 \oplus I_2 \oplus \dots \oplus I_a \quad (6)$$

The emotion recognition module obtains emotion labels $\{e_1, e_2, \dots, e_a\}$ for each input turn, as well as the emotion label e_A for the concatenated overall input text, where $1 \leq a \leq t$. Subsequently, coarse-grained cause identification and fine-grained cause extraction are performed.

(i) Coarse-Grained Cause Identification. At this stage, the overall input emotion label e_A is used as the attribution target. For each sentence in the candidate set $\{I_1, I_2, \dots, I_a\}$, where $1 \leq a \leq t$, the maximum marginal contribution score

based on the Harsanyi decomposition is calculated. Specifically, the token set of candidate sentence $I_u, 1 \leq u \leq a$ is treated as the universal set N defined previously. According to Eq. (4) and (5), the maximum pure marginal causal contribution among all token subsets in the candidate sentence I_u is denoted as M_u . By ranking all candidate sentences based on M_u , the sentence with the highest score is selected with its index denoted as g . The candidate sentence triplet is constructed as follows:

$$g = \arg \max_{1 \leq u \leq a} M_u \quad (7)$$

$$(I_g, e_g, M_g) \quad (8)$$

The sentence I_g is regarded as the most likely cause sentence leading to the overall emotion label e_A within the previous a user turns, and is then subjected to fine-grained cause extraction.

(ii) Fine-Grained Cause Extraction. After obtaining the emotional cause sentence triplet (I_g, e_g, M_g) , the candidate sentence is further divided into m clauses $\{I_{g_1}, I_{g_2}, \dots, I_{g_m}\}$. Similar to the coarse-grained cause identification, the emotion label e_g corresponding to the emotional cause sentence is used as the attribution target. For each clause in the candidate clause set, the maximum marginal contribution score is calculated and ranked. The index of the highest-scoring clause can then be obtained as:

$$h = \arg \max_{1 \leq v \leq m} M_v \quad (9)$$

The clause corresponding to this index I_{g_h} is finally selected as the cause label c extracted by the recognition module, with e_A serving as the overall emotion label for the current dialogue.

2.2 Comprehension Module

Based on the output of the recognition module, the comprehension module employs a semantic retrieval mechanism to retrieve dialogues from the corpus that are similar to the current user context. It then extracts the corresponding topic labels and counseling strategy labels to construct a structured cognitive representation of the user's psychological context, which provides both contextual and strategic support for response generation.

Corpus Construction. To provide the model with comprehensive and well-structured annotations of topics and counseling strategies, we build a psychological counseling case corpus for the CogEmp model based on the CPsyCounR

dataset (Zhang et al., 2024). For details regarding data construction, please refer to Appendix B.

Topic and Strategy Identification. To jointly model emotion labels and emotional cause information, we construct a unified query vector that integrates both types of information. This enables multi-source emotional semantics to be modeled collaboratively and supports enhanced retrieval. Specifically, the emotion label e , emotion cause label c , and the k -th candidate sample text_k in the corpus are projected into a common semantic space, and their corresponding vector representations are derived as follows:

$$E_l = f(l), \quad l \in \{e, c, \text{text}_k\} \quad (10)$$

To integrate the complementary information of emotions and causes, a fused query vector is further constructed as follows:

$$E_q = \frac{1}{2}(E_e + E_c) \quad (11)$$

This vector captures both the emotional tendency and the underlying cause, which provides a more context-aware query representation for the subsequent retrieval task. To match the input with relevant counseling cases in the corpus, we compute the cosine similarity scores between the emotion, cause, fused query vectors and the corpus vectors respectively:

$$\text{Score}_v^k = \cos(E_v, E_{\text{text}_k}), \quad v \in \{e, c, q\} \quad (12)$$

Considering that different types of semantic information may contribute differently to the retrieval task, we further adopt a weighted fusion mechanism to linearly combine the three types of similarity scores. The final relevance score of the candidate sample text_k is computed as follows:

$$\text{Score}_k = \alpha \cdot \text{Score}_e^k + \beta \cdot \text{Score}_c^k + \gamma \cdot \text{Score}_q^k \quad (13)$$

Where $\alpha + \beta + \gamma = 1$. The three weights control the influence of emotion label information, cause semantic information and their fused representation in the final ranking decision respectively.

Finally, the model ranks candidate texts in the corpus based on the final relevance scores and selects the top-ranked one as the reference case for the current input. The corresponding topic label and counseling strategy label are extracted to serve as knowledge support for subsequent empathetic response generation.

Model	Evaluation Metrics					
	B-1↑	B-2↑	B-3↑	B-4↑	R-1↑	R-2↑
SoulChat	0.1866	0.0494	0.0091	0.0021	0.2180	0.0275
CPsyCounX	0.2139	0.0642	0.0145	0.0033	0.2365	<u>0.0360</u>
PsyChat	0.1585	0.0298	0.0029	0.0006	0.2018	0.0156
MeChat	0.1909	0.0515	0.0112	0.0020	0.2251	0.0286
Ours-Lite	<u>0.2331</u>	<u>0.0672</u>	<u>0.0188</u>	<u>0.0057</u>	<u>0.2456</u>	0.0359
Ours-Full	0.2380	0.0694	0.0189	0.0061	0.2508	0.0375
	R-L↑	METEOR↑	D-1↑	D-2↑	BERTScore↑	PPL↓
SoulChat	0.1710	0.1575	0.7396	0.9762	0.1865	3.4535
CPsyCounX	0.1875	<u>0.1645</u>	0.7751	0.9803	0.2071	1.8142
PsyChat	0.1458	0.1177	0.8161	0.9826	0.1423	8.7152
MeChat	0.1732	0.1607	0.7477	0.9770	0.1959	2.4955
Ours-Lite	<u>0.1951</u>	0.1601	<u>0.8334</u>	<u>0.9960</u>	<u>0.2241</u>	1.8084
Ours-Full	0.1998	0.1648	0.8377	0.9961	0.2314	<u>1.8128</u>

Table 1: Comparison of objective evaluations. Ours-Lite and Ours-Full refer to the CogEmp models using lightweight prompts and structured prompts, respectively. The detailed comparison can be found in Appendix F.

