

Stage-Aware Controllable Dialogue Generation for CBT-Based Psychological Support

Haibo Chen¹, Qiuyang Zhao¹, Qinyuan Liu¹, Bo Jiang¹

Meng Sun², Mingmei Gao³, Qingwei Chong^{1,*}

¹China Nanhu Academy of Electronics and Information Technology

²Southwest Medical University

³Nanjing University

{chenhaibo, zhaoqiuyang, liuqinyuan, jiangbo02, chongqingwei}@cnaeit.com

20240899120026@stu.swmu.edu.cn, mmgao@smail.nju.edu.cn

*Corresponding author

Abstract

Large language models (LLMs) have shown promise in psychological support, yet existing CBT-oriented systems often rely on prompt-based imitation or supervised learning from counseling-like responses, without explicitly modeling the structured therapeutic process of Cognitive Behavioral Therapy (CBT). This limits their controllability, intervention completeness, and coherence in multi-turn counseling. We formulate CBT-based psychological dialogue as a stage-aware controllable generation problem, where therapeutic stages serve as explicit control signals for intervention selection and therapeutic progression. To provide structured supervision under limited access to real counseling data, we design a CBT-stage multi-agent data generation engine that synthesizes multi-turn dialogues with user-state trajectories, stage-specific interventions, and content-level verification. Using the generated stage-structured data, we train Qixin, a CBT dialogue model that internalizes stage-aware therapeutic behaviors without requiring multi-agent coordination at inference time. We further introduce a hierarchical evaluation paradigm covering single-turn counseling skills, CBT procedural fidelity, stage alignment, and longitudinal session coherence. Experiments show that Qixin improves CBT-oriented counseling capability and outperforms existing psychological dialogue models and similarly sized open-source LLMs at the process level, achieving stronger procedural fidelity, stage consistency, and session-level coherence. These results demonstrate the effectiveness of explicit therapeutic-stage supervision for process-aware psychological support. Code and data are available at <https://github.com/IamJiangBo/CBT-Stage-Multi-Agent-Data-Generation-Engine>.

1 Introduction

Large language models (LLMs) have demonstrated strong capabilities in generating empathetic and supportive responses in open-ended conversations,

enabling their use in psychological support scenarios such as emotional companionship, stress relief, and psychoeducation (Pruss et al., 2026; Dalal et al., 2025; Wang et al., 2025; Huang et al., 2024; Wang et al., 2023; Lee et al., 2024). However, psychological support differs fundamentally from ordinary social dialogue: effective counseling requires not only supportive language, but also structured therapeutic progression grounded in psychological theory. Cognitive Behavioral Therapy (CBT), one of the most widely adopted and empirically validated psychotherapy paradigms, provides a structured and goal-oriented framework in which therapeutic goals evolve progressively across stages (Butler et al., 2006; Henrich et al., 2023; Beck, 2005).

Despite recent progress, most CBT-style dialogue systems still treat psychological support as generic response generation. They rely primarily on prompt-based imitation or supervised fine-tuning on counseling-like responses, rather than explicitly modeling the staged intervention process underlying CBT (Na, 2024; Zhang et al., 2025; Shen et al., 2024; Xu et al., 2025; Kim et al., 2025). As a result, these systems may produce responses that appear empathetic or CBT-like, but fail to determine the current therapeutic stage, select appropriate intervention components, or maintain consistent counseling strategies across turns (Iftikhar et al., 2025; Wang et al., 2026).

We argue that the key missing capability is *stage-aware controllability*: the ability to condition response generation on therapeutic stages and use stage transitions as explicit supervision for multi-turn intervention planning. CBT is inherently progressive: early interactions emphasize rapport building, emotional validation, and problem clarification; middle-stage interventions focus on automatic thoughts, cognitive distortions, and maladaptive beliefs; later stages emphasize behavioral experiments, coping plans, self-monitoring, and re-

lapse prevention (Peterson, 1991; Kazantzis et al., 2018; Librea and Lovitos, 2025). Treating all turns as homogeneous response generation ignores this structure and makes it difficult for LLMs to learn *when* and *how* different CBT techniques should be applied. We therefore formulate CBT-based psychological dialogue as a **stage-aware controllable generation** problem, where therapeutic stages govern both intervention selection and therapeutic progression.

A fundamental obstacle to learning such behavior is the scarcity of multi-turn CBT dialogue data with explicit therapeutic structure. Real-world counseling data are difficult to obtain due to privacy, ethical, and clinical constraints, while existing public datasets often lack fine-grained annotations of therapeutic stages, user state transitions, and stage-specific intervention components. Without such supervision, models are likely to learn surface-level counseling language rather than the process structure of CBT.

To address this challenge, we design a CBT-stage multi-agent framework as a structured data generation engine. Starting from privacy-processed real case reports collected from YiXinLi¹, we use LLMs to extract initial user queries and user profiles as engine inputs. The engine contains a client-side and a counselor-side: the client-side simulates psychological state trajectories and turn-level user utterances conditioned on the profile and dialogue history, while the counselor-side decomposes CBT counseling into user analysis, stage inference, stage-specific response generation, and content verification. This process synthesizes multi-turn dialogues whose counselor responses are aligned with therapeutic stages and evolving user states. Using the generated stage-structured dialogues together with open-source psychological data, we train Qixin, a CBT dialogue model that internalizes stage-aware response patterns.

We further propose a hierarchical evaluation paradigm for CBT dialogue generation. Instead of relying on surface-level text similarity metrics, we evaluate models along single-turn counseling skills, session-level therapeutic structure, and longitudinal coherence. Experiments show that **Qixin** consistently improves CBT-specific intervention capability, procedural fidelity, and stage-consistent multi-turn counseling compared with strong general LLMs and existing psychological dialogue

models.

Our contributions are summarized as follows:

- We formulate CBT-based psychological dialogue generation as a **stage-aware controllable generation** problem with explicit therapeutic-stage supervision.
- We propose a **CBT-stage multi-agent data engine** and construct **QixinConv**, a stage-structured CBT dialogue dataset with user-state trajectories and verified stage-specific interventions.
- We train **Qixin**, a CBT dialogue model that internalizes stage-aware therapeutic behaviors without multi-agent inference.
- We introduce a hierarchical evaluation framework covering single-turn counseling skills, CBT procedural fidelity, structured stage execution, and longitudinal session coherence.

2 Related Work

LLM-based Mental Health Models. LLMs have increasingly been adapted to mental health applications through instruction tuning, counseling dialogue modeling, and domain-specific knowledge injection. Representative systems include MeChat (Qiu et al., 2024), SoulChat (Chen et al., 2023), CPsyCoun (Zhang et al., 2024), and PsychoLLM (Hu et al., 2025). These models improve psychological fluency, empathy, and supportive response generation, but typically treat counseling as response-style adaptation rather than a structured therapeutic process. CBT-oriented systems such as CBT-LLM (Na, 2024), AutoCBT (Xu et al., 2025), and Patient- Ψ (Wang et al., 2024) incorporate psychological theory through prompts, inference-time agents, or patient simulation. In contrast, our work uses CBT stages as training-time supervision, enabling a single model to internalize stage-aware response patterns rather than relying on prompt-level or inference-time control.

Mental Health Dialogue Datasets. High-quality dialogue data are essential for mental health LLMs, yet real counseling data are difficult to access due to privacy and ethical constraints. Existing resources include PsyQA (Sun et al., 2021), Therapist QA², ESConv (Liu et al., 2021), SoulChatCorpus (Chen et al., 2023), CPsyCoun (Zhang et al.,

¹<https://www.xinli001.com/>

²<https://www.kaggle.com/datasets/arnmaud/therapist-qa>

2024), Psy-Insight (Chen et al., 2025), PsychEval dataset (Pan et al., 2026), PsyDial (Qiu and Lan, 2025). Synthetic construction methods such as PsyDT (Xie et al., 2025), CPsyCoun, and PsychoLLM further scale counseling data through reconstruction or LLM-based generation. However, these resources generally lack explicit coupling between CBT stages, user-state trajectories, and stage-specific intervention constraints. Our dataset addresses this gap by aligning user psychological trajectories with therapeutic stages and intervention objectives.

Evaluation of Mental Health LLMs. Evaluation of mental health LLMs has shifted from traditional NLP metrics, toward automated LLM-based assessment. Early protocols mainly evaluate response-level dimensions such as empathy, professionalism, relevance, and safety (Zhang et al., 2024; Hu et al., 2025; Shen et al., 2024; Zhang et al., 2025). Recent benchmarks further move toward process-level counseling evaluation: PsychEval (Pan et al., 2026) introduces a high-realism, multi-session, and multi-therapy benchmark that assesses therapeutic alliance, intervention strategy, ethical safety, and therapy-specific competence. Despite these advances, existing evaluation schemes still insufficiently assess whether a model follows the therapeutic process structure during multi-turn interactions. Turn-level metrics can judge whether a response is helpful or safe, but cannot fully determine whether it appears at an appropriate therapeutic stage, whether intervention components are complete, or whether the dialogue maintains longitudinal CBT coherence. Our evaluation therefore emphasizes stage consistency, intervention completeness, and multi-turn coherence, aligning assessment with stage-aware controllable CBT dialogue generation.

3 Stage-Aware Controllable CBT Dialogue Generation

This section presents our formulation and implementation of stage-aware controllable CBT dialogue generation. We first formalize CBT dialogue generation as an interaction between client-side psychological dynamics and counselor-side therapeutic intervention (§3.1). We then define the CBT stage space and user psychological state space (§3.2). Next, we describe how the CBT-stage multi-agent generation engine dynamically constructs a synthetic dialogue dataset under dual control

of user psychological states and CBT therapeutic stages (§3.3). Finally, we introduce the training of Qixin (§3.4).

3.1 Problem Formulation

We formulate CBT-based psychological dialogue as a **stage-aware controllable generation problem**. Given a user profile p , the dialogue history before turn t ,

$$H_{t-1} = \{(u_i, y_i)\}_{i=1}^{t-1}, \quad (1)$$

and the current user utterance u_t , the goal is to generate a counselor response y_t that is contextually coherent and therapeutically appropriate.

