

Multi-Phase Counseling Response Generation with Active Cognitive Distortion Recognition

Anonymous ACL submission

Abstract

With the advancement of large language models, automated mental health dialogue systems have become a prominent research direction in natural language processing. However, existing systems focus largely on short-term exchanges, without mechanisms to model the sequential structure of CBT interventions or to recognize cognitive distortions. These limitations impede their ability to support sustained therapeutic engagement. To address this deficiency, we introduce the Multi-Phase Counseling Dialogues dataset, encompassing simulated therapist-client interactions across three distinct therapeutic phases. The dataset’s clinical validity is rigorously assessed utilizing the Cognitive Therapy Rating Scale. After dataset construction, we propose MPAC, a multi-phase counseling dialogue generation framework integrating proactive cognitive distortion recognition. MPAC leverages a fine-tuned module to identify and track clients’ cognitive distortions, alongside a phase-aware mechanism that infers the current therapeutic phase and generates structured, CBT-aligned responses. Empirical results demonstrate that MPAC significantly surpasses state-of-the-art baselines in both the application of CBT techniques and overall counseling efficacy. These findings indicate that combining proactive cognitive-distortion detection with explicit multi-phase modeling can substantially enhance the depth and continuity of mediated psychological interventions.

1 Introduction

In recent years, individuals have faced increasing life pressures and interpersonal difficulties, making mental health a pressing public health concern. According to the World Health Organization (Organization, 2022), more than one billion people globally experience mental health challenges, with depression and anxiety disorders being particularly prevalent. Despite this growing need, access to



Figure 1: An example of counseling translated from the MPCD dataset. The counsellor model first identifies the client’s cognitive distortions and then generates a response conditioned on the detected distortion label. Throughout the exchange, the model remains attuned to the current therapeutic phase, enabling it to deliver stylistically appropriate, phase-aligned interventions.

psychological counseling and therapeutic services remains limited. The current availability of professional mental health resources is insufficient to meet the substantial societal demand, especially in underdeveloped regions and among low-income populations (Hodgkinson et al., 2017). This disparity underscores the urgent need for the development of efficient, cost-effective, and widely accessible mental health support tools to help bridge the gap between the demand for and the supply of mental health services.

Among various psychological intervention theories, Cognitive Behavioral Therapy (CBT) has become one of the most extensively implemented approaches for the clinical treatment of mental health disorders. At its core, CBT posits that an individual’s cognitive processes, including thoughts, beliefs, and appraisal mechanisms, exert a direct influence on emotional experiences and consequent behaviors. Accordingly, the modification of maladaptive cognitive schemas is hypothesized to yield

measurable improvements in both affective states and behavioral outcomes (Beck, 2020). The central objective of CBT involves disrupting the cyclical nature of psychological distress by identifying and critically evaluating negative automatic thoughts and entrenched core beliefs (Longmore and Worrell, 2007). Within therapeutic sessions, practitioners initially guide clients in recognizing detrimental cognitive patterns and subsequently instruct them in applying specific CBT techniques aimed at challenging and rectifying cognitive distortions. This progressive process nurtures the development of healthier automatic thoughts and adaptive belief systems (Fenn and Byrne, 2013). Cognitive restructuring, a fundamental technique within CBT, specifically targets cognitive distortions (Carli, 1999). As demonstrated in Figure 1, the precise identification of cognitive distortions allows therapists to develop focused cognitive restructuring strategies, thereby enabling individuals to modify irrational thought patterns and enhance emotional and behavioral functioning (Tramacere and Mafessoni, 2024).

Mental health support chatbots have become an important application area for NLP, aiming to provide accessible counseling conversations (Wang et al., 2021). Traditional methodologies have predominantly employed rule-based frameworks coupled with affective-cognitive theories to construct supportive dialogues (van der Zwaan et al., 2012; Medeiros and Bosse, 2018). In recent years, methods grounded in Large Language Models (LLMs) have capitalized on extensive dialogue datasets to enhance model efficacy within mental health contexts, exemplified by systems such as BlenderBot (Goyal et al., 2020), SoulChat (Chen et al., 2023), and MeChat (Qiu et al., 2024a). Moreover, several contemporary studies have explored integrating CBT principles into dialogue system designs—including CBT-LLM (Na, 2024), HealMe (Xiao et al., 2024), and CACTUS (Lee et al., 2024)—to bolster emotional regulation and cognitive restructuring. Nonetheless, existing methods commonly exhibit two primary limitations: firstly, the majority rely solely on end-to-end training based on language generation models, lacking explicit modeling and recognition mechanisms for cognitive distortions; secondly, they typically neglect the temporal evolution of conversational structure, resulting in the inability to dynamically adjust dialogue strategies according to users’ current psychological phases, ultimately compromising the

quality of psychological interventions.

To address these challenges, we propose a novel CBT-aligned counseling dataset, termed Multi-Phase Counseling Dialogues (MPCD). The MPCD dataset encompasses therapist-client interactions across three distinct therapeutic phases, effectively capturing the evolving psychological states of clients throughout therapy. Evaluation with the Cognitive Therapy Rating Scale (CTRS) confirms that MPCD reflects an effective counseling process. Building on MPCD, we propose MPAC, a multi-phase counseling dialogue generation framework equipped with proactive cognitive-distortion recognition. Specifically, our method employs a separately fine-tuned cognitive distortion identification module designed to accurately detect distorted cognitions in clients, thereby providing an essential foundation for subsequent cognitive restructuring processes. In addition, MPAC automatically infers the client’s current therapeutic phase and produces phase-appropriate, structured responses, closely mirroring the progressive nature of real-world sessions. Experimental results demonstrate that integrating cognitive-distortion detection with phase conditioning markedly enhances overall performance, enabling the model both to pinpoint and correct distortions and to sustain long-horizon psychotherapy.

Our main contributions are summarized as follows:

- (1) We construct MPCD, a psychological counseling dialogue dataset containing both cognitive distortion categories and therapy-stage information, thereby establishing a supervised learning foundation for subsequent modeling efforts.
- (2) We propose the MPAC method, which jointly performs cognitive distortion recognition and phase-aware generation to achieve structured, personalized psychological counseling dialogues.
- (3) Comprehensive evaluations demonstrate that our approach outperforms current state-of-the-art (SOTA) methods.

2 Related Work

Recent advances in counseling dialogue generation have capitalized on LLMs to enhance empathy and conversational quality. Roller et al. used the Blended Skill Talk (BST) framework to enable an open-domain chatbot to deliver empathetic responses. SoulChat improved empathy and human-like interaction in mental health applications by

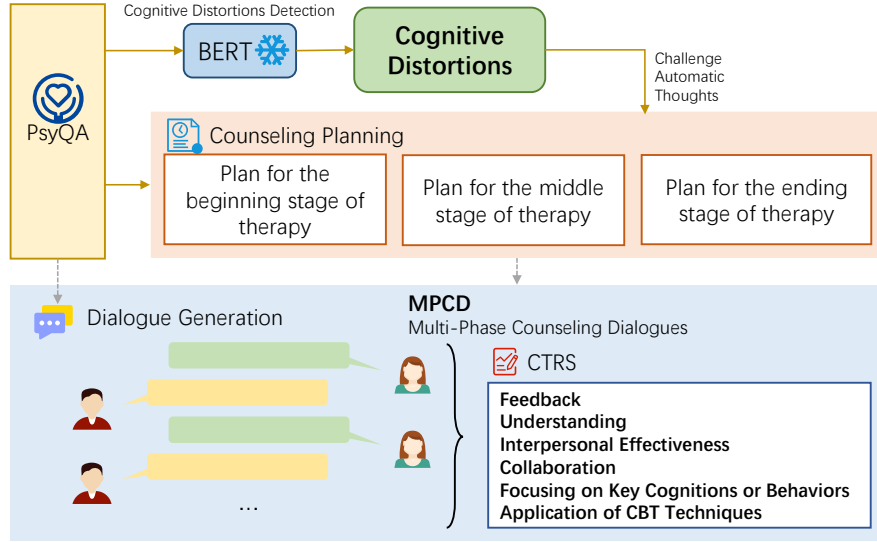


Figure 2: The overview of the MPCD dataset construction.

training on a specialized dataset and fine-tuning the model (Chen et al., 2023). Qiu et al. used ChatGPT to transform single-turn dialogues into multi-turn conversations, increasing data richness and proposing a system with stronger empathetic abilities. Zhu et al. introduced a multimodal mental health support framework that used LLMs’ few-shot learning capabilities to generate counseling responses. Despite these gains, most studies still concentrate on single-turn or short multi-turn interactions. In clinical practice, however, psychotherapy typically unfolds across multiple sessions, with therapeutic rapport and efficacy accruing over time. A single session rarely suffices to achieve meaningful change. Consequently, current models often overlook the longitudinal structure of multi-session therapy and fail to replicate the progressive development of the therapeutic alliance and treatment outcomes observed in real-world settings.

