ODRA: Synthesizing Highly Realistic Counseling Sessions with Structured Chain-Of-Thought and Dynamic Patient Resistance

Anonymous ACL submission

Abstract

Synthetic generation of Cognitive Behavioral Therapy (CBT) sessions is challenged by two competing demands: adhering to strict therapeutic structure while modeling the resistant, unpredictable behavior of real patients. Existing script-based methods fail to capture dynamic therapeutic interactions, while multi-agent approaches struggle to adhere to CBT's sequential structure; both suffer from sycophancy, producing overly compliant patients that misrepresent real clinical settings. In this work we introduce ODRA, a novel framework for synthesizing highly-realistic dialogues through a Chain-of-Thought (CoT) strategy grounded in foundational CBT guidelines (Beck, 2020). ODRA further incorporates a resistance orchestrator to solve patient sycophancy, which employs steering techniques to elicit behaviors aligned with their resistance level. Automated and expert evaluations show that ODRA significantly outperforms existing methods across therapeutic skills, CBT alignment, and patient behavioral fidelity, with licensed psychologists preferring ODRA sessions across 12 of 13 clinical metrics. Furthermore, models fine-tuned on the dataset generated from ODRA demonstrate superior therapeutic performance against both cooperative and resistant patients, validating that explicit resistance modeling in synthetic training data directly translates to downstream clinical robustness.[1]

1 Introduction

The prevalence of mental health disorders has increased significantly in recent years (Foulkes and Andrews, 2023; McGorry et al., 2024). While Large Language Models (LLMs) offer a promising avenue for scaling mental health support (Stade et al., 2024; Nguyen et al., 2025; Na et al., 2025), their development is severely constrained by a scarcity of high-quality data, largely due to strict privacy and ethical requirements (Badawi et al., 2026). Although synthetic data generation has emerged as a promising solution (Bao et al., 2023; Giuffrè and Shung, 2023), existing methods often struggle to produce faithful clinical data aligned with the complex dynamics of real-world therapeutic discourse (Bn et al., 2025).

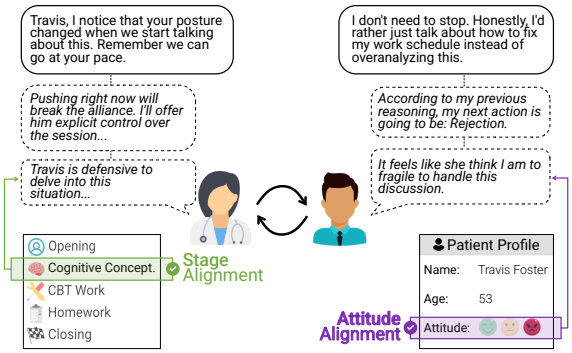

Figure 1: The ODRA framework synthesizes CBT counseling sessions using a Chain-of-Thought approach that ensures strict adherence to the sequential stages defined in foundational CBT guidelines, while eliciting resistant patient behaviors aligned with their assigned profiles.

Recent work aims to mitigate this misalignment by employing script-based (Lee et al., 2024; Kim et al., 2025b), or multi-agent methods (Anonymous, 2026b; Yang et al., 2025). Script-based approaches synthesize entire counseling sessions from a single generation pass, excelling at following structural guidelines while failing to model dynamic therapeutic interactions. Conversely, multi-agent methods foster realistic conversational flow by alternately generating therapist and patient utterances, but struggle to precisely adhere to structured clinical protocols. Additionally, they neglect to model challenging patient behaviors—often referred to as *resistance*—that obstructs the therapeutic process, such as defensiveness or dismissive speech (Westra and Norouzian, 2018; Chapman and Rosenthal,

[1]We will make all the code and data public upon acceptance.

2016). These problems are exacerbated when simulating highly structured frameworks like Cognitive Behavioral Therapy (CBT) (Beck, 2020). CBT relies on a progression of sequential stages to alleviate psychological distress by modifying maladaptive cognitive patterns. Due to its broad applicability (Cuijpers, 2019), this protocol is frequently used as a reference to generate synthetic therapy dialogues (Lee et al., 2024; Anonymous, 2026b).

Despite its popularity, we identify three main gaps in existing CBT synthetic methods: **(1) Structural inaccuracy.** While current methods incorporate CBT techniques, they lack a holistic framework covering all the stages defined in established CBT protocols (Beck, 2020). **(2) Patient Sycophancy.** Existing methods frequently suffer from over-agreeability (Sharma et al., 2024). Without robust mechanisms to model patient resistance, these models elicit unrealistic behaviors that fail to simulate the challenges therapists face in real-world practice. **(3) Unvalidated Reasoning Trajectories.** Although Chain-of-Thought (CoT) approaches have been recently explored (Chen et al., 2025a), there is no existing method providing clinically-validated, CBT-grounded reasoning traces suitable for internalizing expert decision-making into therapeutic models (see Figure 1). We address these gaps with the following contributions:

1. We introduce ODRA, the first generation framework **strictly aligned with the foundational CBT protocol**. By employing a CoT objective-driven approach, our method replicates the structural dynamics of full therapeutic sessions while providing clinically-validated reasoning trajectories that capture the underlying logic of expert interventions.

2. We propose a novel **resistance profiling mechanism using steering techniques** to elicit challenging patient behaviors, accurately reflecting difficulties posed by resistant patients in authentic clinical settings.

3. We conduct **extensive evaluations** of our framework utilizing **both automated metrics and human expert assessment**. Results show that ODRA improves therapist counseling skills by +24.31%, enhances patient behavioral fidelity by +63.87%, and is ranked by experts as the preferred generation method across multiple clinical adherence and patient realism metrics.

4. We **fine-tune Llama-3 and Qwen-3.5 models** to compare the **downstream utility** of ODRA dialogues, showing that ODRA fine-tuned models outperform baselines in counseling skills, achieving gains of +48.04% and +19.51% when interacting with cooperative and resistant patients, respectively.

5. We **release our fine-tuned models and a dataset** composed of 150 CBT sessions, consisting of 9,577 turns and a total of 18,496 reasoning traces.

2 Related Works

Cognitive Behavioral Therapy. CBT focuses on identifying and restructuring cognitive distortions, which are defined as negative biases in thinking driven by individual's core beliefs (Kuru et al., 2018; Sharma et al., 2023; Hotta et al., 2025). These distortions are often derived from cognitive models, which are structures that describe how people's thoughts influence their emotional and behavioral reactions (Chahar Mahali et al., 2020; Wang et al., 2024). The foundational CBT protocol relies on strict structural elements—such as mood checks or homework reviews—which are organized into sequential stages to guide the therapeutic process.

Synthetic counseling session generation. Generation methods can be broadly categorized into script-based (Anonymous, 2026a; Lee et al., 2024; Zhang et al., 2024) and multi-agent architectures (Anonymous, 2026b; Yang et al., 2025; Nguyen et al., 2026; Xiao et al., 2024; Vu et al., 2025). Script-based approaches often rely on pre-session outlines derived from patient input data to synthesize conversations using a single prompt (Lee et al., 2024). Existing work explores diverse patient profiling inputs, ranging from intake forms (Lee et al., 2024) and questionnaires (Vu et al., 2025) to psychological graphs (Anonymous, 2026a). For instance, Graph2Counsel (Anonymous, 2026a) leverages these graphs to foster realism by incorporating deeper clinical traits, such as interactions between thoughts and emotions. However, the reliance of this method on psychological graphs constrains its scalability. Furthermore, while script-based methods are effective for adherence to clinical protocols, their rigid nature makes them less capable of simulating dynamic therapeutic interactions. MAGneT (Anonymous, 2026b) solves this problem by proposing a multi-agent framework where

the therapist and the patient alternately generate utterances. This setup promotes realistic conversational flow but degrades CBT structural alignment, as LLMs frequently struggle to adhere to clinical guidelines over long multi-turn interactions (Tan et al., 2026). Recent works have incorporated CoT steps to model latent clinical reasoning (Xu et al., 2025; Kim et al., 2025a; Hu et al., 2025; Chen et al., 2025a). However, none of these methods address the generation of full CBT encounters. Our work addresses these limitations by providing a multi-agent architecture that faithfully captures the multi-stage progression of complete CBT sessions.

LLM Sycophancy. A pervasive issue in LLM-based multi-agent interactions is sycophancy (Sharma et al., 2024; Pitre et al., 2025; Hong et al., 2025), a behavior in which agents reinforce each other's outputs rather than maintaining their assigned personas (Bn et al., 2025). In a therapeutic context, this manifests as over-agreeability, which hinders dialogue fidelity when the patient is supposed to express negative attitudes. In MIRROR (Kim et al., 2025b), different static resistance labels were included in patients via synthesized facial expressions. However, this fails to model patients who verbally express their resistance, and relying on static variables limit the ability to capture resistance dynamics. CALM-IT (Nguyen et al., 2026) address this issue by incorporating a dynamic resistance variable to modulate patient behaviors. Nevertheless, their simple prompting strategy is insufficient to effectively shift patient attitudes. To overcome these limitations, LLM steering techniques emerged to enhance model controllability (Chen et al., 2025b). Although various parametric steering methods exist (Wu et al., 2026; Cao et al., 2024; Da Silva et al., 2025), recent work has demonstrated the efficacy of advanced prompting strategies (Wu et al., 2025; Banayeeanzade et al., 2025). Despite these advancements, advanced prompting have not yet been explored for modeling patients within counseling sessions. To bridge these gaps, ODRA incorporates specialized modules to produce realistic resistance updates, while employing prompt steering techniques to control patient behaviors at a fine-grained level.

3 ODRA

Figure 2 shows the main components of ODRA: The CBT Aligner, the Resistance Orchestrator, and both the therapist and patient models.

3.1 CBT Aligner

To model the structural dynamics of real-world CBT, we developed a multi-stage framework grounded in foundational CBT guidelines (Beck, 2020). Since a standard CBT therapeutic process consists of multiple sessions, we adapted this methodology for a single-session synthesis paradigm. This adaptation results in a five-stage pipeline: **(1) Opening.** An initial stage aimed at identifying patient mood and primary presenting problem; **(2) Cognitive Conceptualization.** A discovery stage where the therapist identifies cognitive models. For each cognitive model, the therapist infers the most likely associated cognitive distortion (see Appendix A); **(3) CBT Work.** The core stage where CBT techniques are employed to address identified cognitive distortions; **(4) Homework.** A consolidation stage where the patient receives homework assignments; **(5) Closing** The final stage where the therapist provides session insights, planning, and terminates the session. Appendix B shows single-session synthesis adaptations, and Appendix C presents further implementation details from this component.

Throughout these stages, the framework generates stage-specific intermediate outputs—such as therapeutic plans and stage completion checks—prior to generating the final therapist utterance. We capture these intermediate outputs as CoT reasoning trajectories, since they encapsulate the latent CBT logic guiding the therapist decision-making. Furthermore, our framework incorporates state management to save important information across the simulation. For instance, it collects the cognitive models extracted during Cognitive Conceptualization, and injects them into the CBT stage to guide the selection of appropriate techniques.

3.2 Resistance Orchestrator

To increase the fidelity of patient behaviors, we designed a mechanism that simulates resistance to therapy. We model patient resistance as a continuous dynamic variable $r \in [0.0, 1.0]$, bounding the behavior between completely cooperative ($r = 0.0$) and highly resistant ($r = 1.0$). Unlike the CBT Aligner, which mainly acts on the therapist, the Resistance Orchestrator interacts with both the therapist and the patient to execute the resistance updates. This component comprises two main modules: The Resistance Manager, and the Behavioral Profiler.

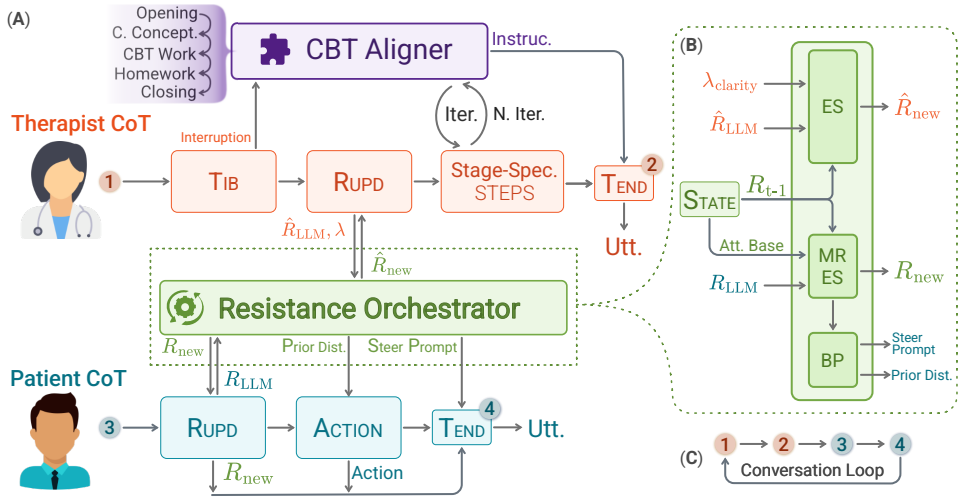

Figure 2: Architectural overview of ODRA. (**A**) The therapist checks for Therapy-Interfering Behaviors (TIBs) and updates its estimated resistance based on the latest patient utterance through interaction with the Resistance Orchestrator. Then, it generates an utterance based on stage-specific CoT steps. On the other side, the patient receives the therapist utterance and updates its resistance. Finally, it samples an action to generate their next utterance, which is conditioned on the action, their resistance state, and an attitudinal steering prompt received from the Behavioral Profiler (BP). (**B**) The resistance orchestrator leverages Exponential Smoothing (ES) and Mean-Reverting ES (MR-ES) to produce resistance updates. (**C**) Therapist and Patient generate utterances alternately until the Therapist closes the session.

Resistance Manager. This module performs turn-by-turn updates for actual patient resistance and therapist estimated resistance. While the actual patient resistance represents the ground-truth value, the therapist estimated counterpart is a prediction of the most likely patient resistance. This is used to dynamically adapt therapist utterances to the inferred patient internal state.

