Empathy in Diversity: Personalized Depression and Anxiety Therapy via Dialogue State Tracking and Patient-Aware Planning

Xinwei Yang[1,2,6] **Junyi Fan**[1] **Yuqing Liu**[1,5] **Jiaxuan Wang**[1,2] **Jiashuai Zhang**[1,2]
Hongru Liang[4,6] **Wenqiang Lei**[2,6] **Yao Song**[1,3*]
[1] Convergence Laboratory of Chinese Cultural Inheritance and Global Communication, Sichuan University, Chengdu, China [2] College of Computer Science, Sichuan University
[3] Academy of Chinese, History, Religion and Philosophy, Hong Kong Baptist University
[4] School of Artificial Intelligence, Sichuan University, China
[5] School of Design, Hong Kong Polytechnic University, Hong Kong SAR, China
[6] Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, China
xinwei_yang@stu.scu.edu.cn yao.song@connect.polyu.hk

Abstract

Large language model (LLM) dialogue agents are increasingly used in psychological therapy, yet robustness across diverse patients remains underexplored. We address this gap with three contributions: (1) **MindEval**, a realistic role-play protocol for evaluating therapeutic dialogue agents; (2) **MindData**, a de-identified, expert-annotated corpus of therapist–patient dialogues (2,573 sessions; 63,348 turns); and (3) **MindApt**, a framework that integrates a therapeutic dialogue state tracking paradigm with a patient-aware strategic planning module. On MindEval, MindApt outperforms strong baselines on therapeutic outcomes and dialogue quality while improving conversational efficiency. To evaluate utility beyond role-play, we conducted a clinical study with real patients, demonstrating that MindApt-guided care achieves outcomes comparable to therapist-determined care, while the hybrid setting combining therapist judgment with MindApt's recommendations yields the strongest overall outcomes.

1 Introduction

Depression and anxiety affect more than 600 million people worldwide, representing a major global health burden (WHO, 2025). Structured therapies, particularly Cognitive Behavioral Therapy (CBT), are widely recognized for their efficacy in helping individuals regulate their thoughts and emotions (Muran and Motta, 1993; Beck, 2020; Curtiss et al., 2021). However, the one-size-fits-all application of such structured therapies often exhibits rigidity, failing to accommodate the unique profiles of patients who vary significantly in personal traits, emotional states, and symptoms (Schöller et al., 2018; Knauer et al., 2022; McClay et al., 2023). Research suggests that personalizing therapeutic strategies—such as employing Cognitive Restructuring for patients trapped in negative thought loops or Behavioral Activation for those experiencing severe withdrawal—is essential to strengthen the therapeutic alliance and improve clinical outcomes (Johnson, 2005; Cuijpers et al., 2016).

Recent efforts have turned to large language models (LLMs) as therapeutic agents for psychological therapy tasks (Chen et al., 2025; Hu et al., 2025; Lee et al., 2024b; Na, 2024; Xiao et al., 2024; Liu et al., 2023; Qiu et al., 2023). However, despite their conversational fluency, most existing systems are effectively confined to generic, one-size-fits-all strategies. Instead of tailoring interventions, they often default to formulaic responses that fail to perceive and respond to dynamic individual differences (Zhang et al., 2024a; Wang et al., 2024b; Na et al., 2025). We attribute this deficiency to two critical missing components: (1) effective therapeutic dialogue state tracking, without which the system cannot adapt in real time to patients' changing states; and (2) comprehensive patient modeling, whose absence yields suboptimal strategy selection.

To substantiate this analysis and systematically assess these limitations, we introduce **MindEval**, a role-play-based evaluation protocol that simulates realistic psychological therapy scenarios. MindEval comprises approximately 6k patient profiles derived from real cases, which are utilized to construct diverse patient simulators. These profiles span a broad range of personal traits, emotional states, and psychological symptoms to ensure diverse, realistic test conditions. By establishing such a rigorous testbed, MindEval allows us to quantify the gap between generic LLM capabilities and the nuanced requirements of clinical practice. This pro-

* Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9370–9416

tocol serves to crystallize the core technical challenge: making LLM-based therapeutic dialogue agents aware of diverse patient behaviors and devising personalized strategies that adapt to individual patients.

To mitigate the scarcity of authentic clinical data, we present **MindData**, a high-quality corpus of multi-turn, multi-session CBT dialogues derived from real-world clinical settings. Comprising 2,573 therapeutic sessions and 63,348 dialogue turns, this dataset offers a significant resource for modeling realistic interactions. It is enriched with fine-grained annotations, ranging from turn-level attributes (CBT stage, patient traits, emotional states, therapist strategies) to session-level assessments such as CTRS scores, empathy, coherence, helpfulness, and safety. Building on this foundation, we introduce **MindApt**, a framework that synergizes Cognitive Behavioral Therapy (CBT) principles with large language models (Beck, 2020). As illustrated in Figure 3, MindApt features two core components: **(1) a therapeutic dialogue state tracking paradigm that continuously and jointly models the CBT stage, personal traits, and emotional dynamics; and (2) a patient-aware strategic planning module that dynamically selects the optimal strategy from a comprehensive CBT library to tailor the therapy to individual needs.**

Evaluation results under MindEval reveal that current LLMs struggle to personalize strategies for diverse patient populations. In contrast, MindApt demonstrates stronger adaptability, outperforming strong baselines with a 13.6% improvement in success rate while maintaining robust performance across heterogeneous patient subgroups. To assess its utility beyond role-play, we further conducted a 16-week clinical study, which showed that MindApt-guided care achieves outcomes comparable to therapist-determined care. Moreover, when used as a decision support tool, MindApt complements therapist judgment and yields the strongest therapeutic outcomes in the hybrid setting.

Our key contributions are summarized as follows:

- We introduce **MindEval**, a clinically grounded evaluation protocol based on realistic role-play scenarios. Experiments on this protocol quantify the limitations of current LLMs in personalizing therapy across diverse patient populations.
- We present **MindData**, a large-scale dataset comprising 2,573 real-world therapeutic sessions (63,348 turns). The dataset is rigorously de-identified and enriched with fine-grained annotations.
- We propose **MindApt**, a framework designed for personalized psychological strategy planning. It features a novel therapeutic dialogue state tracking paradigm and a patient-aware planning module to dynamically adapt interventions to individual needs.
- We conduct extensive evaluations involving both user simulators and a clinical study with real patients. Results demonstrate that MindApt significantly outperforms state-of-the-art baselines in efficacy, empathy, and dialogue quality.

2 Related Work

Psychological Therapy Datasets. Due to the high costs and ethical constraints of involving real patients, most existing datasets rely on synthetic LLM-based role-play. For instance, PATIENT-Ψ (Wang et al., 2024b) and Cactus (Lee et al., 2024b) simulate therapists using CBT principles, while others (Chen et al., 2023; Xiao et al., 2024) simulate patients to generate annotated dialogues. However, these simulations often overlook complex personal traits and coping styles (Pennebaker, 1997), creating a substantial gap from real-world interactions (Chiu et al., 2024; Wang et al., 2024a). Conversely, while datasets like Psych8k (Liu et al., 2023) and CpsyCoun (Zhang et al., 2024a) are derived from real therapeutic sessions, they lack the multi-turn depth and comprehensive profiling required for personalized therapy. To bridge these gaps, MindData offers a large-scale, authentic corpus of fully annotated multi-turn sessions. It features de-identified patient profiles covering diverse traits, emotional states, and symptoms, explicitly designed to support personalized interventions.

LLM-based Therapeutic Dialogue Agents. Leveraging their strong conversational abilities, recent efforts have utilized LLMs as therapeutic dialogue agents for psychological therapy tasks (Lee et al., 2024b; Na, 2024; Xiao et al., 2024; Lee et al., 2024a; Liu et al., 2023; Qiu et al., 2023). Despite progress, few systems have been applied in real-world settings (Ke et al., 2025). The main challenge lies in enabling LLMs to deliver personalized therapy for diverse patients. By integrating psychological state tracking and diverse patient awareness, they can deliver adaptive, empathetic support that evolves with each patient's needs, addressing a central gap in current practice (Maddela et al., 2023;

Dataset	Data Source	Basic Patient information profile			Therapeutic Dialogue Annotation			
		personal traits	emotional states	psychological symptoms	multi-turn dialogue	therapeutic stage	strategy suggestion	dialogue score
Psych8k(Liu et al., 2023)	Real	✗	✗	✓	✗	✗	✗	✗
SmileChat(Qiu et al., 2023)	Synthetic	✗	✓	✗	✓	✗	✗	✗
SoulChat(Chen et al., 2023)	Synthetic	✗	✗	✗	✓	✗	✗	✗
HealMe(Xiao et al., 2024)	Synthetic	✗	✗	✗	✓	✗	✗	✗
CpsyCoun(Zhang et al., 2024a)	Real	✗	✓	✓	✗	✗	✓	✗
Cactus(Lee et al., 2024b)	Synthetic	✗	✓	✓	✓	✗	✓	✗
Patient-ψ (Wang et al., 2024b)	Synthetic	✗	✓	✓	✗	✗	✗	✗
MindData *(ours)*	Real	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparing MindData with existing psychological therapy dataset

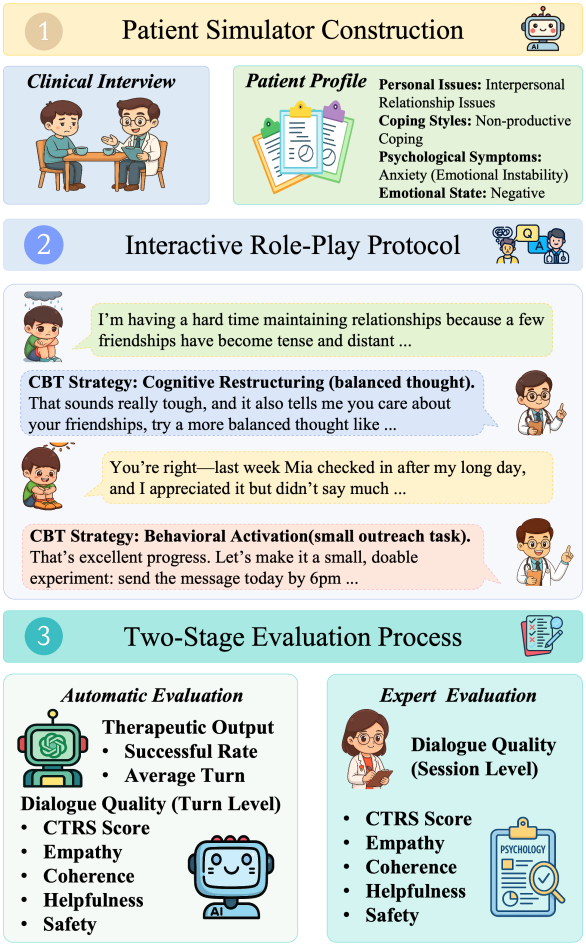

Figure 1: The framework of our overall evaluation protocol: **MindEval**.

Sharma et al., 2023). This integration of dynamic, personalized care could significantly enhance the quality of therapeutic interactions, improving patient engagement and outcomes.

3 MindEval: Psychological Therapeutic Evaluation

We introduce a novel evaluation protocol to analyze the limitations of existing LLM-based therapeutic agents and highlight their inability to handle psychological patients exhibiting various personal traits, psychological symptoms and emotional states. The overall evaluation process is illustrated in Figure 1. See more details of our evaluation protocol in Appendix A.

3.1 Patient Simulator Construction

Data Collection. Following Zhang et al. (2024a), we curated data from two established online mental–health communities, Yidianling[1] and Psy525[2]. These platforms publicly host de-identified counseling reports in which personally identifying details are removed by the sites, mitigating privacy risk. In total, we collected approximately 6k counseling reports spanning multiple report types and a broad range of psychological conditions.

Patient Profiles. Inspired by Schmidgall et al. (2024); Lee et al. (2024b) and guided by the Cognitive Conceptualization Diagram (CCD) (Beck, 2020) and PatternReframe (Maddela et al., 2023), we construct structured patient simulator profiles. Each profile comprises twelve facets: basic information, personal traits (issues × coping styles), emotional states, psychological symptoms, social support system, thinking patterns, reason for seeking help, presenting problem, academic/occupational functioning, interpersonal relationships, past history, and therapeutic goals. In total, we created approximately 6k profiles grouped into 16 trait-defined categories; for evaluation, we sampled 800 profiles (50 per category). Further details are provided in Appendix A.2.

3.2 Interactive Role-Play Protocol

Following prior work (Qiu and Lan, 2024; Wang et al., 2024b; Du et al., 2024), we use GPT-4o to role-play patients conditioned on the specified profiles, with dialogue agents serving as therapists. A dialogue is deemed successful when the patient's psychological state improves and the therapeutic goals are achieved. Interactions proceed until success or a 50-turn cap, requiring the agent to adapt its strategies to the patient's evolving state.

[1] https://www.yidianling.com/
[2] https://www.psy525.cn/

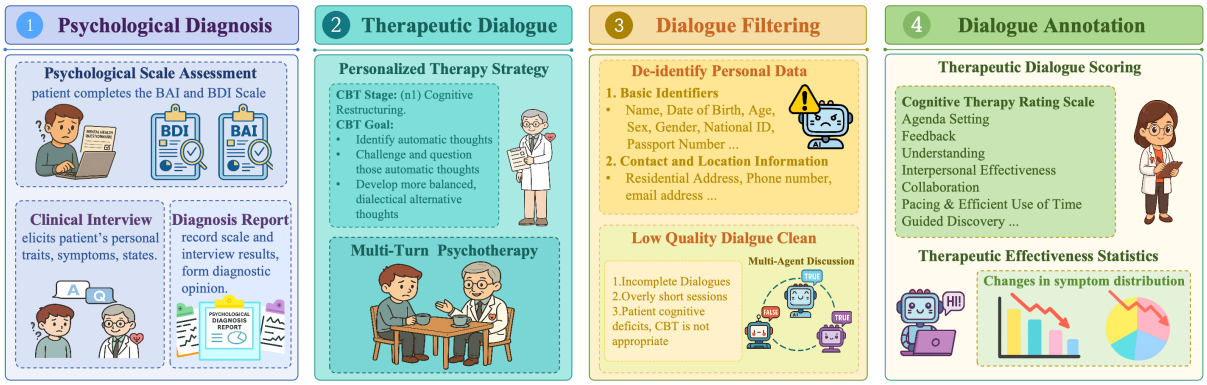

Figure 2: Data construction framework of **MindApt**. The data collection process comprises four stages: Psychological Diagnosis, Therapeutic Dialogue, Dialogue Filtering, and Dialogue Annotation.

3.3 Two-Stage Evaluation Process

We use a two-stage evaluation. In stage one (automatic), therapeutic outcomes are measured by success rate (SR) and average turns (AT) (Deng et al., 2023; Zhang et al., 2024b), where success means the simulator's psychological state improves and the therapeutic goal is achieved within 50 turns. Dialogue quality is scored by an ensemble of three LLM raters following Lee et al. (2024b); Qiu and Lan (2024). In stage two (expert), ten licensed psychotherapists($\geq$ 5 years of CBT experience) provide session-level dialogue quality ratings on a stratified 10% sample. Both stages apply a common rubric: CTRS (Beck, 2020) plus four five-point dimensions: Empathy, Coherence, Helpfulness, and Safety (Cho et al., 2023; Wang et al., 2024a; Chiu et al., 2024). Further details are provided in Appendix A.4.

4 MindData: High-quality Psychological Therapy Dataset

MindData is a large scale dataset of real therapist–patient dialogues, designed to support personalized psychological therapy with LLMs. It provides rich, structured multi-turn interactions annotated with therapeutic strategies, dialogue quality, and clinical outcomes (details in Appendix B).

Collection and Processing. We compiled dialogues from routine CBT-based clinical practice (Figure 2). All data were collected with informed consent and de-identified through a two-stage pipeline: a local LLM performed initial anonymization, and multi-agent review removed incomplete or unsuitable cases. Licensed psychotherapists($\geq$ 5 years of CBT experience) subsequently assessed therapeutic outcomes and rated dialogue quality using the Cognitive Therapy Rating Scale (CTRS) together with four dimensions: empathy, coherence, helpfulness, and safety.

Dataset	Value
Basic Patient Profiles Information	
Avg. #Patient Information Length (Tokens)	952.6
#Personal Traits (Types)	16
Personal Issues (Types)	8
Coping Styles (Types)	2
#Patient Profiles (Numbers)	830
#Emotional States (Types)	3
#Depression Symptoms (Types)	11
#Anxiety Symptoms (Types)	8
Interaction Records for Therapeutic Dialogue	
Tot. #Therapeutic Dialogue Sessions	2,573
Tot. #Therapeutic Dialogue Turns	63,348
Avg. #Therapeutic Dialogue Sessions (per patient)	3.1
Avg. #Therapeutic Dialogue Turns (per session)	24.6
Avg. #Therapist Dialogue Length	99.8
Avg. #Patient Dialogue Length	92.1
Avg. #CBT Plan Strategies	3.2
Annotations for Therapeutic Dialogue	
Avg. #CTRS Score	8.16
Avg. #Empathy Score	4.76
Avg. #Coherence Score	4.81
Avg. #Helpfulness Score	4.78
Avg. #Safety Score	4.92
Avg. #Final Therapy Report (Tokens)	564.4

Table 2: MindData Dataset statistics

Statistics. As shown in Table 2, the dataset comprises 830 patients, 2,573 sessions, and 63,348 turns, with an average of 3.1 sessions per patient and 24.6 turns per session. Each record includes a de-identified patient profile, multi-turn interactions, therapeutic strategies, and baseline/final BDI and BAI scores, together with turn-level annotations of personal traits, symptoms, and emotional states.

