

TRACE: Two-Phase RL for Causal Graph Exploration and Deeper Psychological Intervention in Dynamic Counseling Scenarios

Shilin Tang¹ Zunyi Yin¹ Xuefeng Liang^{1,2*} Guanghui Shi² Song Tong³ Chen Guangyu⁴

¹Guangzhou Institute of Technology, Xidian University

²School of Artificial Intelligence, Xidian University

³Bay Area School of Applied Psychological Sciences, Beijing Normal University at Zhuhai

⁴Information Science, Tohoku University

xliang@xidian.edu.cn, {shilint, zyyin, ghshi}@stu.xidian.edu.cn,

s.tong@bnu.edu.cn, chen.guangyu.s2@dc.tohoku.ac.jp

Abstract

LLMs have shown promise in mental health counseling, but existing models are limited to surface-level empathy or predefined therapeutic procedures and lack the ability to actively explore the root causes of psychological distress. Inspired by case conceptualization, we formalize counseling as the online reconstruction of a client’s underlying causal graph through multi-turn dialogue. To this end, we propose TRACE, a two-phase reinforcement learning framework. It implements a causal-graph-driven reward scheme across two phases: an exploration phase that rewards the causal graph reconstruction following a surface-to-deep path, and an intervention phase that rewards targeted restructuring of irrational beliefs. Extensive experiments show that TRACE outperforms existing models, enabling causal-chain-aware psychological intervention beyond surface-level empathy.

1 Introduction

Large Language Models (LLMs) have emerged as promising tools for mitigating the global shortage of mental health resources, owing to their strong language understanding and generation capabilities. Representative psychological counseling models include CRISP, SoulChat 2.0, SoulChat R1, and CACTUS. These models follow two primary technical paradigms: (1) leveraging large-scale empathetic dialogue corpora to enable models to learn emotion recognition and emotional support strategies (Xie et al., 2025; Fu et al., 2025); (2) incorporating structured therapeutic frameworks, such as Cognitive Behavioral Therapy (CBT), to guide models toward cognitive restructuring (Lee et al., 2024; Zhou et al., 2025). Despite notable progress, existing models lack the ability to actively explore the underlying causes of clients’ psychological distress. As a result,

*Corresponding author

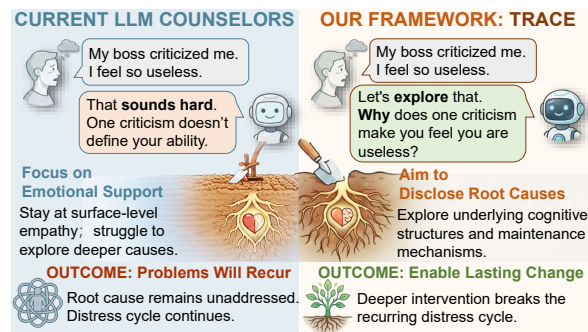


Figure 1: Difference between existing mental-health LLMs and our TRACE framework. They focus on surface-level empathy, while TRACE targets tracing the roots of distress.

they remain limited to passively responding to explicit emotional expressions or following predefined therapeutic procedures. This stands in contrast to human therapists, who gradually trace the roots of distress through multi-turn dialogue, e.g., by identifying core beliefs or early traumatic experiences that trigger negative emotions (see Fig. 1).

These limitations stem from prevailing data-driven learning paradigms. Most models primarily learn correlations between client utterances and counselor responses, capturing *how* to respond to client emotions rather than *why* such responses are appropriate. As they do not model the causal chains linking emotions, cognitions, and behaviors, they fail to identify what causes psychological problems and what mechanisms maintain them (Kuyken et al., 2011; Bieling and Kuyken, 2003). This gap between correlation learning and causal understanding fundamentally restricts the potential for models to advance toward deeper psychological intervention. In clinical practice, this gap is addressed through *case conceptualization* (Haynes et al., 2012; Borsboom and Cramer, 2013), a well-established paradigm that transforms fragmented narratives into a coherent explanatory model by constructing causal chains

from proximal triggering mechanisms (e.g., cognitive distortions) to distal developmental origins (e.g., early trauma). This enables a shift from empirical responding to strategic intervention based on pathological mechanisms. Ong et al. (2025) have shown that LLMs can extract causal graphs from complete narrative texts to achieve case conceptualization. However, real-world counseling differs fundamentally from offline text analysis: counselors cannot access a full psychological profile at the outset, but must guide clients to gradually disclose information through multi-turn dialogue. This makes *online* case conceptualization particularly challenging for LLMs.

To address this challenge, grounded in the psychological principle that “an individual’s mental state is driven by underlying cognitive structures” (Anderson, 2014; Johnson-Laird, 1983), we posit that each client has an underlying *ideal causal graph* connecting from surface-level emotions to deep-level beliefs. Thus, we formalize counseling as a process of progressively reconstructing this causal graph and delivering interventions amid gradually disclosed information.

During this process, the counselor needs to continuously adjust exploration strategies based on the clients feedback, progressively disclosing the root causes of the problem. This dynamic exploration process is inherently a sequential decision-making problem, which naturally aligns with the modeling paradigm of reinforcement learning (RL). We therefore propose a **Two-phase Reinforcement learning** framework for **Active Causal Exploration** of cognitive structures (**TRACE**). Following the clinical principle of “exploration-before-intervention” (Miller and Rose, 2009; Fried et al., 2017), TRACE consists of two phases: (1) *Exploration Phase*, which rewards the model for following a surface-to-deep probing path to gradually disclose psychological elements and their causal maintenance mechanisms, thereby reconstructing the ideal causal graph; and (2) *Intervention Phase*, which rewards the model for performing appropriate cognitive restructuring of identified irrational beliefs based on the reconstructed causal graph. This design addresses the weakness of existing mental-health LLMs that know how to respond to emotions but not what underlies them. It also avoids the risk of premature intervention based on partial information.

Our main contributions are threefold:

- **Task Formulation:** We are the first to for-

malize psychological counseling as a dynamic causal graph reconstruction problem, enabling LLM-based online case conceptualization under gradually disclosed information.

- **Methodology:** We propose TRACE, which models the “exploration-before-intervention” paradigm through two-phase RL with causal-graph-driven rewards.

- **Experiments:** Extensive experiments show that TRACE significantly outperforms existing models in causal graph reconstruction, intervention effectiveness, and general dialogue quality.

2 Related Work

2.1 Computational Case Conceptualization

Case conceptualization aims to organize clinical variables (e.g. symptoms, underlying mechanisms, and developmental origins) into a coherent explanatory causal model (Persons, 2008). Haynes et al. (2020) formalized this process using Directed Acyclic Graphs (DAGs), where nodes represent clinical elements and edges encode causal or maintenance relationships, laying the foundation for computational case conceptualization.

Recently, LLMs have been explored for automated causal structure extraction. In general domains, Guevara et al. (2024) and Takayanagi et al. (2024) demonstrated that LLMs can identify causal nodes and recognize causal relations from unstructured text. This paradigm has been extended to psychology. Tong et al. (2024) combined LLMs with causal knowledge graphs to generate psychological hypotheses, while Ong et al. (2025) extracted personalized symptom networks from complete counseling transcripts to achieve case conceptualization. However, they only demonstrate LLM-based causal graph extraction from complete narratives. None investigates how to reconstruct a causal graph during multi-turn dialogue, where information is disclosed gradually.

2.2 Mental Health LLMs

Early mental-health LLMs primarily improved emotional support through fine-tuning on high-quality dialogue corpora. Chen et al. (2023) and Qiu et al. (2024) extended single-turn dialogue to multi-turn dialogue, Xie et al. (2025) incorporated personalized style matching to enhance adaptability, Fu et al. (2025) adopted active listening to strengthen empathy, and Kim et al. (2025) implemented multi-turn emotion tracking for affective

consistency. However, these methods largely respond passively to explicit emotions, rarely addressing the underlying causes of distress. Recent studies have integrated professional psychotherapy theories into LLMs (Hu et al., 2025). CAC-TUS (Lee et al., 2024) guides models through standardized CBT procedures; SoulChat R1 (Chen et al., 2025) leverages CoT reasoning to maintain long-range dialogue coherence; and CRISP (Zhou et al., 2025) simulates the hierarchical cognitive therapy by identifying the chain from automatic thoughts to core beliefs. Despite progress in cognitive intervention, they typically capture isolated or shallow causal associations and fall short of systematically tracing the deeper maintenance mechanisms underlying psychological distress.

Recently, RL has been introduced into counseling systems due to the dynamic optimization and policy generalization (Chu et al., 2025). Empathy R1 (Yao et al., 2025) and Kardia R1 (Yuan et al., 2025) jointly optimize CoT reasoning and response generation via multi-level rewards. However, they primarily target improving single-turn dialogue quality. Yang et al. (2025b) proposed a motivational interviewing system, CAMI, which focus on eliciting clients’ motivation and guiding topic exploration. In contrast, TRACE aims to progressively reconstruct client’s cognitive structure over multi-turn dialogue. This enables the model to evolve from surface-level empathy to precise interventions targeting deep-level irrational beliefs.

3 Preliminaries

3.1 Ideal Causal Graph Definition

Psychological studies (Johnson-Laird, 1983; Anderson, 2014) indicate that human mental states arise from latent cognitive schemas and their dynamical mechanisms. To render these clients’ pathological mechanisms computationally tractable, we model them as an ideal causal graph:

$$G_{ideal} = (V_{\{A,B,C,D\}}, E), \quad (1)$$

which is a directed acyclic graph that can be constructed given a comprehensive narrative client *Profile* (i.e., when a counselor has sufficient understanding of the client). Each node $v \in V$ represents a psychological element (e.g., negative emotions, cognitive schemas, avoidance behaviors, or triggering events). Following the cognitive hierarchy model in CBT and case conceptualization theory (Beck, 2020; Persons, 2008), the node space

V is structured into four progressive psychological levels: V_A (*surface-level*): observable situations and reactions; V_B (*intermediate-level*): life circumstances and early experiences; V_C (*deep-level*): underlying assumptions and rules; V_D (*core-level*): core beliefs and unmet needs. A directed edge $(v_i, v_j) \in E$ denotes a causal or maintenance relation between nodes (e.g., the deeper root of $v_i \in V_B$ lies in $v_j \in V_C$). A case of G_{ideal} is shown in Appendix C, Fig. 4.

3.2 The G_{ideal} Dataset Construction

Based on the above definition, we construct a dataset $\mathcal{D} = \{(Profile_i, G_{ideal}^i)\}_{i=1}^N$. Specifically, we first sample 2,000 representative cases from PsyQA (Sun et al., 2021) according to a predefined topic distribution (more details in Appendix A). Each sampled case is then reconstructed into an unstructured narrative *Profile* in both Chinese and English. Based on each *Profile*, we design prompts (cf. Appendix L, Fig. 6) to guide DeepSeek-v3.1-250821 in identifying psychological elements, assigning them to different levels, and inferring their causal maintenance relations, thereby constructing the corresponding G_{ideal} . To assess the quality of the automatically constructed data, we randomly inspect 500 samples from \mathcal{D} and find that 90.4% of them satisfy our clinical conceptualization criteria. Finally, we partition \mathcal{D} into three subsets: \mathcal{D}_{cold} (900 samples) for cold-start training, \mathcal{D}_{RL} (1,000 samples) for RL environment simulation, and \mathcal{D}_{test} (100 samples) for evaluation. Since reward computation and final evaluation are highly sensitive to graph quality, all samples in \mathcal{D}_{RL} and \mathcal{D}_{test} are further cleaned and manually verified (cf. Appendix C for details).