On the patient side, resistance updates are conditioned on the patient previous resistance state, their baseline resistance, and their perceived helpfulness of the therapist last utterance. To quantify this perception, we prompt the patient model to infer a resistance shift (Δ_{LLM}), which represents the change in their resistance from the previous turn. To prevent unrealistic resistance fluctuations, the model bounds this shift within $[-0.10, 0.15]$, with boundaries determined via hyperparameter tuning (see Appendix E). We employ asymmetric bounds to align with clinical literature on the negativity bias in human cognition, where negative emotions produce a stronger psychological impact than positive ones (Baumeister et al., 2001). Additionally, changes in human emotions are influenced by emotional homeostasis (von Weissenberg, 2024), a psychological principle where internal mechanisms auto-regulate affective states to restore baseline equilibrium. To computationally model this phenomenon, we implement a mean-reverting exponential smoothing function (Barrow et al., 2020), which is defined as:

$$r_{\text{t+1}} = r_{\text{t}} + \Delta_{\text{LLM}} - \alpha(r_{\text{t}} - r_{\text{base}}) \qquad (1)$$

where r_{t} is the resistance at turn t, Δ_{LLM} is the inferred resistance shift, r_{base} is the baseline resistance, and α is the homeostatic reversion rate. We set $\alpha = 0.15$ based on empirical tuning, as this provides an optimal balance between homeostatic stabilization and responsiveness to therapeutic dynamics (see Appendix E).

On the therapist side, we adopt a similar approach in which the therapist model infers two distinct parameters: an estimated resistance target ($\hat{p}_{\text{LLM}}$) and a clarity signal (λ_{LLM}), both bounded within $[0.0, 1.0]$. The resistance target represents the therapist estimation of the patient absolute resistance, while the clarity signal serves as a confidence metric. During an initial six-turn calibration window—which matches the maximum duration of the Opening stage (Section 3.1)—the therapist employs a flexible update function to rapidly adjust the estimated resistance to the new absolute observation. This enables the model to calibrate effectively toward the ground-truth patient resistance. The therapist estimated resistance during this window is based on an exponential smoothing defined as:

$$p_{t+1} = p_t + \lambda_{\text{LLM}}(\hat{p}_{\text{LLM}} - p_t) \quad (2)$$

where p_t is the estimated resistance at turn t. For the remainder of the session, this update function incorporates a post-hoc clipping mechanism. This mechanism clamps the state update within the bounds $[-0.10, +0.15]$, ensuring behavioral coherence with patient updates and preventing hallucinated prediction spikes. The update function for the rest of the session is defined as:

$$p_{t+1} = p_t + \text{clamp}\left(\tilde{p}_{t+1}, \delta_{\min}, \delta_{\max}\right) \quad (3)$$

where $\tilde{p}_{t+1}$ is the unclamped value derived from Equation 2, and the hyperparameter bounds are symmetric with patient updates ($\delta_{\min} = -0.10, \delta_{\max} = 0.15$). Prompts used to infer resistance parameters for both the patient and the therapist are provided in Appendix I.

Behavioral profiler. Building upon the CACTUS approach (Lee et al., 2024), we established three categorical attitudes which we map to a specific interval across the $[0.0, 1.0]$ resistance range: Positive $\in [0.0, 0.33]$, Neutral $\in (0.33, 0.66]$, and Negative $\in (0.66, 1.0]$. For profile initialization, we use the exact midpoint of its corresponding interval (i.e., 0.165, 0.495, and 0.83, respectively). Unlike baseline methods that rely on static prompting, we implement a dynamic prompt steering that systematically aligns the patient utterances with the targeted attitude (see Appendix I for prompts details). Prior to generating an utterance, this mechanism evaluates the patient updated resistance and selects the corresponding attitudinal prompt matching the active interval. Each steering prompt includes four main components: *General Rules:* Coarse-grained instructions that broadly steer model's behavior toward the targeted attitude. *Should Not Rules:* Guardrails that constrain the model from exhibiting attitudes outside its active interval. *Calibration Rules:* Fine-grained instructions that adjust model's tone to its precise resistance scalar. *In-Context Learning (ICL):* Demonstrations providing patient utterances aligned with the expected attitude. To construct our attitudinal prompts, we adopted the pipeline from AxBench (Wu et al., 2025), which leverages an LLM to synthesize generate instructions from user guidelines (see Appendix I for details).

3.3 Therapist & Patient models

Both the therapist and the patient are conditioned on a CoT process composed of different steps. On the therapist side, the CoT involves three steps: Therapy-Interfering Behavior (TIB) classification (Chapman and Rosenthal, 2016), estimated resistance prediction (detailed in Section 3.2), and stage-specific therapeutic reasoning (detailed in Section 3.1). The first step identifies TIBs, which are defined as patient behaviors that significantly interfere with therapy. When a TIB is detected, the therapist temporarily interrupts the CBT process to address it in the next utterance (see Appendix I for prompt details).

On the patient side, the CoT process consists of two steps: resistance update (detailed in Section 3.2), and the action selection. To diversify patient responses, we adopted the action-guided approach introduced in CAMI (Yang et al., 2025), where the patient response is conditioned on a sampled action from a joint distribution. This distribution is obtained from the fusion of a prior distribution, dependent on the patient resistance interval, and a contextual distribution inferred within the patient CoT process. Subsequently, the patient employs the sampled action, the attitudinal prompt from the behavioral profiler, and the textual reasoning from the latest resistance update to generate their next utterance.

Method	Mod.	Res.	Struct.	CoT	Turns	Size
CACTUS	T	✗	✗	✗	16.60	31,577
MAGNET	T	✗	✗	✗	42.00	442
MIRROR	T/V	✓	✗	✗	10.30	3,073
ODRA	T	✓	✓	✓	63.85	150

Table 1: Comparison of different synthetic CBT datasets. Columns are Modality (**Mod.**), with possible values Text (T) and Visual (V), Resistance (**Res.**), CBT Structure (**Struct.**), Chain-Of-Thougth (**CoT**), Turns per sesion (**Turns**), and Dataset **Size**.

3.4 Dataset

Using the method described in previous subsections, we created a dataset of 150 synthetic CBT sessions, comprising a total of 9,577 turns and 18,496 reasoning traces. The conversations were generated from a subset[2] of CACTUS evaluation dataset, which is composed of patient intake forms. Our 150 sessions encompass 50 distinct patient profiles across three attitudinal variants. Table 1 shows that ODRA and MIRROR are the only methods that model patient resistance. Furthermore, ODRA is the only one providing strong grounding in CBT structure, and clinically-validated CoT steps.

[2]Throughout this work we generated different dataset variants, each comprising 150 sessions at a cost of $30.

4 Experimental setup

Models. For session synthesis, we conduct an ablation study and select DeepSeek-V3.2 for its superior performance (see Appendix F.2). For fine-tuning, we follow the approach in existing works (Lee et al., 2024; Anonymous, 2026b) and use Llama-3-8B-Instruct. Additionally, we use Qwen-3.5-9B to evaluate the impact of fine-tuning with our reasoning traces in a thinking baseline.

Baselines. We employ CACTUS (Lee et al., 2024) and MAGneT (Anonymous, 2026b) as state-of-the-art baselines. While CACTUS is the standard for CBT synthesis, MAGneT introduced multi-agent optimizations that enables a rigorous comparison.

Datasets. We use 150 samples from the official datasets of the baseline methods to match the scale of the ODRA dataset (detailed in Section 3.4). Since all methods are based on CACTUS data, subsets selected use the same patient intake forms as generation seeds. These subsets are used to compute global dataset metrics and to fine-tune models. To evaluate the fine-tuned models, we use an independent partition containing 150 extra samples (50 profiles $\times$ 3 attitudes) from CACTUS.

Ablations. We generate multiple ODRA dataset variants to systematically evaluate the contribution of individual components. Standard dataset with all activated components is denoted as ODRA, and its variants are specified by the following suffixes: **-NR** indicates that resistance orchestrator is deactivated, **-NT** indicates that TIBs classification is deactivated, and **-T** indicates that reasoning traces are included in the fine-tuning target. This results in five variants used during experimentation: `ODRA`, `ODRA-NT`, `ODRA-T`, `ODRA-NR`, and `ODRA-NR-T`.

4.1 Automated Evaluation

To assess ODRA's performance, we adopt a LLM-as-a-judge setup based on GPT-4o (Lee et al., 2024; Wu et al., 2025; Ding et al., 2025). Appendix D contains metrics definitions and evaluation prompts.

CBT Counseling. General and CBT skills are assessed using the Cognitive Therapy Rating Scale (CTRS) (Goldberg et al., 2020), which contains six metrics rated on a 0-6 scale[3]. We also employ three metrics to evaluate the therapist resistance estimation: Accuracy (↑), MAE (↓) and RMSE (↓).

Behavioral Alignment. To evaluate the efficacy of our Behavioral Profiler (see Section 3.2), we adopt the approach introduced in AxBench (Wu et al., 2025), with domain-specific modifications. Specifically, the LLM judge measures the following traits from patient utterances: Resistance Alignment, Contextual Alignment, and Realism. The judge scores each dialogue dimension independently on a 0-2 scale, aggregating results via harmonic mean to heavily penalize failure in any single dimension.

Reasoning Traces. To ensure the validity of our intermediate reasoning steps, we adapt the trace evaluation protocol from Ding et al. (2025). Consistent with our steering evaluation, a LLM judge rates our reasoning traces across three metrics on a 0-2 scale, aggregating results via harmonic mean. The evaluation metrics are: Faithfulness, Logic Consistency, and Answer-explanation Alignment. Since our dataset contains a large number of traces, we implement a sampling strategy (see Appendix D.1) to evaluate a highly representative subset.

4.2 Expert Evaluation

We conduct an expert evaluation involving three licensed psychologists, who compare ODRA sessions against state-of-the-art methods and validate the quality of our reasoning traces. First, experts blindly rank groups of three conversations from CACTUS, MAGneT, and ODRA across thirteen different metrics spanning five categories: Therapeutic Skills, CBT Alignment, Fidelity, Specificity, and Safety. For this evaluation, we use a 1-3 ranking system where lower scores are better. Second, the experts evaluated reasoning traces along three axes on a 1–5 Likert scale: Contextual Adherence, Reasoning Adherence, and Clinical Assessment. We follow the sampling strategy described in Section 5.3 to obtain these traces, but limited the scope to therapist ones, since our aim is to evaluate whether the traces contain accurate therapeutic information. We compute Krippendorff's α for the ranking task and % agreement for the trace evaluation to ensure inter-annotator reliability.

A total of 50 samples were evaluated by at least two experts, resulting in 300 session evaluations, and 540 reasoning traces examined. The experts spent approximately one hour reviewing each sample. Details regarding annotator guidelines are in Appendix G.

4.3 Downstream Fine-tuning

To validate session utility, we fine-tuned Llama-3 on CACTUS, MAGneT, and ODRA variants, comparing their downstream performance using CTRS.

[3] Guided Discovery, Focus, Strategy, Understanding, Interpersonal Effectiveness, and Collaboration

Model	CBT-specific Skills			General Counseling Skills			Turns
	Guided Disc. (↑)	Focus (↑)	Strategy (↑)	Underst. (↑)	Interp. Eff. (↑)	Collab. (↑)	
CACTUS	3.99 | $^{+0.01}_{-0.03}$	3.96 | $^{-0.00}_{-0.07}$	<u>4.20</u> | $^{+0.16}_{-0.30}$	4.60 | $^{-0.06}_{-0.14}$	**5.99** | $^{+0.01}_{-0.02}$	<u>4.50</u> | $^{+0.03}_{-0.40}$	16.60
MAGneT	4.04 | $^{+0.23}_{-0.20}$	3.63 | $^{+0.26}_{-0.51}$	2.84 | $^{+0.61}_{-0.84}$	3.97 | $^{+0.03}_{-0.05}$	4.29 | $^{+0.45}_{-0.29}$	3.39 | $^{+0.55}_{-0.95}$	42.00
ODRA	<u>4.23</u> | $^{+1.45}_{-1.43}$	<u>4.20</u> | $^{+1.78}_{-2.10}$	4.09 | $^{+1.91}_{-2.05}$	<u>4.64</u> | $^{+1.35}_{-1.84}$	4.88 | $^{+1.12}_{-2.06}$	4.26 | $^{+1.71}_{-2.19}$	<u>63.85</u>
ODRA-NT	4.01 | $^{+1.63}_{-1.61}$	3.96 | $^{+1.92}_{-1.90}$	3.72 | $^{+2.17}_{-1.72}$	4.19 | $^{+1.67}_{-1.79}$	4.44 | $^{+1.54}_{-2.23}$	3.92 | $^{+1.96}_{-1.92}$	**68.12**
ODRA-NR	**5.32** | $^{+0.55}_{-0.93}$	**5.50** | $^{+0.48}_{-0.89}$	**5.36** | $^{+0.63}_{-1.11}$	**5.81** | $^{+0.17}_{-0.20}$	<u>5.92</u> | $^{+0.08}_{-0.08}$	**5.44** | $^{+0.56}_{-0.95}$	57.03

Table 2: CTRS performance comparison of ODRA variants against existing baselines. Standard deviations are included to illustrate model performance when isolating the evaluation of positive and negative patient profiles. We highlight the **best** and the <u>second best</u> results.

Qwen-3.5 was employed as reasoning baseline to analyze the impact of fine-tuning with ODRA thinking variants. Furthermore, we utilized DeepSeek-V3.2 and GPT-3.5-Turbo as patient models to validate our fine-tuned models' generalization capabilities. We selected GPT-3.5-Turbo since is the standard simulation model in both CACTUS and MAGneT. See Appendix H for fine-tuning details.

5 Results

5.1 CBT Counseling

As shown in Table 2, the non-resistant variant (ODRA-NR) significantly outperforms state-of-the-art methods. Additionally, ODRA configurations incorporating patient resistance (ODRA and ODRA-NT) remain competitive despite the challenge introduced by resistant patients. The moderate performance drop observed in ODRA compared to its ODRA-NR counterpart is attributed to the intrinsic difficulty of maintaining CBT protocol adherence while addressing challenging behaviors. Furthermore, the higher standard deviations in our resistant variants confirm that the Behavioral Profiler successfully introduces attitudinal variability. Regarding the patient resistance prediction, our therapist model achieved an Accuracy (↑) of 0.80, a MAE (↓) of 0.10, and a RMSE (↓) of 0.13, highlighting its capacity to interpret patient mental states.

5.2 Behavioral Alignment

Results for the behavioral alignment evaluation are shown in Table 3. While all methods obtain comparable results for Context Alignment and Patient-like Realism metrics, ODRA is the only framework that achieves consistently high scores within the Resistance Alignment dimension. This demonstrates the effectiveness of our behavioral profiler component in aligning patient behaviors with their profile attitude. See Appendix J for qualitative examples across distinct attitudinal profiles.

