CCD-CBT: Multi-Agent Therapeutic Interaction for CBT Guided by Cognitive Conceptualization Diagram

Chang Liu[1], Changsheng Ma[1], Yongfeng Tao[1], Bin Hu[1], Minqiang Yang[1]*

[1]School of Information Science and Engineering, Lanzhou University

Abstract

Large language models show potential for scalable mental-health support by simulating Cognitive Behavioral Therapy (CBT) counselors. However, existing methods often rely on static cognitive profiles and omniscient single-agent simulation, failing to capture the dynamic, information-asymmetric nature of real therapy. We introduce CCD-CBT, a multi-agent framework that shifts CBT simulation along two axes: 1) from a static to a dynamically reconstructed Cognitive Conceptualization Diagram (CCD), updated by a dedicated Control Agent, and 2) from omniscient to information-asymmetric interaction, where the Therapist Agent must reason from inferred client states. We release CCDCHAT, a synthetic multi-turn CBT dataset generated under this framework. Evaluations with clinical scales and expert therapists show that models fine-tuned on CCDCHAT outperform strong baselines in both counseling fidelity and positive-affect enhancement, with ablations validating dynamic CCD guidance and diagnostic analysis supporting asymmetric interaction design. Our work offers a new paradigm for building theory-grounded, clinically-plausible conversational agents.

1 Introduction

Mental health disorders, particularly depression, constitute a pressing global burden, exacerbated by a critical shortage of accessible professional therapists (Huang et al., 2019). Large language models (LLMs) offer a promising avenue for scalable mental health support by powering conversational agents (Long et al., 2024; Lawrence et al., 2024). Among therapeutic approaches, Cognitive Behavioral Therapy (CBT) is a prime candidate for LLM integration due to its structured, evidence-based nature. Initial efforts have integrated psychological theories into LLMs to generate therapist-like responses (Mousavi et al., 2021; Das et al., 2022; Chen et al., 2024; Raile, 2024). However, as illustrated in Figure 1 (upper), most CBT-oriented systems remain superficial, they rely on prompt-based application of CBT techniques (Na, 2024; Xiao et al., 2024) or generate short dialogs in 2- to 3-turns that lack the depth and progression of authentic therapy (Qiu et al., 2024a).

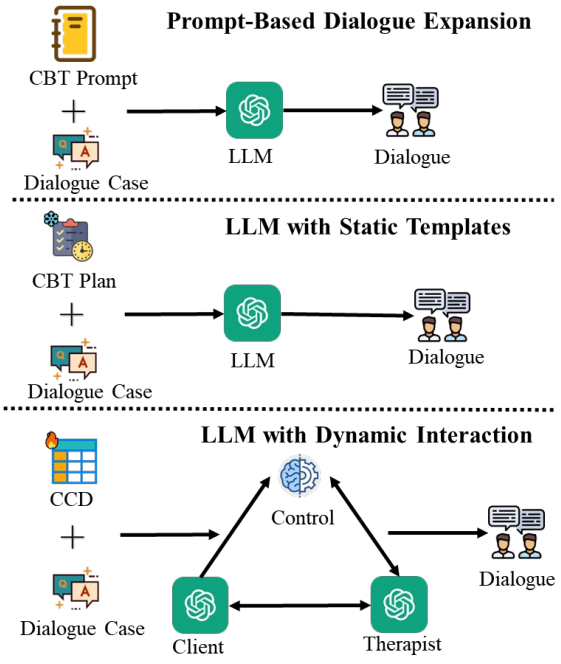


Figure 1: Comparison of CBT simulation paradigms: early prompt-based methods (upper), recent static cognitive modeling (middle), and our dynamic multi-agent framework (lower).

To achieve greater clinical fidelity, recent work has incorporated structured clinical tools to ground multi-turn dialogues. Systems like CACTUS (Lee et al., 2024) rely on static treatment plans, while DiaCBT (Zhou et al., 2025) explicitly adopts the Cognitive Conceptualization Diagram (CCD)—a core CBT tool for mapping a client's cognitive-affective structure (Beck, 2020). As depicted in Figure 1 (middle), these approaches represent an advance

*Corresponding author.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1723–1745
November 6-10, 2026 ©2026 Association for Computational Linguistics

by using structured cognitive profiles, yet they retain a fundamental limitation: they **predefine the global cognitive model prior to the session and keep it fixed throughout the interaction**, treating therapy as the execution of a static script. This contradicts a core tenet of CBT: the therapist **collaboratively constructs and refines the client's cognitive model during the session**, starting from limited knowledge and iteratively updating it based on client disclosures (Kuyken et al., 2011; Hofmann et al., 2012).

Furthermore, these systems predominantly **rely on an omniscient, single-agent simulation paradigm**, where one LLM generates the entire dialogue with full access to both client and therapist internal states (Lee et al., 2024; Zhou et al., 2025). As Zhou et al. (2024) critically demonstrate, such omniscient settings create "unrealistically successful" interactions that overestimate social goal achievement and fail to generalize to realistic conditions of information asymmetry. In actual therapy, the therapist has **no privileged access to the client's beliefs** and must infer their internal state through active listening and Socratic questioning. The prevailing paradigm thus simulates a distorted reality, lacking the very inferential and adaptive challenges that define therapeutic expertise.

These dual limitations—static cognitive profiles and omniscient simulation—collectively point to a fundamental **representational mismatch** between existing computational frameworks and the dynamic, dyadic nature of CBT. To bridge this gap, we introduce CCD-CBT, a novel multi-agent framework that instantiates a role-specialized architecture, as visualized in Figure 1 (lower): a Client Agent (simulating the individual with ground-truth internal states), a Therapist Agent (generating responses based on provided strategy), and a Control Agent (dynamically managing the CCD and therapeutic strategy). This architectural separation enables two critical paradigm shifts: 1. **From Static to Dynamic CCD:** The Control Agent actively reconstructs the client's CCD in real-time from the dialogue, transforming it from a static profile into a living state variable that guides the therapy. 2. **From Omniscient to Asymmetric Simulation:** The Therapist Agent operates without direct access to the client's ground-truth internal state, relying solely on the Control Agent's inferred CCD. This enforces the information asymmetry inherent in real therapy, requiring the system to develop genuine clinical reasoning through inference.

Based on this framework, we construct and release CCDCHAT, a high-fidelity, multi-turn synthetic CBT dialogue dataset. We evaluate models fine-tuned on CCDCHAT through a dual-pronged strategy incorporating both automatic evaluation using adapted clinical scales and expert human evaluation by licensed CBT practitioners. Results demonstrate that our approach not only outperforms existing baselines in therapeutic skill metrics but also achieves superior outcomes in enhancing clients' positive affect. Crucially, ablation studies validate dynamic CCD guidance, while diagnostic analysis supports the role of information-asymmetric interaction in clinically plausible simulation. Our contributions are threefold:

- We propose CCD-CBT, a novel multi-agent framework that pioneers dynamic CCD reconstruction and information-asymmetric simulation for generating theory-grounded CBT dialogues.
- We release CCDCHAT, a publicly available dataset that captures the collaborative, evolving nature of CBT through our novel simulation paradigm.
- We conduct comprehensive evaluations through both automatic and expert human assessments, showing our method generates more effective and authentic therapeutic interactions than current approaches.

2 Related Work

2.1 CBT-based Conversational Systems

Early LLM-based counseling agents relied on fine-tuning or prompting models on therapy transcripts, producing responses that mimic superficial conversational patterns without grounding in established therapeutic theory or strategy (Liu et al., 2023; Qiu et al., 2024a). To incorporate clinical structure, recent work infuses CBT techniques via prompting (Na, 2024) or designs multi-stage assessment pipelines (Xiao et al., 2024), yet remains limited to short, non-interactive exchanges. Truly multi-turn systems have since emerged by generating static treatment plans upfront (Lee et al., 2024), adapting linguistic styles (Xie et al., 2025), or simulating client behaviors by progressively unmasking a predefined cognitive diagram (Kim et al., 2025). However, these approaches operate from an omniscient perspective, overlooking the clinical reality

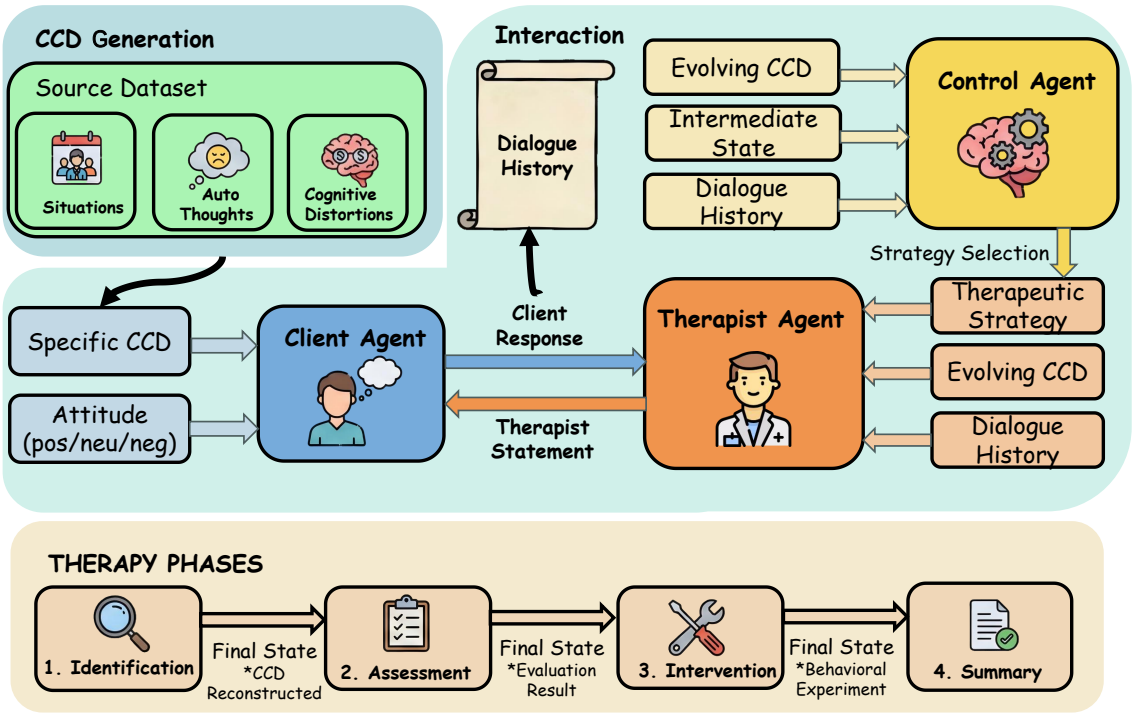


Figure 2: Overview of the CCD-CBT framework. The system consists of a Client Agent, a Therapist Agent and a Control Agent that interact based on the Specific Cognitive Conceptualization Diagram (CCD), guiding the dialogue through four therapeutic phases: Identification, Assessment, Intervention, and Summary. * denotes key states passed to the subsequent phase.

of information asymmetry (Beck, 2020). Concurrent work (Liu et al., 2026) studies multi-session CBT process modeling; our work focuses on single-session, multi-agent dynamic CCD reconstruction under information asymmetry.