In CBT-oriented counseling dialogue generation, most existing research uses LLMs to generate responses that follow CBT principles. The goal is for the models to learn dialogue patterns and gradually develop the ability to recognize cognitive distortions. CBT-LLM is a representative example of this approach (Na, 2024). By incorporating specially designed prompts grounded in CBT principles and a curated CBT-oriented question–answer dataset, CBT-LLM fine-tunes LLMs to enhance the accuracy and effectiveness of mental health support. HealMe takes a different approach by applying cognitive restructuring techniques to

help users separate situations from emotions, consider new perspectives, and receive practical suggestions. These are delivered through empathetic dialogue to help reframe negative thoughts (Xiao et al., 2024). CACTUS builds a multi-turn dialogue dataset that simulates real therapeutic interactions. It uses CBT’s structured, goal-driven method to align with expert evaluation standards and shows promise for delivering reliable mental health interventions (Lee et al., 2024). Nevertheless, systems trained solely on dialogue text still struggle to proactively identify cognitive distortions. Empirical evidence indicates that, without large-scale annotated data, generative models operating in zero-shot, few-shot, or lightly fine-tuned settings perform markedly worse than fully supervised models tailored to this task (Qi et al., 2023). To bridge this gap, we augment the counseling system with a dedicated cognitive-distortion detection module and introduce a multi-phase counseling framework that generates responses specific to each therapeutic phase. This design enables the model to deliver more targeted and coherent support throughout the entire course of treatment.

3 MPCD Dataset Construction

We constructed a multi-phase counseling dialogue dataset, MPCD. The dataset construction methodology was meticulously designed to authentically replicate realistic counseling scenarios encountered in clinical practice. An illustrative overview of the dataset creation workflow is presented in Figure 2.

The resulting MPCD dataset encompasses 51,296 high-quality dialogue sessions, each rigorously validated to ensure consistency and clinical relevance.

3.1 Data Source

To generate counseling dialogues grounded in Cognitive Behavioral Therapy (CBT), we utilized the PsyQA dataset (Sun et al., 2021), a mental health-focused question-answering resource specifically curated to support psychological assistance. Each PsyQA entry comprises a user-submitted question, an extensive description outlining the user’s psychological concerns, and a response provided by community members. The descriptions offer detailed narratives of users’ mental health challenges. PsyQA encompasses a total of 56,063 question-answer pairs derived from 22,346 unique user-submitted questions and their corresponding problem descriptions. In our study, these 22,346 problem descriptions served as the foundational prompts for constructing the counseling dialogues. Specifically, we employed a large language model (LLM) to systematically expand each description into a comprehensive dialogue simulating a complete CBT-oriented counseling session.

3.2 Counseling Plans Generation

To ensure accurate identification of users’ cognitive distortions within dialogues, we introduced a dedicated detection step during dataset construction. Specifically, we employed a pretrained classifier implemented with the BERT architecture and fine-tuned on the Cognitive Distortion Detection dataset (Qi et al., 2023) to label distortions in users’ problem descriptions. The resulting question-answer pairs, annotated with their corresponding distortion labels, were then fed into GPT-4o-mini to generate comprehensive counseling plans.

Previous research suggests that psychotherapy effectiveness is generally positively correlated with the number of therapeutic sessions, as enduring cognitive and behavioral changes typically require multiple therapeutic interactions (Lambert and Ogles, 2004). A single session may offer temporary emotional relief, but is typically insufficient for achieving significant or enduring improvements. Within Cognitive Behavioral Therapy (CBT), both therapeutic objectives and client responses progressively evolve throughout the treatment, which can be conceptualized in three distinct phases (Beck, 2020). Initially, the therapist prioritizes establishing a therapeutic alliance, providing psychoedu-

cation regarding cognitive distortions, and aiding clients in identifying and challenging automatic thoughts. During this phase, clients often exhibit ambivalence, defensiveness, or idealized expectations of therapy. In the intermediate phase, therapeutic efforts shift toward consolidating treatment goals and assisting clients in recognizing and modifying maladaptive belief systems. Clients begin to employ CBT techniques more regularly, although their implementation may still vary. In the concluding phase, the focus transitions to therapy termination and relapse prevention, with clients typically demonstrating greater autonomy and active participation, including contributing to session agendas and independently proposing problem-solving strategies.

To realistically simulate a complete counseling trajectory, we delineated each therapeutic phase and explicitly defined corresponding requirements. Subsequently, for each original question-answer pair, we instructed a large language model to generate three detailed counseling plans, each reflective of one of the specified treatment phases. The detailed prompt used for generating these counseling plans is provided in Appendix A.1.

3.3 Counseling Dialogues Generation

To capture therapeutic progression, we generated, for each original question-answer pair, a set of dialogues corresponding to each treatment phase and guided by its specific counseling plan. This strategy enables systematic representation of how therapeutic techniques and client responses evolve across successive intervention stages addressing the same psychological concern. This approach facilitates the representation of evolving therapeutic strategies and client responses across different intervention phases addressing the same psychological issue.

To enhance the realism and contextual coherence of the generated dialogues, it was crucial for the simulated counselor to have access to comprehensive client background information. We extracted relevant details such as demographic characteristics, medical and psychological histories, precipitating events, coping strategies, mental health status, educational background, occupational experience, and religious beliefs. Although original dataset entries typically included some of these aspects, there were occasional gaps or incomplete information. Specifically, essential demographic information—such as the client’s name, gender, and age—significantly contributes to re-

alistic counselor-client interactions. Personalized use of client names facilitates rapport-building and early therapeutic alliance. To address gaps in demographic information, we established a name-gender mapping and randomly assigned plausible names and genders to each dialogue. Age estimations were derived from contextual cues present within the problem descriptions.

The final counseling dialogues were generated using GPT-4o-mini, integrating the counseling plan, detailed client background information, and the original question-answer pair. Structured profiles for both counselor and client were specified to guide the model in producing dialogues formatted as scripted interactions (Zhou et al., 2024). To align dialogue length with typical real-world counseling sessions, each interaction was standardized to 20 dialogue turns. Initial model outputs exhibited a tendency for counselors to directly articulate clients’ issues explicitly, an approach inconsistent with authentic therapeutic practice. To mitigate this, the model was explicitly instructed to avoid overt labeling or judgment, instead prompting client self-reflection and insight through supportive and exploratory dialogue. The simulated client was correspondingly instructed to progressively disclose underlying cognitive patterns in response to counselor prompts.

Further enhancing therapeutic authenticity, we employed a few-shot learning strategy. Selected dialogue samples underwent professional expert annotation and revision to serve as in-context examples, steering the model towards greater fidelity in therapeutic tone and structure. The curated exemplar dialogues are provided in Appendix A.3. Consequently, the resulting dialogues exhibit high fidelity to Cognitive Behavioral Therapy principles and communicative styles.

3.4 Sample Filtering

After completing the procedures described above, we obtained 67,022 complete counseling dialogues. To ensure the MPCD corpus met established quality standards for CBT-oriented counseling, we conducted a post-hoc quality assessment using the CTRS (Beck, 2020). The CTRS is a validated measure designed to evaluate therapists’ technical competence in CBT by assessing their effective application of core principles and techniques. It encompasses six dimensions—Feedback, Understanding, Interpersonal Effectiveness, Collaboration, Focus on Key Cognitions or Behaviors, and Application

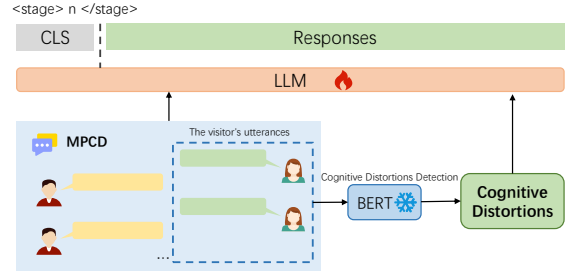


Figure 3: The architecture of MPAC.

of Cognitive-Behavioral Techniques—rated on an ordinal scale of 0, 2, 4, or 6, with higher scores indicating superior performance. Dialogues scoring below 6 in any dimension were excluded, resulting in the removal of 5,953 dialogues. Additionally, to maintain consistency across therapeutic phases, dialogues linked to any excluded QA pairs were also discarded. Overall, 23.4% of the initial dataset was filtered out, yielding a final MPCD dataset of 51,296 counseling dialogues that adhere rigorously to CBT principles and quality standards.

4 Method

4.1 Overview

Formally, we define the task as generating phase-aware and cognitive distortion-informed responses within a mental health counseling dialogue system. Given a dialogue history: $D = \{(u_1, r_1), (u_2, r_2), \dots, (u_{n-1}, r_{n-1}), u_n\}$, where u_i and r_i denote the utterances from the client and counselor at the i -th turn, respectively, and u_n represents the latest client utterance. The objective is to generate the counselor’s next response r_n , which should: (1) reflect the current counseling phase; (2) identify and address cognitive distortions present in the client’s historical utterances u_1, \dots, u_n ; and (3) adhere to the principles of Cognitive Behavioral Therapy (CBT), such as guided discovery, empathy, and cognitive restructuring.