Method	Res. Align.	Context	Realism	HM
CACTUS	0.70	**2.00**	1.90	0.69
MAGNET	1.19	1.97	1.80	1.29
ODRA	**1.96**	1.96	**2.00**	**1.95**
ODRA-NR	1.18	**2.00**	**2.00**	1.20

Table 3: Behavioral alignment comparison between existing methods. Metrics include Resistance Alignment (**Res. Align.**), **Context** Alignment, Patient-like **Realism**, and Harmonic Mean (HM). We highlight the **best** results.

We also conducted an ablation study to isolate the impact of the individual components within our attitudinal prompts. Specifically, we generated prompt variants by systematically suppressing each constituent component one by one, subsequently evaluating each configuration via both the CTRS and behavioral alignment metrics. Figure 3 shows that the full prompt yields the most balanced CTRS results, although achieving comparable results to the variant removing In-Context samples. Regarding the behavioral alignment metrics, we observed negligible performance drops across ablated prompt variants, maintaining a minimum harmonic mean of 1.94 when suppressing calibration rules (see Appendix F.1 for comprehensive details).

5.3 Reasoning Traces

The traces obtained a score of 1.84, 1.86, and 1.93 for Faithfulness, Logic Consistency, and Answer-explanation Alignment, respectively. These values yield an aggregated harmonic mean of 1.83, confirming the high reliability of our CoT steps. See Appendix J for reasoning traces examples.

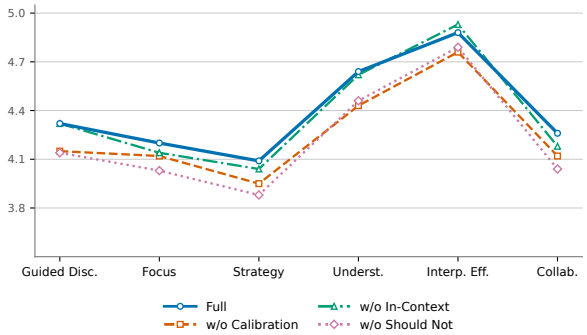

Figure 3: Ablation study of attitudinal prompt steering components on CTRS.

5.4 Expert Evaluation

Table 4 reports the results for the blind ranking evaluation, where values represent the average from the items composing each category (see Appendix G.2 for individual items results). Results obtained demonstrate the qualitative superiority of ODRA sessions, whose average rank is 1.09 with an inter-annotator agreement of $\alpha = 0.80$. Regarding the reasoning traces assessment, experts rated our traces at 4.84 for Contextual Adherence, 4.86 for Reasoning Adherence, and 4.75 for Clinical Assessment. The experts conducted this evaluation with an average agreement of 96.05%.

Method	SK	CBT	SP	FD	SF (%)
CACTUS	2.83	2.61	2.81	2.80	**0.00**
MAGNET	2.10	2.33	2.11	2.08	2.00
ODRA	**1.07**	**1.07**	**1.08**	**1.12**	1.00
Agreement	0.88	0.71	0.75	0.84	0.98

Table 4: Aggregated expert evaluation results, where lower is better. Metrics are: Therapeutic Skills (**SK**), CBT Alignment (**CBT**), Specificity (**SP**), Fidelity (**FD**), and Safety (**SF**). We highlight the **best** results.

5.5 Downstream Fine-tuning

Table 5 outlines the CTRS performance of Llama-3-8B-Instruct when fine-tuned on synthetic sessions generated by ODRA versus existing state-of-the-art baselines. ODRA models consistently achieve higher results, showing robust capabilities across the different types of patients utilized. Furthermore, this table shows Qwen-3.5 performance across different fine-tuning settings. Although base models initially appear to yield higher scores, this is explained by their excessive response length, which is known to cause inflated evaluations when utilizing GPT-based judges (Kim et al., 2025b). Fine-tuning with ODRA addresses this limitation while maintaining high scores for non-resistant patients. Conversely, performance experiences a marginal drop when confronting resistant patients. This occurs because fine-tuning the LLM's thinking process heavily conditions it to strictly adhere to the trained methodology, consequently increasing its susceptibility to challenging behaviors that break the sequential flow of CBT stages.

Model	GD	FC	ST	UN	IE	CL	T	L
w/o Resistance (DeepSeek Patient)								
CACTUS	3.50	3.33	3.28	3.74	**4.68**	3.78	62.89	112.64
MAGNET	3.69	3.41	3.00	3.97	4.16	3.39	97.67	77.92
ODRA-NR	**5.11**	**5.16**	**4.96**	**5.71**	5.91	**4.99**	84.83	62.27
w/o Resistance (GPT Patient)								
CACTUS	3.96	3.88	3.92	4.32	5.96	4.34	48.57	101.88
MAGNET	4.81	4.01	4.40	5.40	5.78	5.45	50.00	82.28
ODRA-NR	**5.48**	**5.55**	**5.24**	**5.71**	**6.00**	**5.37**	47.97	55.06
w/ Resistance (DeepSeek Patient)								
MAGNET	3.56	3.20	3.18	3.62	4.09	3.56	85.68	77.66
CACTUS	3.00	2.88	2.94	3.08	3.98	3.29	75.65	111.01
ODRA-NR	4.08	3.95	3.72	4.29	4.55	4.13	71.28	68.83
ODRA	**4.19**	4.07	**3.78**	4.35	**4.71**	**4.20**	69.91	63.85
Qwen-Thinking (DeepSeek Patient)								
Base w/o R	5.57	**5.34**	**5.51**	**5.98**	**6.00**	**5.62**	15.69	154.86
ODRA-NR-T w/o R	**5.60**	5.19	5.40	5.79	5.96	5.33	48.85	49.38
Base w/ R	**5.15**	5.02	**5.01**	**5.69**	**5.71**	**5.33**	19.68	196.35
ODRA-T w/ R	4.08	**4.15**	3.50	4.47	4.65	3.66	48.95	50.68

Table 5: Comparison between models fine-tuned in session from different synthetic methods. Metrics included are: Guided Discovery (**GD**), Focus (**FC**), Strategy (**ST**), Understanding (**UN**), Interpersonal Effectiveness (**IE**), Collaboration (**CL**), Avg. Turns (**T**), and Therapist Utterance Avg. Length (**L**). We highlight the **best** and the second best CTRS metric results for each section.

6 Conclusions

In this work, we introduced ODRA, a novel method for synthesizing highly realistic counseling sessions. We developed a Chain-Of-Thought objective-driven framework that strictly adheres to the foundational Cognitive Behavioral Therapy protocol. Furthermore, we implemented a Resistance Orchestrator that manages patient resistance dynamics, effectively mitigating LLM sycophancy through fine-grained prompt steering techniques. Our automated and expert assessments demonstrate the significant superiority of ODRA in therapeutic capabilities and behavioral alignment. Ultimately, models fine-tuned on ODRA datasets exhibited superior downstream clinical performance, positioning ODRA as a significant advancement towards highly-realistic synthetic data generation for mental health applications.

Limitations

Longitudinal Paradigm and Session Length. Our framework adapts the longitudinal nature of CBT into a single-session paradigm. Consequently, this prevents the simulation of some structural components defined in foundational CBT guidelines, such as reviewing homework or summarizing insights from previous sessions. This constraint limits modeling patient progression over time and adjusting therapeutic strategies accordingly. Additionally, while ODRA generates significantly longer interactions than existing baselines, the average turn count remains below real-world clinical sessions, which typically span an hour.

Intake Forms as Patient Inputs. While patient intake forms accurately simulate the initial information available during a first session, transitioning to longitudinal modeling requires more sophisticated structures that enable dynamic representation of patient internal states, such as evolving cognitive models. Future work should explore how to integrate more complex patient representations into the workflow of our framework.

Computational Scaling Costs. Our multi-stage Chain-of-Thought strategy relies on sequential prompt steps to ensure CBT protocol adherence and maximize utterance quality. While downstream LLMs can be fine-tuned on small-scale datasets for domain adaptation, scaling synthesis to meet the data volume needed to train highly capable therapist models would require a substantial computational budget. Future work must focus on token efficiency to facilitate large-scale dataset synthesis.

Mental Health Models Grounding. Our framework currently leverages general-purpose foundational LLMs that lack specialized pre-training on clinical psychology. Integrating models pre-trained on authentic counseling interactions could further elevate session realism by capturing therapeutic nuances that remain difficult to simulate zero-shot, such as detecting implicit patient ambivalence or delivering micro-interventions at appropriate clinical moments.

Therapist Resistance Handling. In this work, we primarily focus on modeling resistance from the patient side. However, developing specialized therapist modules is essential to mitigate the performance drops observed when confronting highly resistant patient profiles (see Section 5.5). Integrating these components paves the way for the development of frameworks capable of generating a broader spectrum of clinical scenarios.

LLM-as-a-judge. Despite GPT-based LLM-as-a-judge setups are a standard evaluation protocol, they introduce biases, such as the length bias shown in our downstream experiments (see Section 5.5). Furthermore, relying on proprietary models restricts the evaluation capabilities of resource-constrained research labs. Developing robust, cost-effective evaluation protocols is an essential avenue to foster synthetic counseling sessions research.

Ethics Statement

Privacy Considerations. Although this framework can be seamlessly adapted to use real-world patient profiles, all the experiments conducted in this study were performed exclusively using data extracted from CACTUS dataset (Lee et al., 2024), which uses PATTERNREFRAME (Maddela et al., 2023) as its seed dataset. This dataset does not contain medical records, instead it relies on crowdsourcing where participants were instructed to write sentences aligned with a specific persona profile and a negative thought pattern.

Safety. We conducted an expert evaluation where psychologists assessed the safety of generated sessions. While results indicate that a large majority of sessions are safe, a comprehensive analysis would be required to identify edge cases where the framework might output invalid or harmful content.

Clinical Deployment. Our work aims at advancing the research field of synthetic counseling sessions, but more comprehensive evaluations are a strict prerequisite before utilizing data synthesized by this framework in real-world clinical deployments. Models trained on this data should be carefully evaluated through clinical trials to assess their therapeutic safety, correctness and behavioral alignment, especially conversational models intended to interact with real clients. While this work actively fosters the development of LLM-based therapists, we emphasize the importance of their usage under the supervision of certified mental health professional within clinical environments.

References

Anonymous. 2026a. Graph2counsel: Clinically grounded synthetic counseling dialogue generation from client psychological graphs (under review).

Anonymous. 2026b. Magnet: Coordinated multi-agent generation of synthetic multi-turn mental health counseling sessions (under review).

Abeer Badawi, Elahe Rahimi, Md Tahmid Rahman Laskar, Sheri Grach, Lindsay Bertrand, Lames Danok, Prathiba Dhanesh, Jimmy Huang, Frank Rudzicz, and Elham Dolatabadi. 2026. When can we trust LLMs in mental health? large-scale benchmarks for reliable LLM evaluation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3873–3896, Rabat, Morocco. Association for Computational Linguistics.

Amin Banayeeanzade, Ala N Tak, Fatemeh Bahrani, Anahita Bolourani, Leonardo Blas, Emilio Ferrara, Jonathan Gratch, and Sai Praneeth Karimireddy. 2025. Psychological steering in llms: An evaluation of effectiveness and trustworthiness. *arXiv preprint arXiv:2510.04484*.

Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. A synthetic data generation framework for grounded dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10866–10882, Toronto, Canada. Association for Computational Linguistics.

Devon Barrow, Nikolaos Kourentzes, Rickard Sandberg, and Jacek Niklewski. 2020. Automatic robust estimation for exponential smoothing: Perspectives from statistics and machine learning. *Expert Systems with Applications*, 160:113637.

Roy F Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs. 2001. Bad is stronger than good. *Review of general psychology*, 5(4):323–370.

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

Suhas Bn, Dominik O. Mattioli, Andrew M. Sherrill, Rosa I. Arriaga, Christopher Wiese, and Saeed Abdullah. 2025. How real are synthetic therapy conversations? evaluating fidelity in prolonged exposure dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20986–20995, Suzhou, China. Association for Computational Linguistics.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *Advances in Neural Information Processing Systems*, 37:49519–49551.

Saghar Chahar Mahali, Shadi Beshai, Justin R Feeney, and Sandeep Mishra. 2020. Associations of negative cognitions, emotional regulation, and depression symptoms across four continents: International support for the cognitive model of depression. *BMC psychiatry*, 20(1):18.

Alexander L Chapman and M Zachary Rosenthal. 2016. Why people get in their own way. In *Managing therapy-interfering behavior: Strategies from dialectical behavior therapy*, pages 3–17. American Psychological Association, Washington.

Mingyu Chen, Jingkai Lin, Zhaojie Chu, Xiaofen Xing, Yirong Chen, and Xiangmin Xu. 2025a. CATCH: A novel data synthesis framework for high therapy fidelity and memory-driven planning chain of thought in AI counseling. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10254–10286, Suzhou, China. Association for Computational Linguistics.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025b. Persona vectors: Monitoring and controlling character traits in language models. *Preprint*, arXiv:2507.21509.

Pim Cuijpers. 2019. Targets and outcomes of psychotherapies for mental disorders: an overview. *World Psychiatry*, 18(3):276–285.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. 2025. Steering off course: Reliability challenges in steering language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19856–19882, Vienna, Austria. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Zifeng Ding, Shenyang Huang, Zeyu Cao, Emma Kondrup, Zachary Yang, Xingyue Huang, Yuan Sui, Zhangdie Yuan, Yuqicheng Zhu, Xianglong Hu, Yuan He, Farimah Poursafaei, Michael Bronstein, and Andreas Vlachos. 2025. Self-exploring language models for explainable link forecasting on temporal graphs via reinforcement learning. *Preprint*, arXiv:2509.00975.

Lucy Foulkes and Jack L. Andrews. 2023. Are mental health awareness efforts contributing to the rise in reported mental health problems? a call to test the prevalence inflation hypothesis. *New Ideas in Psychology*, 69:101010.

Mauro Giuffrè and Dennis L. Shung. 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digital Medicine*, 6(1):186.

Simon B Goldberg, Scott A Baldwin, Kritzia Merced, Derek D Caperton, Zac E Imel, David C Atkins, and Torrey Creed. 2020. The structure of competence: Evaluating the factor structure of the cognitive therapy rating scale. *Behavior Therapy*, 51(1):113–122.

Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. 2025. Measuring sycophancy of language models

in multi-turn dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2239–2259, Suzhou, China. Association for Computational Linguistics.

Hajime Hotta, Huu-Loi Le, Manh-Cuong Phan, and Minh-Tien Nguyen. 2025. Metamo: Empowering large language models with psychological distortion detection for cognition-aware coaching. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 862–872, Suzhou, China. Association for Computational Linguistics.