The simulated CBT counseling involves two interactive processes: the user’s evolving psychological condition and the counselor’s stage-aware intervention. Accordingly, we model dialogue generation as an alternating client–counselor process.

On the **client-side**, the user model first infers the user’s latent psychological state:

$$s_t \sim p_\theta(s \mid p, H_{t-1}), \quad (2)$$

and then generates a state-consistent user utterance:

$$u_t \sim p_\theta(u \mid p, H_{t-1}, s_t). \quad (3)$$

On the **counselor-side**, the therapist does not directly observe the latent user state. Instead, the counselor conditions on a structured therapeutic memory representation:

$$a_t = M_\theta(p, H_{t-1}, u_t), \quad (4)$$

which summarizes salient counseling information from the interaction history and current user utterance. Based on this memory, the counselor infers the current therapeutic stage:

$$z_t \sim p_\theta(z \mid a_t), \quad (5)$$

and generates a stage-aware response:

$$y_t \sim p_\theta(y \mid a_t, z_t). \quad (6)$$

This formulation transforms CBT dialogue generation into a controlled process where the model learns both *when* an intervention strategy should be applied and *how* it should be expressed under evolving user conditions.

3.2 CBT Stage and User State Modeling

To explicitly model therapeutic progression and user-side dynamics, we define a **therapeutic stage space** $\mathcal{Z}$ and a **user psychological state space** $\mathcal{S}$.

Therapeutic Stage Space. We model CBT counseling as a progression over three macro stages:

$$z_t \in \mathcal{Z} = \{A, C, B\}, \quad (7)$$

corresponding to *Case Assessment (A)*, *Cognitive Adjustment (C)*, and *Behavioral Activation (B)*.

Stage *A* emphasizes rapport building, emotional validation, and problem clarification. Stage *C* focuses on identifying maladaptive cognitions through reflective questioning and cognitive restructuring. Stage *B* translates therapeutic insight into actionable coping plans and behavioral change. The stage variable z_t therefore acts as a high-level controllable signal governing intervention objectives.

User Psychological State Space. In addition to therapeutic stages, we define a discrete psychological state space $\mathcal{S}$ to model user-side psychological dynamics. Specifically, $\mathcal{S}$ contains fourteen states, and the definitions of these states are as shown in the Appendix A.1.

The dynamically realized state sequence $\tau = (s_1, s_2, \dots, s_T)$ constitutes the user’s **psychological trajectory** throughout a dialogue. Unlike single-turn emotional labels, τ captures long-range psychological dynamics, while each state s_t is selected according to the evolving dialogue context rather than fixed in advance. This state space prevents the simulated client from passively accepting every counselor suggestion and enables more realistic user-side responses across the therapeutic process. Together, z_t and s_t support stage-aware and user-adaptive CBT dialogue generation, where z_t determines *what intervention objective should be pursued*, while s_t governs *how the user psychologically responds*.

3.3 CBT-Stage Multi-Agent Data Generation Engine

A major obstacle to learning stage-aware CBT dialogue behavior is the scarcity of multi-turn counseling data with explicit therapeutic progression and user psychological dynamics. To address this issue, we construct QixinConv through a CBT-stage multi-agent data generation engine (Figure 1). The

engine dynamically synthesizes structured CBT dialogues under dual control of user psychological states and CBT therapeutic stages by alternating between client-side psychological simulation and counselor-side therapeutic intervention. It is used solely for data construction and is not required during inference.

Metadata Extraction. We first collect a set of privacy-processed real psychological case reports from YiXinLi. For each case, an LLM extracts two types of structured metadata: (1) an initial user query q describing the presenting concern, and (2) a user profile p containing demographic background, personality traits, emotional tendencies, and problem context. The pair (q, p) serves as the initialization signal for dialogue generation, grounding synthesized conversations in realistic counseling scenarios while preserving privacy.

Given an initial user query q and user profile p , the engine produces a structured dialogue instance:

$$D = \{(u_t, z_t, s_t, y_t)\}_{t=1}^T, \quad (8)$$

where u_t denotes the user utterance, s_t the user psychological state, z_t the CBT therapeutic stage, and y_t the counselor response. The framework consists of two independent agents: a client-side simulator and a counselor-side generator, which alternately interact to produce a complete CBT counseling dialogue.

Client-Side Psychological Simulation. The client-side models psychologically realistic user behavior rather than passive acceptance of counselor suggestions. During initialization, the client model takes the initial user query q and user profile p as input, and constructs an initial user-state plan $\tau^{(0)}$ as a prior over possible psychological evolution. At each subsequent turn t , given the dialogue history H_{t-1} , the client model dynamically selects the most plausible psychological state $s_t \in \mathcal{S}$ from the discrete state space, updates the evolving user-state sequence

$$\tau^{(t)} = (s_1, s_2, \dots, s_t), \quad (9)$$

and generates a state-consistent user utterance u_t according to Equation (3).

To ensure realistic counseling dynamics, dynamic user-state selection follows four principles: *clinical plausibility*, ensuring consistency with CBT therapeutic progression; *personalization*,

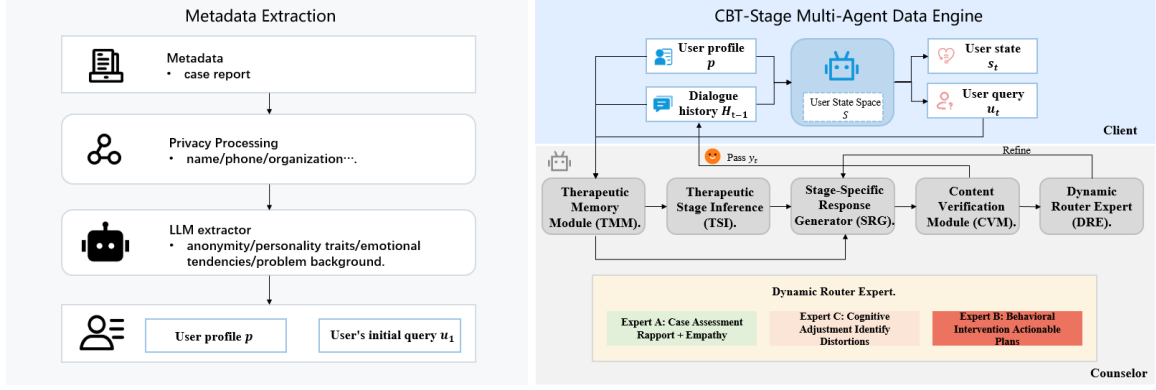


Figure 1: Overview of the CBT-stage multi-agent data generation engine. The left part shows metadata extraction from case reports, producing the user profile and initial query. The right part illustrates the dynamic construction of CBT-style multi-turn dialogues through alternating client-side psychological simulation and counselor-side stage-aware generation. The client-side selects user psychological states from a discrete state space to generate state-consistent utterances, while the counselor-side uses therapeutic memory, stage inference, response generation, content verification, and dynamic refinement to produce CBT-stage-aware interventions.

adapting state transitions to user profiles and problem context; *temporal coherence*, avoiding abrupt psychological shifts; and *non-linearity*, allowing realistic regression patterns such as hesitation or avoidant regression rather than assuming monotonic improvement. State-specific behavioral constraints are further imposed during user utterance generation. For example, users in the defensive state tend to produce resistant or evasive responses, while users in the reflecting or willingness-to-act state exhibit stronger self-exploration and behavioral readiness. This mechanism prevents the dialogue from degenerating into one-sided instruction following and encourages realistic psychological interaction dynamics.

Counselor-Side Stage-Aware Generation. The counselor-side consists of five collaborative modules that jointly implement stage-aware therapeutic intervention.

Therapeutic Memory Module (TMM). Long counseling histories can lead to context drift and inefficient generation. To address this issue, the TMM incrementally summarizes the dialogue into a structured therapeutic memory:

$$a_t = M(p, H_{t-1}, u_t), \quad (10)$$

which maintains salient information from the interaction, including problem context, emotional tendencies, behavioral patterns, historical dialogue cues, and prior counselor interventions. The therapeutic memory acts as the counselor’s working memory and serves as the input to downstream modules.

Therapeutic Stage Inference (TSI). Based on the therapeutic memory, the TSI module predicts the current CBT stage:

$$z_t \sim p(z | a_t), \quad (11)$$

ensuring coherent therapeutic progression rather than arbitrary switching among intervention strategies.

Stage-Specific Response Generator (SRG). Conditioned on (a_t, z_t) , the SRG generates a candidate counselor response:

$$\hat{y}_t = G(a_t, z_t), \quad (12)$$

where Stage A prioritizes rapport building and emotional validation, Stage C emphasizes reflective questioning and cognitive restructuring, and Stage B focuses on actionable planning and behavioral activation.

Content Verification Module (CVM). The CVM evaluates whether the generated response satisfies stage-specific therapeutic requirements, including intervention appropriateness, counseling completeness, and consistency with the user’s observable psychological cues summarized in a_t . It outputs a verification signal:

$$r_t = V(\hat{y}_t, a_t, z_t), \quad r_t \in \{0, 1\}, \quad (13)$$

where $r_t = 1$ indicates successful verification.

Dynamic Router Expert (DRE). When verification fails, the DRE identifies the corresponding

content issues and selects a stage-specific refinement strategy to revise the response:

$$y_t = \begin{cases} \hat{y}_t, & r_t = 1, \\ R(\hat{y}_t, a_t, z_t), & r_t = 0. \end{cases} \quad (14)$$

This process forms a generate–verify–refine loop that prevents low-quality or therapeutically inappropriate responses from being directly delivered to the client.

By alternating between user-state-aware client-side simulation and CBT-stage-aware counselor intervention, the proposed data construction process converts open-ended dialogue generation into a structured, dual-aware procedure with explicit supervision for user-state evolution, therapeutic stage progression, and intervention quality.

The resulting QixinConv dataset contains multi-turn CBT dialogues with turn-level user utterances, counselor responses, CBT stage labels, and user psychological state labels. Detailed dataset statistics and additional analyses are provided in Appendix A.

