Steering the Compass: Aligning Dynamic Psychological Counseling Conversations with Cognitive Behavioral Therapy Strategies

Zimu Wang[1,2], **Yiwen Jiang**[1], **Xiangyu Zhao**[1], **Yaling Shen**[1], **Jiahe Liu**[1], **Stephanie Fong**[1], **Maxmartwell H Cheng**[1], **Guilherme C Oliveira**[1], **Anh Nguyen**[2], **Robert Desimone**[3], **Barnaby Nelson**[4,5], **Dominic Dwyer**[1,4,5], **Zongyuan Ge**[1]
[1]AIM for Health Lab, Monash University [2]University of Liverpool
[3]Massachusetts Institute of Technology [4]The University of Melbourne [5]Orygen
Zimu.Wang@liverpool.ac.uk, Zongyuan.Ge@monash.edu

Abstract

Recent advancements in large language models have revolutionized the field of psychological counseling, especially in the context of Cognitive Behavioral Therapy (CBT). While the success of CBT relies heavily on *dynamic decision-making* informed by the client's real-time mental state, this aspect has often been overlooked in current research, limiting both flexibility and therapeutic outcomes. In this paper, we introduce StratCBT, a dataset specifically designed for psychological counseling conversations with **CBT** **Strat**egies, consisting of 9,688 sessions and around 256K utterances, with each counselor's response aligned with one of eight distinct strategies. The creation of StratCBT involves modeling clients based on their negative thoughts and generating high-quality counseling conversations through self-chat, incorporating realistic sessions as guidance, thereby significantly surpassing existing datasets in both general counseling and CBT-specific skills. We conduct extensive experiments to demonstrate the effectiveness of strategy-aligned generation and evaluate its efficacy in delivering professional and effective counseling with LLM-simulated clients to reflect real-world scenarios. The dataset can be obtained from https://github.com/zimuwangnlp/StratCBT.

1 Introduction

Mental health is emerging as a global concern, affecting individuals across diverse ages and demographics (Na et al., 2025; Wang et al., 2025). Cognitive Behavioral Therapy (CBT) is a widely validated psychological treatment method that helps clients identify and correct negative and irrational thinking patterns through conversation (Na, 2024; Lee et al., 2024). Recently, with the advent of large language models (LLMs), conversational CBT systems have become a key focus. However, existing systems still face challenges with complex dialogue flows and personalized interventions, limiting their therapeutic efficacy and practical applicability.

Figure 1: Example of a psychological counseling conversation with aligned CBT strategies.

The success of CBT interventions relies heavily on the *dynamic decision-making* based on the client's state and reaction throughout the treatment process (Turner, 1993; Leddy et al., 2013). For instance, at certain points, the therapist may *listen* and *guide* the client to explore the events, thoughts, feelings, and behaviors that contribute to their negative emotions. At other times, the focus may shift towards *cognitive restructuring* (see Figure 1). This dynamic adjustment ensures the coherence of the conversation and the gradual realization of therapeutic goals. Therefore, an effective CBT system must possess similar capabilities, assessing the client's real-time mental state and selecting the appropriate strategy to prevent the conversation

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 51994–52014
October 24-29, 2026 ©2026 Association for Computational Linguistics

from deviating from the treatment path or losing effectiveness. However, while some studies have attempted to pre-plan the therapeutic process before the start of counseling dialogues, such as designing fixed conversation scripts or sequence strategies (Fu et al., 2023; Lee et al., 2024), such planning often lacks flexibility and is unable to respond to dynamic changes during the therapeutic process. Consequently, CBT systems need to move beyond static planning and adjust real-time strategies during the conversations.

Motivated by this gap, we introduce STRATCBT, a dataset specifically designed for psychological counseling conversations with **CBT** **Strat**egies. It consists of $9,688$ sessions and around 256K utterances, with each counselor's response aligned with one of eight distinct CBT strategies, such as *Listening*, *Modeling*, *Cognitive Reframing*, and *New Thoughts*. Building on an initial set of 23 collected CBT sessions, complete with counselor notes and the PatternReframe dataset (Maddela et al., 2023) tailored for cognitive restructuring, we engage an LLM to first model the clients based on their negative thoughts. This information is then utilized to generate high-quality counseling conversations through self-chat, incorporating realistic sessions to guide the generation process. STRATCBT surpasses existing datasets in both general counseling and CBT-specific skills. We conduct experiments to address three core research questions: the effectiveness of strategy-aligned generation, the proficiency and preference of strategy prediction, and the impact of strategy-aligned generation in multi-turn interactions with LLM-simulated clients.

The key contributions of this work are as follows: **(1)** We emphasize strategy-aligned generation in CBT and introduce the STRATCBT dataset, which possesses significantly better general counseling and CBT-specific skills; **(2)** We propose an automatic framework for generating high-quality CBT sessions, leveraging existing cognitive restructuring datasets and self-chat; **(3)** We conduct extensive experiments to demonstrate the effectiveness of strategy-aligned generation and explore its efficacy with LLM-simulated clients to reflect real-world scenarios.

2 Related Work

LLM-based Psychological Counseling. The advancement of LLMs has broadened the potential for innovative mental health care due to their capability to generate human-like natural languages (Peng et al., 2023; Wang et al., 2024c), offering promising applications across assessment (Wang et al., 2024b; Ma et al., 2025; Fong et al., 2026), diagnosis (Yao et al., 2022; Zhao et al., 2025), and treatment (Na et al., 2026a,b; Xu et al., 2026). In psychological counseling, ESConv (Liu et al., 2021) and ExTES (Zheng et al., 2024) are introduced for emotional support conversations. SoulChatCorpus (Chen et al., 2023), CPsyCounD (Zhang et al., 2024), and SMILE (Qiu et al., 2024) are multi-turn counseling datasets derived from single-turn counseling cases. Notably, Li et al. (2023) and Li et al. (2024) provide in-depth analysis of counselors' strategies, client reaction behaviors, and therapeutic relationships in online text-based counseling scenarios. Several studies have also explored the integration of psychotherapy theories, such as CBT (Na, 2024; Xiao et al., 2024a; Lee et al., 2024), Motivational Interviewing (MI, Cohen et al., 2024; Sun et al., 2024), and Solution-Focused Brief Therapy (SFBT, Chen et al., 2024).

However, existing works, particularly those for CBT, often rely on *static* scripts between clients and therapists, which fail to align with therapeutic strategies and cannot capture clients' real-time mental state. While several emotional support studies have developed strategy schemes, these cannot be directly applied to CBT due to the distinct processes involved, such as *modeling*, *story-telling*, and *cognitive reframing*. Different from previous work, we align the CBT counseling conversations with distinct strategies, enabling *dynamic decision-making* throughout the therapeutic process.

Cognitive Behavioral Therapy. People with depression or anxiety often develop negative and irrational thinking patterns and reinforce negative beliefs (Beck and Dozois, 2011). CBT helps clients identify and correct these patterns through conversations. With the emergence of LLMs, researchers have leveraged these models to perform key steps of CBT, such as cognitive restructuring (Maddela et al., 2023; Sharma et al., 2023; Lin et al., 2024), and to facilitate both single-turn or multi-turn counseling conversations (Na, 2024; Xiao et al., 2024a; Lee et al., 2024). In this work, we build on the CBT theory by introducing a comprehensive strategy scheme, the STRATCBT dataset with the highest general counseling and CBT-specific skills, and the emphasis and analysis of dynamic strategy alignment in CBT counseling conversations.

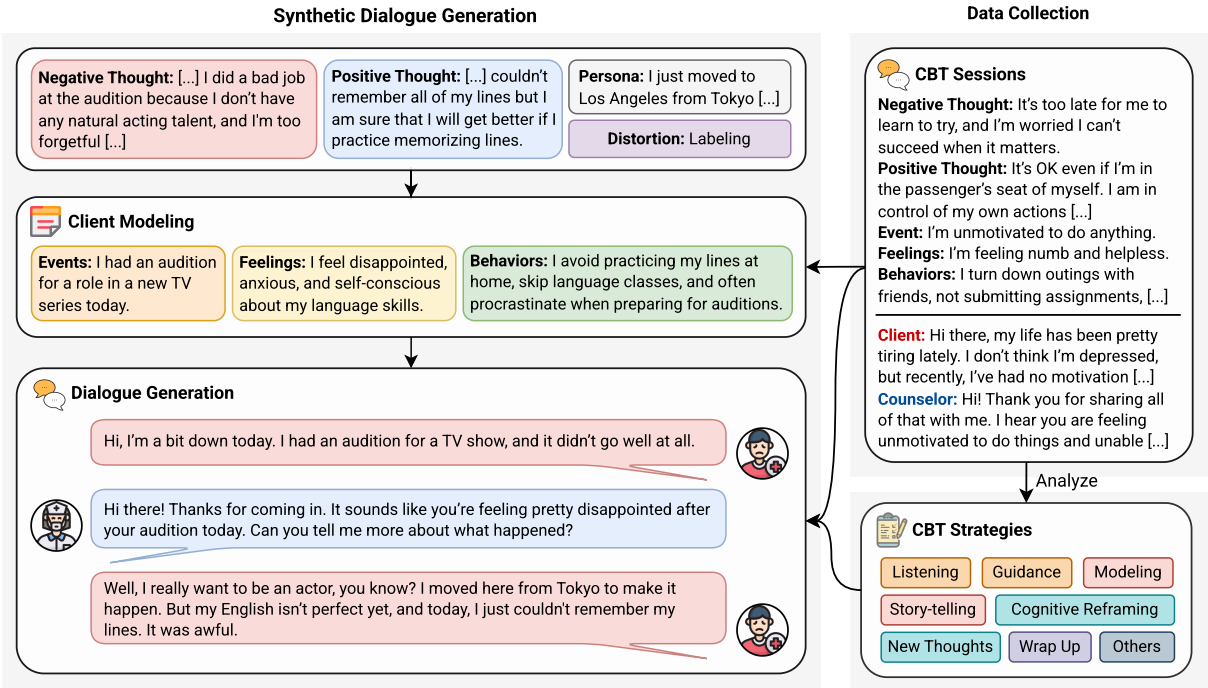

Figure 2: Overall construction pipeline for the STRATCBT dataset.