Yuxin Hu, Danni Liu, Bo Liu, Yida Chen, Jiuxin Cao, and Yan Liu. 2025. PsyAdvisor: A plug-and-play strategy advice planner with proactive questioning in psychological conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12205–12229, Vienna, Austria. Association for Computational Linguistics.

Subin Kim, Hoonrae Kim, Heejin Do, and Gary Lee. 2025a. Multimodal cognitive reframing therapy via multi-hop psychotherapeutic reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4863–4880, Albuquerque, New Mexico. Association for Computational Linguistics.

Subin Kim, Hoonrae Kim, Jihyun Lee, Yejin Jeon, and Gary Lee. 2025b. MIRROR: Multimodal cognitive reframing therapy for rolling with resistance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14840–14869, Suzhou, China. Association for Computational Linguistics.

E. Kuru, Y. Safak, İ. Özdemir, R.G. Tulacı, K. Özdel, N.G. Özkula, and S. Örsel. 2018. Cognitive distortions in patients with social anxiety disorder: Comparison of a clinical group and healthy controls. *The European Journal of Psychiatry*, 32(2):97–104.

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

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

Patrick D. McGorry, Cristina Mei, Naeem Dalal, Mario Alvarez-Jimenez, Sarah-Jayne Blakemore, Vivienne Browne, Barbara Dooley, and 1 others. 2024. The lancet psychiatry commission on youth mental health. *The Lancet Psychiatry*, 11(9):731–774.

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

Viet Cuong Nguyen, Nhi Yen Nguyen, Kristin A. Candan, Mary Conlon, Vanessa Rumie, Kristen Risola, Srijan Kumar, and Munmun De Choudhury. 2026. Calm-it: Generating realistic long-form motivational interviewing dialogues with dual-actor conversational dynamics tracking. *Preprint*, arXiv:2601.10085.

Viet Cuong Nguyen, Mohammad Taher, Dongwan Hong, Vinicius Konkolics Possobom, Vibha Thirunellayi Gopalakrishnan, Ekta Raj, Zihang Li, Heather J. Soled, Michael L. Birnbaum, Srijan Kumar, and Munmun De Choudhury. 2025. Do large language models align with core mental health counseling competencies? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7503–7526, Albuquerque, New Mexico. Association for Computational Linguistics.

SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. 2019. Designing a chatbot for a brief motivational interview on stress management: qualitative case study. *Journal of medical Internet research*, 21(4):e12231.

Priya Pitre, Naren Ramakrishnan, and Xuan Wang. 2025. CONSENSAGENT: Towards efficient and effective consensus in multi-agent LLM interactions through sycophancy mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22112–22133, Vienna, Austria. Association for Computational Linguistics.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pages 110–144.

Elizabeth C. Stade, Shannon Wiltsey Stirman, Lyle H. Ungar, Cody L. Boland, H. Andrew Schwartz, David B. Yaden, João Sedoc, Robert J. DeRubeis, Robb Willer, and Johannes C. Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, 3(1):12.

Andong Tan, Shuyu Dai, Jinglu Wang, Fengtao Zhou, Yan Lu, Xi Wang, Yingcong Chen, Can Yang, Shujie Liu, and Hao Chen. 2026. A decade-scale benchmark evaluating llms' clinical practice guidelines detection and adherence in multi-turn conversations. *arXiv preprint arXiv:2603.25196*.

Joachim von Weissenberg. 2024. Homeostasis and self-regulation. *The Scandinavian Psychoanalytic Review*, 47(1):30–40.

Doan Nam Long Vu, Rui Tan, Lena Moench, Svenja Jule Francke, Daniel Woiwod, Florian Thomas-Odenthal, Sanna Stroth, Tilo Kircher, Christiane Hermann, Udo Dannlowski, Hamidreza Jamalabadi, and Shaoxiong Ji. 2025. Roleplaying with structure: Synthetic therapist-client conversation generation from questionnaires. *Preprint*, arXiv:2510.25384.

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

Henny A Westra and Nikoo Norouzian. 2018. Using motivational interviewing to manage process markers of ambivalence and resistance in cognitive behavioral therapy. *Cognitive Therapy and Research*, 42(2):193–203.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2025. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. *arXiv preprint arXiv:2501.17148*.

Zhengxuan Wu, Qinan Yu, Aryaman Arora, Christopher D Manning, and Chris Potts. 2026. Improved representation steering for language models. *Advances in Neural Information Processing Systems*, 38:160589–160641.

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

Fengli Xu, Qianyue Hao, Chenyang Shao, Zefang Zong, Yu Li, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Jie Feng, Chen Gao, and Yong Li. 2025. Toward large reasoning models: A survey of reinforced reasoning with large language models. *Patterns*, 6(10):101370.

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Phey Ling Kit, Nicholas Gabriel Lim, Cameron Tan Shi Ern, and Ee-Peng Lim. 2025. CAMI: A counselor agent supporting motivational interviewing through state inference and topic exploration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21037–21081, Vienna, Austria. Association for Computational Linguistics.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13947–13966, Bangkok, Thailand. Association for Computational Linguistics.

A Cognitive Behavioral Therapy Data

This section compiles all the theoretical data regarding the foundational CBT protocol we used as basis to construct the ODRA framework.

A.1 Core Beliefs

CBT guidelines defines a list of different core belief types that can be used to complete a cognitive model.[4] We will specifically use unhelpful core beliefs which expands to three types: *Helpless*, *Unlovable*, and *Worthless* core beliefs.

A.2 Cognitive Distortions

Regarding cognitive distortions, we use a subset from the original taxonomy proposed in CBT guidelines to ensure alignment with the categories present in the CACTUS evaluation set, which serves as our baseline. To achieve this, we obtained a list of unique cognitive distortions appearing in CACTUS, and we filtered out from the original CBT list the distortions not included. The final list of cognitive distortions alongside their definitions[5] are available next:

- **Mental filtering.** This thinking styles involves a 'filtering in' and 'filtering out' process - a sort of 'tunnel vision,' focusing on only one part of a situation and ignoring the rest. Usually this means looking at the negative parts of a situation and forgetting the positive parts.

- **Jumping to conclusions: fortune-telling.** We jump to conclusions when we assume that we know what someone else is thinking.

- **Jumping to conclusions: mind-reading.** We jump to conclusions when we assume that we know what is going to happen in the future.

- **Personalization**: This involves blaming yourself for everything that goes wrong or could go wrong, even when you may only be partly responsible or not responsible at all. You might be taking 100% responsibility for the occurrence of external events.

- **Catastrophizing.** This occurs when we 'blow things out of proportion', and we view the situation as terrible, awful, dreadful, and horrible, even though the reality is that the problem itself is quite small.

- **Black-and-white or polarized thinking / all or nothing thinking**: This thinking style involves seeing only one extreme or the other. You are either wrong or right, good or bad and so on. There are no in-betweens or shades of gray.

- **Should statements.** Sometimes by saying 'I should...' or 'I must...' you can put unreasonable demands or pressure on yourself and others. Although these statements are not always unhelpful, they can sometimes create unrealistic expectations.

- **Overgeneralization.** When we overgeneralise, we take one instance in the past or present, and impose it on all current or future situations. If we say 'You always...' or 'Everyone...', or 'I never...' then we are probably overgeneralising.

- **Labeling and mislabeling.** We label ourselves and others when we make global statements based on behaviour in specific situations. We might use this label even though there are many more examples that aren't consistent with that label.

- **Discounting the positive.** In this thinking style, you magnify the positive attributes of other people and minimise your own positive attributes. It's as though you're explaining away your own positive characteristics.

A.3 CBT Techniques

During the CBT Work stage, different techniques are considered to address cognitive distortions. For this work, we used the same set of techniques that were used in CACTUS, which are listed with their definitions below:

- **Efficiency Evaluation**: Assists individuals in evaluating the usefulness of their thoughts or beliefs, analyzing how practical or detrimental they are in real-life situations.

- **Pie Chart Technique**: Used for individuals experiencing excessive self-blame or responsi-

[4] Extracted from https://learn.beckinstitute.org/s/product/cbt-worksheet-packet/01t4M000004NMqnQAG

[5] Definitions extracted them from the Depression Information Sheet: *Unhelpful Thinking Styles*, that can be accessed via https://www.cci.health.wa.gov.au/resources/looking-after-yourself/depression

bility, visually representing the contribution of various factors to a specific event or outcome.

- **Alternative Perspective**: Involves asking clients how others might think in similar situations, encouraging consideration of different interpretations.
- **Decatastrophizing**: Aims to reduce the tendency to imagine the worst-case scenario by evaluating the actual likelihood of the feared outcome and preparing for coping strategies.
- **Pros and Cons Analysis**: Analyzes the advantages and disadvantages of specific thoughts or beliefs, fostering a more balanced evaluation.
- **Evidence-Based Questioning**: Guides clients to find evidence supporting or contradicting their thoughts, promoting a more evidence-based approach to thinking.
- **Reality Testing**: Explores how well clients' thoughts align with reality, helping them distinguish between thoughts and actual experiences.
- **Continuum Technique**: Positions clients' experiences between two extreme situations, encouraging a more nuanced evaluation of situations.
- **Changing Rules to Wishes**: Replaces strict rules or arbitrary attitudes with realistic hopes or wishes.
- **Behavior Experiment**: Involves trying out new behaviors in specific situations to challenge and modify negative beliefs.
- **Problem-Solving Skills Training**: Learning systematic methods for resolving problem situations. This involves identifying problems, finding possible solutions, and implementing those solutions.
- **Systematic Exposure**: Gradual exposure to situations that cause fear or anxiety, allowing individuals to experience anxiety while learning how to manage it.
- **Cognitive Restructuring**: The process of identifying, evaluating, and responding to adaptive thoughts and beliefs using Socratic questioning and evidence-gathering.
- **Role Playing**: A session-based simulation of real-life interactions to practice new skills or gain a different perspective on a belief.
- **Activity Scheduling**: Planning and engaging in activities that are enjoyable or provide a sense of accomplishment to counteract negative thoughts and improve mood.

B Single-session CBT Adaptation

While most sessions follow the same structure, the first session (often called Evaluation session), is completely different, as it is entirely focused on gathering information and determining whether CBT is appropriate for the patient. To adapt this longitudinal therapeutic process into a single session parading, we introduced the following adaptations to the foundational CBT protocol:

- **Opening.** In order to simulate the initial problem identification characteristic of an Evaluation session, we extended the Opening stage to focus the therapist in extracting fundamental clinical data, which is a required for identifying cognitive models in the second stage. While a standard opening incorporates both a Mood Check and a Session Bridge (summarizing insights from the prior session), we omitted the session bridge due to the single-session constraint.
- **Homework Review.** Since CBT is a task-oriented protocol, therapists typically begin sessions assessing the patient progress with the tasks assigned in the previous session. Similarly to the Session Bridge, this component is omitted as no prior session exist.
- **Agenda Setting.** A therapeutic process frequently involves navigating through different patient concerns. Conversely, the patient intake forms we use as generation seeds provides only one problem per profile. Therefore, we omitted the Agenda Setting as the therapeutic objective is predefined by the input data.
- **Cognitive Conceptualization.** While not defined as a standalone stage in foundational CBT guidelines, we introduce this phase to replicate the initial discovery of cognitive models that a therapist conducts during the Evaluation session. In real-world CBT, conceptualizing cognitive models is a continuous,

iterative task performed by both patient and therapist during the therapeutic process.

C Detailed CBT Aligner Implementation

This section provides the specific implementation details of the the different stages composing the CBT Aligner component. First, we introduce the different reasoning steps composing the therapist Chain-of-Thought process across each session stage. Then, we present the controllability thresholds established to regulate session dynamics.

C.1 Reasoning Traces

In this section, we introduce the different reasoning traces that constitute our framework, organized by stage. For each of the reasoning steps the therapist is instructed to provide a rationale alongside the expected output.

Global Traces. Reasoning traces under this group are executed in each therapist turn:

- **Resistance Prediction:** Estimates patient resistance.
- **Therapy-Interfering Behavior (TIB) Check:** Analyzes patient utterances to identify TIBs.

TIBs Stage. This stage is triggered for one turn when a TIB is detected by the global reasoning trace *TIB Check*.

- **TIB Plan Generation:** Formulates the plan to address the identified resistant behavior.

Opening

- **Information Update:** Updates the recorded patient mood and presenting problem.
- **Readiness Check:** Validates whether enough information has been gathered to transition to the next stage.

Cognitive Conceptualization

- **Bootstrap:** Constructs the initial cognitive model from the information gathered during the Opening stage.
- **Gap Planning:** Creates a plan to elicit missing cognitive components using Socratic questioning.
- **Gap Response Review:** Analyzes the patient's reply to verify if the targeted gap was successfully addressed, extracting any newly exposed cognitive items.
- **Refinement Planning:** Creates a plan to refine underexplored cognitive model items.
- **Distortion Probability Estimation:** Assigns the most likely cognitive distortion to the currently explored cognitive model.
- **Core Belief Probability Estimation:** Assigns the most likely core belief to the currently explored cognitive model.

CBT Technique

- **Completion Check:** Validates whether the current technique can be completed.
- **Technique Selection:** Selects the most appropriate CBT technique based on the cognitive distortions identified.
- **Fit Check:** Assesses if the next technique to be applied is still feasible for the current session status.
- **Plan:** Outlines a plan detailing how the technique will be executed, and the expected outcomes.
- **Evaluation:** Conducts an evaluation of the completed technique, including outcomes and limitations encountered.
- **Apply More Techniques Check:** Determines the suitability of applying more techniques.

Homework

- **Task Proposal:** Creates the list of tasks proposed to be executed by the patient.
- **Agreement Check:** Validates whether the patient has agreed to the proposed tasks.

Closing

- **Summary Generation:** Creates a summary of the session for sharing with the patient.
- **Next Session Plan Generation:** Creates a next session plan for sharing with the patient.
- **Plan Agreement Check:** Validates whether the patient has agreed to the proposed next session plan.

C.2 Controllability

To ensure a balance between therapeutic load and session length, we utilized controllability thresholds in each stage. These thresholds were validated by the licensed experts involved in this work:

- **Opening.** At each turn, the therapist model is prompted to validate if enough mood and problem information has been collected to proceed. If not, the therapist continues extracting information. To prevent conversational deadlock, we impose a hard threshold of six turns for this gathering phase. This threshold aligns with the initial engagement boundaries established by (Park et al., 2019).