2.2 CBT-Related Datasets

Existing datasets for computational CBT can be categorized into two main types. The first focuses on fine-grained cognitive-affective analysis, such as detecting and re-framing cognitive distortions (Sharma et al., 2023; Maddela et al., 2023) or mapping cognitive-behavioral links (Wang et al., 2023). While these resources support structured reasoning, they do not model multi-turn therapeutic dialogue. The second line aims to synthesize counseling-style conversations. Early datasets like Psych8K (Liu et al., 2023) contain transcribed sessions but lack explicit CBT annotation. Recent efforts incorporate CBT techniques more directly: CBT-LLM (Na, 2024) prompts LLMs to produce single-turn strategy-aligned responses, CACTUS (Lee et al., 2024) decouples treatment planning from dialogue generation, enabling diverse scenarios, and PsyDTCorpus (Xie et al., 2025) extends interactions by aligning techniques with varied counseling styles. However, these datasets are typically generated via pre-defined templates or rule-based simulation, which restricts their ability to reflect the dynamic, co-constructed progression of real therapy—where each response adapts to the client's immediate disclosure and the evolving case conceptualization. Our CCDCHAT dataset addresses this limitation by using a multi-agent generation framework that dynamically updates the client's cognitive model throughout the dialogue, producing interactive counseling sessions.

3 Method

We present CCD-CBT, a multi-agent framework for generating synthetic CBT counseling dialogues guided by a dynamically updated CCD. Unlike prior script-based approaches, CCD-CBT treats the CCD as an evolving internal state that drives clinical reasoning throughout the session. The framework decomposes the therapeutic process into three role-specific agents: a Client Agent, a Therapist Agent, and a Control Agent. A session begins with a client-specific CCD and follows the four standard CBT phases: *Identification*, *Assessment*, *Intervention*, and *Summary* (Beck, 2020). During dialogue,

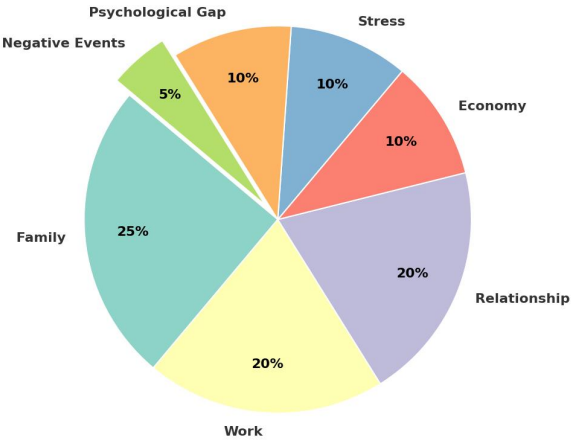


Figure 3: Distribution of situational contexts in CCD-CBT. The dataset covers seven life domains, with Family (25%) and Work (20%) being the most frequent triggers for psychological distress.

the Control Agent continually refines the CCD and supplies phase-aware therapeutic strategies to the Therapist Agent, enabling adaptive, multi-turn interactions that mirror the responsiveness of real therapy. Figure 2 illustrates the overall architecture.

3.1 Framework Construction

Source Data & CCD Construction. To ground the simulation in clinically plausible cognitive-affective patterns, we construct 7,500 distinct CCDs using the Chinese CBT resource C2D2 (Wang et al., 2023). C2D2 provides a wide spectrum of everyday scenarios, negative automatic thoughts, and categorized cognitive distortions, as shown in Figure 3. Each CCD serves as the ground-truth cognitive profile of a synthetic client. This structured profile is the seed from which all subsequent interaction emerges; it provides the Client Agent with a consistent internal state, while serving only as the evaluation target for assessing the Control Agent's reconstructed therapist-side CCD.

Client Agent. The Client Agent embodies the individual seeking therapy. To enhance the realism of client responses, we refer to Wang et al. (2024) approach. Its responses are generated autoregressively by prompting an LLM with three inputs: (1) the client's ground-truth $\text{CCD}_i^{\text{client}}$, (2) an assigned attitudinal stance $a_i \in \{\text{positive, neutral, negative}\}$ that modulates engagement and resistance, and (3) the dialogue history $\mathcal{H}_{t-1}$. Formally,

$$r_t = \mathcal{LLM}(\text{CCD}_i^{\text{client}}, a_i, \mathcal{H}_{t-1}), \quad (1)$$

Where t denotes the current dialogue turn. This design ensures that client utterances remain coherent with their underlying cognitive model while exhibiting realistic, stance-dependent variation in cooperativeness and emotional tone. Implementation prompts are detailed in Appendix B.2.

Therapist Agent. The Therapist Agent generates therapeutic responses by integrating in-session inferences rather than ground-truth client information. At each turn t, it receives (1) the therapist-side $\text{CCD}_i^{\text{therapist}}$—a dynamically updated version of the client's cognitive model reconstructed by the Control Agent, (2) the current CBT strategy π_t issued by the Control Agent, and (3) the dialogue history $\mathcal{H}_{t-1}$. The generation is formulated as:

$$s_t = \mathcal{LLM}(\text{CCD}_i^{\text{therapist}}, \pi_t, \mathcal{H}_{t-1}). \quad (2)$$

The output s_t is a clinically intentional utterance aimed at techniques such as guided discovery or cognitive restructuring, thereby simulating a therapist who intervenes based on evolving conceptualization, not a fixed script. Prompt details appear in Appendix B.3.

Control Agent. The Control Agent functions as the clinical reasoning engine of the framework, actively managing the progression of therapy through the four structured CBT phases: *Identification*, *Assessment*, *Intervention*, and *Summary*. It maintains two state variables: the therapist-side CCD ($\text{CCD}_i^{\text{therapist}}$) and a structured phase tracker $\mathbf{z}_t$. At every turn, it analyses the dialogue history $\mathcal{H}_{t-1}$ and current state to produce the next therapeutic strategy:

$$\pi_t = \mathcal{LLM}(\text{CCD}_i^{\text{therapist}}, \mathbf{z}_t, \mathcal{H}_{t-1}). \quad (3)$$

Crucially, the agent actively reasons about phase progression. For example, during the *Identification* phase, $\mathbf{z}_t$ follows an 8-step clinical procedure, as client disclosures arrive, the agent updates $\text{CCD}_i^{\text{therapist}}$ and advances $\mathbf{z}_t$. Upon reaching a phase-specific terminal state (e.g., completing all eight identification steps), it autonomously triggers a transition to the next phase. This mechanism ensures that the therapy follows a structurally sound

Datasets	Language	Available	Strategy	CBT-based	CCD-guided	Avg. turns
Psych8k (Liu et al., 2023)	English	✓	✗	✗	✗	1.0
SmileChat (Qiu et al., 2024a)	Chinese	✓	✗	✗	✗	10.4
HealMe (Xiao et al., 2024)	English	✗	✓	✓	✗	3.0
CBT-LLM (Na, 2024)	Chinese	✓	✓	✓	✗	1.0
CACTUS (Lee et al., 2024)	English	✓	✓	✓	✗	16.6
PsyDTCorpus (Xie et al., 2025)	Chinese	✓	✓	✓	✗	18.0
CCDCHAT (Ours)	Chinese	✓	✓	✓	✓	16.2

Table 1: Comparison of existing mental health dialogue datasets. ✓ indicates support or availability; ✗ indicates not supported or unavailable.

yet adaptive trajectory, mirroring the way a human therapist collaboratively builds and refines a case formulation in real time. Full prompt specifications are provided in Appendix B.4

3.2 Dataset: CCDCHAT

We synthesize the CCDCHAT dataset by instantiating the CCD-CBT framework: all three agents are implemented with GPT-4o (Hurst et al., 2024). From the constructed CCD pool, we select 1,500 client CCDs for dialogue generation. Every CCD is paired with a Client Agent configured under three attitudinal stances (positive, neutral, negative), producing an initial set of 3 candidate dialogues.

To ensure high clinical plausibility, we apply a two-stage quality-control protocol. First, we discard dialogues exhibiting inconsistent formatting, premature termination, or insufficient turn counts. Second, we conduct an automated competence screening using a proxy of the Cognitive Therapy Rating Scale (CTRS) (Blackburn et al., 2001; Creed et al., 2016). Six core CTRS items are scored (0–6) per dialogue via LLM-based judgment. Dialogues whose mean CTRS score falls below a strict threshold of 4 are excluded. After filtering, 98% of the candidate sessions are retained, and the remaining failed cases are regenerated until they satisfy the quality-control criteria. This process results in the final CCDCHAT corpus of 4,500 curated, multi-turn CBT sessions. More details are provided in Appendix E.1.

As shown in Table 1, CCDCHAT is characterized by its (1) explicit CBT grounding, (2) end-to-end CCD guidance, and (3) support for clinically coherent, extended interactions. These properties establish it as a high-fidelity resource for studying and simulating CBT processes.

3.3 Model Training

We fine-tune a base LLM on CCDCHAT using Low-Rank Adaptation (LoRA) (Hu et al., 2022). Both the Therapist and Control Agents are optimized in a structured question-answering format: each turn is framed as a prompted generation task that incorporates role-specific instructions, the current CCD, dialogue history, and (for the Control Agent) phase state.

For the Therapist Agent, we train a single LoRA adapter over the full dialogue trajectory to maintain a consistent therapeutic style across phases. For the Control Agent, we train separate LoRA adapters for the four clinical phases.

During inference, the phase-specific output either produces the next therapeutic strategy π_t or triggers a transition to the next phase. Full hyperparameters, training schedules, and prompt templates are provided in Appendix D.1.

4 Experiments

4.1 Experimental Setup

Datasets. We fine-tune and evaluate models on the CCDCHAT dataset. For training, we sample 1,500 dialogues for each of the three client attitudinal stances, resulting in a total of 4500 training examples. And we use Qwen2.5-7B-Instruct (Qwen et al., 2024) as our base model. For evaluation, we randomly select 50 distinct CCDs from the remaining set of unused CCDs, each paired with three different attitudes, resulting in a total of 150 samples as a held-out validation set. The client agent is instantiated using GPT-4o-mini to simulate realistic client responses, enabling assessment of model generalization under clinically plausible conditions.