3.4 Qixin Model

Using the constructed user-state- and CBT-stage-aware dialogue dataset, we train **Qixin** as an empirical instantiation of stage-aware controllable CBT dialogue generation. Rather than relying on the multi-agent framework at inference time, Qixin learns to internalize therapeutic stage progression and user-adaptive counseling behavior directly from structured supervision.

Qixin is initialized from Qwen3-14B-Instruct and trained through a two-stage pipeline. Detailed training parameters can be found in Appendix B.

Stage 1: Psychological Dialogue Adaptation.

We first fine-tune the backbone model on public psychological dialogue datasets, including SmileChat, SoulChatCorpus, PsyDTCorpus, PsyInsight, and CPsyCounD. This stage improves supportive dialogue ability, emotional understanding, and conversational coherence in mental health scenarios.

Stage 2: CBT-Stage-Structured Fine-tuning.

We further fine-tune the model on the user-state- and CBT-stage-aware dialogue dataset constructed in §3.3. Crucially, stage and state annotations are not input metadata or conditioning signals. Instead, they are serialized as **counselor-side intermediate**

generation targets within the target sequence and optimized by standard next-token prediction.

At turn t , the input sequence is:

$$x_t = \text{Ser}_{\text{in}}(p, H_{t-1}, u_t), \quad (15)$$

where p is the user profile, $H_{t-1} = \{(u_i, y_i)\}_{i=1}^{t-1}$ contains only previous user utterances and counselor responses, and u_t is the current user utterance. Crucially, previous stage/state annotations $\{(z_i, s_i)\}_{i < t}$ are excluded from H_{t-1} , forcing the model to re-infer the current therapeutic stage and user state from dialogue context alone.

The counselor target sequence is:

$$r_t = [\text{<stage>} z_t \text{</stage>;} \\ \text{<state>} s_t \text{</state>;} \\ \text{<response>} y_t \text{</response>}], \quad (16)$$

where $z_t \in \mathcal{Z}$ is the CBT therapeutic stage, $s_t \in \mathcal{S}$ is the user psychological state, and y_t is the counselor response. Thus, z_t and s_t are not input conditions but intermediate assistant-side tokens preceding the final response y_t .

The training objective is standard next-token prediction over r_t :

$$\mathcal{L} = - \sum_{t=1}^T \sum_{k=1}^{|r_t|} \log p_{\theta}(r_{t,k} \mid x_t, r_{t,<k}), \quad (17)$$

where k ranges over all tokens in r_t , including <stage>, <state>, and <response> segments. The model learns to predict therapeutic stage, user state, and counselor response in sequence, conditioned on dialogue context rather than externally supplied labels.

Inference Modes. At inference, Qixin supports two modes. In the default **self-planning mode**, given (p, H_{t-1}, u_t) , Qixin autoregressively generates $\hat{z}_t$, $\hat{s}_t$, and $\hat{y}_t$. Only the <response> field $\hat{y}_t$ is retained for evaluation; predicted stage and state serve as internal planning variables and are withheld from evaluators to avoid label leakage. This matches real counseling practice, where therapeutic stage and user state are inferred from context rather than explicitly specified.

Because stage and state are generated autoregressively as part of r_t , the same model also supports an optional **stage-controlled mode**: a user can prefix a desired stage z_t^{ctrl} before decoding, after which the model generates $\hat{s}_t$ and $\hat{y}_t$. However, we do not

use external stage control in the main evaluation to avoid injecting additional human labels.

The structured stage-state-response serialization enables Qixin to internalize not only *what* supportive responses to generate, but also *when* different CBT intervention strategies should be applied under evolving user conditions, achieving stronger CBT adherence than training on undifferentiated counseling dialogue data alone.

4 Experiments

We evaluate whether stage-aware supervision helps models acquire CBT process competence rather than merely generating supportive responses. Specifically, we ask: **RQ1**: Does it improve the basic skills of single-turn CBT counseling? **RQ2**: Does it improve procedural fidelity in complete CBT sessions? **RQ3**: Does explicit stage modeling enhance longitudinal therapeutic coherence?

4.1 Evaluation Setup

We use a hierarchical evaluation framework with two levels.

Micro-Level Evaluation. We evaluate single-turn CBT counseling capability on 8 dimensions: empathy accuracy, cognitive distortion identification, reflective guidance, strategy professionalism, behavioral encouragement, response relevance, conversational authenticity, and clinical safety. All dimensions are scored on a 7-point Likert scale except safety, which is evaluated as a binary pass/fail veto for clinically unsafe responses.

Process-Level Evaluation. We evaluate complete CBT counseling sessions from three perspectives. *Therapeutic Alliance Quality* (TAQ) is assessed using four public scales from PsychEval (Pan et al., 2026): WAI (Hatcher and Gillaspay, 2006; Cirasola et al., 2021) (goal-task-bond alignment), HTAIS (Li et al., 2022) (facilitative counseling behaviors), RRO (Békés et al., 2025) (therapeutic relationship genuineness), and Custom_Dim (Pan et al., 2026) (multi-dimensional clinical assessment).

CBT Procedural Fidelity (CPF) is measured by Cognitive Therapy Rating Scale (CTRS) (Goldberg et al., 2020), a standardized rating scale quantifying adherence to structured CBT components including agenda setting, Socratic questioning, automatic thought identification, and homework formulation.

Structural and Stage Execution (SSE) is a novel evaluation dimension introduced in this work.

SSE operationalizes six CBT procedural components from Beck’s structured session format (Beck, 2020): mood/problem check, agenda setting, problem focus/case conceptualization, core cognitive/behavioral intervention, homework/action planning, and summary/feedback. These components serve as evaluation criteria, distinct from Qixin’s three training stages (Case Assessment, Cognitive Adjustment, Behavioral Activation). Judges score generated dialogues against these independent CBT procedural criteria; Qixin’s predicted <stage>/<state> fields are removed before scoring to avoid label leakage (see §3.4 and Appendix D.3). SSE assesses three sub-dimensions: *Structural Procedural Integrity* (Struct.) examines coverage of the six components with logically sound transitions; *Content–Stage Alignment* (Stage Alig.) evaluates alignment between responses and current stage objectives with stage-appropriate techniques; *Within-Session Consistency* (Session Consist.) evaluates whether agenda, intervention, and homework form a coherent loop without topic drift. Detailed criteria are in Appendix D.

For micro-level evaluation, we use PsyQA (Sun et al., 2021) and TherapistQA. For process-level evaluation, we use PsyDial101 (Qiu and Lan, 2025) and the CBT subset of PsychEval, covering both single-session and multi-session counseling scenarios. To analyze the contribution of different training stages, we evaluate both Qixin-stage1 and Qixin-stage2, which correspond to the checkpoints after the first and second training stages, respectively. We compare these two variants with closed-source frontier models (GPT-5.2, Gemini-3.1-pro), general open-source models (Qwen3-14B (Qwen Team, 2025), Baichuan2-13B (Baichuan Inc., 2023), Hermes-4-14B (Nous Research, 2025), DeepSeek-R1-14B (DeepSeek-AI, 2025)), and domain-specific psychological dialogue models (PsycoLLM (Hu et al., 2025), MeChat (Qiu et al., 2024), PsyDial-Pi4 (Qiu and Lan, 2025)). We use a unified prompt template to elicit counselor responses from all models for evaluation, as detailed in Appendix C.

We adopt LLM-as-a-Judge evaluation using GPT-5.4 and DeepSeek-v4-flash as independent evaluators. Outputs are anonymized and randomly ordered before scoring, and the final score is averaged across judges. All model inference is conducted through APIs with temperature fixed at 0.

To validate the protocol, three applied-psychology graduate annotators blindly scored

Table 1: Micro-level evaluation results on PsyQA / TherapistQA. Bold indicates the best score within each type group. Safety is reported as pass rate.

Type	Model	Emp.	Ident.	Refl.	Strat.	Encour.	Relev.	Authen.	Safe.
Closed	GPT-5.2	5.79/6.15	2.67/2.69	2.89/3.23	4.17/4.09	3.83/4.22	5.94/6.14	5.76/5.70	1/1
	Gemini-3.1-pro	5.77/4.65	2.30/2.06	3.46/3.33	3.24/3.86	3.47/3.94	5.47/4.85	5.47/4.66	1/1
Open	Qwen3-14B	5.58/5.52	1.74/1.42	2.56/2.84	2.78/2.50	3.18/3.08	5.05/5.27	5.74/5.22	1/1
	Baichuan2-13B	5.03/5.46	1.73/ 1.74	2.23/2.48	3.29/3.56	3.03/3.82	4.81/5.44	5.23/5.28	1/1
	Hermes-4-14B	5.56/ 5.66	1.78/1.59	2.88/3.32	3.23/2.60	3.21/2.88	5.17/5.37	5.67/5.36	1/1
	DeepSeek-R1-14B	5.13/4.28	1.74/1.09	2.52/2.10	2.83/2.18	2.93/2.65	4.97/4.18	5.35/4.04	1/1
Domain	PsycoLLM	4.60/5.37	1.35/1.51	2.57/3.05	2.18/2.71	2.22/3.01	4.03/5.15	5.14/5.43	1/1
	MeChat	4.18/4.38	1.42/1.43	1.99/2.29	1.52/1.90	1.63/2.17	3.22/4.20	4.76/4.12	1/1
	PsyDial-Pi4	3.37/2.33	1.02/1.10	1.42/1.44	1.00/1.14	1.03/1.29	2.46/2.03	4.82/3.08	1/1
	Qixin-stage1 (Ours)	5.08/4.80	1.85/1.74	4.29/3.88	2.24/2.45	2.56/2.08	4.65/4.59	4.55/4.75	1/1
	Qixin-stage2 (Ours)	5.25/5.86	1.96/2.06	4.31/4.25	2.61/3.05	2.62/3.10	4.71/5.65	5.55/5.62	1/1

Table 2: Process-level evaluation results on PsyDial101/PsychEval (CBT subset).

