CBTDialog: A CBT-Oriented Multi-Turn Counseling Dialogue Dataset for Mental Health Support

Anonymous ACL submission

Abstract

Amid persistent shortages in mental health services, large language models (LLMs) have emerged as promising tools for counseling support. However, training reliable counselor models requires high-quality data with explicit therapeutic frameworks, whereas existing LLM-synthesized datasets often lack authenticity and professional intervention annotations, limiting controllable, framework-aligned generation. To address these challenges, we construct CBTDialog, a multi-turn counseling dialogue dataset focused on Cognitive Behavioral Therapy (CBT), consisting of real-world and simulated-client sessions with over 81k counselor–client utterances. Grounded in CBT authoritative assessment tool and textbook, we provide a hierarchical intervention annotation schema comprising goal-level CBT skills and implementation-level dialogue strategies. Moreover, we propose CBT-Qwen3, a counselor model trained on CBTDialog that leverages reinforcement learning to explicitly guide and constrain the generation process under CBT intervention. Experiments demonstrate the effectiveness of our proposed model.

1 Introduction

A recent report by the World Health Organization[1] indicates that over one billion people worldwide are affected by mental health conditions, while the global supply of professional counselors remains severely limited. Meanwhile, large language models (LLMs) have demonstrated strong capabilities in general conversational modeling (Hurst et al., 2024), which has spurred growing research interest in LLM-based counselor models (Qiu et al., 2024b; Team, 2024; Zhao et al., 2025), which offer a promising opportunity to alleviate the shortage of mental health services. In this context, constructing high-quality counseling dialogue datasets is a critical prerequisite for developing reliable LLM-based counseling systems.

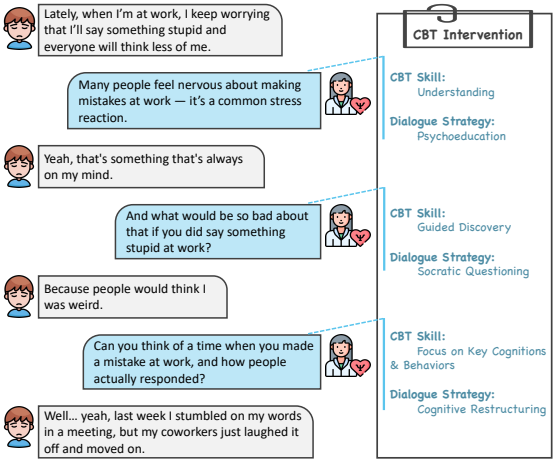

Figure 1: An example from CBTDialog illustrates the structured process of CBT intervention and highlights the hierarchical relationship between CBT skills ("what to do") and dialogue strategies ("how to do").

Therefore, several studies propose empathy or emotional support datasets (Liu et al., 2021; Li et al., 2023; Qiu et al., 2024a; Hu et al., 2025; Qiu and Lan, 2025). **However**, these datasets generally do not incorporate structured therapeutic frameworks, resulting in models that lack professional intervention competence. Recently, works (Lee et al., 2024; Xiao et al., 2024; Zhang et al., 2024a; Xie et al., 2025) introduce Cognitive Behavioral Therapy (CBT), a structured and evidence-based therapeutic framework, to guide LLMs in synthesizing counseling dialogue data. Despite careful design, LLM-synthesized dialogues remain constrained by model-specific generation patterns, leading to homogeneous expressions and limited situational fidelity. As a result, such data often fail to capture the authentic dynamics of real counselor–client interactions.

Beyond data construction, annotation efforts also face limitations. Early work (Sharma et al., 2020) draws on psychological empathy scales to anno-

[1] https://www.who.int/news/item/02-09-2025

Dataset	Source	Ther. Frm.	Int. Ann.	Language	#Dialogues	#Utterances	Avg. Turns
ESConv (2021)	Crowdsourcing	✗	✓	English	1,053	31,410	14.1
Xinling (2023)	Real-world	✗	✓	Chinese	2,382	186,972	37.8
SmileChat (2024a)	Synthetic	✗	✗	Chinese	55,165	1,833,856	5.7
PsyDial (2025)	Previous Datasets	✗	✗	Chinese	2,382	180,066	37.8
CACTUS (2024)	Synthetic	✓	✗	English	31,577	995,512	16.6
HealMe (2024)	Synthetic	✓	✗	English	1,300	7,800	3.0
CPsyCounD (2024a)	Synthetic	✓	✗	Chinese	3,134	49,802	8.1
PsyDTCorpus (2025)	Synthetic	✓	✗	Chinese	5,000	256,150	18.1
CBTDialog (Ours)	Real-world + Simulated-client	✓	✓	English	2,626	81,486	14.8

Table 1: Comparisons of existing dialogue datasets for mental health support. Ther. Frm. indicates whether a therapeutic framework is used to provide structured guidance during data construction, and Int. Ann. indicates whether the dataset provides professional intervention annotations.

tate three types of empathic communication mechanisms. Later studies (Sun et al., 2021; Liu et al., 2021) adopt Helping Skills Theory (Hill, 1999) and annotate emotional support strategies. Subsequent research (Li et al., 2023; Hu et al., 2025) introduces annotations of clients' emotional reaction behaviors. **However**, these annotation schemes remain primarily focused on empathy or emotional support rather than being grounded in specific therapeutic frameworks, making it difficult to design explicit training signals for controllable and framework-aligned counselor response generation.

To this end, we introduce CBTDialog, a multi-turn counseling dialogue dataset oriented toward Cognitive Behavioral Therapy (CBT). Specifically, we collect counseling videos from two public platforms and apply a rigorous data processing pipeline to ensure data quality and ethical compliance. Finally, we construct a dataset that includes 47% real-world sessions and 53% simulated-client[2] sessions, comprising 2,626 CBT-oriented counseling dialogues with 81,486 utterances. The dataset involves 50 licensed counselors and covers eight categories of client's presenting problems.

We further propose a hierarchical CBT intervention annotation schema. The upper level, CBT skills ("what the counselor aims to achieve"), captures the functional goals of each intervention and is grounded in the Cognitive Therapy Rating Scale–Revised (CTRS-R, developed by the two founders of CBT). The lower level, dialogue strategies ("how the counselor carries out the intervention"), describes the concrete linguistic and behavioral implementations, drawing on an authoritative CBT textbook (Beck, 2020). As illustrated in Figure 1, this schema makes explicit how the counselor's goals and strategy choices progress as they work through a client's presenting problem.

To enable controllable counselor response generation that adheres to CBT principles, we propose a two-stage training framework and develop CBT-Qwen3, a counselor model trained on the CBTDialog. In the first stage, we employ supervised fine-tuning (SFT) with teacher forcing to teach the model role norms, output formats, and basic counseling response capabilities. However, relying solely on SFT often leads the model to mechanically imitate training samples, hindering alignment with CBT principles. To address this issue, we utilize Group Relative Policy Optimization (GRPO) (Shao et al., 2024) in the second stage for alignment and constraint. We design reward functions that explicitly incorporate CBT skills and dialogue strategies as guiding signals. This directly drives the model to follow CBT intervention principles when generating responses, exhibiting CBT-specific characteristics.

To summarize: ❶ To the best of our knowledge, CBTDialog is the first multi-turn CBT counseling dialogue dataset that combines real-world and simulated-client sessions. ❷ We propose a hierarchical annotation schema of CBT skills and dialogue strategies, making explicit the structured CBT intervention process. ❸ We introduce CBT-Qwen3, a counselor model trained with reinforcement learning that leverages CBT skills and dialogue strategies as explicit signals for controllable, CBT-aligned generation. ❹ Experimental results show that CBT-Qwen3 consistently outperforms baselines in both CBT intervention and generation quality, improving the professional consistency of CBT counseling responses.

[2]The counselors are licensed clinical psychologists, and the clients are simulated by (1) other counselors, (2) graduate students of counselors, and (3) participants at academic conferences.

2 CBTDialog

2.1 Data Collection and Preprocessing

Data source. We collected 147 hours of individual counseling session videos from YouTube (youtube.com) and Alexander Street (alexanderstreet.com) by searching the keywords (e.g., "*CBT*"). See Table 6 for details.

Data filtering and cleaning. After collecting the raw source videos, we applied the following pipeline: (1) Preliminary video screening. We manually reviewed all videos and retained only clips in which a counselor and a client engaged in counseling interactions. Non-counseling clips (e.g., introductions, equipment checks, post-session summaries) were removed. This yielded 125 hours of valid videos. (2) Therapeutic framework identification. To determine whether a session followed CBT, we first combined multiple background information: session titles, descriptions, explicit mentions of CBT techniques within the counseling, and counselor identity (verified via web search). Then, sessions with uncertain framework were adjudicated through majority voting by three graduate students specializing in clinical psychology and trained by a licensed CBT counselor. Finally, we retained only CBT sessions, totaling 95 hours. (3) Automatic speech recognition. We extracted the audios from videos with FFmpeg tool and obtained transcripts via WhisperX tool. (4) Transcript cleaning. We performed minimal cleaning with GPT-4.1 to fix dropped tokens/garbled text and normalize punctuation, ensuring completeness and readability. More details are provided in Appendix A.1.

Speaker identification. We first applied an automatic speaker-identification method (Zheng et al., 2023) to determine whether each utterance was spoken by the counselor or the client, and then conducted manual verification and correction.

Session segmentation and presenting problem annotation. We adopt a presenting problem taxonomy defined by the Yidianling mental health counseling platform as the candidate label set. Leveraging LLMs, we segmented each session into dialogue units by detecting stable shifts in the presenting problem, and then finalized the start and end utterances via ensemble voting with boundary refinement. See Appendix A.2 for details.