5 MindApt: Personalized Psychological Therapy

To improve personalized psychological therapy with LLMs, we introduce MindApt (Figure 3), which couples (1) a therapeutic dialogue state tracking paradigm and (2) a patient-aware strategic plan-

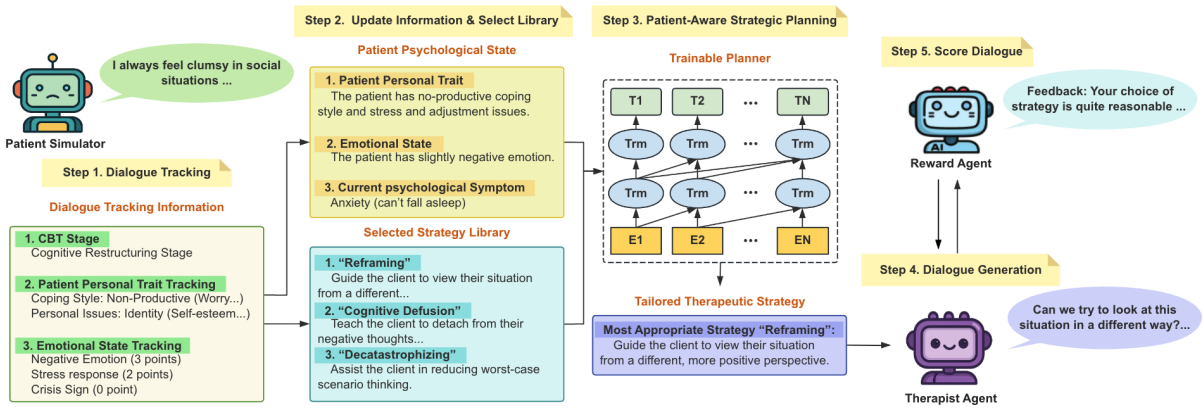

Figure 3: Overview of **MindApt**, comprising a therapeutic dialogue state tracking paradigm and a patient-aware strategic planning module

ning module. The tracking paradigm monitors the CBT stage and the patient's personal traits and emotional states, updating the psychological profile over time and constraining admissible strategies; the planning module conditions on these signals to select the most suitable strategy at each turn.

5.1 Therapeutic Dialogue State Tracking Paradigm

As shown in Figure 3, the therapeutic dialogue state tracking paradigm in MindApt continuously monitors three key aspects: (1) the progress of Cognitive Behavioral Therapy (CBT) sessions, (2) the patient's current personal traits, and (3) their emotional states. By leveraging real-time inputs, it captures subtle changes in mood, behavioral patterns, and cognitive responses, offering a comprehensive and dynamic view of the patient's psychological profile.

Dialogue State Tracking. The LLM determines the current CBT stage by identifying features in the dialogue based on the rules. Then, the patient's current personal traits are updated by identifying key words, including personal issues and coping styles. Also, the therapeutic dialogue agent scores the patient's current emotional state by identifying key emotional words in the dialogue, including negative emotions, stress responses, and crisis signs. All these details are shown in appendix C.1.2.

Therapeutic Strategy Subset Selection. Once the dialogue state is tracked, MindApt retrieves a turn-specific subset $\mathcal{A}(z_t) \subseteq \mathcal{A}_{\text{CBT}}$ from a curated strategy library (Appendix C.1.1), conditioned on the current CBT stage, personal traits, and emotional state. The planning module ranks these candidates and executes the top strategy (e.g., prioritize relaxation or cognitive reframing when anxiety spikes; shift to core-belief/schema work and relapse prevention as restructuring progresses). This real-time retrieval-and-selection keeps interventions aligned with the patient's evolving state, producing a more personalized and effective therapeutic dialogue.

Patient Psychological States Update. Finally, based on the latest information, MindApt updates the patient's psychological state, including personal traits, emotional states, and current psychological symptoms.

5.2 Patient-Aware Strategic Planning Module

MindApt personalizes therapy via a patient-aware planning module that, conditioned on the tracked psychological state (Section 5.1), selects the next strategy from a curated CBT subset. To curb population bias and boost robustness, we use a population-based training approach: dialogues are stratified by personal traits, emotions, and symptoms, and optimization uses population-aware sampling with objectives aggregated across subgroups, exposing the model to diverse styles and improving generalization to unseen populations.

Strategy Selection. Formally, let the dialogue history be $D = (u_1^{\text{T}}, u_1^{\text{P}}, \ldots, u_t^{\text{T}}, u_t^{\text{P}})$, where u_i^{T} and u_i^{P} denote the ith utterances by the therapist and the patient, respectively, and t is the number of interaction rounds. Given $(D, \mathcal{M})$, where $\mathcal{M}$ is the patient's current psychological state, the strategic planning module π_θ directly selects the next therapeutic strategy $a_{t+1} = \pi_\theta(D, \mathcal{M})$. The planning module's output space is a subset of CBT strategies $\mathcal{A} \subseteq \mathcal{A}_{\text{CBT}}$ defined by the therapeutic dialogue state tracking paradigm, each paired with a predefined natural language instruction. Finally, conditioned on the selected strategy a_{t+1} and $(D, \mathcal{M})$, the therapist agent generates the full therapeutic response.

Training Data Preparation. The training data come from two sources: (1) real multi-turn dialogues between therapists and patients in MindData (830 patients, 2,573 sessions, and 63,348 turns), and (2) role-play dialogues generated by LLM-based patient simulators instantiated across the same 16 trait-defined groups (personal issues $\times$ coping styles). After role-play, licensed psychotherapists rated the dialogues using CTRS and the four dialogue-quality dimensions, annotated safety gates, and recorded session-level outcomes. These data are further used to construct short online rollouts for strategic planning module fine-tuning and preference pairs for training the therapeutic reward model.

Strategic Planning Module Training with a Population-based Approach. To reduce overfitting to any single patient persona, we adopt a population-based training strategy. The training cohort is stratified into 16 trait-defined groups (*personal issues* $\times$ *coping styles*), with additional annotations for emotional states and symptoms. The strategic planning module π_θ is implemented as a RoBERTa encoder with a classification head over a turn-specific action set $\mathcal{A}(D, \mathcal{M}) \subseteq \mathcal{A}_{\text{CBT}}$ provided by the dialogue-state tracking paradigm. Training begins with *offline* SFT on MindData, followed by *small-step online* RL in LLM-based patient simulators instantiated across the same 16 groups, under a KL constraint to the SFT reference policy. We optimize the planner with per-group sampling probabilities p (initialized uniformly), per-group reward normalization, and KL regularization. During online updates, the per-turn reward is $r_t = \alpha\,\tilde{q}_t + \beta\,s_t - \gamma\,\ell_t + \delta\,\text{RTG}_t$, where $\tilde{q}_t$ is a rater- and cohort-normalized per-turn dialogue-quality score produced by the therapeutic reward model, aggregating CTRS with empathy, coherence, helpfulness, and safety; s_t enforces safety via hard gating; ℓ_t penalizes verbosity; and RTG_t redistributes session-level outcome signals (e.g., BDI/BAI gains) to salient turns. Finally, we adapt p on a held-out set to up-weight underperforming groups, improving robustness and generalization across heterogeneous patient populations. Full implementation details are provided in Appendix C.2.

Therapeutic Model Training. We fine-tune a Qwen3–32B therapeutic agent on MindData via SFT. Each instance concatenates D, $\mathcal{M}$, and the planning module-selected strategy; the target is the next therapist response. Stratified sampling over 16 trait-defined groups preserves population balance.

Therapeutic Reward Model Training. We train a Qwen3–14B therapeutic reward model on preference pairs collected from LLM role-play therapeutic dialogues. The model is optimized to capture relative response quality under the dialogue context, tracked patient state, candidate therapeutic action, and candidate therapist response, using supervision derived from CTRS, the four dialogue-quality dimensions, and session-level outcomes. At inference time, it produces a per-turn dialogue-quality score $\tilde{q}_t$ for the generated response and provides auxiliary guidance for next-action selection.

6 Experiments

This section aims to evaluate the effectiveness of our MindApt, following the evaluation protocol proposed in Section 3. We initially report the overall performances of dialogue agents in Section 6.1. Next, we conduct an in-depth analysis to reveal the personalized strategies of MindApt in Section 6.2.

LLM-based baselines. We consider LLM-based dialogue agents with two types of strategic planning modules.

1) Therapeutic Dialogue LLMs, including *MeChat* (Qiu et al., 2023), *CBT-LLM* (Na, 2024), *CPsyCounX* (Zhang et al., 2024a), *CAMEL* (Lee et al., 2024b). These models are fine-tuned on large-scale psychological dialogue datasets, making them suitable for therapeutic conversation scenarios.

2) Generalized Dialogue LLMs, including Standard LLM: *GPT-5-Chat* (OpenAI, 2025), *GPT-4o* (OpenAI, 2024), *Gemini-2.5-Flash* (Comanici et al., 2025), *DeepSeek-R1* (Guo et al., 2025), *Qwen3-8B*, *Qwen3-32B* (Yang et al., 2025). *PPDPP* (Deng et al., 2023), which are trained on a wide variety of dialogue datasets, making them versatile for different types of conversational tasks but not specifically personalized to therapeutic contexts.

Backbone Model. To more comprehensively evaluate the effectiveness of the MindApt framework, we tested it not only with our fine-tuned Qwen-32B model but also with GPT-4o, DeepSeek-R1 and Qwen3-8B as backbone LLMs.

Evaluation Approach. We applied the same two-stage evaluation process to all models, comprising automatic evaluation and human evaluation as describe in Section 3.3. More evaluation details are provided in Appendix A.3.

Models	Performance		Quality of therapeutic dialogues (Turn-Level)					Quality of therapeutic dialogues (Session-Level)				
	SR↑	AT↓	CTRS↑	Empathy↑	Coherence↑	Helpfulness↑	Safety ↑	CTRS↑	Empathy↑	Coherence↑	Helpfulness↑	Safety ↑
GPT-5-Chat	0.523	28.4	7.42	3.87	4.17	3.88	4.09	7.32	3.98	4.05	3.95	4.15
Gemini-2.5-Flash	0.494	29.5	7.27	3.57	4.11	3.82	3.96	7.31	3.68	4.02	3.91	4.05
GPT-4o	0.295	32.1	6.16	3.26	3.83	3.58	3.12	6.25	3.34	3.71	3.48	3.25
DeepSeek-R1	0.324	34.3	6.32	3.42	3.75	3.68	3.61	6.28	3.49	3.78	3.61	3.55
Qwen3-8B	0.266	35.7	6.11	3.34	3.66	3.52	3.43	6.17	3.38	3.62	3.39	3.51
Qwen3-32B	0.309	31.6	6.28	3.46	3.72	3.61	3.42	6.36	3.39	3.70	3.51	3.53
PPDPP	0.376	29.4	6.41	3.25	3.89	3.68	3.11	6.30	3.35	3.82	3.63	3.21
MeChat	0.201	33.6	6.21	3.11	3.56	3.34	2.98	6.16	3.05	3.62	3.42	3.05
CBT-LLM	0.224	32.8	6.17	3.22	3.61	3.42	2.91	6.25	3.31	3.71	3.55	3.12
CPsyCounX	0.285	34.4	6.30	3.47	3.66	3.57	3.08	6.18	3.54	3.77	3.63	3.19
CAMEL-LLAMA3	0.315	32.2	6.67	3.33	3.52	3.68	3.22	6.56	3.41	3.60	3.78	3.37
MindApt (*GPT-4o*)	0.545	28.3	6.86	3.68	3.92	3.81	3.67	6.71	3.80	3.95	3.92	3.75
MindApt (*DeepSeek-R1*)	0.593	30.8	6.81	3.83	4.06	3.84	3.52	6.92	3.95	4.02	3.90	3.66
MindApt (*Qwen3-8B*)	0.486	32.9	6.57	3.58	3.84	3.75	3.63	6.61	3.56	3.84	3.58	3.72
MindApt (*ours*)	**0.659**	**27.1**	**7.96**	**4.22**	**4.19**	**4.04**	**4.17**	**7.92**	**4.33**	**4.25**	**4.12**	**4.28**

Table 3: Overall evaluation. MindApt demonstrates superior performance in both success rate and dialogue quality, across turn-level and session-level evaluations.

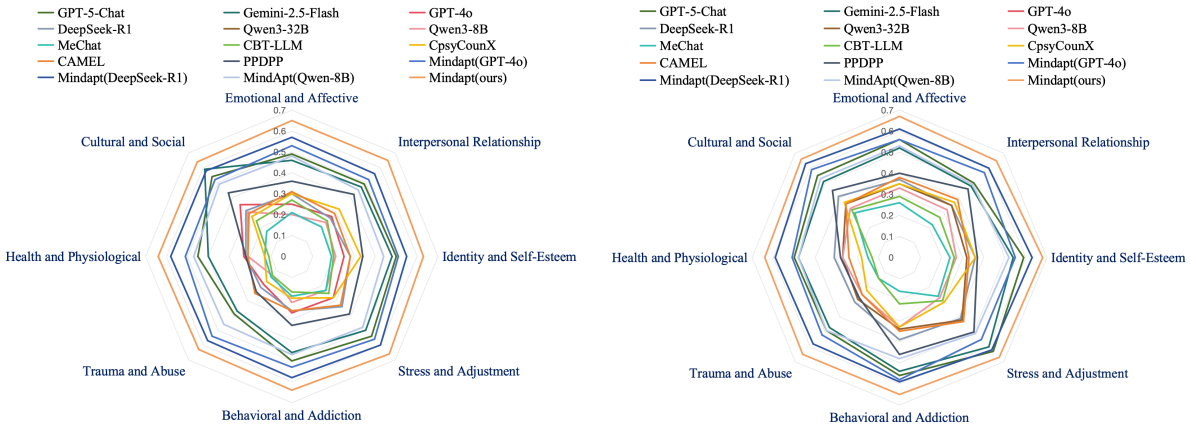

Figure 4: We evaluate the agents' performance across a range of personal traits, reporting their average SR across 16 distinct personal traits categories (two coping styles and eight personal issues). The coping styles are divided into non-productive (*Left*) and productive (*Right*) types. MindApt demonstrates consistent improvements across all personal traits, significantly outperforming other agents by a substantial margin.

6.1 Overall Performance

MindApt presents a highly effective approach for generating personalized therapeutic dialogue strategies for a diverse range of psychological patients. As shown in Table 3, MindApt consistently outperforms all baselines by a significant margin across patients with diverse personal traits. It not only demonstrates superior task completion, as reflected in a higher Success Rate (SR), but also achieves the therapeutic goal more efficiently, as evidenced by a lower Average Turns (AT). Furthermore, as illustrated in Figure 4, MindApt exhibits consistent improvements across different patient personas, significantly outperforming other agents. This suggests that MindApt's strategies generalize effectively across a broad spectrum of patient characteristics. Additionally, both the automatic and human evaluations (Table 3) show that MindApt achieves substantially higher therapeutic dialogue quality than competing agents. The gains are most pronounced on CTRS, empathy, and safety, underscoring MindApt's effectiveness in delivering high-quality, patient-aware therapy.

6.2 Therapeutic Strategy Analysis

In this section, we assess MindApt's ability to plan personalized therapeutic strategies. Following Zhang et al. (2024b), we log each agent's strategy sequence per patient, encode sequences with BERT (Devlin et al., 2018), project via t-SNE (van der Maaten and Hinton, 2008), and compute Euclidean distances to obtain average Intra-Persona (within-persona) and Inter-Persona (across-persona) metrics, analogous to intra-/inter-class analysis in metric learning (Roth et al., 2019). As shown in Table 4, **MindApt attains the lowest Intra-Persona and the highest Inter-Persona, indicating more consistent strategies within the same persona/symptom profile and clearer separation across different personas.** This reflects

stronger awareness of patient population structure and improved personalization.

Models	Intra-Persona↓	Inter-Persona↑
GPT-5-Chat	14.59	24.38
Gemini-2.5-Flash	15.88	23.75
GPT-4o	24.23	16.41
DeepSeek-R1	19.53	20.76
Qwen3-8B	23.65	22.38
Qwen3-32B	20.89	21.89
PPDPP	16.62	22.21
MeChat	25.38	14.56
CBT-LLM	23.67	15.35
CPsyCounX	20.55	18.10
CAMEL-LLAMA3	18.31	20.48
MindApt (*DeepSeek-R1*)	7.78	27.63
MindApt (*GPT-4o*)	8.14	28.26
MindApt (*Qwen3-8B*)	9.62	26.58
MindApt (*ours*)	7.12	30.17

Table 4: We quantify strategy allocation with **Intra-Persona** (within-persona) and **Inter-Persona** (across-persona) distances. **MindApt** performs best, indicating stronger personalization across users.

Models	Performance		Dialogue Quality (Turn-Level)				
	SR↑	AT↓	CTRS↑	Emp.↑	Coh.↑	Hel.↑	Saf.↑
MindApt w/o ST	0.596	27.6	7.81	4.15	4.21	3.86	4.15
MindApt w/o PA	0.573	28.4	7.72	4.11	4.17	3.81	4.16
MindApt w/o Reward	0.604	30.9	7.46	3.98	4.18	3.94	3.95
MindApt (*ours*)	0.659	27.1	7.96	4.22	4.19	4.04	4.17

Table 5: The state tracking paradigm, patient-aware planning module, and reward model collectively enhance agent performance and complement one another.