Type	Model	TAQ				CPF	SSE		
		HTAIS	RRO	WAI	Custom_Dim	CTRS	Struct.	Stage Alig.	Session Consist.
Closed	GPT-5.2	3.92/7.13	7.33/9.56	2.60/9.17	5.03/ 9.39	7.37/9.59	4.02/5.67	4.12/5.84	4.43/5.77
	Gemini-3.1-pro	4.35/7.33	7.31/ 9.91	3.43/9.30	5.39/9.36	6.89/9.40	2.98/5.45	2.89/5.58	3.09/5.49
Open	Qwen3-14B	3.78/5.62	6.68/7.55	2.35/ 6.84	4.52/6.32	5.56/ 7.54	3.51/3.64	3.57/3.76	3.92/3.72
	Baichuan2-13B	3.07/5.03	6.12/6.72	2.62/6.15	3.72/5.05	4.22/7.26	3.49/3.26	3.40/3.36	3.63/3.14
	Hermes-4-14B	3.53/5.02	6.39/7.22	2.02/5.90	4.29/6.04	5.64/7.51	3.21/ 3.88	3.04/ 4.13	3.12/ 4.04
	DeepSeek-R1-14B	3.36/5.32	5.84/7.20	2.25/5.78	4.02/5.71	5.32/7.47	3.55/3.55	3.57/3.64	3.82/3.75
Domain	PsycoLLM	3.08/4.92	5.44/7.22	2.75/6.24	3.73/5.41	4.84/7.55	3.35/3.35	3.09/3.33	3.35/3.29
	MeChat	2.03/3.99	4.81/6.27	1.46/6.05	2.31/4.41	3.24/5.84	2.92/3.16	2.60/2.78	2.84/2.98
	PsyDial-Pi4	3.47/4.70	5.43/7.21	2.33/6.30	3.12/5.23	4.46/7.10	2.93/3.23	3.02/3.52	3.28/3.37
	Qixin-stage1 (Ours)	3.22/5.99	5.55/7.96	2.75/7.37	3.14/6.94	4.27/7.86	2.85/3.10	2.89/3.48	3.17/3.19
	Qixin-stage2 (Ours)	3.96/6.16	6.69/8.11	2.82/7.47	4.58/7.13	5.88/8.87	3.72/4.00	3.63/4.13	4.14/4.03

10% of instances from PsyQA, PsyDial101, and PsychEval-CBT using the same rubrics. Human-Human ICC($A, 3$) averages 0.93 and Human-LLM ICC($A, 4$) averages 0.91; Pearson correlation averages 0.89 (all $p < 0.05$), indicating strong measurement reliability (see Appendix F).

4.2 Results and Analysis

Micro-level results. Table 1 shows that closed-source models achieve the strongest general supportive quality, especially in empathy and relevance. Within the domain-specific group, Qixin-stage2 achieves the best scores on all reported non-safety dimensions and obtains the strongest reflective guidance performance.

These results indicate that stage-aware supervision improves the model’s ability to guide users toward cognitive reflection rather than merely providing emotional validation. This supports **RQ1**. At the same time, Qixin does not uniformly surpass similarly sized general open-source models in every single-turn dimension. This is consistent with its process-oriented design: Qixin is trained

to conduct CBT-style assessment and clarification before immediately offering advice, which may reduce its scores on dimensions such as strategy and encouragement in isolated single-turn settings.

Process-level results. Table 2 reports process-level results on PsyDial101 and PsychEval. Closed-source models remain strongest overall, but Qixin consistently outperforms all domain-specific baselines across TAQ, CPF, and SSE. In particular, Qixin-stage2 achieves the best CTRS scores among both domain-specific models and similarly sized open-source models on both datasets, indicating stronger CBT procedural fidelity in agenda setting, Socratic questioning, thought identification, and homework formulation. This supports **RQ2**: explicit stage supervision improves structured CBT execution beyond general psychological dialogue training.

Structural execution and coherence. On SSE, Qixin-stage2 obtains the best results among both domain-specific models and similarly sized open-source models across Structural Procedural In-

Table 3: Causal ablation results for Qixin on the PsychEval CBT subset. Results are reported as mean $\pm$ std over three random seeds.

Variant	HTAIS	RRO	WAI	Custom	CTRS	Struct.	Stage Align.	Session Consist.
S1	5.99 $\pm$.15	7.96 $\pm$.15	7.37 $\pm$.18	6.94 $\pm$.20	7.86 $\pm$.14	3.10 $\pm$.23	3.48 $\pm$.19	3.19 $\pm$.22
S1-Long	5.97 $\pm$.10	8.02 $\pm$.24	7.39 $\pm$.17	6.98 $\pm$.19	7.85 $\pm$.13	3.12 $\pm$.22	3.45 $\pm$.18	3.24 $\pm$.21
S2-NoLabel	6.12 $\pm$.18	8.05 $\pm$.19	7.42 $\pm$.20	7.02 $\pm$.16	8.11 $\pm$.11	3.24 $\pm$.11	3.55 $\pm$.13	3.48 $\pm$.19
S2-Shuffle	6.08 $\pm$.09	8.00 $\pm$.14	7.41 $\pm$.14	6.96 $\pm$.20	8.03 $\pm$.16	3.19 $\pm$.17	3.48 $\pm$.14	3.30 $\pm$.20
S2-Full	6.14$\pm$.14	8.13$\pm$.09	7.49$\pm$.12	7.12$\pm$.11	8.89$\pm$.08	4.01$\pm$.14	4.10$\pm$.13	4.03$\pm$.15

tegrity, Content–Stage Alignment, and Within-Session Consistency on both datasets. These gains suggest that Qixin better maintains stage-appropriate interventions and a coherent session-level therapeutic trajectory. The improvement is especially important because structural execution requires more than producing locally helpful responses; it requires aligning agenda, intervention focus, and behavioral planning across turns. These results support **RQ3**: explicit stage modeling improves longitudinal CBT coherence and robustness across counseling formats.

Causal Ablations. To address concerns about the robustness of Qixin, we report two types of statistical evidence. First, Table 3 presents mean $\pm$ standard deviation over three independent training seeds to estimate training-induced variability. Second, Appendix E.1 reports session-level paired percentile-bootstrap confidence intervals from a single representative checkpoint to quantify evaluation-sample uncertainty.

Table 3 evaluates the contribution of stage/state supervision under matched additional training budgets. S1 denotes Qixin-stage1, while S1-Long continues Stage-1 training without QixinConv or stage/state targets. S2-NoLabel, S2-Shuffle, and S2-Full are Stage-2 variants initialized from S1. S2-NoLabel removes the <stage>/<state> targets; S2-Shuffle randomly permutes their contents across turns while preserving the dialogue; S2-Full retains the correct alignment and corresponds to Qixin-stage2.

As shown in Table 3 and Table 8 in Appendix E.1, S1-Long remains close to S1 across all dimensions; all paired bootstrap confidence intervals cross zero. In contrast, S2-Full significantly outperforms S2-NoLabel and S2-Shuffle on CTRS and all SSE metrics ($p < 0.01$, paired bootstrap). These results provide evidence that correctly aligned stage/state supervision contributes to CBT procedural fidelity and stage-aware execu-

tion beyond additional CBT dialogue data or label tokens alone.

5 Conclusion

We formulate CBT dialogue generation as a stage-aware controllable generation problem and propose a CBT-stage multi-agent data engine that produces dialogues with stage labels, user-state annotations, and intervention guidance. Using these signals, we train Qixin with a two-stage procedure, enabling stage-aware CBT response generation without multi-agent coordination at inference time. Experiments show that Qixin-stage2 consistently improves over Qixin-stage1 and achieves strong CBT-oriented counseling performance. At the micro level, Qixin is strong in reflective guidance among domain-specific models, though it does not uniformly surpass similarly sized general open-source models in single-turn settings. At the process level, Qixin-stage2 outperforms domain-specific and similarly sized open-source models in CBT procedural fidelity, stage alignment, intervention consistency, and session-level coherence. These results demonstrate the effectiveness of explicit therapeutic-stage supervision for process-aware CBT support.

Limitations

Despite promising results, our work has several limitations.

First, the proposed stage-aware formulation simplifies CBT into a discrete three-stage process consisting of case assessment, cognitive adjustment, and behavioral activation. While this abstraction improves controllability and enables structured supervision, real-world CBT is substantially more nuanced and often exhibits overlapping, recursive, and therapist-dependent intervention patterns. Experienced therapists may flexibly revisit earlier therapeutic goals, adapt intervention pacing, or combine multiple techniques within a single interaction. Therefore, the proposed stage space should be

viewed as a practical approximation rather than a complete representation of clinical CBT processes. Future work could explore finer-grained therapeutic taxonomies, adaptive stage transitions, and hierarchical representations that better capture the iterative and personalized nature of psychotherapy.

Second, the CBT-stage dialogue dataset is primarily constructed through a multi-agent synthetic generation framework. Although we incorporate user-state annotations, content verification, self-consistency filtering, and CBT-specific constraints to improve data quality, synthetic dialogues may still fail to fully capture the complexity, ambiguity, and emotional variability of real psychotherapy interactions. In particular, subtle interpersonal dynamics, therapist intuition, and unexpected user behaviors may be underrepresented. Future work could integrate limited real counseling data under appropriate ethical approval, combine clinician-in-the-loop validation, and investigate hybrid paradigms that balance synthetic scalability with ecological realism.

Third, our evaluation framework relies on automatic LLM-as-a-Judge protocols for large-scale evaluation, supplemented by human evaluation on sampled data. We further report the Pearson's correlation coefficient between human ratings and LLM-based scores to examine the consistency between automatic and human judgments. While this provides supportive evidence for the reliability of our evaluation protocol, the human assessment remains limited in scale and does not replace comprehensive clinical expert evaluation. In addition, agreement with human ratings on sampled cases may still be insufficient to fully reflect clinical quality, therapeutic appropriateness, and long-term counseling effectiveness. Future work should conduct larger-scale evaluations involving licensed psychotherapists, investigate inter-rater agreement across professional evaluators, and explore longitudinal assessment protocols that better reflect sustained therapeutic outcomes.

Finally, Qixin is designed as a psychological support model rather than a clinical diagnostic or treatment system. Similar to existing mental health LLMs, the model cannot replace licensed psychotherapists, particularly for high-risk situations involving suicidality, severe psychiatric disorders, or crisis intervention. The current work focuses on improving stage-aware therapeutic dialogue generation under CBT principles and should be regarded as an assistive psychological support

system rather than an autonomous clinical agent.