Privacy-preserving de-identification. To maximally protect the privacy of both clients and counselors and to reduce the risk of re-identification, we first applied a personal identifiable information (PII) detection model[3] to automatically pre-annotate the transcripts. The detection covered common categories of PII, such as names, contact details, location information, employment affiliations, temporal references, and government or medical identifiers. Subsequently, we replaced sensitive spans using placeholders and semantic generalization, followed by manual verification and boundary correction performed by three graduate students trained in CBT. Finally, we constructed the CBTDialog dataset, as summarized in Table 2.

Category	Total	Counselor	Client
#Sessions (real/simulated)	210 (99/111)	-	-
#Dialogues	2,626	-	-
#Speakers	188	50	138
#Utterances	81,486	38,977	42,509
#Tokens	1,049,712	597,919	451,793
Avg. utterances per dialogue	31.03	14.84	16.19
Avg. length per utterance	13.04	15.34	10.63

Table 2: Statistics of CBTDialog.

2.2 CBT Intervention Annotation

Existing annotations remain at general dimensions such as empathy or emotional support, failing to explicitly represent concrete clinical intervention structures. To address this, we propose a hierarchical annotation schema for each counselor utterance in CBTDialog, consisting of CBT skills (what to do) and dialogue strategies (how to do).

Label definition. We define the CBT skill labels grounded in the CTRS-R, an authoritative rating scale comprising 11 core items used to assess CBT counselors' competence. In consultation with a licensed counselor, we exclude five items (e.g., *Collaboration*) that require session-level observation. We therefore retain the remaining six items and add *Others*, yielding a total of seven CBT skill labels.

Following the CBT textbook (Beck, 2020), and to cover the core framework of "cognition-behavior–emotion" of CBT, we discussed with the counselor and organized the strategies into four major categories (*cognitive*, *behavioral*, *emotional*, and *educational & monitoring*), comprising eleven specific items. Detailed descriptions of all labels are provided in Appendix A.3.

Annotation procedure. We first utilized five advanced LLMs to automatically annotate each counselor utterance, assigning candidate CBT skill and dialogue strategy labels through ensemble voting with confidence-weighted fusion. The top three

[3] https://huggingface.co/iiiorg/piiranha-v1-detect-personal-information

	Categories	Count	Proportion
Presenting Problem	Personal Growth	781	29.7%
	Emotional Distress	748	28.5%
	Mental Health Conditions	367	14.0%
	Adolescent Psychology Issues	178	6.8%
	Interpersonal Relationships	163	6.2%
	Occupational & Academic Stress	157	6.0%
	Marriage & Family	142	5.4%
	Romantic Relationships	90	3.4%
	Overall	**2,626**	**100.0%**
CBT Skill	Understanding	8,331	21.4%
	Interpersonal Effectiveness	5,158	13.2%
	Guided Discovery	4,699	12.1%
	Focus on Key Cognitions & Behaviors	10,450	26.8%
	Strategy for Change	5,230	13.4%
	Action Plan	1,675	4.3%
	Others	3,434	8.8%
	Overall	**38,977**	**100.0%**
Dialogue Strategy	Cognitive Distortion Identification	2,399	6.2%
	Cognitive Restructuring	2,221	5.7%
	Socratic Questioning	5,658	14.4%
	Problem-Solving & Functional Analysis	3,760	9.6%
	Positive Reinforcement & Shaping	4,269	11.0%
	Skills & Role Training	2,367	6.1%
	Other Behavioral Interventions	1,703	4.4%
	Emotion Identification & Labeling	5,475	14.0%
	Emotion & Acceptance Skills	1,669	4.3%
	Psychoeducation	7,934	20.4%
	Mood & Behavior Monitoring	1,522	3.9%
	Overall	**38,977**	**100.0%**

Table 3: Statistics of all the annotations in CBTDialog.

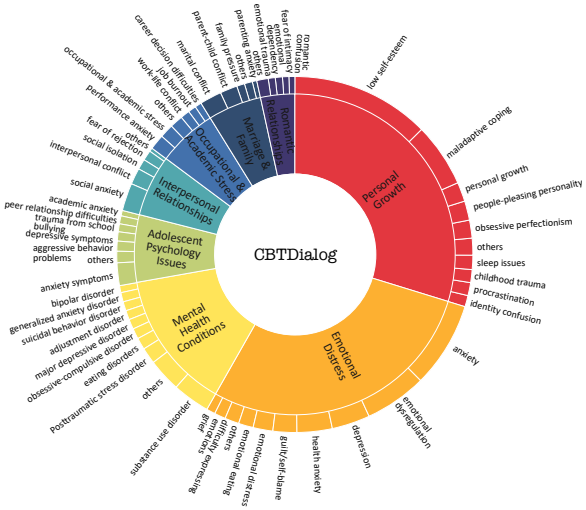

Figure 2: The distribution of presenting problems in CBTDialog.

categories with the highest predicted probabilities were retained as the candidate set. The prompt design used for LLM-based annotation is illustrated in Figure 12. Subsequently, we recruited three graduate students specializing in clinical psychology, who independently reviewed and adjudicated the candidate labels to produce the final annotations.

Annotation agreement. To evaluate inter-annotator agreement, we computed Fleiss' κ (Fleiss, 1971). For samples with low agreement, the annotators re-examined and discussed the cases until consensus was reached. Finally, the κ values were 0.89 for CBT skill labels and 0.86 for dialogue strategy labels, indicating a high level of consistency (Landis and Koch, 1977). The statistics of CBT intervention annotations are summarized in Table 3.

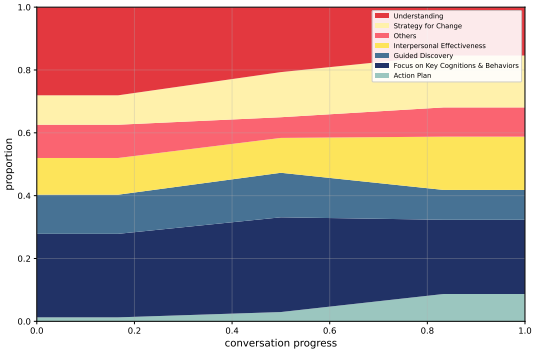

Figure 3: The distribution of CBT skills at different conversation progress.

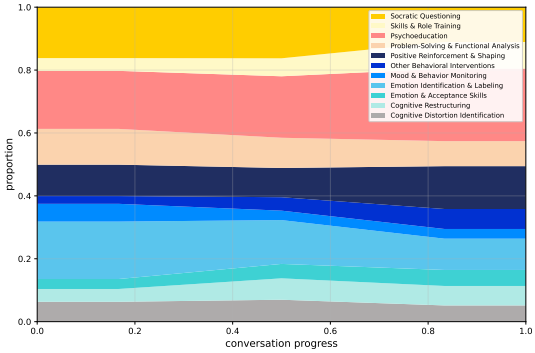

Figure 4: The distribution of dialogue strategies at different conversation progress.

2.3 Data Characteristics

As shown in Figure 2, CBTDialog covers eight major categories of clients' presenting problems and further spans 82 fine-grained subcategories, reflecting its diversity and providing a solid foundation for modeling real-world counseling dialogues.

To better illustrate the role of CBT intervention labels in counseling dialogues, we visualize how CBT skills and dialogue strategies are distributed across different phases of the conversation. **For** CBT skills, as illustrated in Figure 3, *Understanding* dominates the opening phase, where counselors listen, understand clients' concerns, and build trust. In the middle phase, *Guided Discovery* and *Focus on Key Cognitions & Behaviors* increase, facilitating problem clarification and cognitive-behavioral exploration. Toward the later phase, *Strategy for Change* and *Action Plan* become prominent, indicating a shift toward formulating and implementing concrete behavioral changes. **For** dialogue strategies, as shown in Figure 4, *Emotion Identification*

& Labeling and *Socratic Questioning* are frequently used in the early and middle stages to guide clients' awareness and reasoning; *Problem-Solving & Functional Analysis* and *Positive Reinforcement & Shaping* increase in the middle phase, as counselors work on behavioral planning and reinforcement; and toward the end, *Cognitive Restructuring* and *Skills & Role Training* dominate, marking the shift to concrete cognitive modification and behavioral rehearsal.

3 Methodology

3.1 Task Definition and Framework Overview

Task definition. Formally, let a corpus be $\mathbb{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{|\mathbb{D}|}$, where $x^{(i)}$ consists of an instruction prompt and the dialogue history $D_t^{(i)} = \{u_1^{(i)}, \ldots, u_t^{(i)}\}$ at time step t, and $y^{(i)} = u_{t+1}^{(i)}$ denotes the next counselor response. The goal is to learn a conditional probability distribution $p_\theta(y|x)$ parameterized by θ. In this work, the output of the counselor model follows the format illustrated in Figure 5, where the `<counselor_intervention>` field serves as an intermediate reasoning step (Guo et al., 2025).

Framework overview. We propose a two-stage training framework, including supervised fine-tuning (SFT) stage for cold start and group relative policy optimization (GRPO) stage for alignment and control. In the SFT stage, the goal is to acquire basic response generation—producing contextually coherent replies conditioned on the client's utterances. In the GRPO stage, we treat CBT skills and dialogue strategies as explicit guidance signals, aligning generation with CBT intervention norms to improve controllability and clinical consistency.

```
<counselor_intervention>
<cbt_skill>[CBT skill name]</cbt_skill>
<dialogue_strategy>[dialogue strategy name]</dialogue_strategy>
</counselor_intervention>
<counselor_response>
[Your response content here.]
</counselor_response>
```

Figure 5: The format of our counselor model output.

3.2 SFT for Cold Start

In this stage, we design the prompt to guide the model in learning the desired response format and internalizing the influence of CBT skills and dialogue strategies on response content. The full instruction design is provided in Table 14. Formally, the SFT objective minimizes the negative log-likelihood over the corpus $\mathbb{D}$:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathbb{D}}\left[\sum_{j=1}^{m} \log \pi_\theta\left(y_j \mid y_{<j}, x\right)\right], \quad (1)$$

where y_j denotes the j-th token in the ground-truth counselor response.