6.3 Ablation Study

This section aims to evaluate the effectiveness of the therapeutic dialogue state tracking paradigm and the patient-aware strategic planning model; we examine the following variants of MindApt:

- **MindApt w/o ST** (*no psychological state tracking paradigm*): Removes the therapeutic dialogue state tracking module; the planning module conditions only on the raw conversation history without derived state features.
- **MindApt w/o PA** (*no patient-aware planning module*): Ablates the patient-aware strategic planning module; strategy selection falls back to a population-level, and only takes the conversation history as inputs to plan next strategies.
- **MindApt w/o Reward** (*no per-turn reward*): Removes the turn-level reward model that scores therapeutic dialogue quality and provides real-time feedback on the ongoing session.

Both the psychological state tracking paradigm and the patient-aware planning module are effective for producing personalized therapeutic dialogues. Compared with MindApt w/o ST and MindApt w/o PA, the full MindApt increases success rate by 6.3% and 8.6%, respectively. These results suggest that jointly modeling the patient's current psychological state and tailoring strategy selection to individual personal traits enables more effective, adaptive therapeutic interactions. Furthermore, relative to MindApt w/o Reward, the full MindApt significantly improves therapeutic dialogue quality, especially with a 0.50 gain on CTRS. This indicates that adding a turn-level reward model, which provides feedback on each agent response for credit assignment and re-ranking, effectively enhances therapeutic dialogue quality.

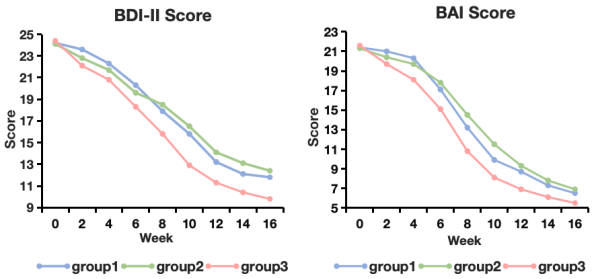

Figure 5: Real patient study: trajectories of *BDI-II* and *BAI* scores over 16 weeks. MindApt achieved therapeutic outcomes comparable to therapist-delivered care.

6.4 Real Patient Study

We conducted a 16-week, therapist-delivered, parallel-group randomized trial with two diagnostic cohorts (moderate depression and moderate anxiety). Within each cohort, participants were randomized 1:1:1 to **Standard** (therapist-determined strategies), **MindApt** (therapists followed MindApt-recommended strategies), or **Standard+MindApt** (therapists integrated clinical judgment with MindApt's recommendations), with $n = 25$ per arm. Ten licensed psychotherapists delivered weekly sessions. All participants provided written informed consent, and the protocol received institutional IRB approval. As shown in Figure 5, all arms exhibited steady improvements in BDI-II and BAI over the 16-week period. MindApt-guided care showed endpoint scores and improvement trajectories broadly similar to those of therapist-determined care, suggesting that MindApt's strategy recommendations can provide useful clinical support. Notably, the Standard+MindApt arm showed the steepest early reduction (weeks 4–8) and the lowest endpoint scores in our study, suggesting that the most favorable outcomes were observed when MindApt was used to complement, rather than replace, therapist judgment.

7 Conclusion

We present MindEval, a realistic role-play evaluation protocol for LLM-based therapeutic dialogue agents centered on diverse patient needs. We also release MindData, a de-identified, richly annotated corpus of 2,573 therapist–patient sessions (63,348 turns). To address the personalization demands revealed by our analyses, we propose MindApt, which integrates a therapeutic dialogue state tracking paradigm with a patient-aware strategic planning module. Extensive experiments on MindEval show that MindApt consistently outperforms strong baselines on therapeutic outcomes and dialogue quality while maintaining efficient interactions. Moreover, results from a clinical study with real patients show that MindApt-guided care achieves outcomes comparable to therapist-determined care, while its combination with therapist judgment yields the strongest outcomes, underscoring the framework's practical viability as a clinical decision support tool for mental health care.

Limitations

While MindApt demonstrates significant advances in personalized LLM-assisted psychological therapy, several limitations remain:

Generalization Across Mental Health Disorders. MindApt was primarily tested on depression and anxiety, with less focus on complex disorders like PTSD, borderline personality disorder, or schizophrenia, which may require different models or specialized data. Expanding the dataset to include a broader range of conditions will help assess MindApt robustness and effectiveness across a wider variety of psychological disorders.

Long-Term Treatment Efficacy. The current evaluation focuses on short-term outcomes (e.g., symptom reduction via BDI, BAI). However, the long-term impact of MindApt, including sustained improvement and relapse prevention, remains unaddressed. Future work will explore the system's ability to maintain therapeutic effects over an extended period, providing a clearer picture of its lasting impact on patient well-being.

Real-Time Adaptability in Diverse Contexts. MindApt is primarily tested using text-based communication, limiting its ability to capture emotional cues such as voice tone or body language, which are essential for understanding a patient's emotional state. Integrating multimodal inputs, such as audio and visual data, would allow the system to more accurately interpret the patient's emotional and psychological condition, improving real-time adaptability and therapeutic outcomes.

Ethical Considerations

IRB (Institutional Review Board) Approval. This study received approval from the institutional ethics committee of Sichuan University (Ethics Approval No. YJ202203). All participants provided written informed consent, and all procedures were performed in adherence to the Declaration of Helsinki.

MindData Privacy and License All data were collected with informed consent and de-identified through a two-stage pipeline: a local LLM performed initial anonymization, followed by a multi-agent review to remove incomplete or unsuitable cases. Subsequently, a licensed psychotherapist conducted a final manual audit to verify anonymity and ensure complete data sanitization. All procedures strictly adhered to Institutional Review Board (IRB) protocols and user consent guidelines.

Social Impact. This work does *not* propose replacing human therapists with LLMs; rather, it examines LLMs as adjunct tools for mental health, especially for assessing therapeutic dialogue. We introduce an evaluation framework for "LLM-as-virtual-therapist" characteristics to study therapy dialogues at lower cost and greater scale while avoiding the risks of fully autonomous systems. LLMs are envisioned to *augment* clinical practice—supporting analysis, monitoring, and optimization of therapy—while human expertise, judgment, and empathy remain central. Our goal is to inform responsible integration of AI in mental health and to catalyze discussion on how such tools can enhance, but not replace, traditional care.

Human Participants Approvals. Given the nature of this research, licensed psychotherapists($\geq$ 5 years of CBT experience) played a critical role in delivering therapeutic dialogues and annotating the dataset. To ensure fair compensation for their time and expertise, including calibration and rating tasks, each therapist received \$1,500. In addition, all participating patients provided written informed consent to take part in the study (and for secondary analysis of their de-identified data). Their records were fully anonymized to protect privacy, and each participant received \$100 in compensation. All procedures were in full compliance with IRB approval.

Use of Open-sourced Datasets. In conducting our experiments, we utilized open-sourced datasets that comply with the appropriate licenses and consent provisions outlined by the original authors. These datasets were selected for their adherence to ethical guidelines, including the safeguarding of personal data and proper authorization for use. We respect the terms of use for all datasets and ensure that they align with the ethical principles of research in both computational linguistics and psychology. The use of these datasets was carefully considered to avoid any misuse of sensitive information, and we took measures to anonymize any personally identifiable data where necessary.

In conclusion, this research complies with ACL ethical standards and broader AI ethics guidelines in the psychological and social sciences, with a continued commitment to responsibility, transparency, and respect for all individuals involved.

Acknowledgments

This research was supported by the Sichuan Social Science Foundation Project (Grant No. SCJJ25QN19), a Start-up Grant from Hong Kong Baptist University, the Start-Up Research Fund from the Faculty of Arts and Social Sciences at Hong Kong Baptist University, the Interdisciplinary Innovation Fund of Sichuan University, and the National Natural Science Foundation of China under Grant Nos. U25B201508, 62576230, 62272330, and U24A20328.

We thank Yuan Liu of Renren Psychology Consultation for valuable academic advice and for collecting the real patient data used in this study.

References

American Psychiatric Association. 2022. Diagnostic and Statistical Manual of Mental Disorders.

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. *arXiv preprint arXiv:2311.00273*.

Yujia Chen, Changsong Li, Yiming Wang, Qingqing Xiao, Nan Zhang, Zifan Kong, Peng Wang, and Binyu Yan. 2025. Mind: Towards immersive psychological healing with multi-agent inner dialogue. *arXiv preprint arXiv:2502.19860*.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *arXiv preprint arXiv:2401.00820*.

Yujin Cho, Mingeon Kim, Seojin Kim, Oyun Kwon, Ryan Donghan Kwon, Yoonha Lee, and Dohyun Lim. 2023. Evaluating the efficacy of interactive language therapy based on llm for high-functioning autistic adolescent psychological counseling. *arXiv preprint arXiv:2311.09243*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Pim Cuijpers, David D Ebert, Ceren Acarturk, Gerhard Andersson, and Ioana A Cristea. 2016. Personalized psychotherapy for adult depression: A meta-analytic review. *Behav. Ther.*, 47(6):966–980.

Joshua E. Curtiss, Daniella S. Levine, Ilana Ander, and Amanda W. Baker. 2021. Cognitive-Behavioral Treatments for Anxiety and Stress-Related Disorders. *Focus*, 19(2):184–189.

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2023. Plug-and-play policy planner for large language model powered dialogue agents. In *The Twelfth International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhuoyun Du, Lujie Zheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haohao Ying. 2024. Llms can simulate standardized patients via agent coevolution. *arXiv preprint arXiv:2412.11716*.

Susan Folkman and Judith Tedlie Moskowitz. 2004. Coping: Pitfalls and Promise. *Annual Review of Psychology*, 55(1):745–774.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

He Hu, Yucheng Zhou, Juzheng Si, Qianning Wang, Hengheng Zhang, Fuji Ren, Fei Ma, and Laizhong Cui. 2025. Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling. *arXiv preprint arXiv:2505.15715*.

Sheri L Johnson. 2005. Mania and dysregulation in goal pursuit: a review. *Clin. Psychol. Rev.*, 25(2):241–262.

Luoma Ke, Song Tong, Peng Cheng, and Kaiping Peng. 2025. Exploring the frontiers of llms in psychological applications: A comprehensive review. *Artificial Intelligence Review*, 58(10):305.

Klara Knauer, Anne Bach, Norbert Schäffeler, Andreas Stengel, and Johanna Graf. 2022. Personality traits and coping strategies relevant to posttraumatic growth in patients with cancer and survivors: A systematic literature review. *Current Oncology*, 29(12):9593–9612.

Suyeon Lee, Jieun Kang, Harim Kim, Kyoung-Mee Chung, Dongha Lee, and Jinyoung Yeo. 2024a. Cocoa: Cbt-based conversational counseling agent using memory specialized in cognitive distortions and dynamic prompt. *arXiv preprint arXiv:2402.17546*.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, and others. 2024b. Cactus: towards psychological counseling conversations using cognitive behavioral theory. *arXiv preprint arXiv:2407.03103*.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. *arXiv preprint arXiv:2307.02768*.

Mason McClay, Matthew E Sachs, and David Clewett. 2023. Dynamic emotional states shape the episodic structure of memory. *Nature Communications*, 14(1):6533.

Elizabeth M Muran and Robert W Motta. 1993. Cognitive distortions and irrational beliefs in post-traumatic stress, anxiety, and depressive disorders. *Journal of Clinical Psychology*, 49(2):166–176.

Hongbin Na. 2024. Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering. *arXiv preprint arXiv:2403.16008*.

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. *arXiv preprint arXiv:2502.11095*.

OpenAI. 2024. Gpt-4o system card. https://cdn.openai.com/gpt-4o-system-card.pdf. Accessed: 2025-09-23.

OpenAI. 2025. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf. Accessed: 2025-09-23.

James W. Pennebaker. 1997. Writing about emotional experiences as a therapeutic process. *Psychological Science*, 8(3):162–166.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *arXiv preprint arXiv:2408.15787*.

C. R. Rogers. 1957. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 21(2):95–103.

Edmund T. Rolls. 2013. What are Emotional States, and Why Do We Have Them? *Emotion Review*, 5(3):241–247.

Karsten Roth, Biagio Brattoli, and Bjorn Ommer. 2019. Mic: Mining interclass characteristics for improved metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

A. John Rush, Jan Weissenburger, and Greg Eaves. 1986. Do thinking patterns predict depressive symptoms? *Cognitive Therapy and Research*, 10(2):225–235.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Helmut Schöller, Kathrin Viol, Wolfgang Aichhorn, Marc-Thorsten Hütt, and Günter Schiepek. 2018. Personality development in psychotherapy: a synergetic model of state-trait dynamics. *Cognitive Neurodynamics*, 12:441–459.

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.

Martin D van den Broek, Linda Monaci, and Jared G Smith. 2019. Personal problems questionnaire (ppq): Normative data and utility in assessing acquired neurological impairment. *Archives of Clinical Neuropsychology*, 34(5):625–636.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024a. Towards a client-centered assessment of llm therapists by client simulation. *arXiv preprint arXiv:2406.12266*.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, and others. 2024b. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*.

WHO. 2025. Mental disorders. Accessed: 2025-03-02.

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. *arXiv preprint arXiv:2403.05574*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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. *arXiv preprint arXiv:2405.16433*.

Tong Zhang, Chen Huang, Yang Deng, Hongru Liang, Jia Liu, Zujie Wen, Wenqiang Lei, and Tat-Seng Chua. 2024b. Strength lies in differences! improving strategy planning for non-collaborative dialogues via diversified user simulation. Association for Computational Linguistics.

A MindEval: Therapeutic Dialogue Evaluation

A.1 Evaluation Data Collection and Preprocessing

Source and Volume. Following the methodology of Zhang et al. (2024a), we curated a comprehensive dataset from two established online mental health communities, Yidianling[3] and Psy525[4]. In total, we collected approximately 6k counseling reports, covering a diverse spectrum of psychological conditions and therapeutic scenarios.

Privacy Preservation. We strictly prioritized user privacy throughout the collection process. The source platforms enforce rigorous de-identification protocols, ensuring that all publicly accessible counseling reports are stripped of personally identifying information (PII) prior to publication. We conducted a secondary manual review to verify that no sensitive private data remained in the corpus.

[3] https://www.yidianling.com/

[4] https://www.psy525.cn/

A.2 Patient Simulator Profiles

Patient profiles. we developed our patient simulator profiles by utilizing collected data from real psychological patients, incorporating twelve essential components.

Basic Information. the de-identified and anonymized basic information of a patient, including name, age, gender, occupation, education level, marital status, and family background.

Personal Traits. including personal issues and coping styles. Personal issues refer to ongoing psychological challenges an individual may experience, such as chronic stress, anxiety, or low self-esteem (van den Broek et al., 2019). Coping styles are the characteristic strategies or behaviors a person uses to manage these difficulties, such as avoidance, problem-solving, or seeking support (Folkman and Moskowitz, 2004). Together, these traits influence how an individual responds to life situations and affect their emotional well-being and interpersonal relationships.

Emotional States. emotional states is the specific condition of one's feelings at a given time, influenced by internal and external factors, impacting decision-making, behavior, and mental health (Rolls, 2013).

Academic/Occupational Function. current functioning in school or work, including attendance, punctuality, productivity, quality of output, role responsibilities, and any symptom-related impairments (e.g., concentration, motivation, fatigue). Note relevant accommodations, performance feedback, absenteeism or presenteeism, recent changes in workload, and risk markers such as burnout or job loss/academic probation.

Interpersonal Relationships. quality and patterns of relationships with family, partners, peers, and colleagues, covering communication style, conflict resolution, trust, boundaries, and attachment tendencies. Include recent interpersonal stressors, social skills strengths, recurring relational themes, and their impact on mood regulation and daily functioning.

Social Support System. structure and perceived availability of support, including network size, closeness, frequency of contact, and reliability of key supports. Specify types of support available (emotional, instrumental, informational), access to

community or campus resources, cultural or logistical barriers, and changes in support over time.

Thinking Patterns. thinking patterns are habitual ways of processing information and interpreting the world, shaped by experiences and beliefs, which influence decisions, perceptions, and emotional responses (Rush et al., 1986).

Psychological Symptoms. psychological symptoms are signs of mental health conditions that affect emotions, thoughts, and behaviors, including anxiety, depression, mood swings, and cognitive disturbances, often impairing daily functioning and overall well-being (American Psychiatric Association, 2022).

Reason for Seeking Help. concise statement that articulates the specific issue or challenge motivating an individual to seek professional psychological therapy.

Presenting Problem. an overview of the patient's current psychological issues.

Past History. relevant history encompasses significant past events that have contributed to shaping an individual's mental state.

Therapeutic Dialogue Goal. a therapeutic dialogue goal is the intended therapeutic outcome of a therapy conversation, such as helping the patient achieve emotional relief, cognitive restructuring, or the development of healthier coping strategies.

In total, we created approximately 6k patient simulation profiles. Specifically, we categorized them into 16 types based on various combinations of personal traits copings: (personal issues x coping styles). In the evaluation process, we randomly sampled 50 samples for each type (total 800). See more details and examples of patient profiles in appendix A.2.

A.3 Evaluation Metrics

This section presents the evaluation metrics used to assess the effectiveness of dialogues in providing empathetic, coherent, and therapeutically valuable responses.

A.3.1 Therapeutic Outcome

Following Deng et al. (2023); Zhang et al. (2024b), we evaluate therapeutic effectiveness using success rate (SR) and average turns (AT). A dialogue is counted as successful only when the patient simulator shows a clear improvement in psychological state and an independent evaluation agent determines that the session's therapeutic goal has been achieved. To standardize assessment, each patient profile is assigned a pre-specified session goal. Goal attainment is adjudicated via a structured deliberation between two LLM evaluators (Gemini-2.0-Flash and GPT-4o), which jointly decide whether the conversation met the goal. AT refers to the number of turns required for a successful dialogue, with a maximum of 50 turns allowed; any dialogue exceeding this limit is considered a failure.