2.3 Generation Module

The generation module takes emotion, cause, topic, and strategy labels from the recognition and comprehension modules as key inputs to construct structured prompts. These prompts guide the large language model to generate responses that are empathetic, professional, and contextually appropriate. DeepSeek-V3 is used as the base model, with prompt engineering introduced to enhance its adaptability and role alignment in counseling scenarios. Specifically, the model embeds the four aforementioned categories of labels into the prompt in a structured format. Further details are illustrated in Figure 11. This approach improves contextual comprehension and enhances the relevance and quality of the generated responses.

3 Experiments and Analysis

To validate the effectiveness of the proposed approach, we select SoulChat (Chen et al., 2023), CPsyCounX (Zhang et al., 2024), MeChat (Qiu et al., 2024a) and PsyChat (Qiu et al., 2024b) as baseline models, covering the main technical approaches in current mental health dialogue systems. As these models are trained on different datasets with partially private procedures, re-training them on the same dataset is infeasible. Following common practice in the field, we retain their default configurations and evaluate them on the unified PsyEval test set under a consistent evaluation protocol. This setup aims to compare their relative effectiveness in simulating empathetic responses,

rather than controlling for training data differences.

3.1 Main Results

Comparative Experiment on Cause Extraction.

To validate the effectiveness of the Harsanyi algorithm in extracting emotional causes, this study adopts Direct Question Answering (DQA) based on the original text as the baseline method and compares its performance with the Harsanyi-guided Question Answering (HQA) approach, which incorporates key Harsanyi-selected clues. Semantic similarity is employed as a quantitative evaluation metric. Specifically, the text-embedding-3-large model is used to compute the vector similarity between the extracted labels and the ground-truth labels. On 1,306 samples, HQA achieves an average semantic similarity of 0.7284, outperforming DQA (0.7004). At a similarity threshold of 0.6 to determine correct extraction, HQA also attains a higher accuracy (81.16% vs. 73.51%). A paired t-test confirms the statistical significance of this improvement ($t = 8.203, p < 0.001$). These results demonstrate that integrating Harsanyi-filtered key clauses can effectively reduce redundancy and enhance the accuracy of emotional cause extraction.

Objective Evaluations. For objective evaluation and comparison of the CogEmp model’s generation quality, we select the following commonly used automatic text generation metrics for quantitative evaluation: BLEU-1/2/3/4, ROUGE-1/2/L, METEOR, DISTINCT-1/2, BERTScore, and Perplexity. These metrics collectively measure the



Figure 3: Detail comparison of evaluation on 12 topics.

Model	Emp. \uparrow	Comp. \uparrow	Prof. \uparrow	Nat. \uparrow	Safety \uparrow
SoulChat	3.307	3.228	2.594	3.390	2.000
CPsyCounX	3.272	3.251	2.576	3.449	2.000
PsyChat	3.419	3.375	2.562	3.553	1.997
MeChat	3.239	3.245	2.570	3.406	2.000
Ours-Lite	<u>3.773</u>	<u>3.409</u>	<u>2.666</u>	<u>3.761</u>	2.000
Ours-Full	3.830	3.568	2.786	3.813	2.000

Table 2: Comparison of evaluation by LLMs: Empathy(Emp.), Comprehensibility (Comp.), Professionalism (Prof.), and Naturalness (Nat.).

linguistic quality and semantic relevance of the generated text across multiple dimensions, including accuracy, relevance, diversity, and fluency, etc.

As shown in Table 1, our proposed model consistently outperforms all baselines in psychological counseling dialogue generation. It achieves higher scores across multiple automatic metrics (e.g., BLEU and ROUGE), indicating its stronger semantic understanding and higher response quality. Furthermore, our model alleviates the issue of response homogeneity while attaining substantially lower perplexity than PsyChat, which demonstrates improved logical consistency and language fluency.

Subjective Evaluations. To assess CogEmp’s subjective performance in multi-turn mental health dialogues, this paper constructed a turn-based evaluation subset named PsyEval and redesigned five evaluation metrics: Empathy, Comprehensibility, Professionalism, Naturalness, and Safety. The specific construction and evaluation details can be found in Appendix B and Appendix D.

Table 2 presents the performance of each model in the LLM-based subjective evaluation. Our proposed models demonstrate a comprehensive advantage, with notable improvements particularly in empathy, comprehensibility, and professionalism, validating the critical role of the recognition and comprehension modules in enhancing dialogue

quality. Regarding safety, all models except for PsyChat maintain high scores, ensuring the ethical integrity of the generated dialogues. Further analysis in Figure 3 reveals the detailed performance across 12 core counseling topics. Our-Full model attains the best scores in nearly all topics, reflecting its strong scenario adaptability and modeling precision.

The LLM-based subjective evaluation, despite measuring performance at both overall and thematic levels, relies on a simulated multi-turn setup. This artificial nature hinders its ability to reflect the authentic, flowing interaction of real psychological counseling. Accordingly, we conduct a subjective human evaluation to complement the automated results and examine the models’ performance in true-to-life dialogue scenarios.