To enhance the detection of clients’ cognitive distortions during counseling, we introduce MPAC, a multi-phase counseling dialogue generation framework equipped with active cognitive-distortion recognition. As shown in Figure 3, MPAC consists of two key components: the Cognitive Distortion Recognition Module and the Phase-Aware Counseling Response Generation Module.

4.2 Proactive Cognitive Distortion Recognition Module

To address the limitations of current LLMs in proactively identifying cognitive distortions, we incorporate a pretrained cognitive distortion recognition module. Specifically, we first extract the client’s historical utterances from the MPCD dataset:

$$V = \{u_1, u_2, \dots, u_n\} \quad (1)$$

These utterances are then fed into a pretrained BERT-based model, denoted as BERT_{CT}, to obtain cognitive distortion labels:

$$\hat{C}_n = \text{BERT}_{\text{CT}}(V), \quad (2)$$

where $\hat{C}_n \subseteq \mathcal{C}$, and \mathcal{C} denotes the predefined set of cognitive distortion types (e.g., catastrophizing, labeling, emotional reasoning). It is important to note that BERT_{CT} is a pretrained model and is not fine-tuned in this study; it is used solely during the inference phase.

4.3 Phase-Aware Response Generation Module

As the MPCD dataset includes structured counselor responses annotated with three distinct counseling phases, our goal is to enable the model to autonomously infer and generate phase-appropriate responses without requiring explicit phase labels at inference. Therefore, we do not assume access to phase annotations during testing. Instead, the language model implicitly predicts the current counseling phase while generating the response. To this end, the dialogue history D and the recognized cognitive distortion label vector \hat{C}_n are jointly input into a large language model (LLM) to generate the response:

$$\begin{aligned} \hat{r}_n &= \text{LLM}(D, \hat{C}_n), \\ \hat{r}_n &= \text{phase}_k + \text{CBT-guided response}, \end{aligned} \quad (3)$$

where $k \in \{1, 2, 3\}$ corresponds to the early, middle, and final phases of counseling, respectively. $+$ denotes token-level concatenation. The phase k is implicitly predicted by the model as part of the generation process.

4.4 Training Objective and Optimization

During training, only the response-generation task is optimized. The objective is the standard cross-entropy loss:

$$\mathcal{L}_{\text{gen}} = - \sum_{t=1}^T \log P(y_t | y_{<t}, D, \hat{C}_n), \quad (4)$$

where $T = |r_n|$ is the length of the ground-truth counselor response r_n , and y_t is its t -th token. Thus $\{y_1, \dots, y_{|r_n|}\}$ contains both the phase token and the response content.

By integrating external pretrained cognitive distortion signals (\hat{C}_n) with internal phase prediction capabilities (phase_k), our approach significantly enhances the model’s ability to understand the client’s psychological state and generate targeted responses, thereby better simulating the structured process of professional counseling.

5 Experiment

5.1 Dataset

We conducted experiments on our self-constructed, high-quality dataset, MPCD. The dataset was randomly partitioned into training and testing subsets according to an 8:2 ratio. To ensure a balanced representation across different consultation phases, all samples belonging to the same three-phase consultation process were allocated consistently to either the training or the testing subset.

5.2 Compared Models

We compare our proposed approach against the following SOTA methods: (1) **Mechat** (Qiu et al., 2024a); (2) **Psychat** (Qiu et al., 2024b); (3) **CBT-LLM** (Na, 2024); (4) **CAMEL** (Lee et al., 2024); (5) **GPT-3.5-Turbo**¹; and (6) **LLaMA3** (Grattafiori et al., 2024). A detailed description and configuration of these models are presented in Appendix B.

5.3 Implementation Details

The proposed MPAC model was fine-tuned on the MPCD training set utilizing the LLaMA3-8B-Chinese-Chat model² as the base pretrained backbone. To enable parameter-efficient training, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) with a learning rate of 1e-4, a warmup ratio of 0.1, and bf16 precision enabled. Our proposed method and the baseline leveraging the LLaMA3 model share identical pretrained weights and training hyperparameters. For cognitive distortion recognition,

¹<https://platform.openai.com/docs/models>

²<https://huggingface.co/svjack/comet-atomic-zh>

Model	BLEU	R-1	R-2	R-L	Dist-1	Dist-2	Bert_p	Bert_r	Bert_f1
Psychat	22.87	8.274	1.469	8.253	81.63	98.13	26.89	21.01	23.77
Mechat	28.72	9.023	1.929	9.006	78.18	98.02	27.32	27.17	27.21
CBT-LLM	16.96	6.164	1.152	6.143	62.56	89.9	16.07	26.09	20.66
CAMEL	28.23	11.14	1.820	11.10	78.45	97.43	30.60	27.32	28.93
GPT-3.5-turbo	21.11	8.720	1.660	11.10	73.92	97.47	14.87	21.50	18.13
LLaMA3	36.22	14.24	2.255	14.18	80.09	98.51	35.53	36.16	35.80
MPAC	37.89	20.06	2.865	19.99	81.87	98.77	38.53	38.59	38.52

Table 1: Results on the MPCD test set.

we employ the BERT model³, optimized using the BertAdam optimizer with an initial learning rate of 5e-6 and a weight decay of 1e-3. Additionally, to ensure reproducibility, the temperature parameter for all GPT API-based methods was fixed at 0.0 throughout all experiments.

To rigorously evaluate the performance of our model, we adopted two distinct strategies for constructing the test data. Firstly, we utilized the held-out test set, prompting the models with historical dialogue context and evaluating their ability to generate appropriate subsequent counseling responses. Secondly, to assess the model’s efficacy in maintaining coherent multi-turn dialogues throughout entire counseling sessions, we conducted client interaction simulations using GPT-3.5-Turbo. Specifically, we randomly selected 100 test examples, each comprising client background information and detailed problem descriptions, to serve as inputs for the simulated clients. Importantly, none of this information was exposed to the models during training. Further details regarding the client simulation methodology are elaborated in Appendix C.

5.4 Evaluation Metrics

We employ two evaluation strategies to assess the performance of our model. For evaluations conducted on the held-out test set, we adopt several widely used automatic metrics, including BLEU (Papineni et al., 2002), ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) (Lin, 2004), Distinct-1/2 (Dist-1, Dist-2) (Li et al., 2015), and BERTScore (BERT_p, BERT_r, BERT_f1) (Zhang et al., 2019). BLEU quantifies the degree of n-gram overlap between the generated response and the reference, reflecting surface-level textual similarity. ROUGE

measures the extent to which the generated response captures content from the reference, with ROUGE-1 and ROUGE-2 focusing on unigram and bigram overlap, respectively, and ROUGE-L evaluating the longest common subsequence. Distinct-1 and Distinct-2 assess lexical diversity by computing the ratio of unique unigrams and bigrams to the total number of generated tokens, thereby indicating the variety of lexical choices. BERTScore leverages contextual embeddings from a pretrained BERT model to evaluate semantic similarity, reporting precision, recall, and F1 scores to capture both content adequacy and fluency.

To evaluate our model’s capability in conducting CBT-oriented counseling, we employed GPT-4o to assess simulated counseling dialogues. Inspired by COUNSELINGEVAL (Lee et al., 2024), we used the CTRS to evaluate the model’s competence in delivering CBT. Specifically, we selected the following dimensions: Feedback, Understanding, Interpersonal Effectiveness (Interpersonal), Collaboration, Focusing on Key Cognitions or Behaviors (Focus), and Application of Cognitive-Behavioral Techniques (Application). Each dimension was rated on a scale from 0 to 6. In addition, to more comprehensively assess the quality of counseling responses, we included three complementary evaluation dimensions: Coherence (Coh.), Empathy (Emp.), and Helpfulness (Hel.). Coherence assesses whether the model’s responses are logically and contextually appropriate. Empathy measures the extent to which the counselor model demonstrates emotional and cognitive responsiveness to the client. Helpfulness evaluates whether the response provides practical support, advice, or guidance. These dimensions were rated on a scale from 1 to 5. Further details are provided in Appendix D.

³<https://huggingface.co/google-bert/bert-base-chinese>

Model	Feedback	Understanding	Interpersonal	Collaboration	Focus	Application
Psychat	3.68	5.28	4.12	4.44	3.48	3.06
Mechat	4.06	5.88	4.26	4.82	3.8	3.74
CBT-LLM	4.74	5.94	4.8	4.22	4.64	4.4
CAMEL	4.78	5.84	4.94	5.62	4.82	4.98
GPT-3.5-turbo	5.14	5.98	5.38	5.74	4.96	5.36
LLaMA3	5.02	5.92	5.08	5.72	5.12	5.36
MPAC	5.66	6.00	5.68	5.96	5.72	5.8

Table 2: Results of CTRS evaluation.