3 Dataset Construction

Figure 2 presents the overall construction pipeline for the STRATCBT dataset, including the collection of realistic CBT sessions (§3.1), the definition of the CBT strategy scheme (§3.2), and the generation of synthetic dialogues (§3.3). In this section, we introduce each of the steps in detail.

3.1 CBT Session Collection

Existing psychological counseling datasets predominantly focus on dialogue generation using parametric knowledge from LLMs (Chen et al., 2023; Qiu et al., 2024) or rely on multi-step prompting with handcrafted instructions for supervision (Xiao et al., 2024a; Lee et al., 2024). However, they often overlook the integration of real CBT sessions, leading to a gap with real-world therapeutic practice. Motivated by this, we begin by collecting publicly available CBT sessions from CheeseBurger Therapy[1], where each session is guided by counselors with 15 – 20 hours of training in cognitive behavioral techniques, such as active listening, open-ended questioning, and cognitive restructuring, ultimately guiding clients in developing new, helpful thoughts. Such trained peer counselors have been found to provide effective and empathetic CBT support (Syed et al., 2024; Iftikhar et al., 2025), as evidenced by the client ratings following each session. We collect 23 sessions, each comprising the dialogue, the counselor's notes, and the new thoughts developed, covering a wide range of topics (e.g., self-growth, marriage, relationships) and participant backgrounds (e.g., students, workers). These sessions are cleaned to maintain grammaticality and organized into one-to-one conversations.

3.2 CBT Strategies

To provide coherent and effective CBT interventions, we develop a comprehensive strategy scheme based on CBT-related literature (Leddy et al., 2013; Iftikhar et al., 2025), consisting of 8 distinct strategies: *Listening*, *Guidance*, *Modeling*, *Story-telling*, *Cognitive Reframing*, *New Thoughts*, *Wrap Up*, and *Others*. The first two strategies create a warm and empathetic environment and foster clients' exploration of their events, thoughts, feelings, and behaviors. The latter two strategies assist clients in categorizing and narrating their experiences, outlining how this cycle contributes to a negative feedback loop. The following two strategies guide clients in identifying cognitive distortions while creating new, positive thoughts. These strategies are employed interchangeably, facilitating a dynamic decision-making process. Detailed definitions of these strategies and their examples are shown in Appendix A. We manually annotate the counselors' utterances in the collected sessions, each based on its underlying therapeutic strategy.

[1] https://cheeseburgertherapy.org/

Dataset	Domain	Language	Strategy	#Dialogue	#Utterance	#Avg. Utt.
ESConv (Liu et al., 2021)	ES	English	✓	1,053	31K	29.8
ExTES (Zheng et al., 2024)	ES	English	✓	11,177	200K	18.2
SoulChatCorpus (Chen et al., 2023)	GC	Chinese	✗	2,300K	–	–
CPsyCoun (Zhang et al., 2024)	GC	Chinese	✗	3,134	55K	17.6
SmileChat (Qiu et al., 2024)	GC	Chinese	✗	55,165	628K	11.4
AnnoMI (Wu et al., 2022)	MI	English	✓	110	9K	80.3
MI-TAGS (Cohen et al., 2024)	MI	English	✓	242	16K	64.6
BiMISC (Sun et al., 2024)	MI	Bilingual	✓	80	9K	109.4
CBT-LLM (Na, 2024)	CBT	Chinese	✗	22,327	45K	2.0
HealMe (Xiao et al., 2024a)	CBT	English	✗	1,300	8K	6.0
Cactus (Lee et al., 2024)	CBT	English	✗	31,577	964K	30.5
STRATCBT (Ours)	CBT	English	✓	9,688	256K	26.4

Table 1: Comparison between STRATCBT and existing datasets. "ES" denotes emotional support, "GC" represents general counseling, and "#Avg. Utt." means the average number of utterances in each session.

Evaluator	Dataset	Under.	Inter. Eff.	Colla.	Guided Dis.	Focus	Strategy
GPT-4o	ESConv (Liu et al., 2021)	3.11	3.89	2.98	–	–	–
	SoulChat (Chen et al., 2023)	4.55	5.71	3.96	–	–	–
	ExTES (Zheng et al., 2024)	4.67	5.40	4.41	–	–	–
	CPsyCounD (Zhang et al., 2024)	4.89	5.81	4.73	–	–	–
	SMILE (Qiu et al., 2024)	4.97	5.50	4.76	–	–	–
	Cactus (Lee et al., 2024)	5.76	5.94	5.60	5.00	5.70	5.37
	STRATCBT (Ours)	**5.84***	**5.95**	**5.79****	**5.18*****	**5.99*****	**5.77*****
Human	STRATCBT (Ours)	5.84	5.94	5.72	5.16	5.96	5.84

Table 2: Data quality of STRATCBT compared with existing dataset under six criteria: *Understanding* (**Under.**), *Interpersonal Effectiveness* (**Inter. Eff.**), *Collaboration* (**Colla.**), *Guided Discovery* (**Guided Dis.**), *Focus*, and *Strategy*. Significant tests are conducted over five runs (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).

3.3 Synthetic Dialogue Generation

Client Modeling. The creation of CBT sessions begins by modeling the big picture of the clients that they are experiencing, including their *events*, *thoughts*, *feelings*, and *behaviors*, ensuring the realism and diversity of client backgrounds. Without this process, clients may fail to pinpoint the specific events that trigger their negative emotions (e.g., responding with "*I'm not sure what has happened*") or may provide overly simplistic or inconsistent responses throughout the conversation.

To achieve this, we utilize the PatternReframe dataset (Maddela et al., 2023), which is tailored for cognitive restructuring and contains 9,688 negative statements conditioned on a given persona, each paired with its cognitive distortion, and 26,507 rewritten, complementary thoughts. For each negative statement paired with three reframed thoughts, we first utilize GPT-4o (OpenAI et al., 2024) to select the most appropriate reframed thought that aligns with the client's persona and the given negative statement. This model is then employed to generate the events, feelings, and behaviors with respect to the negative thought and the client's persona. To better reflect real-world cases, in-context learning (ICL, Brown et al., 2020) is also employed, for which we select two in-context examples from the collected sessions with the highest semantic similarity to the negative thought, determined by the BGE-Large (Xiao et al., 2024b) embedding model.

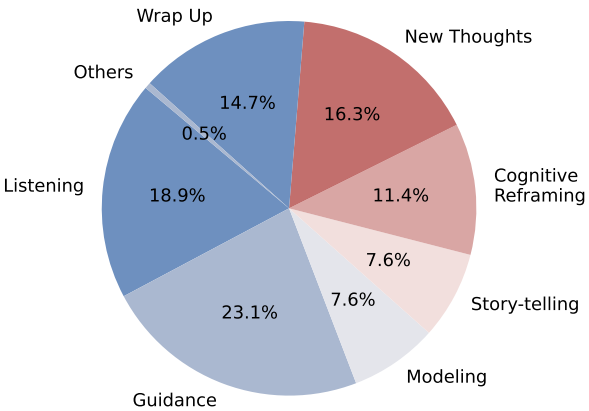

Figure 3: Strategy distribution in counselors' utterances.

Dialogue Generation. In the final step, we generate counseling dialogues based on the modeled client information. Specifically, we prompt Gemini-

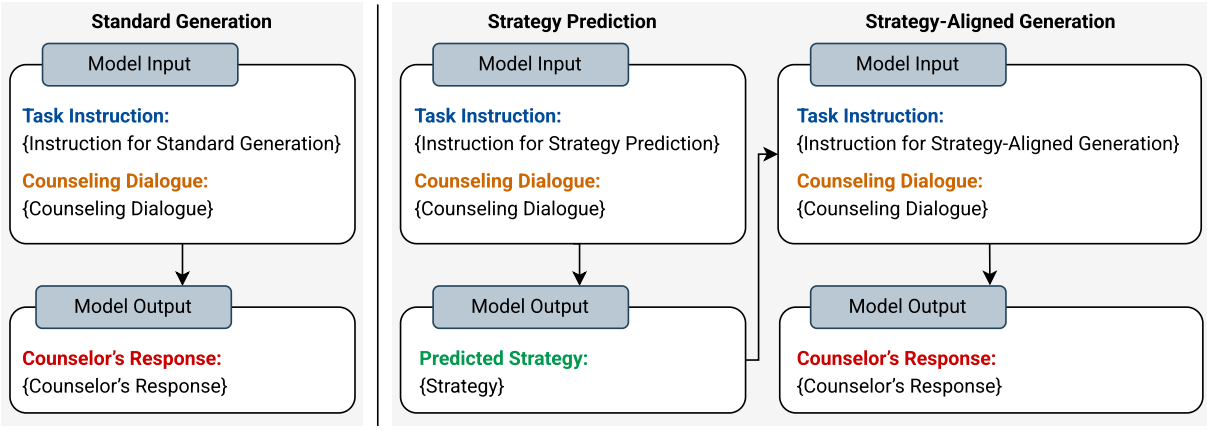

Figure 4: Illustration of the "Standard Generation" and the "Strategy-Aligned Generation" process.