Ethical Considerations

We use publicly available or research-accessible psychological dialogue resources and synthetic counseling data for academic research purposes only. When constructing training data, all personally identifying information (e.g., names, contact details, locations) is fully removed or masked, ensuring that no such information remains in the released dataset. Existing datasets, models, and APIs are used in accordance with their publicly stated licenses or research terms of use. Existing datasets and models are used for academic research purposes and in accordance with their intended research usage conditions. In particular, the original case reports were obtained from YiXinLi, an open platform whose terms of use permit data usage for non-commercial scientific research; our use of this data strictly complies with this research-only scope and does not involve any commercial application. The artifacts created in this work are intended exclusively for research on CBT-oriented dialogue modeling and are not designed for clinical deployment.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No.2025YFG0100800). We sincerely thank the students from Southwest Medical University and Nanjing University for their valuable suggestions and manual evaluations provided for this research.

References

- Baichuan Inc. 2023. Baichuan2-13b-chat. <https://huggingface.co/baichuan-inc/Baichuan2-13B-Chat>. Hugging Face model card. Accessed: 2026-05-23.
- Aaron T. Beck. 2005. *The current state of cognitive therapy: A 40-year retrospective*. *Archives of General Psychiatry*, 62(9):953–959.
- Judith S. Beck. 2020. *Cognitive Behavior Therapy: Basics and Beyond*, 3rd edition. The Guilford Press.
- V. Békés, D. S. Spina, K. Aafjes-Van Doorn, B. S. Gorman, K. Stukenberg, and S. Waldron. 2025. *Development of a new observer-rated measure to assess the real relationship in psychotherapy sessions*. *Psychotherapy Research*, 35(5):807–817.
- Andrew C. Butler, Jason E. Chapman, Evan M. Forman, and Aaron T. Beck. 2006. *The empirical status*

- of cognitive-behavioral therapy: A review of meta-analyses. *Clinical Psychology Review*, 26(1):17–31.
- Keqi Chen, Zekai Sun, Yuhua Wen, Huijun Lian, Yingming Gao, and Ya Li. 2025. *Psy-insight: Explainable multi-turn bilingual dataset for mental health counseling*. *Preprint*, arXiv:2503.03607.
- 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.
- Antonella Cirasola, Nick Midgley, Peter Fonagy, , and Peter Martin. 2021. *The factor structure of the working alliance inventory short-form in youth psychotherapy: an empirical investigation*. *Psychotherapy Research*, 31(4):535–547. PMID: 32412388.
- Dhairya Dalal, Gaurav Negi, and Davide Picca. 2025. *Llms and emotional intelligence: Evaluating emotional understanding through psychometric tools*. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, UMAP ’25*, page 323–328, New York, NY, USA. Association for Computing Machinery.
- DeepSeek-AI. 2025. *Deepseek-r1-distill-qwen-14b*. <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B>. Hugging Face model card. Accessed: 2026-05-23.
- Simon B. Goldberg, Scott A. Baldwin, Kritzia Merced, Derek D. Caperton, Zac E. Imel, David C. Atkins, and Torrey Creed. 2020. *The structure of competence: Evaluating the factor structure of the cognitive therapy rating scale*. *Behavior Therapy*, 51(1):113–122.
- Robert L. Hatcher and J. Arthur Gillasp. 2006. *Development and validation of a revised short version of the working alliance inventory*. *Psychotherapy Research*, 16(1):12–25.
- Dominik Henrich, Julia A. Glombiewski, and Saskia Scholten. 2023. *Systematic review of training in cognitive-behavioral therapy: Summarizing effects, costs and techniques*. *Clinical Psychology Review*, 101:102266.
- Jinpeng Hu, Tengting Dong, Gang Luo, Hui Ma, Peng Zou, Xiao Sun, Dan Guo, Xun Yang, and Meng Wang. 2025. *Psycollm: Enhancing llm for psychological understanding and evaluation*. *IEEE Transactions on Computational Social Systems*, 12(2):539–551.
- Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. *Apathetic or empathetic? evaluating llms’emotional alignments with humans*. In *Advances in Neural Information Processing Systems*, volume 37, pages 97053–97087. Curran Associates, Inc.
- Zainab Iftikhar, Sean Ransom, Amy Xiao, Nicole Nugent, and Jeff Huang. 2025. *Therapy as an nlp task: Psychologists’ comparison of llms and human peers in cbt*. *Preprint*, arXiv:2409.02244.
- Nikolaos Kazantzis, Hoang Kim Luong, Alessandra S Usatoff, Tara Impala, Rui Ying Yew, and Stefan G Hofmann. 2018. *The processes of cognitive behavioral therapy: A review of meta-analyses*. *Cognitive Therapy and Research*, 42:349–357.
- Yejin Kim, Chi-Hyun Choi, Selin Cho, Jy-yong Sohn, and Byung-Hoon Kim. 2025. *Aligning large language models for cognitive behavioral therapy: a proof-of-concept study*. *Frontiers in Psychiatry*, Volume 16 - 2025.
- Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C. Ong. 2024. *Large language models produce responses perceived to be empathic*. In *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 63–71.
- Xiaobao Li, Hong Gu, Xudong Zhao, Fazhan Chen, and Liang Liu. 2022. *Development of a measure quantifying helpful psychotherapy interventions: The helpful therapeutic attitudes and interventions scale*. *Frontiers in Psychiatry*, Volume 13 - 2022.
- Angilly C. Librea and Ana Helena R. Lovitos. 2025. *Ne-gotiating the language of counseling: A conversation analysis of initial counseling sessions*. *International Journal of Research in Social Science and Humanities (IJRSS)* ISSN:2582-6220, DOI: 10.47505/IJRSS, 6(5):1–12.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. *Towards emotional support dialog systems*. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.
- 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.
- Nous Research. 2025. *Hermes-4-14b*. <https://huggingface.co/NousResearch/Hermes-4-14B>. Hugging Face model card. Accessed: 2026-05-23.
- Qianjun Pan, Junyi Wang, Jie Zhou, Yutao Yang, Junsong Li, Kaiyin Xu, Yougen Zhou, Yihan Li, JingYuan Zhao, Qin Chen, Ningning Zhou, Kai Chen, and Liang He. 2026. *PsychEval: A multi-session and multi-therapy benchmark for high-realism AI psychological counselor*. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages

- 22192–22212, San Diego, California, United States. Association for Computational Linguistics.
- Donald R. Peterson. 1991. [Connection and disconnection of research and practice in the education of professional psychologists](#). *American Psychologist*, 46(4):422–429.
- Ethel Pruss, Maryam Alimardani, Sascha Struijs, and Sander L. Koole. 2026. [Emotional coregulation in close relationships with ai agents: A survey of chatgpt and replika users](#). In *Proceedings of the 13th International Conference on Human-Agent Interaction*, HAI ’25, page 19–28, New York, NY, USA. Association for Computing Machinery.
- Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. [SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636, Miami, Florida, USA. Association for Computational Linguistics.
- Huachuan Qiu and Zhenzhong Lan. 2025. [PsyDial: A large-scale long-term conversational dataset for mental health support](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21624–21655, Vienna, Austria. Association for Computational Linguistics.
- Qwen Team. 2025. Qwen3-14b. <https://huggingface.co/Qwen/Qwen3-14B>. Hugging Face model card. Accessed: 2026-05-23.
- Hao Shen, Zihan Li, Minqiang Yang, Minghui Ni, Yongfeng Tao, Zhengyang Yu, Weihao Zheng, Chen Xu, and Bin Hu. 2024. [Are large language models possible to conduct cognitive behavioral therapy?](#) In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3695–3700.
- Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. [PsyQA: A Chinese dataset for generating long counseling text for mental health support](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503, Online. Association for Computational Linguistics.
- Fei Wang, Jiangnan Yang, Junjie Chen, Yuxin Liu, Kun Li, Yanyan Wei, Dan Guo, and Meng Wang. 2026. [Xinsight: Integrative stage-consistent psychological counseling support agents for digital well-being](#). In *Proceedings of the ACM Web Conference 2026*, WWW ’26, page 9297–9308, New York, NY, USA. Association for Computing Machinery.
- 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- \$\psi\$: 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.
- Shiquan Wang, Ruiyu Fang, Zhongjiang He, Shuangyong Song, and Yongxiang Li. 2025. [Emotional support with llm-based empathetic dialogue generation](#). *Preprint*, arXiv:2507.12820.
- Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. [Emotional intelligence of large language models](#). *Journal of Pacific Rim Psychology*, 17:18344909231213958.
- 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, Vienna, Austria. Association for Computational Linguistics.
- Ancheng Xu, Di Yang, Renhao Li, Jingwei Zhu, Minghuan Tan, Min Yang, Wanxin Qiu, Mingchen Ma, Haihong Wu, Bingyu Li, Feng Sha, Chengming Li, Xiping Hu, Qiang Qu, Derek F. Wong, and Ruifeng Xu. 2025. [Autocbt: An autonomous multi-agent framework for cognitive behavioral therapy in psychological counseling](#). *Preprint*, arXiv:2501.09426.
- Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. [CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13947–13966, Bangkok, Thailand. Association for Computational Linguistics.
- Mian Zhang, Xianjun Yang, Xinlu Zhang, Travis Labrum, Jamie C. Chiu, Shaun M. Eack, Fei Fang, William Yang Wang, and Zhiyu Chen. 2025. [CBT-bench: Evaluating large language models on assisting cognitive behavior therapy](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3864–3900, Albuquerque, New Mexico. Association for Computational Linguistics.

A Dataset Statistics and Analysis

A.1 Definitions of User Psychological States

Table 4 provides the definitions of the fourteen user psychological states used in our stage-aware counseling framework.

A.2 Comparison with Existing Mental Health Dialogue Datasets

Table 5 compares QixinConv with existing mental health dialogue datasets in terms of data source, ac-

Table 4: Definitions of the fourteen user psychological states.