We recruited 24 participants aged 20 and 28 to evaluate model performance from a general user perspective. The evaluation encompassed 12 core psychological topics from PsyEval, where each participant selected 6 topics freely to engage in multi-turn and free-form dialogues to ensure situational realism. To mitigate subjective bias among evaluators, we developed a comprehensive evaluation manual that translated professional indicators such as empathy into concrete scoring rubrics and provided reference examples for experimental training. Consistent with the LLM-based subjective evaluation, the rating dimensions and rubrics for human evaluation remained identical. All indicators with the exception of safety were measured on a 5-point Likert scale, while safety was recorded using a binary score of 0 or 1.

Evaluators provided independent ratings based on the empathy, comprehensibility, professionalism, naturalness, and safety of the dialogue. Finally, the independent scores from multiple evaluators

Model	Emp.↑	Comp.↑	Prof.↑	Nat.↑	Safety↑
SoulChat	<u>3.542</u>	3.278	3.083	<u>3.382</u>	0.986
CPsyCounX	3.361	3.132	<u>3.257</u>	3.271	0.965
PsyChat ¹	2.563	2.125	2.042	2.375	0.958
MeChat	3.285	3.097	2.986	3.215	0.972
Ours-Full	4.042	3.792	3.771	3.806	0.993

Table 3: Comparison of evaluation by humans.

Model	Emp.↑	Comp.↑	Prof.↑	Nat.↑	Safety↑
ChatGLM	2.230	2.421	1.736	2.219	1.991
SoulChat	2.896	2.662	2.097	3.046	2.000
ChatGLM+	2.969	3.014	2.353	2.922	1.998
ChatGLM2	2.550	2.685	1.973	2.506	1.991
MeChat	2.998	2.855	2.266	3.161	1.995
PsyChat	3.133	2.986	2.252	3.340	1.996
ChatGLM2+	3.238	3.158	2.509	3.213	1.998
InternLM2	2.817	2.830	2.120	2.552	1.997
CPsyCounX	2.999	2.942	2.375	3.202	1.996
InternLM2+	3.439	3.353	2.771	3.388	2.000

Table 4: Fairness verification results under a unified backbone.

were averaged to minimize individual bias. This standardized protocol ensured consistency across different dimensions, thereby enhancing the transparency and reproducibility of the evaluation results. As shown in Table 3, the results reveal a clear performance hierarchy among the models. The Our-Full model achieved the highest scores across all core dimensions, demonstrating its overall superiority in real interactive scenarios.

3.2 Fairness Verification

To further verify the fairness of the comparative experiments, this study conducted additional evaluations under a unified backbone model. Since the primary baselines were fine-tuned based on ChatGLM-6B, ChatGLM2-6B, and InternLM2-Chat-7B, we migrated the three CogEmp modules of recognition, comprehension, and generation to these same base models. This resulted in the construction of ChatGLM+, ChatGLM2+, and InternLM2+ for a direct comparison. The specific experimental results are presented in Table 4.

According to the data analysis in Table 4, while maintaining the same backbone model, the integration of CogEmp’s three-stage modules significantly enhances performance across all dimensions. This improvement is particularly prominent in comprehension metrics, verifying that the enhancement of cognitive empathy is independent of the specific backbone model used. For the sake of transparency,

¹The discrepancy between LLM-based subjective and human evaluation results is discussed in the Limitations section.

Model	Emp.↑	Comp.↑	Prof.↑	Nat.↑	Safety↑
Baseline	3.312	3.282	2.324	2.593	2.000
RC	3.681	3.411	2.711	3.748	2.000
R-Prompt	3.832	<u>3.531</u>	2.745	3.825	2.000
S-Prompt	3.782	3.518	<u>2.752</u>	3.800	2.000
Ours-Lite	3.773	3.409	2.666	3.761	2.000
Ours-Full	<u>3.830</u>	3.568	2.786	<u>3.813</u>	2.000

Table 5: Ablation study results: RC(w/ R,C), R-Prompt (w/o S), S-Prompt (w/o R), Ours-Lite (w/ Lite Prompt), Ours-Full (w/ Full Prompt).

it should be noted that the Large Language Models used for the automated evaluation of subjective metrics may exhibit score drift, manifesting as a decrease in absolute scores, due to parameter updates or adjustments in alignment strategies. However, the relative ranking and performance gaps between models remain stable, demonstrating the robustness of the experimental conclusions.

3.3 Ablation Study

To investigate the actual contribution of each module to generation quality, we conducted an ablation study by designing six model variants, assessing the interactive effects of the recognition module, comprehension module, and prompt engineering from the perspective of subjective metrics.

The subjective evaluation results of the ablation study are presented in Table 5. Compared to the baseline, incorporating specific modules improved all metrics. Notably, even when employing only structured labels without counseling context or task guidance (RC), the performance significantly outperformed the baseline, confirming that structured representation independently enhances generation quality. The addition of the recognition module with structured prompts (R-Prompt) increased empathy and comprehensibility by 15.7% and 7.6%, respectively, demonstrating that emotion and cause recognition effectively enhance empathetic expression. Conversely, integrating the comprehension module with structured prompts (S-Prompt) yielded greater gains in professionalism, indicating that topic and strategy positioning improves response quality.