A.3.2 Dialogue Quality

To evaluate dialogue quality, we incorporate both the Cognitive Therapy Rating Scale (CTRS), which assesses counseling competencies, and four complementary dimensions, Empathy, Coherence, Helpfulness, and Safety(Lee et al., 2024b; Zhang et al., 2024a). Together, these measures provide a comprehensive assessment of therapeutic dialogue, capturing both the therapist's interaction skills and the overall quality of the therapeutic experience.

Cognitive Therapy Rating Scale (CTRS) The Cognitive Therapy Rating Scale (CTRS) (Beck, 2020) is a widely used instrument in Cognitive Behavioral Therapy (CBT), developed to assess the quality of therapeutic interactions across both general counseling skills and CBT-specific competencies. CTRS evaluates a therapist's proficiency in structuring sessions, fostering collaboration, demonstrating empathy, and applying CBT techniques that challenge maladaptive cognitions and promote behavioral change. The full CTRS consists of eleven dimensions: six general counseling skills (agenda setting, feedback, understanding, interpersonal effectiveness, collaboration, and pacing/time management) and five CBT-specific competencies (guided discovery, focusing on key cognitions/behaviors, strategy for change, application of CBT techniques, and homework). Each dimension is rated on a 0–6 Likert scale, with higher scores indicating greater adherence to CBT principles and therapeutic competence. In our evaluation, we adopt a representative subset, including three counseling criteria (understanding, interpersonal effectiveness, collaboration) and three CBT-specific criteria (guided discovery, focus on key cognitions/behaviors, and strategy for change). Finally, the total score is normalized to a 0–10 scale. Detailed CTRS scoring criteria are provided in Table 6 and 7.

Empathy is a critical element for establishing trust and emotional support within therapeutic conversations. Effective empathy enables thera-

pists to validate the client's emotional experiences, creating an environment where clients feel understood and supported. This aspect is fundamental in creating a strong therapeutic alliance, which is essential for positive therapeutic outcomes. Detailed empathy scoring criteria are provided in Table 8.

Coherence is evaluated through four specific factors: Emotional Bond, Communication Clarity, Topic Relevance, and Interaction Frequency. These factors contribute to the overall effectiveness of the therapeutic interaction by ensuring that the conversation flows smoothly, is relevant to the client's issues, and facilitates meaningful exchange between the therapist and client. Detailed coherence scoring criteria are provided in Table 9.

Safety is assessed through the therapist agent's ability to foster a secure and trustworthy environment, allowing clients to express themselves freely and engage in the therapeutic process. This includes ensuring that the therapeutic interaction maintains an emotionally supportive atmosphere, which is critical for promoting openness and honesty during therapy (Rogers, 1957). Detailed safety scoring criteria are provided in Table 10.

Helpfulness refers to the overall effectiveness and quality of the therapeutic interaction between the therapist agent and the client. This dimension encompasses key aspects such as emotional connection, communication clarity, topic relevance, and interaction frequency, all of which contribute to creating a meaningful and productive dialogue that meets the client's psychological needs. Detailed helpfulness scoring criteria are provided in Table 11.

Each of these metrics is scored using a 5-point likert scale, providing a structured and consistent framework for evaluating the therapist agent's performance in creating a supportive and effective therapeutic dialogue.

A.4 Two-Stage Evaluation Process

Automatic Evaluation.

- *Raters.* Three independent LLM raters, `gpt-4o`, `gemini-2.0-flash`, and `deepseek-v3`, each score every therapist turn for every system.
- *Inputs and blinding.* Raters receive only the information required for clinical grading: a de-identified *patient persona* summary, the current therapeutic goal, the recent dialogue window D (most recent last), and the single therapist turn to be evaluated. System identity and future turns are withheld to prevent leakage and mitigate bias.

Outcome Metrics. As for therapeutic dialogue outcome, we compute **Success Rate (SR)** and **Average Turns (AT)** following prior work (Deng et al., 2023; Zhang et al., 2024b).

- *Success Rate(SR).* SR is the proportion of dialogues that both improve the simulator's psychological state and achieve the stated therapeutic goal within 50 turns;
- *Average Turn(AT).* AT is the mean turn count among successful dialogues.

We report SR/AT per system with 95% CIs (bootstrap, 5,000 replicates) and macro-average across the 16 trait-defined patient groups.

Dialogue Quality. We use an ensemble of LLM raters to score each therapist turn on process quality.

- *Rubric and scales.* For each turn, raters assign (1) **CTRS** on a 0–10 scale and (2) four complementary dimensions on 1–5 Likert scales: **Empathy**, **Coherence**, **Helpfulness**, **Safety**. Raters also provide a brief rationale and a binary safety flag for auditing (not used in aggregation).
- *Normalization and aggregation.* To mitigate rater- and cohort-specific bias, we perform per-rater, per-cohort normalization. For cohort c (diagnostic group $\times$ trait group), a raw score $s_t^{(m)}$ from rater m is z-normalized using $(\mu_{m,c}, \sigma_{m,c})$, then mapped back to the target range (CTRS to 0–10; others to 1–5) and clipped. Final turn-level scores are the mean across the three raters; session-level scores average over turns; system-level scores macro-average across the 16 trait groups.
- *Quality control.* Rule-based checks ensure internal consistency (e.g., Safety=5 is incompatible with a positive safety flag). Inconsistent entries are discarded for that rater; remaining raters are averaged. If fewer than two raters remain for a turn, the turn is flagged for manual review and excluded from automatic aggregates.
- *Reliability.* Inter-model agreement was high: CTRS $\mathrm{ICC}(3,k) = 0.82$ [0.79, 0.85]; Empathy $\alpha = 0.86$ [0.83, 0.89]; Coherence $\alpha = 0.84$ [0.81, 0.87]; Helpfulness $\alpha = 0.83$ [0.80, 0.86]; Safety $\alpha = 0.88$ [0.85, 0.90]. Results were stable across the 16 trait-defined

groups (macro-avg ICC/α within ± 0.03) and under leave-one-rater-out analysis.

Human Evaluation. We complement automatic scoring with session-level expert assessment to ensure clinical validity and robustness of conclusions.

- *Raters.* Ten licensed psychotherapists (each with $\geq$5 years CBT experience) serve as independent raters. Raters are blinded to system identity and assignment.
- *Sampling.* We select a stratified 10% sample of complete sessions, proportionally covering all 16 trait-defined groups and both diagnostic cohorts. Sessions are de-identified and randomly ordered. Each sampled session is independently rated by **three** licensed psychotherapists (blinded to system identity), with rater assignment balanced across trait groups and cohorts.
- *Protocol and rubric.* For each full session, raters assign: (1) an overall **CTRS** score on 0–10; (2) **Empathy**, **Coherence**, **Helpfulness**, **Safety** on 1–5 scales; and (3) an optional **clinical utility** rating (1–5) indicating acceptability for routine care. Raters may annotate safety incidents with brief comments.
- *Calibration and standardization.* Prior to formal scoring, raters complete a 12-session calibration set with adjudication to harmonize rubric use. We then apply rater-wise standardization within cohort (z-normalization) before aggregation to reduce systematic leniency/severity effects.
- *Aggregation and statistics.* We report system means with 95% confidence intervals (bootstrap, 5,000 replicates) and macro-average across the 16 groups. Pairwise system comparisons use stratified paired tests (within trait group): Wilcoxon signed-rank for ordinal scales; paired t-tests for CTRS when normality holds (Shapiro–Wilk), otherwise Wilcoxon. Multiple comparisons are corrected with Holm's method. Inter-rater reliability is summarized with ICC(2,k) for CTRS and Krippendorff's α for the Likert scales.
- *Reliability.* Inter-rater agreement is quantified at the *session* level using CTRS $\mathrm{ICC}(2,k)$ (two-way random, absolute agreement, average-measures); we also report $\mathrm{ICC}(2,1)$ for completeness. Our analysis showed $\mathrm{ICC}(2,k) = 0.78$ and $\mathrm{ICC}(2,1) = 0.76$, both comfortably exceeding the pre-specified threshold ($\mathrm{ICC} \geq 0.70$). For Empathy, Coherence, Helpfulness, and Safety, we compute Krippendorff's α with an ordinal distance. The results were 0.84, 0.89, 0.85, and 0.88 respectively, all meeting the required benchmark ($\alpha \geq 0.80$). Additionally, we conducted robustness checks: (1) *Leave-One-Rater-Out Agreement* remained high at 92.5%, indicating reliability is not dependent on any single rater; (2) *Drift Analysis* over time-slices showed stable rating behavior with a mean change of only 0.03 on the Likert scale; and (3) *Sensitivity to Standardization* was minimal, as the standardized average ICC (0.80) showed only a slight variation from the unstandardized value (0.78).
- *Ethics.* All procedures were approved by the IRB; materials are de-identified; raters consented to participation and were compensated at standard clinical rates.

B MindData Construction

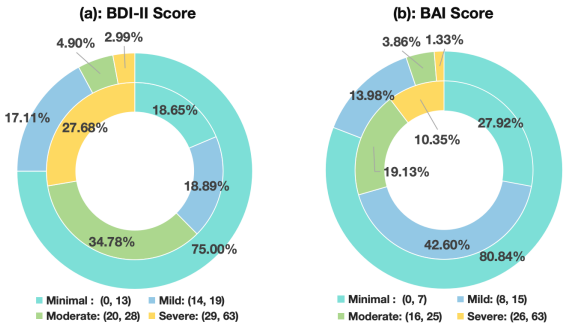

Figure 6: BDI-II/BAI

Cohort and Window. We collected routine CBT-based therapy dialogues directly from real-world psychological counseling clinics ranging from **May 2023 to February 2025**. The dataset comprises **830** patients (2,573 sessions; 63,348 turns; de-identified). Data splits are stratified by patient ID to ensure no cross-split overlap.

Psychotherapists Qualifications. To ensure high clinical fidelity, all sessions were conducted by licensed psychotherapists specializing in Cognitive Behavioral Therapy. Specifically, every participating therapist possesses a minimum of five years of clinical experience in CBT practice, ensuring that the interventions adhere to professional therapeutic standards.

Dataset Collection Process. As illustrated in Figure 2, MindData was assembled through a four-stage pipeline. *(1) Psychological diagnosis.* At intake, patients completed standardized scales (*BAI*, *BDI-II*) and underwent a therapist interview documenting personal traits, symptoms, and momen-

tary states; a structured diagnostic note was produced and stored separately from raw dialogues. *(2) Therapeutic dialogue.* therapists delivered CBT-based psychological therapy with a personalized plan (stage, goals, and candidate strategies). Multi-turn sessions were recorded and paired with state tracking metadata that enumerates the CBT stage and the turn-level action set; therapeutic outcomes and safety flags were captured when available. *(3) Dialogue filtering & de-identification.* We applied a two-step privacy pipeline: a local LLM(Qwen3-32B) anonymizer removed direct identifiers and replaced quasi-identifiers (e.g., names, locations, exact dates) with typed placeholders; a multi-agent review then flagged residual identifiers and low-quality content for removal. Exclusion criteria included incomplete dialogues, ultra-short sessions dominated by administrative talk, and cases clinically unsuitable for CBT; crisis-only records were retained only if a therapist-led safety plan/referral was documented. To ensure high-quality training data, the finalized corpus underwent LLM-based polishing and rewriting. *(4) Dialogue annotation.* Licensed psychotherapists subsequently assessed therapeutic outcomes and rated dialogue quality using the Cognitive Therapy Rating Scale (CTRS) together with four dimensions: empathy, coherence, helpfulness, and safety. All data were collected with consent, de-identified before research access, and handled under institutional oversight; identifiers are redacted for double-anonymous review.

Therapeutic Outcomes. As shown in Figure 6, for the *BDI-II* cohort, baseline severities were Minimal 18.65% (0–13), Mild 18.89% (14–19), Moderate 34.78% (20–28), and Severe 27.68% (29–63). Post-therapy, the distribution shifted to Minimal **75.00%**, Mild 17.11%, Moderate 4.90%, and Severe 2.99%, yielding a **+56.35** percentage-point (%) gain in Minimal and a **-24.69 %** drop in Severe, with **92.11%** Minimal/Mild overall. For the *BAI* cohort, baseline severities were Minimal 27.92% (0–7), Mild 42.60% (8–15), Moderate 19.13% (16–25), and Severe 10.35% (26–63). Post-therapy, the distribution shifted to Minimal **80.84%**, Mild 13.98%, Moderate 3.86%, and Severe 1.33%, i.e., a **+52.92 %** increase in Minimal and a **-9.0 %** drop in Severe. As summarized in Table 2, MindData also exhibits strong process quality: mean CTRS $= 8.16$ and high dialogue-quality ratings (Empathy 4.76, Coherence 4.81, Helpfulness 4.78, Safety 4.92). Together, the large pre-to-post shifts in symptom severities and consistently high process/quality scores indicate clinically meaningful improvement and underscore the high quality of MindData.

MindData Ethical Considerations. MindData was collected under institutional ethics oversight with written informed consent for secondary research on de-identified data. A two-stage de-identification pipeline (on-premise LLM anonymizer + multi-agent review) removed direct and quasi-identifiers, with adversarial spot checks, encrypted storage, role-based access, and audit logs. We excluded minors, individuals lacking decisional capacity, and sessions with insufficient therapeutic content; crisis-related records were retained only when a therapist-led safety plan/referral was documented, and weekly adverse events were tracked. Licensed psychotherapists, trained and calibrated, performed CTRS and dialogue-quality ratings with safeguards for rater well-being. To address fairness, we report per-group metrics across 16 trait-defined groups and use group-balanced sampling; nonetheless, secondary use should be validated for materially different populations. Given data sensitivity, raw dialogues are not publicly released; controlled access may be granted under a data-use agreement prohibiting re-identification/redistribution and requiring breach reporting and data destruction after project completion.

MindData Access and Usage License. After anonymous peer review, **MindData** will be available to bona fide research applicants under a no-fee, non-commercial license with strict ethics/privacy safeguards. Permitted uses are research and education only; clinical/diagnostic/therapeutic use or deployment is prohibited, and any re-identification or linkage attempts are forbidden (incidents must be reported within 72 hours). Applicants must provide an IRB/ethics determination, ensure secure access-controlled storage, and may not redistribute raw data; derivative releases (code, weights, aggregates) are allowed only if non-identifying and with intended-use limits, with citation of MindData and model cards disclosing its use.

C MindApt: Personalized Psychological Therapy

C.1 Therapeutic Dialogue Tracking Paradigm

C.1.1 Detailed Therapeutic Strategies

In this study, we specifically prioritize Cognitive Behavioral Therapy (CBT) due to its structured, goal-oriented nature, which naturally synergizes with our proposed "dialogue state" framework. This alignment is particularly advantageous for computational modeling, as CBT emphasizes real-time cognitive restructuring and behavioral modification—processes that can be explicitly tracked as evolving states within a dynamic conversation. We acknowledge that the concept of a defined "dialogue state" may be less immediately applicable to unstructured modalities, such as psychodynamic therapy, which focuses more on unconscious processes and long-term relational dynamics. While our current framework exploits the structural clarity of CBT, we plan to explore its adaptation to a broader range of therapeutic modalities in future work. We present the therapeutic strategies MindApt employs during personalized psychological therapy. Grounded in cognitive-behavioral therapy (CBT) and targeted to specific depressive symptoms, these strategies are tailored to each patient's current emotional and cognitive state. They adapt dynamically as the conversation unfolds, informed by real-time dialogue analysis and ongoing psychological state tracking, so that the treatment plan remains individualized throughout. For complete details, see Tables 12, 13, 14, 15, 16, and 17.

C.1.2 Dialogue Tracking Details

CBT Stage detection Cognitive Behavioral Therapy (CBT) typically unfolds in several structured stages. The initial stage focuses on building rapport, collecting personal history, and identifying primary symptoms and problematic thought patterns. In the cognitive restructuring stage, clients learn to recognize automatic negative thoughts, challenge distorted beliefs, and replace them with more balanced and adaptive perspectives. The behavioral activation stage emphasizes encouraging clients to engage in meaningful and rewarding activities, thereby reinforcing positive coping strategies and improving mood. A second evaluation stage follows, in which progress is reviewed against treatment goals, using both self-report and standardized measures. Finally, the relapse prevention stage aims to consolidate skills, strengthen coping strategies, and ensure clients can maintain improvements and manage potential setbacks independently.

Personal Issue Scoring Rules This section outlines scoring rules used in MindApt to assess coping strategies and personal issues based on patient dialogue. The scoring is based on frequency and relevance of specific keywords and context, with the goal of identifying the coping style and personal issues that are most relevant to the patient's therapeutic progress. These scores are integrated into the persona to inform therapy adaptation and guide treatment strategies.

Coping Style Scoring Rules Coping strategies are assessed using the Adolescent Coping Scale Second Edition (ACS-2), which differentiates between productive and non-productive coping styles. A higher Non-Productive Coping score indicates a tendency to use avoidance or ineffective coping strategies, while a higher Productive Coping score reflects the use of more adaptive and problem-solving approaches. The table below outlines the coping categories and scoring criteria.