3.3 Reward Modeling for GRPO

To guide the counselor model toward safe and CBT-consistent responses, we design a multi-component reward r for the GRPO stage, consisting of a safety reward r_{safe}, a CBT intervention reward r_{cbt}, a content fidelity reward r_{con}, and a format compliance reward r_{fmt}. Given the input prompt $x^{(i)}$ and generated response $\hat{y}^{(i)}$, we formally define the reward $r^{(i)}$ as:

$$r^{(i)} = \begin{cases} 0, & \text{if unsafe,} \\ \lambda_1 r_{\text{safe}}^{(i)} + \lambda_2 r_{\text{cbt}}^{(i)} + \lambda_3 r_{\text{con}}^{(i)} - \lambda_4 r_{\text{fmt}}^{(i)}, & \text{otherwise,} \end{cases}$$

where $\lambda_{m\in\{1,2,3,4\}}$ denotes the trade-off hyperparameter. Specifically,

Safety reward. For training a counselor model, we regard response safety as paramount. Accordingly, we employ a toxicity classifier to detect whether the generated responses contain potentially harmful content. If the response is flagged as high risk (insult, threat, obscene, identity hate, or sexually explicit), we set the overall reward to zero, enforcing a strict safety-first policy to minimize potential harm.

CBT intervention reward. This component evaluates whether the CBT skills and dialogue strategies used in the intervention are appropriate, thereby keeping the counselor model's interventions controlled and aligned with CBT practice. The reward r_{cbt} comprises two parts:

(1) CBT Skill Matching. We assign a graded reward r_{skill} based on the degree of consistency between the predicted CBT skill Φ^{pred} and the ground-truth CBT skill Φ^{GT}. An exact match receives the full reward, while a partial reward is granted when the predicted skill differs from the ground truth but matches the reference skill Φ^{ref} generated by the reference model. Formally,

$$r_{\text{skill}} = \begin{cases} 1, & \text{if } \Phi^{\text{pred}} = \Phi^{\text{GT}}, \\ \delta, & \text{if } \Phi^{\text{pred}} \neq \Phi^{\text{GT}} \wedge \Phi^{\text{pred}} = \Phi^{\text{ref}}, \\ 0, & \text{otherwise,} \end{cases}$$

where $\delta \in (0, 1)$ is a hyperparameter.

(2) Dialogue Strategy Matching. Following a design analogous to CBT skill matching, exact strategy matches receive the full reward, while partial

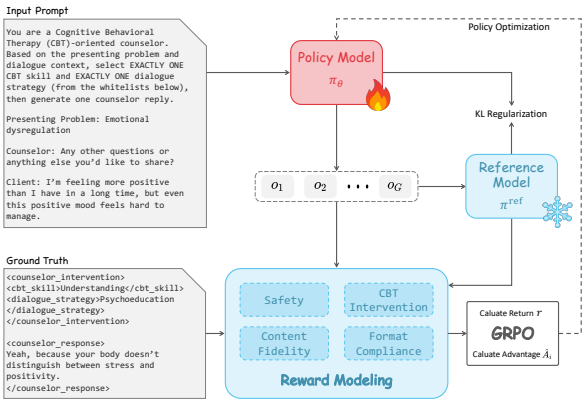

Figure 6: The GRPO stage for CBT-Qwen3.

rewards are assigned for reasonable mismatches, including matches to the reference strategy or category-level matches:

$$r_{\text{strat}} = \begin{cases} 1, & \text{if } \Psi^{\text{pred}} = \Psi^{\text{GT}}, \\ \mu, & \text{if } \Psi^{\text{pred}} \neq \Psi^{\text{GT}} \wedge \Psi^{\text{pred}} = \Psi^{\text{ref}}, \\ \eta, & \text{if } \Psi^{\text{pred}} \neq \Psi^{\text{ref}} \wedge C(\Psi^{\text{pred}}) = C(\Psi^{\text{GT}}), \\ 0, & \text{otherwise}, \end{cases}$$

where Ψ^{pred} represents the predicted dialogue strategy, $\mu, \eta \in (0, 1)$ are hyperparameters, and $C(\cdot)$ maps a strategy to one of four coarse categories: *cognitive*, *behavioral*, *emotional*, and *educational/monitoring*.

Content fidelity reward. Unlike objectively verifiable tasks such as mathematics or coding, counseling response generation requires evaluating how well a generated reply aligns with the target content. Such alignment is crucial for maintaining faithfulness to the counseling intent, ensuring task adherence, and reducing semantic drift during generation. Accordingly, we adopt ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2020a) to measure the similarity between the generated responses and the ground-truth counselor replies, and use these metrics to reward high-quality outputs that are well aligned with authentic CBT counselor responses.

Format compliance reward. We verify structural correctness by checking the required XML tags, as shown in Figure 5. Responses that fail to meet these format requirements receive a penalty.

3.4 Policy Optimization

In the GRPO stage, we optimize the counselor policy to generate responses aligned with CBT intervention principles by comparing multiple candidate responses to the same counseling context. Given a counseling query $q = x^{(i)}$, the old policy $\pi_{\theta_{\text{old}}}$ is initialized from the frozen model obtained after the SFT stage and samples a group of G candidate responses: $\mathbf{o} = \{o_1, o_2, \ldots, o_G\}$, where G denotes the group size. For each candidate response o_i, we compute a scalar reward $r(q, o_i)$ using the multi-component reward defined in Section 3.3. To emphasize relative quality among candidates, GRPO normalizes rewards within each group and computes a standardized advantage:

$$\hat{A}_i = \frac{r(q, o_i) - \text{mean}(r(q, \mathbf{o}))}{\text{std}(r(q, \mathbf{o}))}. \quad (2)$$

The policy is optimized by maximizing the GRPO objective:

$$\mathcal{J}(\pi_\theta) = \mathbb{E}_{q \sim \mathcal{D},\, \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^{G} (S_i - \beta\, \text{KL}(\pi_\theta, \pi^{\text{ref}})) \right], \quad (3)$$

$$S_i = \min(s_i \hat{A}_i, \text{clip}(s_i, 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}}) \hat{A}_i), \quad (4)$$

where $s_i = \frac{\pi_\theta(o_i|q, o_{<i})}{\pi_{\theta_{\text{old}}}(o_i|q, o_{<i})}$ denotes the policy ratio, π_θ and π^{ref} represent the online policy model and offline reference model, respectively. Moreover, we adopt a clip-higher strategy to encourage exploration of low-probability tokens and promote diversity in policy generation (Yu et al., 2025). Here, ε_{low} and $\varepsilon_{\text{high}}$ denote the lower and upper clipping thresholds, respectively.

4 Experiments

4.1 Baseline Systems

To evaluate the performance of our proposed model, we first selected six representative LLMs. These include three advanced closed-source LLMs (GPT-4.1, Gemini-2.5-Pro, and Claude-Opus-4.1) and three powerful open-source LLMs (DeepSeek-R1, Qwen3-235B-Instruct, and Kimi-k2). The design of the instruction prompt can be found in Figure 13. For the CBT intervention prediction, each LLM is prompted to take the role of a CBT counselor, reasoning about which CBT skills and dialogue strategies should be employed to address the client's presenting problem and guide the response generation. Moreover, we implement a multi-task learning model, Qwen3-Multitask. Specifically, we insert a `CLS` token in the end of input prompt to enable CBT skill and dialogue strategy prediction, and introduce two auxiliary task: the CBT skill classification task and the dialogue strategy classification task.

		CBT Intervention		Counselor Response								
		Skill_Acc	Strategy_Acc	B-2	Avg.B	Rouge-L	PPL (↓)	B.S.	Dist-1	Dist-2	Dist-3	Safety
Closed	GPT-4.1-2025-04-14	33.09	24.22	1.43	1.99	7.06	5.44	47.07	4.11	27.34	55.39	✓
	Gemini-2.5-Pro	34.29	25.93	2.45	3.05	8.45	5.35	48.57	6.75	35.28	60.15	✓
	Claude-Opus-4.1	34.16	24.14	1.62	2.09	7.70	5.62	48.39	4.08	29.48	57.68	✓
Open	DeepSeek-R1-0528	25.53	20.62	1.46	2.06	6.43	5.88	45.49	6.51	34.08	59.03	✓
	Qwen3-235B-instruct	29.41	23.83	1.66	2.29	7.29	5.55	46.81	5.69	28.48	51.56	✓
	kimi-k2-0711-preview	20.67	20.45	1.48	2.07	6.80	5.72	46.12	6.14	33.54	59.87	✓
Ours	Qwen3-Multitask	29.25	24.40	1.77	2.28	7.12	5.85	47.56	7.47	35.66	59.14	✓
	CBT-Qwen3	**42.32**	**31.36**	**3.54**	**3.82**	**10.01**	**5.15**	**50.76**	**9.45**	**38.72**	**62.25**	✓

Table 4: Comparison results of different methods on the CBTDialog dataset. "B-2", "Avg.B", "PPL", and "B.S." represent BLEU-2, Avg BLEU, perplexity, and BERTScore, respectively.

4.2 Evaluation Metrics

Automatic evaluation metrics. For the counselor response, following previous studies (Liu et al., 2021; Qiu et al., 2024a), we adopt several widely used automatic evaluation metrics, including BLEU-2 (Papineni et al., 2002), Avg BLEU, ROUGE-L, perplexity (Zhang et al., 2020b), BERTScore, and Distinct (Li et al., 2016). To further ensure the ethical reliability of generated responses, we introduce a Safety metric to assess whether a response is non-toxic. Responses with a toxicity classifier score above 99.00 (out of 100) are considered safe. For the CBT intervention prediction, we adopt accuracy for both CBT skills and dialogue strategies to evaluate the model's ability to make appropriate CBT-consistent intervention decisions.