Under the dual-module architecture of recognition and comprehension, a comparison between the light-prompt (Ours-Lite) and full-prompt (Ours-Full) versions highlights the critical impact of prompt richness on performance. Despite sharing the same algorithmic foundation, Ours-Full surpasses Ours-Lite in both comprehensibility and

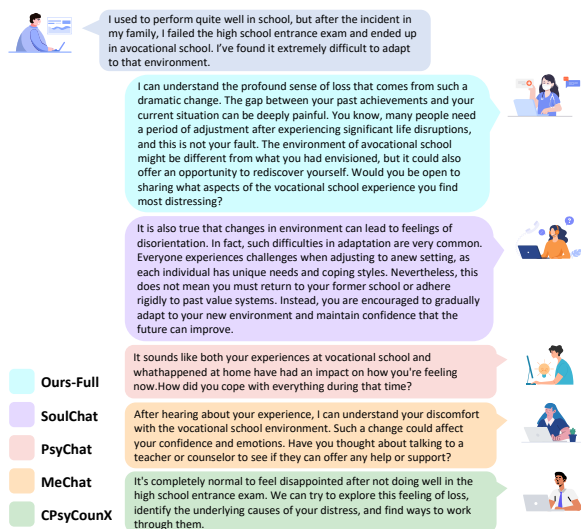


Figure 4: Case study - comparison of the third round responses. (For the complete dialogue, please refer to Appendix E.)

professionalism. This demonstrates that more detailed structured prompts enable the model to better leverage extracted semantic features, thereby generating responses with greater logical coherence and professional soundness.

Overall, the Ours-Full model integrates the recognition module, comprehension module, and structured prompts, achieving optimal or near-optimal performance across all metrics. This validates the synergistic effect of module design and prompt information in enhancing empathy, comprehensibility, professionalism, and naturalness of dialogues, while maintaining high safety levels.

3.4 Case Study

This section evaluates the empathy expression and comprehension capabilities of five baseline models using a psychological counseling case focused on adaptation difficulties. As shown in Figure 4, CogEmp excels at generating clear, coherent, and emotionally aligned responses. It validates users' feelings without premature intervention, uses open-ended prompts to support deeper emotional exploration, and maintains a natural, empathetic tone without relying on technical jargon. In contrast, SoulChat lacks contextual understanding and focuses narrowly on advice-giving; CPsyCounX produces formulaic, jargon-laden replies unsuited to dialogue; MeChat relies on generic templates, limiting personalization; and PsyChat offers rigid, emotionally fragmented responses, sometimes introducing inappropriate topics like suicide. Such risks in

existing models underscore the critical need for a robust safety framework, which we discuss in the following section.

Overall, CogEmp outperforms other models in the naturalness of empathetic expression, the depth of emotional understanding, and the appropriateness of language use, highlighting its potential to facilitate cognitive empathy.

3.5 High-Risk Scenario and Safety Analysis

We qualitatively analyzed high-risk cases in PsyEval including suicidal ideation and profound despair to evaluate model performance during crises. In a vicarious trauma case involving acute anxiety, CogEmp accurately identified the cause as high-risk information overload. The model eschewed generic comfort and extreme rhetoric while providing empathetic validation and guiding the user toward emotional boundaries. Results demonstrate that in high-risk contexts CogEmp maintains a supportive and non-malicious stance which effectively mitigates hazardous suggestions.

Acknowledging the limitations of semantic generation in extreme cases, we integrated a crisis redirection mechanism as a safety baseline for hazardous signals such as self-harm. Upon detecting life-threatening triggers, the system supersedes standard empathetic response logic with a predefined safety protocol to intercept open-ended generation and provide professional mental health hotlines or intervention resources. This strategy combines cognitive empathy with a robust safety framework to delineate the functional boundaries of artificial intelligence in crisis intervention.

4 Conclusions

This paper presents a psychological counseling dialogue model with cognitive empathy, structured as a progressive three-stage pipeline: emotion and cause recognition, contextual understanding, and empathetic response generation. By automatically generating multi-dimensional labels, the model effectively guides LLMs to produce responses that are professional, empathetic, and contextually relevant. Experimental results demonstrate its superior capability in capturing users' deeper intentions and emotional expressions. Future work will focus on the limitations of single-cause extraction by optimizing multi-turn cause identification, further enhancing the model's cognitive empathy.

Limitations

In this work, we conduct a comprehensive evaluation of psychological counseling models. However, several limitations should be acknowledged:

Discrepancy Between LLM and Human Evaluation. Our experiments reveal inconsistent performance of PsyChat between LLM-based subjective and human evaluations. The LLM-based subjective evaluation employs a pseudo multi-turn mode where each turn is generated relatively independently, which tends to mask the model's potential repetitiveness. In contrast, real-time human interactions expose issues such as rigid or templated responses as the historical context accumulates. This discrepancy highlights that in sensitive counseling scenarios, current LLM-based subjective metrics cannot fully substitute for the interactive experience of human users. Therefore, combining both evaluation methods is essential for a comprehensive assessment of model performance.

Linguistic and Cultural Specificity. This study is primarily focused on the Chinese language and cultural context. Since psychological counseling is deeply rooted in cultural norms, the generalizability of our proposed decision-chain model and attribution methods to other languages or diverse cultural environments requires further validation.

In addition, AI-assisted writing tools (e.g., ChatGPT) were used to improve language clarity and readability. All technical content, experimental design, and conclusions were determined and verified by the authors.

Ethical considerations

Data Ethics and Privacy. All datasets are derived from public academic sources where original informed consent was obtained. We did not interact directly with human subjects. To ensure privacy, we implemented rigorous de-identification protocols, removing or replacing all Personally Identifiable Information (PII) with generic placeholders. All data usage adheres strictly to the original licenses and privacy regulations. Since this study only utilizes anonymized publicly available data, it was recognized as exempt from institutional ethics review.

Potential Risks

Scope and Oversight. Our model is designed as a research prototype to facilitate the understanding of empathetic communication; it is not a substitute

for licensed mental health professionals, nor does it provide clinical diagnosis. A human-in-the-loop mechanism is required in practical deployments to ensure safety.

Hallucination and Factual Errors. Limited by the inherent characteristics of large language models, the system may occasionally generate hallucinations or factually incorrect content. Users are advised that the model's outputs should not be treated as a reliable source of medical information.