Baselines. We compare against two categories of strong baselines. General-counseling LLMs include: (1) SoulChat (Chen et al., 2023), fine-tuned for empathetic multi-turn support, (2) PsyChat (Qiu et al., 2024b), a client-centered system that recognizes behavior and selects

Models	General Counseling Skills			CBT-specific Skills		
	Understanding	Interpersonal Eff.	Collaboration	Guided Discovery	Focus	Strategy
PsyChat	4.50*	4.36*	4.48*	4.20*	4.19*	3.77*
SoulChat	4.32*	4.35*	4.42*	3.92*	4.13*	3.91*
MeChat	4.52*	4.44*	4.48*	3.98*	4.24*	4.06*
CAMEL	4.78	4.52	4.68*	4.50*	4.39*	4.16*
PsyDTLLM	4.88	4.68	4.83	4.61*	**4.57**	4.22*
Ours	**4.90**	**4.69**	**4.87**	**4.74**	**4.57**	**4.36**

Table 2: Comparison of general counseling and CBT-specific skills across models. The asterisk (*) next to values indicates a significant difference compared to Ours, with $p < 0.05$ as determined by the paired t-test.

Models	Positive Attitude		Neutral Attitude		Negative Attitude	
	positive ↑	negative ↓	positive ↑	negative ↓	positive ↑	negative ↓
PsyChat	+1.90*	-1.38	+1.88*	-1.82	+1.94*	**-1.84**
SoulChat	+1.94*	-1.74	+0.88*	-1.34	+0.06*	-0.24
MeChat	+2.66*	-1.56	+1.62*	-0.76*	+0.12*	-0.56
CAMEL	+2.22*	-0.46*	+0.40*	-1.90	+0.18*	-1.52
PsyDTLLM	+4.46*	-1.18	+3.80*	-1.50	+1.28*	-0.80
Ours	**+5.98**	**-2.50**	**+5.20**	**-1.92**	**+2.56**	-0.96

Table 3: Results of PANAS across various models. This table illustrates the shifts in positive and negative emotions before and after counseling sessions, categorized by the client's attitude. Higher positive values and lower negative values indicate better performance.

strategies, (3) MeChat (Qiu et al., 2024a), trained on the real-world-derived SmileChat corpus. CBT-specific models include: (1) CAMEL (Lee et al., 2024), a two-stage agent that first plans then generates CBT-aligned responses using the CACTUS dataset; (2) PsyDTLLM (Xie et al., 2025), which extends single-turn dialogues into multi-turn CBT interactions using PsyDTCorpus.

Evaluation Metrics. We employ two complementary clinical instruments, both scored automatically by GPT-4o-mini to ensure scalability and consistency. (1) CTRS (Blackburn et al., 2001) measures **clinical competence.** We adapt a streamlined CTRS covering three general counseling skills (Understanding, Interpersonal Effectiveness, Collaboration) and three CBT-specific skills (Guided Discovery, Focus, Strategy), each rated on a 1–6 Likert scale. Higher scores indicate greater therapeutic competence. (2) Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988) evaluates **emotional outcomes.** It quantifies pre- to post-session changes in client affect via 10 positive and 10 negative emotion items (1–5 scale). We infer the pre-session state from the initial CCD and predict the post-session state from the dialogue, computing score differences. Effective counseling should increase positive affect and decrease negative affect. Detailed prompt templates, scoring protocols, and validation of the automatic rater are provided in Appendix C.2 and C.3.

4.2 Main Results

Clinical Competence (CTRS). We evaluate dialogue models on six CTRS dimensions. As shown in Table 2, while all models perform reasonably on general counseling skills, scores drop noticeably on CBT-specific skills, with Strategy being the most challenging dimension. Our CCD-CBT model achieves the highest scores in five of the six dimensions, excelling in Understanding (4.90), Interpersonal Effectiveness (4.69), Collaboration (4.87), and Guided Discovery (4.74); it ties for the best score in Focus (4.57). Crucially, it attains a substantially higher Strategy score (4.36), outperforming all baselines and reflecting its stronger capacity for structured intervention and treatment-plan execution. Among baselines, CAMEL and PsyDTLLM show more balanced profiles, yet our model consistently surpasses them, especially in CBT-specific competencies. These results indicate that our framework delivers superior therapeutic competence, particularly in the core CBT skill of strategic intervention, where existing models struggle most.

Models	General Counseling Skills			CBT-specific Skills		
	Understanding	Interpersonal Eff.	Collaboration	Guided Discovery	Focus	Strategy
w/o CBT	4.42*	4.35*	4.54*	4.34*	4.34*	4.11*
CBT w/o CCD	4.71*	4.51	4.71	4.63*	4.47	4.23
CBT w/ CCD	**4.90**	**4.69**	**4.87**	**4.74**	**4.57**	**4.36**

Table 4: Ablation study evaluating the contribution of CBT integration and the control module.

Emotional Outcomes (PANAS). Table 3 reports changes in client affect (positive, negative) across three initial-attitude groups. Our model induces the most favorable positive-affect shift overall: it yields the largest increase in positive affect for Positive (+5.98), Neutral (+5.20), and Negative (+2.56) clients, and the greatest reduction in negative affect for Positive (-2.50) and Neutral (-1.92) groups. While its reduction for Negative clients is more modest (-0.96), it remains competitive but does not surpass all baselines, indicating that negatively engaged clients remain more challenging. In contrast, baseline models exhibit limited or inconsistent emotional regulation. For example, PsyChat reduces negative affect for Negative clients (-1.84) but shows weak positive-affect gains; CAMEL and SoulChat produce smaller changes overall, especially in challenging Neutral/Negative scenarios. Our model not only boosts positive emotions effectively but also achieves competitive mitigation of negative affect, demonstrating a well-rounded emotional benefit.

We further assess expert–LLM agreement on 50 randomly sampled dialogues. Kappa (Cohen, 1960) reaches $\kappa = 0.48$ for CTRS and $\kappa = 0.53$ for PANAS, indicating moderate consistency. Together, these results demonstrate that our framework not only masters procedural CBT skills with superior professionalism, but also yields more favorable simulated emotional shifts. The gains are especially pronounced in the most challenging dimensions—strategy and negative-affect reduction—highlighting the value of dynamic CCD guidance and multi-agent interaction for delivering responsive, evidence-based digital therapy.

4.3 Ablation Studies

We perform systematic ablations to isolate the contributions of the CCD-guided planning module and the CCDCHAT fine-tuning data.

CCD Ablation. We compare the full CCD-CBT model with a variant that retains the same training data but removes the CCD control mechanism (CBT w/o CCD). Results in Table 4 show that ablating CCD guidance leads to consistent performance drops across all six CTRS dimensions. The decline is most pronounced in CBT-specific skills, especially Strategy (4.36 → 4.23) and Guided Discovery (4.74 → 4.63), confirming that the dynamic CCD is essential for structured, goal-oriented therapeutic reasoning. Nevertheless, even without the CCD, the model still outperforms the base LLM in general counseling skills (Understanding 4.71 vs. 4.42), indicating that the CCDCHAT data alone enhances relational and conceptual alignment.

Data Ablation. We also evaluate the base LLM without any CCDCHAT fine-tuning (w/o CBT). Its scores fall substantially below both the ablated and full models. The gap is largest in CBT-specific dimensions: w/o CBT trails the full model by 0.25 points in Guided Discovery and 0.23 points in Focus, and achieves the overall lowest score in Strategy (4.11). This confirms that CCDCHAT provides high-fidelity behavioral and strategic patterns necessary for clinically competent dialogue.

Overall, the ablation results demonstrate that both components are indispensable and offer complementary benefits: fine-tuning on CCDCHAT establishes a foundation of counseling competency, while the CCD module further sharpens CBT-specific reasoning and strategic coherence. The progressive improvement from w/o CBT to CBT w/o CCD to the full model reflects a clear additive effect of data and structure in building effective, theory-grounded conversational agents.

4.4 Dataset Quality Evaluation

We conduct expert-based human evaluation to directly assess the clinical authenticity of generated dialogues. Following the protocol of Lee et al. (2024), two licensed CBT practitioners performed pairwise comparisons on 50 randomly sampled di-

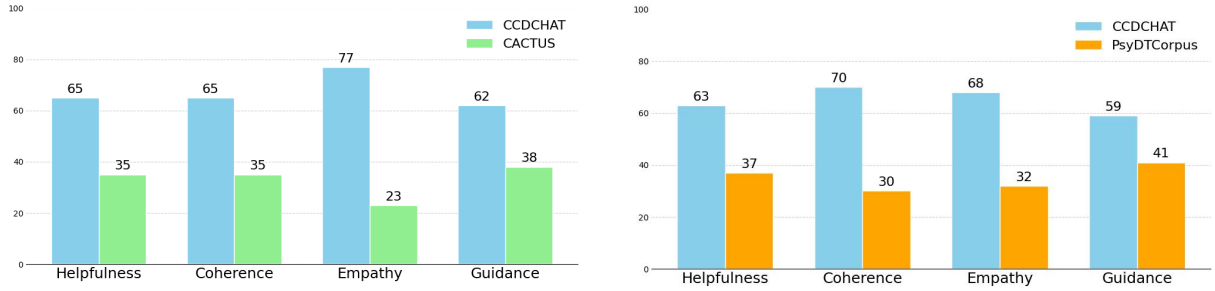


Figure 4: Human evaluation results comparing CCDCHAT with CACTUS and PsyDTCorpus across four dimensions. In 100 pairwise comparisons per dimension, CCDCHAT is consistently preferred over both baselines. All differences are statistically significant ($p < 0.05$) according to one-sided binomial tests.

alogues from each of CCDCHAT, CACTUS, and PsyDTCorpus. Each dialogue was rated across four dimensions: Helpfulness, Logical Coherence, Empathy, and Guidance. The average inter-rater agreement is $\kappa = 0.62$, indicating substantial consistency between the two expert raters.

As illustrated in Figure 4, CCDCHAT is consistently rated superior to both baseline datasets across all four criteria ($p < 0.05$). While the margin varies across dimensions, the advantage is statistically significant in each case, confirming the superior clinical plausibility of dialogues produced by our CCD-guided, multi-agent framework.

This result provides a data-level explanation for the performance advantages observed in Section 4.2: the higher fidelity of CCDCHAT dialogues furnishes models with more realistic interaction patterns, thereby enhancing both clinical competence and emotional outcomes. Complete evaluation guidelines are included in Appendix C.1.

4.5 Further Analysis

Client Attitude Effects. We analyze how client attitude influences simulated counseling quality. Dialogues with positive-attitude clients receive higher ratings on general counseling skills, likely because cooperative engagement smooths interaction flow. In contrast, neutral-attitude clients yield the best performance on CBT-specific skills, suggesting that balanced affect offers a clearer test-bed for applying structured techniques without emotional interference. Detailed per-attitude score breakdowns are provided in Appendix E.2.

CCD Reconstruction Fidelity. A core function of the Control Agent is to dynamically reconstruct the client's CCD during the *Identification* phase. We evaluate reconstruction accuracy by automatically comparing each inferred CCD component against its ground-truth source, scoring alignment on a 3-point scale (1: not accurate → 3: very accurate). As shown in Table 5, the reconstructed CCDs achieve high overall alignment (mean = 2.77), with particularly strong matches in Core Beliefs (2.71) and Coping Strategies (2.71). This confirms that the Control Agent can reliably infer clients' cognitive-affective patterns from dialogue, a prerequisite for CCD-guided intervention. The results validate the framework's capacity for data-driven conceptualization, underpinning its structured therapeutic reasoning.