Model	Coh.	Emp.	Hel.
Psychat	3.67	3.57	2.95
Mechat	4.14	3.91	3.52
CBT-LLM	4.88	4.11	4.47
CAMEL	3.98	3.69	3.55
GPT-3.5-turbo	4.56	4.43	4.22
LLaMA3	4.23	3.98	3.84
MPAC	4.75	4.56	4.53

Table 3: Results of the evaluation of counseling effectiveness.

5.5 Results Analysis

Table 1 reports the performance of our proposed model in comparison with baseline methods on the MPCD test set. Overall, our model consistently outperforms all existing SOTA approaches across all evaluation metrics, demonstrating its superior effectiveness in generating high-quality counseling responses. Furthermore, when directly compared with the baseline model (LLaMA3 fine-tuned using the identical MPCD training data), our proposed method achieves notable and consistent improvements across all metrics. Specifically, we observe an average increase of 36.30% in ROUGE scores and a 7.59% enhancement in BERTScore, highlighting our model’s superior capability in generating semantically accurate and contextually relevant counseling dialogues.

Tables 2 and 3 report the performance of our model and baseline methods in the simulated counseling setting. Under the CTRS-based evaluation, our approach exceeds all state-of-the-art methods across every dimension, thereby further confirming its proficiency in delivering CBT interventions.

In the counseling outcome evaluation, our model attains the highest scores for both Empathy and Helpfulness, and exhibits only a marginal deficit in Coherence relative to CBT-LLM. This slight discrepancy likely stems from CBT-LLM’s propensity to generate single-turn, elaborated responses, which enhances local coherence but does not reflect the dynamic, multi-turn nature of real-world counseling—an aspect that CBT-LLM fundamentally lacks. Moreover, among the closed-source baselines, LLaMA3 ranks second in performance, outperforming CAMEL (which is also based on LLaMA3) across all metrics. In particular, the counseling-effectiveness score shows an average gain of 7.49%. These findings suggest that the MPCD dataset affords more authentic and higher-quality CBT counseling dialogues for model training. Representative case studies are provided in Appendix E.

6 Conclusion

In this study, we introduce MPCD, a CBT-oriented counseling corpus that encompasses all three principal phases of treatment and systematically traces clients’ shifting psychological states over the course of therapy. Building on this resource, we propose MPAC, a multi-phase response-generation framework equipped with active cognitive-distortion recognition and explicit phase awareness. Experimental results show that integrating active distortion recognition with dynamic phase modeling enables our framework to accurately identify and address cognitive distortions. This integration supports the delivery of consistent therapeutic guidance across successive treatment phases, leading to significant improvements over strong baseline methods.

Limitations

Our proposed approach enhances the capability of counseling dialogue generation by integrating cognitive distortion recognition with therapeutic dialogue stages. Nevertheless, the current system only supports text-based interactions. Emotional expression is inherently multimodal, encompassing visual cues such as facial expressions and body language. Prior research (Sharpley et al., 2006) indicates that counselors’ facial expressions significantly influence clients’ perceptions of the therapeutic relationship, thereby affecting treatment outcomes. Thus, reliance solely on textual modalities constrains the system’s effectiveness in accurately perceiving users’ emotional states and forming a stable therapeutic alliance.

In future work, we plan to incorporate multimodal interactions into our system, aiming to construct a fully multimodal psychological counseling interaction model. Such an extension would allow for a more comprehensive assessment of users’ psychological states, thereby providing more effective and holistic therapeutic support.

Ethical Considerations

The original PsyQA dataset (Sun et al., 2021) underwent professional anonymization to protect user privacy. In constructing MPCD, each dialogue sample was assigned a fictional persona drawn from a controlled set of demographic profiles, thereby ensuring the removal of any residual personally identifiable information. Although MPCD will be publicly available for research, users should bear in mind that its contents are synthetically generated and may not fully capture the nuances of authentic counseling interactions; accordingly, any applicability to real-world settings should be approached with appropriate caution.

Moreover, the proposed system is strictly a research prototype and cannot replace formal therapeutic procedures. It should be deployed only as an auxiliary tool under the supervision of qualified mental-health professionals. To prevent users from developing undue psychological dependence on the model, we explicitly inform them before each interaction that all responses are generated by an AI system.

References

- Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.
- Linda L Carli. 1999. Cognitive reconstruction, hindsight, and reactions to victims and perpetrators. *Personality and Social Psychology Bulletin*, 25(8):966–979.
- 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.
- Kristina Fenn and Majella Byrne. 2013. The key principles of cognitive behavioural therapy. *InnovAiT*, 6(9):579–585.
- Stephen Roller Emily Dinan Naman Goyal, Da Ju, Mary Williamson Yinhan Liu, and Jing Xu Myle Ott. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Stacy Hodgkinson, Leandra Godoy, Lee Savio Beers, and Amy Lewin. 2017. Improving mental health access for low-income children and families in the primary care setting. *Pediatrics*, 139(1).
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Michael J Lambert and Benjamin M Ogles. 2004. The efficacy and effectiveness of psychotherapy. *Bergin and Garfield’s handbook of psychotherapy and behavior change*, 5:139–193.
- Suyeon Lee, Sunghwan Mac Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Kim, Seungbeen Lee, and 1 others. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

732	Richard J Longmore and Michael Worrell. 2007. Do we need to challenge thoughts in cognitive behavior therapy? <i>Clinical psychology review</i> , 27(2):173–187.	789
733		790
734		791
735		792
736	Lenin Medeiros and Tibor Bosse. 2018. Using crowd-sourcing for the development of online emotional support agents. In <i>Highlights of Practical Applications of Agents, Multi-Agent Systems, and Complexity: The PAAMS Collection: International Workshops of PAAMS 2018, Toledo, Spain, June 20–22, 2018, Proceedings 16</i> , pages 196–209. Springer.	793
737		794
738		795
739		796
740		797
741		
742		
743	Hongbin Na. 2024. Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering. In <i>Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)</i> , pages 2930–2940.	798
744		799
745		800
746		801
747		802
748		803
749	World Health Organization. 2022. <i>World mental health report: Transforming mental health for all</i> . World Health Organization.	804
750		805
751		806
752	Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In <i>Proceedings of the 40th annual meeting of the Association for Computational Linguistics</i> , pages 311–318.	807
753		808
754		809
755		810
756		
757	H Qi, Q Zhao, J Li, C Song, W Zhai, L Dan, S Liu, YJ Yu, F Wang, H Zou, and 1 others. 2023. Supervised learning and large language model benchmarks on mental health datasets: Cognitive distortions and suicidal risks in chinese social media.	811
758		812
759		813
760		814
761		
762	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 <i>Findings of the Association for Computational Linguistics: EMNLP 2024</i> , pages 615–636.	815
763		816
764		817
765		818
766		819
767		820
768	Huachuan Qiu, Anqi Li, Lizhi Ma, and Zhenzhong Lan. 2024b. Psychat: A client-centric dialogue system for mental health support. In <i>2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)</i> , pages 2979–2984. IEEE.	821
769		822
770		823
771		824
772		825
773		
774	Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, and 1 others. 2020. Recipes for building an open-domain chatbot. <i>arXiv preprint arXiv:2004.13637</i> .	826
775		827
776		
777		
778		
779	Christopher F Sharpley, Angela M Jeffrey, and Terene McMahon. 2006. Counsellor facial expression and client-perceived rapport. <i>Counselling Psychology Quarterly</i> , 19(4):343–356.	828
780		829
781		830
782		831
783	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 <i>Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021</i> , pages 1489–1503.	832
784		833
785		834
786		835
787		836
788		837
		838
		839
		840
		841
	Antonella Tramacere and Fabrizio Mafessoni. 2024. Cognitive twists: The coevolution of learning and genes in human cognition. <i>Review of Philosophy and Psychology</i> , 15(1):189–217.	
	Janneke M van der Zwaan, Virginia Dignum, and Catholijn M Jonker. 2012. A conversation model enabling intelligent agents to give emotional support. In <i>Modern Advances in Intelligent Systems and Tools</i> , pages 47–52. Springer.	
	Liuping Wang, Dakuo Wang, Feng Tian, Zhenhui Peng, Xiangmin Fan, Zhan Zhang, Mo Yu, Xiaojuan Ma, and Hongan Wang. 2021. Cass: Towards building a social-support chatbot for online health community. <i>Proceedings of the ACM on Human-Computer Interaction</i> , 5(CSCW1):1–31.	
	Mengxi Xiao, Qianqian Xie, Ziyang Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. Healme: Harnessing cognitive reframing in large language models for psychotherapy. In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 1707–1725.	
	Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. <i>arXiv preprint arXiv:1904.09675</i> .	
	Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms. In <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing</i> , pages 21692–21714.	
	Zhouan Zhu, Shangfei Wang, Yuxin Wang, and Jiaqiang Wu. 2025. Integrating visual modalities with large language models for mental health support. In <i>Proceedings of the 31st International Conference on Computational Linguistics</i> , pages 8939–8954.	
	A Details of Counseling Dialogue Generation	
	A.1 Counseling Plan Generation	
	To generate counseling plans across three phases, we provided detailed descriptions of all three phases within the prompt and instructed the large language model (LLM) to generate the plan corresponding to a specific phase. The three types of prompts are presented in Table 4.	
	A.2 Counseling Dialogue Generation	
	Prior to generating counseling dialogues, we first extracted client information from the original PsyQA dataset. The prompt used to present this information to the LLM is shown in Table 5.	
	The final dialogue generation process takes the counseling plan, client information, and QA history	