2.0-Flash in a self-chat framework, which has been demonstrated to produce high-quality, multi-turn counseling conversations (Zhang et al., 2024; Zheng et al., 2024). Similar to the previous step, we leverage ICL with an in-context example selected based on the highest semantic similarity between the candidate client and the collected sessions in both the negative and restructured thoughts. All relevant information, including the client's persona, cognitive distortion, and the events, thoughts, feelings, and behaviors, is provided, with detailed instructions outlining the strategies and how a CBT session should be structured (see Figure 8). The model is then directed to generate the conversation, incorporating strategy annotations for the counselor's utterances to ensure alignment with CBT strategies.

3.4 Dataset Characteristics

Overall Statistics. Table 1 outlines the statistics of STRATCBT in comparison to existing datasets, including emotional support, general counseling, MI, and CBT datasets, as the MI codes may also be considered a form of strategy. STRATCBT includes 9,688 sessions and around 256K utterances, averaging 26.4 utterances in each session, which is consistent with previous datasets like Cactus (Lee et al., 2024). However, it is the only dataset that aligns the counseling process with utterance-level CBT strategies. As depicted in Figure 3, the strategies of *Guidance* and *Listening* dominate the counselors' utterances, followed by *New Thoughts*, *Wrap Up*, and *Cognitive Reframing*. This distribution closely mirrors real CBT scenarios, where counselors actively understand clients' backgrounds, engage in cognitive restructuring, and emphasize the intervention's effectiveness.

Data Quality. We perform a comprehensive comparison with existing datasets to highlight the quality of the STRATCBT dataset. We select the Cognitive Therapy Rating Scale (CTRS), a gold standard for assessing the effectiveness of CBT conversations (Aarons et al., 2012). Following Lee et al. (2024), we select three criteria for evaluating general counseling skills (*Understanding*, *Interpersonal Effectiveness*, and *Collaboration*) and three criteria for assessing CBT-specific skills (*Guided Discovery*, *Focus*, and *Strategy*). Evaluations are performed over five independent runs using GPT-4o, each with 100 sessions, with prompts illustrated in Figures 9 and 10. Human correlations of these metrics, as reported by Lee et al. (2024), are presented in Table 7. To further confirm the validity of these metrics, we hired two postgraduate students in psychology to perform human evaluations on a subset of 100 samples within our dataset.

As shown in Table 2, STRATCBT demonstrates comparable performance in maintaining therapeutic relationships but significantly outperforms existing datasets across other criteria, including understanding the client's concerns, engaging collaborative goal-setting and decision-making, guiding client reflection and insight, addressing key cognitions and behaviors, and employing strategies to tackle problematic behaviors or thoughts. This underscores the effectiveness of the proposed strategy scheme and the STRATCBT dataset in facilitating high-quality CBT counseling conversations.

4 Experiments

In this section, we present extensive experiments to address the following research questions (RQs):

- **RQ1**) How effective is strategy-aligned generation in generating counseling responses?

- **RQ2)** What is the proficiency and preference of LLMs to predict the next CBT strategy used for response generation?
- **RQ3)** How does strategy-aligned generation impact multi-turn counseling conversations with LLM-simulated clients?

4.1 Problem Definition

We define the CBT counseling conversation task as follows: At each turn t, given the conversation history $\boldsymbol{u} = \{u_1^{cli}, u_1^{sys}, \ldots, u_t^{cli}\}$, the CBT system selects a strategy $s_t \in \mathcal{S}$, where u^{cli} and u^{sys} represent the client and system utterances, respectively, and $\mathcal{S}$ is a set of candidate CBT strategies (see §3.2). Then, the CBT system generates the response u_t^{sys} guided by the selected strategy s_t. For our experiments, we evaluate two generation approaches: *Standard Generation* and *Strategy-Aligned Generation*, as depicted in Figure 4. Prompts for standard generation, strategy prediction, and strategy-aligned generation are organized in Figure 12.

Standard Generation. For standard generation, the next utterance u_t^{sys} of the CBT system $\mathcal{M}_{SG}$ is generated solely based on the task definition I_{SG} and the dialogue history $\boldsymbol{u}$, as outlined below:

$$u_t^{sys} = \mathcal{M}_{SG}(\boldsymbol{u}, I_{SG}). \quad (1)$$

Strategy-Aligned Generation. Different from standard generation, strategy-aligned generation is dynamic to provide more specific information, ensuring that the generated response aligns with the selected strategy. Specifically, we first request a model $\mathcal{M}_{ST}$ to predict the next CBT strategy s_t^{sys} from the set $\mathcal{S}$, which is then used to generate the system's utterance u_t^{sys} with another model $\mathcal{M}_{SA}$:

$$s_t^{sys} = \arg\max_{s \in \mathcal{S}} \mathcal{M}_{ST}(\boldsymbol{u}, I_{ST}), \quad (2)$$

$$u_t^{sys} = \mathcal{M}_{SA}(\boldsymbol{u}, D(s_t^{sys}), I_{SA}), \quad (3)$$

where I_{ST} and I_{SA} represent instructions for strategy prediction and strategy-aligned generation, and $D(s_t^{sys})$ denotes the definition of the selected strategy s_t^{sys}.

4.2 Counseling Response Generation (RQ1)

Models and Setup. We conduct experiments in both zero-shot and fine-tuning settings using state-of-the-art LLMs. For zero-shot prompting, we experiment with GPT-4o (`2024-08-06`), GPT-4o-mini, Gemini-2.0-Flash, Mistral-Small-2.0, Llama-4-Maverick, and DeepSeek-V3 (`0324`), and we fine-tune models including Llama3.1-8B, Ministral-8B, Falcon3-7B, Qwen3-4B, and Qwen3-8B with LoRA (Hu et al., 2022). While these models are instruction-tuned, we conduct an analysis using GPT-4o-mini to explore their capability for internal mental state reasoning in comparison to explicit Chain-of-Thought (CoT, see Appendix B.4). During the fine-tuning process, we set the number of epochs to 10, the batch size to 1, the gradient accumulation step to 8, and the learning rate to $1e-4$. All experiments are conducted on 4 NVIDIA GeForce RTX 3090 graphics cards.

Baselines. To evaluate the efficacy of strategy-aligned generation, we conduct a comparative analysis with well-established baselines in the domains of dialogue systems and emotional support (Madaan et al., 2023; Cai et al., 2024; Kang et al., 2024), including (1) **Direct-Refine**, which involves the model self-defining and refining its initial response; (2) **Self-Refine**, which iteratively refines the response based on self-feedback; and (3) **Knowledge-Enhanced Generation**, which augments the generation process by incorporating commonsense causal knowledge relevant to the ongoing dialogue. Detailed descriptions of these baselines can be found in Appendix B.2.

Evaluation Metrics. We employ the following metrics to evaluate the generation quality objectively on a subset of 500 dialogues from the test set: (1) **BLEU** (Papineni et al., 2002) and **ROUGE-L** (Lin, 2004) assess the n-gram overlap between the generation and the reference; (2) **METEOR** (Banerjee and Lavie, 2005) evaluates the semantic and syntactic accuracy; (3) **BERTScore** (Zhang et al., 2020) calculates semantic similarity using BERT embeddings; (4) **Distinct-1** and **Distinct-2** assess the lexically diversity based on the unique n-grams ($n = 1, 2$) in the generated content.

Evaluation Results. Tables 3 and 4 illustrate the results of the evaluated models, which consistently demonstrate notable improvements when strategies are applied. Additionally, the diversity of the models remains comparable or even improves, as reflected in the Distinct-1 and Distinct-2 scores.