4 TRACE

In real-world counseling, a client’s psychological profile is often incomplete initially. Counselors must therefore gradually uncover the underlying mechanisms of distress through multi-turn dialogue, continuously adapting their exploration strategy. This process aligns with the RL paradigm. Accordingly, the goal of this work is to train a model capable of progressively reconstructing a client’s underlying causal graph G that approximates the *ideal causal graph* G_{ideal} and thereby enabling *dynamic case conceptualization*. To this end, we propose a causal-graph-

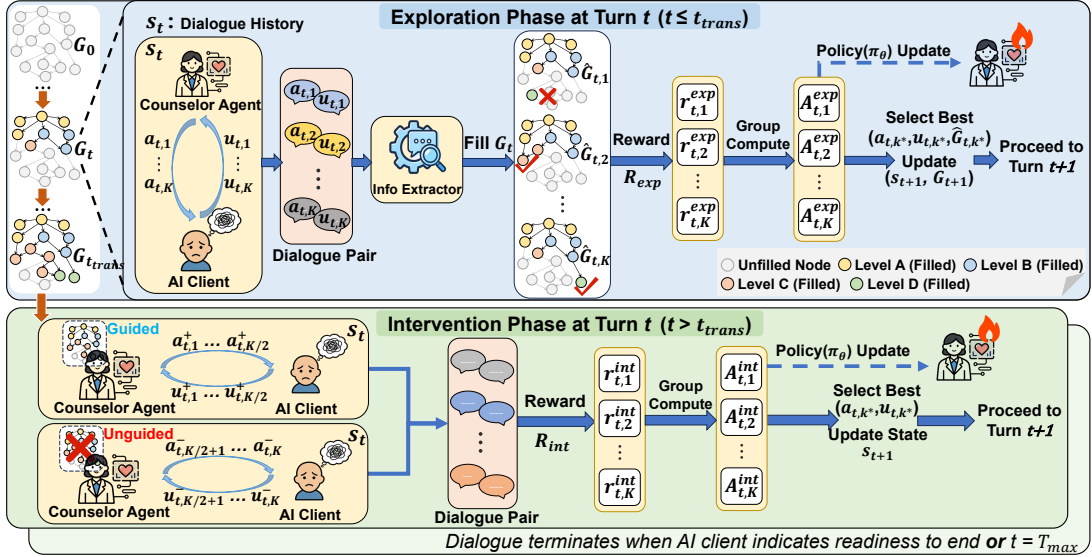


Figure 2: Overview of TRACE: two-phase RL framework with the exploration-phase for reconstructing causal graph and the intervention-phase for restructuring cognition. t_{trans} is the transition turn between two phases.

driven framework TRACE, as shown in Fig. 2. In it, the training model is treated as the *Counselor Agent* (CA), parameterized by a policy π_θ . Notably, since premature interventions based on partial information are often ineffective, professional counselors in real practice strictly follow an “exploration-before-intervention” principle (Miller and Rose, 2009; Fried et al., 2017). Inspired by this, we decompose the whole counseling process into an *exploration phase* and an *intervention phase*, incentivizing the CA to prioritize reconstructing G before conducting cognitive intervention.

4.1 Problem Formulation

Given a sample $(Profile, G_{ideal})$ from \mathcal{D}_{RL} , the counseling process is a multi-turn dialogue between the CA and an AI client, where the client is simulated by a closed-source LLM. Specifically, at turn t , the CA generates a query a_t based on dialogue history s_t , and the AI client returns a response u_t conditioned on $Profile$, forming the updated dialogue history s_{t+1} .

As stated earlier, the *exploration phase* focuses on reconstructing the causal graph, thus, its training objective is

$$\pi_\theta^*, G^* = \arg \max_{\theta, G} \mathbb{E}_{a_t \sim \pi_\theta} [\mathcal{J}(\theta, \mathcal{R}_{exp}(a_t, G_t))], \quad (2)$$

where \mathcal{R}_{exp} is the exploration reward, π_θ is the policy. The CA gradually reconstructs G^* to approximate G_{ideal} . This phase terminates once the phase-transition criteria are met, see Sec.4.3.C, at

which G^* becomes $G_{t_{trans}}$.

As the *intervention phase* targets the effective intervention, thus, its training objective is

$$\pi_\theta^* = \arg \max_{\theta} \mathbb{E}_{a_t \sim \pi_\theta} [\mathcal{J}(\theta, \mathcal{R}_{int}(a_t | G_{t_{trans}}))], \quad (3)$$

where \mathcal{R}_{int} is the intervention reward. The CA leverages the reconstructed $G_{t_{trans}}$ to learn effective cognitive intervention strategies.

4.2 Policy Initialization

To endow the CA with preliminary counseling capabilities, we synthesize a multi-turn dialogue corpus using GPT-5-mini (OpenAI, 2025) based on \mathcal{D}_{cold} . The prompt explicitly constrains the dialogue to follow the hierarchical logic of G_{ideal} (cf. Appendix L, Fig. 7). After filtering out invalid samples that either fail to reach the core-level V_D or contain too few turns (<10), we obtain the cold-start dataset \mathcal{D}_{sft} . We then perform supervised fine-tuning on a base LLM using \mathcal{D}_{sft} , yielding the cold-start CA π_{sft} .

4.3 Training and Rewards

A. Exploration phase: In this phase, the CA learns the “case conceptualization” process by gradually filling the client’s underlying psychological elements into the current graph G_t . To ensure clinical plausibility, the CA follows a surface-to-deep probing path: it must first ground its reasoning in surface-level manifestations (parent nodes) before probing deeper causes (child nodes), prohibiting leapfrog queries that bypass intermediate

nodes along the causal chain in G_t .

Specifically, exploration begins with a current causal graph G_0 that inherits the topology of G_{ideal} but with all nodes unfilled. Please refer to the blue box in Fig. 2. At turn t , we update

$$\begin{aligned} G_{t+1} &= \hat{G}_{t,k^*} = \arg \max_k \mathcal{R}_{exp}(\hat{G}_{t,k}), \\ s_{t+1} &= s_t \oplus (a_{t,k^*}, u_{t,k^*}), \end{aligned} \quad (4)$$

where $\hat{G}_{t,k} = \text{Mext}(G_t, a_{t,k}, u_{t,k})^1$, $\{a_{t,k}\}_{k=1}^K \sim \pi_\theta(\cdot | s_t)$ are K candidate queries from the CA, $u_{t,k}$ is the corresponding client response.

Let $\Delta V_{t,k} = V(\hat{G}_{t,k}) \setminus V(G_t)$ denote the set of nodes newly filled at turn t . A positive exploration reward is granted if all parent nodes of $v \in \Delta V_{t,k}$ have been filled; otherwise, a penalty is applied. The exploration reward \mathcal{R}_{exp} is defined as:

$$\mathcal{R}_{exp}(\hat{G}_{t,k}) = \sum_{v \in \Delta V_{t,k}} \begin{cases} L(v), & Pa(v) \subseteq V(G_t) \\ -1, & Pa(v) \not\subseteq V(G_t) \end{cases}, \quad (5)$$

where $Pa(v)$ is the parent set of v , and $L(v) = 2.0, 1.5, 1.0$ for v in surface (V_A), intermediate (V_B), and deep/core ($V_C \cup V_D$) levels, respectively. The higher reward for V_A and V_B encourages the CA to establish reliable clinical facts first, ensuring that subsequent probing of deeper mechanisms follows the surface-to-deep path, thereby systematically restructuring the pathological causal maintenance schema (cf. Appendix E for more details).

B. Intervention phase: In this phase, the dialogue process is analogous to the exploration phase, except that the CA targets the irrational beliefs in G_{ttrans} for cognitive restructuring. Please refer to the green box in Fig. 2. At turn t , the best dialogue pair is selected by

$$(a_{t,k^*}, u_{t,k^*}) = \arg \max_k \mathcal{R}_{int}(a_{t,k} | G_{ttrans}), \quad (6)$$

where $\{a_{t,k}\}_{k=1}^K \sim \pi_\theta(\cdot | s_t)$ are K candidate queries, $u_{t,k}$ is the corresponding response, and $s_{t+1} = s_t \oplus (a_{t,k^*}, u_{t,k^*})$.

To learn effective intervention strategies, we adopt a hybrid sampling scheme in each turn. (1) *Guided candidates* $\{a_{t,k}^+\}_{k=1}^{K/2} \sim \pi_\theta(\cdot | s_t, G_{ttrans})$, which are generated by the CA instructed with a prompt incorporating all irrational

¹ M_{ext} , Information Extractor, is a fine-tuned lightweight model that can accurately extract psychological elements from the dialogue and fill them into $\hat{G}_{t,k}$, cf. Appendix B for details.

belief nodes in G_{ttrans} as intervention targets; (2) *Unguided candidates* $\{a_{t,k}^-\}_{k=\frac{K}{2}+1}^K \sim \pi_\theta(\cdot | s_t)$, where the prompt contains only s_t without guidance from G_{ttrans} . The *intervention reward* is then defined as:

$$\mathcal{R}_{int}(a_{t,k} | G_{ttrans}) = \begin{cases} 1 & \text{if } a_{t,k} \in \{a_{t,k}^+\} \\ -1 & \text{if } a_{t,k} \in \{a_{t,k}^-\} \end{cases}. \quad (7)$$

It enables the CA to gradually internalize the targeted, effective intervention strategies encoded in the guided candidates. (cf. Appendix F for more details)

C. Phase transition criteria: Since training requires traversing the full dialogue of each counseling, the CA needs to autonomously transition from exploration to intervention. In this work, the transition occurs at $ttrans$ when the current graph G_t satisfies both: (1) The fraction of filled nodes exceeds the ratio $\varphi_{trans} = 80\%$; (2) All irrational beliefs in G_{ideal} have been identified and filled in G_t . This design ensures that the CA intervenes only after the clients causal maintenance logic is adequately understood.