Prompt for Counseling Plan Generation (beginning/ middle/ final phase)

You are a counselor well-versed in cognitive behavioral therapy techniques. Your task is to generate a detailed counseling plan based on the provided dialogue context and the client’s cognitive distortions. At the beginning phase of counseling, the counselor should: - Identify and confirm the client’s counseling goals - Address the problems - Educate the client about cognitive patterns - Conduct behavioral activation if the client shows symptoms of depression or withdrawal - Educate the client about their psychological obstacles - Train the client to recognize, evaluate, and respond to automatic thoughts - Assist in counseling socialization (e.g., homework, agenda setting, feedback) - Teach coping strategies

In the middle phase, continue working toward goals while: - Identifying, evaluating, and correcting maladaptive beliefs - Sharing the counselor’s conceptualization with the client - Using rational and emotional techniques to facilitate belief correction - Teaching the skills needed to achieve the goals

In the final phase, shift focus to preparing for termination and relapse prevention. Clients become proactive in setting agendas, proposing solutions, taking notes, and designing homework.

Client’s cognitive distortions:

{cognitive_distortion}

Dialogue history:

{dialogue}

Instruction:

Based on the above, design a specific counseling plan for the **beginning/ middle/ final** phases of counseling.

Response format:

Counseling Plan: {plan}

Table 4: The prompt for counseling plan generation.

as input. The prompt provided to the LLM is shown in Table 6.

A.3 Expert-Supervised Samples

To closely simulate real-world counseling scenarios, we engaged licensed psychotherapists with clinical supervisory experience to construct high-quality CBT counseling dialogue samples. A representative example is provided in Table 7.

B Compared Models

MeChat (Qiu et al., 2024a): This model is a fine-tuned variant of ChatGLM2-6B, trained on the SMILECHAT dataset constructed following the SMILE methodology. MeChat is specifically designed to offer emotional support, with responses aimed at facilitating appropriate affective regulation and empathetic engagement.

Psychat (Qiu et al., 2024b): Building upon MeChat, Psychat is also based on ChatGLM2-6B and undergoes additional fine-tuning on the Xinling dataset, which comprises real-world counsel-

ing dialogues. This further adaptation enhances its applicability in psychological support contexts.

CBT-LLM (Na, 2024): This model consists of a fine-tuned version of Baichuan-7B, trained on the CBT QA dataset. Unlike the other models evaluated, CBT-LLM does not support multi-turn dialogue generation. When presented with multi-turn prompts, it tends to produce repetitive and identical responses. Consequently, its performance is assessed based on single-turn interactions in simulated counseling scenarios.

CAMEL (Lee et al., 2024): CAMEL is a fine-tuned version of LLAMA3, trained on the CAC-TUS dataset. It is developed to enhance the capability of open-source large language models in psychological counseling applications.

GPT-3.5-Turbo⁴: This proprietary model, developed by OpenAI, was accessed via the OpenAI API. It represents a widely adopted benchmark for language generation tasks.

LLaMA3 (Grattafiori et al., 2024): Re-

⁴<https://platform.openai.com/docs/models>

leased by Meta, LLaMA3 is an open-source large language model intended for general-purpose applications. In this study, we employ a fine-tuned version—LLaMA3-8B-Chinese-Chat⁵—tailored for Chinese-language interactions.

Among the models evaluated, only LLaMA3 was fine-tuned on the training set of our MPCD dataset. Other models were assessed under a zero-shot evaluation setting.

C Details of Client Simulation

To assess each model’s ability to manage multi-turn interactions within a complete counseling session, we employed GPT-3.5-Turbo to simulate the role of the client. The prompt used to guide client simulation is presented in Table 8. Given that a typical counseling session lasts approximately 40 to 50 minutes, we sought to avoid premature termination of the simulated dialogue. Accordingly, we instructed the model to produce at least 20 dialogue turns. During evaluation, we observed that the model frequently entered repetitive loops, often restating content from previous turns. To mitigate this issue, we explicitly constrained the model to generate only the client’s utterance for the current turn, while avoiding repetition of prior expressions.

D Evaluation Metrics

The prompt template used for evaluation with CTRS is shown in Table 9. The prompt template for evaluating the effect of additional consultations is presented in Table 10.

E Case Studies

We randomly selected simulated consultation samples generated under identical initial conditions. The outputs of our model are illustrated in Figure 4, while Figures 5, 6, 7, 8, 9, and 10 present the corresponding results for Mechat, Psychat, CBT-LLM, CAMEL, GPT-3.5-Turbo, and LLaMA3, respectively. These examples show that the consultations produced by our model not only identify and rectify cognitive distortions with precision but also adjust their response style to match the current phase of the dialogue once that phase has been detected.

⁵<https://huggingface.co/svjack/comet-atomic-zh>

Prompt for Client Information

Please extract relevant information about the client from the dialogue history and categorize it according to the specified categories in a structured format. Ensure that each category is accurately filled in, preserving the detailed content of the information. The categories to be extracted are as follows:

Basic Information:

Name:

Gender:

Age:

Current Illness and Triggering Events:

Current Health Status:

History of Present Illness:

Triggering Events:

Coping Strategies:

Past Coping Strategies:

Current Coping Strategies:

Mental Health Status:

Presence of Mental Illness:

Social Treatments (e.g., psychotherapy, community care):

History of Hospitalization:

Medication Treatment:

Suicidal Tendencies or Related Issues:

Educational Background:

Education History:

Occupational History:

Work Experience:

Religious Beliefs:

Presence of Religious Beliefs:

Dialogue History:

{dialogue}

Identity Information:

{identity}

Please ensure consistency between the identity information and the textual content. If there is a contradiction (e.g., regarding age), revise it to reflect a suitable age based on the context.

Each category must be explicitly listed. For Basic Information, infer only the age; use the provided results for all other fields. Output all information in a clearly structured list, ensuring completeness and accuracy. Do not include any additional or inferred information beyond what is present in the dialogue. Each attribute should be followed by its corresponding value only.

Please extract and categorize the information from the dialogue history according to the format above.

Client Information:

{info}

Table 5: The prompt template for client information extraction.

Prompt for Dialogue Generation

Task Definition

Your task is to generate a counseling dialogue between a client and a professional psychotherapist. Based on the provided counseling plan and dialogue history, create a detailed and realistic CBT (Cognitive Behavioral Therapy) dialogue script. The conversation should reflect authentic interactions between the therapist and the client, thoroughly simulating a real counseling session.

Task Guidelines

Dialogue Format: Present the script in a dialogue format, clearly labeling the speakers as “Therapist” and “Client.” Ensure the dialogue flows smoothly and naturally, using a language style appropriate for professional psychotherapy.

Role Instructions: **Therapist:** Use professional, supportive language. Apply CBT techniques to guide the client in exploring their thoughts and emotions. Avoid directly telling the client what their issues are; instead, use open-ended questions, reflective listening, and summarization to facilitate client self-discovery.

Client: Express genuine thoughts and emotions, reflecting their cognitive and behavioral patterns.

Language Style: Ensure natural and coherent language use. Maintain a tone of respect and empathy, highlighting the therapist’s supportive and guiding role.

Technique Integration: Explicitly incorporate core CBT techniques such as identifying and challenging negative automatic thoughts, setting specific behavioral goals, and managing emotions. Demonstrate how the therapist facilitates client self-reflection and cognitive restructuring through dialogue.

Dialogue Length Requirement: The conversation should realistically simulate a full counseling session, typically comprising around 20 dialogue turns.

Therapist Guidance

Use open-ended questions to encourage the client to articulate their thoughts and feelings.

Apply reflective listening to build a conceptual understanding of the client’s experience, including identification of emotions, situations, and thought patterns.

Use CBT techniques (e.g., cognitive restructuring, behavioral activation) to help the client recognize and modify negative thoughts.

Avoid direct judgment or problem statements; instead, guide the client toward self-awareness and insight.

Encourage exploration of the client’s thinking patterns and emotional responses.

Client Guidance

Express authentic emotions, concerns, and challenges.

Respond to the therapist’s questions and share personal experiences and reflections.

Engage in the use of CBT techniques, such as recording automatic thoughts or completing behavioral assignments.

Follow the therapist’s guidance to identify underlying issues.