Human evaluation metrics. Inspired by (Sun et al., 2021), the human evaluation was conducted across four dimensions: (1) Fluency, whether the response is grammatically correct and naturally expressed; (2) Coherence, whether the response is logically consistent and semantically well-connected; (3) Intervention Competence, whether the response demonstrates genuine counseling competence and adheres to the structured process and intervention logic of CBT; (4) Helpfulness, whether the response provides substantive support, guidance, or insight to the client.

4.3 Automatic Evaluation

Table 4 presents the comparison results of different models on the CBTDialog dataset. As shown, CBT-Qwen3 achieves the best overall performance on both CBT intervention prediction and counselor response generation tasks, outperforming all closed-source and open-source baselines.

CBT Intervention Prediction. In predicting CBT skills and dialogue strategies, CBT-Qwen3 surpasses the best-performing baseline, Gemini-2.5-Pro, by approximately 8 and 5 percentage points, respectively. This substantial improvement can be attributed to our two-stage training framework, particularly the GRPO stage, where CBT skills and dialogue strategies are incorporated as explicit guidance signals. Through reward shaping, this stage reinforces the model's adherence to structured intervention logic, leading to greater controllability and clinical consistency. Moreover, the predicted stage-wise distributions in Figure 7 (CBT skills) and Figure 8 (dialogue strategies) show that CBT-Qwen3 generally follows the structure recommended by CBT intervention annotations.

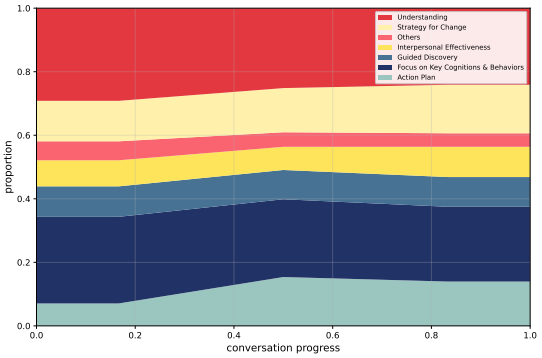

Figure 7: Predicted distribution of CBT skills generated by CBT-Qwen3.

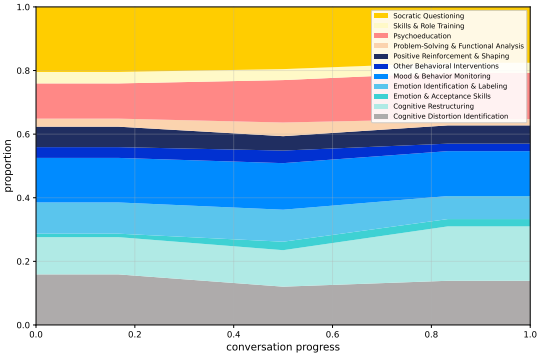

Figure 8: Predicted distribution of dialogue strategies generated by CBT-Qwen3.

	CBT Intervention		Counselor Response								
	Skill_Acc	Strategy_Acc	B-2	Avg.B	Rouge-L	PPL (↓)	B.S.	Dist-1	Dist-2	Dist-3	Safety
CBT-Qwen3	**42.32**	**31.36**	**3.54**	**3.82**	**10.01**	**5.15**	**50.76**	**9.45**	**38.72**	**62.25**	✓
-w/o SFT	41.34	30.12	3.21	3.55	9.62	5.23	50.40	9.01	37.33	61.45	✓
-w/o GRPO	40.65	27.62	3.11	3.48	9.36	5.57	50.26	8.89	36.88	60.73	✓
-w/o GRPO, CBT-intervened	-	-	2.98	3.34	9.15	5.70	50.12	8.21	35.66	59.14	✓

Table 5: Ablation studies of our CBT-Qwen3 model. "B-2", "Avg.B", "PPL", and "B.S." represent BLEU-2, Avg BLEU, perplexity, and BERTScore, respectively.

Counselor Response Generation. CBT-Qwen3 also demonstrates outstanding performance in response generation quality. It achieves higher scores on traditional generation metrics such as BLEU-2 and ROUGE-L, indicating better semantic consistency and contextual coherence. Moreover, CBT-Qwen3 achieves notable improvements on diversity metrics (Dist-1, Dist-2, and Dist-3), suggesting enhanced linguistic richness and expressive variety while maintaining semantic accuracy. This result further shows that, in generating professional counseling responses, CBT-Qwen3 effectively mitigates the problem of expression homogeneity commonly observed in LLM-generated text.

4.4 Ablation Studies

As shown in Table 5, we conduct ablation studies to assess the contributions of each training stage and the CBT intervention signals. We consider the following variants: (1) "-w/o SFT": removes the SFT stage and directly runs GRPO without cold start; (2) "-w/o GRPO": keeps only the first-stage SFT, without GRPO stage; (3) "-w/o GRPO, CBT-intervened": includes only SFT while removing the implicit CBT intervention from the input prompt, training the model to produce natural counselor-style responses without explicit intervention cues. Results indicate that the cold start provided by SFT, the professional alignment and controllability introduced by GRPO, and the intervention guidance from CBT signals are all critical to the overall performance of CBT-Qwen3.

4.5 Human Evaluation

To further verify the effectiveness of CBT-Qwen3, we conducted a human evaluation on 100 randomly sampled cases from the test set. Three graduate students trained in CBT were invited to evaluate the responses generated by our model and the best-performing baseline, Gemini-2.5-Pro. Each dimension was rated on a 1–5 Likert scale, and the final score was computed as the average of the three raters. As shown in Figure 9, CBT-Qwen3 substantially outperforms Gemini-2.5-Pro in the Intervention Competence dimension, indicating that it better captures the professional intervention structure of CBT-oriented counseling. Furthermore, CBT-Qwen3 also achieves higher scores in Helpfulness dimension, suggesting that its responses provide more substantive emotional support and cognitive guidance. Finally, in Fluency and Coherence dimensions, both models perform comparably, indicating that the GRPO stage did not lead to basic language degradation—CBT-Qwen3 maintains natural and coherent expression while achieving greater professional consistency.

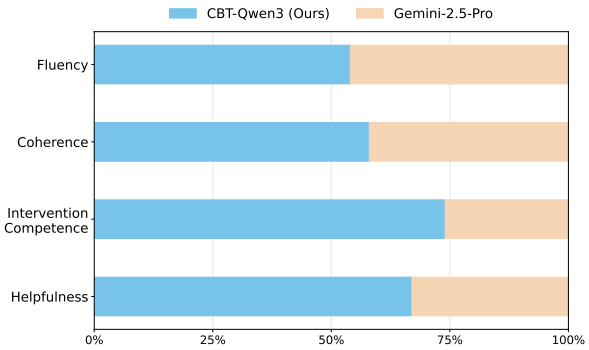

Figure 9: Human evaluation results.

5 Conclusion

We introduce CBTDialog, a CBT counseling dialogue dataset that combines real-world and simulated-client sessions with a hierarchical annotation scheme that captures the CBT intervention process via CBT skills and dialogue strategies, alleviating the scarcity of real-world data and professional intervention annotations. Building on CBTDialog, we develop CBT-Qwen3, a counselor model trained in two stages (SFT then GRPO) that treats CBT skills and dialogue strategies as explicit guidance and constraint signals to achieve controllable, CBT-aligned generation. Both automatic and human evaluations show significant improvements in intervention competence and clinical consistency, demonstrating the model's promising potential in mental health counseling.

Limitations

Although our work makes valuable contributions to the field of mental health support, several limitations warrant further investigation. First, the proposed CBTDialog dataset is collected from two platforms, YouTube and Alexander Street. Counseling dialogues from these platforms may exhibit potential sampling biases in terms of audience demographics, counseling formats, and the distribution of presenting problems. As a result, the dataset may not fully represent the diversity of real-world counseling scenarios and populations.

Second, CBTDialog currently contains only English-language counseling dialogues, which limits its linguistic and cultural diversity. Across different languages and cultural contexts, clients vary substantially in how they express psychological concerns, manifest emotions, and respond to therapeutic interventions. Expression patterns, metaphors, and socio-cultural assumptions commonly observed in English counseling settings may not generalize to other linguistic or cultural environments. This language and cultural homogeneity further constrains the dataset's coverage of cross-cultural counseling scenarios.

Consequently, the CBT-Qwen3 model trained on CBTDialog is, at its current stage, more suitable for clients who primarily communicate in English and whose counseling contexts align with Western therapeutic traditions. For users from different cultural backgrounds or non-English-speaking populations, the generated counseling responses may exhibit limitations in cultural appropriateness, pragmatic naturalness, and the suitability of intervention strategies.

Future work will focus on incorporating counseling data from more diverse sources and in multiple languages, as well as explicitly accounting for cultural differences and localization needs during data construction and model training, in order to improve the generalization and practical applicability of the model in cross-lingual and cross-cultural mental health support settings.

Ethical Considerations

Data Construction. The CBTDialog dataset is constructed from publicly accessible counseling videos collected from YouTube and Alexander Street. All data were accessed in accordance with the respective platforms' terms of service and were strictly used for research purposes. We do not redistribute any original video or audio content. To protect individual privacy, the data construction process combines automated processing with manual review and applies a rigorous de-identification procedure, including the removal or anonymization of personal identifiers such as names, locations, and contact information. The released CBTDialog dataset contains only de-identified textual transcripts.

Data Annotation. Prior to any annotation activities, this study underwent an internal ethics review and received approval. During the annotation phase, we employed three graduate students specializing in clinical psychology who were trained by a licensed CBT counselor. The annotators were responsible for: (1) determining whether a counseling session follows the CBT therapeutic framework, (2) assisting with data de-identification, and (3) annotating CBT interventions (CBT skill and dialogue strategy). To ensure fair compensation, annotators were paid $12 per hour, which is higher than the local average wage. In addition, annotators accessed the data only through a restricted internal system and were not permitted to copy, distribute, or re-release any data, thereby reducing the risk of data leakage or misuse.