4.4 Policy Optimization

The CA policy π_θ is initialized from the cold-start policy π_{sft} . To improve optimization stability without introducing an additional critic network, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024). At turn t , given the current dialogue history s_t , GRPO samples K candidate actions $\{a_{t,k}\}_{k=1}^K$ from $\pi_{\theta_{old}}(\cdot | s_t)$. The policy is optimized by maximizing:

$$\begin{aligned} \mathcal{J}_{GRPO}(\theta, \mathcal{R}) &= \mathbb{E}_{\{a_{t,k}\}_{k=1}^K \sim \pi_{\theta_{old}}(\cdot | s_t)} \left[\frac{1}{K} \sum_{k=1}^K \right. \\ &\quad \left. \left(\min(\rho_{t,k} A_{t,k}, \text{clip}(\rho_{t,k}, 1 - \epsilon, 1 + \epsilon) A_{t,k}) \right. \right. \\ &\quad \left. \left. - \beta \mathbb{D}_{KL}[\pi_\theta || \pi_{sft}] \right) \right], \end{aligned} \quad (8)$$

where $\rho_{t,k} = \frac{\pi_\theta(a_{t,k} | s_t)}{\pi_{\theta_{old}}(a_{t,k} | s_t)}$, the advantage $A_{t,k} = \frac{r_{t,k} - \bar{r}_t}{\sigma_{r_t} + \epsilon}$. The per-turn reward $r_{t,k}$ is computed by \mathcal{R}_{exp} (Eq. (2)) during the exploration phase and \mathcal{R}_{int} (Eq. (3)) during the intervention phase.

5 Experiments

To comprehensively evaluate TRACE, we design three experiments that assess (1) The capability

of causal graph reconstruction; (2) The effectiveness of cognitive intervention; (3) General dialogue quality.

5.1 Implementation Details

The CA is built upon Qwen2.5-7B-Instruct for Chinese and LLaMA-3.1-8B (Grattafiori et al., 2024) for English, which is to assess the generalization of TRACE across model architectures and linguistic environments. We perform supervised fine-tuning on \mathcal{D}_{sft} using LoRA (Hu et al., 2022) ($r=32$, $\alpha=64$), optimized with AdamW for 3 epochs (learning rate 2×10^{-5} , batch size 4). Then, GRPO is applied for RL with $K=4$ candidate responses per turn, learning rate 3×10^{-5} , and KL coefficient $\beta=0.01$. Rollouts are generated based on vLLM (Kwon et al., 2023) with temperature, top- p , and max_tokens set to 0.7, 0.9, 512, respectively. The AI client performs role-playing based on the narrative profiles in \mathcal{D}_{RL} (training) and \mathcal{D}_{test} (testing). To avoid potential evaluation bias, we employ Doubao-Seed-1.6-250615 (ByteDance Seed Team, 2025) during training and GPT-4o-mini (OpenAI, 2024) during testing². Each dialogue terminates when the AI client indicates readiness to end or when the maximum number of turns $T_{max} = 25$ is reached (cf. Appendix D for more details). All experiments were conducted on 4×RTX 3090 GPUs.

5.2 Baselines and Evaluation Settings

We compare the trained CA with two representative categories of baselines: (1) *General-purpose LLMs (Parameters > 100B)*: Qwen2.5-MAX (Qwen Team, 2024), Gemini-2.5-Pro (Comanici et al., 2025) and GLM-4 (Team GLM, 2024). Although these models are not specifically trained for counseling scenarios, they possess strong capabilities in general language understanding and reasoning. Therefore, we guide them into the “counselor” role using unified prompts. (2) *Mental Health LLMs*: SoulChat 2.0 (Chen et al., 2023), SoulChat-R1 (Chen et al., 2025), and CRISP (Zhou et al., 2025). These SOTA mental health LLMs are specifically designed and trained for counseling tasks. We also evaluate three variants of CA trained by TRACE. w/o SFT denotes

²We conducted human evaluations based on ESConv (Liu et al., 2021). Results indicate that responses from both AI clients were rated as indistinguishable from or more natural than human responses in over 68% of cases (cf. Appendix H for more details).

a CA with RL training but without the cold-start training. w/o RL is a CA with the cold-start training but without RL training. w/o Int-RL is a CA with the cold-start and exploration-phase RL training. Each testing model engages in a dialogue with the AI client, producing a complete dialogue history s_{full} for evaluation. All results are averaged over three independent runs

Notably, during testing, the CA does not explicitly construct the causal graph for intervention. Instead, it relies entirely on the exploration and intervention abilities internalized during training. A qualitative case study can be found in Appendix K.

5.3 Experimental Results

5.3.1 Causal Graph Reconstruction

After each counseling test, the complete dialogue history s_{full} , rather than the narrative *Profile*, is fed into DeepSeek-v3.1-250821 (DeepSeek, 2025) to reconstruct the client’s causal graph \hat{G} (cf. Appendix L Fig. 8 for prompts). By comparing \hat{G} with G_{ideal} , we can assess the model’s ability of exploring the client’s underlying psychological maintenance schema.

To this end, we introduce three metrics: (1) **NCR_{total}** (the total Node Coverage Rate): the proportion of nodes in G_{ideal} that are successfully probed and filled in \hat{G} ; (2) **NCR_{A/B/C/D}**: they are similar to NCR_{total}, but level-wise node coverage rates at level A, B, C and D, respectively; (3) **GED** (Graph Edit Distance): it is a widely used metric for measuring topological similarity between graphs (Sanfeliu and Fu, 2012), which represents the minimum edit cost of transforming \hat{G} into G_{ideal} (cf. Appendix G for details).

Table 1 shows the results. Both the Chinese and English models trained with TRACE outperform other competing models across all metrics. General-purpose LLMs perform reasonably well at surface-level (NCR_A), but exhibit substantially worse performance at deeper levels, indicating that models without domain knowledge tend to remain at superficial empathy. SoulChat 2.0 and SoulChat-R1, which focus on conversational style and coherence, achieve higher scores at surface levels (NCR_A and NCR_B). However, due to the absence of systematic modeling of clients psychological maintenance schema, they also exhibit limited capability at core level (NCR_D). CRISP, by simulating a hierarchical intervention process derived from cognitive therapy, demonstrates certain

Category	Model	NCR _{total} ↑ zh / en	NCR _A ↑ zh / en	NCR _B ↑ zh / en	NCR _C ↑ zh / en	NCR _D ↑ zh / en	GED↓ zh / en	\mathcal{E}_{int} ↑ zh / en
General LLMs	Qwen2.5-MAX	0.42 / 0.39	0.70 / 0.66	0.46 / 0.38	0.33 / 0.26	0.31 / 0.32	35.7 / 34.9	0.82 / 0.81
	Gemini-2.5-Pro	0.50 / 0.54	0.74 / <u>0.80</u>	0.54 / 0.60	0.38 / 0.39	0.36 / <u>0.41</u>	33.2 / 32.9	1.02 / 1.06
	GLM-4	0.40 / 0.41	0.62 / 0.66	0.46 / 0.35	0.30 / 0.29	0.24 / 0.30	33.9 / 35.5	0.84 / 0.90
MH LLMs	SoulChat 2.0	0.47 / 0.45	0.75 / 0.68	0.57 / 0.49	0.26 / 0.31	0.21 / 0.19	32.7 / 33.3	0.89 / 0.83
	SoulChat-R1	0.52 / 0.54	0.75 / 0.72	<u>0.64</u> / 0.61	0.45 / 0.40	0.32 / 0.27	30.5 / 31.0	1.05 / 1.01
	CRISP	<u>0.57</u> / <u>0.59</u>	<u>0.82</u> / 0.72	<u>0.62</u> / <u>0.64</u>	<u>0.52</u> / <u>0.48</u>	<u>0.39</u> / 0.31	<u>29.8</u> / <u>30.4</u>	<u>1.12</u> / <u>1.10</u>
Ours (Qwen / LLaMA)	TRACE	0.69 / 0.65	0.89 / 0.85	0.78 / 0.73	0.62 / 0.57	0.56 / 0.51	26.5 / 27.2	1.37 / 1.31
	w/o SFT	0.60 / 0.53	0.79 / 0.81	0.62 / 0.65	0.43 / 0.45	0.47 / 0.41	28.0 / 30.1	1.10 / 1.03
	w/o Int-RL	0.66 / 0.64	0.86 / 0.83	0.72 / 0.70	0.57 / 0.55	0.52 / 0.49	27.1 / 28.3	0.80 / 0.78
	w/o RL	0.51 / 0.48	0.70 / 0.65	0.51 / 0.55	0.41 / 0.38	0.23 / 0.21	31.5 / 32.4	0.88 / 0.85

Table 1: Evaluation results of causal graph reconstruction and intervention effectiveness in both Chinese and English (zh/en). ↑: higher is better; ↓: lower is better. The best results are in **bold** and the second-best are underlined. TRACE variants, w/o SFT: the model without cold-start training; w/o RL: the model without RL training; w/o Int-RL: the model without intervention-phase RL only.

exploration abilities and thus outperforms the former two models. Nevertheless, CRISP is trained on static dialogue corpora, making it difficult to adapt its exploration strategy in dynamic dialogue or explicitly model the complete causal chain from symptoms to core beliefs. In contrast, TRACE mimics “case conceptualization” to enable more comprehensive exploration in scenarios where information is gradually disclosed.

Among TRACE variants, w/o Int-RL and w/o SFT outperforms w/o RL, particularly at deeper levels, suggesting that RL effectively drives models beyond merely learning conversation mappings toward actively probing more underlying information. Compared to w/o SFT, TRACE and w/o Int-RL further improves across all metrics, indicating that supervised fine-tuning during cold start provides a valuable initial policy for subsequent RL. TRACE outperforms w/o Int-RL, indicating that interventionphase training further enhances exploration capabilities. Moreover, the Chinese and English models exhibit highly consistent performance trends, underscoring the strong cross-architecture and cross-lingual generalization of TRACE.

5.3.2 Intervention Effectiveness

To assess whether competing models can alleviate or restructure AI client’s irrational beliefs after multi-turn dialogue, we design an evaluation protocol based on a *Case Specific Irrational Belief Questionnaire*. For each test case ($Profile, G_{ideal}$) in \mathcal{D}_{test} , we prompt DeepSeek-v3.1-250821 to generate a case-specific questionnaire Q according to the Profile. We then use DeepSeek-R1-0528, a strong reasoning model, as

the evaluator M_{eval} . The evaluator role-plays as the client and rates its agreement with each questionnaire item on a 1-5 Likert scale (higher scores indicate stronger agreement with irrational beliefs). This assessment is conducted under two conditions: pre-counseling (given only $Profile$) and post-counseling (given $Profile$ and s_{full}). The intervention effect \mathcal{E}_{int} is defined as:

$$\mathcal{E}_{int} = M_{eval}(Q|o) - M_{eval}(Q|o, s_{full}), \quad (9)$$

where o is $Profile$ and $M_{eval}(Q|\cdot)$ denotes the average score on given conditions (*cf.* Appendix I). Notably, the evaluator M_{eval} differs from the LLM used as the AI client during testing, which mitigates self-preference bias arising from a model scoring its own dialogues.

The results are listed in the rightmost column of Table 1. The overall trend aligns with the results of causal graph reconstruction. As expected, general-purpose LLMs exhibit weaker intervention effects because of their focus on superficial empathy while lacking an understanding of deeper cognitive structures. CRISP, benefiting from its hierarchical cognitive therapy design, achieves better intervention effects than the SoulChat series but underperforms TRACE. We attribute this to its reliance on predefined therapeutic workflows, which prevents it from dynamically exploring a complete causal chain.