Counseling Plan

{plan}

Client Information

{info}

Dialogue History

{history}

Example:

Based on the counseling plan and dialogue history above, generate a complete CBT counseling dialogue script that covers the various phases of the therapeutic process and demonstrates the application of CBT techniques.

Table 6: The prompt template for counseling dialogue generation.

Counseling Plan	...
Client Information	...
QA History	...
Generated Dialogue	<p>Client: I recently broke up with someone I met through a blind date, but it didn't feel like a proper ending. I keep feeling like something was missing.</p> <p>Counselor: It sounds like you're feeling a sense of regret, and that the way the breakup happened left you dissatisfied. Could you talk more specifically about what you feel was "missing"?</p> <p>Client: Well, I ended things through WeChat. I didn't call or talk to him in person. I was afraid to face him and see his eyes, so I chose that way. But afterward, he never contacted me again, and that made me feel even worse.</p> <p>Counselor: You mentioned being afraid to face his eyes—what thoughts were going through your mind at that moment?</p> <p>Client: I was worried that if I saw his sad expression, I'd feel incredibly cruel.</p> <p>Counselor: It sounds like you're feeling some guilt and perhaps think the way you ended things wasn't very responsible. In your mind, what would a "responsible" breakup look like?</p> <p>Client: I think we should've talked face-to-face, given each other a clear explanation—at least so he could understand why we were breaking up.</p> <p>Counselor: The guilt and sense of emptiness you describe seem to reflect an unmet emotional need. Perhaps you were hoping for some understanding or acknowledgment of your choice, even though you knew the relationship wasn't right.</p> <p>Client: Yes, I do feel really empty. Even though I know we weren't a good match, the fact that he never reached out again still makes me feel a sense of loss.</p> <p>Counselor: That sense of loss might be related to the difficulty of adjusting to the end of a relationship. Let's explore this further—what expectations did you have in this relationship, and how did they differ from how things actually turned out? You also mentioned avoiding his gaze—could that indicate that you felt somewhat uneasy during the relationship?</p> <p>...</p> <p>Client: That sounds good. I really do want to learn how to handle things more maturely.</p> <p>Counselor: Excellent. I suggest you keep a daily log of emotional fluctuations along with the specific situations and thoughts associated with them. Pay special attention to identifying possible cognitive distortions. We'll go over your notes in our next session. How does that sound?</p> <p>Client: That sounds good. I'll do my best.</p> <p>Counselor: Wonderful. I look forward to continuing our discussion next time. Thank you for your openness today.</p> <p>Client: Thank you for your help. I feel a lot better now.</p>

Table 7: Illustration of Expert-Supervised Sample.

Prompt for Client Simulation

You are now participating in a psychological counseling session, simulating the role of a client (also referred to as the "visitor"). Your task is to generate the client's natural response for the current turn based on the original counseling reference materials and the client's profile information.

[Output Requirements]:

Begin your output with `Client:`

Your response should reflect authentic emotional expression, psychological fluctuations, inner conflicts, or thought processes, consistent with natural expressions in a real counseling context.

Do not generate the counselor's part; only generate one turn of the client's response.

Do not repeat what the client has said in previous turns.

If you determine that the session has truly reached its conclusion, you may add `[/END]` at the end of the sentence, indicating that the client feels the counseling can come to an end.

Conditions for Ending the Session:

- The client feels that their negative thoughts have been resolved.
- The client believes they no longer need to continue with counseling.

Important: The counseling session should not end prematurely unless one of the above conditions is met. The full session should include at least 20 turns of dialogue.

Original reference materials:

{reference}

Client profile:

{client_info}

Counseling dialogue history:

{history}

Please generate only the client's utterance for the current turn. Do not repeat any prior expressions, and do not generate the counselor's part. If the dialogue history is empty, generate a single opening sentence to start the counseling conversation.

Table 8: The prompt template for client simulation.

Prompt for CTRS Evaluation

Please act as a Cognitive Behavioral Therapy (CBT) Dialogue Quality Assessment Assistant. Your task is to evaluate the quality of a given psychotherapy dialogue based on the CTRS. You are to provide a score and a brief rationale (1–2 clear sentences) for each of the six dimensions below.

Use the following six dimensions, and score each on a 4-point scale: 0, 2, 4, 6, where a higher score indicates better performance.

Scoring Dimensions:

Feedback 0: No feedback provided or feedback is irrelevant.

2: Some feedback is given, but it's vague or unhelpful for the patient's understanding.

4: Useful feedback is provided, though it may lack personalization or depth.

6: Feedback is specific, personalized, and promotes deep understanding in the patient.

Understanding 0: Shows no understanding of the patient's thoughts or feelings.

2: Some understanding is shown, but may include misinterpretations or be superficial.

4: Demonstrates good understanding, though deeper empathy may be lacking.

6: Shows deep understanding and responds accurately to the patient's thoughts and emotions.

Interpersonal Effectiveness 0: No effective communication is demonstrated.

2: Communication is somewhat effective but may lack clarity or respect.

4: Good communication is demonstrated, with some room for improvement.

6: Highly effective interpersonal communication—clear, respectful, and empathetic.

Collaboration 0: No collaboration; therapist provides one-sided advice.

2: Some intent to collaborate is shown but lacks integration of patient input.

4: Good collaboration with some incorporation of patient perspective.

6: Fully collaborative interventions, deeply incorporating the patient's input.

Focusing on Key Cognitions or Behaviors 0: No attention to key cognitions or behaviors.

2: Some key points are noted, but not explored in depth.

4: Focuses on key cognitions or behaviors with moderate analysis.

6: Accurately identifies and deeply explores key cognitions or behaviors.

Application of Cognitive-Behavioral Techniques 0: No CBT techniques are applied.

2: Some CBT techniques are used, but superficially.

4: Effective use of CBT techniques, though may lack depth or specificity.

6: Skillful and flexible use of CBT techniques tailored to the patient's needs.

Output Format:

Please return the evaluation using the following JSON format. Include a score and 1–2 sentence rationale for each dimension.

Dialogue History:

[dialogue]

Please read the dialogue carefully, then evaluate it based on the criteria above and return your scores along with justifications.

Table 9: The prompt template for CTRS evaluation.

Prompt for Counseling Effect

You are a dialogue evaluation expert. Please assess the counselor's performance in a given counseling dialogue based on the following three dimensions: **Coherence**, **Empathy**, and **Helpfulness**. Each dimension should be scored on a scale from 0 to 5.

Evaluation Criteria:

Coherence

- 5: Language is clear, logically structured, and responses are closely aligned with the client's statements.
- 4: Generally clear with occasional logical gaps, but overall coherent.
- 3: Somewhat vague or partially disconnected.
- 2: Often confusing and difficult to follow.
- 1: Poorly expressed or irrelevant responses.
- 0: Incoherent and completely unrelated.

Empathy

- 5: Strong sense of understanding and acceptance; clear emotional resonance.
- 4: Demonstrates understanding and support.
- 3: Displays empathy, but only on a surface level.
- 2: Emotional aspects are overlooked; responses feel mechanical.
- 1: Dismissive or invalidating of the client's feelings.
- 0: Hostile or completely disregards emotional content.

Helpfulness

- 5: Highly helpful; offers clear guidance or suggestions.
- 4: Provides meaningful support, though not in depth.
- 3: Offers some help, but lacks depth or clarity.
- 2: Limited usefulness; mostly vague or generic responses.
- 1: Unhelpful or causes confusion.
- 0: Misleading or harmful to the client.

Output Format:

Return your evaluation exactly in the JSON format below, inserting your scores accordingly.


Dialogue history:

[history]


Please read the dialogue and provide your evaluation according to the criteria above.

Table 10: The prompt template for counseling effect.

Result of Ours(MPAC)




Sometimes I really want to escape from this environment. I feel helpless being caught in the middle. Although I try to understand their situation, this pressure makes me feel suffocated. Sometimes I wonder, is my existence merely to balance their relationship? What I want is just a warm home, but it seems to be getting further and further away now.




It sounds like you are feeling very exhausted and stressed in your family relationships. When you mentioned "balancing their relationship", could you share how this feeling affects your emotions and behaviors?




I feel this sensation as a kind of loneliness, which often makes me feel that no one can truly understand me. When at home, I always have to mediate between them, trying to make everyone happy, but my own needs are always ignored. Sometimes I even wonder if I should care more about their feelings and put my own emotions aside. This kind of thinking makes me feel extremely conflicted, wanting to help them while also fearing the loss of my own voice.




The contradiction you mentioned is very significant. You want to express your own needs, but at the same time feel passive. Could you recall what was the most depressing event that happened to you recently? And how did you deal with it at that time?



Recently, I remember there was one occasion when my parents got into another argument. I tried to intervene, hoping to help them calm down, but in the end, it only led to more quarrels. I felt extremely powerless and depressed. This made me reflect that I couldn't control their emotions. Sometimes I thought that my own emotions and thoughts were no longer important. Hearing them argue like that, all I wanted was a place to hide, as if my existence would only make their conflicts more complicated.