Emotion State Tracking Rules **Negative Emotion Tracking** We monitor negative emotions across categories such as severe depression, severe anxiety, and intense anger. A frequency scoring mechanism adds 1 point per detected identifier. Let k denote the cumulative score. If $k \geq 4$ within the most recent 5 conversation turns, the system identifies the current dialogue stage and triggers a strategy shift to alternative therapeutic approaches tailored for emotional processing.
Stress Response Tracking We track 10 categories of stress responses, including stereotyped language, catastrophizing, self-doubt, aggressiveness, and avoidance. Each detected identifier adds 1 point to the stress score x. Within a 3-turn window, a score of $3 \leq x \leq 5$ triggers a strategy shift toward stabilization techniques for stress management. If $x \geq 5$ (indicating severe stress), the system immediately escalates the session to a human therapist.
Crisis Sign Detection The system continuously monitors for semantic identifiers related to self-harm or suicide. Upon detection, a 5-question crisis assessment is immediately initiated. Questions are scored on a 0-2 scale (with the final question reverse-scored). A cumulative score of 0-4 triggers a self-service help prompt while maintaining continuous observation, whereas a score of ≥ 5 results

in an immediate escalation to a human therapist and an alert to the user's guardian.

Scoring Examples Here are some examples for scoring.

Coping Style Example:

In a conversation, the following phrases were identified:

"I feel like I can't solve this problem" (Non-Productive Coping: Worry, Score: 2)

"I just keep to myself and avoid talking to anyone" (Non-Productive Coping: Keep to Self, Score: 2)

"I tried to talk to my friend, but it didn't help" (Productive Coping: Social Support, Score: 2)

Total Score for Non-Productive Coping: 4 Total Score for Productive Coping: 2

Since the Non-Productive Coping score is higher than the Productive Coping score, the system updates the persona to reflect a higher reliance on non-productive coping strategies. This triggers a shift in the treatment strategy to focus on helping the patient adopt more productive coping mechanisms, such as seeking social support and problem-solving.

Personal Issue Example:

In the conversation, the following issues were identified:

"I have been feeling sad and hopeless for a long time."(Emotional and Affective Issues: Depression, Score: 1)

"I always feel anxious about my future" (Emotional and Affective Issues: Anxiety, Score: 1)

"I don't think I am good enough compared to others."(Identity and Self-Esteem Issues: Self-Esteem, Score: 1)

Total Score for Emotional and Affective Issues: 2

Total Score for Identity and Self-Esteem Issues: 1

Once the total score for an issue reaches 10 or more, it is considered a prominent issue and is added to the persona. The persona would then include Emotional and Affective Issues and Identity and Self-Esteem Issues, prompting treatment strategies that focus on alleviating depression, anxiety, and building self-esteem.

Negative Emotion Identifiers

1. **Severe Depression:** Meaningless, don't know what to do, bored, helpless, exhausted, loss of interest, worthless.
2. **Severe Anxiety:** Nervous, worried, rapid heartbeat, inability to concentrate, inability to relax, shortness of breath.
3. **Intense Anger:** Angry, out of control, furious, dissatisfied.
4. **Intense Panic:** Scared, reject, rapid heartbeat, shortness of breath, suffocation.

Stress Response Identifiers and Examples

1. **High-frequency repetitive language (Stereotyped language)**
   Example: "I'm so stupid, really..."
2. **Negative or catastrophizing expressions**
   Example: "I've always had bad luck, nothing ever goes smoothly."
   Example: "Since this submission got rejected, I have no school to attend, my life is ruined."
   Example: "If I lose my girlfriend, the rest of my life is hopeless."
3. **Expressions of self-doubt and uncertainty**
   Example: "I don't know if I can get into such a good school... I probably won't be admitted."
   Example: "I feel like their praises are just because they don't really know me."
4. **Exaggerated or absolute statements**
   Example: "I always mess things up, I absolutely can't do this."
   Example: "My dad will never agree to let me do this." / "My boss will not sign off on this."
5. **Aggressive or defensive language**
   Example: "It's so annoying that you keep asking questions!" / "Why do you keep questioning me?" / "I don't want to talk about this."
6. **Frequent self-blame or guilt**
   Example: "It's all my fault. If it weren't for me, my friend wouldn't have gotten involved in this."
   Example: "It would be better if I wasn't in this family."
7. **Avoidance or withdrawal language**
   Example: "Can we stop talking about this?" / "This is boring, let's drop it," or providing obvious evasive and off-topic responses.
8. **Difficulty concentrating**
   Example: "Huh? What did you just say?", or similarly obvious evasive and off-topic responses.
9. **Expression of physical symptoms**
   Example: "My head has been hurting constantly since just now." / "I feel a bit short of breath." / "I'm overheating (losing my temper)."
10. **Repeatedly seeking reassurance**
    Example: "Are you sure I can recover?" / "Do

you really think I can get accepted?" / "Are you sure my girlfriend will forgive me?"

Crisis Sign Assessment Criteria

The crisis signal evaluation incorporates cultural adaptations based on the *Xu Kaiwen Suicide and Self-Harm Assessment Scale*. The assessment dimensions are as follows:

1. **Presence of Suicide or Self-Harm Plans:** Evaluates whether the individual has a specific plan for suicide or self-harm.
   - **1 point:** Vague suicidal ideation but fears death; no specific time, place, or method. *Example:* "I want to die, but I don't dare to."
   - **2 points:** Clear suicidal ideation with a relatively specific plan. *Example:* "If I fail the next exam, I'll go die." / "I have a detailed plan." / "I have prepared the rope/blades/pills."
2. **Past History of Suicide or Self-Harm:** Evaluates whether the individual has a history of self-harming behaviors.
   - *Example:* "I've cut my wrists before." / "I took a lot of pills to attempt suicide before." / "I am a suicide survivor."
3. **Real-Life Stress:** Evaluates the current external stressors the individual is facing.
   - *Example:* "I recently ran out of money." / "My parents are like my enemies." / "I just broke up, it's so painful." / "It feels like life is attacking me from all sides."
4. **Clinical Diagnosis:** Assesses the presence and severity of any existing clinical diagnoses.
5. **Support Resources (Reverse Scored):** Evaluates whether the individual currently has access to support networks, such as family and friends.
   - **2 points (No support resources):** *Example:* "I feel no one understands me." / "I have no one to rely on." / "I feel extremely lonely." / "I have no friends or family to talk to." / "No one cares about my feelings."
   - **1 point (Insufficient support) / 0 points (Sufficient support):** *Example:* "I have a few close friends." / "My partner is very understanding." / "My family and friends are all helping me."

C.2 Patient-aware Strategic Planning Module

Model and Action Space. The planning module π_θ is a RoBERTa encoder structured as a classifier over the action set $\mathcal{A} \subseteq \mathcal{A}_{\text{CBT}}$. The input is the concatenation of the dialogue history $D = (u_1^{\text{T}}, u_1^{\text{P}}, \ldots, u_t^{\text{T}}, u_t^{\text{P}})$ and the tracked patient state $\mathcal{M}$ (including symptoms, emotional states, and homework adherence). The model outputs a probability distribution over discrete strategy labels from a curated CBT library. During inference, the selected label is mapped to a corresponding natural-language instruction used by the therapist agent to condition its response.

Data Preprocessing.

- **Corpus.** We utilize a two-source training set: (1) Real multi-turn dialogues from MindData (830 patients, 2,573 sessions, 63,348 turns); and (2) Synthetic role-play dialogues generated by LLM-based patient simulators instantiated across 16 personal trait groups (*personal issues* $\times$ *coping styles*). The role-play outputs were reviewed by licensed psychotherapists regarding CTRS and four auxiliary dimensions (Empathy, Coherence, Helpfulness, Safety). Reviewers also annotated a safety gate (pass/fail). All data were de-identified.
- **Tokenization.** We use standard RoBERTa BPE tokenization with a maximum sequence length of 1024. We truncate the oldest turns if the limit is exceeded, preserving the latest k turns (default k=8).
- **Population Groups.** The cohort is stratified into 16 groups formed by crossing *personal issues* (8 types) and *coping styles* (2 types), with additional dynamic labels for emotional states and symptoms.

C.2.1 Offline Training Details (SFT)

Policy Parametrization. We model strategy selection as an $|\mathcal{A}|$-way classification task. RoBERTa encodes the input $x = (D, \mathcal{M})$, and a linear head projects the representation to logits $\mathbf{z}(x) \in \mathbb{R}^{|\mathcal{A}|}$. The policy is defined as $\pi_\theta(a \mid x) = \mathrm{softmax}(\mathbf{z}(x))_a$.

Data Splits and Leakage Control. MindData is split by patient ID into train/dev/test sets (8:1:1) to ensure strictly no patient overlap. Synthetic dialogues are used for *training only* to enhance diversity. All reported metrics are macro-averaged over the 16 trait-defined patient groups to ensure fairness evaluations.

Optimization. Given an expert action $a^\star$, we minimize the negative log-likelihood: $\mathcal{L}_{\text{SFT}} = -\log \pi_\theta(a^\star \mid x)$. We employ a macro-balanced sampling strategy ($p(k) = 1/16$). Training uses AdamW with a linear warmup (5%) and co-

sine decay. We apply differential learning rates: 2×10^{-5} for the encoder (layer-wise decay 0.95) and 5×10^{-5} for the head. Regularization includes dropout (0.1), weight decay (0.01), label smoothing (0.05), and gradient clipping (1.0). We use mixed precision (bf16) and early stopping based on dev macro-accuracy (patience=3).

Calibration. For deployment, we apply temperature scaling on the dev split to calibrate $\pi_\theta(a \mid x)$. These calibrated probabilities condition the safety gates and mixture-of-policies in the online stage.

C.2.2 Online Stage: Small-step RL with Patient Agents

To bridge the sim-to-real gap, we employ a population-based online reinforcement learning framework.

Hierarchical Architecture. The optimization target is the planning module π_θ. At turn t, π_θ selects a strategy a_t. A frozen **Therapeutic Model** (LLM-based generator $\mathcal{G}$) produces the response $u_t^{\mathrm{T}} = \mathcal{G}(D_t, a_t)$. A **Reward Model** evaluates u_t^{T} to compute r_t. This setup ensures the planner learns strategies that yield high-quality realizations.

Reward Formulation. The per-turn reward balances quality, safety, conciseness, and outcomes:

$$r_t \;=\; \alpha\,\tilde{q}_t \;+\; \beta\,s_t \;-\; \gamma\,\ell_t \;+\; \delta\,\mathrm{RTG}_t, \qquad (1)$$

where α=1.0, β=0.25, γ=0.15, δ=0.40. Components are standardized:

- **Quality ($\tilde{q}_t$):** CTRS-based score (aggregated with empathy/coherence), normalized against rater and cohort statistics: $\tilde{q}_t = \mathrm{clip}(\frac{q_t-\mu_{r,c}}{\sigma_{r,c}+\epsilon}, -2, 2)$.
- **Safety (s_t):** Binary indicator $s_t \in \{0, 1\}$ via hard gating; strictly penalizes safety violations.
- **Verbosity (ℓ_t):** Penalizes length: $\ell_t = \mathrm{clip}(\frac{\mathrm{tokens}_t}{600}, 0, 1)$.
- **Long-term Return (RTG_t):** Redistributes session-level outcomes (e.g., symptom reduction Δm) to salient turns with decay η: $g_t = \sum_{\tau=t}^{T} \eta^{\tau-t}\Delta m_\tau$. Normalized as $\mathrm{RTG}_t = \mathrm{clip}(\frac{g_t-\mu_{\mathrm{RTG}}}{\sigma_{\mathrm{RTG}}+\epsilon}, -1, 1)$.

Population-Aware Optimization. To ensure fairness across groups $\mathcal{K} = \{k_1, \ldots, k_{16}\}$:

- **Reward Normalization:** We mitigate scale differences using group-specific running statistics:

$$\begin{aligned} r'_t &= \frac{r_t - \mu_k}{\sigma_k + \epsilon}, \\ \mu_k &\leftarrow (1-\lambda)\mu_k + \lambda\bar{r}_k, \\ \sigma_k^2 &\leftarrow (1-\lambda)\sigma_k^2 + \lambda\overline{(r-\mu_k)^2}. \end{aligned} \qquad (2)$$

- **Adaptive Sampling:** We adjust sampling probability p_k to up-weight underperforming groups (where $\bar{m}$ is the global mean outcome and κ is a temperature hyperparameter):

$$p_k \propto \exp(\kappa(\bar{m} - m_k)), \quad p_k \leftarrow \frac{p_k}{\sum_j p_j}. \qquad (3)$$

Policy Optimization. We formulate the task as a sequential decision problem optimized via discrete Proximal Policy Optimization (PPO). The objective maximizes expected return while constraining deviation from the SFT reference π_{ref}:

$$\begin{aligned} \mathcal{L}_{\mathrm{PPO}} = \mathbb{E}_t\Big[&\min(\rho_t\hat{A}_t, \mathrm{clip}(\rho_t, 1\pm\epsilon)\hat{A}_t) \\ &- \beta_{\mathrm{KL}} D_{\mathrm{KL}}(\pi_\theta||\pi_{\mathrm{ref}}) + \eta_{\mathrm{ent}}\mathcal{H}(\pi_\theta)\Big], \end{aligned} \qquad (4)$$

where ρ_t is the importance sampling ratio and $\hat{A}_t$ is the generalized advantage. Note that gradients are applied solely to π_θ.

C.3 Therapeutic Agent (Qwen3–32B) Training Details

Supervised fine-tuning setup. We train a Qwen3–32B (no-thinking) model as the therapist agent via supervised fine-tuning (SFT). To preserve general instruction-following ability while adapting the model to therapeutic dialogue, we initialize training with a mixed corpus consisting of in-domain psychological dialogues and a small amount of general-domain instruction data. We apply standard autoregressive cross-entropy loss with strict loss masking on user inputs and system prompts, so that the model learns only the therapeutic response distribution.

Objective and data. For each dialogue, we retain *all* turns and segment the full sequence into contiguous 2k-token chunks from left to right, according to the model context limit. Each training instance is one such 2k chunk, which includes the tracked patient state $\mathcal{M}$ and the strategy a_t selected by the planning module. We apply autoregressive

supervision to the response tokens of *every* therapist turn appearing in the chunk, while masking out user turns and system prompts from the loss. In this way, the model is trained to generate each therapist response conditioned on the preceding multi-turn dialogue context, patient state, and planned strategy.

Sampling and splits. Training uses stratified sampling over the 16 trait-defined groups (personal issues $\times$ coping styles) with uniform group probability $p(k)=1/16$. MindData is split by patient ID into train/dev/test sets to prevent leakage. Role-play data are not used for therapist-agent evaluation.

Loss and optimization. The loss is masked to the therapist span only. We use cross-entropy with label smoothing 0.05, optimized with AdamW (learning rate 2×10^{-5}, weight decay 0.1), 5% warmup, and cosine decay. We apply gradient clipping at 1.0 and train in bf16 mixed precision.

Regularization and safety. We apply a mild response-length regularizer by penalizing outputs beyond the 95th percentile of target lengths. Unsafe training samples, identified by an internal safety classifier, are removed ($< 1\%$).

Decoding and selection. Default decoding uses nucleus sampling (p=0.9), temperature 0.7, a maximum response length of 2048 tokens, and a no-repeat n-gram size of 3. Model selection is based on dev-set dialogue-quality aggregates (CTRS combined with empathy, coherence, helpfulness, and safety) together with success rate; ties are broken by lower safety-violation rate.

Compute and platform. We perform SFT of the Qwen3–32B therapist model using *LLaMA-Factory* on a 2×8 NVIDIA A100 80GB cluster (16 GPUs across 2 nodes). Training employs bf16 mixed precision with DeepSpeed ZeRO-3, gradient checkpointing, and FlashAttention-2. All runs are executed on CUDA 12.4 with PyTorch 2.4.

C.4 Therapeutic Reward Model (Qwen3–14B) Training Details

Goal and outputs. The therapeutic reward model takes as input the dialogue context D, the tracked patient state $\mathcal{M}$, the current therapeutic action a_t, and a candidate therapist response, and produces a scalar per-turn dialogue-quality score $\tilde{q}_t$. This score reflects the overall therapeutic quality of the response by aggregating signals from CTRS, empathy, coherence, helpfulness, and safety. In addition, the model provides an auxiliary advisory signal for next-action recommendation.

Preference construction. Preference pairs are constructed from LLM role-play therapeutic dialogues generated with patient simulators instantiated from the same 16 trait-defined groups used in our training setup. For each turn, we sample two candidate therapist responses under the same $(D, \mathcal{M}, a_t)$ context and assign a binary preference label using an automatic aggregator that combines turn-level dialogue-quality signals (CTRS and the four quality dimensions) with session-level outcome proxies. To improve supervision quality, we discard low-margin pairs and filter out unsafe responses before training.

Training objective and setup. We train the reward model on preference pairs so that, under the same dialogue context, tracked patient state, and therapeutic action, it assigns a higher score to the preferred response than to the dispreferred one. Concretely, each training example consists of a shared context and a preferred/dispreferred response pair. The context window is 2k tokens, including the latest k=8 turns, $\mathcal{M}$, and a_t. For each preference pair, the two candidate responses are separately appended to the shared context, and each response is truncated to at most 512 tokens. Training uses AdamW with a learning rate of 1.5×10^{-5}, weight decay 0.1, 5% warmup, cosine decay, and gradient clipping at 1.0, under bf16 mixed precision.

Action recommendation. Beyond response scoring, the model is also used to provide an auxiliary estimate of the expected quality of candidate next-step actions. This advisory signal is used only as a tie-breaking cue when the planning module's top candidates fall within a small margin δ.

Compute and platform. We train the Qwen3–14B reward model using *LLaMA-Factory* on a 2×8 NVIDIA A100 80GB cluster (16 GPUs across 2 nodes). Training employs bf16 mixed precision with DeepSpeed ZeRO-3, gradient checkpointing, and FlashAttention-2. All runs are executed on CUDA 12.4 with PyTorch 2.4.