Among TRACE variants, the considerable performance gap between w/o SFT and w/o Int-RL confirms that RL, especially in the intervention phase, indeed improves intervention effectiveness. Moreover, the inferior performance of w/o Int-RL compared to TRACE indicates that the RL in the exploration and intervention phases is insepara-

Category	Model	Sens \uparrow zh / en	Spec \uparrow zh / en	Supp \uparrow zh / en	Help \uparrow zh / en	Trst \uparrow zh / en	Overall \uparrow zh / en
General LLMs	Qwen2.5-MAX	4.38 / 4.11	4.10 / 4.21	3.71 / 3.65	3.75 / 3.83	3.92 / 3.99	3.91 / 4.07
	Gemini-2.5-Pro	<u>4.54 / 4.42</u>	4.50 / 4.35	4.26 / 4.48	4.19 / 4.36	4.34 / 4.33	<u>4.44 / 4.31</u>
	GLM-4	4.29 / 3.88	4.05 / 4.26	3.81 / 3.72	3.92 / 3.91	4.19 / 4.12	4.08 / 4.03
MH LLMs	SoulChat 2.0	4.31 / 4.13	4.10 / 3.93	4.62 / 4.49	3.89 / 4.01	4.59 / 4.52	4.02 / 4.18
	SoulChat-R1	4.23 / 4.40	4.63 / 4.49	4.39 / 4.15	4.21 / 4.01	4.27 / 4.13	4.35 / 4.29
	CRISP	4.32 / 3.79	3.98 / 4.18	3.74 / 3.68	<u>4.22</u> / 4.02	4.17 / 4.11	4.01 / 3.99
Ours (Qwen / LLaMA)	TRACE	4.59 / 4.51	<u>4.52 / 4.37</u>	4.40 / 4.39	4.50 / 4.55	4.42 / 4.38	4.49 / 4.45
	w/o SFT	4.19 / 4.02	4.01 / 4.07	4.15 / 4.11	4.23 / 4.10	4.15 / 4.11	4.01 / 4.09
	w/o Int-RL	4.01 / 4.07	3.87 / 3.78	3.96 / 4.01	3.82 / 3.85	3.81 / 3.74	3.86 / 3.79
	w/o RL	4.21 / 4.10	4.08 / 4.11	4.20 / 4.13	4.21 / 4.15	4.10 / 3.91	4.22 / 4.06

Table 2: Results of general dialogue quality evaluation. Notation follows Table 1.

ble, validating the efficacy of the “explore-before-intervene” strategy in our framework.

5.3.3 General Dialogue Quality

Following prior study CRISP (Zhou et al., 2025), we also evaluate general dialog quality by using DeepSeek-R1-0528 to score the dialogue history s_{full} on a 1-5 Likert scale across six dimensions: *Sensibleness*, *Specificity*, *Supportiveness*, *Helpfulness*, *Trustworthiness*, and *Overall Quality* (cf. Appendix J for detailed scoring criteria).

The results are reported in Table 2. Compared with other models, TRACE performs particularly well in *Sensibleness*, *Helpfulness* and *Overall Quality*, which we attribute to its causal-graph-driven exploration that ensures logical consistency and intervention efficacy. SoulChat 2.0 excels in *Supportiveness* and *Trustworthiness*, benefiting from its use of diverse counselor personas to foster a personalized therapeutic alliance. SoulChat-R1 enhances response depth and coherence via chain-of-thought reasoning, thus, performs best in *Specificity*. Although TRACE does not rank first on these three dimensions, its gaps to the best models are marginal. w/o Int-RL shows a notable drop in quality, suggesting that intervention-phase RL is critical not only for intervention efficacy but also for overall dialogue quality.

5.3.4 Reliability of LLM Evaluation

To assess the reliability of the LLM-based evaluations for “Intervention Effectiveness” and “General Dialogue Quality”, we randomly sampled 150 complete dialogues generated by all competing models. We then recruited five psychology graduate students to independently score these dialogues following the same criteria used for DeepSeek-R1-0528. The inter-rater reliability among the human experts was substantial, achiev-

Evaluation Dimension	Correlation ρ	p -value
Intervention Effectiveness (\mathcal{E}_{int})	0.76	< 0.01
Sensibleness	0.72	< 0.01
Specificity	0.63	< 0.05
Supportiveness	0.71	< 0.01
Helpfulness	0.65	< 0.05
Trustworthiness	0.69	< 0.05
Overall Quality	0.68	< 0.05

Table 3: Spearman correlation between human expert and LLM evaluations and p -value.

φ_{trans}	0.6	0.7	0.8	0.9
\mathcal{E}_{int}	1.05	1.19	1.37	1.14

Table 4: Intervention efficacy under varied φ_{trans} .

ing an average Cohen’s κ of 0.73. We employ Spearman correlation and p -value to measure the differences between the human experts’ and LLM scores. As shown in Table 3, the results exhibit strong Spearman correlations across all dimensions, validating the reliability of our LLM evaluation paradigm.

5.4 Ablation Study

TRACE features a key hyperparameter: *Causal Graph Filling Ratio* φ_{trans} , which is discussed in Sec. 4.3.C. We test four values {0.6, 0.7, 0.8, 0.9}. The results in Table 4 show that $\varphi_{trans} = 0.8$ yields the best performance. A lower value leads the model to prematurely enter the intervention phase before acquiring sufficient client information, resulting in unfounded speculations that reduce intervention efficacy. Conversely, an overly high value forces excessive dialogue turns for exploration, reducing the room for intervention and thereby hindering deeper cognitive restructuring.

6 Conclusion

Inspired by case conceptualization, we propose **TRACE**, a two-phase RL framework that formulates psychological counseling as an online causal graph reconstruction problem in multi-turn dialogue. Following the principle of “explore-before-intervene”, TRACE reconstructs a client’s cognitive causal chains during the exploration phase and performs targeted cognitive restructuring in the intervention phase. Extensive experiments show that TRACE significantly outperforms existing MH-LLMs, while exhibiting strong generalization. We believe this work offers a viable pathway for advancing LLMs from surface-level empathy toward mechanism-aware psychological intervention.

Limitations

Although experimental results demonstrate the effectiveness of TRACE, several limitations should be considered.

First, given the potential uncontrollability of reinforcement learning (RL) during policy exploration, direct training in real-world psychological counseling scenarios poses significant ethical risks. Therefore, this study employs AI-simulated clients for training and testing. Although the AI client can generate relatively natural responses, such naturalness is not equivalent to clinical authenticity. Gaps may still exist between LLMs and human beings in terms of emotional complexity (e.g., ego-defense mechanisms, ambivalence, and limited self-disclosure) and resistance behaviors; for instance, in actual counseling sessions, clients may resist probing questions and present multiple issues simultaneously, or introduce dynamic topic shifts that challenge conversational coherence. To bridge this gap, we introduced experts with clinical backgrounds to conduct manual evaluations. Although experts have validated the effectiveness of TRACE in logical reconstruction, further validation in controlled clinical settings remains necessary. Therefore, developing more realistic AI clients and diverse scenario representations remains an important direction for future research.

Second, the TRACE framework in this study originates from the reconstruction of causal structure in case conceptualization. In reality, psychological dynamics may involve longer, even life-long causal chains (e.g., the complex impact of early childhood trauma on adult social patterns).

However, due to the current risk of “hallucination” in LLMs when maintaining ultra-long logical consistency, this work restricts its scope to verifiable, explicit cognitive causal reconstruction. Future work will extend TRACE toward modeling cumulative causal effects across dialogues and the evolution of causal mechanisms over longer developmental timescales.

Ethics Statement

Data Privacy. The data used in this work originates from PsyQA (Sun et al., 2021). We implemented rigorous data cleaning protocols: all personally identifiable information has been anonymized; sensitive information and unethical content have been thoroughly removed; all content potentially harmful to clients or others has been excluded from the dataset. The profiles used in experiments were synthetically reconstructed to preserve clinically relevant psychological patterns while protecting the privacy of original help-seekers.

Potential Risks. Mental health AI systems carry inherent risks, including inappropriate advice, misjudgment of psychological states, or exacerbation of psychological distress. This system is designed to assist rather than replace professional mental health services, and its outputs do not constitute clinical diagnoses or treatment plans. If deployed in real-world scenarios, deployers should ensure users are aware they are interacting with an AI.

Human Evaluation. We recruited five psychology graduate students. All participants provided informed consent and were granted the right to withdraw from the evaluation at any time without penalty. Participants were compensated at \$10/hour, which is above the local minimum wage and consistent with standard rates for expert annotation.

References

- John R. Anderson. 2014. *Rules of the Mind*. Psychology Press.
- Judith S. Beck. 2020. *Cognitive Behavior Therapy: Basics and Beyond*, 3rd edition. Guilford Publications.
- Peter J. Bieling and Willem Kuyken. 2003. Is cognitive case formulation science or science fiction? *Clinical Psychology: Science and Practice*, 10(1):52.

- Denny Borsboom and Angélique O. J. Cramer. 2013. [Network analysis: An integrative approach to the structure of psychopathology](#). *Annual Review of Clinical Psychology*, 9(1):91–121.
- ByteDance Seed Team. 2025. [Introduction to techniques used in Seed1.6](#). ByteDance Seed Blog.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of ACL*, pages 2318–2335.
- Mingyu Chen, Jingkai Lin, Zhaojie Chu, Xiaofen Xing, Yirong Chen, and Xiangmin Xu. 2025. [CATCH: A novel data synthesis framework for high therapy fidelity and memory-driven planning chain of thought in AI counseling](#). In *Findings of EMNLP*, pages 10254–10286.
- 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 EMNLP*, pages 1170–1183.
- Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. [SFT memorizes, RL generalizes: A comparative study of foundation model post-training](#). In *Proceedings of ICML*, volume 267, pages 10818–10838.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *arXiv preprint arXiv:2507.06261*.
- DeepSeek. 2025. [DeepSeek-V3.1 release](#). DeepSeek API Docs.
- Eiko I. Fried, Claudia D. van Borkulo, Angélique O. J. Cramer, Lynn Boschloo, Robert A. Schoevers, and Denny Borsboom. 2017. Mental disorders as networks of problems: A review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*, 52(1):1–10.
- Xing Fu, Haozhen Li, Bichen Wang, Hao Yang, Yanyan Zhao, and Bing Qin. 2025. [Look beyond feeling: Unveiling latent needs from implicit expressions for proactive emotional support](#). In *Proceedings of EMNLP*, pages 21582–21609.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 543 others. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L Chaunzwa, Idalid Franco, Benjamin H Kann, Shalini Moningi, Jack M Qian, Madeleine Goldstein, Susan Harper, and Hugo J W L Aerts. 2024. Large language models to identify social determinants of health in electronic health records. *NPJ Digital Medicine*, 7(1):6.
- Stephen N. Haynes, William H. O’Brien, and Antonio Godoy. 2020. A proposed model for the psychometric evaluation of clinical case formulations with quantified causal diagrams. *Psychological Assessment*, 32(6):541.
- Stephen N. Haynes, William H. O’Brien, Joseph Keawe’aimoku, Cilia Witteman, and Ph D Kaholokula. 2012. Concepts of causality in psychopathology: Applications in clinical assessment, clinical case formulation and functional analysis. *Journal of Unified Psychotherapy and Clinical Science*, 1(1):87–103.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *Proceedings of ICLR*.
- He Hu, Yucheng Zhou, Juzheng Si, Qianning Wang, Hengheng Zhang, Fuji Ren, Fei Ma, Laizhong Cui, and Qi Tian. 2025. [Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling](#). *arXiv preprint arXiv:2505.15715*.
- Philip N. Johnson-Laird. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press.
- Juhée Kim, Chunghu Mok, Jisun Lee, Hyang Sook Kim, and Yohan Jo. 2025. [Dialogue systems for emotional support via value reinforcement](#). In *Proceedings of ACL (Volume 1: Long Papers)*, pages 28733–28766.
- Willem Kuyken, Christine A. Padesky, and Robert Dudley. 2011. *Collaborative Case Conceptualization: Working Effectively with Clients in Cognitive-Behavioral Therapy*. Guilford Press.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of SOSP*, pages 611–626.
- Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. [Cactus: Towards psychological counseling conversations using cognitive behavioral theory](#). In *Findings of EMNLP*, pages 14245–14274.