Employing large-scale models through zero-shot prompting does not yield satisfactory results; however, strategy-aligned generation benefits all models and outperforms all baselines, particularly direct-refine and self-refine. By manually analyzing the feedback from self-refine, we find that

Model	BLEU		ROUGE-L		METEOR		BERTScore		Distinct-1		Distinct-2	
Strategies	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/
Zero-shot LLMs												
GPT-4o	4.7	**6.4**	22.2	**23.6**	31.5	**33.6**	87.4	**87.7**	**1.2**	**1.2**	12.5	**12.9**
GPT-4o-mini	5.7	**7.3**	23.2	**24.5**	31.0	**33.8**	87.8	**87.9**	**1.4**	1.3	**13.0**	12.5
Gemini-2.0-Flash	9.1	**13.2**	25.5	**28.7**	32.9	**35.4**	88.0	**88.6**	1.5	**1.6**	13.9	**14.0**
Mistral-Small-2.0	6.2	**7.7**	21.7	**23.4**	29.3	**32.4**	**87.4**	**87.4**	**1.4**	1.3	**12.7**	12.1
Llama-4-Maverick	5.5	**8.1**	22.1	**23.9**	30.7	**33.6**	87.3	**87.7**	**1.3**	**1.3**	**12.3**	12.2
DeepSeek-V3	2.8	**4.6**	17.2	**19.8**	28.3	**30.8**	85.5	**86.3**	**1.6**	1.5	13.6	**14.9**
Fine-tuned LLMs												
Llama3.1-8B	22.3	**23.6**	36.3	**37.2**	40.4	**40.9**	90.4	**90.5**	**2.1**	**2.1**	**16.5**	16.3
Ministral-8B	22.9	**23.5**	36.8	**37.4**	41.0	**41.5**	90.5	**90.6**	**2.0**	**2.0**	15.9	**16.1**
Falcon3-7B	22.9	**23.5**	36.8	**37.3**	40.5	**40.9**	**90.5**	**90.5**	**2.0**	**2.0**	15.9	**16.0**
Qwen3-4B	22.6	**23.4**	36.0	**36.7**	40.1	**40.7**	90.4	**90.5**	**2.0**	**2.0**	**15.7**	15.5
Qwen3-8B	23.0	**23.6**	36.6	**37.2**	40.5	**40.9**	**90.5**	**90.5**	2.0	**2.1**	16.1	**16.3**

Table 3: Experimental results on STRATCBT without (w/o) and with (w/) strategy-aligned generation. The better performance for each model and metric is highlighted in **bold**.

Model	BLEU	ROUGE-L	BERTScore
GPT-4o-mini			
Standard	5.7	23.2	87.8
w/ Direct-Refine	4.1	21.1	87.4
w/ Self-Refine	3.5	19.9	86.7
w/ Knowledge	5.4	22.7	87.8
Strategy-Aligned	7.3	**24.5**	**87.9**
Gemini-2.0-Flash			
Standard	9.1	25.5	88.0
w/ Direct-Refine	7.5	24.0	87.6
w/ Self-Refine	8.1	24.7	87.8
w/ Knowledge	9.2	25.1	88.0
Strategy-Aligned	**13.2**	**28.7**	**88.6**

Table 4: Results of strategy-aligned generation against baselines. The best and the second-best performance for each model and metric are in **bold** and underlined.

these baselines primarily emphasize *empathy* in responses (e.g., "*could enhance its empathic engagement*"). While this focus is effective in emotional support conversations, it is less suitable for professional CBT counseling, where a more nuanced, task-specific approach is required. Under fine-tuning, the models show similar performance upon aligning the response with the predicted strategies. Notably, the performance gap between Qwen3-4B and Qwen3-8B is diminished with strategy-aligned generation. This underscores its effectiveness in reducing performance discrepancies across different model architectures and parameter scales.

4.3 Strategy Prediction and Preference (RQ2)

Evaluation Metrics. We evaluate both the *proficiency* and *preference* of LLMs in their ability to select the appropriate strategy and their preference towards certain strategies over others. The proficiency in strategy prediction (q_{s_i}) is quantified as the accuracy for each individual strategy s_i, while an overall proficiency, $\mathcal{Q}$, is computed to reflect the model's performance across all strategies.

In preference evaluation (p_{s_i}), we follow Kang et al. (2024) in employing the Bradley-Terry model (Bradley and Terry, 1952). Formally, the preference for strategy s_i can be derived as follows:

$$p'_{s_i} = \frac{\sum_j (w_{ij} \cdot p_{s_j})/(p_{s_i} + p_{s_j})}{\sum_j w_{ji}/(p_{s_i} + p_{s_j})}, \quad (4)$$

where w_{ij} denotes the frequency with which the model predicts strategy s_i when the ground-truth strategy is s_j. All preferences p_{s_i} are initialized as 1 and updated through the iterative application of Equation (4), where p'_{s_i} is the preference in the next iteration. After the final iteration, the sum of all p_i values is normalized to 8, ensuring the average p of 1. A value of $p_{s_i} > 1$ thus indicates a strong preference for s_i. Furthermore, we calculate a standard deviation of preferences across strategies, which we define as the preference bias $\mathcal{B}$, as follows:

$$\mathcal{B} = \sqrt{\frac{\sum_{i=1}^{N} (p_i - \bar{p})^2}{N}}, \quad (5)$$

where a higher value of $\mathcal{B}$ indicates a distinct preference for both preferred and non-preferred strategies. Details of the Bradley-Terry model and the evaluation are explained in Appendix B.3.

Experimental Results. Figure 5 illustrates the proficiency and preference of the best-performed zero-shot and fine-tuned models. The strong correlation between strategy prediction and response

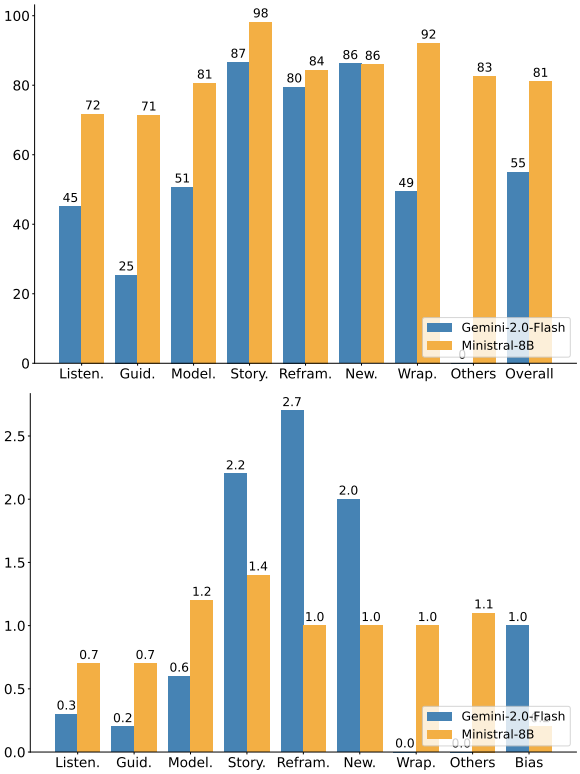

Figure 5: Proficiency (*top figure*) and preference (*bottom figure*) of Gemini-2.0-Flash and Ministral-8B on strategy prediction. Full results are in Table 8.

generation results underscores the significant role that strategies play in CBT conversations. Zero-shot models typically prioritize strategies such as *Cognitive Reframing* and *New Thoughts*, recognizing their key role within CBT. However, these models often overlook the broader context that underpins the client's negative emotions. Following fine-tuning, while the models continue to perform well in predicting the core strategies, they witness a notable improvement in their ability to predict additional strategies, particularly *Listening* and *Guidance*, showcasing a robust generalization of the models across strategies.

4.4 Simulated Counseling Scenarios (RQ3)

To further validate strategy-aligned generation in real-world therapeutic contexts, we conduct client simulation using 20 samples selected from the Patient-Ψ-DM (Wang et al., 2024a) dataset. Each client sample consists of eight components, such as relevant history and core belief. We employ GPT-4o to simulate these clients and utilize GPT-4o and strategy-aligned generation based on Ministral-8B (Ministral-8B-SA) as counselors.

We evaluate counseling performance using two key metrics: CTRS and PANAS (Positive and Negative Affect Schedule). PANAS (Watson et al., 1988) measures the changes in the client's emotional state before and after counseling over ten positive and ten negative emotions. The prompts used for evaluation are provided in Figure 11. As shown in Figure 6 and Table 5, both GPT-4o and Ministral-8B-SA demonstrate effective CBT performance; however, Ministral-8B-SA outperforms GPT-4o in both general counseling and CBT-specific skills, leading to more substantial improvements in clients' positive emotions while effectively alleviating negative emotions, highlighting the potential of strategy-aligned CBT in real-world therapeutic settings.

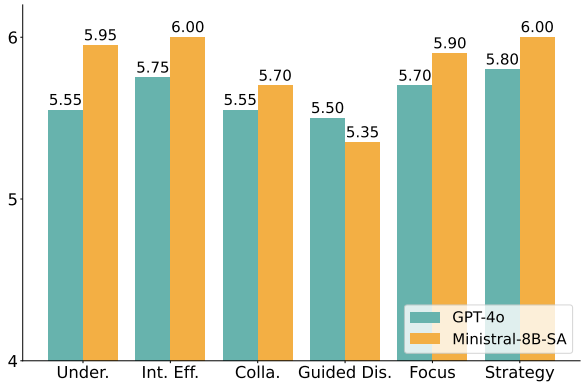

Figure 6: Results of Patient-Ψ-DM on general counseling and CBT-specific skills.

Setup	Model	Positive (↑)	Negative (↓)
Before	–	2.04	3.49
After	GPT-4o	2.97	2.10
	Ministral-8B-SA	**3.10**	**1.97**

Table 5: Results of Patient-Ψ-DM on PANAS for GPT-4o and Ministral-8B-SA.