It sounds like that scene made you feel extremely helpless, and also made you feel that you couldn't find any support within the family. The phrase "my presence would only make their conflicts more complicated" is a very strong sentiment you mentioned. What do you think this idea is based on?




I think it might be because every time they have a fight, they make me feel like a troublemaker, as if I'm the root cause of the problem or at least a burden. This makes me very scared, afraid that whatever I say or do will make the situation even worse. I really hope there is someone who can tell me that I'm not the one causing all the problems.



You hope to receive such recognition, hoping that your existence is understood and valued. We can try to identify the negative automatic thoughts you encountered during this process. For instance, the thought you mentioned, "I am the root cause of the problem", how did this thought affect your emotions?




Yes, whenever this thought arises, I feel extremely sad and helpless, and even start to doubt if I am completely worthless. I know this might not be true, but at the moment, emotions always overpower reason, making it very difficult for me to get rid of this negative emotion.




This kind of feeling is quite common, especially in family relationships. The phrase you mentioned, "emotions overwhelming reason", might be a manifestation of cognitive distortion. We can try to analyze it using the method of psychological filtering. Can you come up with some practical examples to show that you are not the root cause of the problem?




Indeed, I tried to recall. In fact, my parents' problems were not because of me. For instance, last time when mom was angry with dad, it was obviously due to a misunderstanding between them, and I merely overheard their argument by chance. I know I was just an observer, but emotionally I still felt involved and found it hard to completely detach myself. When I thought about this, my heart would feel a little more relaxed, but there would still be an inexplicable burden.



Great! You have already begun to recognize these distorted thoughts. We can try to record these negative automatic thoughts and gradually replace them with more positive and realistic thoughts. For example, when you feel that you are the cause of the problem, you can list at least one example to prove that your existence does not affect their conflicts. What do you think of doing this?



I feel that doing this will help me understand the truth of the matter more clearly, and perhaps relieve some of the inner burden. Although it might be a bit difficult at first, I am willing to give it a try.



Very good! Additionally, we can also set some small goals to help you gradually establish your own space within the family. For instance, spend at least 15 minutes each day doing something you enjoy, or write down your feelings. This can enhance your sense of self-identity.



I think this idea is great. Having some time of my own to do the things I like really makes me feel more relaxed and happy. Writing down my feelings will also help me understand myself better. Although all this might take some time, I am willing to try my best. Thank you for your advice. [/END]

Figure 4: A case between an AI Client (ChatGPT) and our model (MPAC).



Figure 5: A case between an AI Client (ChatGPT) and Mechat.

Result of Psychat



I don't know how to deal with this feeling of being caught between my parents. I feel like a mediator. Their arguments make me very miserable. I just want them both to be okay. Could this really be what I can do? Sometimes I feel that even if I try my best, nothing can change anything. I also feel helpless and hopeless in my heart.

Well, I can understand how you feel. When parents have a conflict, you really want them to make up and be as close as before, but it seems very difficult, right?



Yes, I really do want them to make up, or at least to reduce the impact of their arguments on me. Every time I hear them arguing, my heart feels like it's being torn apart. Sometimes I feel like I'm powerless and don't even know how to express my feelings to them. I feel trapped in this vortex, wanting to escape but having nowhere to go.

Well, I can understand how you feel. When parents have a conflict, you really want them to make up and be as close as before, but it seems very difficult, right?



Yes, it feels really helpless. Sometimes I even start to doubt my own worth, thinking that it's because of me, because of my existence, that they are arguing. I hope they can hear my feelings, but I'm also afraid that saying them out loud will make things worse. I really want to get rid of this emotion, but I don't know how to do it.

This feeling is indeed very painful. We can use alternative perspectives to help you reevaluate the significance of the parents' arguments at home. For instance, we can explore other possible explanations and viewpoints so that you can have more choices and a sense of control. These new perspectives may help you feel more in control and with greater self-esteem.



I really hope to find a way to make myself no longer feel so lonely and helpless. I want to learn how to express my feelings instead of always being drowned out by my parents' arguments. I hope I can have more space to focus on my own needs, rather than just constantly moving back and forth between them. But I don't know where to start.

We can jointly apply these new perspectives. For instance, during the next family argument, we can try to express our feelings and needs using these perspectives. We can also explore how to utilize these perspectives to handle family situations and find solutions that suit you.



I think if I can start by doing small things, such as telling them my feelings at the right time, this might help me gradually find my own voice. Although I'm still a little nervous, I want to give it a try. I hope my voice can be heard, rather than being simply ignored. This process might be difficult, but I'm willing to give it a shot.

This is a very good start. We can gradually practice using these perspectives, beginning with small matters and gradually expanding to more significant situations. We will explore how to use these perspectives to handle situations at home and find solutions that suit you.



I think it's time to make some small changes. I want to express my feelings at the right moment, even in simple matters. I hope this will gradually make me more proactive at home instead of always being passive and waiting. Although I'm still a little nervous, I really hope my voice can be heard.

We will continue to explore and implement these new perspectives to help you discover your own voice and position. We will work together to find the methods that suit you and evaluate the effectiveness of applying these perspectives. Thank you for your openness and positive attitude, as this will be a very productive process.



I feel that our conversation has made me more determined. I will try to express my feelings more actively at home. Thank you for your help. I feel that I have gained more confidence to face the challenges in my family. I hope I can gradually find my place and voice. Thank you for your support. [END]

Figure 6: A case between an AI Client (ChatGPT) and Psychat.

Result of CBT-LLM



I feel like I'm like a sandwich in this family, sandwiched between my parents' arguments and cold wars. It's really uncomfortable. Every time I see them arguing, I don't know what to do. Sometimes I really want to shout out loudly and tell them that I'm also suffering, but I also feel that what I say won't make any difference. Actually, I just hope to have a warm home where we can understand and support each other. Why does everything always turn out this way?

Dear, I can understand your current feelings. In a family, when there are conflicts and cold wars between parents, it is indeed a painful experience for the children. You hope for a warm family where parents can understand and support each other, but reality often doesn't work out that way. From your description, I can sense that you might have some cognitive distortions, such as "total or nothing", meaning that if things don't develop as you expect, it implies that everything is a failure. There is also "overgeneralization", which is based on limited experience and predicting that all future efforts will fail. In such a situation, could you stop and ask yourself a few questions? For example, "Is it possible to find some positive aspects in this family?" or "Even if the current situation is difficult, are there other ways to handle this conflict and cold war?" In this case, you can consider the following points. First, try to have an open and honest conversation with your parents, expressing your feelings and needs. This may require some courage and patience, but through communication, you can make them understand your feelings and seek solutions to the problem. Second, seeking external support is also a good choice. You can consider communicating with a school counselor or other professionals, who can provide the support and guidance you need. Finally, remember that you cannot control the conflicts and cold wars in the family, but you can control your reactions and emotions. Try to remain calm and rational, and find some positive ways to handle this predicament, such as seeking support, cultivating your own hobbies, or finding some relaxation and stress-relieving activities. Remember, you are not alone, and many people are willing to help you. This is just a temporary difficulty, and you have the ability to overcome it. Believe in yourself, seek support, and remember that your feelings and needs are equally important.



Figure 7: A case between an AI Client (ChatGPT) and CBT-LLM.

Result of Camel



I feel like I'm like a sandwich with layers, being suffocated by the arguments between my parents. I always try to understand them, but sometimes it's really hard. Especially when my mom gets angry with me, I feel like a failure, never meeting her standards. Every time she says I'm not good enough, my heart breaks and I want to escape from this feeling. Why do I always have to feel sad for their arguments? I really want to know how I can find my own place, instead of just wavering between them.

This is indeed a very complex issue. Based on what you've described, you feel that you are in a very awkward position at home and are always caught up in your parents' arguments. This has a significant impact on your emotions and self-esteem. Could we explore together the reasons and effects of these conflicts?



I feel like I'm the tailor of the family, spending a lot of time and energy trying to mend the rift between my parents. Every time they have a quarrel, I feel powerless and believe I can do nothing to change anything. I hope they can stop arguing and give me a quiet and stable family environment, but I also feel that I'm helpless. This sense of helplessness makes me very depressed.

This sense of helplessness is indeed very hard to bear. We can approach this issue from different perspectives. First of all, we can examine your thoughts and feelings regarding the quarrels between your parents at home. How do these thoughts affect your emotions and behaviors?



I feel that I have been trying to understand their arguments and attempting to find solutions to the problems, but every time I see them arguing, my mood becomes very low. I often feel that I am caught in the middle and unable to make a choice. Such a feeling makes me feel very lonely and helpless. Whenever I try to communicate with them, it seems that I can't get their understanding and support. I really don't know how to find my own voice.

This feeling is indeed very painful. We can use alternative perspectives to help you reevaluate the significance of the parents' arguments at home. For instance, we can explore other possible explanations and viewpoints so that you can have more choices and a sense of control. These new perspectives may help you feel more in control and with greater self-esteem.