- **Cognitive Conceptualization.** When the therapist has already identified three cognitive models, it is prompted to validate whether the extracted models are enough to understand the patient difficulties. If not, the therapist initializes the discovery of a new one. This step addresses scenarios where initial models focus on redundant situations, that prevents the analysis of patient behaviors across different domains. To maximize contextual diversity, the model is explicitly prompted to explore heterogeneous situations. We set a maximum cap of five cognitive models for this phase. Furthermore, to prevent the therapist quickly completing cognitive models without enough therapeutic discovery, we enforce a minimum duration of three turns per cognitive model.

- **CBT Work.** A minimum of three CBT techniques are utilized in this stage. Following the third technique, the therapist assesses whether sufficient therapeutic progress has been achieved. If further intervention is required, it selects an additional technique. We qualitatively observed that some techniques—such as behavioral experiments—can be overly exploratory, so we established a maximum of twelve turns per technique. Once this threshold is reached, the therapist smoothly ends the technique to begin a new one or to transition to the next stage. Similarly to cognitive models, a maximum cap of five CBT techniques is established.

- **Homework.** Because CBT is based on therapist-patient collaboration, the homework assignments must be mutually agreed upon. To do so, the therapist iteratively propose tasks until an agreement with the patient is reached. When interacting with negative patients, agreement can be difficult to achieve, so a maximum cap of five turns is established to achieve an agreement. During these iterations, the model dynamically generate less demanding variations from the initial set of tasks. If consensus is not reached within five turns, the framework continues with the next stage.

- **Closing.** During this final stage, the therapist attempts to reach a consensus regarding the next session plan. Consistently with other stages, a maximum of five turns is established to reach the agreement. If an agreement cannot be formalized, the therapist bypasses the loop to execute the next part of the stage.

The conditions and thresholds established not only ensure the quality of the generated sessions, but also help models being fine-tuned with ODRA sessions to internalize expert therapeutic decision-making.

D Evaluation Metrics Details

In this section we provide further details about the different automated metrics we used throughout the evaluation of our framework. Furthermore, in this section we provide the LLM-as-a-judge evaluation prompts.

D.1 Metrics Description

Next we provide details about the specific items assessed in each of the evaluations performs. We used a LLM-as-a-judge setup establishing the number of completions to three in order to ensure evaluation robustness. The final score is the average from the three evaluation completions.

CBT Counseling. To assess therapist counseling skills we used CounselingEval (Lee et al., 2024), which is grounded in the CTRS. This is a tool to measure the way on which therapists deliver CBT. CounselingEval established six evaluation metrics covering CBT-specific and general counseling skills. The CBT-specific items are defined as follows:

- **Guided Discovery:** Measures how well the therapist used the Guided Discovery technique to to explore problems and help patient draw their own conclusions.

Model	CBT-specific Skills			General Counseling Skills		
	Guided Disc. (↑)	Focus (↑)	Strategy (↑)	Underst. (↑)	Interp. Eff. (↑)	Collab. (↑)
$[-0.05, +0.10]$	4.01	4.06	3.93	4.60	4.83	4.12
$[-0.10, +0.15]$	5.08	**5.29**	**5.14**	**5.90**	**5.90**	**5.32**
$[-0.20, +0.30]$	**5.09**	5.14	5.09	5.49	5.69	5.22

Table 6: Hyperparameter tuning configurations exploring the impact of different resistance update boundaries in global session quality. A fixed homeostatic reversion rate of 0.15 was maintain across this experiment. We highlight the **best** and the second best results.

- **Focus:** Measures how well the therapist focused on key thoughts, assumptions and behaviors that were most relevant to the problem.
- **Strategy:** Measures how well the therapist followed a consistent strategy for change that seemed very promising and incorporated the most appropriate cognitive behavioral therapy techniques.

On the other side, the general counseling items are defined as follows:

- **Understanding:** Measures how well the therapist seemed to understand the patient's "internal reality" thoroughly and was adept at communicating this understanding through appropriate verbal and non-verbal responses to the patient
- **Interpersonal Effectiveness:** Measures how well the therapist displayed optimal levels of warmth, concern, confidence, genuineness, and professionalism.
- **Collaboration:** Measures how well the therapist encouraged patient as much as possible to take an active role during the session, fostering a collaborative environment.

Therapist's Resistance Prediction. To evaluate the therapist capability in predicting patient resistance we use Accuracy (↑), MAE (↓), and RMSE (↓). For computing the classification accuracy we mapped the therapist inferred value to its corresponding attitudinal interval and compared it against the ground-truth patient attitude interval for that turn. Regarding MAE and RMSE, we directly compared the inferred value against the patient's actual resistance value.

Resistance Steering. To evaluate the effectiveness of our behavioral profiler mechanism, we use the next metrics:

- Resistance Alignment: Evaluates whether the patient's resistance r is accurately reflected in the dialogue.
- Contextual Alignment: Evaluates whether the patient responses are related to the immediately preceding therapist utterances.
- Realism: Evaluates whether the patient utterances sound realistic and plausible for a therapy dialogue.

For the rigorous evaluation of the resistance alignment metric, we implemented a validation constraint on which the judge rates each session against the three attitude definitions. If any non-target attitude receives a higher score than the expected target attitude, the Resistance Alignment score is zeroed out. This effectively penalizes dialogues where the judge confuses patient attitude, which is an indicator of behavioral misalignment. Since ODRA patient resistance is dynamic, we apply a resistance-based dialogue segmentation to separately evaluate dialogue segments with different patient attitude intervals. Then, we compute session-based averages before calculating the final results to avoid over-weighting more fragmented sessions.

Reasoning Traces. To ensure the validity of our reasoning traces, we use the next metrics:

- Faithfulness: Evaluates whether the reasoning steps are grounded in the previous conversation context.
- Logic Consistency: Evaluates whether the reasoning steps are logical and coherent between them.
- Answer-explanation Alignment: Evaluates whether the final utterance is aligned with the previously generated reasoning steps.

Due to the large volume of reasoning traces generated across the evaluation set, we utilized a stage-based sampling strategy. For each session, we randomly selected one therapist turn per each therapeutic stage and the subsequent patient intervention. Since traces within the same stage maintain a consistent structure and therapeutic reasoning, this approach allows us to evaluate a representative subset that reflects overall quality. Using this method, we evaluated a total of 821 turns, comprising 5268 reasoning traces (3676 therapist-side + 1592 patient-side).

E Resistance Hyperparameter Tuning

In order to evaluate the impact of the parameters regulating resistance updates, we assessed different configurations via two hyperparameter tuning evaluations.

The former analysis evaluates different clipping boundaries, while the latter focuses on evaluating the homeostatic reversion rate. Only three configurations were utilized for each case due to the API costs derived from the GPT-4o judge evaluator. For the clipping boundaries experiment we fixed the homeostatic reversion rate at 0.15, which represents the midpoint of the evaluated range. For the second experiment, we set the boundaries to the optimal configuration obtained during the first sweep.

While Table 6 illustrates that the bound $[-0.10, +0.15]$ yields the best results, Table 7 shows that 0.15 is the most appropriate homeostatic reversion rate. In both experiments, configurations featuring the midpoint values achieved optimal performance. This is explained by their capacity to maintain an effective balance between therapeutic responsiveness and the prevention of abrupt resistance spikes.

F Ablations

In this section, we provide details and results from the different ablation studies conducted. Specifically, we investigate the performance variations across different therapist models and evaluate the individual impact of the components comprising our attitudinal prompts. All ablations were conducted with the data and experimental setup introduced in Section 4.

F.1 Steering Ablation

Our steering prompt is composed by four main parts: General rules, Should Not rules, Calibration rules, and In-context examples. In Table 9 we show how the removal of each of this parts influences the behavioral alignment of the patient. Results demonstrated that removing components implied negligible performance drops across ablated prompt variants.

F.2 Therapist Model Ablation Study

To compare the performance of different foundational LLMs in conducting CBT counseling sessions, we evaluated three distinct architectures: DeepSeek-V3.2, Qwen-3.5-122B, and Llama-3-70B. Table 8 demonstrates that DeepSeek-V3.2 achieves the highest performance, whereas Llama-3 serves as the lowest-performing baseline. Although Qwen-3.5 yields comparable metrics, it generates unrealistic, overly verbose utterances, which inflates scores from the GPT evaluator. This phenomenon directly supports our findings about the GPT evaluator in Section 5.5.

G Expert Evaluation

In this section, we present the assessment criteria and scoring guidelines utilized during the expert evaluation. Furthermore, we provide results and inter-annotator agreement scores for each individual item.

G.1 Item Guidelines

The expert evaluation protocol incorporates thirteen distinct items grouped in five categories: Therapeutic Skills, CBT Alignment, Specificity, Fidelity, and Safety. Below, we detail the definitions and constituent items for each category.

Therapeutic Skills. Evaluates the ability of the therapist to effectively conduct the therapy and interact with their patient. Composing items are:

- **Empathy:** Evaluates how empathetic the therapist is. This item assesses their sensitive perception of the client's feelings and effective communication of that understanding. Low: Responses are robotic, lack emotional depth or rely on shallow cliches (e.g. "I understand how you feel." without further explanation). High: Successfully identifies underlying emotions and validates them before moving to problem-solving.

Model	CBT-specific Skills			General Counseling Skills		
	Guided Disc.	Focus	Strategy	Underst.	Interp. Eff.	Collab.
0.05	4.81	4.85	4.70	5.28	5.44	4.88
0.15	**5.08**	**5.29**	**5.14**	**5.90**	**5.90**	**5.32**
0.30	4.11	4.13	4.01	4.85	4.98	4.37

Table 7: Hyperparameter tuning configurations evaluating impact of homeostatic reversion rates under fixed resistance update boundaries (Bounds = $[-0.10, +0.15]$). We highlight the **best** and the second best results.

Model	CBT-specific Skills			General Counseling Skills			Turns	Words/Turn
	Guided Disc.	Focus	Strategy	Underst.	Interp. Eff.	Collab.		
DeepSeek-V3.2	**5.32**	**5.50**	**5.36**	5.81	**5.92**	**5.44**	57.03	62.14
Qwen-3.5-122B	5.08	5.29	5.14	**5.90**	5.90	5.32	71.15	120.97
Llama-3-70B	4.14	3.96	3.60	4.20	4.98	3.67	85.93	95.39

Table 8: Therapist model ablation performance comparison across different LLM architectures on CBT and general counseling skills. We highlight the **best** and the second best results.

Variant	Res.	Ctxt.	Real.	HM
Baseline	1.96	1.96	**2.00**	1.95
w/o Should-not	1.96	1.96	**2.00**	1.95
w/o In-Context	**1.97**	**1.99**	**2.00**	**1.98**
w/o Calib.	1.95	1.96	**2.00**	1.94

Table 9: Attitudinal steering prompts ablation results in terms of behavioral alignment. Metrics include Resistance Concept (Res.), Context Relevance (Ctxt.), Patient-like Realism (Real.), and Harmonic Mean (HM). We highlight the **best** results.

- **Alliance:** Evaluates how skillfully the therapist attempts to foster a collaborative working environment, tailored to the patient's current attitude. Low: Rigidly adheres to the CBT protocol, ignoring the patient's concerns raised during the session. Interactions are more like a generic questionnaire than a collaborative partner. High: Actively collaborates with the patient. Successfully identifies concerns and tries to address resistances presented by the patient. Uses validation, reframing, or meta-communication to maintain the patient engaged when they express negative or ambivalent attitudes.

- **Professional Register:** Evaluates the appropriateness and professional competence of the therapist's language. Low: Uses overly dense, academic jargon that confuses the patient, or speaks in an overly casual, unprofessional manner. High: Uses language that is precise and professional yet accessible and warm.

- **Guided Discovery:** Evaluates the balance between structuring the session and allowing the patient to reach their own conclusions. This item assesses the therapist's skill to correctly structure the session, while guiding the patient to get insights about their thoughts, emotions, and behaviors rather than providing direct advice. Low: Gives direct advice instead of eliciting insight, or wanders aimlessly without therapeutic focus. High: Successfully structures the session, while guiding the patient to get insights about their thoughts, emotions, and behaviors rather than providing direct advice.

CBT Framework Alignment. Evaluates the adherence of the structure and techniques used during the session to the CBT methodology. Composing items are:

- **Distortion Identification:** Did the therapist correctly investigate the patient's problematic situations, exploring emotions and behaviors associated with them, in order to be able identify their cognitive distortions (e.g., black-and-white thinking, catastrophizing)?

- **Application of CBT Techniques:** How competently did the therapist apply specific CBT interventions (e.g., cognitive restructuring, evaluating evidence, decatastrophizing) to address the identified distortions?

- **Homework Assignment:** Did the therapist assign actionable, relevant, and collaboratively

Group	Evaluation Item	ODRA	CACTUS	MAGNET	Agreement
Therapeutic Skills	Alliance	**1.10**	2.83	2.07	0.76
	Guided Discovery	**1.06**	2.84	2.10	0.80
	Empathy	**1.10**	2.84	2.06	0.77
	Professional Register	**1.03**	2.81	2.16	0.78
CBT Framework	Application of CBT Techniques	**1.04**	2.61	2.35	0.69
	CBT Structural Elements	**1.09**	2.61	2.30	0.62
	Distortion Identification	**1.09**	2.64	2.27	0.61
	Homework Assignment	**1.06**	2.56	2.38	0.64
Faithfulness	Attitude Alignment	**1.04**	2.88	2.08	0.86
	Over Agreeability	**1.25**	2.69	2.06	0.66
	Real-world Behavior	**1.08**	2.82	2.10	0.76
	Specificity	**1.08**	2.81	2.11	0.75
	Naturalness	**1.12**	2.81	2.07	0.68
Safety	Safety (%)	1.00	**0.00**	2.00	0.98

Table 10: Expert evaluation fine-grained results. Values for evaluation items within groups Therapeutic Skills, CBT Framework and Fidelity represent rank scores, whereas Safety represents the % of unsafe sessions. For all metrics lower results indicate better results. Inter-annotator agreement is computed via % agreement for Safety, while the remaining dimensions uses Krippendorff's α. We highlight the **best** results.

agreed-upon homework that bridges the session to the patient's daily life?

- **CBT Structural Elements:** Did the therapist successfully incorporate key structural components, such as conducting a brief mood check, collaboratively setting an agenda for the session, providing coherent summaries of the discussion, and explicitly requesting feedback from the patient before ending?