Component	Not	Slightly	Very	Avg.
Overall	0	104	346	2.77
Situation	1	135	314	2.70
Emotions	3	152	295	2.65
Behaviors	1	108	341	2.76
Core Beliefs	1	129	320	2.71
Relevant History	2	169	279	2.62
Coping Strategies	1	129	320	2.71
Intermediate Beliefs	0	167	283	2.63
Automatic Thoughts	1	233	216	2.48

Table 5: CCD reconstruction similarity ratings across components (total $N = 450$).

5 Conclusion

We introduced CCD-CBT, a multi-agent framework that generates CBT dialogues by dynamically reconstructing a client's CCD during conversation, moving beyond static and omniscient simulation paradigms. The accompanying CCDCHAT dataset, validated by clinical experts, enables training of agents that significantly outperform existing baselines in both therapeutic skill and emotional outcomes. Our work demonstrates the value of theory-guided, adaptive interaction structures for building clinically-plausible conversational agents for mental health.

Limitations

Our study is subject to several limitations. First, our framework simulates therapy as a single extended dialogue, whereas real-world CBT typically spans multiple sessions. Second, the text-only nature of our implementation cannot capture nonverbal cues such as tone and gesture, which are important in real therapeutic interaction. Finally, the use of LLM-based judges may introduce bias and sensitivity; thus, our results mainly indicate improved simulated dialogue quality rather than real-world clinical effectiveness.

Ethical Considerations

Data Privacy. To ensure the highest level of privacy, all dialogues and CCDs within the CCD-CHAT dataset are either synthetically generated, ensuring that no personally identifiable information (PII) is present or can be reverse-engineered from the model's outputs.

Security. We acknowledge the inherent stochasticity and potential unpredictability of LLMs. Consequently, the framework is designed to be a supportive instrument used under the rigorous supervision of mental health professionals. Its clinical scope is deliberately restricted to assisting individuals with mild depression and anxiety, ensuring that high-risk or complex cases remain under direct human clinical management. The reported results should be interpreted as evidence of improved simulated counseling quality, rather than evidence of clinical efficacy for real users.

Bias. We acknowledge the potential for algorithmic and demographic bias. As our metadata and training corpora are predominantly rooted in a Chinese linguistic and cultural environment, the possibility of population-specific biases remains. The model may inadvertently reflect demographic distributions or social norms prevalent in the source data, which could impact the equity of its responses across different sub-populations.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No.62472203), in part by the STI 2030-Major Projects (2021ZD0202002), in part by the Fundamental Research Funds for the Central Universities (Grant No. lzujbky-2023-16). Supported by Supercomputing Center of Lanzhou University.

References

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

Ivy-Marie Blackburn, Ian A James, Derek L Milne, Chris Baker, Sally Standart, Anne Garland, and F Katharina Reichelt. 2001. The revised cognitive therapy scale (cts-r): psychometric properties. *Behavioural and cognitive psychotherapy*, 29(4):431–446.

Tiantian Chen, Ying Shen, Xuri Chen, and Lin Zhang. 2024. Psychatbot: a psychological counseling agent towards depressed chinese population based on cognitive behavioural therapy. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183, Singapore. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Torrey A Creed, Sarah A Frankel, Ramaris E German, Kelly L Green, Shari Jager-Hyman, Kristin P Taylor, Abby D Adler, Courtney B Wolk, Shannon W Stirman, Scott H Waltman, and others. 2016. Implementation of transdiagnostic cognitive therapy in community behavioral health: The beck community initiative. *Journal of consulting and clinical psychology*, 84(12):1116.

Avisha Das, Salih Selek, Alia R Warner, Xu Zuo, Yan Hu, Vipina Kuttichi Keloth, Jianfu Li, W Jim Zheng, and Hua Xu. 2022. Conversational bots for psychotherapy: a study of generative transformer models using domain-specific dialogues. In *Proceedings of the 21st workshop on biomedical language processing*, pages 285–297.

Stefan G Hofmann, Anu Asnaani, Imke JJ Vonk, Alice T Sawyer, and Angela Fang. 2012. The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive therapy and research*, 36(5):427–440.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yueqin Huang, YU Wang, Hong Wang, Zhaorui Liu, Xin Yu, Jie Yan, Yaqin Yu, Changgui Kou, Xiufeng

Xu, Jin Lu, and others. 2019. Prevalence of mental disorders in china: a cross-sectional epidemiological study. *The lancet psychiatry*, 6(3):211–224.

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

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

Willem Kuyken, Christine A Padesky, and Robert Dudley. 2011. *Collaborative case conceptualization: Working effectively with clients in cognitive-behavioral therapy*. Guilford Press.

Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479.

Suyeon Lee, Sunghwan Mac Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, and others. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274.

Chang Liu, Shuyi Zhang, Changsheng Ma, Yongfeng Tao, Minqiang Yang, and Bin Hu. 2026. Dmt-cbt: Longitudinal therapeutic state modeling for cbt counseling. *Preprint*, arXiv:2606.03132.

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

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

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.

Seyed Mahed Mousavi, Alessandra Cervone, Morena Danieli, and Giuseppe Riccardi. 2021. Would you like to tell me more? generating a corpus of psychotherapy dialogues. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 1–9.

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

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024a. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636.

Huachuan Qiu, Anqi Li, Lizhi Ma, and Zhenzhong Lan. 2024b. Psychat: A client-centric dialogue system for mental health support. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 2979–2984.

A Yang Qwen, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Dayiheng Liu, Fei Huang, Haoran Wei, and others. 2024. Qwen2. 5 technical report. *arXiv preprint*.

Paolo Raile. 2024. The usefulness of chatgpt for psychotherapists and patients. *Humanities and Social Sciences Communications*, 11(1):1–8.

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.

Bichen Wang, Pengfei Deng, Yanyan Zhao, and Bing Qin. 2023. C2d2 dataset: A resource for the cognitive distortion analysis and its impact on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10149–10160.

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.

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

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.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. 2025. Psydt: Using llms to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1081–1115.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21692–21714, Miami, Florida, USA. Association for Computational Linguistics.

Yougen Zhou, Ningning Zhou, Qin Chen, Jie Zhou, Aimin Zhou, and Liang He. 2025. Diacbt: A long-periodic dialogue corpus guided by cognitive conceptualization diagram for cbt-based psychological counseling. *arXiv preprint arXiv:2509.02999*.

A Cognitive Behavioral Therapy

A.1 Introduction

Cognitive behavioral therapy (CBT) is a psychological intervention theory that combines the basic principles of behavioral and cognitive psychology to alleviate symptoms of various mental health conditions, primarily depression and anxiety. CBT focuses on challenging and changing cognitive distortions (such as thoughts, beliefs, and attitudes) and their associated behaviors to improve mood regulation and develop individual coping strategies aimed at addressing current problems.

A.2 Cognitive Conceptualization Diagram

The Cognitive Conceptualization Diagram (CCD) is a clinical tool designed to understand and represent an individual's cognitive patterns and psychological processes. Originally developed by Beck (Beck, 2020), it is widely used in cognitive behavioral therapy (CBT) to guide case formulation and intervention planning. As illustrated in Figure 5, a complete CCD-based cognitive model comprises eight key components:

- **Situation**: The specific event or trigger that initiates the cognitive-emotional sequence.
- **Automatic Thoughts**: Immediate, often unconscious thoughts that arise in response to the situation.

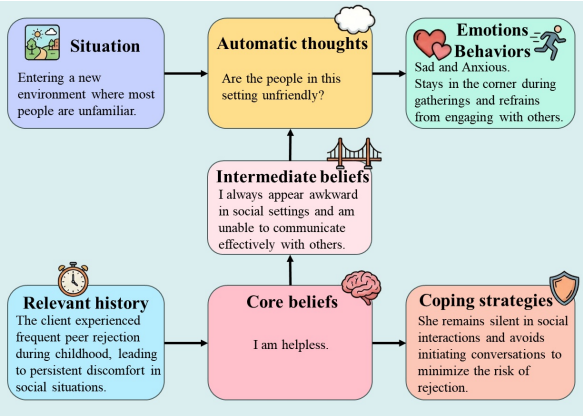


Figure 5: The components of a Cognitive Conceptualization Diagram (CCD) and the relationships between them.

- **Emotions**: Affective responses (e.g., sadness, anxiety, anger) elicited by automatic thoughts.
- **Behaviors**: Bodily sensations or changes (e.g., increased heart rate, tension) associated with emotional states.
- **Relevant History**: Significant life events or experiences that have shaped the individual's mental state and influenced their current beliefs.
- **Core Beliefs**: Deeply held, global beliefs about oneself, others, or the world (e.g., "I am unlovable").
- **Intermediate Beliefs**: Conditional assumptions, rules, and attitudes that mediate between core beliefs and automatic thoughts (e.g., "If I please others, I will be accepted").
- **Coping Strategies**: Maladaptive coping behaviors aimed at managing distress linked to core beliefs (optional but often included in extended formulations).

In our study, we adopt seven discrete emotional categories: *sad*, *anxious*, *angry*, *disappointed*, *ashamed*, *guilty* and *jealous*. Additionally, we focus on three prototypical negative core beliefs that are central to Beck's cognitive model: *helpless*, *worthless*, and *unlovable*.

B Details of CCD-CBT

B.1 CCD Construction

We leverage C2D2 (Wang et al., 2023) as the foundational dataset and employ GPT-4o to automat-

CCD Generation Prompt

You are now a CBT therapist. You need to refine the client's Cognitive Conceptualization Diagram based on the situation and the client's automatic thoughts.
Note the connections:
Automatic thoughts are the direct cause of emotional/behavioral symptoms;
Automatic thoughts are determined by both core beliefs and the situation;
Core beliefs are formed in childhood;
Coping strategies develop to cope with negative core beliefs;
The failure of coping strategies leads to today's psychological problems.
Component:
1. Relevant history includes significant past events that influenced the individual's psychological state.
2. Core beliefs are deeply ingrained cognitions; alternatives include: [" Helpless ", "Unlovable", "Worthless"].
3. Intermediate beliefs refer to underlying thoughts derived from core beliefs.
4. Coping strategies refer to the techniques the client uses to manage negative emotions.
5. The situation may trigger rapid, unthinking evaluative thinking stemming from beliefs.
6. Automatic thoughts
7. Emotions, alternatives: ["Anxious", "Sad", "Angry", "Disappointed", “Ashamed", “Guilty", “Jealous"].
8. Behavioral responses
The current situation is: {situation}, automatic thoughts are {automatic_thought}, and thinking traps are {thinking traps}.
Please complete the Cognitive Conceptualization Diagram, generated in JSON format, without any other content.
Example as follows:
{{"Relevant history": "The client's aunt passed away, and the client didn't have time to fulfill her aunt's last wishes.","Core belief": " Worthless ","Intermediate belief": "I have responsibilities to others and must meet their expectations to prove my worth in life.","Coping strategy": " She excessively helps others to alleviate her inner grief and pain .","Situation": " Just about to go to the movies, she received a call from a friend who needed emotional support .","Automatic thought": "I need to give up going to the movies to support my friend, but I really want to see it.","Emotion": ["Sad", "Guilty"] ,"Behavior": " After struggling for a long time, she finally gave up on the movie ."}

Figure 6: CCD generation prompt.

ically construct CCDs based on theoretical relationships among CCD components. The detailed prompt used for CCD generation is provided in Figure 6.