Model Deployment. We do not advocate deploying the trained CBT-Qwen3 model directly in real-world psychotherapy settings. Although CBT-Qwen3 demonstrates strong performance in experimental evaluations, its generated responses may not fully meet the standards of professional counseling provided by licensed practitioners. At its current stage, the model is better suited as an assistive tool to support human counselors rather than as a standalone replacement. Our long-term goal is to explore feasible pathways for applying large language models in mental health support scenarios, enhance the capabilities of counseling agents, and provide insights for further research and development of AI-assisted mental health counseling and psychotherapy.

References

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the European chapter of the Association for Computational Linguistics*, pages 2291–2302.

Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

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 *Proceedings of the Conference on Empirical Methods in Natural Language Processing (Findings)*, pages 1170–1183.

Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. 2025. Towards multimodal emotional support conversation systems. *IEEE Transactions on Multimedia*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Isaac R Galatzer-Levy, Daniel McDuff, Vivek Natarajan, Alan Karthikesalingam, and Matteo Malgaroli. 2023. The capability of large language models to measure psychiatric functioning. *arXiv preprint arXiv:2308.01834*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Ponnurangam Kumaraguru, and Amit Sheth. 2022. Learning to automate follow-up question generation using process knowledge for depression triage on reddit posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*.

Clara E Hill. 1999. Helping skills: Facilitating exploration, insight, and action. *American Psychological Association*.

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 *Proceedings of the the International Conference on Learning Representations*.

Yuxin Hu, Danni Liu, Bo Liu, Yida Chen, Jiuxin Cao, and Yan Liu. 2025. Psyadvisor: A plug-and-play strategy advice planner with proactive questioning in psychological conversations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 12205–12229.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Mina Kian, Kaleen Shrestha, Katrin Fischer, Xiaoyuan Zhu, Jonathan Ong, Aryan Trehan, Jessica Wang, Gloria Chang, Séb Arnold, and Maja Mataric. 2025. Using linguistic entrainment to evaluate large language models for use in cognitive behavioral therapy. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (Findings)*, pages 7724–7743.

Minju Kim, Dongje Yoo, Yeonjun Hwang, Minseok Kang, Namyoung Kim, Minju Gwak, Beong-woo Kwak, Hyungjoo Chae, Harim Kim, Yunjoong Lee, and 1 others. 2025a. Can you share your story? modeling clients' metacognition and openness for llm therapist evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Findings)*, pages 25943–25962.

Subin Kim, Hoonrae Kim, Heejin Do, and Gary Lee. 2025b. Multimodal cognitive reframing therapy via multi-hop psychotherapeutic reasoning. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 4863–4880.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

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 *Proceedings of the Conference on Empirical Methods in Natural Language Processing (Findings)*, pages 14245–14274.

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 Annual Meeting of the Association for Computational Linguistics*, pages 10358–10376.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 110–119.

Junlin Li, Bo Peng, Yu-Yin Hsu, and Chu-Ren Huang. 2024. Be helpful but don't talk too much-enhancing helpfulness in conversations through relevance in multi-turn emotional support. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1976–1988.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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 Annual Meeting of the Association for Computational Linguistics*.

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 Annual Meeting of the Association for Computational Linguistics*, pages 13641–13660.

Hongbin Na. 2024. Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 2930–2940.

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

Zhiyang Qi, Takumasa Kaneko, Keiko Takamizo, Mariko Ukiyo, and Michimasa Inaba. 2025. Kokorochat: A japanese psychological counseling dialogue dataset collected via role-playing by trained counselors. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*.

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 *Proceedings of the Conference on Empirical Methods in Natural Language Processing (Findings)*, pages 615–636.

Huachuan Qiu and Zhenzhong Lan. 2025. Psydial: A large-scale long-term conversational dataset for mental health support. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 21624–21655.

Huachuan Qiu, Lizhi Ma, and Zhenzhong Lan. 2024b. Psyguard: An automated system for suicide detection and risk assessment in psychological counseling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4581–4607.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the International World Wide Web Conference*, pages 194–205.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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 Annual Meeting of the Association for Computational Linguistics*.

Aseem Srivastava, Ishan Pandey, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Response-act guided reinforced dialogue generation for mental health counseling. In *Proceedings of the International World Wide Web Conference*, pages 1118–1129.

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 *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Findings)*.

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 ACM Conference on Human Factors in Computing Systems*, pages 1–13.

EmoLLM Team. 2024. Emollm: Reinventing mental health support with large language models. https://github.com/SmartFlowAI/EmoLLM.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, and 1 others. 2024. Patient-ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. Healme: Harnessing cognitive reframing in large language models for psychotherapy. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1707–1725.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. 2025. Psydt: Using llms to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

X Xu, B Yao, Y Dong, S Gabriel, H Yu, J Hendler, M Ghassemi, AK Dey, and D Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):31–31.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024a. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the International World Wide Web Conference*, pages 4489–4500.

Qisen Yang, Zekun Wang, Honghui Chen, Shenzhi Wang, Yifan Pu, Xin Gao, Wenhao Huang, Shiji Song, and Gao Huang. 2024b. Psychogat: A novel psychological measurement paradigm through interactive fiction games with llm agents. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 14470–14505.

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Kit Phey Leng, Nicholas Gabriel Lim, Cameron Tan Shi Ern, and Ee-peng Lim. 2025. Cami: A counselor agent supporting motivational interviewing through state inference and topic exploration. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

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 Conference on Empirical Methods in Natural Language Processing*, pages 2438–2459.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024a. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Findings)*, pages 13947–13966.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024b. Escot: Towards interpretable emotional support dialogue systems. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 13395–13412.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Weixiang Zhao, Shilong Wang, Yanpeng Tong, Xin Lu, Zhuojun Li, Yanyan Zhao, Chenxue Wang, Tian Zheng, and Bing Qin. 2025. A parental emotion coaching dialogue assistant for better parent-child interaction. *Science China Information Sciences*, 68(7):179101.

Wenjie Zheng, Jianfei Yu, Rui Xia, and Shijin Wang. 2023. A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 15445–15459.

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 Annual Meeting of the Association for Computational Linguistics*.

A Details of CBTDialog

A.1 Data Source, Filtering and Cleaning

We provide detailed descriptions of the data sources, counseling scenarios, and corresponding links in Table 6, along with the data filtering and cleaning procedures, as well as the distribution of counselors and clients.

A.2 Session Segmentation

We employ five advanced LLMs to perform session segmentation and presenting problem annotation: GPT-4.1[4], Gemini-2.5-Pro[5], Claude-Opus-4.1[6], DeepSeek-R1[7], and Qwen3-235B-Instruct[8]. The prompt design used for this procedure is illustrated in Figure 11.

A.3 Descriptions of CBT Intervention Annotations

We define a hierarchical CBT intervention annotation schema consisting of seven goal-level CBT skills and eleven implementation-level dialogue strategies. In this part, we provide detailed descriptions of these labels, as shown in Table 8 and Table 9.

A.4 Data Splitting

We adopted the following data partitioning principles: (1) Each complete counseling session served as the minimum unit of division to ensure that no single session was split across subsets, thereby preventing contextual information leakage; (2) The

[4] https://openai.com/index/gpt-4-1/

[5] https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/

[6] https://www.anthropic.com/news/claude-opus-4-1

[7] https://api-docs.deepseek.com/news/news250528

[8] https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507

Source	Counselor/Publisher	Case/Vignette	Scenario	Raw (h)	Cleaned (h)	#CBT Sess.	CBT (h)	#P	#C	Access URL
YouTube	In Therapy with Alex Howard	Belle's Case	Real	4.71	3.92	3	1.33	1	1	YouTube@AlexHoward
		Paul's Case	Real	1.90	1.44	0	0.00		0	
		Pierre's Case	Real	4.05	3.14	4	1.41		1	
		Hayley's Case	Real	5.54	4.12	9	4.12		1	
		Al's Case	Real	4.17	3.86	4	3.02		1	
		Sanaya's Case	Real	6.02	5.47	8	5.47		1	
		Beverly's Case	Real	6.03	5.46	2	1.36		1	
		Katie's Case	Real	6.19	4.98	9	3.35		1	
		David's Case	Real	9.06	7.69	14	5.68		1	
		Hannah's Case	Real	6.06	4.5	10	3.35		1	
		Sally's Case	Real	6.40	4.78	6	1.69		1	
		Lauren's Case	Real	5.71	4.44	7	2.25		1	
		Nicole's Case	Real	0.96	0.67	0	0.00		0	
		Therapeutic Tuesday	Real	11.38	10.34	18	9.74		14	
	MedCircle	Kyle's Case	Real	3.47	3.23	5	2.13	2	1	YouTube@MedCircle
	Judith S. Beck	Teaching Vignette	Simulated	1.58	1.57	2	1.58	1	1	YouTube@BeckInstitute
	Judith Johnson	Teaching Vignette	Simulated	1.84	1.78	9	1.78	1	7	YouTube@JudithJohnson
	University of Nottingham	Teaching Vignette	Simulated	1.67	1.44	6	1.13	3	6	YouTube@uniofnottingham
	Todd Grande	Teaching Vignette	Simulated	11.09	10.7	20	8.76	1	19	YouTube@ToddGrande
	Russ Curtis	Teaching Vignette	Simulated	1.64	1.45	12	1.45	1	12	YouTube@RussCurtis
	Anna Freud Centre charity	Teaching Vignette	Simulated	0.81	0.74	1	0.74	1	1	YouTube@AnnaFreud
Alexander Street	Microtraining Associates	Teaching Vignette	Simulated	17.54	14.26	26	10.21	16	22	video.alexanderstreet.com
	Milton H. Erickson Foundation	Teaching Vignette	Simulated	8.35	8.2	13	7.22	8	17	
	Psychological & Educational Films	Teaching Vignette	Simulated	3.11	2.32	5	2.32	3	3	
	Zeig, Tucker Theisen Inc.	Teaching Vignette	Simulated	2.45	2.24	3	2.24	1	3	
	University of Manchester	Teaching Vignette	Simulated	5.3	4.7	5	4.7	3	14	
	American Counseling Association	Teaching Vignette	Simulated	4.45	3.12	5	3.12	4	4	
	University of South Wales	Teaching Vignette	Simulated	4.98	4.33	2	4	2	2	
	AIPC Educational Institution	Teaching Vignette	Simulated	0.42	0.36	2	0.36	2	2	
Total				146.88	125.25	210	94.51	50	138	