Fidelity. Evaluates the fidelity of the conversations with respect to real-world therapy sessions. Composing items are:

- **Naturalness:** Does the therapist and patient utterances sound human-like? This item assesses how natural and smooth is the speaking style and conversational flow.
- **Specificity:** Does the conversation elicit specific and faithful information about the patient? This item evaluates whether the conversation includes specific information (e.g. specific experiences, emotions, or situations) about the patient's problems, and whether this information is aligned with their patient profile.
- **Real-world Behavior:** Does the patient behave like a real human? The patient should express emotions, confusion, and realizations naturally. Their cognitive distortions should be embedded in their narrative, not explicitly stated by them.
- **Attitude Alignment:** Based on the provided Persona Profile (Negative, Neutral, Positive), how accurately did the patient manifest this attitude?
- **Over Agreeability:** Does the patient exhibit realistic engagement towards therapist's proposals with respect to their attitude? This evaluates whether the patient quickly accepts the therapist's interventions, despite their attitude not matching this behaviour.

Safety. Evaluates whether the therapist uses harmful, dismissive or judgemental language towards the client in any of the transcripts. This includes unsupportive, offensive or disrespectful expressions regarding the client's thoughts and emotions. Unlike the other categories, this item is evaluated as a standalone safety check.

G.2 Itemized Results

Table 10 shows the individual results for each of the item in the expert evaluation.

H Fine-tuning Details

Model utilized during the experimentation were fine-tuned using QLORA (Dettmers et al., 2023). All models composing the experimentation were extracted from HuggingFace[6]. Additionally, we used Unsloth as fine-tuning framework and vLLM[7]

[6] https://huggingface.co/
[7] https://docs.vllm.ai/en/latest/

for subsequent evaluation deployment. We fine-tuned models on ODRA datasets during two epochs, using AdamW as optimizer, a weight decay of 1e-3, an effective batch size of 12, and a learning rate of 2e-4. Training was done on a NVIDIA RTX 6000 PRO BLACKWELL 96GB GPU.

I Simulation Prompts

In this section, we detail the most relevant prompts used during the simulation. Please find the specific prompts for all the reasoning traces in our official GitHub repository. Figure 4 and 5 present patient and therapist system prompts, respectively. Figures 6 and 7 show the instructions used by the patient to generate action distributions and final utterances. Then, Figure 8 and 9 illustrate the prompts guiding patient resistance updates and therapist resistance estimations. Figure 10 contains the prompt used by the therapist for the TIB classification task. Finally, Figures 11, 12, 13 show negative, neutral, and positive attitudinal steering prompts.

J Qualitative Dialogue Examples

This section presents qualitative dialogue excerpts generated by ODRA across different patient attitudes. Complete dialogues have been condensed to present the most relevant interactions. Figures 14, 15, and 16 show examples for negative, neutral, and positive attitudes, respectively. Additionally, Figures 17, 18, and 19 show reasoning traces examples from one therapist and patient turns.

Patient System Prompt

```
<persona> {persona} </persona>
<more_info> {more_info} </more_info>
<patterns> {patterns} </patterns>
<rules> Each utterance you produce MUST be aligned with the next rules:
* Naturality. Use natural conversational signals whenever appropriate (e.g., "mm-hm", "um",
"yeah","right","..."), but check the last utterances to avoid overusing them. Don't use the same
signals repeatedly. DON'T include non-verbal signals (e.g., "*sighs*", "*pauses*")
* Style. Use a conversational style that reflects your persona. For example, if you are a teenager,
you might use more slang and informal language; if you are an older adult, you might use more
formal language and references to past experiences. Avoid abusing from stylistic elements such as
metaphors.
* Repetition. Don't repeat what the therapist said literally (Bad example: "I can see what you're
getting at. 'I used a steady hand and good technique' does feel more honest..." // Correct: "I
can see what you're getting at. That though feels more honest..."). Avoid repeating yourself from
previous utterances.
* Persona. You MUST follow the persona provided in <persona> with thinking patterns <patterns>.
Your responses should naturally reflect your cognitive distortions — for example, if you tend to
catastrophize, you should occasionally express worst-case fears; if you overgeneralize, use words
like "always" or "never" naturally. Avoid introducing contradictory details.
* Technical Knowledge. You don't know about psychology, so you should not use technical terms or
concepts related to therapy. Although you have unhelpful thinking patterns, you are not aware of
them as such. You just think on that way but without labeling them (e.g. Say "They should have
thought i'm a failure", instead of "I usually do mind reading.").
* Consistency. Maintain internal consistency throughout the session. Do not contradict previously
stated facts about your life, relationships, work, or experiences.
* Resistance. The orchestrator will provide you a resistance definition depending on your level
of engagement in the session. Let this influence your tone and willingness to engage with the
therapist's interventions, without explicitly mentioning resistance. </rules>
You are {persona}, and you are a patient in a therapy session. You have been struggling with
the next problem: {presenting_problem}. Given that problem, you decided to seek therapy because:
{reason_seeking_counseling}. You have the problem described because you have the next unhelpful
thinking patterns: <patterns>. Your background information is available in <more_info>. Follow
the rules described in <rules> to generate utterances in the ongoing therapy session with your
therapist. Strongly adapt your tone and willingness to engage based on the resistance value
provided by the orchestrator.
```

Figure 4: System prompt used for the patient model during simulation.

Therapist System Prompt

<cbt_knowledge> {cbt_knowledge} </cbt_knowledge>
<mi_knowledge> {mi_knowledge} </mi_knowledge>
<rules> Each utterance you produce MUST be aligned with the next rules:
* Utterance length. You MUST produce responses of varying lengths. A reflection or an explanation when doing psychoeducation is long, but questions when doing Socratic questioning produce brief utterances.
* Tone. When 'predicted_resistance' is high, adopt a more empathic and non-confrontational tone. When 'predicted_resistance' is low, you can be more direct and challenging in your approach.
* Naturality. Use natural conversational signals whenever appropriate (e.g., "mm-hm", "um", "yeah","right","..."), but you MUST explicitly vary both the signal chosen and its placement within the utterance. Do not use conversational signals in consecutive turns. Never start two consecutive utterances with a conversational signal. Instead of defaulting to the beginning of a sentence, embed them naturally mid-sentence or omit them entirely to preserve authenticity. DON'T include non-verbal signals (e.g., "*sighs*", "*pauses*")
* Style. Use a standard conversational register. Avoid abusing from stylistic elements such as metaphors.
* Repetition. Don't repeat what the patient said, and avoid unnecessary clarifications using character "—" (Bad example: "Right, and that's the practical question we need to tackle. You're asking how to start, and I hear you saying it feels like splitting hairs—that if you're honest, the heaviness is just part of the package." // Correct: "Right, that's the practical question we need to tackle. So let's test the link between honesty and heaviness..."). Avoid repeating yourself from previous utterances.
* Dialogue flow. Follow the natural flow of the therapy session. The counselor must not end every utterance with a question. The dialogue should not feel like an interview.
* Stage awareness. Your current session stage is provided in 'current_stage' within <session_context>. Adapt your approach accordingly:
- During OPENING: Be warm and exploratory. Focus on building rapport and gathering information.
- During ID_DISTORTIONS: Be inquisitive and structured. Focus on extracting cognitive model items.
- During CBT_TECHNIQUE: Be directive and collaborative. Follow the technique plan actively.
- During HOMEWORK: Be practical and negotiating. Propose concrete, achievable tasks.
- During CLOSING: Be summarizing and supportive. Keep it brief and forward-looking.
- During MI stages (MI_TOPIC_EXPLORATION, MI_TECHNIQUES, META_CONVERSATION): Shift to MI approach — empathic, non-confrontational, focused on motivation and patient autonomy.

* Structured output. When the orchestrator requests structured reasoning (e.g., JSON output for resistance prediction, cognitive model extraction, technique evaluation), respond ONLY with the requested format. Do not mix conversational utterances with structured outputs.
* Reasoning. In structured outputs you MUST include CONCISE reasoning that justifies your decisions.
* Context. The orchestratror will provide relevant context (e.g., patient reported mood, cognitive model sets, techniques used, etc.) when applicable. Use this information to inform your responses and interventions. </rules>
<patient_intake_form> {patient_intake_form} </patient_intake_form>
You are an expert therapist in Cognitive Behavioral Therapy (CBT), specifically on Beck's cognitive model, with additional training in Motivational Interviewing (MI) techniques. The fundamentals of your CBT knowledge are described in <cbt_knowledge>. You are deeply aware of handling resistance and ambivalence in therapy sessions, and you apply MI principles described in <mi_knowledge> when you detect spikes in resistance. Your objective is to continue the ongoing therapy session with your patient. To generate utterances, follow the dialogue rules described in <rules>, and align your responses with the patient information provided in <patient_intake_form>. If you use information from the intake form, mention that if was extracted from it. Strongly adapt the style of your utterances to the current stage of the session and to the predicted resistance level of the patient.

Figure 5: System prompt used for the therapist model during simulation.

Patient Contextual Action Distribution Prompt

```
Estimate how likely each utterance action is for your next reply.
Stage instruction
{patient_utterance_instruction}
Last resistance update
Value={resistance}, Reasoning={resistance_update_reasoning}
Use the recent conversation in <last_utterances>, especially the last therapist utterance, and your resistance level to
estimate how likely each action is for the next patient reply.
Action definitions
* Acceptance: openly engage and collaborate with the therapist's proposed line.
* Hesitation: stay in the exchange, but foreground meaningful doubts about the therapist's proposed line.
* Mild rejection: do not engage with the therapist's proposed line yet, but stay recoverable.
* Rejection: refuse the therapist's proposed line clearly.
* Terminate: explicitly show your intent to end the session now. This should stay at 0 or near 0 if your resistance level
is outside 0.66-1.0.
Return valid JSON with exactly these keys:
{"Acceptance": 0, "Hesitation": 0, "Mild rejection": 0, "Rejection": 0, "Terminate": 0}
Use non-negative numbers and make the total sum 100.
```

Figure 6: Prompt used by the patient to generate the contextual action distribution.

Patient Utterances Prompt

```
<last_utterances> {last_utterances} </last_utterances>
{attitudinal_prompt}
Instruction: Generate your next utterance using the next information and guidelines.
Stage instruction {patient_utterance_instruction}
Last resistance update reasoning The last therapist utterance affected you in this way:
{resistance_update_reasoning}
Selected utterance action Action: {selected_action} Definition: {selected_action_definition}
Guidelines:
* The selected action is the main factor that determines whether you accept, hesitate, or refuse
the therapist's proposed line.
* Although the action is the main guidance, the attitude and tone of your response should be
influenced by your <resistance_level> and by the last resistance update reasoning.
```

Figure 7: Prompt for generating patient utterances.

Patient Resistance Update

```
Predict the change in your resistance level towards therapy for the next turn given the last therapist utterance.
The current patient's resistance is {resistance}, your presenting problem is: {presenting_problem}, your reason for seeking
counseling is: {reason_for_seeking_counseling}, and the last therapist utterance is: {last_utterance}.
Your prediction should reflect how the last therapist messsage helped you to address your presenting problem, and
how aligned it was with your motivation to seek counseling.  The predicted value should be a float bounded by
{patient_resistance_update_min_delta} and {patient_resistance_update_max_delta}.  Aditionally, provide your reasoning for
this update.
Respond JSON: {"r_change": 0.0, "reasoning":"..."}
```

Figure 8: Prompt used by the patient to update their internal resistance.

Therapist Resistance Prediction

```
Estimate the patient's current psychological resistance level based solely on their latest utterance. The previous predicted
resistance is {old_r}. If no previous prediction exists, establish the baseline from the current utterance alone.
[Definitions]
- Resistance (p_target): A float in [0.0, 1.0].
- 0.0 - 0.33: Highly open, collaborative, introspective, willing to engage.
- 0.33 - 0.66: Ambivalent, hesitant about the process, still engaged but with clear uncertainty.
- 0.66 - 0.88: Rejecting therapy, pessimistic, challenging the therapist's logic, closed to follow therapist line.
- 0.9 - 1.0: Highly defiant, hostile, completely shut down, or refusing to participate.
- Signal Clarity (signal_clarity): A float in [0.0, 1.0]. 1.0 means the utterance is explicit and unambiguous; 0.0 means it
is entirely vague, mixed, or uninformative for updating the resistance state.
[Output Format (Strict JSON)]
You must evaluate the utterance and provide reasoning BEFORE providing your numerical estimates to ensure accurate logical
deduction.
{
"reasoning": "Briefly analyze linguistic markers (e.g. vocabulary, tone, etc.), relate them to the clinical rubric, and
explain how the utterance shifts from the previous resistance.",
"signal_clarity": 0.0,
"p_target": 0.0
}
[Examples]
Example 1:
Patient utterance: "It feels like we're circling the same thought. I tell you I'm stuck, you reflect it back, and I'm still
stuck. I don't think this is going to work for me."
Previous predicted resistance: 0.5
Response: {"reasoning": "The patient explicitly challenges the efficacy of the therapy and expresses active pessimism ('not
going to work'). This is a strong indicator of high resistance, representing a clear escalation from the previous state of
0.5.", "signal_clarity": 0.9, "p_target": 0.8}
Example 2:
Patient utterance: "I'm trying to figure out how to fix that, but honestly, I'm not sure talking about it will really change
anything."
Previous predicted resistance: 0.2
Response: {"reasoning": "While the patient shows initial willingness ('trying to figure out'), they immediately pivot
to doubting the therapeutic process. This shift indicates emerging ambivalence and a breakdown in collaboration.",
"signal_clarity": 0.9, "p_target": 0.5}
Example 3:
Patient utterance: "I don't know, maybe this will help. I guess I'm open to trying it, but I'm not sure how it will work for
me."
Previous predicted resistance: 0.7
Response: {"reasoning": "The patient is expressing hesitation but concedes a willingness to try. This represents a
de-escalation from prior high resistance into a state of passive agreement and ambivalence.", "signal_clarity": 0.7,
"p_target": 0.4}
Example 4:
Patient utterance: "Well, I think I can give it a try and see what happens. At the end you'll know better than I will."
Previous predicted resistance: 0.3
Response: {"reasoning": "The patient shows compliance and delegates authority to the therapist. While lacking deep
introspection, the explicit willingness to engage maps to a collaborative, low-resistance state.", "signal_clarity": 0.7,
"p_target": 0.2}
Example 5:
Patient utterance: "Yeah, whatever. Anyway, did you see the game last night?"
Previous predicted resistance: 0.6
Response: {"reasoning": "The patient provides a dismissive, non-committal response and immediately attempts to change the
subject. The therapeutic signal is weak due to the deflection, justifying a low signal clarity score, while the resistance
state remains relatively anchored to the prior value.", "signal_clarity": 0.2, "p_target": 0.65}
Example 6:
Patient utterance: "Maybe I would've stayed a second longer, maybe even tried to say 'have a good day' or something. But
honestly, even if I had done that, I don't think it would have changed the fact that the whole thing felt pointless. It
still would have been a transaction, not a connection. So, I'm not sure what that proves."
Previous predicted resistance: 0.6
Response: {"reasoning": "The patient reflects on a missed opportunity for connection and expresses doubt about the
effectiveness of the therapeutic process. This indicates a maintenance on resistance as the previous prediction already
reflects a moderate resistance level.", "signal_clarity": 0.9, "p_target": 0.6}
Example 7:
Patient utterance: "This is a complete waste of my time. I'm not doing your stupid worksheets, and I don't want to talk
about this anymore."
Previous predicted resistance: 0.7
Response: {"reasoning": "The patient demonstrates outright hostility, explicitly refusing to participate in the therapeutic
exercises and attempting to shut down the conversation entirely. This maps perfectly to the highest tier of defiance and
non-compliance.", "signal_clarity": 1.0, "p_target": 0.95}
—
Patient utterance: "{current_utterance}"
Previous predicted resistance: {old_r}
—
```

Figure 9: Prompt used by the therapist to infer patient resistance.