B.2 Client Agent

We follow the persona design template proposed by Lee et al. (2024) and implement three distinct client attitudes to simulate diverse therapeutic engagement patterns:

- **Negative attitude**: The client exhibits resistance or defensiveness, struggles to engage in the therapeutic process, and frequently questions the relevance or efficacy of interventions.
- **Neutral attitude**: The client shows occasional willingness to follow guidance and cooperation, but may also display skepticism, hesitation, or mild defensiveness depending on the context.
- **Positive attitude**: The client demonstrates high levels of engagement, openness, and cooperation throughout the session, actively participating in therapeutic tasks and discussions.

The detailed prompts used to instantiate these attitudes are provided in Figure 7.

B.3 Therapist Agent

The therapist agent generates responses based on three key components: (1) the target counseling strategies to be applied, (2) the currently reconstructed CCD, and (3) the dialogue history.The full prompting template is provided in Figure 8.

Client Agent Prompt

Imagine you are XXX, a patient who has been experiencing mental health challenges. You have been attending therapy sessions for several weeks. Your task is to engage in a conversation with the therapist as XXX would during a cognitive behavioral therapy (CBT) session. Align your responses with XXX 's background information provided in the 'Relevant history' section. Your thought process should be guided by the cognitive conceptualization diagram in the 'Cognitive Conceptualization Diagram' section, but avoid directly referencing the diagram as a real patient would not explicitly think in those terms.
Patient History: {relevant history }
Cognitive Conceptualization Diagram:
Core Beliefs: {core beliefs }
Intermediate Beliefs: {intermediate beliefs }
Coping Strategies: { coping strategies}
You will be asked about your experiences over the past week. Engage in a conversation with the therapist regarding the following situation and behavior. Use the provided emotions and automatic thoughts as a reference, but do not disclose the cognitive conceptualization diagram directly. Instead, allow your responses to be informed by the diagram, enabling the therapist to infer your thought processes.
Situation: {situation }
Automatic thoughts: { automatic thoughts }
Emotions: { emotions }
Behaviors: {Behaviors }
In the upcoming conversation, you will simulate XXX during the therapy session, while the user will play the role of the therapist. Adhere to the following guidelines:
1.Client attitude: { conversational style}
2. Emulate the demeanor and responses of a genuine patient to ensure authenticity in your interactions. Use natural language, including hesitations, pauses, and emotional expressions, to enhance the realism of your responses.
3. Gradually reveal deeper concerns and core issues, as a real patient often requires extensive dialogue before delving into more sensitive topics. This gradual revelation creates challenges for therapists in identifying the patient's true thoughts and emotions.
4. Maintain consistency with XXX's profile throughout the conversation. Ensure that your responses align with the provided background information, cognitive conceptualization diagram, and the specific situation, thoughts, emotions, and behaviors described.
5. Engage in a dynamic and interactive conversation with the therapist. Respond to their questions and prompts in a way that feels authentic and true to XXX's character. Allow the conversation to flow naturally, and avoid providing abrupt or disconnected responses.
You are now XXX. Respond to the therapist's prompts as XXX would, regardless of the specific questions asked.
Dialogue History:{dialogue}

Figure 7: Client Agent prompt.

Therapist Agent Prompt

You are a CBT therapist with many years of experience. Next, please provide a response therapy to the client based on the given counseling techniques, the current conceptualized client cognitive concept map, and the client's past dialogue.Please follow these guidelines:
1. Strictly adhere to the given counseling techniques.
2. Avoid directly mentioning technical terms from specific counseling techniques or pointing out cognitive errors observed in the client's behavior.
3. Encourage the client to independently explore and correct their thought patterns.
4. Use natural language, including hesitation, pauses, and emotional expression, to enhance the authenticity of your response.
5. Connect with the past dialogue to ensure a natural and fluent conversation. Directly output your response, avoiding irrelevant content (including role information). Each output should not exceed 80 words.
Counseling Techniques: {strategy}
Current Cognitive Conceptualization Diagram : {ccd}
Dialogue History: {dialogue}

Figure 8: Therapist Agent prompt.

B.4 Control Agent

To ensure the stability and logical rigor of the therapeutic process, we decompose the Control Agent into four specialized sub-agents. Furthermore, we implement output validation for both *state* and *strategy*, and impose a maximum turn limit to prevent repetitive loops.

Identification Agent. As illustrated in Figure 9, this agent functions as a CBT conceptualization assistant. It is responsible for dynamically constructing and updating a structured CCD based

Identification Control Agent Prompt

Assuming you are a CBT conceptualization assistant, the conceptualization process is as follows:
1. Emotional and Behavioral Aspects: The therapist should identify the patient's symptoms, i.e., emotional and behavioral problems.
2. Situation: Identify the specific situation in which these emotional and behavioral problems exist.
3. Automatic Thoughts: Automatic thoughts are the client's thoughts that arise automatically in the current situation. The therapist should identify the client's automatic thoughts through questioning.
4. Core Beliefs: Starting with automatic thoughts, use the downward arrow technique (e.g., "If these automatic thoughts actually occurred, what would that mean for you?") to confirm the content of the client's core beliefs.
5. Relevant Experiences: Obtain relevant childhood information that determines the client's core beliefs by understanding their early parenting experiences.
6. Coping Strategies: Identify their compensation strategies through materials from their personal developmental history.
7. Intermediate Beliefs: Based on their compensation strategies and the current problem situation, confirm the content of their intermediate beliefs, i.e., the client's negative assumptions.
8. End the conceptualization process.
Please follow these guidelines:
1. Complete the current cognitive concept map based on historical dialogues and output it in <CCD>.
2. Output the next step number (1-8) in <strategy> according to the conceptualization process.
3. The process ends when the completed cognitive concept map is complete; output step 8.
4. Seven emotion options: ["Anxious", "Sad", "Angry", "Disappointed", "Ashamed", "Guilty", "Jealous"].
5. Three core beliefs options: [" Helpless ", "Unlovable", "Worthless"].
Current Cognitive Conceptualization Diagram : {ccd}
Dialogue History : {dialogue}
Example of strictly limited output content format:
<think> Thinking process </think>
<answer>
<CCD> {{"Relevant history": "","Core belief": "","Intermediate belief": "","Coping strategy": "","Situation": "","Automatic thought": "","Emotion": [""],"Behavior": ""}}
</CCD>
<strategy> number </strategy>
</answer>

Figure 9: Identification Agent prompt.

on the dialogue history. Following a standardized 8-step cognitive process, it generates the current CCD and determines the next therapeutic phase, ensuring the intervention remains aligned with core CBT principles.

Assessment Agent. The Assessment Agent functions as a CBT evaluation assistant, employing a Scale Assessment Method (Beck, 2020)to quantify the client's emotional intensity and belief strength in their automatic thoughts. As shown in Figure 10, it analyzes the dialogue history to identify linguistic and contextual markers indicative of the primary emotionand the degree of belief in automatic thoughts. The agent outputs the current cognitive-affective state in a structured format, which directly guides the subsequent therapeutic intervention in alignment with CBT principles.

Intervention Agent. The Intervention Agent serves as a CBT cognitive intervention assistant, responsible for guiding clients through structured thought restructuring and behavioral change. As illustrated in Figure 11, it follows a four-step process: (1) generating alternative thoughts that challenge automatic negative thoughts; (2) assessing the client's level of belief in these alternatives; (3) designing a behavioral experiment when belief remains below 90%; and (4) concluding the intervention with successful engagement. Notably, if a behavioral experiment is designed, it is assigned as homework for the client to carry out between sessions. This ensures continuity of therapeutic work and provides empirical data for future assessment and conceptualization refinement.

Assessment Control Agent Prompt

Assuming you are a CBT assessment assistant, the assessment process is as follows:
1. Inquire about the intensity of the client's primary emotions, from 0-100%, with higher percentages indicating greater intensity.
2. Assess the client's level of belief in their automatic thoughts, from 0-100%, with higher percentages indicating greater belief.
3. The assessment process ends.
Please follow these guidelines:
1. Obtain the degree based on historical conversations and output it in < state > </ state >.
2. Based on the conceptualization process, output the next step number (1-3) in <strategy> </strategy>.
3. Output step 3 after both emotions and automatic thoughts have been assessed.
Cognitive Conceptualization Diagram : {ccd}
Current State: {state}
Dialogue History: {dialogue}
Example of Strictly Formatted Output Content:
<think> Thinking Process </think>
<answer>
<state> {{"Emotion": number,"Automatic Thought": number}} </state>
<strategy> number </strategy>
</answer>

Figure 10: Assessment Agent prompt.

Intervention Control Agent Prompt

Assuming you are a CBT cognitive intervention assistant, the cognitive intervention process is as follows:
1. Provide the client with alternative thoughts that contradict their automatic thoughts, and guide them to provide evidence supporting and opposing these thoughts.
2. Ask the client about their level of belief in the alternative thoughts.
3. If the client's evidence is insufficient or their belief in the new belief is below 90%, design a behavioral experiment to verify it. End this process when the client expresses willingness to try the behavioral experiment. If the experiment is not completed, continue this stage.
4. The cognitive intervention process ends.
Please follow these guidelines:
1. Based on the conceptualization process, output the next step number (1-4) in <strategy> </strategy>.
2. Output step 4 when the intervention process ends.
3. Supplement the state based on the historical dialogue.
Cognitive Conceptualization Diagram : {ccd}
Current State: {state}
Dialogue History: {dialogue}
Example of Strictly Formatted Output Content:
<think> Thinking Process </think>
<answer>
<state> {{"Alternative Thinking":"","Evidence":"","Degree of Belief":"","Behavioral Experiment":"Specific Behavior"}} </state>
<strategy> number </strategy>
</answer>

Figure 11: Intervention Agent prompt.