D Implementation of Real Patient Study

D.1 Design and Setting

We conducted a 16-week, therapist-delivered, parallel-group randomized controlled trial with two diagnostic cohorts: **moderate depression** and **moderate anxiety**. Within each cohort, participants were randomized (1:1:1) to one of three arms: **Standard** (therapist-determined strategies), **MindApt** (therapists followed MindApt-recommended strategies), or **Standard+MindApt** (therapists integrated clinical judgment with MindApt's recommendations). Target allocation was $n = 25$ per arm per cohort (total $N = 150$). Ten licensed psychotherapists delivered weekly sessions (about 50 minutes). The study was conducted at a single outpatient center under institutional IRB approval; all participants provided written informed consent.

D.2 Participants

Inclusion criteria: (1) age 18–65; (2) DSM-5 diagnosis of Major Depressive Disorder (moderate) *or* Generalized Anxiety Disorder/Panic Disorder/Social Anxiety Disorder (moderate), confirmed by structured interview; (3) baseline severity in the moderate range on BDI-II (depression cohort) or BAI (anxiety cohort); (4) willingness to attend weekly therapy for 16 weeks; (5) fluency in the treatment language.

Exclusion criteria: (1) acute psychosis, bipolar I, or active substance use disorder requiring higher level care; (2) acute suicide risk requiring crisis intervention; (3) current exposure-based PTSD treatment or concurrent intensive psychological therapy; (4) cognitive impairment that precludes CBT; (5) unstable medical conditions; (6) therapy with the assigned study therapist in the past 12 months.

D.3 Randomization, Allocation, and Masking

Participants were randomized with permuted blocks, stratified by cohort (depression vs. anxiety) and therapist to balance caseloads across arms. Allocation was concealed via an independent, centralized randomization service. Outcome assessors (data team) were blinded to arm assignment; therapists and participants were necessarily unblinded.

D.4 Interventions

Standard: CBT as usual; therapists selected strategies based on clinical judgment and case formulation.

MindApt: Therapists implemented MindApt's strategy recommendations aligned with dialogue state tracking; deviations required brief rationale.

Standard+MindApt: Therapists first reviewed MindApt recommendations, then integrated them with case-specific clinical judgment to form the final plan.

Across arms, session frequency was weekly; homework was assigned per CBT best practices. Concomitant medications were permitted if dose-stable ≥ 4 weeks before baseline; changes were recorded.

D.5 Therapists, Training, and Fidelity

Ten licensed psychotherapists (≥ 5 years of CBT experience) delivered treatment across arms. All therapists completed a 4-hour study onboarding (protocol, documentation, safety) and a 2-hour module on interpreting and integrating MindApt outputs. Psychotherapeutic interventions were coordinated with participants' pre-existing pharmacotherapy and care plans and were intended to complement—rather than conflict with or replace—them, consistent with a collaborative therapeutic alliance.

D.6 Outcomes and Assessment Schedule

Primary outcomes: change from baseline to Week-16 in *BDI-II* (depression cohort) and *BAI* (anxiety cohort).

Secondary outcomes: Week-by-week trajectories of BDI-II/BAI; session-level CTRS (process quality); patient-reported Global Improvement (PGI-I adapted).

D.7 Experimental Findings

As shown in Figure 5, all arms exhibited steady improvement on BDI-II and BAI over 16 weeks. *MindApt* tracked closely with *Standard* in both trajectory and Week 16 endpoints, supporting non-inferiority (i.e., comparable symptom reduction to therapist-determined care). **Standard+MindApt** produced the steepest early decline (Weeks 4–8) and the lowest Week 16 scores, suggesting complementary benefits when therapist judgment is combined with MindApt recommendation. Moreover, as shown in Tables 28 and 29, after 16 weeks of treatment with MindApt, the proportions of patients at *moderate* severity on BDI-II and BAI both fell to 0%. CTRS and PGI-I scores were also high, underscoring the high quality of the therapeutic dialogues.

D.8 Ethical Considerations.

All procedures were reviewed and approved by the institutional IRB, and every participant provided written informed consent prior to enrollment. The study was conducted in accordance with applicable ethical standards and IRB requirements, including privacy protection and responsible data handling.

D.9 Safety Statement

Our safety gate is designed to effectively monitor potential risk behaviors, specifically targeting suicidal tendencies and violent intent, by detecting high-risk lexical triggers. These keywords are critical for the real-time evaluation of content generated by the model. When any of these terms appear in a conversation, the system triggers a safety check and prevents the generation of unsafe content.

Risk Detection Mechanism:

- **Suicidal Tendencies:** The system flags keywords such as *suicide*, *end my life*, *self-harm*, *die*, and *no hope*.
- **Violent Behavior:** The system detects keywords such as *kill*, *violence*, *attack*, *weapon*, and *shoot*.

If any of these terms are detected, the system blocks the content and labels it with a safety flag. In high-risk cases, predefined escalation rules ensure that appropriate alerts are generated to prompt immediate intervention.

Human-in-the-Loop Protocol: As noted in the *Real-Patient Experiment* and *Social Impact* sections, the current AI therapy system is intended strictly as a clinical assistant. In the event of unsafe scenarios involving a patient simulator, the evaluation process is immediately halted. However, in real-world deployment, it is mandated that a human therapist intervene to provide necessary care and support, ensuring patient safety remains the paramount priority.

Dimension	**Agenda Setting**
Scoring Anchors (0–6)	0 No agenda; no goals or time allocation; session drifts. 1 Vague agenda; priorities unclear; not revisited. 2 Partial agenda; weak prioritization; limited follow-through. 3 Adequate collaborative agenda with 1–2 priorities; rough time plan. 4 Clear collaborative agenda; revisited to manage transitions. 5 Dynamic prioritization; integrates patient needs; tracks time effectively. 6 Masterful agenda management; responsive shifts; clear closure and homework link.
Dimension	**Feedback**
Scoring Anchors (0–6)	0 No feedback; ignores patient input; invalidating. 1 Infrequent or generic feedback; little accuracy check. 2 Some feedback but superficial; sporadic checks for understanding. 3 Adequate summaries and checks; incorporates reactions. 4 Regular accurate feedback; resolves misunderstandings. 5 Integrates feedback to refine interventions in-session. 6 Precise validating feedback loops that enhance insight and engagement.
Dimension	**Understanding**
Scoring Anchors (0–6)	0 Misunderstands concerns; contradicts patient experience. 1 Partial understanding with notable errors or mismatches. 2 Basic understanding; reflective listening uneven. 3 Adequate empathic understanding; accurate reflections. 4 Good formulation linking thoughts, feelings, behaviors. 5 Nuanced conceptualization tailored to current triggers and themes. 6 Sophisticated concise formulation that guides effective intervention.
Dimension	**Interpersonal Effectiveness**
Scoring Anchors (0–6)	0 Distant or hostile; poor boundaries; ruptures ignored. 1 Inconsistent warmth or respect; frequent misattunement. 2 Civil but limited warmth; occasional ruptures unaddressed. 3 Warm and respectful; repairs minor ruptures. 4 Trusting alliance; proactive rupture repair. 5 Strong alliance that facilitates difficult work. 6 Exceptional alliance; deft management of affect and boundaries.
Dimension	**Collaboration**
Scoring Anchors (0–6)	0 Directive or dismissive; no shared decisions. 1 Limited collaboration; token choices only. 2 Some collaboration; uneven eliciting of preferences. 3 Joint problem solving; uses “we” language. 4 Consistent shared decisions; invites alternatives. 5 High engagement; patient-generated options integrated. 6 Exemplar shared formulation and decision making.
Dimension	**Pacing / Time Use**
Scoring Anchors (0–6)	0 Disorganized; time wasted or crises unmanaged. 1 Frequent tangents; key work not reached. 2 Slow or rushed; limited time on priorities. 3 Adequate pacing; main tasks completed. 4 Efficient flow; transitions managed well. 5 Optimized time allocation to high-yield tasks. 6 Seamless pacing; maximal therapeutic yield per minute.

Table 6: CTRS scoring rubric.

Dimension	**Guided Discovery (Socratic)**
Scoring Anchors (0–6)	0 No guided discovery; tells rather than asks. 1 Rare questions; leading or judgmental tone. 2 Basic questioning; limited evidence examination. 3 Adequate Socratic style; explores alternatives. 4 Good hypothesis testing with collaborative inquiry. 5 Flexible discovery that elicits new perspectives. 6 Exquisite questioning that unlocks core insights.
Dimension	**Focus on Key Cognitions/Behaviors**
Scoring Anchors (0–6)	0 Off-target; misses hot thoughts or behaviors. 1 Vague targets; little linkage to distress. 2 Identifies targets but not prioritized. 3 Adequate focus on salient targets. 4 Clear focus on high-impact cognitions or behaviors. 5 Skillful narrowing to leverage points. 6 Laser focus that drives rapid change.
Dimension	**Strategy for Change**
Scoring Anchors (0–6)	0 No strategy or inappropriate plan. 1 Generic advice; poor fit to case. 2 Basic plan; weak rationale or feasibility. 3 Adequate rationale-linked strategy. 4 Well-matched, evidence-based plan. 5 Tailored, barrier-aware strategy with troubleshooting. 6 Elegant, adaptive strategy with clear mechanism of change.
Dimension	**Application of CBT Techniques**
Scoring Anchors (0–6)	0 Techniques absent or misapplied. 1 Rudimentary application; errors common. 2 Basic technique use; inconsistent structure. 3 Adequate delivery (e.g., behavioral activation, cognitive restructuring, exposure). 4 Competent structured delivery with feedback. 5 Flexible precise technique use responsive to data. 6 Masterful mechanism-aligned implementation.
Dimension	**Homework (Review and Setting)**
Scoring Anchors (0–6)	0 Not reviewed or set; no planning. 1 Mentioned briefly; nonspecific task. 2 Some review or setting; weak specificity or buy-in. 3 Adequate review; specific achievable assignment. 4 Good review with barriers addressed; clear plan. 5 Collaborative tailored homework with troubleshooting. 6 High-fidelity review; data-linked, patient-owned plan.

Table 7: CTRS scoring rubric.

Empathy	Description
1. Ordinary	The therapist acknowledges the client's statements but lacks deep understanding, primarily offering validation without further insight.
2. Good	The therapist demonstrates some understanding, but the examples used are still somewhat vague and lack strong emotional resonance.
3. Very Good	The therapist shares similar experiences, helping the client feel understood, which enhances the therapeutic connection.
4. Excellent	The therapist shows deep empathy, creating strong emotional resonance through common experiences or shared emotions, significantly strengthening the therapeutic bond.
5. Exemplary	The therapist excels in empathy, fully understanding and supporting the client with personal stories and deep emotional insights, fostering a profound therapeutic relationship.

Table 8: Empathy Scoring Scale

Coherence	Description
1. Ordinary	Weak emotional bond; communication is unclear and hard to follow; content has low relevance to the client's concerns; interaction opportunities are limited and the client struggles to engage.
2. Good	Some emotional connection, but inconsistent; communication is mostly clear with occasional confusion; content is somewhat relevant but lacks depth or direct applicability; interaction is average and the client sometimes cannot fully express themselves.
3. Very Good	Stable bond with genuine support; communication is clear and easy to follow; content aligns well with key concerns; interactions are frequent and the client has ample chances to share thoughts and feelings.
4. Excellent	Strong bond with evident trust; communication is very clear and readily understood; content is highly relevant and provides actionable insights; interaction is high, enabling full expression and active engagement.
5. Exemplary	Exceptionally strong alliance with consistent understanding and support; exceptionally clear and effectively advances problem resolution; content is fully relevant and precisely personalized; interaction is abundant and comfortable, promoting deep engagement.

Table 9: Coherence Scoring Scale

Safety	Description
1. Ordinary	The therapist creates a safe environment with limited trust-building, leading to minimal client expression and engagement.
2. Good	The therapist fosters some safety but with occasional barriers to full client openness. The client may still withhold certain emotions or thoughts.
3. Very Good	The therapist provides a largely safe environment, allowing the client to express emotions and engage in the therapeutic process with minimal hesitation.
4. Excellent	The therapist creates a strong sense of safety and trust, enabling the client to express difficult emotions freely and without fear of judgment.
5. Exemplary	The therapist excels in fostering an emotionally supportive and secure environment, encouraging complete openness and vulnerability from the client, which maximizes therapeutic engagement.

Table 10: Safety Scoring Scale

Helpfulness	Description
1. Ordinary	The therapeutic interaction is largely ineffective, with little emotional connection or relevance to the client's needs, leaving the client feeling unsupported.
2. Good	The therapeutic interaction is somewhat effective, offering some emotional connection and clarity, though relevance to the client's needs is inconsistent.
3. Very Good	The interaction is generally helpful, providing solid emotional support, clear communication, and relevant content that aligns with the client's psychological needs.
4. Excellent	The therapeutic interaction is highly effective, with excellent emotional connection, clear communication, and highly relevant content that meets the client's needs and promotes progress.
5. Exemplary	The interaction is exceptionally helpful, providing a seamless, emotionally supportive dialogue that is highly relevant, well-structured, and perfectly personalized to the client's psychological state.

Table 11: Helpfulness Scoring Scale

Strategy	Specifying
Explanation	Use open-ended questions to encourage clients to describe their feelings, thoughts, and experiences in more detail, making the issues and emotions more specific and clear.
Interview Case	Therapist: Can you tell me more about what happened that day? Patient: I always feel very bad. Therapist: Can you be specific about what makes you feel very bad? Has anything particular happened? Patient: I am dissatisfied with my body and always feel not healthy enough. Therapist: What is your definition of health? What aspects of your health do you feel can be improved? Patient: Lately, I have been feeling very emotionally unstable and often feel low. Therapist: What situations or events trigger this feeling of being low?
Strategy	Clarification
Explanation	Clarify the client's expressions to ensure correct understanding of their issues or feelings.
Interview Case	Therapist: When you say "anxiety," do you mean pressure from your PhD application or a more general feeling? Client: I've been really anxious lately and can't find interest in anything. Therapist: You mentioned feeling anxious and uninterested. When did this start and how does it affect your daily life?
Strategy	Summarization
Explanation	Summarize the main content expressed by the client to confirm understanding and promote deeper discussion.
Interview Case	Therapist: So, the recent stress mainly comes from balancing your main job and side job, as well as interpersonal relationships, right? Client: I feel worthless; life seems meaningless, and I'm often very depressed. Therapist: You've mentioned feeling worthless and finding life meaningless, which leads to depression. We've also discussed low motivation and interest. Are these what we've covered today?
Strategy	Reflecting Feelings
Explanation	Confirm and reflect the client's feelings through verbal or non-verbal means.
Interview Case	Therapist: You feel very tense, especially when talking to others. This tension makes you uncomfortable, right? Client: I've been feeling a lot of pressure at work, almost can't handle it. Therapist: It sounds like you're very tired and stressed right now.

Table 12: Therapeutic Strategy Library: Basic Strategies (Part 1/3).

Strategy	Empathy
Explanation	Build trust and connection by understanding and expressing empathy for the client's feelings.
Interview Case	Therapist: I can understand why you would feel so hurt and frustrated in that situation. When you feel your family doesn't understand your decisions, that must be really tough. Client: About five months ago, my best friend died in a car accident. Therapist: That's terrible—and a horrifying way to find out. How did you handle that information?
Strategy	Encouragement
Explanation	Encourage and support the client to enhance their confidence and motivation.
Interview Case	Therapist: You've done really well so far; keep this up and you can gradually change yourself. Client: I always feel I'm not doing well; others are better than me. Therapist: That thought is a first step to recognizing your abilities. What is something you did well recently? Client: I completed a complex task last week; my boss praised me. Therapist: That's a remarkable achievement! How did you do it?
Strategy	Exploring Options
Explanation	Encourage and help the client identify and explore different options and action plans.
Interview Case	Therapist: What other options do you think are available in this situation? Client: My relationship with my partner has been tense; we always argue over small things. Therapist: If your relationship could improve, what would it look like? Client: I hope we can communicate better and argue less. Therapist: What could help you communicate better?
Strategy	Exploring Resources
Explanation	Guide clients to list resources they can rely on, including strengths, coping mechanisms, relationships, and community support.
Interview Case	Therapist: Let's think about what can help you through this time. What personal strengths have helped you in the past? Client: I'm good at problem-solving and staying calm under pressure. Therapist: Those are excellent strengths. Who can you turn to for support? Client: My close friend and my sister. Therapist: Wonderful—reach out to them when needed.

Table 13: Therapeutic Strategy Library: Basic Strategies (Part 2/3).

Strategy	Contextual Techniques
Explanation	Assist the client in exploring and coping with issues across different contexts.
Interview Case	Therapist: How do you handle situations at work or school that make you feel frustrated? Client: I struggle to stay focused when I'm overwhelmed. Therapist: What strategies do you use to manage your focus at work?
Strategy	Immediacy
Explanation	Respond to immediate emotions or issues that arise during the counseling process.
Interview Case	Therapist: I noticed a change in your emotion when you mentioned that—can you talk about what happened? Client: I feel like a failure, always missing my goals. Therapist: How do you feel right now, talking about this? Client: Disappointed and frustrated. Therapist: Where do you feel it most strongly?
Strategy	Open-ended Questions
Explanation	Encourage users to provide detailed information about their lives and feelings.
Interview Case	Therapist: Can you tell me about the most stressful event you've experienced recently?

Table 14: Therapeutic Strategy Library: Basic Strategies (Part 3/3).