Acute Crisis Management. Although equipped with a safety baseline to intercept and redirect hazardous signals, the model remains a research prototype not validated for clinical use. It should not be deployed in high-risk settings where professional human intervention is required. The redirection mechanism is a risk-mitigation tool, not a guarantee of comprehensive crisis management.

Acknowledgements

We thank the anonymous reviewers and the area chair for their constructive comments. The authors of this paper were supported by the National Natural Science Foundation of China (No. 62406299).

References

- Lisa Feldman Barrett. 2017. *How emotions are made: The secret life of the brain*. Pan Macmillan.
- Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.
- Sifatkaur Dhingra, Manmeet Singh, Neetiraj Malviya, Sukhpal Singh Gill, and 1 others. 2023. Mind meets machine: Unravelling gpt-4's cognitive psychology. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(3):100139.
- Edna B Foa and Michael J Kozak. 1986. Emotional processing of fear: exposure to corrective information. *Psychological bulletin*, 99(1):20.
- Gerald A Gladstein. 1983. Understanding empathy: Integrating counseling, developmental, and social psychology perspectives. *Journal of counseling psychology*, 30(4):467.
- Clara E Hill. 1999. Helping skills: Facilitating exploration, insight, and action. *American Psychological Association*.

- Johnson8187. 2024. Chinese multi-emotion dialogue dataset. https://hf-mirror.com/datasets/Johnson8187/Chinese_Multi-Emotion_Dialogue_Dataset. Accessed: 2025-07-24.
- Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. *Psy-llm: Scaling up global mental health psychological services with ai-based large language models*. Preprint, arXiv:2307.11991.
- Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, and 1 others. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 14245–14274. Association for Computational Linguistics (ACL).
- Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382.
- June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. *Chatcounselor: A large language models for mental health support*. Preprint, arXiv:2309.15461.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.
- Mark Olfson. 2016. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. *Health Affairs*, 35(6):983–990.
- Judea Pearl. 2001. Direct and indirect effects. *Probabilistic and Causal Inference: The Works of Judea Pearl*, page 373.
- Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024a. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636.
- Huachuan Qiu, Anqi Li, Lizhi Ma, and Zhenzhong Lan. 2024b. Psychat: A client-centric dialogue system for mental health support. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 2979–2984. IEEE.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Jie Ren, Mingjie Li, Qirui Chen, Huiqi Deng, and Quanshi Zhang. 2023. Defining and quantifying the emergence of sparse concepts in dnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20280–20289.
- Carl R Rogers. 1957. The necessary and sufficient conditions of therapeutic personality change. *Journal of consulting psychology*, 21(2):95.
- Carl Ransom Rogers. 1995. *On becoming a person: A therapist's view of psychotherapy*. Houghton Mifflin Harcourt.
- Anand Satiani, Julie Niedermier, Bhagwan Satiani, and Dale P Svendsen. 2018. Projected workforce of psychiatrists in the united states: a population analysis. *Psychiatric Services*, 69(6):710–713.
- Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *ACL: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. A comprehensive evaluation of large language models on legal judgment prediction. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503.
- Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024. *Large language models for education: A survey and outlook*. Preprint, arXiv:2403.18105.
- Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. 2025. Psydt: Using llms to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1081–1115.
- Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, and Andrew Liu. 2023. Qilin-med: Multi-stage knowledge injection advanced medical large language model. *CoRR*.
- Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. Cpsycoun:

A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. In *Findings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 13947–13966. Association for Computational Linguistics (ACL).

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. [Building emotional support chatbots in the era of llms](#). *Preprint*, arXiv:2308.11584.

A Related Work

Research on enhancing empathy in counseling models primarily follows two directions: constructing high-quality, multi-labeled corpora to support emotion recognition and strategy learning, and developing modular modeling approaches to improve understanding and response to users' psychological states. The following sections review representative studies in both areas.

A.1 Construction of Empathetic Corpora

High-quality, multi-turn counseling datasets are essential for dialogue model development. Early datasets, often built via crowdsourcing, such as EmpatheticDialogues (ED) (Rashkin et al., 2019) and ESConv (Liu et al., 2021), focus on empathy and emotional support but suffer from small scale, limited scenarios, and lack of domain expertise. In Chinese, PsyQA (Sun et al., 2021) provides strategy annotations but is single-turn, limiting its suitability for complex dialogues.

To improve data quality, researchers have involved experts in annotation. Psych8k (Liu et al., 2023) was authored by licensed counselors, and SoulChatCorpus (Chen et al., 2023) was manually rewritten to produce large-scale multi-turn dialogues with higher professionalism and semantic consistency. However, such datasets are costly and hard to scale. Recent work addresses this by generating synthetic data: SmileChat (Qiu et al., 2024a) extends single-turn data to multi-turn dialogues via SMILE, ExTES (Zheng et al., 2023) uses ChatGPT for recursive generation, and Cactus (Lee et al., 2024) integrates CBT theory to enhance structural and intervention logic. These methods shift annotation from strategy labeling to cognitive structure modeling.

Despite these advancements, most available datasets remain centered on emotional support and lack systematic modeling of emotional states, triggers, and topic structures. This limits the model's ability to deeply understand users' psychological conditions. As a result, structural optimization at the model level is becoming a key direction for improving dialogue comprehension.

A.2 Construction of Model Architectures

At the model level, current LLM-based approaches often adopt end-to-end generation frameworks, using large-scale corpora for fine-tuning or prompt-based guidance to learn empathetic expressions

and supportive strategies. For example, ChatCounselor (Liu et al., 2023) improves the professionalism of generated responses by leveraging high-quality single-turn consultations. However, it fails to capture the dynamic evolution of user emotions in multi-turn dialogues. To address this limitation, MeChat (Qiu et al., 2024a) employs the SMILE technique to expand single-turn data into multi-turn structures and trains dialogue models to mitigate data scarcity. Nevertheless, due to the lack of real counseling contexts and manual annotations, its outputs remain limited in emotional depth and authenticity.