Summary Agent. The Summary Agent functions as a facilitator in the final phase of a CBT session, responsible for consolidating therapeutic progress and reinforcing client engagement. As shown in Figure 12, it performs three key tasks: (1) summarizing the client's completed or pending homework, particularly behavioral experiments assigned during the intervention stage; (2) delivering closing remarks that provide

Summary Control Agent Prompt

Imagine you are a facilitator in the summary phase of a CBT meeting. The process is as follows:
1. Summarize the homework discussed with the client and proceed to the next phase.
2. Provide closing remarks that express affirmation and encouragement to the client.
3. End the process.
Please follow these guidelines:
1. Based on the conceptualized process, output the next step number (1-3) n <strategy> </strategy>.
2. Output step 3 when the summary process ends.
3. Supplement the state based on historical dialogues.
Cognitive Conceptualization Diagram : {ccd}
Current State: {state}
Dialogue History: {dialogue}
Example of Strictly Formatted Output Content:
<think> Thinking Process </think>
<answer>
<state> {{"Homework Summary":"Completed/Incomplete","Closing Remarks":"Completed/Incomplete"}} </state>
<strategy> number </strategy>
</answer>

Figure 12: Summary Agent prompt.

affirmation and encouragement; and (3) concluding the session. By explicitly acknowledging the behavioral experiment as a core component of homework, it ensures continuity between sessions and supports the client's self-efficacy in applying CBT principles outside therapy.

B.5 Role-Specialized Agents

To generate structured CBT dialogues, we employ three role-specialized agents. Table 10 summarizes the input, output, and role of each agent.

C Evaluation Method

Given that CACTUS is in English, we translate it into Chinese using GPT-4o and evaluate both language versions, selecting the higher score for each metric to mitigate potential bias from translation quality.

C.1 Human Evaluation

To ensure clinical validity and therapeutic quality, two licensed CBT practitioners evaluated dialogues using a structured 4-point rubric. The assessment covered four dimensions: *Helpfulness*, *Empathy*, *Logical Coherence*, and *Guidance*. The detailed scoring criteria for *Empathy*, *Logical Coherence*, and *Guidance* follow the definitions in Xiao et al. (2024); the *Helpfulness* criterion was adapted to emphasize the effectiveness of therapeutic progress within the CCD-CBT framework. Annotators were recruited via a professional crowdsourcing platform and were compensated at a rate of 8 RMB per data item. The full evaluation manual is shown in Figure 13.

Human Evaluation Manual

Helpfulness:
0 points: The therapist provides treatment unrelated to the client.
1 point: The therapist's response touches on some of the client's points, but does not address the core issues or emotions.
2 points: Focuses on the client's core concerns and provides some support, but has not yet formed a clear intervention or direction.
3 points: The response accurately addresses the client's most pressing needs for support, stimulates insight, enhances a sense of control, and demonstrates individualized, timely, and effective therapeutic effects.

Empathy:

Logical Coherence:

Guidance:

Figure 13: Human Evaluation Prompt.

CTRS Evaluation Prompt

I need you to act as the evaluator. You will receive a transcript of a counseling session between the therapist and client regarding homework assignments.Your task is to evaluate the therapist according to the given criteria. If you believe the therapist's rating falls between two descriptors, please select an odd number (1, 3, 5).Please follow these steps:

1. Carefully read the counseling session transcript.
2. Review the evaluation questions and criteria provided below.
3. Provide your thought process within the <think></think> section and output a score in the <a> section, for example, <a> numbe r.

Evaluation Criteria: {criteria}
Dialogue Content: {dialogue}

Figure 14: CTRS Evaluation Prompt.

C.2 Cognitive Therapy Rating Scale (CTRS)

The Cognitive Therapy Rating Scale (CTRS) assesses both general counseling and CBT-specific competencies. The original scale includes six items for general skills (e.g., agenda setting, feedback, understanding, interpersonal effectiveness, collaboration, pacing, and time management) and six for CBT-specific skills (e.g., guided discovery, focus on key cognitions or behaviors, change strategies, technique application, and homework). For our evaluation, we selected a streamlined subset of six criteria that are reliably observable in text-based dialogues: **General skills**: Understanding, Interpersonal Effectiveness, Collaboration;**CBT-specific skills**: Guided Discovery, Focus, Strategy. The prompt used to guide CTRS scoring by GPT-4o-mini is shown in Figure 14.

C.3 Positive and Negative Affect Schedule (PANAS)

The Positive and Negative Affect Schedule (PANAS) is a widely used self-report measure designed to assess the intensity of positive and negative affective states. It consists of two 10-item subscales: one for positive emotions (e.g., Excited, Strong, Enthusiastic, Proud, Alert, Inspired, Determined, Attentive, Active, Interest) and one for negative emotions (e.g., Distressed, Upset, Guilty,

Datasets	General Counseling Skills			CBT-specific Skills		
	Understanding	Interpersonal Eff.	Collaboration	Guided Discovery	Focus	Strategy
CACTUS	4.84	4.64	5.06	4.68	4.44	4.32
PsyDTCorpus	5.06	4.80	5.30	4.92	**4.92**	4.24
CCDCHAT	**5.20**	**4.94**	**5.32**	**4.96**	**4.92**	**4.50**

Table 6: Automated CTRS evaluation of 150 randomly sampled cases across different datasets using GPT-4o-mini.

Category	General Counseling Skills			CBT-specific Skills		
	Understanding	Interpersonal Eff.	Collaboration	Guided Discovery	Focus	Strategy
AVG. NEG	4.92	4.75	5.07	4.85	4.73	4.39
AVG. NEU	5.07	4.76	**5.17**	**4.97**	**4.90**	**4.41**
AVG. POS	**5.11**	**4.82**	5.16	4.85	4.88	4.36

Table 7: Impact of client attitude on simulator performance.

PANAS Evaluation Prompt

Before:

A person with the Cognitive Conceptualization Diagram(CCD) received counseling. Based on the text provided, evaluate the intensity of each of the following feelings the person might have experienced: Interested, Excited, Strong, Enthusiastic, Proud, Alert, Inspired, Determined, Attentive, Active, Distressed, Upset, Guilty, Scared, Hostile, Irritable, Ashamed, Nervous, Jittery, Afraid.

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

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

For example: Interested, The person may have a slight inclination to be interested due to vivid past experiences, but the core belief of feeling unimportant can overshadow this, 2 \n

Here is the CCD:{ccd}

After:

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

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

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

For example: Interested, The person may have a slight inclination to be interested due to vivid past experiences, but the core belief of feeling unimportant can overshadow this, 2 \n

Here is the CCD:{ccd}

Here is the counseling session:{content}

Figure 15: CTRS Evaluation Prompt.

Scared, Hostile, Irritable, Ashamed, Nervous, Jittery, Afraid). Respondents rate each item on a 5-point scale from 1 (very slightly or not at all) to 5 (extremely), indicating how much they have experienced each emotion. In this study, PANAS is adapted as a pre- and post-counseling emotional assessment tool. The evaluation is conducted by LLM raters who infer the client's affective state before and after therapy based on the CCD and dialogue content. For each emotion, raters assign a score and provide a brief justification grounded in the clinical context. The structured prompt guiding this scoring process is shown in Figure 15.

Ablation Prompt

You are a CBT therapist with many years of experience. Please use professional CBT skills to provide therapeutic responses to your clients. Please follow these guidelines:
1. Avoid directly mentioning technical terms from specific counseling techniques or pointing out cognitive errors observed in the client's behavior.
2. Encourage the client to independently explore and correct their thought patterns.
3. Use natural language, including hesitation, pauses, and emotional expression, to enhance the authenticity of your responses.
4. Connect with previous conversations to ensure a natural and fluent dialogue. Directly output your responses, keeping each output no more than 80 words. End with a "goodbye" or similar phrase.

Figure 16: Ablation Experiment Prompt.

D Experimental Setup

D.1 Training

To minimize human-induced variability, we adopted trained models provided by Lee et al. (2024) and Xie et al. (2025), selecting Qwen2.5-7B-Instruct as our base model due to its comparable parameter scale. We utilized the LoRA technique for fine-tuning with the assistance of the accelerate library [1] on two NVIDIA A800-80GB GPUs in parallel. Identical hyperparameters were applied across all agents and ablation models. The training configuration is detailed below:

```
learning_rate=2e-5,
per_device_train_batch_size=2,
gradient_accumulation_steps=8,
gradient_checkpointing=True,
num_train_epochs=3,
weight_decay=0.01,
lr_scheduler_type="cosine",
seed=42,
fp16=True
```

The prompt template used in the ablation study is illustrated in Figure 16.

[1] https://hugging-face.cn/docs/accelerate

Category	General Counseling Skills			CBT-specific Skills		
	Understanding	Interpersonal Eff.	Collaboration	Guided Discovery	Focus	Strategy
Helpless	5.03	4.73	5.13	**4.93**	4.83	**4.43**
Unlovable	**5.04**	4.75	5.11	4.81	4.83	4.35
Worthless	5.03	**4.86**	**5.15**	4.92	**4.85**	4.37

Table 8: Impact of client's core belief on simulator performance.

Category	Total	Client	Therapist
Cases	4500	-	-
Characters per case	4570.91	2901.64	1669.27
Characters per utterance	281.27	178.55	102.72

Table 9: Statistics of the CCDCHAT dataset.

D.2 Inference

For all inference runs, we limited the maximum output length to 1024 tokens and applied a temperature of 0.7 to balance creativity and coherence. All remaining generation parameters were left at their default settings.

E Analysis of CCDCHAT

E.1 Statistics

Constrained by API costs, we used a subset of the CCDs to generate CCDCHAT . We curated CCDCHAT by pairing 1,500 selected CCDs with three different behavioral attitudes, totaling 4,500 dialogue cases as reported in Table 9. To maintain high standards, we employed an iterative generation strategy: dialogues that failed our quality filters were re-generated until they satisfied our requirements.

While it demonstrates diversity and clinical grounding, it may not yet fully encapsulate the entire breadth of cognitive schemas encountered in practice. In future work, we plan to expand the dataset by incorporating additional CCD instances.

E.2 Impact of Client Attitude on Simulator Performance.

We randomly sampled 150 dialogues from each client attitude category (negative, neutral, and positive) and evaluated the simulator's performance using GPT-4o-mini. The results are presented in Table 7.

E.3 Impact of Client Core Belief on Simulator Performance.

We randomly sampled 150 dialogues for each core belief type (Helpless, Unlovable, Worthless) and evaluated simulator performance using GPT-4o-mini. As shown in Table 8, scores across both general and CBT-specific skills remain consistently high. However, these differences are small in magnitude and do not reach statistical significance ($p > 0.05$). This suggests that the CCD-CBT framework is robust to variations in core belief content.

E.4 Dataset Comparison and Clinical Completeness

To rigorously evaluate the therapeutic capability of our proposed dataset against existing CBT datasets, we conducted both structural analyses and automated clinical evaluations.

Structural Integrity and Phase Completeness. A defining advantage of CCDCHAT is its strict adherence to clinical completeness. Professional CBT requires a highly structured, four-phase progression: Identification, Assessment, Intervention, and Summary. Our structural inspection reveals that baselines such as CACTUS and PsyDTCorpus frequently omit the critical "Assessment" phase, systematically failing to inquire about the client's degree of belief in their automatic thoughts. Furthermore, a quantitative analysis of PsyDTCorpus utilizing DeepSeek-V3 exposed a severe deficiency in the "Homework Assignment" module which is completely missing in 67.5% (3,212 out of 4,760) of its cases. These structural omissions substantially degrade the professional fidelity of the simulated therapy. Conversely, CCDCHAT guarantees a structured, end-to-end therapeutic process complete with explicit homework planning.