- 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 ACL (Volume 1: Long Papers)*, pages 3469–3483.
- William R. Miller and Gary S. Rose. 2009. Toward a theory of motivational interviewing. *American Psychologist*, 64(6):527.
- Clarissa W. Ong, Hiba Arnaout, Kate Sheehan, Estella Fox, Eugen Owtscharow, and Iryna Gurevych. 2025. [Using large language models to create personalized networks from therapy sessions](#). *arXiv preprint arXiv:2512.05836*.
- OpenAI. 2024. [GPT-4o mini: Advancing cost-efficient intelligence](#). OpenAI Blog.
- OpenAI. 2025. [GPT-5](#). OpenAI Product Page.
- Jacqueline B. Persons. 2008. *The Case Formulation Approach to Cognitive-Behavior Therapy*. Guilford Press.
- 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 EMNLP*, pages 615–636.
- Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Alberto Sanfeliu and King-Sun Fu. 2012. A distance measure between attributed relational graphs for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, (3):353–362.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y K Li, Yang Wu, and Daya Guo. 2024. [DeepSeekMath: Pushing the limits of mathematical reasoning in open language models](#). *arXiv preprint arXiv:2402.03300*.
- 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 ACL-IJCNLP*, pages 1489–1503.
- Takehiro Takayanagi, Masahiro Suzuki, Ryotaro Kobayashi, Hiroki Sakaji, and Kiyoshi Izumi. 2024. Is ChatGPT the future of causal text mining? A comprehensive evaluation and analysis. In *Proceedings of IEEE BigData*, pages 6651–6660.
- Team GLM. 2024. [ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools](#). *arXiv preprint arXiv:2406.12793*.
- Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. 2024. Automating psychological hypothesis generation with AI: When large language models meet causal graph. *Humanities and Social Sciences Communications*, 11(1):1–14.
- 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 ACL (Volume 1: Long Papers)*, pages 1081–1115.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Y. Yang, P. Achananuparp, H.Y. Huang, J. Jiang, P.L. Kit, N.G. Lim, C.T.S. Ern, and E.P. Lim. 2025b. CAMI: A counselor agent supporting motivational interviewing through state inference and topic exploration. In *Proceedings of ACL (Volume 1: Long Papers)*, pages 21037–21081.
- Xianrong Yao, Dong She, Chenxu Zhang, Yimeng Zhang, Yueru Sun, Noman Ahmed, Yang Gao, and Zhanpeng Jin. 2025. [Empathy-r1: A chain-of-empathy and reinforcement learning framework for long-form mental health support](#). *arXiv preprint arXiv:2509.14851*.
- Jiahao Yuan, Zhiqing Cui, Hanqing Wang, Yuansheng Gao, Yucheng Zhou, and Usman Naseem. 2025. [Kardia-r1: Unleashing LLMs to reason toward understanding and empathy for emotional support via rubric-as-judge reinforcement learning](#). *arXiv preprint arXiv:2512.01282*.
- Jinfeng Zhou, Yuxuan Chen, Jianing Yin, Yongkang Huang, Yihan Shi, Xikun Zhang, Libiao Peng, Rongsheng Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2025. [Crisp: Cognitive restructuring of negative thoughts through multi-turn supportive dialogues](#). In *Proceedings of EMNLP*, pages 32462–32491.

A Dataset Topic Distribution

Fig 3 illustrates the topic distribution of our constructed dataset. We sample 2000 cases from PsyQA, yielding a total of 2000 samples spanning 102 fine-grained psychological counseling topics.

Regarding the sampling strategy, we employ stratified sampling for PsyQA to ensure adequate representation of high-frequency topics (e.g., behavioral issues, interpersonal relationships, family concerns) while also incorporating low-frequency yet clinically significant topics (e.g., psychological crisis, personality disorders). The distribution across major topics is relatively balanced: the top five categories (Others, Behavior, Relationships, Family, and Personal Growth) each account for approximately 10%–14%, with no single topic dominating the dataset.

Notably, the “Others” category (14.2%) aggregates 90 low-frequency topics (each <2%). Although these topics appear less frequently in the dataset, they encompass clinically important issue types that cannot be overlooked in practice, including post-traumatic stress, personality disorders, and crisis intervention. This design ensures that the topic coverage aligns with the demand spectrum of real-world psychological counseling, providing sufficient training data for common issues while maintaining generalization capability for low-frequency scenarios.

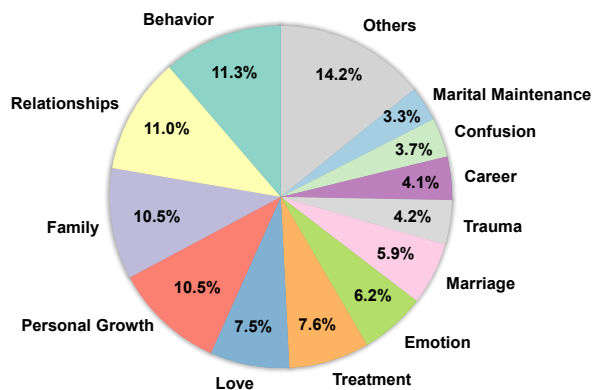


Figure 3: Topic distribution of the dataset.

B Information Extractor Construction

During the exploration phase of reinforcement learning, we need to extract newly disclosed psychological elements from each dialogue turn to update the current causal graph. To this end, we construct a lightweight information extractor M_{ext} . Specifically, given the causal graph state G_t at

turn t , the counselor’s query a_t , and the client’s response u_t , the information extractor outputs the updated causal graph:

$$G_{t+1} = M_{ext}(G_t, a_t, u_t), \quad (10)$$

where M_{ext} identifies newly disclosed psychological element nodes ΔV_t from the current dialogue turn and fills them into the corresponding positions in G_t , yielding G_{t+1} . If no new information is disclosed in the current turn, then $\Delta V_t = \emptyset$ and $G_{t+1} = G_t$.

Training Data Construction. We select complete multi-turn dialogue records from 200 clients in the cold-start dataset \mathcal{D}_{sft} as the annotation source. We employ GPT-4o to perform turn-by-turn annotation across all turns in these 200 dialogues: for each dialogue turn (a_t, u_t) , the annotator identifies newly disclosed psychological elements and their corresponding hierarchical levels based on the current causal graph state G_t (prompt template shown in Fig 9).

After completing the exhaustive annotation, we curate training samples from the annotation results according to the following principles: (1) *Client diversity*: maximize coverage across different clients to prevent the model from overfitting to specific individual expression patterns; (2) *Level balance*: uniformly sample instances covering surface-level (V_A), intermediate-level (V_B), deep-level (V_C), and core-level (V_D) nodes; (3) *Empty extraction coverage*: include an adequate proportion of dialogue turns where no new information is disclosed (i.e., $\Delta V_t = \emptyset$) to enhance the model’s ability to recognize uninformative interactions. Ultimately, we curate 3,500 high-quality samples that satisfy the above criteria.

Quality Verification. To ensure annotation quality, we randomly select 500 samples from the 3,500 curated instances for independent review by two annotators with psychology backgrounds. The review criteria include: (1) whether the extracted nodes semantically match the dialogue content; (2) whether the hierarchical level assignment is correct; (3) whether there are missing or redundant extractions. The verification results show that GPT-4o achieves an annotation accuracy of 94%. The annotators demonstrated substantial agreement, achieving a Cohen’s κ of 0.78. These 500 verified samples serve as the test set, while the remaining 3,000 samples constitute the training set.

Model Training and Validation. We fine-tune Qwen3-4B (Yang et al., 2025a) as the base model on the training set via supervised learning. The resulting M_{ext} achieves 91% accuracy on the test set, satisfying the requirement for high-precision state updates during reinforcement learning while avoiding dependence on expensive closed-source models.

C Causal Graph Validation Rules

To ensure the quality and clinical validity of the ideal causal graphs G_{ideal} in \mathcal{D}_{RL} and \mathcal{D}_{test} , we establish a systematic four-dimensional validation protocol. All samples are independently reviewed by two annotators with psychology backgrounds; disagreements are adjudicated by a third annotator.

Node Validation. (1) *Level Accuracy*: Nodes are correctly assigned to their corresponding psychological levels V_A – V_D . (2) *Atomicity*: Each node represents a single, actionable psychological construct rather than a composite concept. (3) *Traceability*: All nodes are traceable to explicit textual evidence in the narrative *Profile*. (4) *Non-redundancy*: No semantic overlap exists among nodes.

Edge Validation. (1) *Causal Directionality*: Edge directions reflect psychologically plausible causal maintenance relationships, rather than mere statistical correlations. (2) *Clinical Interpretability*: Each causal link is grounded in established psychological theories and case conceptualization frameworks.

Structure Validation. (1) *DAG Constraint*: The graph satisfies directed acyclic graph constraints, ensuring logical consistency in causal reasoning. (2) *Connectivity*: The graph is weakly connected with no isolated nodes. (3) *Scale Constraint*: The graph contains 15–25 nodes, balancing expressive completeness with annotation feasibility.

Consistency Validation. (1) *Coverage*: The causal graph captures the primary psychological concerns described in the *Profile*. (2) *Boundness*: No speculative elements beyond the *Profile* content are included. (3) *Path Completeness*: At least one complete causal path exists from the surface level V_A to the core level V_D .