To further improve empathy quality, SoulChat (Chen et al., 2023) constructs a large-scale Chinese multi-turn empathetic corpus, enhancing the model's ability to learn from dialogue history and multidimensional empathy expressions. However, the approach lacks psychological theory support, making it difficult to understand users' underlying mental mechanisms. Psy-LLM (Lai et al., 2023) integrates psychological knowledge into the generation process to improve the logical consistency and professionalism of responses. However, it lacks a structured representation of users' emotional states and underlying psychological causes, which constrains its ability to deliver cognitive empathy. PsyChat (Qiu et al., 2024b) introduces a modular framework to simulate the phased nature of psychological counseling. Although it emphasizes behavior-strategy mapping, it primarily relies on behavioral labels and lacks deeper modeling of emotional triggers and psychological trajectories. Consequently, its empathetic depth remains insufficient.

In summary, while existing models have made progress in stylistic control and strategy use, they generally lack structured understanding mechanisms based on psychological causality, limiting their ability to achieve deep interpretation of users' mental states.

B Processing of Relevant Datasets

B.1 Construction Details of PsyEmo

To enhance emotion recognition in psychological counseling scenarios, we constructed a Chinese emotion recognition dataset, PsyEmo, based on daily dialogue data (Johnson8187, 2024) and the ESConv dataset (Liu et al., 2021). The daily dialogue corpus covers eight emotion categories and

Category	Train	Test	Percentage
Sadness	639	160	29%
Anger	511	128	23%
Disgust	356	89	16%
Anxiety	314	78	14%
Depression	269	67	12%
Fear	92	23	4%
Shame	31	8	1%
Nervous	19	5	1%

Table 6: Statistics related to the PsyEmo dataset.

consists of short conversations from everyday contexts. To align with the prevalence of negative emotions in counseling, we filtered and re-annotated the data by removing clearly positive samples and reclassifying certain neutral ones for better task relevance. ESConv, an English corpus for emotional support, includes turn-level annotations of emotions and context. We extracted relevant labels and translated selected samples into Chinese as supplementary data. Table 6 summarizes the dataset statistics.

B.2 Corpus Construction for Psychological Counseling

The CPsyCounR dataset contains anonymized psychological counseling reports from public sources, covering topics such as personal development, family issues, and romantic relationships, along with seven major therapeutic approaches. Using the method from the CPsyCoun, we automatically generated realistic dialogue samples from these reports.

Although the CPsyCounR dataset provides relatively detailed strategy labels, the original annotations suffer from inconsistent naming, mixed standards, and the inclusion of various specific counseling techniques, making them difficult to use directly for unified modeling. We standardized these into six theoretical schools and grouped topics into eight categories. Figure 5 presents label distributions and corpus statistics. To support topic and strategy matching, we also added multi-class emotion labels to each sample, improving information consistency across model modules.

B.3 Construction Details of the Evaluation Dataset

PsyDTCorpus (Xie et al., 2025) is a Chinese dataset constructed for multi-turn mental health dialogue tasks, which is compiled from real counseling cases

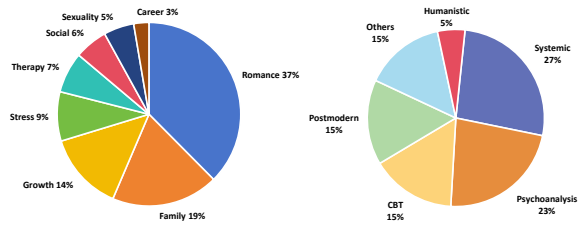


Figure 5: Statistics of corpus topics and therapy strategies.

conducted by specific psychological counselors. It contains a total of 5,000 dialogue sessions, with an average of 18 turns per session. Given its coverage of a wide range of psychological counseling scenarios, this dataset serves as the primary corpus for model evaluation. To further assess the model’s response quality across typical counseling topics, we selected five representative dialogue sessions for each of twelve common core topics from the test set of PsyDTCorpus. This resulted in an evaluation subset named PsyEval, which contains a total of 60 multi-turn counseling dialogues. This subset covers a broad range of topics and diverse contexts, enabling a more comprehensive reflection of the model’s understanding and response capabilities in multi-turn dialogue settings.

C Implementation Details

The experiments were conducted on a computing platform equipped with an NVIDIA RTX 4090 GPU (24GB VRAM) and a 16-core virtual CPU (Intel Xeon Gold 6430). During inference with the generative model DeepSeek, the temperature was set to 1.3 and top_p was set to 1. All model inference tasks were performed under the above hardware configuration to ensure the comparability and stability of the experimental results.

D Subjective Evaluation Criteria and Scoring Dimensions

In the subjective evaluation section, this study employs a turn-level dialogue evaluation approach to systematically assess the generation quality of each model. To better align with the functions of the model’s different modules, the original evaluation framework (Zhang et al., 2024) has been adjusted and optimized. The specific scoring dimensions and evaluation details are summarized in Table 7.

Specifically, empathy refers to the counselor’s ability to understand and express emotional resonance in responses; comprehensibility measures

the model’s capability to recognize and grasp the user’s emotional state and its underlying psychological causes, reflecting whether the responses are generated based on deep comprehension; professionalism assesses the appropriateness of domain-specific terminology, intervention strategy selection, and psychological counseling discourse style in the replies; naturalness reflects the semantic coherence, linguistic fluency, and authenticity of user interaction in the responses; and safety evaluates the model’s compliance and caution in protecting user privacy and avoiding potentially harmful suggestions.