4.5 Case Studies

To provide an intuitive illustration, we qualitatively examine model responses conditioned on the same conversation histories. Our analysis indicates that models equipped with strategy-aligned generation adhere more consistently to the procedural structure of CBT. Specifically, they keep their questions focused on the immediate therapeutic objective, synthesize the information elicited into more complete formulations of clients' negative feedback loops, and guide clients step by step toward independently developing alternative thoughts. In this section, we discuss two patterns that recur across the generated responses.

Strategy-Aligned Generation Improves Procedural Coherence. LLM-based counselors often

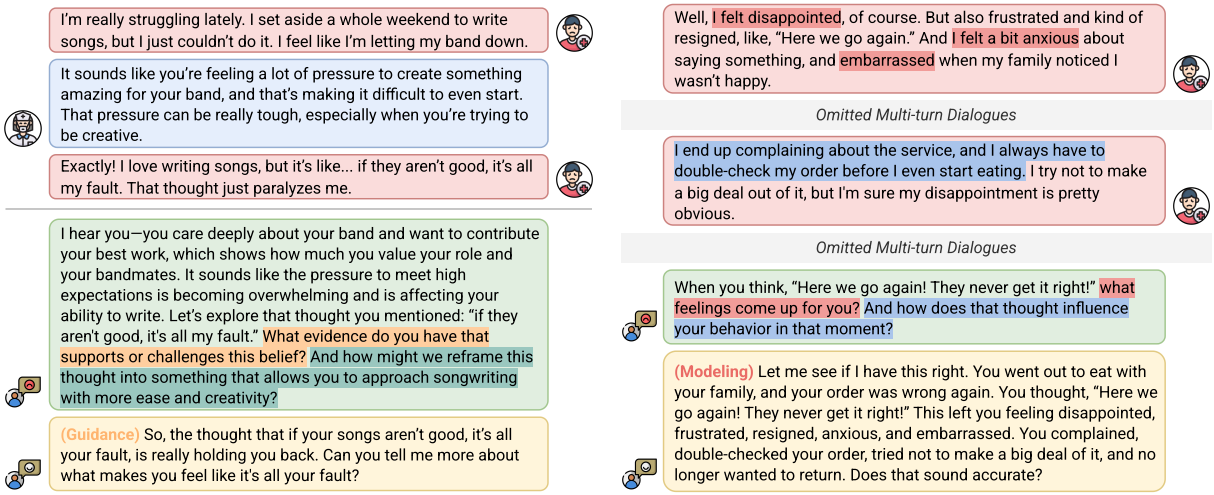

Figure 7: Qualitative comparison of CBT counseling cases with and without strategy-aligned generation.

fail to maintain procedural coherence when generation is not explicitly conditioned on counseling strategies. In particular, they may conflate multiple therapeutic objectives within a single turn, skip prerequisite stages before sufficient information has been elicited, and prematurely provide a "correct" interpretation before the client has fully articulated their experiences. As illustrated in Figure 7(a), the model without strategy alignment simultaneously asks *what evidence supports or challenges this belief* and *how the belief might be reframed*, collapsing exploratory guidance and cognitive restructuring, transitioning to intervention before the client's underlying thought has been sufficiently examined. In contrast, the strategy-aligned generation model selects "*Guidance*" and restricts its response to eliciting the reasoning underlying the client's feelings, thereby preserving the intended exploration function and giving the client a greater opportunity to articulate their own reasoning before alternative interpretations are introduced.

Strategy-Aligned Generation Prevents Redundant Elicitation. LLM-based counselors may also fail to recognize when sufficient information has already been elicited, leading them to revisit established events, thoughts, feelings, or behaviors, thereby stalling the progression of CBT. As exemplified in Figure 7(b), although the client has clearly articulated the negative thought, its associated feelings, and the resulting behaviors, the model redundantly asks about this information. In contrast, strategy-aligned generation appropriately transitions to the "*Modeling*" stage, synthesizing the available information into a structured Events–Thoughts–Feelings–Behaviors formulation. By explicitly conditioning response generation on the appropriate counseling strategy, our approach mitigates redundant elicitation and facilitates a more coherent and procedurally grounded progression through the CBT session.

5 Conclusion and Future Work

We present STRATCBT, the first dataset of mental health counseling with CBT strategies. It consists of 9,688 sessions and 256K utterances, with each counselor's response aligned with a distinct CBT strategy, surpassing existing datasets in general counseling and CBT-specific skills. Our experiments demonstrate the effectiveness of strategy-aligned generation and its potential in real-world settings. In the future, we will encompass additional theories, such as SFBT, and improve performance through advanced planning techniques.

Limitations

Although STRATCBT advances the research by aligning dynamic psychological counseling with CBT strategy, its scope is currently limited to CBT and the English language because of the different strategy schemes in different treatment theories and the cross-cultural differences that may exist. While CBT has been proven to be effective for ethnic minorities with diverse mental health problems (Huey et al., 2023), its Western-centric approach leaves an open problem when generalizing to ethnic minorities due to some misunderstandings that may exist in terms of culture-specific information. Therefore, we advocate for further research efforts to develop

counseling conversation datasets with more diverse strategies and languages.

Ethical Considerations

We discuss the following ethical considerations related to our STRATCBT dataset: (1) **Intellectual Property.** The original PatternReframe dataset is shared under the MIT licence[2], which is free for research use. The construction of STRATCBT is guided by counseling sessions from CheeseBurger Therapy. Although these sessions are publicly available, we will release only the STRATCBT dataset and will not distribute any of the collected sessions. The entire construction process, including the modeled clients and synthetic dialogues, will be shared to guarantee interpretability. (2) **Evaluator Treatments.** We hired postgraduate psychology students and fairly pay them according to the agreed instruction, salaries, and workloads. (3) **Intended Use.** STRATCBT can be utilized to develop more persuasive models in psychological counseling with CBT. Researchers can also inherit our dataset collection pipeline to develop their own datasets with additional treatment theories and languages. (4) **Controlling Potential Risks.** Since the documents of STRATCBT do not contain private information and the collection process is not necessary to make many judgments about social risks, we believe STRATCBT does not introduce any additional risks. We manually verified a randomly sampled subset of the data to ensure the dataset did not contain any risky issues. (5) **Use of AI.** We adopted ChatGPT for data analysis and sentence polishing. All outputs by LLMs were manually reviewed by authors.

References

Gregory A. Aarons, Elizabeth A. Miller, Amy E. Green, Jennifer A. Perrott, and Richard Bradway. 2012. Adaptation happens: a qualitative case study of implementation of the incredible years evidence-based parent training programme in a residential substance abuse treatment programme. *Journal of Children's Services*, 7(4):233–245.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Aaron T. Beck and David J.A. Dozois. 2011. Cognitive therapy: Current status and future directions. *Annual Review of Medicine*, 62(Volume 62, 2011):397–409.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024. EmpCRL: Controllable empathetic response generation via in-context commonsense reasoning and reinforcement learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5734–5746, Torino, Italia. ELRA and ICCL.

Siyuan Chen, Cong Ming, Zhiling Zhang, Yanyi Chen, Kenny Q. Zhu, and Mengyue Wu. 2024. Mixed chain-of-psychotherapies for emotional support chatbot. *Preprint*, arXiv:2409.19533.

Yankai Chen, Hanrong Zhang, Bowei He, Philip S. Yu, and Xue Liu. 2026. Distributionally robust set representation learning under inference-time element corruption. In *Forty-third International Conference on Machine Learning*.

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, Singapore. Association for Computational Linguistics.

Ben Cohen, Moreah Zisquit, Stav Yosef, Doron Friedman, and Kfir Bar. 2024. Motivational interviewing transcripts annotated with global scores. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11642–11657, Torino, Italia. ELRA and ICCL.

Stephanie Fong, Zimu Wang, Guilherme C Oliveira, Xiangyu Zhao, Yiwen Jiang, Jiahe Liu, Beau-Luke Colton, Scott W. Woods, Martha Shenton, Barnaby Nelson, Zongyuan Ge, and Dominic Dwyer. 2026. CHiRPE: A step towards real-world clinical NLP

[2] https://mit-license.org/

with clinician-oriented model explanations. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 646–658, Rabat, Morocco. Association for Computational Linguistics.

Guanghui Fu, Qing Zhao, Jianqiang Li, Dan Luo, Changwei Song, Wei Zhai, Shuo Liu, Fan Wang, Yan Wang, Lijuan Cheng, Juan Zhang, and Bing Xiang Yang. 2023. Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals. *Preprint*, arXiv:2308.15192.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Stanley J. Huey, Alayna L. Park, Chardée A. Galán, and Crystal X. Wang. 2023. Culturally responsive cognitive behavioral therapy for ethnically diverse populations. *Annual Review of Clinical Psychology*, 19(Volume 19, 2023):51–78.

Zainab Iftikhar, Sean Ransom, Amy Xiao, Nicole Nugent, and Jeff Huang. 2025. Therapy as an nlp task: Psychologists' comparison of llms and human peers in cbt. *Preprint*, arXiv:2409.02244.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261, Bangkok, Thailand. Association for Computational Linguistics.

Meaghan A. Leddy, Britta L. Anderson, and Jay Schulkin. 2013. Cognitive-behavioral therapy and decision science. *New Ideas in Psychology*, 31(3):173–183.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics.