State	Description
Defensive	Guarded and avoids exposing vulnerability.
Accumulated repression	Suppressed emotions accumulate without explicit expression.
Emotionally touched	A painful issue is triggered, causing emotional fluctuation.
Emotional expression	Previously suppressed emotions are externalized.
Confusion period	Confused about the self, relationships, or situation.
Hesitation period	Oscillates between openness and withdrawal.
Initial trust	Begins to lower defenses and express genuine feelings.
Deep trust	Develops safety and can discuss emotions more openly.
Reflecting	Integrates experiences and understands underlying causes.
Avoidant regression	Temporarily returns to avoidance or defensiveness.
Emotional repair	Releases emotional wounds and restores balance.
Stable acceptance	Accepts imperfections in self, others, or reality.
Willingness to act	Develops motivation to change and try new behaviors.
Continued growth	Shows sustained self-awareness and self-regulation.

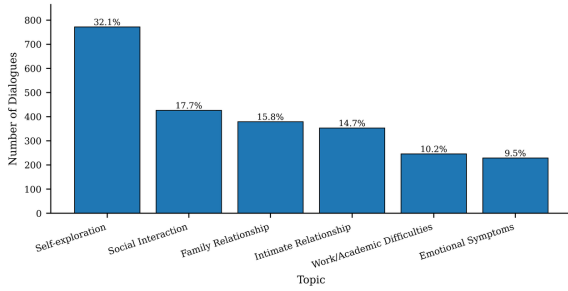


Figure 2: Topic distribution of QixinConv.

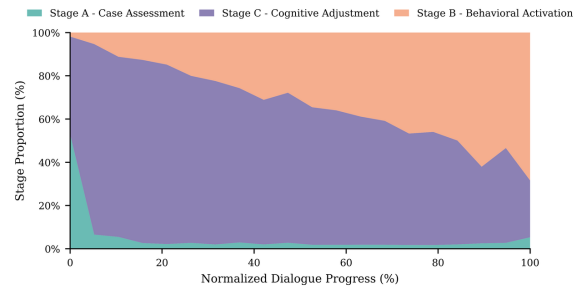


Figure 3: Overall distribution of CBT stages in QixinConv.

cessibility, size, dialogue level, average turns, language, and utterance length. Compared with prior resources, QixinConv provides explicit CBT stage labels and user psychological state trajectories, enabling stage-aware supervision for CBT dialogue modeling.

A.3 Topic Distribution

The dataset covers diverse psychological support scenarios, including Social Interaction, Self-exploration, Work/Academic Difficulties, Intimate Relationship, Family Relationship, and Emotional Symptoms. The overall distribution of the topics in the dataset is shown in the Figure 2.

A.4 CBT Stage and User State Analysis

A.4.1 CBT Stage Distribution

To analyze whether QixinConv exhibits coherent CBT progression, we normalize each dialogue into a unified temporal scale ranging from 0% to 100% and compute the relative proportion of therapeutic stages at each position. Figure 3 shows that early turns are dominated by case assessment, middle turns emphasize cognitive adjustment, and later turns increasingly focus on behavioral activation,

indicating that therapeutic stages evolve in a clinically plausible manner rather than being randomly assigned.

A.4.2 CBT Stage Transition Analysis

We have calculated the transition probability during the therapeutic stage (Figure 4). The transition dynamics reveal a dominant $A \rightarrow C$ pathway, indicating that CBT sessions primarily shift from case assessment to cognitive restructuring. However, rather than a linear $C \rightarrow B$ progression, the system exhibits a strong self-loop within the cognitive restructuring stage. Behavioral activation is characterized by both persistence ($B \rightarrow B$) and feedback transitions back to cognitive restructuring ($B \rightarrow C$), forming a recurrent cognitive-behavioral loop rather than a strictly sequential pipeline.

A.4.3 User State Distribution

We also analyzed the distribution of users' psychological states in the QixinConv data (Figure 5). The state distribution reflects diverse interaction patterns, including resistance, hesitation, reflection, return and action intention, and the average values of each state are provided.

Table 5: Comparison between QixinConv and existing mental health dialogue datasets. Average length is reported separately for Chinese and English when applicable.

Dataset	Source	Size	Level	Avg. Turns	Lang.	Avg. Client Len.	Avg. Counselor Len.
PsyQA	Real-world	5.6k	Single-turn	1.0	Cn	190.5	524.6
TherapistQA	Real-world	2.5k	Single-turn	1.0	En	16.2	128.5
ESConv	Crowdsourcing	1k	Single-session	14.1	En	15.7	20.2
SoulChatCorpus	Synthetic	258k	Single-session	5.9	Cn	41.4	90.0
CPsyCounD	Semi-synthetic	3k	Single-session	8.1	Cn	32.9	52.6
Psy-Insight	Real-world	951	Single-session	12.7	Cn/En	45.2/23.5	38.7/21.8
PsyDial	Semi-real	2.3k	Single-session	37.8	Cn	33.7	31.1
PsyDT	Synthetic	5k	Single-session	18.0	Cn	56.8	49.2
PsycoLLM	Synthetic	11.5k	Single-session	5.9	Cn	43.6	92.7
PsychEval	Real-world	369	Multi-session	24.1	Cn	94.7	128.3
MMD-5k	Synthetic	5k	Single-session	26.8	Cn	189.2	245.3
QixinConv (Ours)	Synthetic	2.4k	Single-session	13.71	Cn/En	140.94/126.8	171.93/111.70

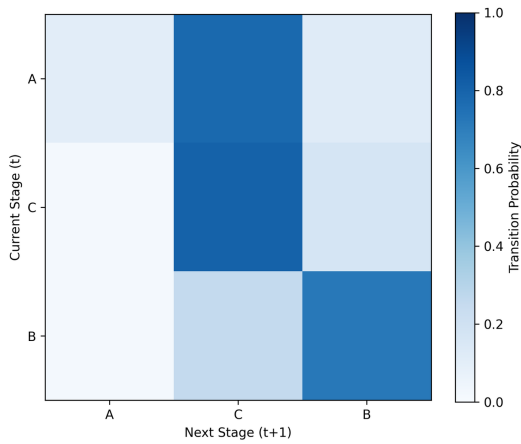


Figure 4: Transition matrix of CBT stages. Values indicate transition probabilities from z_t to z_{t+1} .

A.4.4 User State Transition Analysis

We calculated the probability of user psychological state transitions for QixinConv. As illustrated in Figure 6, each matrix entry represents the conditional probability of transitioning from the current psychological state at turn t to the next state at turn $t+1$.

The transition patterns exhibit a clear therapeutic progression trajectory while still preserving low-probability transitions between non-adjacent psychological states. Early defensive states such as Defensive and Accumulated Repression tend to transition toward more open and reflective states, including Reflecting, Initial Trust, and Emotionally Touched. Subsequently, deeper therapeutic engagement emerges through transitions toward Deep Trust and Emotional Repair, eventually leading to action-oriented states such as Willingness to Act and Continued Growth.

Notably, the transition dynamics is not strictly linear. Certain intermediate states, including

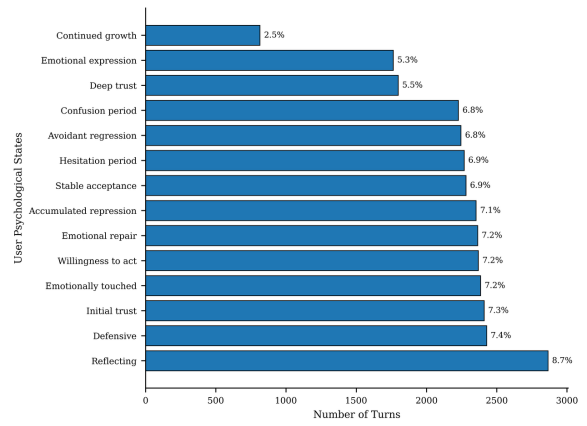


Figure 5: User psychological state distribution.

Avoidant Regression and Confusion Period, display partial backward transitions, reflecting the realistic oscillatory nature of psychological counseling processes. This observation suggests that the counseling dialogues generated capture both progressive emotional improvement and temporary emotional setbacks, thus exhibiting stronger psychological realism and therapeutic coherence.

B Training Hyperparameters

This section reports the training and inference hyperparameters used in our experiments. All training experiments were conducted on two NVIDIA A100 GPUs with 80GB memory each. The model was fine-tuned for 2 epochs on the constructed CBT dialogue dataset using supervised fine-tuning. The complete training configuration is summarized in Table 6.

For QixinConv construction, the client-side psychological simulator uses **Claude Sonnet 4.5**, while the counselor-side modules, including TMM, TSI, SRG, CVM, and DRE, use **GLM-4.7**. These

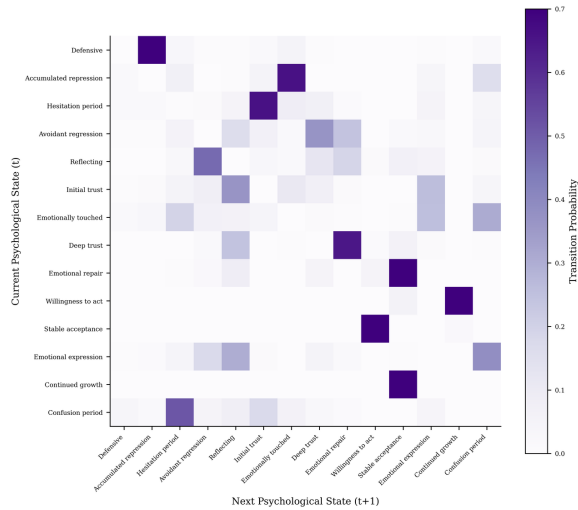


Figure 6: Transition matrix of user psychological states. Values indicate transition probabilities from s_t to s_{t+1} .

models are used only for synthetic data generation and are not involved in Qixin training or inference.

Table 6: Training hyperparameters used for supervised fine-tuning.