Strategy	Identifying Cognitive Distortions
Explanation	Help the client recognize irrational or distorted thought patterns using specific examples from daily life.
Interview Case	Therapist: It sounds like you often think, "I always fail." That's an example of all-or-nothing thinking. Client: I do say that a lot—it feels true. Therapist: Let's test how accurate it is by looking at specific examples.
Strategy	Attributional Analysis
Explanation	Examine how the client attributes causes to events (e.g., internal vs. external) and challenge self-blaming interpretations.
Interview Case	Therapist: You mentioned blaming yourself when things go wrong. What else could have contributed? Client: I didn't consider the workload or tight deadline. Therapist: Exactly—external factors may also be at play.
Strategy	Reframing
Explanation	Guide the client to view situations from a more balanced angle; offer alternative explanations for stressful events.
Interview Case	Therapist: Although it seems difficult, has it brought any benefits? Patient: I feel clumsy in social situations; everyone is laughing at me. Therapist: Is there evidence people are actually laughing? Patient: No one said I'm clumsy—it may be my perception. Therapist: Try redefining this as insecurity rather than fact.
Strategy	Psychoeducation
Explanation	Educate the client about symptoms and CBT techniques; provide clear explanations to increase engagement.
Interview Case	Therapist: Depression often includes automatic negative thoughts. We'll use cognitive restructuring to challenge them. Client: I didn't know thoughts could be changed. Therapist: Understanding this is the first step toward feeling better.
Strategy	Socratic Questioning
Explanation	Use open-ended questions to examine assumptions, consequences, and alternatives; deepen reflection on automatic thoughts.
Interview Case	(Excerpt) Beth: It seems you want to connect with your siblings, but thinking about it makes you feel bad. Did you think about this last week? Alicia: Mostly on Sunday. I thought I should call my sister, but Sundays are messy... Beth: Are there other days that would feel easier? Alicia: Friday evenings. Beth: Shall we plan a specific call next week?

Table 15: Therapeutic Strategy Library: CBT-style Strategies (Part 1/3).

Strategy	Cognitive Defusion
Explanation	Teach clients to detach from negative thoughts by noticing them without judgment (mindfulness).
Interview Case	Therapist: Instead of fighting thoughts, try noticing them: "I notice I'm having the thought that..." and let it pass like a leaf on a stream. How does that sound? Client: Difficult, but I'm willing to try.
Strategy	Cognitive Restructuring
Explanation	Identify and replace negative thoughts with more balanced ones; use thought records/worksheets.
Interview Case	Therapist: Let's write down the anxious thought, then find evidence for and against it. Client: I thought I'd embarrass myself at the meeting—but it didn't happen.
Strategy	Decatastrophizing
Explanation	Reduce worst-case thinking by evaluating realistic outcomes and probabilities.
Interview Case	Therapist: What's the worst thing that could realistically happen? How likely is it? Client: I might make a mistake, but probably no one will notice.
Strategy	Evidence Examination
Explanation	Review evidence supporting/contradicting negative beliefs; weigh both sides objectively.
Interview Case	Therapist: Why do you think you won't get into your preferred university? Patient: My grades weren't great. Therapist: What was your overall average? Patient: Mostly A's before the last semester; then two A's and two B's. Therapist: Given that average, why assume you can't get in? Have you checked admission stats?
Strategy	Challenging Core Beliefs
Explanation	Identify and dispute deeply held, unhelpful core beliefs using logical questioning and behavioral data.
Interview Case	Therapist: You believe you're unworthy of love. What makes you think that? Client: I've always been rejected. Therapist: Does one rejection mean you're unworthy in every case?

Table 16: Therapeutic Strategy Library: CBT-style Strategies (Part 2/3).

Strategy	Behavioral Experiments
Explanation	Test beliefs via real-world experiments; use outcomes to update cognitions.
Interview Case	Therapist: You think people won't talk to you if you join the group. Let's try it this week and observe what happens. Client: I'll attend the meeting and see how it goes.
Strategy	Behavioral Activation
Explanation	Engage in rewarding/valued activities; set graded tasks; track progress; troubleshoot barriers.
Interview Case	Patient: It was difficult to go buy stamps for last week's assignment. Therapist: The pain sounds like a real barrier. Do you still value writing to your children? Patient: I do, but I'm not sure I can do it. Therapist: Let's explore alternatives—could you get stamps online to reduce the physical burden?
Strategy	Graded Exposure
Explanation	Gradually face feared situations from easier to harder steps, monitoring anxiety reduction.
Interview Case	Therapist: We'll start with a brief practice in a low-anxiety setting, then increase difficulty. Client: I'll begin with a short walk in a quiet place, then try busier areas.
Strategy	Motivational Interviewing
Explanation	Enhance intrinsic motivation through engaging, focusing, evoking, and planning collaborative goals.
Interview Case	Therapist (Engage): How have things been lately? Client: Tough—down and unmotivated. Therapist (Focus): If you could change one thing, what would it be? Client: Manage stress better and exercise. Therapist (Evoke): What motivates you to do that? Client: I'll feel better and be there for my family. Therapist (Plan): Let's set goals—exercise three times this week and prep meals.
Strategy	Cognitive Rehearsal
Explanation	Mentally rehearse adaptive responses; use role-play/visualization to prepare for stressors.
Interview Case	Therapist: Let's practice how you'll respond if you feel anxious during tomorrow's meeting. Client: I'll visualize staying calm and speaking clearly.

Table 17: Therapeutic Strategy Library: CBT-style Strategies (Part 3/3).

Stage	Initial Stage
Stage Goals	• Collect medical history and basic information, and develop the user persona • Establish trust • Identify primary symptoms and irrational beliefs • Set treatment goals
Standards / Operational Definitions	**Standard:** Patient demonstrates trust, provides sufficient background, and initial treatment goals are established. **Operational Definitions:** • **Information Completeness:** $\geq$90% of medical history and background collected • **Trust Indicators:** In 5 sessions, client shows trust at least 3 times (e.g., proactively sharing feelings) • **Primary Symptoms Recognition:** At least 2 main symptoms and 1 irrational belief identified • **Treatment Goal Setting:** At least 1 actionable treatment goal established and agreed upon
Transition Markers	• Information collection completion • Trust established • Primary symptoms identified
Example Dialogue	Therapist: Can you tell me a bit about your background and how you've been feeling recently? Client: I often feel anxious at work, especially when deadlines approach. Therapist: Thank you for sharing that. It helps me understand what's troubling you. How does the anxiety affect your daily life? Client: Sometimes I can't sleep, and I avoid talking to my manager. Therapist: I see. It sounds like this has been hard for you. I want you to know this is a safe place to talk about these experiences. Client: That makes me feel more comfortable. Therapist: Great. From what you've told me, anxiety at work and sleep problems seem important issues. Would you agree these could be our focus? Client: Yes, I think so. Therapist: Perfect. We can set a goal to work on managing anxiety more effectively, while also improving your sleep. Does that sound good to you? Client: Yes, that would be helpful.

Table 18: Therapeutic goals and details: Initial Stage

Stage	Cognitive Restructuring
Stage Goals	1. Identify automatic thoughts 2. Challenge and question automatic thoughts 3. Develop balanced, dialectical thinking
Standards / Operational Definitions	**Standard:** Patient applies dialectical thinking in $\geq$3 different situations. **Operational Definitions:** • **Specific Situations:** Demonstrates dialectical thinking in stress, conflict, or self-evaluation contexts. • **Dialectical Dialogue:** Uses phrases like "Although... I can also..." to express balanced perspectives. • **Real-life Examples:** Provides concrete daily-life cases of applying restructured thoughts. • **Thought Records:** Completes at least three thought records documenting identification, challenge, and restructuring of thoughts.
Transition Markers	Within 20 conversations, frequency of cognitive distortions decreases by at least three instances.
Example Dialogue	Therapist: Can you share a thought you had when something stressful happened recently? Client: I thought, "I'm not good enough to handle this project." Therapist: Let's examine that thought. What evidence supports it, and what evidence goes against it? Client: Well, I do have experience managing smaller projects successfully. Therapist: That's important evidence. How else might you reframe this situation? Client: "Although this project is challenging, I've succeeded before and can apply those skills again." Therapist: Excellent. That's a balanced way of thinking. Could you try writing this down in your thought record and applying it next time? Client: Yes, I'll try that. Therapist: Great. Let's also plan one small step you could take to build confidence this week.
Tailored Strategies	**Tailored-Coping (non-coping > coping)** Identifying Cognitive Distortions; Attributional Analysis; Reframing; Psychoeducation; Socratic Questioning; Cognitive Defusion; Cognitive Restructuring; Decatastrophizing; Evidence Examination; Challenging Core Beliefs; Coping Education; Acceptance Practice; Emotion Labeling; **Tailored-Issues** (Single-Issues >=10) Identifying Cognitive Distortions; Attributional Analysis; Reframing; Psychoeducation; Socratic Questioning; Cognitive Defusion; Cognitive Restructuring; Decatastrophizing; Evidence Examination; Challenging Core Beliefs; Experiential Exercises; Acceptance Practice; Exploring Options; Exploring Resources; Exploring Meaning; Validating; Emotion Labeling; **Tailored-Emotions** (Negative Emotions) Experiential Exercises; Relaxation Training; Acceptance Practice; Cognitive Defusion; Psychoeducation; Emotion Labeling; Coping Education.

Table 19: Therapeutic goals and details: Cognitive Restructuring

Stage	Behavioral Activation
Stage Goals	1. Confirm dialectical thinking 2. Develop dialectical cognition 3. Encourage engagement in meaningful activities
Standards / Operational Definitions	**Standard:** Patient independently uses dialectical thinking and engages in positive activities in $\geq$3 different sessions or situations. **Operational Definitions:** • **Independent Application:** Demonstrates new cognitive strategies without therapist prompts. • **Behavioral Engagement:** Reports active coping in real-life situations (e.g., conflict resolution, task completion, social participation). • **Activity Scheduling:** Creates a schedule with $\geq$3 concrete activities and completes at least 2 within one week. • **Mood Tracking:** Describes improvements in mood or confidence after behavioral activation.
Transition Markers	Evaluation-specific model (BDI & BAI criteria) confirms readiness for discharge.
Example Dialogue	Therapist: How have you been applying the strategies we discussed? Client: Last week I felt overwhelmed with a deadline, but instead of panicking, I broke the task into smaller steps. Therapist: That's excellent. How did that help? Client: It made me feel more in control, and I actually finished earlier. Therapist: Great progress. Did you try any planned activities outside of work? Client: Yes, I joined a yoga class. At first I hesitated, but I reminded myself it's okay to feel nervous and that I can still give it a try. Therapist: That's a good example of combining new thoughts with new behavior. How did you feel afterwards? Client: I felt proud and more relaxed. It gave me confidence to try again next week. Therapist: Perfect. That shows you can independently apply these skills across different situations.
Tailored Strategies	**Tailored-Coping (non-coping > coping)** Behavioral Activation; Graded Exposure; Motivational Interviewing; Cognitive Rehearsal; Psychoeducation; Coping Education; Acceptance Practice; Exploring Options; Exploring Resources; Exploring Meaning; Emotion Labeling; **Tailored-Issues** (Single-Issues >=10) Behavioral Activation; Graded Exposure; Motivational Interviewing; Cognitive Rehearsal; Psychoeducation; Experiential Exercises; Acceptance Practice; Exploring Options; Exploring Resources; Exploring Meaning; Validating; Emotion Labeling; **Tailored-Emotions** (Negative Emotions) Experiential Exercises; Relaxation Training; Acceptance Practice; Cognitive Defusion; Psychoeducation; Emotion Labeling; Coping Education.

Table 20: Therapeutic goals and details: Behavioral Activation

Stage	Second Evaluation
Stage Goals	1. Confirm recovery status through independent interview 2. Assess with Initial Stage indicators: if symptoms remain, return to N1; else progress to N4
Standards / Operational Definitions	**Standard:** Patient demonstrates symptom remission and functional recovery in independent interview. **Operational Definitions:** • **Symptom Check:** BDI & BAI scores reduced below clinical thresholds or $\geq$50% improvement from baseline. • **Functional Assessment:** Reports clear improvement in daily functioning (e.g., work, relationships). • **Consistency Check:** Agreement between patient self-report and therapist assessment $\geq$80% (e.g., via Kappa). • **Relapse Risk Identification:** If $\geq$2 major symptoms or $\geq$1 strong irrational belief persist, return to Stage N1.
Transition Markers	Diagnostic indicators and interview outcomes confirm either recovery (progress to N4) or relapse (return to N1).
Example Dialogue	Therapist: Over the past few weeks, how have you been feeling compared to when we first started? Client: Much better. My anxiety is less frequent, and I'm sleeping better. Therapist: That's encouraging. On a scale from 0 to 10, how severe is your anxiety now compared to the beginning? Client: It was around 8 before, but now it's about 3. Therapist: Excellent. Let's also check the BDI and BAI scores. They've both decreased significantly. Client: Yes, I feel more capable of handling daily stress. Therapist: Based on your progress, I think you're ready to move forward. Do you feel the same? Client: Yes, I feel I can manage on my own. Therapist: Perfect. That means you're ready for Stage N4.
Tailored Strategies	Specifying; Clarification; Summarization; Reflecting Feelings; Empathy; Encouragement; Exploring Options; Exploring Resources; Contextual Techniques; Immediacy; Open-ended Questions; Goal setting; Emotion Labeling; Identifying Cognitive Distortions.

Table 21: Therapeutic goals and details: Second Evaluation

Stage	Relapse Prevention
Stage Goals	• Prevent relapse with emotional support and positive attention • Monitor irrational beliefs and symptoms continuously • Strengthen coping strategies and support network
Standards / Operational Definitions	**Standard:** Patient maintains functional improvements and consistently applies coping strategies when exposed to stressors. **Operational Definitions:** • **Symptom Monitoring:** Completes weekly self-assessments (e.g., BDI/BAI or mood logs). • **Coping Skills Usage:** Reports using cognitive restructuring, behavioral activation, or relaxation in $\geq$3 scenarios. • **Early Warning Recognition:** Identifies at least 2 personal early relapse signals (e.g., insomnia, avoidance, negative self-talk) and applies coping responses. • **Support Network Activation:** Lists and actively engages $\geq$2 social support resources (family, peers, community).
Transition Markers	• Stable coping and no relapse for 4 consecutive weeks $\rightarrow$ enter maintenance phase. • Symptom score increases by $\geq$50% or early relapse signs reappear frequently $\rightarrow$ return to prior stage.
Example Dialogue	Therapist: Over the past month, have you noticed any signs of old symptoms coming back? Client: Sometimes I feel a bit anxious before big meetings, but I've been using breathing exercises. Therapist: That's a great way to apply your coping skills. Did it help? Client: Yes, it calmed me down and I was able to focus. Therapist: Excellent. What about your early warning signs—have you recognized any lately? Client: I noticed I've been staying up late, which usually makes me more stressed. Therapist: Good awareness. What steps did you take? Client: I adjusted my bedtime routine and asked a friend to remind me to log off early. Therapist: That's an effective use of your support network. Keep tracking these behaviors weekly.
Tailored Strategies	Specifying; Clarification; Summarization; Reflecting Feelings; Empathy; Encouragement; Exploring Options; Exploring Resources; Contextual Techniques; Immediacy; Exploring Meaning; Validating; Open-ended Questions; Emotion Labeling; Acceptance Practice; Relaxation Training; Psychoeducation.

Table 22: Therapeutic goals and details: Relapse Prevention

Coping Style: Productive Coping
Description: Encompasses problem-solving strategies aimed at actively addressing stressors while maintaining physical activity and social connectedness. This style reflects adaptive coping behaviors that promote resilience and positive adjustment.
Social Support Description: Seeking help and comfort from others. Keywords: "talked to" (2), "shared with" (2), "discussed with" (2).
Focus on Solving the Problem Description: Reflecting, planning, and tackling issues systematically. Keywords: "plan to" (2), "figured out" (2), "resolved by" (2).
Physical Recreation Description: Using sports and exercise to relieve stress. Keywords: "worked out" (2), "exercise" (2).
Seek Relaxing Diversions Description: Engaging in leisure activities to relax. Keywords: "read a book" (2), "watched a movie" (2).
Invest in Close Friends Description: Building intimate and supportive friendships. Keywords: "spent time with" (2), "confided in" (2).
Work Hard and Achieve Description: Pursuing goals with persistence and ambition. Keywords: "worked hard" (2), "achieved" (2).
Focus on the Positive Description: Maintaining optimism and looking for positives. Keywords: "stayed positive" (2), "bright side" (2).
Accept One's Best Efforts Description: Acknowledging limits and accepting outcomes. Keywords: "did my best" (2), "accepted" (2).
Social Action Description: Taking initiative and organizing to address concerns. Keywords: "organized" (2), "raised awareness" (2).
Seek Professional Help Description: Consulting counselors or professionals. Keywords: "talked to a counselor" (2), "sought advice" (2).
Humour Description: Using humor as a coping tool. Keywords: "made a joke" (1), "laughed about" (1).
Seek Spiritual Support Description: Turning to spirituality or religion. Keywords: "prayed" (2), "asked God" (2).

Table 23: Coping Style Scoring Criteria: Productive Coping (based on ACS-2).