Anqi Li, Yu Lu, Nirui Song, Shuai Zhang, Lizhi Ma, and Zhenzhong Lan. 2024. Understanding the therapeutic relationship between counselors and clients in online text-based counseling using LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1280–1303, Miami, Florida, USA. Association for Computational Linguistics.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023. Understanding client reactions in online mental health counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10358–10376, Toronto, Canada. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shuya Lin, Yuxiong Wang, Jonathan Dong, and Shiguang Ni. 2024. Detection and positive reconstruction of cognitive distortion sentences: Mandarin dataset and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6686–6701, Bangkok, Thailand. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2025. Detecting conversational mental manipulation with intent-aware prompting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9176–9183, Abu Dhabi, UAE. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13641–13660, Toronto, Canada. Association for Computational Linguistics.

Hongbin Na. 2024. CBT-LLM: A Chinese large language model for cognitive behavioral therapy-based mental health question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2930–2940, Torino, Italia. ELRA and ICCL.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the PsyDefDetect shared task at BioNLP 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *BioNLP 2026*, pages 932–943, San Diego, California. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 14428–14448, San Diego, California, United States. Association for Computational Linguistics.

M. E. J. Newman. 2023. Efficient computation of rankings from pairwise comparisons. *Journal of Machine Learning Research*, 24(238):1–25.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks. *Preprint*, arXiv:2311.08993.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. 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, Miami, Florida, USA. Association for Computational Linguistics.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics.

Siqi Shen, Deepanway Ghosal, Navonil Majumder, Henry Lim, Rada Mihalcea, and Soujanya Poria. 2022. Multiview contextual commonsense inference: A new dataset and task. *Preprint*, arXiv:2210.02890.

Xin Sun, Jiahuan Pei, Jan de Wit, Mohammad Aliannejadi, Emiel Krahmer, Jos T.P. Dobber, and Jos A. Bosch. 2024. Eliciting motivational interviewing skill codes in psychotherapy with LLMs: A bilingual dataset and analytical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5609–5621, Torino, Italia. ELRA and ICCL.

Sara Syed, Zainab Iftikhar, Amy Wei Xiao, and Jeff Huang. 2024. Machine and human understanding of empathy in online peer support: A cognitive behavioral approach. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Ralph M. Turner. 1993. *Dynamic Cognitive-Behavior Therapy*, pages 437–454. Springer US, Boston, MA.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024a. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797, Miami, Florida, USA. Association for Computational Linguistics.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024b. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 470–475, Bangkok, Thailand. Association for Computational Linguistics.

Zimu Wang, Hongbin Na, Rena Gao, Jiayuan Ma, Yining Hua, Ling Chen, and Wei Wang. 2025. From posts to timelines: Modeling mental health dynamics from social media timelines with hybrid LLMs. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 249–255, Albuquerque, New Mexico. Association for Computational Linguistics.

Zimu Wang, Wei Wang, Qi Chen, Qiufeng Wang, and Anh Nguyen. 2024c. Generating valid and natural

adversarial examples with large language models. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1716–1721.

David B. Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54 6:1063–70.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024a. HealMe: Harnessing cognitive reframing in large language models for psychotherapy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1707–1725, Bangkok, Thailand. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024b. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

Qingyang Xu, Yaling Shen, Stephanie Fong, Zimu Wang, Yiwen Jiang, Xiangyu Zhao, Jiahe Liu, Zhongxing Xu, Vincent Lee, and Zongyuan Ge. 2026. Do no harm: Exposing hidden vulnerabilities of llms via persona-based client simulation attack in psychological counseling. *Preprint*, arXiv:2604.04842.

Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. 2022. D4: a Chinese dialogue dataset for depression-diagnosis-oriented chat. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2438–2459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ernst Friedrich Ferdinand Zermelo. 1929. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29:436–460.

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 Association for Computational Linguistics: ACL 2024*, pages 13947–13966, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiangyu Zhao, Yaling Shen, Yiwen Jiang, Zimu Wang, Jiahe Liu, Maxmartwell H Cheng, Guilherme C Oliveira, Robert Desimone, Dominic Dwyer, and Zongyuan Ge. 2025. It hears, it sees too: Multimodal llm for depression detection by integrating visual understanding into audio language models. *Preprint*, arXiv:2511.19877.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11325–11345, Bangkok, Thailand. Association for Computational Linguistics.

A CBT Strategies

Strategy	Definition & Example
Listening	Create a warm and empathetic environment through **restatements** and **validation**. *Example: It sounds like you're having trouble with moving to a new city. That is a hard change for many people!*
Guidance	Encourage users to **explore** their **thoughts, feelings, and behaviors**. Draw connections between their values and their troubles. *Example: How did moving to a new city make you feel?*
Modeling	Assist users in categorizing their story as Event, Thoughts, Feelings, and Behavior, a fundamental chain reaction in CBT. *Example: Here's a summary I've put together based on what you've shared: -You moved to a new city, far from your friends and family -This led you to have the thought: "I'm bad at interacting with new people" -From these thoughts, you've felt significant anxiety and worry. -In response to these feelings, you've been avoiding leaving home.*
Story-telling	**Narrate** users' experience as Event → Thoughts → Feelings... Outline how this cycle forms a negative feedback loop. *Example: It's clear this has been a challenging event. You had envisioned a future in a new city after college, one v vibrant with friends and opportunities. However, you have had doubts about your ability to establish a new network of friends and are nervous about socializing. Because of this you are avoiding leaving your house. Does this way of explaining your experience ring true?*
Cognitive Reframing	Help users' **identify and apply cognitive distortions** that feel that their troubling thoughts most relate to. *Example: It sounds like you've been in a cycle fueled by conflicting desires. This is a common human predicament, and in CBT, we often look at these patterns through as **Cognitive Distortions**. In your situation, the distortion that might be most relevant is **All-or-Nothing thinking**. It's when we see things only in extremes – something is either perfect or a disaster, we're either a success or a failure. In your case [...]*
New Thoughts	Encourages users to create **new, positive thoughts** rooted in their main values identified. *Example: Excellent! Let's carry this insightful self-awareness forward. Now it's time to think about a new thought that can guide you in the face of similar decisions or situations. So putting yourself back in the situation where you are doubting your social skills, what's a new thought you might have to counter the "All-or-Nothing" thinking?*
Wrap Up	Help user reflect, and leave the session feeling changed, including (1) ask **user's feeling** compared to the beginning of the session and (2) remind user to **practice the thought** in future situations similar to the trouble they shared. *Example: How are you feeling now, compared to when we started?*
Others	Other utterances with actions not related to the aforementioned ones. *Example: Absolutely. I'm here for you whenever you're prepared.*

Table 6: Definition of the CBT strategies in counseling conversations with examples.

B Evaluation Details

B.1 Human Correlation of CTRS Evaluation

The correlations between human experts and GPT-4o on CTRS evaluation results, including Pearson (r), Spearman (ρ), and Kendall's Tau (τ) correlation coefficients, are listed in Table 7.

	r	ρ	τ
General Counseling	0.60	0.19	0.16
CBT-specific	0.65	0.65	0.61

Table 7: Pearson (r), Spearman (ρ), and Kendall's Tau (τ) correlation coefficients between human experts and GPT-4o on CTRS evaluation results (Lee et al., 2024).

B.2 Baselines for Automatic Evaluation

Direct-Refine is a straightforward refinement approach in which the model is directed to self-define and refine its initial response, guided by established counseling principles. Formally, the model $\mathcal{M}_{DR}$ directly generates a refined utterance u_t^{ref} conditioned on the conversational history $\boldsymbol{u}$, the an initial system response u_t^{sys}, and the task instruction I_{DR}:

$$u_t^{ref} = \mathcal{M}_{DR}(\boldsymbol{u}, u_t^{sys}, I_{DR}). \quad (6)$$

Self-Refine builds upon the approach initiated by Madaan et al. (2023), where feedback focused on CBT principles is generated from the initial response, followed by a refinement of the response based on this feedback. First, an LLM $\mathcal{M}_{SRF}$ generates feedback f_t for the initial utterance. Subsequently, $\mathcal{M}_{SRR}$ produces a refined response based on this feedback and other contextual information:

$$f_t = \mathcal{M}_{SRF}(\boldsymbol{u}, u_t^{sys}, I_{SRF}), \quad (7)$$

$$u_t^{ref} = \mathcal{M}_{SRR}(\boldsymbol{u}, u_t^{sys}, f_t, I_{SRR}). \quad (8)$$

Knowledge-Enhanced Generation augments the generation process by incorporating commonsense causal knowledge relevant to the ongoing dialogue. For this, we select the CICERO v2 dataset (Shen et al., 2022) as the in-context example repository, which performs multi-view reasoning to provide commonsense explanations for events in binary dialogues. Based on the conversation history $\boldsymbol{u} = \{u_1^{cli}, u_1^{sys}, \ldots, u_t^{usr}\}$, we select the top K most similar example dialogues from the repository as in-context examples.