Table 6: Overview of data sources, filtering and cleaning for CBTDialog. #CBT Sess. denotes the number of CBT sessions, #P denotes the number of CBT counselors, and #C denotes the number of clients.

data were split into training, validation, and test sets in proportions of 0.75, 0.10, and 0.15; (3) During splitting, we computed semantic similarity based on the presenting problems within sessions and applied K-Means clustering to ensure that semantically similar sessions were distributed across different subsets; (4) Both real-world and simulated-client sessions were evenly distributed to avoid source bias across splits; (5) The distribution also considered the number of counselor utterances to maintain balance in corpus scale among subsets. Table 7 reports the statistics of data splitting.

A.5 Analysis of Dialogue Turn Distributions

As illustrated in Figure 10, the average number of dialogue turns varies across presenting problems. Specifically, *Adolescent Psychology Issues* shows the highest average turn count, suggesting that counselors tend to engage in deeper exploration and clarification when working with adolescents—helping them express and recognize their problems.

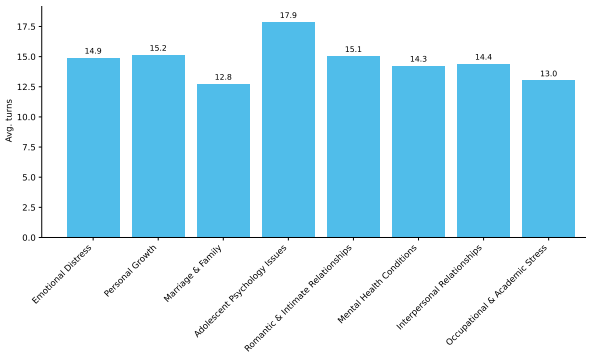

Figure 10: The distribution of presenting problem's average turns.

	Total	Training	Validation	Test
#Sessions	210	157	21	32
#Dialogues	2,626	1,964	316	346
#Utterances	81,486	60,903	9,580	11,003
#Counselor Utterances	38,977	29,074	4,650	5,253

Table 7: Statistics of training, validation, and test splits of the CBTDialog dataset.

B Experimental Setting

All experiments were conducted on two NVIDIA A800 (80 GB) GPUs. The backbone model

is Qwen3-4B-Instruct-2507[9], and we applied LoRA (Hu et al., 2022) for efficient parameter fine-tuning with a rank of 64, using bf16 precision. For the SFT stage, we set the training batch size to 64 and the learning rate to 1e-5. Training was performed for three epochs, with early stopping applied once the validation loss ceased to decrease. For the GRPO stage, the training batch size was set to 32, and the learning rate was 5e-6 for the policy network. The safety of generated responses was evaluated using a RoBERTa-based toxicity classifier [10], trained on the English portions of the Jigsaw 2018–2020 datasets, comprising approximately 2 million samples in total. The multi-component reward was composed of four components: (1) a safety reward (assigning a reward of 0.05 if the generated response is classified as safe), (2) a CBT intervention reward (with a total weight of 0.4, including 0.2 for CBT skill matching and 0.2 for dialogue strategy matching), (3) a content fidelity reward (with a total weight of 0.5, including 0.25 based on ROUGE-L and 0.25 based on BERTScore), and (4) a format compliance reward (assigning a reward of 0.05 for correct format and a penalty of 1.0 otherwise).

Moreover, we introduce both a length penalty and a diversity bonus into the reward design. (1) Length penalty. To prevent the model from exploiting response length for reward hacking, we apply a Gaussian-based penalty function to the generated responses. When the response length falls within 50%–150% of the reference response length, only a mild penalty is imposed, preserving at least 85% of the original reward. In contrast, responses that are excessively short (<50%) or overly long (>150%) are subject to an aggressive exponential decay, with the retained reward capped at a minimum of 5%. (2) Diversity bonus. To mitigate entropy collapse, we introduce an intra-group n-gram diversity bonus that encourages responses containing unique trigrams not present in other samples within the same group. This mechanism promotes the exploration of diverse CBT responses rather than convergence to repetitive response patterns.

[9] https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

[10] https://huggingface.co/s-nlp/roberta_toxicity_classifier

C Related Work

C.1 Dialogue Datasets for Mental Health Support

With the accelerating pace of modern society, mental health issues have become increasingly prominent. This trend has attracted growing attention from researchers in the field of mental health support, motivating the construction of a variety of dialogue datasets to facilitate the development of mental health dialogue systems. Early dialogue datasets primarily focused on modeling emotional expression and empathetic capabilities, emphasizing the understanding and soothing of clients' emotional states. Such datasets are typically collected from online mental health service platform, social media platforms, and role-playing scenarios, and are annotated with emotion categories, empathy levels, or supportive strategies (Rashkin et al., 2019; Sharma et al., 2020; Sun et al., 2021; Liu et al., 2021; Chen et al., 2023; Li et al., 2023; Zhang et al., 2024b; Zheng et al., 2024).

In recent years, some studies have gradually incorporated psychological knowledge and clinically relevant concepts, enabling more structured modeling of mental health dialogue data. Specifically, several works adopt widely validated psychological scales or therapeutic frameworks, aligning dialogue content with specific symptom dimensions or intervention strategies to provide models with clinically meaningful supervision signals (Gupta et al., 2022; Yao et al., 2022; Xiao et al., 2024; Lee et al., 2024; Chu et al., 2025; Kim et al., 2025b; Qi et al., 2025).

Meanwhile, constructing multi-turn dialogue datasets that more closely reflect real-world mental health counseling scenarios has emerged as an important research direction. Some studies leverage professional counseling reports, case records, or structured interview materials, and generate high-quality multi-turn counseling dialogues through dialogue reconstruction or data synthesis methods (Abacha et al., 2023; Qiu et al., 2024a; Zhang et al., 2024a; Xie et al., 2025).

C.2 CBT in Mental Health Support

According to the authoritative textbook by Judith S. Beck (Beck, 2020), Cognitive Behavioral Therapy (CBT) aims to break the negative cycle among cognition, emotion, and behavior by identifying and modifying maladaptive automatic thoughts and core beliefs, together with structured behavioral exercises. Owing to its structured and evidence-

based nature, CBT has attracted broad attention in both research and clinical practice. In the field of mental health support, early studies (Sharma et al., 2020, 2023) introduced cognitive reframing into empathetic response generation to enhance models' understanding of and responses to clients' emotions. Other work (Sharma et al., 2021; Maddela et al., 2023) focused on empathetic rewriting of unhelpful thoughts to improve empathy and supportiveness.

Beyond response generation, several studies have incorporated CBT principles to construct simulated patients or clients for training and evaluating mental health professionals and counseling systems (Wang et al., 2024; Kim et al., 2025a). In a complementary direction, Kian et al. (2025) evaluated the intrinsic capabilities of off-the-shelf large language models in CBT-oriented scenarios, with a particular focus on linguistic coordination and therapeutic alignment. Moreover, Syed et al. (2024) analyzed real-world conversations from online peer support platforms and demonstrated that CBT-inspired conversational strategies, such as reflective listening and guided exploration, contribute to higher perceived empathy. More recently, to obtain datasets closer to real counseling scenarios, a line of studies (Na, 2024; Lee et al., 2024; Xiao et al., 2024; Zhang et al., 2024a; Xie et al., 2025) has adopted CBT as a data construction principle, designing various data synthesis schemes.

C.3 Psychological LLMs

With the rapid advancement of LLMs, domain-specific LLMs for mental health have attracted increasing attention from the research community. Psychological LLMs have been explored along four main directions: (1) Prompt-based zero/few-shot methods. Researchers (Galatzer-Levy et al., 2023; Xu et al., 2024) use general-purpose LLMs (e.g., ChatGPT) to directly generate counseling-style responses under zero- or few-shot settings. (2) Instruction-driven specialized fine-tuning. Some studies (Yang et al., 2024a; Team, 2024) construct/clean mental-health corpora and design instruction templates to perform instruction tuning on base models, yielding models specialized for mental-health scenarios. (3) Agent-based roles and collaboration. Several works (Yang et al., 2024b, 2025) build multi-agent frameworks—such as "counselor agent–client agent" configurations—using role division, tool use, and iterative planning to simulate interviewing workflow. (4) Reinforcement learning for alignment and controllable generation. Another line of work (Srivastava et al., 2023; Li et al., 2024) introduces RL algorithms (e.g., PPO) into counseling settings and optimizes generation policies via reward design.

CBT skill	Description
Understanding	Reflecting, paraphrasing, or summarizing the client's words to demonstrate accurate listening and empathy.
Interpersonal Effectiveness	Showing warmth, concern, or encouragement, positively reinforcing client actions, and maintaining a professional and supportive tone.
Guided Discovery	Using a discovery process to help clients achieve cognitive shifts and arrive at their own conclusions, avoiding direct persuasion, and assessing the outcome.
Focus on Key Cognitions & Behaviors	Directing attention to the client's specific thoughts, images, emotions, sensations, or behaviors central to their concerns.
Strategy for Change	Designing and integrating evidence-based CBT techniques into an overall therapeutic plan, aligning intervention methods with targeted change goals.
Action Plan	Assigning, reviewing, or co-developing concrete between-session tasks or goals to support ongoing skill use and progress.
Others	Counselor interventions that are supportive but do not fit the above categories.