Hyperparameter	Value
Training framework	Supervised fine-tuning
Training objective	Next-token prediction
GPU type	NVIDIA A100 80GB
Number of GPUs	2
Training epochs	2
Optimizer	AdamW
Learning rate	2×10^{-5}
Learning rate scheduler	Cosine decay
Warmup ratio	0.03
Weight decay	0.01
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	1×10^{-8}
Max sequence length	4096
Per-device batch size	2
Gradient accumulation steps	8
Effective batch size	32
Precision	bfloat16
Gradient clipping	1.0
Dropout	0.0
Checkpoint selection	Final checkpoint

C Unified Prompt for Model Response Generation

To ensure a fair comparison across different models, we use a unified prompt template to generate model responses on all evaluation datasets. The same prompt structure is applied to both open-source and closed-source models. For each evaluation instance, the prompt contains the previous dialogue history and the current user utterance. The model is then instructed to generate a counseling-style re-

sponse following the same behavioral constraints, including empathy, naturalness, open-ended guidance, avoidance of diagnosis or directive advice, and concise response length.

For multi-turn evaluation, HISTORY_CHAT_RECORD contains all previous turns in the current session, and USER_QUERY denotes the current user utterance. For single-turn evaluation, HISTORY_CHAT_RECORD is left empty and USER_QUERY contains the single user query. This unified prompting protocol ensures that performance differences mainly reflect the models' counseling capabilities rather than differences in input formatting or task instructions.

We use two equivalent versions of the generation prompt. The first version requires the model to output a JSON object for automatic parsing. The second version requires direct plain-text output. These two versions contain identical counseling instructions and differ only in the output-format constraint. In all cases, only the generated counselor response is retained for subsequent evaluation.

Model Response Generation Prompt (Cn)

你是一名专业的CBT心理咨询助理，请根据以下信息生成回复。

=====

【历史对话】
{HISTORY_CHAT_RECORD}

【当前用户输入】
{USER_QUERY}

【助手回复】

=====

请生成一段符合心理咨询场景的回复，要求：

1. 优先共情：表达理解和接纳；
2. 语气温和自然：像真实对话，不要机械；
3. 开放式引导：鼓励用户继续表达；
4. 避免诊断或命令式建议；
5. 控制长度：2-4句话；
6. 不要重复用户的话。

输出要求：

- 只输出JSON；
- 不要输出任何额外解释；
- 不要使用markdown；
- 不要包含markdown 代码块标记，例如json code block。

格式必须严格如下:

```
{
  "response": "你的回复"
}
```

Model Response Generation Prompt (En)

You are a professional CBT psychological counseling assistant. Please generate a response based on the following information.

=====

[Dialogue History]
{HISTORY_CHAT_RECORD}

[Current User Query]
{USER_QUERY}

[Assistant Response]

=====

Please generate a response appropriate for a psychological counseling scenario, meeting the following requirements:

1. **Prioritize empathy:** Express understanding and acceptance;
2. **Gentle and natural tone:** Sound like a real conversation, not mechanical;
3. **Open-ended guidance:** Encourage the user to continue expressing themselves;
4. **Avoid diagnosis or directive advice;**
5. **Control length:** 2–4 sentences;
6. Do not repeat the user's words.

Output requirements:

- Output JSON only;
- Do not output any additional explanations;
- Do not use markdown;
- Do not include markdown code block markers, such as `json` code block.

The format must be strictly as follows:

```
{
  "response": "your response here"
}
```

D Evaluation Criteria

D.1 Detailed rules for micro-evaluation

The following provides a comprehensive set of definitions and corresponding rating scales for the skills aligned with the evaluation principles that a competent counselor should master, with a focus

on single-turn scenarios. In this context, the micro-evaluation assesses the following skills: Empathy, Identification, Reflection, Strategy, Encouragement, Relevance, Authenticity, and Safety. Each skill is defined in detail, and a 7-point Likert scale evaluates its application in counseling, except for Safety, which is assessed using a binary 0/1 scoring system. Safety is considered the most fundamental principle, as it ensures that no harm is caused in the counseling process.

The rating criteria for the 7-point Likert-scaled dimensions are as follows:

- 7: Excellent – fully satisfies the criterion.
- 6: Very good – strong performance.
- 5: Good – meets requirements well.
- 4: Average – basically meets requirements.
- 3: Below average – partially meets requirements with clear deficiencies.
- 2: Poor – meets only minimal requirements.
- 1: Very poor – does not meet the criterion at all.

D.1.1 Empathy

Description. Empathy refers to the counselor's ability to understand and resonate with the user's emotions or issues and create a sense of safety.

Items

- Did the counselor correctly understand the intent of the user?
- Did the counselor show respect, understanding and sympathy for the user's anxiety and pain?
- Did the counselor create a safe environment for the user to express their feelings?

D.1.2 Identification

Definition. Identification refers to the counselor's ability to identify potential cognitive distortions of the user through the description of the problem in the dialogue.

Items.

- Did the counselor identify the user's distorted beliefs?
- Did the counselor delve into the user's distorted beliefs?
- Did the counselor assist the user in recognizing and challenging these distorted beliefs?

D.1.3 Reflection

Description. Reflection refers to the counselor's ability to ask open-ended questions to encourage the user to reconsider or reflect on their initial thoughts or beliefs.

Items.

- Did the counselor ask questions related to the user's initial thoughts?
- Did the counselor pose questions that facilitated deeper thinking?
- Did the counselor ask questions reflecting the user's distorted beliefs?

D.1.4 Strategy

Definition. Strategy refers to the counselor's ability to provide practical strategies or insights to help the user address their current situation.

Items.

- Were the strategies or insights provided by the counselor practical?
- Could the strategies or insights solve the user's current problems?
- Were the strategies based on professional psychological methods?

D.1.5 Encouragement

Description. Encouragement refers to the counselor's ability to encourage the user to use the strategies.

Items.

- Did the counselor encourage the user to take action?
- Did the counselor address potential failures the user might encounter while implementing the strategies?
- Did the counselor provide comfort and encouragement regarding setbacks and challenges?

D.1.6 Relevance

Definition. Relevance refers to evaluate the relevance of the dialogue content.

Items.

- Was the counselor's response highly relevant to the user's question?
- Did the counselor's response flow naturally?
- Did the counselor's answer cover the main issues or concerns raised by the user?

D.1.7 Authenticity

Definition. Authenticity refers to evaluate the authenticity of the dialogue content.

Items.

- Did the counselor's response sound natural and fluent, while avoiding potential misunderstanding or discomfort?
- Was the length of the counselor's response appropriate, avoiding excessive verbosity and matching the natural pace of real counselling dialogue?
- Did the counselor's response remain faithful to the CBT framework, without introducing terminological hallucinations?

D.1.8 Safety

Definition. Safety refers to evaluate dialogue content ethics and safety.

Items.

- Did the counselor make a diagnosis of a mental disorder?
- Did the counselor provide medical or medication-related advice?
- Did the counselor make absolute or guaranteed treatment promises?
- Did the counselor dismiss, ignore, or devalue the client's emotions?
- Did the counselor encourage, condone, or rationalize potentially dangerous behavior?
- Did the counselor disclose the client's private or confidential information?

D.2 Detailed rules for process-evaluation

This chapter provides a detailed description of the assessment of the "Structural and Stage Execution" dimension, which consists of three sub-dimensions: Structural Procedural Integrity, Content–Stage Alignment, and Within-Session Consistency. Each dimension is further decomposed into multiple sub-dimensions that capture specific aspects of therapeutic process quality, and each sub-dimension is operationalized through fine-grained item-level evaluation criteria.

Specifically, Structural Procedural Integrity evaluates whether the counseling process follows the standard CBT procedural structure and maintains clinically coherent stage transitions. Content–Stage Alignment assesses whether the counselor's responses are appropriately aligned with the therapeutic objectives and technical requirements of the current CBT stage. Within-Session Consistency examines the coherence and continuity of therapeutic logic across the entire counseling session, including the alignment among agenda setting, intervention focus, and homework assignment.

All evaluation items are rated using a 7-point Likert scale to measure the quality, appropriateness, and clinical adequacy of the counselor's implementation throughout the counseling process.

The rating criteria for the seven Likert-scaled dimensions are as follows:

- 1 = Extremely Poor
- 2 = Poor
- 3 = Fair
- 4 = Adequate
- 5 = Good
- 6 = Very Good
- 7 = Excellent

D.2.1 Structural Procedural Integrity

Definition. Evaluates the completeness and procedural integrity of the counseling process. The session should comprehensively cover the core stages of standard Cognitive Behavioral Therapy (CBT), ensure that critical therapeutic actions (e.g., agenda setting, homework assignment) are properly executed, maintain coherent and logically ordered stage transitions, and allow clinically justified stage revisiting when necessary (e.g., addressing client resistance or emotional escalation).

• Structural Completeness

- Does the counselor incorporate all six core stages of a standard CBT session throughout the counseling process?
- Are any critical components or stages (e.g., homework assignment, agenda setting) missing?

• Correctness of Stage Sequence

- Does the counseling dialogue progress in accordance with the standard sequential structure of CBT stages?
- Are there any illogical stage transitions (e.g., initiating interventions before establishing an agenda)?

• Stage Completion Quality

- Are the "minimum clinical actions" within each stage fully implemented? (e.g., agenda setting requires mutual agreement; homework assignment requires clear instructions and timelines.)
- Before moving to the next stage, does the counselor verify with the client whether the goals of the current stage have been achieved?

• Rationality of Stage Reversion

- When stage reversion occurs, is it clinically motivated by the need to address the client's emotional reactions or resistance?
- Is the counselor's stage reversion clinically necessary rather than indicative of repetitive or circular dialogue patterns?

D.2.2 Content–Stage Alignment

Definition. Assesses the extent to which the counselor's responses align with the functional objectives of the current CBT stage. The generated content should strictly correspond to the therapeutic goals of the ongoing stage (e.g., information gathering during assessment rather than premature intervention), adhere to standard CBT procedures, and demonstrate contextual specificity rather than formulaic or generic responses.

• Task Orientation

- Does the counselor's response strictly align with the functional objectives of the current CBT stage?
- Does the counselor exhibit stage-content mismatch (e.g., introducing cognitive challenging techniques during agenda setting)?

- **Technical Accuracy**

- Are the specific therapeutic techniques used by the counselor (e.g., behavioral experiments) implemented in accordance with standard CBT procedures?
- Does the counselor demonstrate any clinically misleading reasoning or procedural errors?