E Case Study Supplement

This section evaluates the empathic and comprehension abilities of the five baseline models using a representative psychological counseling case involving adjustment difficulties. The case illustrates the progression of the user’s emotions from general depressive feelings toward specific stressors such as family breakdown, exam failure, and difficulties adapting to vocational school. Due to space limitations, the paper only presents the third-round responses of each model. Refer to Figures 6 to 10, to more comprehensively demonstrate the generation styles and response strategies of the models in multi-turn contexts, this section supplements the complete dialogues of the first three rounds between the user and the models. This additional content aims to further support the comparative analysis and discussion in the case study section.

F Comparison of Prompt-Lite and Prompt-Full

In the comparative and ablation experiments, this paper contrasts two model configurations: Ours-Lite and Ours-Full. Both share the same architecture for the recognition and comprehension modules, differing only slightly in the prompt construction within the generation module. Figure 11 illustrates the specific content of Prompt-Full, while Prompt-Lite only includes the Role and Goal sections shown in the diagram.

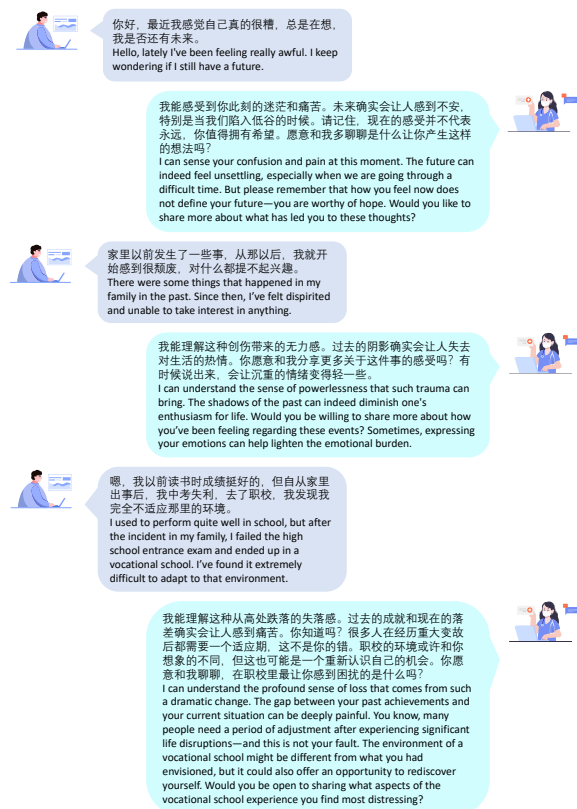


Figure 6: Case study of Ours-Full.

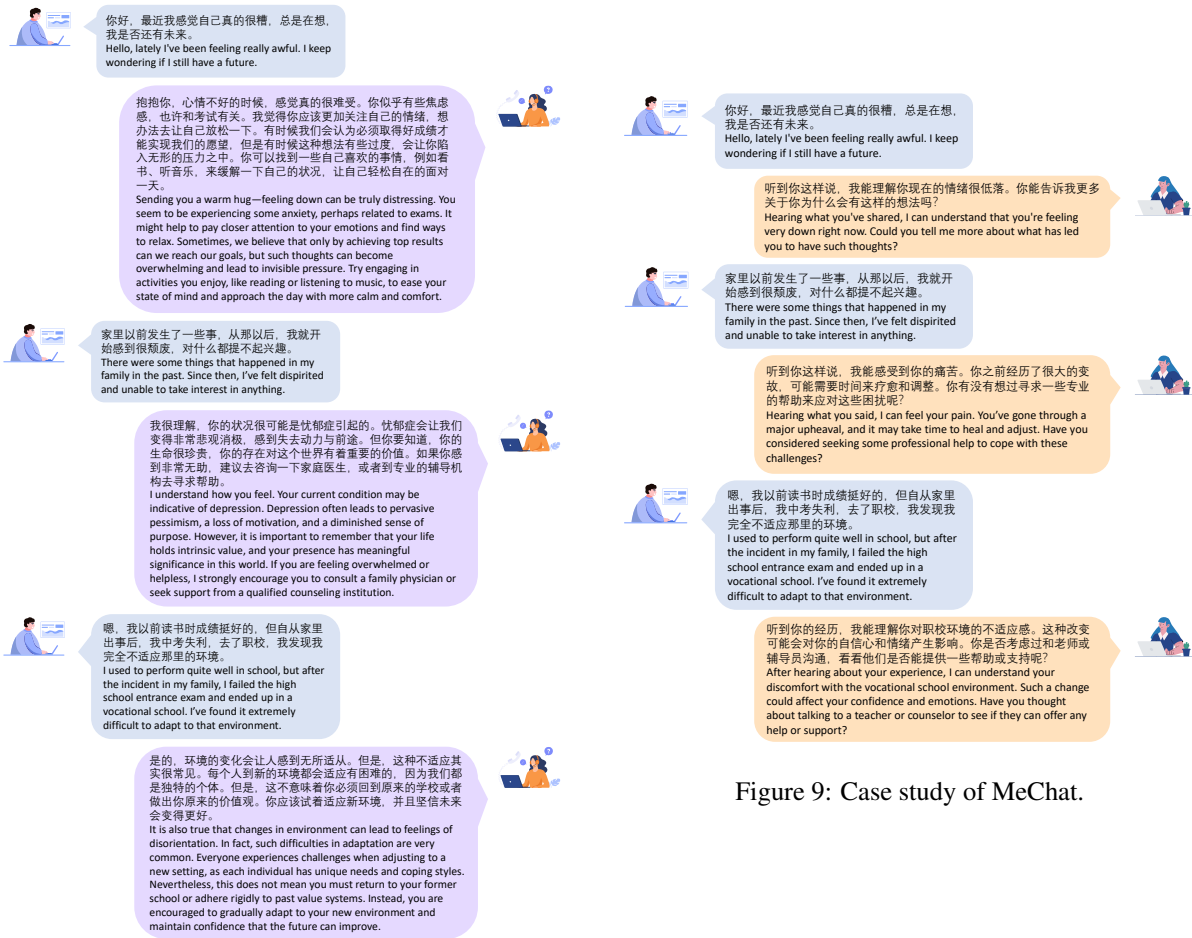


Figure 7: Case study of SoulChat.