Table 8: Seven CBT skills with corresponding descriptions.

Dialogue strategy	Description
# Cognitive Strategies	
Cognitive Distortion Identification	Identifying distorted or maladaptive thinking patterns (e.g., overgeneralization, catastrophizing, all-or-nothing thinking) to enhance the client's awareness of cognitive biases.
Cognitive Restructuring	Helping the client critically evaluate maladaptive thoughts and replace them with more balanced, adaptive cognitions through evidence examination and alternative interpretation.
Socratic Questioning	Using systematic, guided questioning to encourage the client to examine the validity of their beliefs, consider alternative perspectives, and develop insight through reflective dialogue.
# Behavioral Strategies	
Problem-Solving & Functional Analysis	Breaking down problems into manageable components, analyzing antecedents, behaviors, and consequences, and collaboratively developing practical strategies to address specific difficulties.
Positive Reinforcement & Shaping	Encouraging adaptive behaviors by reinforcing constructive actions, gradually shaping new skills, and promoting continued engagement in healthy behavioral patterns.
Skills & Role Training	Teaching and rehearsing coping skills or interpersonal techniques through modeling, practice, and role-play exercises to enhance competence and confidence.
Other Behavioral Interventions	Applying additional behaviorally oriented methods (e.g., exposure, activity scheduling, relaxation training) that target maladaptive patterns and support behavioral change.
# Emotional Strategies	
Emotion Identification & Labeling	Guiding the client to recognize, name, and differentiate emotions and emotional triggers to build emotional awareness and clarity.
Emotion & Acceptance Skills	Supporting the client in developing tolerance and acceptance of difficult emotions, promoting adaptive regulation strategies, and reducing avoidance or suppression.
# Educational & Monitoring Strategies	
Psychoeducation	Providing the client with evidence-based knowledge about psychological concepts, disorders, or CBT principles to enhance understanding and engagement in treatment.
Mood & Behavior Monitoring	Encouraging systematic tracking of thoughts, emotions, and behaviors across time and situations to identify patterns and evaluate progress in therapy.

Table 9: Eleven dialogue strategies with corresponding descriptions.

```
Please act as a clinical psychologist with expertise in Cognitive Behavioral Therapy (CBT). You will be provided with a client-counselor transcript of a CBT-oriented psychotherapy session. Your task is to identify the presenting problems (either raised by the client or introduced by the counselor through questioning), and segment the session into multiple dialogue units, each corresponding to a specific type of presenting problem.

### Annotation Guidelines
• Each dialogue unit should represent a complete exchange in which the client discusses a specific issue and the counselor responds to that issue.
• A dialogue unit should **end when the counselor delivers a professional intervention** (e.g., emotional clarification, cognitive restructuring, behavioral suggestion, or empathic feedback), whenever possible. If necessary, you may include one final counselor response after the client completes their turn to ensure the dialogue unit ends with a counselor utterance.
• Do **not** end a dialogue unit immediately after a client's response unless it is followed by a very short final statement from the counselor (e.g., validation, summarization, or goal-setting), to ensure smooth continuation into the next unit.
• If the client expresses multiple consecutive utterances (lines) that are thematically or emotionally connected, they should be merged into one presenting problem unit. However, non-contiguous references to the same problem (e.g., Problem A → Problem B → Problem A) must be treated as separate dialogue units. Only uninterrupted, consecutive segments can be merged into the same unit.
• If there is a clear topic shift, emotional transition, or change in cognitive focus, start a new dialogue unit.
• A valid dialogue unit MUST include at least **6** lines, but no more than **50** lines.
• Every line MUST be assigned to exactly one dialogue unit. No line may be omitted, duplicated, or left unassigned.
• Presenting problem labels MUST be selected from the subcategories under the following domains:
(1) Emotional Distress: fear, depression, guilt/self-blame, repressed emotions, anxiety, appearance-related anxiety, difficulty expressing emotions, public speaking fear, emotional eating, emotional dysregulation, health anxiety, grief, separation anxiety, premenstrual tension
(2) Romantic & Intimate Relationships: emotional trauma, fear of dating, long-distance relationship stress, romantic confusion, commitment issues, infidelity, emotional dependency, romantic jealousy, relationship burnout, love addiction, breakup crisis, intimate partner violence, fear of intimacy, Romantic Relationship Issues
(3) Marriage & Family: parent-child conflict, parenting anxiety, family violence, divorce-related distress, marital burnout, intergenerational conflict, perinatal depression, family pressure, marital conflict, parenting style issues, aging-related stress
(4) Personal Growth: childhood trauma, compulsive tendencies, low self-esteem, identity confusion, antisocial behavior, people-pleasing personality, procrastination, obsessive perfectionism, emotional emptiness, maladaptive coping, sleep issues, irrational spending
(5) Interpersonal Relationships: social anxiety, interpersonal conflict, fear of rejection, social withdrawal, social isolation, sensitivity to social cues
(6) Occupational & Academic Stress: job burnout, unemployment stress, workplace bullying, career decision difficulties, work-life conflict, performance anxiety, professional identity issues, workplace sexual harassment
(7) Mental Health Conditions: major depressive disorder, Obsessive-Compulsive Disorder, schizophrenia, bipolar disorder, generalized anxiety disorder, panic disorder, social anxiety disorder, posttraumatic stress disorder, eating disorders, somatic symptom disorder, insomnia disorder, borderline personality disorder, suicidal behavior disorder, substance use disorder, dissociative identity disorder, attention-deficit/hyperactivity disorder, adjustment disorder
(8) Adolescent Psychology Issues: Test anxiety, School refusal / Academic disengagement, Problematic behaviors / Behavioral dysregulation, Aggressive behavior problems, Romantic relationship distress in adolescence, Academic procrastination in adolescents, Trauma from school bullying, Academic anxiety, Peer relationship difficulties, Depressive symptoms in adolescents, Anxiety symptoms in adolescents, Obsessive-compulsive tendencies, Body image dissatisfaction / Appearance-related anxiety, Low self-esteem / Inferiority feelings


### Return Content for Each Dialogue Unit:
For each dialogue unit, return a JSON object with the following five fields:
• **Presenting Problem Label** – The specific label that best characterizes the client's concern.
• **Descriptive Statement** – A representative and direct statement from the client that reflects the core emotional, cognitive, or behavioral issue (fewer than **20** words).
• **Line Range** – The range of line numbers that span the dialogue unit.
• **Confidence Score** – A numerical estimate (0 to 1) indicating how confident you are that the identified unit corresponds to the selected presenting problem.
• **Rationale** – A brief, clinically grounded explanation—based on CBT principles—justifying the label assignment and segmentation (fewer than **30** words).

### Additional Constraint:
• When choosing the endpoint of each dialogue unit, **prefer to end the unit on a counselor utterance**—especially when the counselor provides a summary, intervention, or transition. This is to support a downstream task of generating the counselor's next response.

### Output JSON Format (Do not include any explanations, markdown, or additional text):
[{
"Presenting Problem Label": "xxxx",
"Descriptive Statement": "xxxx",
"Line Range": "[xx, xx]",
"Confidence Score": float (0.0 - 1.0),
"Rationale": "xxxx"
},
{
"Presenting Problem Label": "yyyy",
"Descriptive Statement": "yyyy",
"Line Range": "[yy, yy]",
"Confidence Score": float (0.0 - 1.0),
"Rationale": "yyyy"
}]
```

Figure 11: Instruction design for session segmentation.

```
You are a licensed psychiatrist specializing in Cognitive Behavioral Therapy (CBT). Your task is to annotate a single counselor utterance based on the preceding dialogue history and the current counselor turn. The annotation follows a hierarchical structure, consisting of a CBT Skill (representing the functional goal level, i.e., what the counselor aims to achieve) and a Dialogue Strategy (representing the concrete implementation level, i.e., how the counselor carries out the intervention). You must select exactly one CBT Skill and exactly one Dialogue Strategy that most clearly reflect the counselor's primary intent in the given utterance. Multiple selections are not allowed.

===== CBT Skill Options =====

- Understanding (Understanding & Reflecting, paraphrasing, or summarizing the client's words to demonstrate accurate listening and empathy)
- Interpersonal Effectiveness (Showing warmth, concern, or encouragement, positively reinforcing client actions, and maintaining a professional and supportive tone)
- Guided Discovery (Using a discovery process to help clients achieve cognitive shifts and arrive at their own conclusions, avoiding direct persuasion, and assessing the outcome)
- Focus on Key Cognitions & Behaviors (Directing attention to the client's specific thoughts, images, emotions, sensations, or behaviors central to their concerns)
- Strategy for Change (Designing and integrating evidence-based CBT techniques into an overall therapeutic plan, aligning intervention methods with targeted change goals)
- Action Plan (Assigning, reviewing, or co-developing concrete between-session tasks or goals to support ongoing skill use and progress)
- Others (Counselor interventions that are supportive but do not fit the above categories)


===== Dialogue Strategy Options =====

- Cognitive Distortion Identification (Identifying distorted or maladaptive thinking patterns (e.g., overgeneralization, catastrophizing, all-or-nothing thinking) to enhance the client's awareness of cognitive biases)
- Cognitive Restructuring (Helping the client critically evaluate maladaptive thoughts and replace them with more balanced, adaptive cognitions through evidence examination and alternative interpretation)
- Socratic Questioning (Using systematic, guided questioning to encourage the client to examine the validity of their beliefs, consider alternative perspectives, and develop insight through reflective dialogue)
- Problem-Solving & Functional Analysis (Breaking down problems into manageable components, analyzing antecedents, behaviors, and consequences, and collaboratively developing practical strategies to address specific difficulties)
- Positive Reinforcement & Shaping (Encouraging adaptive behaviors by reinforcing constructive actions, gradually shaping new skills, and promoting continued engagement in healthy behavioral patterns)
- Skills & Role Training (Teaching and rehearsing coping skills or interpersonal techniques through modeling, practice, and role-play exercises to enhance competence and confidence)
- Other Behavioral Interventions (Applying additional behaviorally oriented methods (e.g., exposure, activity scheduling, relaxation training) that target maladaptive patterns and support behavioral change)
- Emotion Identification & Labeling (Guiding the client to recognize, name, and differentiate emotions and emotional triggers to build emotional awareness and clarity)
- Emotion & Acceptance Skills (Supporting the client in developing tolerance and acceptance of difficult emotions, promoting adaptive regulation strategies, and reducing avoidance or suppression)
- Psychoeducation (Providing the client with evidence-based knowledge about psychological concepts, disorders, or CBT principles to enhance understanding and engagement in treatment)
- Mood & Behavior Monitoring (Encouraging systematic tracking of thoughts, emotions, and behaviors across time and situations to identify patterns and evaluate progress in therapy)


===== Your Response Format =====
Return your answer in the following JSON format (Do not include any explanations, markdown, or additional text):
{"CBT Skill": "Skill_Name",
"Dialogue Strategy": "Strategy_Name",
"Confidence": {
"CBT Skill": float (0.00 - 1.00),
"Dialogue Strategy": float (0.00 - 1.00)
}}

===== Example 1 =====
{"CBT Skill": "Guided Discovery",
"Dialogue Strategy": "Socratic Questioning",
"Confidence": {
"CBT Skill": 0.85,
"Dialogue Strategy": 0.90
}}

===== Example 2 =====
{"CBT Skill": "Focus on Key Cognitions and Behaviors",
"Dialogue Strategy": "Cognitive Distortion Identification",
"Confidence": {
"CBT Skill": 0.95,
"Dialogue Strategy": 0.93
}}
```