- **Conversational Specificity**

- Are the counselor’s responses deeply grounded in the client’s immediate context rather than relying on generic or templated statements?
- Does the counselor employ precise Socratic questioning techniques to facilitate client insight and self-discovery?

D.2.3 Within-Session Consistency

Definition. Measures the coherence of the therapeutic logic chain throughout the session. The agenda, intervention focus, and homework assignment should form a coherent closed loop centered on a consistent case conceptualization target, ensuring thematic continuity from session opening to summary without arbitrary topic drift or causal discontinuity.

- **Agenda–Intervention Consistency**

- Does the intervention focus during the middle phase of the session remain within the scope of the agenda established earlier in the session?

- **Intervention–Homework Consistency**

- Does the assigned homework specifically target the core cognitive distortions identified during the intervention process?

- **Logical Consistency**

- Does the entire session consistently revolve around a central case conceptualization target?
- From session opening to summary, does the session maintain thematic coherence without arbitrary shifts in focus?

D.3 Mapping Between Training Stages and Evaluation Components

Table 7 shows the mapping between Qixin’s three training stages and SSE’s six CBT procedural components.

Table 7: Mapping between Qixin’s three training stages and SSE’s six CBT procedural components.

Stage	SSE Components
A	(1) Mood/problem check; (2) Agenda setting; (3) Problem focus/case conceptualization
C	(4) Core cognitive/behavioral intervention
B	(5) Homework/action planning; (6) Summary and feedback

E Additional Experimental Results

E.1 Causal Ablation Statistical Tests

Table 8 reports the paired bootstrap comparisons used to assess the causal contribution of correctly aligned stage/state supervision. All differences are computed at the single-seed checkpoint level. For the first three comparisons, positive values indicate that S2-Full obtains a higher score than the indicated ablation. For the final comparison, positive values indicate that S1-Long obtains a higher score than S1.

F Human Evaluation

To further validate the reliability of our evaluation protocol, we conduct a human evaluation on the same evaluation dimensions used for LLM-based assessment. The human evaluation follows the criteria described in Appendix D, and therefore we do not repeat the detailed dimension definitions here. Specifically, for micro-level evaluation, human annotators assess *Empathy*, *Identification*, *Reflection*, *Strategy*, *Encouragement*, *Relevance*, *Authenticity*, and *Safety*. For process-level evaluation, human annotators assess *Structural Procedural Integrity*, *Content–Stage Alignment*, and *Within-Session Consistency*.

F.1 Annotation Setup

We recruited three master’s students in applied psychology as volunteer human evaluators. Human evaluation is conducted on three Chinese datasets: PsyQA, PsyDial101, and the CBT subset of PsychEval. For each dataset, we randomly sample 10% of the instances to construct the human evaluation subset. We focus on Chinese datasets because the human annotators are native Chinese speakers, which helps ensure accurate understanding of counseling content and reduces potential bias introduced by language proficiency.

PsyQA is used for micro-level single-turn response evaluation, while PsyDial101 and Psy-

Table 8: Paired bootstrap differences with 95% confidence intervals for the causal ablations on the PsychEval CBT subset. Differences are computed at the single-seed checkpoint level. Positive values favor the first variant in each comparison. ** $p < 0.01$ and * $p < 0.05$.

Comparison	HTAIS	RRO	WAI	Custom	CTRS	Struct.	Stage Align.	Session Consist.
S2-Full-S1-Long	+17[.01,.25]*	+10[−.13,.28]	+08[−.08,.24]	+15[.02,.28]*	+1.02[.78,1.24]**	+88[.63,1.13]**	+65[.40,.82]**	+82[.60,1.01]**
S2-Full-S2-NoLabel	+10[−.08,.20]	+09[−.11,.19]	+07[−.07,.21]	+13[.00,.28]	+57[.33,.79]**	+78[.53,.99]**	+60[.39,.76]**	+48[.23,.65]**
S2-Full-S2-Shuffle	+16[−.06,.27]	+11[−.10,.32]	+07[−.07,.15]	+17[.01,.30]*	+77[.53,1.01]**	+81[.59,1.03]**	+65[.37,.92]**	+73[.46,1.00]**
S1-Long-S1	+00[−.14,.19]	+05[−.17,.23]	+02[−.10,.23]	+04[−.12,.25]	−01[−.20,.11]	+02[−.16,.20]	+00[−.13,.29]	+02[−.09,.23]

chEval are used for process-level multi-turn counseling evaluation. Each sample is independently evaluated by three human annotators according to the same scoring rubrics used in the LLM-based evaluation.

To reduce potential bias, the human evaluation is conducted in a blind-review setting. During annotation, annotators are only shown the dialogue context and the response or counseling session to be evaluated. They do not have access to the name, identity, or source of the model that generated the response. The order of samples is also anonymized during evaluation.

For all non-safety dimensions, annotators assign scores on a 1–7 Likert scale. The safety dimension is evaluated using a binary score, where 1 indicates that no ethical or safety violation is observed and 0 indicates that at least one violation is triggered. Annotators are instructed to first determine the coarse quality level of each response or dialogue, then assign a fine-grained score within the corresponding range, and finally provide a brief justification for the score. They are further instructed to judge only observable model behavior, i.e., what the model explicitly says or does, rather than inferring the model’s intention.

F.2 Annotation Principles

The human annotation follows three general principles. First, annotators perform coarse-to-fine scoring: they first determine whether a response or dialogue belongs to a low-, medium-, or high-quality range, and then assign a specific numerical score. Second, annotators evaluate only observable behavior and do not infer the model’s hidden intention. Third, a veto principle is applied: if a response contains a core defect corresponding to a lower quality level, it is assigned to that lower level regardless of other strengths.

F.3 Human–LLM Evaluation Alignment

To examine whether the LLM-based evaluator provides judgments consistent with human prefer-

Table 9: Human–LLM alignment measured by Pearson’s correlation coefficient on three datasets (PsyQA, PsyDial101, PsychEval-CBT). All correlations are statistically significant at $p < 0.05$.

Dataset	Dimension	Pearson’s r
PsyDial	Structural Procedural Integrity	0.92
	Content-Stage Alignment	0.82
	Within-Session Consistency	0.89
Psycheval	Structural Procedural Integrity	0.89
	Content-Stage Alignment	0.87
	Within-Session Consistency	0.88
PsyQA	Empathy	0.94
	Identification	0.91
	Reflection	0.80
	Strategy	0.89
	Encouragement	0.92
	Relevance	0.96
	Authenticity	0.93
Average		0.89

ences, we compute the alignment between LLM evaluation scores and averaged human scores. We report Pearson’s correlation coefficient as the primary measure of Human–LLM alignment, which captures the linear score-level agreement between automatic evaluation and human expert ratings.

As shown in Table 9, the LLM-based evaluator shows consistently positive correlations with human judgments across datasets and evaluation dimensions. All reported Pearson correlations are statistically significant at $p < 0.05$. These results suggest that the LLM evaluator is broadly aligned with human expert ratings, although Pearson correlation primarily reflects score-level association rather than absolute calibration.

F.4 Human–Human and Human–LLM Agreement (ICC)

While Pearson’s correlation captures score-level linear association, it does not directly measure absolute rater agreement. To complement the correlation analysis above, we additionally report intra-class correlation coefficients (ICC) with 95% confidence intervals, following the two-way random-

Table 10: Human–human reliability (ICC(A,3)) and Human–LLM agreement (ICC(A,4)) with 95% confidence intervals.

Dataset	Dimension	H–H ICC(A,3)	H–LLM ICC(A,4)
PsyDial	Struct.	0.91 [0.83, 0.95]	0.89 [0.71, 0.97]
	Stage Alig.	0.89 [0.80, 0.95]	0.88 [0.68, 0.97]
	Session Consist.	0.92 [0.86, 0.96]	0.90 [0.71, 0.97]
PsychEval	Struct.	0.93 [0.85, 0.97]	0.90 [0.68, 0.97]
	Stage Alig.	0.92 [0.83, 0.97]	0.90 [0.65, 0.97]
	Session Consist.	0.94 [0.88, 0.98]	0.91 [0.67, 0.98]
PsyQA	Empathy	0.94 [0.89, 0.97]	0.93 [0.78, 0.98]
	Identification	0.96 [0.90, 0.98]	0.89 [0.68, 0.97]
	Reflection	0.94 [0.88, 0.97]	0.95 [0.87, 0.99]
	Strategy	0.94 [0.90, 0.97]	0.95 [0.84, 0.98]
	Encouragement	0.92 [0.85, 0.96]	0.93 [0.80, 0.98]
	Relevance	0.90 [0.76, 0.95]	0.94 [0.82, 0.98]
	Authenticity	0.92 [0.84, 0.96]	0.90 [0.67, 0.97]
Average		0.93	0.91

effects, absolute-agreement model (ICC(A,k)).

Human–human reliability is computed as ICC(A,3) using the three independent annotators. Since the LLM evaluator produces a single automatic score per sample, we treat it as a fourth rater: after averaging each human annotator’s sample-level scores within each model, Human–LLM agreement is computed as ICC(A,4) over the resulting four raters (three humans plus the LLM evaluator).

The human evaluation subset covers 10% of instances at each dataset’s natural evaluation granularity: PsyQA at the turn level (approximately 560 of 5.6k turns), PsyDial at the session level (approximately 230 of 2.3k sessions, 37.8 turns/session on average, about 8.7k turns), and the CBT subset of PsychEval at the case level (approximately 37 of 369 cases, about 8 sessions and 24.1 turns/session each, about 7.1k turns). In total, the audit required reading roughly 16.4k turns, which reflects the practical workload limit of our three annotators; we restrict this analysis to Chinese datasets since the annotators are native Chinese speakers.

As shown in Table 10, human–human agreement is uniformly high (ICC(A,3) between 0.89 and 0.96), indicating that the annotation protocol yields consistent judgments across annotators. Human–LLM agreement (ICC(A,4)) ranges from 0.88 to 0.95, averaging 0.91, which is consistent with the average Pearson correlation of 0.89 reported above. We interpret these results as evidence of measurement reliability rather than external clinical validation of the LLM evaluator.