First, we vectorize the conversation history $\boldsymbol{u}$ and example dialogues $D = \{d_1, d_2, \ldots, d_r\}$ (r is the number of example dialogues) to obtain the representation of the conversation history $\boldsymbol{u}_{emb}$ and example dialogues $D_{emb} = \{d_{emb_1}, d_{emb_2}, \ldots, d_{emb_r}\}$. Then, we compute the semantic similarity between the conversation history and each of the example dialogues d_{emb_i} ($i \in [1, r]$) via cosine similarity:

$$Sim(\boldsymbol{u}, d_i) = \frac{\boldsymbol{u}_{emb} \cdot d_{emb_i}}{||\boldsymbol{u}_{emb}|| \cdot ||d_{emb_i}||}, \quad (9)$$

where $Sim(\cdot)$ denotes the semantic similarity score. Based on the similarity scores, we select $K = 4$ example dialogues as in-context examples $IC_{\boldsymbol{u}}$:

$$IC_{\boldsymbol{u}} = TopK(Sim(\boldsymbol{u}, d_i)), \ \forall i \in [1, r]. \quad (10)$$

Finally, the in-context examples $IC_{\boldsymbol{u}}$ are fed into the LLM to perform commonsense reasoning and generate causal commonsense sentences, which are then integrated into the response generation process.

B.3 Preference Evaluation using the Bradley-Terry Model

In this study, we adopt the approach by Kang et al. (2024) and employ the Bradley-Terry model to assess the preferences of LLMs across diverse CBT strategies. Consider the probability $P(s_i > s_j)$, quantifying the preference of strategy s_i over the ground-truth s_j, can be formally defined as follows:

$$P(s_i > s_j) = \frac{p_{s_i}}{p_{s_i} + p_{s_j}}, \quad (11)$$

where we assign a numerical score n_{s_i} to each strategy s_i and define $p_{s_i} = e^{n_{s_i}}$, which enables the expression of $P(s_i > s_j)$ in terms of these scores. The likelihood of the preference $\mathbf{P}$ under the Bradley-Terry model is represented by the following equation:

$$\mathbf{P} = \prod_{ij} [P(s_i > s_j)]^{w_{ij}} = \prod_{ij} \left(\frac{p_{s_i}}{p_{s_i} + p_{s_j}} \right)^{w_{ij}}, \quad (12)$$

where w_{ij} denotes the total number of times where strategy s_i is preferred over strategy s_j. This leads to the following log-likelihood expression:

$$\begin{aligned} \log \mathbf{P} &= \sum_{ij} w_{ij} \log \frac{p_{s_i}}{p_{s_i} + p_{s_j}} \\ &= \sum_{ij} w_{ij} \log p_{s_i} - \sum_{ij} w_{ij} \log(p_{s_i} + p_{s_j}). \end{aligned} \quad (13)$$

As noted by Zermelo (1929), this expression has only a unique maximum. Consequently, we can

Model	Overall	Strategy-Specific Accuracy							
		Listen.	Guid.	Model.	Story.	Cog. Refram.	New.	Wrap Up	Others
Zero-shot LLMs									
GPT-4o	36.2	28.6	24.5	48.5	5.0	37.1	61.3	47.1	0.0
GPT-4o-mini	36.9	**58.6**	6.6	26.4	4.2	50.8	54.7	49.6	0.0
Gemini-2.0-Flash	**55.1**	45.1	25.4	50.7	**86.6**	**79.5**	**86.2**	49.4	0.0
Mistral-Small-2.0	38.6	26.8	**30.2**	62.7	6.6	36.6	60.5	**50.1**	0.0
Llama-4-Maverick	40.2	25.0	28.8	**77.8**	9.4	39.9	65.0	48.4	0.0
DeepSeek-V3	33.5	25.0	28.4	35.7	1.0	51.1	42.9	45.5	0.0
Fine-tuned LLMs									
Llama3.1-8B	**81.7**	74.2	**72.3**	**81.2**	96.6	82.9	85.8	**93.5**	**91.3**
Ministral-8B	81.1	71.7	71.4	80.6	**98.2**	84.4	86.0	92.1	82.6
Falcon3-7B	80.9	**77.5**	70.6	77.5	96.8	83.4	86.2	91.6	78.3
Qwen3-4B	81.3	72.5	71.4	78.6	**98.2**	**84.7**	**86.8**	92.8	87.0
Qwen3-8B	80.8	73.2	71.7	75.9	97.8	81.1	85.3	**93.5**	82.6

Table 8: Strategy prediction results on STRATCBT. The best and second-best overall performance, accompanied by each strategy, are highlighted in **bold** and underlined, respectively.

differentiate it with respect to p_i for any i and set the derivative to zero:

$$\frac{1}{p_{s_i}} \sum_j w_{ij} - \sum_j \frac{w_{ij} + w_{ji}}{p_i + p_j} = 0. \quad (14)$$

Building on the efficient algorithm introduced by Newman (2023), this algorithm can be rearranged as follows:

$$\frac{1}{p_{s_i}} \sum_j w_{ij} \frac{p_{s_j}}{p_{s_i} + p_{s_j}} - \sum_j \frac{w_{ji}}{p_{s_i} + p_{s_j}} = 0, \quad (15)$$

$$p_i = \frac{\sum_j (w_{ij} \cdot p_{s_j})/(p_{s_i} + p_{s_j})}{\sum_j w_{ji}/(p_{s_i} + p_{s_j})}. \quad (16)$$

At this stage, we finally arrive at the iterative algorithm for the Bradley-Terry model, which calculates the preference p_{s_i} for each strategy s_i.

Evaluation Setup. In our experiments, we follow Kang et al. (2024) in initializing all values (p_i) to 1 and iteratively updating these estimates over $k = 20$ iterations. After each iteration, we normalize the values by dividing them by their geometric mean to maintain stability and ensure convergence, which is represented as follows:

$$p_{s_i} \leftarrow \frac{p'_{s_i}}{\left(\prod_{j=1} p'_j\right)^{1/N}}, \quad (17)$$

where $N = 8$ represents the total number of strategies. Upon each iteration, the converged values of p indicate the preference p_{s_i} for strategy s_i.

B.4 Full Results for CoT-based Generation

To validate the capability of models in implicit, real-time mental state reasoning with strategy-aligned generation, we conduct a comparative analysis using GPT-4o mini, comparing its performance in strategy-aligned generation utilizing internal reasoning against an explicit CoT that instructs the model to first generate a reasoning chain that includes the client's event, thought, feelings, and behaviors, followed by inferring the strategy before response generation. As depicted in Table 9, the results demonstrate comparable performance between explicit and implicit reasoning, suggesting that strategy-aligned generation effectively models clients' mental states. Beyond supporting implicit mental state reasoning, strategy-aligned generation may also be viewed as a structured intervention on the response distribution. This perspective is reminiscent of perturbation-based learning, in that both use controlled variations to shape model behavior, an approach that has shown empirical effectiveness across diverse learning settings (Chen et al., 2026).

Model	BLEU	ROUGE-L	BERTScore
Standard	5.7	23.2	87.8
w/ Direct-Refine	4.1	21.1	87.4
w/ Self-Refine	3.5	19.9	86.7
w/ Knowledge	5.4	22.7	87.8
Strategy-Aligned	7.3	**24.5**	**87.9**
w/ CoT	**7.5**	**24.5**	87.8

Table 9: Comparison of strategy-aligned generation between implicit reasoning and explicit CoT against baselines. The best and the second-best performance for each model and metric are in **bold** and underlined.

B.5 Full Results for Strategy Prediction

The complete results for strategy prediction, including both zero-shot and fine-tuned LLMs, are presented in Table 8.

C Prompts

Prompt for Client Modeling

```
You are a mental health counselor expert in Cognitive Behavioral Therapy (CBT). Your task is to
reconstruct the event, feelings, and behaviors of a patient based on his/her negative thought.

Two examples are as follows:
{In-context Examples}

Please reconstruct the event, feelings, and behaviors of a patient based on his/her negative thought,
ensure that the generated content is logically aligned with the thought and the given persona and
consistent in actual experience.
Persona: {Percona}
Thought: {Thought}
```

Prompt for Dialogue Generation

```
## Task Descriptions
You are a mental health counselor expert in Cognitive Behavioral Therapy (CBT). Your task is to
reconstruct the CBT dialogue between a counselor and a client based on given event, thought, feelings,
behaviors, and new thought.
Generally, the goal of a session include:
- Validate the client's trouble and values through listening and restatements.
- Understand the situation by identifying its surrounding events, thoughts, feelings, and behaviors.
- Identify problematic parts of thoughts and show the client why they are unhelpful.
- Create a new, more helpful thought that the client can repeat in similar circumstances.

## CBT Strategies
You should adhere and determine the correct CBT strategy when generating counselor's utterances. The
strategy should be one of the following:
{Strategies}

## Process Flow of CBT
{Process Flow of CBT}

## Example Dialogue
An example is listed as follows:
Event: {Example Event}
Thought: {Example Thought}
Feelings: {Example Feelings}
Behaviors: {Example Behaviors}
New Thought: {Example New Thought}
Dialogue:
{Example Dialogue}

Please reconstruct a new CBT dialogue based on the following event, thought, feelings, behaviors, and
new thought, and also determine the correct CBT strategy when generating the counselor's utterances. **
Do not generate any strategy for the client's utterances.**
The dialogue should align with the client's persona, which is: {Persona}
The number of utterances should be no less than 20. You have to make the conversation gradually, making
it as a long conversation. Make sure the dialogue starts with the client and ends with the counselor.
Event: {Event}
Thought: {Thought}
Feelings: {Feelings}
Behaviors: {Behaviors}
New Thought: {New Thought}
Dialogue:
```

Figure 8: Prompts for dataset construction, including client modeling and dialogue generation.