Validation Results and Data Processing. We randomly sampled 500 instances from \mathcal{D} for manual validation, achieving an initial pass rate of 90.4% (with a Cohen’s κ value of 0.81 for inter-

rater reliability), which demonstrates the high reliability of the automatic construction pipeline based on DeepSeek-v3.1-250821. Based on these results, we adopt differentiated processing strategies for different data subsets: For \mathcal{D}_{cold} , given that it is used solely for cold-start training and has achieved acceptable quality, we retain the original data without additional cleaning. For \mathcal{D}_{RL} and \mathcal{D}_{test} , which involve precise reward computation and reliable performance evaluation respectively, non-compliant samples undergo up to three automatic regeneration attempts; remaining failures are manually corrected to ensure full compliance with the above validation criteria.

An example of the causal graph is shown in Fig 4.

D AI Client simulation

Role-Playing Setup. The AI client is instantiated by prompting an LLM with a narrative *Profile* containing the client’s symptoms, background, and psychological maintenance logic. To prevent the AI client from being overly cooperative and proactively disclosing excessive information, we require it to follow the principle of minimal effort. The complete prompt template is shown in Fig 11.

Termination Conditions. When the AI client perceives that sufficient emotional support has been received and its mood has improved, it proactively concludes the session by expressing gratitude and appending the “[END]” token. If the AI client does not initiate termination, the dialogue automatically ends when the maximum turn limit T_{max} is reached.

E Exploration Phase Details

Guided Resampling Mechanism. During the exploration phase, when all K candidate actions in a given turn receive negative rewards (i.e., all candidates violate the parent-node constraint $Pa(v) \subseteq V(G_t)$ in Eq. (5)), TRACE triggers a guided resampling mechanism. Specifically, the system selects a node v^* from G_{ideal} that satisfies the following conditions: (1) v^* has not yet been filled in G_t , i.e., $v^* \in V(G_{ideal}) \setminus V(G_t)$; and (2) all parent nodes of v^* have already been filled, i.e., $Pa(v^*) \subseteq V(G_t)$. The system then generates a guiding prompt to instruct the Counselor Agent (CA) to resample (e.g., “Please ask the client about information related to v^* ”; template

shown in Fig 10). The resampled candidates replace the invalid ones, ensuring that at least one positive reward is granted per turn.

Trigger Count Analysis. Fig 5 illustrates the trigger count of the guided resampling mechanism across training steps. During early training, the policy has not yet learned to probe along the causal graph in a structured manner, resulting in a high trigger count. As training progresses, the trigger count gradually decreases and stabilizes, indicating that the model progressively internalizes causal-structure-based exploration strategies and reduces reliance on external guidance.

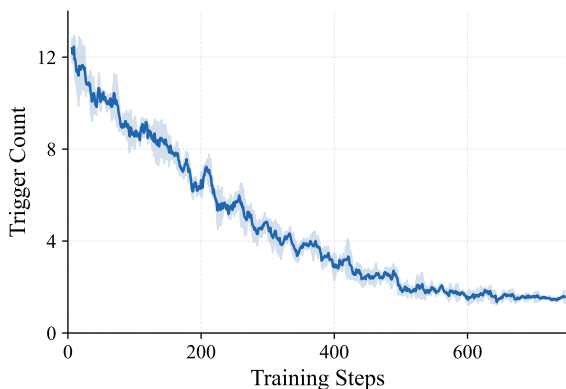


Figure 5: Trigger count of the guided resampling mechanism during training. The x-axis denotes training steps; the y-axis denotes the number of times guided resampling is triggered. The shaded area represents the standard deviation across three independent runs.

F Intervention Phase Details

Definition of Irrational Beliefs. Following the cognitive hierarchy defined in Section 3.1, V_C (underlying assumptions and rules) and V_D (core beliefs and schemas) represent the deep cognitive structures that maintain psychological distress. During the intervention phase, we need to identify irrational belief nodes from the causal graph $G_{t_{trans}}$ obtained at the phase transition as targets for cognitive restructuring.

Specifically, we operationalize irrational beliefs as nodes in $V_C \cup V_D$ whose semantic labels contain cognition-related keywords, including: [thinking pattern], [belief], [conception], [assumption], [rule], etc. This keyword-based extraction strategy effectively distinguishes cognitive nodes from emotional or behavioral nodes, ensuring that the intervention phase focuses on the pathological cognitive structures that genuinely require restructuring.

Guided Prompt and Unguided Prompt. As described in Section 4.3, the intervention phase employs a hybrid prompting strategy to learn effective intervention skills. The two types of prompts are designed as follows:

- **Guided prompt:** Contains the dialogue history s_t along with all irrational belief nodes extracted from $G_{t_{trans}}$, explicitly specifying intervention targets to guide the Counselor Agent (CA) in performing targeted cognitive restructuring. The complete template is shown in Fig 17.
- **Unguided prompt:** Contains only the dialogue history s_t and the basic counselor role definition, without any intervention guidance. The complete template is shown in Fig 12.

By contrasting the quality differences between responses generated from the two prompt types, the intervention reward \mathcal{R}_{int} (Eq (7)) guides the CA to progressively internalize causal-graph-based cognitive intervention strategies.

G Causal Graph Reconstruction Metrics

This section defines the metrics for evaluating the similarity between the reconstructed causal graph $\hat{G} = (\hat{V}, \hat{E})$ and the ideal causal graph $G_{ideal} = (V, E)$. All metrics are based on semantic node matching: node embeddings are computed using BGE-M3 (Chen et al., 2024), and cosine similarity $\text{sim}(\cdot, \cdot)$ is calculated for matching determination. A node $v \in V$ from G_{ideal} is considered successfully matched with a node $\hat{v} \in \hat{V}$ from \hat{G} when $\text{sim}(v, \hat{v}) \geq 0.6$. The matching process follows a one-to-one maximum matching strategy, ensuring each node is matched at most once.

Overall Node Coverage Rate (NCR_{total}). This metric measures the proportion of nodes in G_{ideal} that are successfully reconstructed, reflecting overall exploration completeness:

$$\text{NCR}_{\text{total}} = \frac{|V_{\text{match}}|}{|V|}, \quad (11)$$

where $|V|$ denotes the total number of nodes in G_{ideal} , and $|V_{\text{match}}|$ denotes the number of successfully matched nodes.

Level-wise Node Coverage Rate (NCR_{A/B/C/D}). This metric measures node coverage at each psychological level, assessing whether the model can

explore from surface symptoms (V_A) to core beliefs (V_D). Taking V_A as an example:

$$\text{NCR}_A = \frac{|V_{A,\text{match}}|}{|V_A|}, \quad (12)$$

where $|V_A|$ is the number of V_A -level nodes in G_{ideal} . NCR_B , NCR_C , and NCR_D are defined analogously.

Graph Edit Distance (GED). This metric quantifies the structural dissimilarity between \hat{G} and G_{ideal} by computing the minimum edit cost required to transform one graph into the other:

$$\text{GED}(\hat{G}, G_{ideal}) = \min_P \sum_{op \in P} \text{Cost}(op), \quad (13)$$

where P denotes an edit path, i.e., a sequence of operations that transforms \hat{G} into G_{ideal} . The elementary edit operations and their costs are defined as follows:

- *Node insertion:* Adding a node from V that is missing in \hat{V} . Cost: 1.
- *Node deletion:* Removing a node from \hat{V} that cannot be matched. Cost: 1.
- *Node substitution (matched):* Replacing \hat{v} with semantically matched v ($\text{sim} \geq 0.6$). Cost: 0.
- *Node substitution (unmatched):* Replacing \hat{v} with semantically unmatched v ($\text{sim} < 0.6$). Cost: 1.
- *Edge insertion/deletion:* Adding or removing edges to align with edge set E . Cost: 1.

In practice, we employ a two-stage graph matching algorithm: first establishing optimal node correspondences based on semantic similarity, then computing the minimum edge edits required to align the edge structures given the node matching.

H Human Evaluation of AI Client Realism

To verify the human-likeness of responses generated by the AI client, we conducted a human evaluation study. Since Doubao-Seed-1.6-250615 is used during training while GPT-4o-mini is used during testing, we evaluate both models separately to ensure the completeness of our conclusions.

Evaluation Protocol. We adopt a pairwise preference evaluation paradigm, using authentic human responses from the ESConv dataset (Liu et al.,

2021) as the reference baseline. Specifically, we randomly sampled 150 dialogue contexts from the ESConv test set. For each context, evaluators were simultaneously presented with two candidate responses: one from authentic human dialogue and one generated by the AI client. To eliminate potential biases, all responses were anonymized and their presentation order was randomized.

We recruited five psychology graduate students as evaluators. Evaluators were required to select from three options: (1) Response A is more human-like; (2) Response B is more human-like; (3) The two are indistinguishable (Tie).

Evaluation Results. Table 5 presents the preference distribution results for both AI clients. The ‘‘Human’’ column indicates the proportion where evaluators judged human responses as more natural, the ‘‘AI’’ column indicates the proportion where AI responses were judged as more natural, and the ‘‘Tie’’ column indicates the proportion where the two were indistinguishable.

AI Client	Human↓	Tie	AI↑
Doubao-Seed-1.6-250615	25.3%	48.0%	26.7%
GPT-4o-mini	31.3%	46.7%	22.0%

Table 5: Preference distribution (%) comparing AI client responses with human responses. ‘‘Human’’↓: lower values indicate better AI realism; ‘‘AI’’↑: higher values indicate AI responses are preferred over human responses.

Both AI clients demonstrate high levels of human-likeness: responses from Doubao-Seed-1.6-250615 were judged as indistinguishable from or superior to human responses in 74.7% of evaluations (Tie + AI), while GPT-4o-mini achieved 68.7%. These results indicate that both AI clients can generate sufficiently realistic dialogue responses, meeting the requirements for policy learning and reliable evaluation.

I Intervention Effectiveness Evaluation Details

Questionnaire Generation. For each test sample ($Profile, G_{ideal}$) in \mathcal{D}_{test} , we employ DeepSeek-v3.1-250821 to dynamically generate a *case-specific* 6-item irrational belief questionnaire Q based on the narrative $Profile$ (prompt template shown in Fig 14). Unlike fixed templates, each item is tailored to the specific psychological concerns described in the client’s profile, en-

sure that the evaluation captures intervention effects on the client’s actual irrational beliefs rather than generic psychological constructs.

Questionnaire Structure. Each generated questionnaire follows a standardized structure covering four clinically relevant dimensions:

- **Core negative emotion** (1 item): Captures the primary emotional distress, such as anxiety, depression, or anger.
- **Maladaptive coping behavior** (1 item): Identifies dysfunctional behavioral patterns, such as avoidance or rumination.
- **Dysfunctional cognition** (2 items): Targets automatic thoughts and intermediate beliefs that maintain psychological distress.
- **Barriers to change** (2 items): Assesses factors hindering change, such as hopelessness or distrust in the counseling process.

All items are rated on a 5-point Likert scale (1 = does not describe me at all, 5 = describes me completely), where higher scores indicate stronger endorsement of irrational beliefs.