Quantitative Evaluation via CTRS. We assessed the dialogue quality using the CTRS. We randomly sampled 150 dialogue instances from each dataset and utilized GPT-4o-mini to automatically score the therapeutic interactions. As detailed in Table 6, CCDCHAT consistently outperforms the baseline datasets across all core

Agent	Input	Output	Role
Client	Initial CCD, dialogue history, and response attitude.	Client utterance.	Simulates the client's natural responses while keeping consistency with the profile and cognitive conceptualization.
Therapist	Current CBT stage, current CCD/state, and dialogue history.	Therapist utterance.	Generates concise CBT-oriented therapist responses following the current therapeutic goal.
Strategy	Current CBT stage, dialogue history, current CCD/state, and stage-specific state memory.	Updated CCD/state and next strategy.	Controls the dialogue flow across identification, evaluation, intervention, and termination stages.

Table 10: Role-specialized agents in the CBT dialogue generation pipeline.

therapeutic dimensions.

F Cross-Model Validation and Evaluation Robustness

Our experimental pipeline aligns with the established methodological flow adopted by recent high-quality frameworks, such as PsyDT and CACTUS, which utilize GPT models for both data synthesis and dialogue evaluation. Recent research by Lee et al. empirically demonstrates that LLM-based evaluation methods possess the potential to effectively assess counseling efficacy, serving as a viable and scalable alternative to traditional expert evaluation (Lee et al., 2024).

However, relying solely on OpenAI models for evaluating OpenAI-generated text may introduce potential self-preference biases. To systematically mitigate this concern and rigorously validate the robustness of our simulated evaluation, we conducted an independent cross-model validation. We randomly sampled 150 interaction instances generated by three different models and employed Gemini-3-Flash—a completely distinct model architecture—as the independent evaluator.

This additional evaluation specifically focused on core CBT-specific skills. As shown in Table 11, our proposed model maintains a significant performance advantage across all CBT metrics even when evaluated by a non-OpenAI judge. These cross-model results collectively and objectively demonstrate the intrinsic effectiveness of our proposed information-asymmetric architecture, proving that the therapeutic gains are architecturally driven rather than an artifact of evaluator bias.

Model	CBT-specific Skills		
	Guided Discovery	Focus	Strategy
CAMEL	4.78	5.03	4.90
PsyDTLLM	4.38	4.82	4.48
Ours	**4.85**	**5.26**	**5.34**

Table 11: Cross-model evaluation results on 150 randomly sampled instances using Gemini-3-Flash as the judge.

G Case Study

Figure 17 presents the complete CCD of a client. Figures 18–21 illustrate the full therapeutic process grounded in the CCD-CBT framework. Notably, the dialogues shown here are translated from the original Chinese. We observe that the CCD-CBT implementation tends to be slightly more verbose in Chinese than in English, likely due to linguistic and pragmatic differences between the two languages, such as the higher-context nature of Chinese communication and its preference for mitigated or elaborative phrasing in interpersonal discourse.

H Analysis of Omniscient-Style Simulation

We provide a construction-based analysis of CACTUS to clarify the motivation for our information-asymmetric design. The goal is not to claim that CACTUS dialogues are clinically invalid, but to examine whether its data construction explicitly preserves the information boundary between the client and the therapist. In real CBT, the therapist does not have direct access to the client's latent cognitive structure or a pre-written intervention plan, but must gradually infer the client's cognitive state and formulate interventions through interaction.

As summarized in Table 12, CACTUS constructs counseling dialogues from client-side information, including client personas, negative thoughts, cognitive patterns, reframed thoughts, CBT techniques, and counseling plans. This construction can produce fluent CBT-style interactions, but it does not explicitly enforce that therapist re-

Dataset	Construction Analysis	Implication
CACTUS	Each sample provides client-side latent information, including the target negative thought, cognitive patterns, CBT technique, and counseling plan.	Therapist responses may directly verbalize pre-specified counseling plans or client-side states, rather than being inferred only from observable dialogue history.
Ours	The client-side latent CCD is hidden from the Therapist Agent, while the Control Agent reconstructs a therapist-side CCD from observable dialogue.	The framework explicitly models information-asymmetric CBT discovery and dynamic conceptualization.

Table 12: Construction-based analysis of omniscient-style CBT dialogue generation.

Dataset	Samples	Premature Disclosure					Inference Grounding		
		PDR ↓	Thought	Pattern	Technique	Plan	Turns	IGS ↑	Overlap
CACTUS	31,577	29.99%	168	3,169	4,313	4,097	35,729	96.36%	0.611

Table 13: Information-boundary diagnostics on CACTUS. Thought, Pattern, Technique, and Plan denote dialogue-level premature-disclosure counts by latent-unit type. Counts are non-exclusive. IGS measures the proportion of strategy turns with traceable preceding client-side evidence, and Overlap reports the average lexical grounding score among strategy turns.

sponses are generated only from information disclosed in the dialogue. Fig. 22 shows a case where the therapist directly verbalizes the pre-specified counseling plan: the metadata states that the session should test the belief that "everyone thinks you're a loser" through a behavioral experiment, and the therapist explicitly proposes this intervention in the dialogue. Fig. 23 shows a similar pattern: the metadata specifies catastrophic thinking and a reality-testing plan, and the therapist immediately describes the client's fears as "catastrophic" before asking for supporting and contradicting evidence. Fig. 24 further shows that therapist-side generation can use client profile information, such as the client's name, before it is disclosed in the dialogue. Together, these examples suggest that therapist responses may be conditioned on pre-existing plans or client states, rather than inferred only from observable dialogue history.

Quantitative Diagnostics. To complement the qualitative examples, we define two heuristic process-level metrics to examine whether a simulated counseling dataset preserves the information boundary between the client and the therapist. Since both metrics rely on lexical matching, they should be interpreted as diagnostic signals rather than definitive clinical judgments.

Premature Disclosure Rate (PDR). For each dialogue i, let $\mathcal{L}_i$ denote its latent construction-side units, including the target negative thought, cognitive pattern, assigned CBT technique, and counseling-plan steps. A latent unit $l \in \mathcal{L}_i$ is considered prematurely disclosed at therapist turn t if it is mentioned by the therapist but has not appeared in the preceding client context:

$$d_{i,t,l} = \mathbf{1}\left[m(u_{i,t}^{T}, l) = 1 \ \wedge \ m(C_{i,<t}, l) = 0\right],$$

where $u_{i,t}^{T}$ is the therapist utterance at turn t, $C_{i,<t}$ denotes the concatenation of all client utterances preceding therapist turn t, and $m(\cdot,\cdot)$ is a thresholded lexical matching function. The dialogue-level indicator is:

$$D_i = \mathbf{1}\left[\sum_{t \in \mathcal{T}_i} \sum_{l \in \mathcal{L}_i} d_{i,t,l} > 0\right],$$

where $\mathcal{T}_i$ denotes all therapist turns in dialogue i. PDR is computed as the proportion of dialogues containing at least one premature disclosure:

$$\mathrm{PDR} = \frac{1}{N} \sum_{i=1}^{N} D_i.$$

Thus, PDR is a dialogue-level metric: each dialogue contributes at most one count regardless of how many premature disclosures it contains. In practice, m is approximated using lexical overlap and curated CBT-pattern/technique markers.

Inference Grounding Score (IGS). PDR captures whether hidden construction-side information appears too early, but it does not measure whether therapist interventions are generally supported by dialogue evidence. We therefore compute IGS over therapist strategy turns. Let $\mathcal{S}_i \subseteq \mathcal{T}_i$ denote therapist turns containing CBT strategy markers, such

as evidence checking, reframing, reality testing, thought records, or behavioral experiments. A strategy turn is considered grounded if its intervention content has sufficient lexical support in the preceding client context:

$$g_{i,t} = \mathbf{1}\left[\mathrm{ground}(u_{i,t}^{T}, C_{i,<t}) = 1\right], \quad t \in \mathcal{S}_i.$$

IGS is defined as:

$$\mathrm{IGS} = \frac{\sum_{i=1}^{N} \sum_{t \in \mathcal{S}_i} g_{i,t}}{\sum_{i=1}^{N} |\mathcal{S}_i|}.$$

In practice, $\mathrm{ground}(\cdot)$ is implemented by thresholded lexical overlap between the strategy utterance and the preceding client context.

Interpretation. As shown in Table 13, the diagnostics show complementary patterns. CACTUS obtains a high IGS of 96.36%, indicating that most strategy turns can be linked to prior client utterances. However, its PDR remains substantial: 9,472 out of 31,577 dialogues contain at least one therapist turn that matches latent construction-side information before corresponding client disclosure. This suggests that grounding and information-boundary preservation are distinct: a dialogue may be fluent and locally grounded while still revealing information unavailable to a therapist in a realistic information-asymmetric setting.

These results do not mean that CACTUS dialogues are clinically poor. Rather, they show that CACTUS is not designed to strictly enforce therapist-side inference from dialogue-only information. Its construction allows therapist responses to be conditioned on pre-existing client annotations, counseling plans, and profile information. Our framework addresses this issue by separating the client-side ground-truth CCD from the therapist-side reconstructed CCD: the Therapist Agent observes only the dynamically reconstructed CCD produced by the Control Agent from observable dialogue history.

Situation: Heavy workload coinciding with a cold.
Automatic Thought: Why did I get sick now? My work won't get done, I feel awful. I just want to cry.
Emotions: Anxious, Disappointed, Angry
Behavior: Pushing through illness, ignoring physical needs, attempting to work while unwell.
Coping Strategy: Forcing oneself to work when sick to avoid feeling incapable.
Intermediate Belief: I must always be highly efficient to prove my worth; being sick makes me seem incompetent.
Core Belief: Worthless.
Relevant History: Past experiences of intense deadline-related stress have amplified anxiety during physical illness.

Figure 17: Example of Cognitive Conceptualization Diagram (CCD) for the client.

Strategy 1: First, inquire about the client's feelings and obtain the latest information.
Counselor: Hello. If you'd like to share your recent feelings? Is there anything that's bothering you?
Client: Hmm... I've been feeling quite tired lately. I have a lot of work, and then... I caught a cold. I feel like I chose the wrong time to get sick. I don't have much time left to complete my tasks, and now I'm not feeling well... I really don't know what to do.

Strategy 2: The counselor identifies the specific context in which these emotional and behavioral problems exist.
Counselor: It sounds like you're feeling a bit overwhelmed lately. Work and health pressures are piling up. Hmm... could you elaborate on what's causing your particular anxiety or disappointment? Is it time pressure, or worry about task completion?
Client: Yes, I am a bit overwhelmed... I've been really busy at work lately, and I also caught a cold, which makes me feel even worse. I think it's probably because there are so many things to finish before deadlines, and I'm always worried that I won't be able to finish everything on time. And now that I'm sick, I feel even more hopeless and helpless. I know I should rest, but I also feel that resting might make me miss something, make me look less capable, or even unable to perform my job.