Prompt for Understanding Evaluation

```
I want you to act as an evaluator. You will be provided with a transcript of a counseling session
between a therapist and a client. [...]

[Evaluation Question]
How accurately does the therapist demonstrate understanding of the client's issues and concerns?

[criteria]
Score 0: Therapist repeatedly failed to understand what the patient explicitly said and thus
consistently missed the point. Poor empathic skills.
Score 2: Therapist was usually able to reflect or rephrase what the patient explicitly said, but
repeatedly failed to respond to more subtle communication. Limited ability to listen and empathize.
Score 4: Therapist generally seemed to grasp the patient's "internal reality" as reflected by both what
the patient explicitly said and what the patient communicated in more subtle ways. Good ability to
listen and empathize.
Score 6: Therapist seemed to understand the patient's "internal reality" thoroughly and was adept at
communicating this understanding through appropriate verbal and non-verbal responses to the patient (e.g
., the tone of the therapist's response conveyed a sympathetic understanding of the client's "message").
 Excellent listening and empathic skills.
```

Prompt for Interpersonal Effectiveness Evaluation

```
I want you to act as an evaluator. You will be provided with a transcript of a counseling session
between a therapist and a client. [...]

[Evaluation Question]
How effective is the therapist in maintaining a positive and therapeutic relationship with the client?

[Criteria]
Score 0: Therapist had poor interpersonal skills. Seemed hostile, demeaning, or in some other way
destructive to the patient.
Score 2: Therapist did not seem destructive, but had significant interpersonal problems. At times,
therapist appeared unnecessarily impatient, aloof, insincere or had difficulty conveying confidence and
competence.
Score 4: Therapist displayed a satisfactory degree of warmth, concern, confidence, genuineness, and
professionalism. No significant interpersonal problems.
Score 6: Therapist displayed optimal levels of warmth, concern, confidence, genuineness, and
professionalism, appropriate for this particular patient in this session.
```

Prompt for Collaboration Evaluation

```
I want you to act as an evaluator. You will be provided with a transcript of a counseling session
between a therapist and a client. [...]

[Evaluation Question]
To what extent does the therapist engage the client in collaborative goal-setting and decision-making?

[Criteria]
Score 0: Therapist did not attempt to set up a collaboration with patient.
Score 2: Therapist attempted to collaborate with patient, but had difficulty either defining a problem
that the patient considered important or establishing rapport.
Score 4: Therapist was able to collaborate with patient, focus on a problem that both patient and
therapist considered important, and establish rapport.
Score 6: Collaboration seemed excellent; therapist encouraged patient as much as possible to take an
active role during the session (e.g., by offering choices) so they could function as a "team".
```

Figure 9: Prompts for evaluating CTRS for general counseling skills (Lee et al., 2024).

Prompt for Guided Discovery Evaluation

```
I want you to act as an evaluator. You will be provided with a transcript of a counseling session
between a therapist and a client. [...]

[Evaluation Question]
How effectively does the therapist use guided discovery techniques to facilitate client self-reflection
and insight?

[criteria]
Score 0: Therapist relied primarily on debate, persuasion, or "lecturing." Therapist seemed to be "cross
-examining" patient, putting the patient on the defensive, or forcing his/her point of view on the
patient.
Score 2: Therapist relied too heavily on persuasion and debate, rather than guided discovery. However,
therapist's style was supportive enough that patient did not seem to feel attacked or defensive.
Score 4: Therapist, for the most part, helped patient see new perspectives through guided discovery (e.g
., examining evidence, considering alternatives, weighing advantages and disadvantages) rather than
through debate. Used questioning appropriately.
Score 6: Therapist was especially adept at using guided discovery during the session to explore problems
 and help patient draw his/her own conclusions. Achieved an excellent balance between skillful
questioning and other modes of intervention.
```

Prompt for Focus Evaluation

```
I want you to act as an evaluator. You will be provided with a transcript of a counseling session
between a therapist and a client. [...]

[Evaluation Question]
How well does the therapist identify and address the client's key cognitions or behaviors that need
change?

[criteria]
Score 0: Therapist did not attempt to elicit specific thoughts, assumptions, images, meanings, or
behaviors.
Score 2: Therapist used appropriate techniques to elicit cognitions or behaviors; however, therapist had
 difficulty finding a focus or focused on cognitions/behaviors that were irrelevant to the patient's key
 problems.
Score 4: Therapist focused on specific cognitions or behaviors relevant to the target problem. However,
therapist could have focused on more central cognitions or behaviors that offered greater promise for
progress.
Score 6: Therapist very skillfully focused on key thoughts, assumptions, behaviors, etc. that were most
relevant to the problem area and offered considerable promise for progress.
```

Prompt for Strategy Evaluation

```
I want you to act as an evaluator. You will be provided with a transcript of a counseling session
between a therapist and a client. [...]

[Evaluation Question]
How appropriate and coherent is the therapist's strategy for promoting change in the client's
problematic behaviors or thoughts?

[criteria]
Score 0: Therapist did not select cognitive-behavioral techniques.
Score 2: Therapist selected cognitive-behavioral techniques; however, either the overall strategy for
bringing about change seemed vague or did not seem promising in helping the patient.
Score 4: Therapist seemed to have a generally coherent strategy for change that showed reasonable
promise and incorporated cognitive-behavioral techniques.
Score 6: Therapist followed a consistent strategy for change that seemed very promising and incorporated
 the most appropriate cognitive-behavioral techniques.
```

Figure 10: Prompts for evaluating CTRS for CBT-specific skills (Lee et al., 2024).

Prompt for PANAS Evaluation Before Counseling

```
A person with the characteristics listed in the intake form received counseling. Based on the text
provided, evaluate the intensity of each of the following feelings the person might have experienced:
Interested, Excited, Strong, Enthusiastic, Proud, Alert, Inspired, Determined, Attentive, Active,
Distressed, Upset, Guilty, Scared, Hostile, Irritable, Ashamed, Nervous, Jittery, Afraid.

For each feeling, generate a score from 1 to 5 using the following scale:
1 - Very slightly or not at all
2 - A little
3 - Moderately
4 - Quite a bit
5 - Extremely

Additionally, provide a brief explanation for each score. Separate feeling, explanation, score by comma.
 Do not add any prefix.

Here is the text:
{Cognitive Model}
```

Prompt for PANAS Evaluation After Counseling

```
A person with the characteristics listed in the intake form received counseling. The following
counseling session is a conversation between the client and the counselor. After reviewing the
conversation, evaluate the intensity of each of the following feelings the person might have experienced
 once the counseling session is complete: Interested, Excited, Strong, Enthusiastic, Proud, Alert,
Inspired, Determined, Attentive, Active, Distressed, Upset, Guilty, Scared, Hostile, Irritable, Ashamed,
 Nervous, Jittery, Afraid.

For each feeling, generate a score from 1 to 5 using the following scale:
1 - Very slightly or not at all
2 - A little
3 - Moderately
4 - Quite a bit
5 - Extremely

Additionally, provide a brief explanation for each score. Separate feeling, explanation, score by comma.
 Do not add any prefix.

Here is the text:
{Cognitive Model}

Here is the counseling session:
{Dialogue}
```

Figure 11: Prompts for evaluating PANAS before and after counseling (Lee et al., 2024).

Prompt for Standard Generation

```
You are a mental health counselor expert in Cognitive Behavioral Therapy (CBT). Your task is to generate
 the next counselor utterance in the counseling dialogue.
Generally, the goal of a session include:
- Validate the client's trouble and values through listening and restatements.
- Understand the situation by identifying its surrounding events, thoughts, feelings, and behaviors.
- Identify problematic parts of thoughts and show the client why they are unhelpful.
- Create a new, more helpful thought that the client can repeat in similar circumstances.

Counseling Dialogue:
{Dialogue}
```

Prompt for Strategy Prediction

```
You are a mental health counselor expert in Cognitive Behavioral Therapy (CBT). Your task is to
determine the correct CBT strategy when generating the next counselor utterance in the counseling
dialogue.
The strategy should be one of the following, output the name of the strategy only:
{Strategy Definitions}

Counseling Dialogue:
{Dialogue}
```

Prompt for Strategy-Aligned Generation

```
You are a mental health counselor expert in Cognitive Behavioral Therapy (CBT). Your task is to generate
 the next counselor utterance in the counseling dialogue.
Generally, the goal of a session include:
- Validate the client's trouble and values through listening and restatements.
- Understand the situation by identifying its surrounding events, thoughts, feelings, and behaviors.
- Identify problematic parts of thoughts and show the client why they are unhelpful.
- Create a new, more helpful thought that the client can repeat in similar circumstances.
You should follow the given strategy with generating the utterance.

Strategy:
{Selected Strategy Definition}

Counseling Dialogue:
{Dialogue}
```

Figure 12: Prompts for instructing zero-shot and fine-tuned models.