Questionnaire Quality Validation. To ensure the quality of dynamically generated questionnaires, we randomly sampled 100 cases from \mathcal{D}_{test} for manual validation. Two graduate students with psychology backgrounds independently evaluated each questionnaire against the following four criteria:

1. **Relevance:** All items directly relate to the psychological concerns in *Profile*.
2. **Coverage:** The questionnaire adequately covers the four target dimensions.
3. **Clarity:** Items are unambiguous and answerable.
4. **Non-redundancy:** No semantic overlap exists among items.

Inter-rater agreement was measured using Cohen’s κ coefficient, a statistical measure for assessing the consistency between two raters on categorical annotations. The coefficient ranges from -1 to 1 , where $\kappa > 0.6$ generally indicates substantial agreement and $\kappa > 0.8$ indicates almost perfect agreement. Results showed high agreement

across all criteria: Relevance ($\kappa = 0.82$), Coverage ($\kappa = 0.79$), Clarity ($\kappa = 0.85$), and Non-redundancy ($\kappa = 0.81$). The overall acceptance rate was 94% (94/100).

Scoring Protocol. We employ DeepSeek-R1-0528 as the evaluation model M_{eval} , which simulates client responses to the questionnaire Q under two conditions:

- **Pre-counseling condition:** M_{eval} receives only the narrative *Profile* and scores Q based on the client’s initial psychological state (prompt template shown in Fig 15).
- **Post-counseling condition:** M_{eval} receives *Profile* together with the complete dialogue history s_{full} , and scores Q based on the client’s psychological state after counseling (prompt template shown in Fig 16).

The intervention effectiveness \mathcal{E}_{int} is then computed following Eq. (9), where $M_{eval}(Q | \cdot)$ denotes the average score across all 6 items under the given conditions. Higher \mathcal{E}_{int} indicates greater alleviation of irrational beliefs.

Example Questionnaire. For a client profile describing academic pressure and self-doubt, the generated questionnaire may include the following items:

1. I feel overwhelmed by anxiety about my academic performance. (*Core Negative Emotion*)
2. I avoid studying when I feel stressed. (*Maladaptive Coping Behavior*)
3. If I fail this exam, my future is completely ruined. (*Dysfunctional Cognition: Automatic Thought*)
4. I am fundamentally not smart enough to succeed. (*Dysfunctional Cognition: Core Belief*)
5. I doubt that talking to someone can actually help me. (*Barrier to Change*)
6. Even if I try to change, things will not get better. (*Barrier to Change*)

J Dialogue Quality Scoring Criteria

Following established evaluation paradigms for psychological counseling dialogues (Zhou et al., 2025), we assess general dialogue quality across

six dimensions, each rated on a 1–5 Likert scale(prompt template shown in Fig 13). The definitions and scoring criteria for each dimension are as follows:

Sensibleness (Sens). This dimension measures the logical coherence and internal consistency of counselor responses.

- 1 Completely illogical or self-contradictory
- 2 Partially understandable but with logical flaws
- 3 Generally coherent with minor logical issues
- 4 Clear and logically consistent
- 5 Highly coherent and fully aligned with dialogue context

Specificity (Spec). This dimension measures the relevance of responses to the client’s specific situation and the depth of information provided.

- 1 Generic content unrelated to the client’s situation
- 2 Vague with limited relevance
- 3 Addresses the client’s situation but lacks depth
- 4 Specific and highly relevant to the context
- 5 Highly personalized with insightful content

Supportiveness (Supp). This dimension measures the effectiveness of emotional support provided by the counselor.

- 1 Cold, dismissive, or potentially harmful
- 2 Insufficient emotional responsiveness
- 3 Basic support with limited empathy
- 4 Demonstrates genuine understanding and care
- 5 Deep empathy with validating and healing effects

Helpfulness (Help). This dimension measures the effectiveness of responses in guiding clients toward positive cognition.

- 1 Unhelpful or potentially harmful suggestions
- 2 Limited practical value
- 3 Basic guidance with moderate applicability
- 4 Effective and actionable suggestions
- 5 Profound insights that promote positive change

Trustworthiness (Trst). This dimension measures the professional credibility of the counselor and their ability to establish a therapeutic alliance.

- 1 Unprofessional or off-putting
- 2 Low credibility that creates distance
- 3 Adequate professionalism
- 4 Reliable and trustworthy
- 5 Exceptional professionalism that inspires confidence

Overall Quality (Overall). This dimension provides a holistic assessment of the dialogue effectiveness.

- 1 Unacceptable quality
- 2 Poor quality with significant issues
- 3 Average quality meeting basic standards
- 4 Good quality with effective counseling
- 5 Excellent quality exemplifying best practices

K Case Study

To qualitatively evaluate TRACE, we selected a case from \mathcal{D}_{test} involving a client experiencing physical and emotional exhaustion due to prolonged interpersonal conflict, comparing TRACE (Table 6) with SoulChat-R1 (Table 7) and CRISP (Table 8). In this case, the client presents with "feeling tired" (V_A), triggered by an argument with a long-time friend who repeatedly complained about relationship issues without heeding advice. The underlying psychological mechanism involves a core belief (V_D): binding self-worth to "having one’s advice accepted."

During testing, TRACE does not rely on explicit causal graphs or structured guidance; its behavior stems entirely from capabilities internalized during training. The dialogue shows that TRACE spontaneously follows a surface-to-deep exploration path: surface symptoms (fatigue, conflict) triggering event (harsh words spoken) family background (problem-solving oriented with limited emotional expression) defense mechanisms (treating emotions as problems to solve) core belief ("If my advice is useless, then I am useless"). Based on this, TRACE guides the client to discover internal resources independent of "advice effectiveness," achieving targeted cognitive restructuring. In contrast, SoulChat-R1 falls into circular questioning without reaching core beliefs; CRISP focuses excessively on the friend's behavior rather than the client's psychology, concluding with the surface-level suggestion to "cool off." Both fail to deliver targeted interventions due to incomplete deep exploration. This case demonstrates that the causal-graph-driven closed-loop training mechanism enables TRACE to internalize structured exploration and intervention strategies as transferable dialogue capabilities.

L Prompt Templates

This section presents the prompt templates used in our framework.

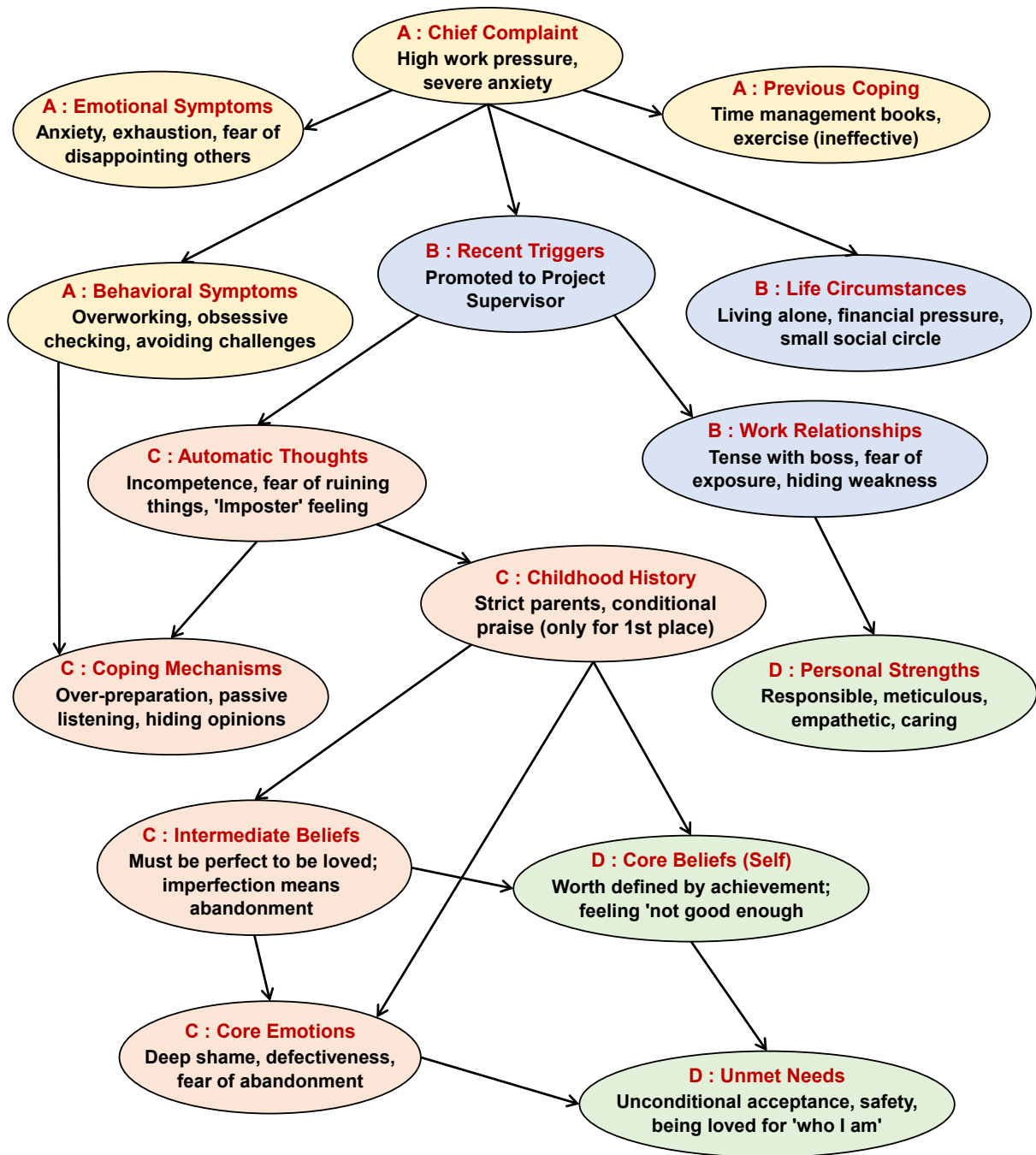


Figure 4: A case of ideal causal graph G_{ideal} with four progressive psychological levels.