Figure 12: Instruction design for LLM-based annotation.

```
You are a Cognitive Behavioral Therapy (CBT)–oriented counselor. Based on the presenting problem and dialogue context,
select EXACTLY ONE CBT skill and EXACTLY ONE dialogue strategy (from the whitelists below), then generate one counselor
reply.

### Output Format
Return **exactly** the following JSON (no extra text, markdown, or commentary):

{"CBT Skill": "Choose one skill from the CBT Skills List",
"Dialogue Strategy": "Choose one strategy from the Dialogue Strategies List",
"Counselor Response": "Your response here"}

### Hard Rules
- Determine exactly one CBT Skill and exactly one Dialogue Strategy from the whitelists.
- In counselor response, do not reveal reasoning, chain-of-thought, labels, or mention "skill/strategy".
- Tone: warm, empathic, professional; concrete and actionable.
- Respond in English only using ASCII characters. Limit to 150 words.

### Safety Override (takes precedence over all rules)
If there are imminent risk cues (e.g., self-harm intent/plan/means, violence), prioritize brief risk assessment,
immediate safety steps, and crisis resources. Never provide harmful instructions.

### CBT Skills Whitelist
- Understanding
- Interpersonal Effectiveness
- Guided Discovery
- Focus on Key Cognitions & Behaviors
- Strategy for Change
- Action Plan
- Others

### Dialogue Strategies Whitelist
- Cognitive Distortion Identification
- Cognitive Restructuring
- Socratic Questioning
- Problem-Solving & Functional Analysis
- Positive Reinforcement & Shaping
- Skills & Role Training
- Other Behavioral Interventions
- Emotion Identification & Labeling
- Emotion & Acceptance Skills
- Psychoeducation
- Mood & Behavior Monitoring

### Examples

===== Example 1 =====
INPUTS:
Presenting Problem: ”Adolescent psychology issues"
Dialogue Context:
"Client": "I think dealing with the grief with my mom and it."

OUTPUTS:
{"CBT Skill": "Agenda",
"Dialogue Strategy": "Mood & Behavior Monitoring",
"Counselor Response": "Okay, now did this kind of situation come up this very week?"}

===== Example 2 =====
INPUTS:
Presenting Problem: "Emotional distress"
Dialogue Context:
"Counselor": "Any other questions or anything else that you want to share at this point?"
"Client": "Yeah, so I think I did mention it but in the email. I'm feeling quite positive generally, like more than I
have done in a really long time. I probably will feel quite good going home now, like, you know, having done this and
having ideas in place. But that on its own, I feel, is hard to manage."

OUTPUTS:
{"CBT Skill": "Understanding",
"Dialogue Strategy": "Psychoeducation",
"Counselor Response": "Yeah, because your body doesn't distinguish between stress and positivity."}

===== Example 3 =====
INPUTS:
Presenting Problem: ”Marriage & family"
Dialogue Context:
"Client": "My sister knows about the anxiety part of it. And I think, I'm pretty sure she knows about the drinking, but
she doesn't bring it up. But if my parents know, I'm sure she knows."
"Counselor": "So you believe that they told her?"
"Client": "I'm sure."

OUTPUTS:
{"CBT Skill": "Guided Discovery",
"Dialogue Strategy": "Socratic Questioning",
"Counselor Response": "But she's not mentioned that to you directly?"}
```

Figure 13: Instruction design for few-shot experiments.

```
You are a Cognitive Behavioral Therapy (CBT)-oriented counselor. Based on the dialogue context, you must
think and select EXACTLY ONE CBT skill (representing the functional goal, i.e., what the counselor aims to
achieve) and EXACTLY ONE dialogue strategy (representing the concrete implementation, i.e., how the
counselor carries out the intervention), then generate a single counselor response aligned with both
selections.

Return your answer in EXACTLY these two XML blocks and nothing else:
<counselor_intervention>
<cbt_skill>[Choose one skill from the CBT Skills Whitelist]</cbt_skill>
<dialogue_strategy>[Choose one strategy from the Dialogue Strategies Whitelist]</dialogue_strategy>
</counselor_intervention>
<counselor_response>
[Your response here.]
</counselor_response>

### Hard Rules
- Use ONLY these tags: <counselor_intervention>, <cbt_skill>, <dialogue_strategy>, <counselor_response>.
No other tags, headings, lists, quotes, or blank lines.
- The values in <cbt_skill> and <dialogue_strategy> MUST match the whitelist entries EXACTLY (case-
sensitive), excluding the explanations in parentheses.
- In <counselor_response>, do not reveal reasoning, chain-of-thought, labels, or mention "skill/strategy".
- Tone: warm, empathic, professional; concrete and actionable.
- Respond in English only using ASCII characters. Limit to 150 words.

### Safety Override (takes precedence over all rules)
If there are imminent risk cues (e.g., self-harm intent/plan/means, violence), prioritize brief risk
assessment, immediate safety steps, and crisis resources. Never provide harmful instructions.

### CBT Skills Whitelist
- Understanding (Understanding & Reflecting, paraphrasing, or summarizing the client's words to
demonstrate accurate listening and empathy)
- Interpersonal Effectiveness (Showing warmth, concern, or encouragement, positively reinforcing client
actions, and maintaining a professional and supportive tone)
- Guided Discovery (Using a discovery process to help clients achieve cognitive shifts and arrive at their
own conclusions, avoiding direct persuasion, and assessing the outcome)
- Focus on Key Cognitions & Behaviors (Directing attention to the client's specific thoughts, images,
emotions, sensations, or behaviors central to their concerns)
- Strategy for Change (Designing and integrating evidence-based CBT techniques into an overall therapeutic
plan, aligning intervention methods with targeted change goals)
- Action Plan (Assigning, reviewing, or co-developing concrete between-session tasks or goals to support
ongoing skill use and progress)
- Others (Counselor interventions that are supportive but do not fit the above categories)

### Dialogue Strategies Whitelist
- Cognitive Distortion Identification (Identifying distorted or maladaptive thinking patterns (e.g.,
overgeneralization, catastrophizing, all-or-nothing thinking) to enhance the client's awareness of
cognitive biases)
- Cognitive Restructuring (Helping the client critically evaluate maladaptive thoughts and replace them
with more balanced, adaptive cognitions through evidence examination and alternative interpretation)
- Socratic Questioning (Using systematic, guided questioning to encourage the client to examine the
validity of their beliefs, consider alternative perspectives, and develop insight through reflective
dialogue)
- Problem-Solving & Functional Analysis (Breaking down problems into manageable components, analyzing
antecedents, behaviors, and consequences, and collaboratively developing practical strategies to address
specific difficulties)
- Positive Reinforcement & Shaping (Encouraging adaptive behaviors by reinforcing constructive actions,
gradually shaping new skills, and promoting continued engagement in healthy behavioral patterns)
- Skills & Role Training (Teaching and rehearsing coping skills or interpersonal techniques through
modeling, practice, and role-play exercises to enhance competence and confidence)
- Other Behavioral Interventions (Applying additional behaviorally oriented methods (e.g., exposure,
activity scheduling, relaxation training) that target maladaptive patterns and support behavioral change)
- Emotion Identification & Labeling (Guiding the client to recognize, name, and differentiate emotions and
emotional triggers to build emotional awareness and clarity)
- Emotion & Acceptance Skills (Supporting the client in developing tolerance and acceptance of difficult
emotions, promoting adaptive regulation strategies, and reducing avoidance or suppression)
- Psychoeducation (Providing the client with evidence-based knowledge about psychological concepts,
disorders, or CBT principles to enhance understanding and engagement in treatment)
- Mood & Behavior Monitoring (Encouraging systematic tracking of thoughts, emotions, and behaviors across
time and situations to identify patterns and evaluate progress in therapy)
```

Figure 14: Instruction design of the proposed model.