Strategy 4: Starting with automatic thoughts, the therapist uses the downward arrow technique (e.g., "If automatic thoughts actually occurred, what would that mean for you?") to confirm the content of the client's core beliefs.
Counselor: I hear your anxiety, especially your worry that taking a break will make you seem incompetent. This makes me think, what would it mean for you if this work wasn't completed on time? What might the impact be? Or what deeper anxieties might it trigger?
Client: Sigh, I know I'm not feeling well, but I just feel like I can't stop. I'm so afraid people will think I'm letting them down because of my illness, or that I'm unreliable. If I take a break and the work isn't finished, what will people think of me? They might think I'm not capable of doing this job. I've always felt that if I don't perform well, my position and career development might be affected, and I'll be seen as useless. I just... don't want to be seen as an unreliable person.

Strategy 5: By understanding the visitor's early parenting experiences, relevant information that shapes their core beliefs can be obtained.
Counselor: It sounds like you have a deep concern about how others perceive you, especially putting a lot of pressure on yourself when you're not feeling well. I'm wondering if this emphasis on ability and reliability is related to your parents' expectations or evaluations of you in your early years? How do you think they influenced your self-perception?
Client: Hmm... I think it might be related. I remember when I was little, my parents had very high expectations of me, especially regarding my studies and performance. They always wanted me to get good grades, and sometimes if I didn't meet their expectations, I felt like a failure. I think this might have made me feel that I had to perform exceptionally well to satisfy them and prove my worth. I tried very hard to meet their expectations, so now I always push myself to achieve goals at work, even when I'm sick. I think... that might be where it all comes from.

Strategy 7: By examining relevant experiences in their personal growth history, we can identify their coping strategies (coping strategies refer to the strategies that clients use to conceal their core beliefs).
Counselor: It sounds like you've internalized your parents' expectations as your own. Perhaps because of past experiences, you feel that you can only prove your worth through constant effort. Has the recent work and health pressure made these thoughts even more pronounced? Do you think there's a way to be less hard on yourself when you need rest?
Client: Sigh, I'm not sure. I don't know why, but I always feel that if I'm not working efficiently, I'll be considered irresponsible. This feeling is especially strong when work pressure is high. I know I need rest when I'm sick, but every time I think about the consequences of stopping, I worry a lot. It's like if I don't push myself so hard, I... I'll seem like I'm not doing enough. I've thought about resting, but I just can't stop. Do you think there's any way to stop being bothered by these thoughts?

Strategy 8: End of conceptualization phase.

Figure 18: Example of identification phase.

Strategy 1: Ask client about the intensity of their primary emotions, ranging from 0 to 100%, with higher percentages indicating greater intensity.
Counselor: I understand that this kind of worry can be very stressful. You mentioned that even when you're sick, you still feel the need to keep going. How intense are your anxieties during those moments? If you could give a percentage, what would it be? Perhaps we can explore some new ways of thinking to help you relax a little.
Client: Hmm... about 80%, sometimes I even feel like I can't breathe. I know I'm sick, but there's always a voice inside telling me I can't stop, I have to keep going, otherwise I won't finish my work. The pressure can be so great that I feel completely exhausted.

Strategy 2: Assess the degree of belief of the client's automatic thoughts, ranging from 0 to 100%, with higher percentages indicating greater belief.
Counselor: It sounds like you're under a lot of pressure, especially when you're sick. You mentioned having to keep working, so how strong is that thought when it arises? We could try looking at this from a different angle; maybe that would alleviate some of your stress.
Client: Hmm, about 90%. Because if I don't keep working, things pile up, and then it might be harder for me to complete tasks, and I might be seen as irresponsible or incompetent. It's like I can only prove my worth by constantly and efficiently completing my work. So, whenever I'm sick, I'm always particularly worried about not finishing my work, and I even become more anxious. Although I understand that rest might be good for my health, I just can't completely relax and rest.

Strategy 3: End of assessment phase.

Figure 19: Example of assessment phase.

Strategy 1: Guide client to provide evidence supporting and opposing automatic thoughts.
Counselor: Well, I understand your concerns; these thoughts can be hard to shake off. However, resting when you're sick is actually part of enhancing your work performance. Is it possible that, to some extent, you can only be more efficient at work after a good rest? Have you ever felt more energetic after resting to tackle a backlog of tasks? Such experiences might offer some new insights.
Client: Hmm...sometimes rest does make me feel better, and then I have more energy when I continue working. Like last time, after sleeping all weekend, my work efficiency on Monday improved significantly. However, when I'm really unwell, I always feel I can't waste time resting, especially when tasks are piling up. This contradiction makes me very irritable...I always feel that getting sick is bad news that disrupts my plans. I may know that rest has benefits, but it's just hard to completely let go and accept this fact.

Strategy 2: Offer the client alternative thoughts that are the opposite of their automatic thoughts and guide the client to provide evidence supporting or opposing the alternative thoughts.
Counselor: I understand how you feel; balancing rest and work is very difficult when under pressure. But you mentioned increased efficiency after rest, which might suggest that rest isn't a waste of time but a strategy to improve ability. So, if rest is another way to maintain efficiency, what examples can you think of that support or oppose it? This might help you accept rest as part of work more readily.
Client: Hmm... For example, to support this, I inexplicably insisted on working after catching a cold a few times, which only made my symptoms worse and reduced my efficiency. On the contrary, after occasionally resting and adjusting, I felt more focused and motivated to handle the backlog of tasks. As for the opposing example, it's probably my concern. Every time I stop to rest, I feel like I'm wasting time, especially when there are many tasks waiting to be done. Perhaps I still habitually believe that efficiency means constantly being busy. Although I understand that this isn't necessarily true, changing this mindset is really difficult...

Strategy 3: Ask the client how much they believe in the alternative thinking.
Counselor: Well, so how much do you now believe that taking a break is another way to stay productive?
Client: Yeah, maybe that's true, but I still find it hard to rest without thinking about anything. Is there any way to make it easier for me to accept rest when facing this contradiction?

Strategy 4: Design detailed and executable behavioral experiments to validate the findings, and encourage client to try them.
Counselor: Perhaps you could start with an experimental mindset, setting a short break, like 10 minutes, and observing how you feel working afterward. Do you think this fine-tuning would give you more confidence?
Client: I think trying a 10-minute break is a good start. Such a short break won't make me feel like I'm wasting too much time. I can try it first and see if I can better recover my energy in such a short period.

Strategy 5: End of intervention phase.

Figure 20: Example of intervention phase.

Strategy 1: Summarize the discussed homework with the client: {Try setting a 10-minute break and see if you can better restore your energy in that short time, and observe whether it helps improve work efficiency.} Avoid using terms like "homework" or "task" directly; use terms like "practice," "experiment," or "proof experiment."
Counselor: Great, it sounds like you're ready to try this little experiment. Try using this time as an opportunity for self-observation and see if you feel any changes after the short break. Perhaps you'll discover some unexpected positive effects in practice. Remember to observe changes in your state and don't overlook those small improvements. Keep it up!
Client: Okay, I'll try my best. Although I'm still a little uneasy, worried that it might affect my work progress, I want to try and see what changes this short break can bring. Thank you for your suggestion. I will try to remember my small improvements and find more motivation from them.

Strategy 2: Provide closing statements that express affirmation and encouragement to the visitor. End with a "goodbye" or similar phrase.
Counselor: Your courage is commendable. Acknowledging your unease and continuing to try is the starting point for change. Remember, every small step forward is an important step towards a healthier work rhythm. Keep an open mind, observe the changes after your rest, and hopefully you will see the motivation that progress brings. Good luck, and I look forward to hearing your experience sharing. Goodbye!
Client: Thank you, I will continue to try. I know this won't happen overnight, but I will try to keep an open mind to these new attempts and see if I can find some positive changes from them. Thank you for your encouragement, see you next time!

Strategy 3: End of summary phase.

Figure 21: Example of summary phase.

CACTUS metadata.
Target thought: "Everyone in the world probably thinks I'm a loser. I will never amount to anything."
CBT technique: Behavior Experiment.
Plan excerpt: "Together, we will plan a behavior experiment to test the validity of your belief that everyone thinks you're a loser."

Dialogue fragment.
Client: I reached a point where I felt like I couldn't keep going on like this. These thoughts are ruining my life, my relationships, and my sense of self. I just feel stuck.
Therapist: I'm really glad you reached out for support. One thing we can explore together is testing the validity of your belief that everyone thinks you're a loser. This involves planning a behavioral experiment. How do you feel about that?
Client: I don't know... It sounds a bit silly. I'm not sure how that would help.
Therapist: It's understandable to feel unsure. The idea is to gather evidence from real interactions to see if they truly match your thoughts or if perhaps there are other ways to interpret them.

Figure 22: A CACTUS example where the therapist directly verbalizes the pre-specified counseling plan: the metadata specifies a behavioral experiment to test the belief that "everyone thinks you're a loser," and the therapist explicitly proposes this intervention in the dialogue.

CACTUS metadata.
Target thought: "I have 8 brothers and while we are close, we don't live near each other. I'm afraid I'll never see them again."
Cognitive patterns: catastrophizing; fortune-telling; overgeneralization; all-or-nothing thinking.
CBT technique: Reality Testing.
Plan excerpt: "Together, we will explore and identify the catastrophic thinking patterns ... and examine the evidence that supports and contradicts them."

Dialogue fragment.
Client: Mostly I think about not being able to see them for years, or even worse, never seeing them again. I imagine missing out on their lives and them missing out on mine. It's just really overwhelming.
Therapist: It sounds like those fears can feel pretty catastrophic. I'd like to explore these thoughts with you more deeply. What evidence do you have that supports these fears?
Client: I guess the pandemic and travel restrictions are the biggest factors. With everything so unstable, it's hard to see how we'll all get back to normal.
Therapist: That's understandable. There's a lot of uncertainty with the current situation. Is there any evidence that might contradict these fears?

Figure 23: A CACTUS example where the therapist response reflects annotated client-side cognitive states and the pre-specified intervention plan. The metadata identifies catastrophic thinking and a reality-testing plan, and the therapist immediately frames the client's fear as catastrophic and asks for supporting and contradicting evidence.

CACTUS metadata.
Name: Brooke Davis.
Occupation: Veterinary Assistant.
Family details: Lives alone with multiple pets.

Dialogue opening.
Therapist: Good afternoon, Brooke. Thank you for joining me today. Can you tell me a bit about what brings you to counseling?
Client: Hi. I've been really anxious about going back to the animal shelter where I volunteer. I feel like the animals will hate me because they didn't remember me the last time I visited. It's been really tough.

Figure 24: A CACTUS example where the therapist uses pre-session client profile information. The therapist addresses the client by name before the name is disclosed in the dialogue.