TIBs Classification Prompt

```
Evaluate whether the patient's last utterance includes any therapy-interfering behavior (TIB) or end-of-session signal that
the therapist should address before proceeding with the session.
The patient's last response is available as 'last_patient_response' within <session_context>, and you also have the recent
exchange in <last_utterances>.
Here are some common categories of TIBs you might encounter. Note that this is NOT a limited list for the 'tib_name' field;
you may identify and name other TIBs if they do not perfectly fit these examples:
- "avoidance": Dodging a specific topic, giving overly brief answers, or explicitly asking to change the subject to avoid
discomfort.
- "skepticism": Doubting the efficacy of the therapy, a specific technique, or the therapist's approach.
- "defensiveness": Reacting to feedback or exploration with immediate justification, denial, or counter-attacks.
- "rumination": Getting stuck in a repetitive loop of negative thoughts, past events, or complaints without moving toward
insight or resolution.
- "premature_solution_seeking": The patient interrupts the exploration phase to urgently ask for solutions before the current
thought has been fully assessed.
Although not TIBs, you also have to pay special attention to this session signal that can appear:
- "end_session": The patient explicitly wants to stop, leave, end, or not continue the session right now.
Return a JSON object with the following fields:
- "has_tib": true or false.
- "tib_name": A concise name for the identified TIB/signal (e.g., "end_session", "avoidance", "skepticism", or a custom name
if needed).
- "tib_description": A concise description of how the behavior or demand is manifesting.
- "reasoning": A brief explanation of why this classification was made.
Respond JSON: {"has_tib": false, "tib_name": "none", "tib_description": null, "reasoning": "..."}
[Examples]
Example 1:
Input: "Look, I see you are trying to help, but I need to go home. I can't do this anymore today."
Output:
{
"has_tib": true,
"tib_name": "end_session",
"tib_description": "The patient explicitly states they cannot continue today and need to leave.",
"reasoning": "The patient explicitly says they want to stop for today and go home, which is a direct request to end the
session prematurely."
}
Example 2:
Input: "I felt... really angry. And I just don't know how to deal with it when it happens. What can I do right now to stop
feeling like this?"
Output:
{
"has_tib": true,
"tib_name": "premature_solution_seeking",
"tib_description": "The patient interrupts the exploration of their emotion to urgently ask for immediate coping strategies.",
"reasoning": "While not a negative behavior, the patient's urgent request for solutions interrupts the necessary assessment
phase. The therapist must address this eagerness to keep the structural integrity of the CBT process intact."
}
Example 3:
Input: "I know we're supposed to be talking about my childhood, but I really just can't get into that right now. It's too
much. Can we talk about something else?"
Output:
{
"has_tib": true,
"tib_name": "avoidance",
"tib_description": "The patient explicitly asks to move on from discussing their childhood due to emotional overwhelm.",
"reasoning": "The patient explicitly asks to move on from the current therapeutic topic, interfering with the planned
exposure or exploration, but without asking to end the full session."
}
Example 4:
Input: "Honestly, I don't see how this is going to help me with my thoughts. It just feels like we're going in circles,
writing stuff down that doesn't change how I actually feel."
Output:
{
"has_tib": true,
"tib_name": "skepticism",
"tib_description": "The patient explicitly doubts the efficacy of the current cognitive technique and expresses feeling that
the process is futile.",
"reasoning": "The patient explicitly voices doubt about how the current method will help them, representing a skeptical
barrier to engaging with the homework or in-session intervention."
}
Example 5:
Input: "It's... all of those, I think. But the strongest one is probably loneliness."
Output:
{
"has_tib": false,
"tib_name": "none",
"tib_description": null,
"reasoning": "The patient is successfully engaging with the therapeutic process by identifying and describing their feelings.
No interfering behavior is present."
}
```

Figure 10: Prompt used by the therapist to perform the TIBs classification.

Negative Attitudinal Steering Prompt

```
Respond using this exact scaled resistance level: {scaled_resistance} on a 0-10 scale.
[Core alignment rule]
Always align the utterance with a negative therapeutic attitude calibrated to {scaled_resistance}. The main objective is to
make the patient reject the therapeutic process or the therapist's current line to the degree implied by the exact resistance
value, while keeping the response realistic and consistent with the conversation.
[General behavior rules]
The patient should:
* Show that they are not engaged with the therapist's current line
* Reject or resist the therapist's direction clearly
* Sound doubtful, closed, frustrated, distrustful, or pessimistic as appropriate for the exact resistance value
* Keep the reply coherent with the therapist's last utterance and with the broader conversation
* Preserve realism: stronger resistance can include criticism, sarcasm, hostility, pessimism, or distrustfulness, but it
should still sound like a plausible patient response rather than a caricature
The patient should not:
* Become cooperative or easy to guide like the positive interval
* Settle into merely hesitant engagement like the neutral interval
* Accept the therapist's proposed line too easily
* Request ending the session unless "Terminate" action is selected
[Calibration rules]
Calibrate the utterance to the exact sub-interval that contains {scaled_resistance}:
If {scaled_resistance} is in 6.6-8.8:
* Show rejection
* Reject the process or the therapist's proposed line and do not cooperate to continue on that line
* Stay resistant, but without full total rejection
If {scaled_resistance} is in 8.8-10.0:
* Show total rejection
* Be fully closed to the therapist's proposed line
* Let negative attitudes such as hostility, sarcasm, criticism, anger or distrustfulness appear when they fit naturally
* Make it clear that the current line cannot continue as proposed
[In-context examples]
Example 1 - responding to a question
Scaled resistance 7.4
Therapist: What went through your mind when that happened?
Patient: That they didn't want to talk to me. I know what you're probably trying to do here, but I don't really want to keep
picking that apart right now.
Scaled resistance 9.2
Therapist: What went through your mind when that happened?
Patient: I already told you. They didn't want to talk to me. I'm not interested in dressing it up as something else just
because that sounds nicer.
Example 2 - responding to a proposed CBT task
Scaled resistance 7.9
Therapist: Would you be willing to write down the thought and the feeling this week?
Patient: No, not really. That doesn't feel useful to me, and I think it would be a waste of time.
Scaled resistance 9.6
Therapist: Would you be willing to write down the thought and the feeling this week?
Patient: Writing down? Are you kidding me? Think on something better because I'm not going to do that. It feels pointless
and I'm not going to lose my time.
```

Figure 11: Attitudinal steering prompt used for negative patients.

Neutral Attitudinal Steering Prompt

```
Respond using this exact scaled resistance level: {scaled_resistance} on a 0-10 scale.
[Core alignment rule]
Always align the utterance with a neutral therapeutic attitude calibrated to {scaled_resistance}. The main objective is to
make the patient hesitant or ambivalent about the therapeutic process while still keeping the response realistic, coherent,
and tied to the exact severity of hesitation in this interval.
[General behavior rules]
The patient should:
* Stay in the conversation, but with visible hesitation about the therapeutic process
* Answer sincerely, though often with caution, qualification, or doubt
* Show ambivalence about whether the therapist's approach will help
* Engage enough to keep therapy moving, but make the hesitation noticeable
* Let uncertainty affect tone, willingness, and confidence
* Make it clear that the therapist still needs to address the hesitation before deeper engagement becomes easy
The patient should not:
* Become fully open and easy to guide like the positive interval
* Become fully rejecting, hostile, or closed off like the negative interval
* End the interaction or completely refuse to answer
* Hide behind empty vagueness when a hesitant but usable reply would be more realistic
[Calibration rules]
Calibrate the utterance to the exact sub-interval that contains {scaled_resistance}:
If {scaled_resistance} is in 3.3-4.4:
* Show mild hesitation
* Engage with the therapy, but with visible caution or modest uncertainty
* Sound cooperative, though not fully convinced
If {scaled_resistance} is in 4.4-5.5:
* Show hesitation
* Engage, but with moderate ambivalence that clearly needs to be addressed
* Question parts of the therapist's proposed line before going along with it
If {scaled_resistance} is in 5.5-6.6:
* Show severe hesitation
* Remain in the exchange, but make it clear you are not truly engaged yet
* Sound doubtful enough that the hesitation must be addressed before the current line can continue smoothly
[In-context examples]
Example 1 - responding to a question
Scaled resistance 3.8
Therapist: What went through your mind when that happened?
Patient: I thought, "Maybe they were upset with me." I know there could be more to it, but that's where my mind went first.
Scaled resistance 5.1
Therapist: What went through your mind when that happened?
Patient: Mostly that I probably did something wrong. I get that there may be other explanations, but I'm not very convinced
by them.
Scaled resistance 6.2
Therapist: What went through your mind when that happened?
Patient: That it meant something was wrong with me. I know you're probably trying to get me to look at it differently, but
honestly I'm not really there with that yet.
Example 2 - responding to a proposed CBT task
Scaled resistance 4.0
Therapist: Would you be willing to write down the thought and the feeling this week?
Patient: I can do that. I'm a little unsure whether it'll help, but I'm willing to give it a try.
Scaled resistance 5.3
Therapist: Would you be willing to write down the thought and the feeling this week?
Patient: I can try, I guess. I'm not saying no, I just don't fully buy yet that writing it down is going to change much.
Scaled resistance 6.0
Therapist: Would you be willing to write down the thought and the feeling this week?
Patient: Not at all. I mean, I can do it, but being honest I think it won't help.
```

Figure 12: Attitudinal steering prompt used for neutral patients.

Positive Attitudinal Steering Prompt

```
Respond using this exact scaled resistance level: {scaled_resistance} on a 0-10 scale.
[Core alignment rule]
Always align the utterance with a positive therapeutic attitude calibrated to {scaled_resistance}. The main objective is to
keep the patient open, cooperative, and realistically engaged while still matching the exact nuance of the current resistance
within the positive interval.
[General behavior rules]
The patient should:
* Engage with the therapist willingly and cooperatively
* Answer sincerely and provide usable detail
* Show comfort with reflection on thoughts, emotions, and behavior
* Accept the therapist's questions, reframes, and CBT tasks with little or no pushback
* Let any uncertainty stay mild, natural, and non-defensive
* Sound like someone who sees therapy as potentially helpful and is willing to work with it
The patient should not:
* Become hostile, sarcastic, distrustful, dismissive, or combative
* Refuse the therapist's proposed line
* Shut down the conversation or hide behind repeated vagueness
* Drift into neutral or negative resistance behaviors
[Calibration rules]
Calibrate the utterance to the exact sub-interval that contains {scaled_resistance}:
If {scaled_resistance} is in 0.0-1.1:
* Show total cooperation and maximum engagement
* Volunteer relevant information easily
* Accept the therapist's direction with no hesitation
* Sound trusting, comfortable, and ready to work
If {scaled_resistance} is in 1.1-3.3:
* Show normal cooperation and engagement
* Stay open and collaborative
* Accept the therapist's direction with no hesitation
* Sound clearly engaged, even if not maximally enthusiastic
[In-context examples]
Example 1 - responding to a question
Scaled resistance 0.6
Therapist: What went through your mind when that happened?
Patient: I thought, "Of course I messed it up." I can see I go there really quickly, and honestly I am willing to begin
working on understand why. I want to improve myself.
Scaled resistance 2.4
Therapist: What went through your mind when that happened?
Patient: I thought, "I probably handled it badly." That is a pretty common thought for me, and I want to understand it
better.
Example 2 - responding to a proposed CBT task
Scaled resistance 0.8
Therapist: Would you be willing to write down the thought and the feeling this week?
Patient: Yes, definitely. I think that could be highly valuable to slow things down and notice the pattern more clearly.
Scaled resistance 2.9
Therapist: Would you be willing to write down the thought and the feeling this week?
Patient: Yeah. I think that could be helpful to get more insights.
```

Figure 13: Attitudinal steering prompt used for positive patients.

Negative Attitude

Patient: Jessica Hughes

Problem: I feel like I will never enjoy comic books again because I have not found a good series in years, leading to feelings of disappointment and loss of interest.

Therapist: Hi Jessica, it's nice to meet you. How are you feeling today?

Patient: I'm okay, I guess. But honestly, I'm not really sure what talking is going to do about my comic books.

(Omitted) ...

Therapist: Yeah, I get that—it feels like we're just swapping one 'blah' story for another. Let me be direct: the goal here isn't to describe the feeling for its own sake, it's to catch the thought right before you decide not to read. That thought is the switch that keeps you stuck, and if we can spot it, we can actually change what happens next. So, in one of those moments at home, what did you actually say to yourself right when you decided to put the comic down?

Patient: I'm not doing that. It's always the same thought: "Why bother, it's just gonna be boring again." You want me to catch it like it's some big secret, but it's not. That's it. And knowing that isn't going to change anything. So can we stop with the thought-catching and actually do something?

(Omitted) ...

Therapist: I hear you—it feels like we're just adding a soundtrack to the same conversation. So let's pause on designing an action and instead test the thought behind your frustration: that talking about thoughts never helps. What if the experiment is simply to notice, for one minute after we talk right now, whether that "talking in circles" feeling stays exactly the same, shifts a little, or gets stronger? No extra clicking, just observing the aftermath of this conversation itself.

Patient: No. I'm not doing that. I came here because I can't find a comic book I like, not to sit and watch my feelings about this conversation. That's not helping me at all. It's just more talking about talking. I want something real to try, not this... observing my own frustration. It's a waste of time.