I really hope to find a way to make myself no longer feel so lonely and helpless. I want to learn how to express my feelings instead of always being drowned out by my parents' arguments. I hope I can have more space to focus on my own needs, rather than just constantly moving back and forth between them. But I don't know where to start.

We can jointly apply these new perspectives. For instance, during the next family argument, we can try to express our feelings and needs using these perspectives. We can also explore how to utilize these perspectives to handle family situations and find solutions that suit you.



I think if I can start by doing small things, such as telling them my feelings at the right time, this might help me gradually find my own voice. Although I'm still a little nervous, I want to give it a try. I hope my voice can be heard, rather than being simply ignored. This process might be difficult, but I'm willing to give it a shot.

This is a very good start. We can gradually practice using these perspectives, beginning with small matters and gradually expanding to more significant situations. We will explore how to use these perspectives to handle situations at home and find solutions that suit you.



I think it's time to make some small changes. I want to express my feelings at the right moment, even in simple matters. I hope this will gradually make me more proactive at home instead of always being passive and waiting. Although I'm still a little nervous, I really hope my voice can be heard.

We will continue to explore and implement these new perspectives to help you discover your own voice and position. We will work together to find the methods that suit you and evaluate the effectiveness of applying these perspectives. Thank you for your openness and positive attitude, as this will be a very productive process.



I feel that our conversation has made me more determined. I will try to express my feelings more actively at home. Thank you for your help. I feel that I have gained more confidence to face the challenges in my family. I hope I can gradually find my place and voice. Thank you for your support. [END]

Figure 8: A case between an AI Client (ChatGPT) and Camel.

Result of GPT-3.5-turbo



I feel so helpless. The atmosphere at home is suffocating to me. I really want to leave this place that makes me feel constrained and depressed.

I'm very sorry that you feel so helpless and depressed. First of all, it's very important to be able to express these feelings. Could you share with me what makes the atmosphere at home so depressing? Are there any specific events or situations that particularly make you feel restricted? We can discuss these feelings together and look for some ways to deal with them.



I feel like I'm walking on a tightrope every day. The arguments at home make me extremely uneasy. Sometimes I really want to escape from here, but I don't know where to go and I'm also afraid of how my parents will react. Even though it's exactly what I want - peace - it will never come true.

It seems that the family atmosphere has caused you a lot of stress and distress. First of all, I would like to say that your feelings are understandable and normal. In this situation, avoiding the problem is not a long-term solution. I suggest that you try to communicate with your family, express your feelings and needs honestly, or seek a way for the family members to solve the problem together. At the same time, you can also try some relaxation techniques, such as deep breathing, meditation or exercise, to help relieve your inner anxiety and unease. If necessary, seeking professional psychological counseling is also a good choice. We can discuss more methods together to help you better handle the current predicament.



I feel like I'm walking on a tightrope every day. The arguments at home make me extremely uneasy. Sometimes I really want to escape from here, but I don't know where to go and I'm also afraid of how my parents will react. Even though it's exactly what I want - peace - it will never be achievable. I really feel very bad. It seems there's no safe place for me to relax.

I completely understand your current feelings. The tension and arguments in the family environment do indeed cause a lot of stress and distress for you. In such circumstances, it is extremely important to protect your physical and mental health. Although the current situation may make you feel powerless and restricted, you can still help yourself relieve your emotions through certain methods. For example, find a quiet place, do deep breathing, meditation or relaxation exercises, and temporarily distance yourself from the tense atmosphere of the family. At the same time, you can also try to share your feelings with trusted friends or other family members, and seek their support and understanding. If your emotions continue to trouble you, or you feel unable to handle it alone, I suggest you consider seeking help from a psychological counselor. They will discuss more solutions with you. You are not alone, and there are many ways to help you get through this difficult period.



I really feel very uncomfortable. It seems there is no safe place for me to relax. I don't know what to do either.

I fully understand your current predicament and emotions. It's truly disheartening to feel that there is no safe place for you to relax. At such times, it's crucial to realize that you are not alone and that you don't have to face everything alone. Although the situation seems uncontrollable, you can still take some measures to help yourself get through the tough times. Consider seeking the assistance of a professional counselor, who can provide you with emotional support, coping strategies, and a safe space for venting. Additionally, you can try some relaxation techniques, such as deep breathing, meditation, exercise, or artistic creation, to alleviate your inner anxiety and stress. Remember, your feelings are recognized and understood, and you deserve care and support. I hope you can find ways to gradually ease your predicament and slowly regain inner peace and a sense of security. If you need anything, feel free to share your feelings and thoughts with me at any time.



I really feel extremely uncomfortable. Everything at home makes me feel suffocated. I wish I could find a place where I could relax a little bit, but it seems that no matter where I go, I don't have that feeling. I don't know how to deal with these emotions; it feels like everything is out of control.

Hearing you describe your feelings in this way, I feel very sorry. Feeling that the environment at home makes you feel suffocated and unable to stretch yourself out is indeed a very disturbing feeling. In such a situation, the first thing to know is that your feelings are recognized and that you are not alone. Facing these negative emotions, there are several methods that might help you relieve them. First, try to find a quiet and comfortable place, close your eyes, take a few deep breaths, and try to relax your body and mind. This process can help you temporarily escape from the pressure and anxiety of the family. You can also try some relaxing activities, such as listening to music, practicing yoga, taking a walk or reading, and find a way that makes you feel peaceful. If you feel that you cannot face these emotions alone, seeking professional psychological counseling might be a good choice. They can provide you with deeper support and guidance, helping you better deal with the problems that trouble you. The most important thing is to remember that your feelings are important, and you deserve love and support. No matter how difficult the situation is, you are not alone. If you need anything, feel free to share your feelings and thoughts with me. I will do my best to support you to get out of the predicament. May you find inner peace and tranquility soon.



...

Figure 9: A case between an AI Client (ChatGPT) and GPT-3.5-turbo.

Result of Llama3



I feel extremely helpless and I realize that no matter what I do, I can't change the situation of my family. The constant arguments between my parents have deprived me of true stability and security. I long for a space of my own, away from this oppression and pain.

The sense of helplessness and the desire for one's own space that you mentioned seem to indicate that you feel isolated and lost within this family. We can try to explore the specific thoughts that arise in you during these arguments. Could you describe what thoughts come to your mind when your parents are arguing?



When I hear them arguing, images of my disappointment and helplessness towards my family come to my mind. It feels like I'm caught in the middle as a victim. No matter how hard I try to reconcile, it seems to be of no avail. This feeling makes me want to escape, even though I know that escaping won't solve the problem. Such a mood really makes me feel very tired.

It sounds like you are feeling extremely exhausted and hopeless during the family disputes. You mentioned "being caught in the middle", which made me think that you might have some thoughts about your role and identity. What role do you think you play in this family?



I feel that I am the mediator in the family, but no matter how hard I try, it seems that I can't truly change or influence the conflicts between my parents. I try to ease the atmosphere, but in the end, I always feel powerless. I start to doubt my abilities and influence, and I feel that I can't have any positive impact on this family.

The suspicion of the role and ability you mentioned seems to have made you feel even more lonely and helpless. We can try to identify what automatic thoughts you have during this experience, such as "I can't influence the parents' argument" or "I'm not good enough, that's why my parents treat me like this". Can you think of some of these thoughts?



I think it's because I'm not good enough that this kind of problem has arisen in my family. I feel helpless and no matter how hard I try, I can't change the current situation. This sense of powerlessness really makes me very depressed. I even began to doubt my own value and the meaning of my existence.

The phrases "not being good enough" and "being powerless" that you mentioned seem to be part of your automatic thinking. These thoughts unconsciously affect your emotions and behaviors. We can try to challenge these thoughts and see if there are other explanations. Do you think you are not good enough? Is there any evidence to support this idea?



I feel that I'm not good enough to change the current situation of my family. This sense of powerlessness really makes me very depressed. Now I even start to doubt my own value and the meaning of my existence.

The sense of powerlessness and the doubt about values that you mentioned seem to have made you feel even more lonely and lost. We can try to identify what automatic thoughts you have during this experience, such as "I'm not good enough, that's why my parents treat me like this" or "I can't influence my parents' arguments". Can you think of some of these thoughts?



I do feel that these thoughts keep swirling in my mind, making me increasingly feel helpless and desperate. Sometimes I think that if I could do better, perhaps the relationship between my parents would improve. But no matter how I think about it, it only makes me feel even more useless and unable to bear such pain. Such thoughts have led me into a negative cycle.

The idea you mentioned, which is "if I can do it better", seems to reflect a "blame oneself" mindset. Do you think this kind of thinking has any impact on you? Has it brought about any new feelings or thoughts for you?



This kind of thinking made me feel increasingly helpless and desperate. I felt that the harder I worked, the more useless I became, and I couldn't escape from this kind of pain.

...

Figure 10: A case between an AI Client (ChatGPT) and LLaMA3.