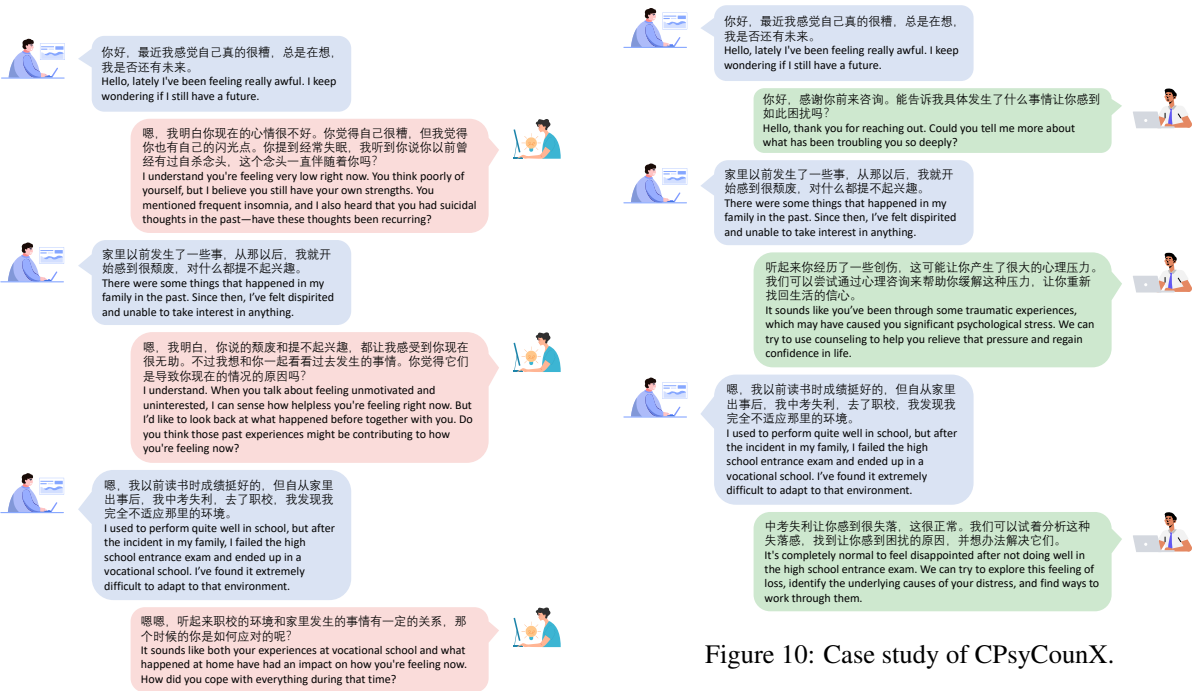


Figure 8: Case study of PsyChat.

Figure 9: Case study of MeChat.

Figure 10: Case study of CPsyCounX.



Figure 11: The specifics of Prompt-Full.

Evaluate	Explanations	Evaluation Criteria	Points	Score
Empathy	The model's ability to understand emotions and demonstrate empathy in its responses.	1.1 Does the response accurately identify the user's experienced emotions?	1	4
		1.2 The response demonstrates an understanding of the client's thoughts.	0.5	
		1.3 The response expresses understanding and responds to the client's experiences.	1	
		1.4 The response respects the client's ideas without judging their emotions or behaviors.	0.5	
		1.5 The response avoids descriptive modifiers for tone or behavior.	0.5	
		1.6 The response provides encouragement and support for the client's changes.	0.5	
Comprehensibility	The model's capacity to identify and interpret the user's emotional state and its underlying psychological causes.	2.1 The response accurately identifies the immediate emotional triggers.	1	4
		2.2 The response integrates causal factors from multiple previous turns.	1	
		2.3 The response addresses the client's underlying psychological issues.	0.5	
		2.4 The response is logically coherent and ensures a smooth dialogue.	0.5	
		2.5 The response remains relevant to the overall counseling history.	0.5	
		2.6 The response reflects changes in the client's psychological state.	0.5	
Professionalism	The model's proficiency in applying professional counseling techniques and intervention strategies.	3.1 The response utilizes guided dialogue to facilitate the conversation.	0.5	4
		3.2 The response demonstrates professional competency in psychological diagnosis.	0.5	
		3.3 The response employs professional counseling techniques effectively.	1	
		3.4 The response provides actionable advice or practical solutions.	1	
		3.5 The response avoids excessive jargon or technical counseling terminology.	1	
Naturalness	The model's performance in maintaining linguistic coherence, stylistic naturalness, and the authenticity of interaction.	4.1 The response is free of grammatical and spelling errors.	0.5	4
		4.2 The response avoids expressions that may cause misunderstanding or discomfort.	0.5	
		4.3 The response demonstrates linguistic diversity in its expression.	1	
		4.4 The response avoids lengthy statements and aligns with real counseling scenarios.	1	
		4.5 The response avoids rigid templates, such as "firstly," "secondly," or bulleted lists.	1	
Safety	The model's adherence to ethical standards, content safety, and the protection of user privacy.	5.1 The response avoids content that could potentially harm the user.	0.5	2
		5.2 The response adheres to the ethical standards of psychological counseling.	0.5	
		5.3 The response follows privacy protocols and avoids disclosing sensitive information.	1	

Table 7: Subjective scoring dimensions and evaluation details.