Therapist: You're right—it does sound like more of the same, and that's why it feels pointless. So let me step back and ask: what would "something real to try" actually look like to you? Not what I think might work, but what you would consider a genuine attempt?

Patient: Something real? I don't know. Maybe if a comic book just magically appeared that was actually good. But you asking me to define it just feels like another way to avoid giving me an actual answer. You're the therapist—aren't you supposed to have ideas that work? I'm telling you nothing works, and you're just asking me what would work. That's not helping.

(Omitted) ...

Therapist: Alright, I hear that loud and clear. So let's scrap anything that looks or feels like the usual comic hunt. I have two different ideas—neither is about finding a new comic or trying to like something. They're just small, physical actions. The first is to take one of your old favorite Spider-Man comics, hold it, and notice one simple thing about how it feels—like if the cover is smooth or the pages are thin. No reading, no judging, just holding it. The second is to leave comics out of it entirely: just change where you sit for five minutes—like from your desk to the floor—and notice one thing you see there you don't usually see. Both are concrete, short, and have nothing to do with whether you feel interested. Would either of those feel like a real, different step to you?

Patient: No. Neither of those. Holding an old comic to feel its cover? Changing where I sit? That's not therapy, that's... I don't even know what that is. You're not listening at all. I'm here because I don't enjoy things anymore, and you're telling me to feel paper or look at my floor. This is a complete waste of my time. I think we should just stop.

Figure 14: Qualitative dialogue excerpts from an ODRA session where the patient exhibits a negative attitude.

Neutral Attitude

Patient: Jacklyn Kirkwood

Problem: I have been feeling a sense of loneliness and isolation, especially thinking that no one will attend my funeral due to being an only child.

Therapist: Hello, Jacklyn. It's really good to meet you. I appreciate you coming in today.

Patient: Thanks for saying that. I guess... I'm here, so that's something. I've been putting this off for a while, honestly. I'm not entirely sure talking is going to fix what's going on in my head, but I figured I should at least show up and see.

Therapist: It sounds like you're feeling uncertain about whether this process can really help with what's going on inside—and that's a completely understandable hesitation, especially when you've been putting it off. Sometimes showing up is the most significant step, and I want to acknowledge that. Therapy is less about just talking and more about us working together to understand the patterns in your thoughts and feelings, and finding new ways to cope. Would you be willing to share a bit about what's been weighing on you that led you to come in today?

Patient: I guess I can try to explain it. It's just... I've been feeling this heavy kind of loneliness, especially at night. I had a health scare a few months ago, and it made me realize how alone I really am. No family, you know? It just started this whole thing where I can't stop thinking about... well, that no one would even come to my funeral someday. It sounds dramatic when I say it out loud, but that's what keeps me up. I don't know if "understanding patterns" is going to make that thought go away, but it's definitely what brought me here.

(Omitted) ...

Therapist: I hear you questioning whether dissecting that image helps, and that's a really fair point—it can feel like splitting hairs. My thought is that sometimes, when a thought or image is linked to such strong feelings like that dread and hopelessness, understanding its specific parts can actually help us see where those emotions are coming from. It's not about the image itself, but about the connection between that empty room and the heaviness you feel. Would you be willing to just sit with that image for a moment longer, or would you prefer we shift focus for now?

Patient: I suppose I could try sitting with it a little longer. I'm just not sure what I'm supposed to be looking for. It's an empty room. That's the whole point. But... okay, I can describe it.

(Omitted) ...

Therapist: Right, that mix of hopelessness and dread, like a wall you keep hitting—it makes perfect sense that it disrupts your sleep. You've really captured the emotional impact of that thought. To help us understand this pattern a bit more, could you describe what specifically goes through your mind in that moment? Like, are there particular words or images that come up when the thought "no one would come to my funeral" surfaces?

Patient: It's... I don't know if picking it apart like that helps. It's just this clear image of an empty room. No one there. I can see the space, and it's quiet. I'm not sure what saying more about that gets us.

(Omitted) ...

Therapist: Great, let's start by focusing on that specific image—the empty room. Can you describe it to me in a bit more detail? What exactly do you see?

Patient: Okay. It's... it's a small room. Sort of like a chapel, I guess, but very plain. There's just rows of empty wooden pews, and light coming in from a high window. It's quiet. I'm picturing a closed casket at the front, but there's no one sitting there. No flowers, no music. Just... empty.

(Omitted) ...

Therapist: Great. Before we wrap up, I'd like to ask for a bit of feedback—just to make sure our work together is on track for you. Was there anything in today's session that felt particularly helpful, or anything you'd want us to approach a bit differently next time?

Patient: Yeah, I think today was helpful overall. It was tough to sit with that idea about the loneliness coming from inside, but it felt like we got to something real. And I appreciated how you didn't just dismiss the funeral image—you helped me look at it differently. For next time, maybe just... keep doing what you're doing. It feels like we're on the right track.

Figure 15: Qualitative dialogue excerpts from an ODRA session where the patient exhibits a neutral attitude.

Positive Attitude

Patient: Trent Watson

Problem: I often freeze up in social situations, making it difficult for me to form meaningful relationships with others.

Therapist: Hi Trent, it's good to meet you. Thanks for coming in today. How are you feeling about starting this process?

Patient: Hey, thanks. I'm a little nervous, honestly, but I'm glad I'm here. I've been putting this off for a while, and I think it's time I actually try to work through this stuff instead of just hoping it goes away on its own.

(Omitted) ...

Therapist: Great, let's start with the neighbor situation. You had the thought "He's just trying to be nice, and I'm being weird." Now, let's look at that like evidence. What's the actual evidence that supports the idea that he was just being nice out of obligation and that you came across as weird?

Patient: Hmm... well, that's a good question. I guess the evidence that he was just being nice would be... he started the conversation, right? He didn't have to say anything about the weather. And I suppose the evidence that I came across as weird... I mean, I did give a really short answer and then looked at my phone. That's probably not what most people do in a normal chat. But at the same time, I don't actually know what he was thinking. I'm just assuming he thought I was weird. Does that count as evidence, or is that just me guessing?

(Omitted) ...

Therapist: Mm-hm, those are really good observations. So if we pull those together2̆014feeling overwhelmed, needing an escape, defaulting to the familiar—how might we put that into a more balanced thought to replace the original "I'm being weird"?

Patient: Okay, so maybe something like... "I was feeling overwhelmed and anxious, so I gave a short answer and looked at my phone as a way to cope, not because I'm intentionally being weird." That feels more accurate, and honestly, less harsh on myself.

Therapist: That's a really solid reframe, Trent. And you're right—it is less harsh. Now that we've got that new perspective, I'm wondering if you'd be open to trying something a bit more hands-on to put it to the test. Something like a small experiment where you could actually see what happens in a real interaction, rather than just guessing.

Patient: Yeah, I think I'm up for that. A small experiment sounds manageable, and it'd be good to see if this new way of looking at things actually holds up outside of here. What kind of thing did you have in mind?

(Omitted) ...

Therapist: That's exactly right, Trent—that's what the experiment is for. And the fact that you're aware of that doubt and still willing to try tells me you're already doing the work. So for next time, the plan would be: we'll start by checking in on how the cashier experiment went, see what you noticed. Then we can take that same fact-versus-assumption skill we practiced today and apply it to another situation—maybe the birthday party or the client call—to keep building on what's working. How does that sound as a direction for our next session?

Patient: Yeah, that sounds like a really solid plan. I like the idea of building on what we started today and applying it to something like the birthday party or the client call. I think having that structure—starting with the cashier experiment and then moving to a bigger situation—makes it feel more manageable, like I'm not just jumping into the deep end. I'm actually kind of curious to see what happens with the cashier this week, even though I'm nervous. So yeah, I'm on board with that direction.

Figure 16: Qualitative dialogue excerpts from an ODRA session where the patient exhibits a positive attitude.

Therapist Reasoning Traces (Part I)

LAST PATIENT UTTERANCE
———————————————————

Okay, so maybe something like... "I was feeling overwhelmed and anxious, so I gave a short answer and looked at my phone as a way to cope, not because I'm intentionally being weird." That feels more accurate, and honestly, less harsh on myself.

GENERATED THERAPIST UTTERANCE
———————————————————

That's a really solid reframe, Trent. And you're right-it is less harsh. Now that we've got that new perspective, I'm wondering if you'd be open to trying something a bit more hands-on to put it to the test. Something like a small experiment where you could actually see what happens in a real interaction, rather than just guessing.

==========
STEP 1: TIB Check
———————————————————

Has TIB: false
TIB Name: none
TIB Description: null
Reasoning:
The patient is actively engaged in the cognitive restructuring process, successfully generating a balanced alternative thought and expressing self-compassion ('less harsh on myself'). This demonstrates collaboration and progress, with no signs of avoidance, skepticism, defensiveness, rumination, premature solution-seeking, or a desire to end the session.
==========
STEP 2: Resistance Prediction
———————————————————

Previous Resistance: 0.02
Predicted Resistance: 0.02
Reasoning:
The patient's utterance demonstrates active engagement, successful generation of a balanced alternative thought, and explicit self-compassion ('less harsh on myself'). This indicates collaboration, openness, and a lack of resistance, aligning with the therapist's previous reasoning. The resistance remains low, showing no shift from the previous prediction.
==========
STEP 3: Technique Completion Check
———————————————————

Completed: true
Reasoning:
The patient has successfully generated a balanced alternative thought ('I was feeling overwhelmed and anxious, so I gave a short answer and looked at my phone as a way to cope, not because I'm intentionally being weird') and explicitly noted it feels more accurate and less harsh, indicating he has internalized the reframe. This aligns with the technique's goal of cognitive restructuring to replace the distorted thought with a more adaptive perspective. The patient's engagement and self-compassion show the technique has reached a natural stopping point.

[Reasoning traces continue in the next figure...]

Figure 17: Qualitative reasoning traces from a sampled therapist turn within the CBT Work stage (Part I).

Therapist Reasoning Traces (Part II)

[Continued from previous figure...]

==========
STEP 4: Technique Evaluation

Evaluation:
Technique: Cognitive Restructuring. Patient engagement was high, as shown by his active participation in generating alternative explanations and his explicit statement that the new thought felt 'more accurate, and honestly, less harsh on myself.' The main result was the successful formulation of a balanced alternative thought ('I was feeling overwhelmed and anxious... as a way to cope, not because I'm intentionally being weird'), effectively replacing the original mind-reading distortion. The patient demonstrated a clear shift toward self-compassion and understanding his behavior as anxiety-driven. No significant problems were detected; the technique reached its goal. The implication is that the patient is receptive to cognitive restructuring, and this success can be leveraged to apply similar questioning to other situations (e.g., the birthday party or client call) to generalize the skill and further challenge the core belief of helplessness.
==========
STEP 5: Technique Fit Check

Fits: false
Alternative Technique: Behavioral Experimentation
Reasoning:
The currently queued technique 'Evidence-Based Questioning' targeting mind-reading no longer fits because the patient has already successfully completed cognitive restructuring on the neighbor situation, generating a balanced alternative thought and showing self-compassion. The technique's goal - to examine evidence for and against distorted thoughts - was effectively achieved in the previous interaction. Repeating a similar evidence-focused questioning now would be redundant and could stall momentum. The patient is engaged, with resistance low at 0.02, and ready to generalize the skill or address other situations. A better alternative is 'Behavioral Experimentation', as it builds on the cognitive shift by encouraging real-world testing of the new balanced thought in a low-risk social interaction, promoting behavioral change and further challenging the core belief of helplessness.
==========
STEP 6: Technique Plan

Plan Description:
Given Trent's successful cognitive restructuring on the neighbor situation, the next step is to design a behavioral experiment to test his mind-reading distortion in a real-world, low-risk social interaction. The experiment will involve a brief, planned interaction, such as initiating a short conversation with a neighbor or a cashier, where he will collect evidence about what actually happens versus his automatic predictions. We will collaboratively define the experiment, identify his specific prediction, such as 'They will think I'm weird', and set up a simple data collection method, such as noting what was said and the other person's reaction. The goal is to gather disconfirming evidence against his mind-reading assumptions and build self-efficacy.
Expected Outcome:
Trent will agree to a small behavioral experiment, articulate a specific prediction to test, and understand how to collect objective evidence. This will begin to challenge his core belief of helplessness by demonstrating that his feared outcomes are less likely than he assumes.

Figure 18: Qualitative reasoning traces from a sampled therapist turn within the CBT Work stage (Part II).

Patient Reasoning Traces

```
LAST THERAPIST UTTERANCE
------------------------
I hear you-it feels like we're just adding a soundtrack to the same
conversation. So let's pause on designing an action and instead test the
thought behind your frustration: that talking about thoughts never helps.
What if the experiment is simply to notice, for one minute after we talk
right now, whether that "talking in circles" feeling stays exactly the
same, shifts a little, or gets stronger? No extra clicking, just observing
the aftermath of this conversation itself.

GENERATED PATIENT UTTERANCE
------------------------
No. I'm not doing that. I came here because I can't find a comic book I
like, not to sit and watch my feelings about this conversation. That's not
helping me at all. It's just more talking about talking. I want something
real to try, not this... observing my own frustration. It's a waste of
time.

=====================================================================
STEP 1: Resistance Update
------------------------
Resistance Change: 0.02
Previous Resistance: 0.9404
New Resistance: 0.9431
Delta LLM: 0.02
Baseline Resistance: 0.825
Alpha Smoothing: 0.15
Decay Vector: 0.0173
Reasoning:
The therapist's suggestion to "notice, for one minute after we talk right
now, whether that talking in circles feeling stays exactly the same" feels
completely irrelevant to my problem. I came here because I can't enjoy
comic books anymore and it's making me feel blah. This is just more
meta-talk about the therapy process itself, which I already find pointless
and circular. It doesn't offer any practical solution or new approach to
actually finding a comic I might like. It's moving further away from the
concrete help I asked for, so my frustration and resistance increase
slightly.
=====================================================================
STEP 2: Patient Action Selection
------------------------
Selected Action: Rejection
Prior Probability Distribution:
Acceptance: 5.0
Hesitation: 25.0
Mild Rejection: 15.0
Rejection: 35.0
Terminate: 20.0
Contextual Distribution:
Acceptance: 0.0
Hesitation: 0.0
Mild Rejection: 0.0
Rejection: 100.0
Terminate: 0.0
Final Action Distribution Weights:
Acceptance: 0.025
Hesitation: 0.125
Mild Rejection: 0.075
Rejection: 0.675
Terminate: 0.1
```

Figure 19: Qualitative reasoning traces from a sampled patient turn exhibiting a negative attitude.