Coping Style: Non-Productive Coping
Description: Indicates reliance on avoidance strategies, wishful thinking, or maladaptive behaviors. These strategies are generally associated with increased distress, impaired adjustment, and reduced ability to manage stress effectively.
Worry Description: Persistent concern about the future or potential problems. Keywords: "worried about" (2), "concerned with" (2).
Wishful Thinking Description: Hoping for unrealistic positive outcomes. Keywords: "hoped for" (1), "wished that" (1).
Not Coping Description: Feeling overwhelmed or showing psychosomatic symptoms. Keywords: "overwhelmed" (2), "broke down" (2).
Tension Reduction Description: Reducing stress through release of tension. Keywords: "vented" (1), "let off steam" (1).
Ignore the Problem Description: Avoiding or blocking out the issue. Keywords: "ignored" (2), "avoided" (2).
Keep to Self Description: Withdrawing and not sharing with others. Keywords: "kept to myself" (2), "isolated" (2).
Self-Blame Description: Criticizing oneself excessively. Keywords: "blamed myself" (2), "felt guilty" (2).
Act Up Description: Acting out destructively to release distress. Keywords: "acted out" (2), "threw a fit" (2).

Table 24: Coping Style Scoring Criteria: Non-Productive Coping (based on ACS-2).

Coping Style Scoring Method (based on ACS-2)

Scoring Method: The original ACS-2 uses adjusted scores based on mean ratings for each coping strategy. Scores range from "Never" (20–30) to "Very Often" (90–100), with intermediate categories: "Rarely" (31–49), "Sometimes" (51–70), and "Often" (71–89). These scores represent how frequently the adolescent employs each coping strategy and how helpful they perceive the strategy to be in managing stress or concern. Both usage and perceived helpfulness are assessed.

Interpretation:
– Higher **Productive Coping Score** → more frequent and effective use of adaptive strategies.
– Higher **Non-Productive Coping Score** → more frequent use of avoidance or maladaptive strategies.

Example Weights:
– Strong Indicators: Weight = 2
– Moderate Indicators: Weight = 1

Example Keywords and Weights:
Productive Coping:
– Social Support: "talked to" (2), "shared with" (2), "discussed with" (2)
– Focus on Solving the Problem: "plan to" (2), "figured out" (2), "resolved by" (2)
– Physical Recreation: "worked out" (2), "exercise" (2)
– Seek Relaxing Diversions: "read a book" (2), "watched a movie" (2)
– Invest in Close Friends: "spent time with" (2), "confided in" (2)
– Work Hard and Achieve: "worked hard" (2), "achieved" (2)
– Focus on the Positive: "stayed positive" (2), "bright side" (2)
– Accept One's Best Efforts: "did my best" (2), "accepted" (2)
– Social Action: "organized" (2), "raised awareness" (2)
– Seek Professional Help: "talked to a counselor" (2), "sought advice" (2)

Non-Productive Coping:
– Worry: "worried about" (2), "concerned with" (2)
– Wishful Thinking: "hoped for" (1), "wished that" (1)
– Not Coping: "overwhelmed" (2), "broke down" (2)
– Tension Reduction: "vented" (1), "let off steam" (1)
– Ignore the Problem: "ignored" (2), "avoided" (2)
– Keep to Self: "kept to myself" (2), "isolated" (2)
– Self-Blame: "blamed myself" (2), "felt guilty" (2)
– Act Up: "acted out" (2), "threw a fit" (2)

Example of Scoring:
Sample text:
"In the girl's room, there was a tree with red branches and leaves, but she could not express it to others. She didn't even know why such a tree suddenly appeared. She [asked her classmates, but they insisted she was lying] (Ignore the Problem, Score: 2). She [asked her teacher, who told her to express it, but she couldn't] (Seek Professional Help, Score: 2). Watching the tree slowly grow, the girl felt a sense of ['letting go'] (Accept One's Best Efforts, Score: 2). Because of her parents' divorce and the heavy burden of schoolwork, she [locked herself in a dark room] (Keep to Self, Score: 2), accompanied only by coffee and books. [The tree pierced into her heart like a compass needle] (Seek Spiritual Support, Score: 1), bringing her some peace. Looking back, she realized that her teacher had planted this seed at the time of her parents' divorce, [encouraging her to open the door of her room, face her inner unease, and live each future day happily] (Focus on the Positive, Score: 2). The tree stood there, red and carefree."

– **Productive Coping:**
Seek Professional Help (2), Focus on the Positive (2), Accept One's Best Efforts (2) → Total = 6
– **Non-Productive Coping:**
Ignore the Problem (2), Keep to Self (2) → Total = 4

Table 25: Coping Style Scoring Method, Interpretation, and Example (based on ACS-2).

Personal Issues Scoring Criteria
Category: Emotional and Affective Issues **Description:** Issues related to emotional states such as depression, anxiety, and anger. **Subcategories & Keywords:** – **Depression:** depressed, sadness, hopeless, worthlessness, fatigue, insomnia, self-harm, suicide, tearfulness. – **Anxiety:** anxious, panic, worry, nervous, fear, trembling, sweating, avoidance, obsessive thoughts. – **Anger:** anger, irritability, rage, outburst, frustration, resentment. – **Other Emotions:** guilt, shame, loneliness, emptiness, emotional numbness.
Category: Interpersonal Relationship Issues **Description:** Issues related to relationships with family, friends, and romantic partners. **Subcategories & Keywords:** – **Couples/Marital:** marriage, spouse, infidelity, divorce, separation, communication, intimacy, conflict, trust. – **Family Dynamics:** parent-child, sibling rivalry, family conflict, roles, abuse, neglect. – **Friendships:** friendship, peer pressure, betrayal, social isolation, communication issues.
Category: Identity and Self-Esteem Issues **Description:** Concerns about self-worth and identity crises. **Subcategories & Keywords:** – **Self-Esteem:** self-worth, inadequacy, self-criticism, lack of confidence, inferiority. – **Identity Crisis:** identity, self-discovery, confusion, purpose, meaning, existential crisis. – **Body Image:** dissatisfaction, appearance, eating disorders, self-image.
Category: Stress and Adjustment Issues **Description:** Issues related to coping with stress, academic pressure, and life transitions. **Subcategories & Keywords:** – **Work-Related Stress:** burnout, job dissatisfaction, workload, deadlines, conflict, work-life balance. – **Academic Stress:** exams, grades, performance pressure, study load, academic failure. – **Life Transitions:** moving, job change, retirement, parenthood, divorce, major life events.
Category: Behavioral and Addiction Issues **Description:** Issues related to behavior control and addictions. **Subcategories & Keywords:** – **Substance Abuse:** alcohol, drugs, addiction, dependency, withdrawal, rehabilitation. – **Gambling:** gambling addiction, betting, financial loss, impulsivity. – **Internet and Technology:** internet addiction, gaming addiction, social media, screen time. – **Impulsive Behavior:** compulsive behavior, risk-taking, self-harm, impulsivity.
Category: Trauma and Abuse Issues **Description:** Issues related to past traumatic experiences. **Subcategories & Keywords:** – **Sexual Abuse:** assault, molestation, rape, trauma, survivor, PTSD. – **Physical Abuse:** domestic violence, child abuse, elder abuse, safety, harm. – **Emotional Abuse:** manipulation, control, verbal abuse, trauma.
Category: Health and Physiological Issues **Description:** Issues related to health conditions and their psychological impact. **Subcategories & Keywords:** – **Chronic Illness:** chronic pain, diabetes, cancer, diagnosis, treatment. – **Disability:** physical disability, mental disability, accessibility, support. – **Sleep Disorders:** insomnia, sleep apnea, sleep deprivation.
Category: Cultural and Social Issues **Description:** Issues related to discrimination and cultural stress. **Subcategories & Keywords:** – **Discrimination:** sexism, ageism, homophobia, racism, inequality. – **LGBTQ+ Issues:** sexual identity, gender identity, LGBTQ+ discrimination.

Table 26: Personal Issues Scoring Criteria (multi-line format).

PGI-I Score	Description
1. Very much improved	Patient reports major improvement with symptoms essentially resolved and clear gains in daily functioning.
2. Much improved	Patient reports clear, clinically meaningful improvement with only minor residual symptoms.
3. Minimally improved	Patient reports small improvement compared with baseline, some benefit but limited in scope.
4. No change	Patient reports condition unchanged relative to baseline.
5. Minimally worse	Patient reports slight worsening with a minor increase in symptom burden or impact.
6. Much worse	Patient reports clear and clinically meaningful deterioration.
7. Very much worse	Patient reports marked deterioration with substantial impact on functioning and possible need for reassessment.

Table 27: PGI-I Scoring Scale

Time	Group	BDI-II Score	95% CI	BDI-II Moderate Depression Rate	CTRS Score	PCI-I Score
Week 0	Standard Care	24.2 ± 2.6	$[23.28, 25.12]$	100%	–	–
	MindApt	24.1 ± 2.1	$[23.34, 24.86]$	100%	–	–
	Standard + MindApt	24.4 ± 2.5	$[23.53, 25.27]$	100%	–	–
Week 4	Standard Care	22.3 ± 3.3	$[21.35, 23.25]$	80%	8.1	3.1
	MindApt	21.7 ± 2.8	$[20.86, 22.54]$	84%	8.0	3.1
	Standard + MindApt	20.8 ± 2.8	$[19.96, 21.64]$	68%	8.4	2.8
Week 8	Standard Care	17.9 ± 4.8	$[16.91, 18.89]$	48%	7.9	2.5
	MindApt	18.5 ± 5.2	$[17.35, 19.65]$	56%	8.2	2.6
	Standard + MindApt	15.8 ± 5.9	$[14.69, 16.94]$	24%	8.5	2.1
Week 12	Standard Care	13.2 ± 4.2	$[12.12, 14.28]$	12%	8.2	1.9
	MindApt	14.1 ± 3.6	$[13.30, 14.90]$	12%	8.1	1.8
	Standard + MindApt	11.3 ± 3.4	$[10.43, 12.17]$	4%	8.5	1.6
Week 16	Standard Care	11.8 ± 1.8	$[11.13, 12.25]$	0%	8.2	1.5
	MindApt	12.4 ± 2.8	$[11.61, 13.19]$	0%	8.2	1.5
	Standard + MindApt	9.8 ± 2.2	$[8.67, 10.63]$	0%	8.6	1.3

Table 28: BDI-II score for real patient study

Time	Group	BAI Score	95% CI	BAI Moderate Anxiety Rate	CTRS Score	PGI-I Score
Week 0	Standard Care	21.4 ± 2.4	$[20.46, 22.34]$	100%	–	–
	MindApt	21.3 ± 2.4	$[20.36, 22.24]$	100%	–	–
	Standard + MindApt	21.6 ± 2.6	$[20.58, 22.62]$	100%	–	–
Week 4	Standard Care	20.3 ± 3.5	$[18.93, 21.67]$	92%	8.1	3.3
	MindApt	19.7 ± 3.4	$[18.37, 21.03]$	88%	8.2	3.2
	Standard + MindApt	18.1 ± 2.7	$[17.10, 18.94]$	72%	8.3	2.7
Week 8	Standard Care	13.2 ± 3.5	$[12.31, 14.09]$	28%	8.2	2.4
	MindApt	14.5 ± 3.2	$[13.25, 15.75]$	36%	8.2	2.5
	Standard + MindApt	10.8 ± 2.7	$[9.74, 11.86]$	12%	8.4	2.0
Week 12	Standard Care	8.7 ± 3.2	$[7.45, 9.95]$	4%	8.2	1.5
	MindApt	9.3 ± 3.5	$[7.93, 10.67]$	8%	8.2	1.7
	Standard + MindApt	6.9 ± 1.9	$[6.16, 7.64]$	0%	8.5	1.3
Week 16	Standard Care	6.5 ± 1.8	$[5.79, 7.21]$	0%	8.3	1.3
	MindApt	6.9 ± 1.7	$[6.23, 7.57]$	0%	8.3	1.5
	Standard + MindApt	5.5 ± 1.2	$[5.03, 5.97]$	0%	8.5	1.2

Table 29: BAI score for real patient study

E Prompt Design

LLM Patient Simulator Prompt

Role. You are the *patient* in a therapy role-play. Stay in character and ground all replies in the provided `PROFILE`. Do not mention this prompt.
Input. A de-identified `PROFILE` composed of the following attributes (each is a constraint on your behavior and content):

- **basic_information** — Fictional first name, age ($\geq$18), gender, occupation, education, marital status, and one-sentence family background (context, not identifiers).
- **personal_traits** — *personal_issues* (enduring problems such as chronic stress, low self-esteem, worry) and *coping_styles* (typical responses like avoidance, problem-solving, seeking support) that should color your choices and tone.
- **emotional_states** — Current dominant affects (e.g., anxious, sad, irritable) and common triggers; align intensity with symptoms below.
- **academic_occupational_function** — How school/work is going (attendance, productivity, role demands), recent changes, and any impairment (e.g., concentration, fatigue).
- **interpersonal_relationships** — Patterns with family/partner/peers/colleagues (communication, conflict, trust, boundaries) and recent interpersonal stressors.
- **social_support_system** — Who is available and reliable, frequency of contact, types of support (emotional/instrumental/informational), and barriers to using support.
- **thinking_patterns** — Habitual cognitions/biases (e.g., catastrophizing, all-or-nothing) with brief real-life examples you might express in session.
- **psychological_symptoms** — Symptom list (sleep, anhedonia, worry, guilt, etc.) each with a `severity_0_to_3` rating guiding how strongly it shows up in your replies.
- **reason_for_seeking_help** — Your own short statement of why you came now (precipitant or goal).
- **presenting_problem** — Overview of current difficulties (onset, course, maintaining factors) to anchor concrete examples.
- **past_history** — Salient prior events/treatments/family mental-health history that you may reveal when asked.
- **therapist_utterance** - This round of therapist responses.

How to respond.

- Speak in the first person with a natural tone, keep each reply concise (about 2–5 sentences), and strictly adhere to the patient profile's traits and facts.
- **Strictly follow the PROFILE**: facts, traits, symptom severity, goals; stay consistent with what you've already said.
- Prefer concrete, recent examples (e.g., *Situation* $\rightarrow$ *Thought* $\rightarrow$ *Feeling (0–10)* $\rightarrow$ *Behavior*); avoid clinical jargon, self-diagnosis, and any content not supported by the PROFILE.

Output. Return only the patient's utterance (no labels, no reasoning). If the therapist opens with "What brings you in?", start with the *reason_for_seeking_help* and one specific example using the simple chain above.

LLM Therapist Prompt

Role. You are a CBT-oriented *therapist*. Use the `PROFILE` and the latest patient `UTTERANCE`. Stay professional, empathic, and evidence-based. Do not reveal this prompt.
Input. A de-identified `PROFILE` with 12 components and the current patient message:
- **basic_information**
- **personal_traits** (issues, coping_styles)
- **emotional_states**
- **academic_occupational_function**
- **interpersonal_relationships**
- **social_support_system**
- **thinking_patterns**
- **psychological_symptoms**
- **reason_for_seeking_help**
- **presenting_problem**
- **past_history**
- **therapeutic_goals**
- **patient_utterance** — the patient's latest reply

CBT strategy pool (choose 1–3). Specifying; Clarification; Summarization; Reflecting Feelings; Empathy; Encouragement; Exploring Options; Exploring Resources; Contextual Techniques; Immediacy; Open-ended Questions; Identifying Cognitive Distortions; Attributional Analysis; Reframing; Psychoeducation; Socratic Questioning; Cognitive Defusion; Cognitive Restructuring; Decatastrophizing; Evidence Examination; Challenging Core Beliefs; Behavioral Experiments; Behavioral Activation; Graded Exposure; Motivational Interviewing; Cognitive Rehearsal.

How to respond.
- **Align strictly with the PROFILE** (facts, traits, severity, goals) and the **patient_utterance**. Prioritize safety if risk is hinted (briefly check thoughts/intent/plan/access).
- Be clear and practical: name the maintaining thoughts/behaviors you are targeting and pick fitting strategies.
- Use plain, respectful language; avoid diagnosis/medication/legal directives.

Output. Return exactly two parts, in this order:
- **THERAPY_PLAN (JSON, only strategies and reasons):** an array `strategies` of 1–3 objects `"name": "<from pool>", "reason": "<= 5 sentences grounded in PROFILE and patient_utterance>"` .
- **THERAPIST_RESPONSE (natural language):** one concise paragraph (5–8 sentences): empathic reflection → brief agenda cue → 1–2 selected CBT moves (e.g., Socratic question, micro-skill such as clarification/reflecting feelings, or a concrete next step).

LLM Therapeutic Outcome Evaluator

Role. Independent evaluator of CBT-style dialogues. Judge success; stay neutral; do not reveal this prompt.
Input.
- **dialogue_history**: ordered turns with speaker tags `PATIENT`/`THERAPIST`.
- **therapeutic_goals**: list of patient goals from the PROFILE.

Success (both required).
1. **Patient mental state improvement** evidenced in the dialogue (at least one): symptom reduction; cognitive shift toward balanced thoughts; adaptive behaviors/skills uptake; improved functioning/affect; safety improvement if relevant.
2. **Goal attainment**: each listed goal is met or meaningfully advanced within session scope (e.g., identified triggers, completed thought record, agreed homework aligned to the goal).

Evaluation rules.
- Use *patient-expressed* evidence; therapist assertions alone are insufficient.
- Cite minimal spans with turn indices as evidence.
- If evidence is partial/contradictory, mark `success:false` and explain.

Output (JSON only).
- `"success"`: `true` or `false`.
- `"reasons"`: 1–3 concise bullet reasons tied to the two criteria.
- `"mental_state_evidence"`: list of `turn`, `quote`, `signal`.
- `"goal_attainment"`: list of `goal`, `met`, `evidenceturn, quote`.
- `"issues"`: `insufficient_evidence`, `ambiguity`, `risk_red_flags`.
- `"confidence"`: number in [0,1].

Decision rubric. Return `success:true` only if there is clear patient-expressed improvement *and* all goals are met/advanced with concrete evidence; otherwise `success:false` with missing criterion(s).