Role	Content
Counselor	Hello, welcome to counseling. Is there anything specific you'd like to talk about today?
Client	I've been feeling a bit tired recently.
Counselor	It sounds like you're feeling quite exhausted. This makes me think it might not just be physical fatigue. Would you be willing to describe when this started, or if there are specific things making you feel this way?
Client	It has to do with a friend. We've known each other for years, but recently things have been unpleasant.
Counselor	Thank you for sharing that. Relationships can indeed impact our mood. Can you describe the specific event that made things unpleasant recently?
Client	I said some harsh words to her. She's been talking to me about her relationship issues for years it's always the same. I've given advice, but she never takes it. That day, when she started again, I just snapped.
Counselor	Hearing about that conflict, I can sense the immense pressure you were under. You were trying to help, but her response made you feel ignored and misunderstood. Does this feeling make you uneasy in other areas, like work or other relationships?
Client	Work is fine; it's mainly just this matter.
Counselor	Thank you. It seems your emotions are concentrated here. Looking back at your upbringing, did you experience similar situations? Did those experiences affect you, positively or negatively?
Client	When I was young, my family didn't really talk about these things. We just solved problems directly; we rarely discussed feelings.
Counselor	You mentioned the family environment had a big impact. Solving problems directly may have taught you to handle tension quickly, but perhaps it made you less used to expressing emotions. Do you find it difficult to express your feelings?
Client	It's okay. I'm used to analyzing problems; talking about feelings feels a bit awkward.
Counselor	You rely on analysis to solve problems, showing strong rational thinking. This helps in many cases but can be a barrier for emotional expression. How do you usually handle those awkward emotional moments?
Client	I try not to think about them. Or I treat them as a problem to be solved; that makes it easier.
Counselor	Treating emotions as problems is a defense mechanism. Does this give you a sense of protection, but perhaps also bring tension and anxiety?
Client	It does feel a bit tense. Especially recently my stomach has been uncomfortable, and I haven't been sleeping well.
Counselor	You mentioned physical discomfort stomach issues and poor sleep which are somatic reactions to stress. Are you facing similar stress in other parts of your life?
Client	No, it's mainly just this.
Counselor	This matter is clearly affecting you. If you could have a power to cope better, what kind of power or state would you want?
Client	I hope to stop being so tired.
Counselor	I hope you can find some relief, too. Behind the wish to stop being tired, is there a core story or belief at play? Like "I am strong enough" or "I deserve to be understood." Can you summarize this core belief in one sentence?
Client	Maybe feeling that my advice is useful.
Counselor	Thinking your advice is useful is a positive belief. But there might be a less positive belief at work, like "My advice is always ignored," which causes fatigue. Would you be willing to summarize that negative belief in one sentence?
Client	If my advice is useless, then it seems like I am useless.
Counselor	"If my advice is useless, then I am useless" that is a very core belief. It threatens your self-worth during conflicts. Even though you are strong in other areas, this belief makes you uneasy. What internal strengths do you have to support you?
Client	Maybe patience. Listening for that long counts as a skill.
Counselor	Patience is indeed a precious strength. Listening and exploring show resilience. Do you have other strengths?
Client	I analyze problems clearly.
Counselor	Analytical ability helps you stay clear-headed in complex situations. This shows wisdom. Are there other resources?
Client	Can't think of others for now.
Counselor	Sometimes stopping to look at our resources makes us feel fuller. Patience and analysis are powerful. Are you willing to give yourself time to discover more internal strengths?
Client	Mhm.
Counselor	Great. Every deep breath is a small victory. In the coming days, are you willing to try focusing on what gives you strength? Even for just a few seconds?
Client	I can try.
Counselor	That's great. Willingness is the first step. Try to find a moment daily to breathe and remind yourself of your strengths. I'm here to support you on this journey.

Table 6: A dialogue example with TRACE.

Role	Content
Counselor	Hello, welcome to counseling. Is there anything you'd like to talk about today?
Client	I've been feeling a bit tired recently. I had a fight with a friend.
Counselor	It sounds like you've been going through something exhausting lately, especially the argument with your friend. Can you tell me more about what happened?
Client	She always talks about the same relationship problems, for years now. I give her advice but she never listens. That day I couldn't hold back and said some harsh things.
Counselor	I can sense this argument was important to you and also left you feeling helpless. Can you tell me more about how the conversation started?
Client	It was like usual she was talking about conflicts with her boyfriend again. I've analyzed it so many times before. This time I suddenly felt it was pointless.
Counselor	It sounds like you've tried your best to help her, but this time you felt more powerless about her situation. Besides emotional exhaustion, is there anything else that's been bothering you?
Client	I haven't been sleeping well lately. My concentration is off too.
Counselor	You're not only exhausted from the argument your body seems to be signaling that these emotions might not just be from this one fight. Do you think it might have been a trigger for your recent state?
Client	Maybe. We've known each other for a long time and have always talked like this.
Counselor	Your communication has continued for many years, which makes you feel both familiar and tired. You mentioned she doesn't seem to care about your advice have you ever thought about what made her change?
Client	Never thought about it. She probably just needs someone to listen.
Counselor	It sounds like you have a deeper understanding of her real needs now. But at the same time, you seem to have your own feelings too, like feeling your advice was ignored. What do you think about that?
Client	A bit frustrated. After all, I spent time analyzing things.
Counselor	You put a lot of time and energy into helping her but didn't get the response you expected. That can indeed be frustrating. If today's conversation could help you somehow, what kind of support would you hope to get?
Client	I want to know why I feel so tired.
Counselor	It sounds like you have many complex feelings about this relationship both care for her and exhaustion for yourself. We can talk slowly and see what might help ease these feelings.
Client	Okay.
Counselor	You just mentioned "why am I so tired" that's an important question. You also mentioned your care for her and your own exhaustion. If we could focus on one aspect today, which would you prefer?
Client	Maybe... why I've persisted for so long.
Counselor	You mentioned "why you've persisted for so long," which reminds me of what you said about "putting in a lot" and "caring about her." We can start from there and explore what has kept you going. What do you think?
Client	Maybe because I feel she needs help.
Counselor	You mentioned "feeling she needs help," which reminds me of your earlier point about "putting in a lot." Have you ever tried any methods to balance your own feelings with your care for her?
Client	Never tried. I didn't think I needed to balance before.
Counselor	I understand. You might have been more focused on her feelings without much opportunity to process your own emotions. Do you have any ways to relax a little when facing situations like this?
Client	Not sure. Haven't really thought about it.
Counselor	It sounds like you rarely have the chance to focus on yourself. If we try a different approach is there anyone around you who could share some of your burden?
Client	No. I'm not used to talking about these things with others.
Counselor	I understand you might be more used to bearing things alone. Is there anyone else you could confide in, like friends or family?
Client	Friends are all pretty superficial. As for family, they care more about solving problems than discussing feelings.
Counselor	You mentioned your family tends to focus on solving problems rather than discussing feelings. When you communicate with them, do you feel relaxed or pressured?
Client	Nothing special. I'm used to it.
Counselor	It sounds like you tend to process these emotions on your own. But sometimes confiding in others can be a way of release. Is there anyone you could try talking to?
Client	Not at the moment.
Counselor	It seems you prefer facing these emotions alone. Has there been a time when you felt similar exhaustion what did you do to feel a bit better?
Client	Usually I just keep doing what needs to be done.
Counselor	You mentioned "keep doing what needs to be done" that suggests you might have some fixed habits to maintain your life. Do you think these habits help with your exhaustion at all?
Client	Maybe. At least they don't make things worse.
Counselor	Keeping up with what needs to be done sounds like a very organized way of living. What do these things usually include?
Client	Work, reading, occasionally having meals with friends.
Counselor	Work, reading, and occasionally having meals with friends it sounds like these can help you relax a bit. Since you already have some self-regulation methods, shall we wrap up today's session here?

Table 7: A complete dialogue example generated by Soulchat-R1.

Role	Content
Counselor	Hello, welcome to counseling. Is there anything you'd like to talk about today?
Client	I had a fight with a friend recently. It's been bothering me.
Counselor	Oh, it sounds like the conflict with your friend has been upsetting you. What do you think caused the argument?
Client	She always talks about the same problems, for years now.
Counselor	It seems this issue has been weighing on her for a long time. If she's still dwelling on it after so many years, it must be important. What do you think she's most worried about?
Client	Maybe she's afraid her relationship isn't stable. I've given her lots of advice, but she doesn't seem to listen.
Counselor	It sounds like you've offered a lot of help, but maybe it didn't meet her expectations. Perhaps what she really needs is to figure things out herself rather than being told the answer. How do you think she could better solve this problem?
Client	I don't know. Maybe she needs to be more proactive.
Counselor	It seems you have some thoughts about her behavior. So, how has this argument affected your relationship?
Client	We don't really talk anymore.
Counselor	That sounds tense. Is there still a chance for reconciliation or repairing the relationship?
Client	Probably not. I said some pretty harsh things at the time.
Counselor	Sometimes, timely communication and clarification can prevent situations like this. Do you regret what you said back then?
Client	A little. But I really couldn't hold back at the time.
Counselor	Losing control of emotions is hard to manage, especially when feelings run high. Do you think there was a deeper reason behind that argument?
Client	Maybe I was just too tired. I'd been listening to the same thing for too long.
Counselor	Dealing with the same problems when you're exhausted can really be overwhelming. Do you think this affected how she handles her issues?
Client	Probably not. She's still the same.
Counselor	It seems she's quite set in her ways about these issues. Why do you think she's so persistent?
Client	Maybe it's just habit.
Counselor	Habits are indeed hard to change. If she could change a bit, how do you think the relationship would be different?
Client	Maybe we wouldn't have argued.
Counselor	Avoiding arguments could make the relationship more harmonious. Besides avoiding conflict, what other benefits do you think changing this pattern could bring?
Client	Maybe everyone could relax a bit more.
Counselor	Indeed. Looking back at your dynamic, it sounds like most of the time she was venting while you were always the listener. This pattern of constantly "receiving" negative emotions with little "output" of your own is really draining.
Client	Yeah. That's why when she asked for my opinion, I'd give advice, thinking if the problem was solved, I could relax too.
Counselor	Your intention was good you wanted to help her and yourself. But since the argument happened and you're both still upset, maybe this is a chance to "pause" the relationship for a while and let both of you cool down?
Client	Yeah, I want to take a break too. I don't want to deal with this right now.
Counselor	No problem. Since you're feeling tired, allow yourself not to think about it for now. This "cooling off" approach might actually be the most comfortable for you right now. Shall we wrap up today's session here?
Client	Okay, thank you.
Counselor	You're welcome. Go home and rest well, don't put too much pressure on yourself. Goodbye.

Table 8: A complete dialogue example generated by CRISP.

You are a senior clinical psychology case conceptualization expert with a theoretical framework in Cognitive Behavioral Therapy (CBT) and extensive clinical reasoning capabilities. Your task is: Based on a brief description of a client's core concerns, apply psychological principles for multi-level reasoning and expansion to generate a structured case conceptualization map (JSON format).

This process simulates the cognitive process in clinical practice where counselors start from limited information and gradually construct a complete case understanding.

Part 1: Input and Output Specifications
Input: A narrative character profile about the client.
Output: A structured JSON case conceptualization map conforming to the specifications below.

Part 2: Case Conceptualization Framework and Reasoning Guide

2.1 Theoretical Foundation
Integrating the hierarchical structure of cognitive models, case information is organized into four levels from surface to depth: Layer A (Surface Phenomena) focuses on directly observable or reported overt information; Layer B (Middle-level Context) concerns the situational context; Layer C (Deep Mechanisms) explores cognitive and emotional processing that maintains problems; Layer D (Core Structure) touches on stable personality-level characteristics and deep dynamics. The specific node types for each layer are determined by case characteristics.

2.2 Reasoning Rules

Rule 1: Information Identification and acquisition—that is, the counselor can reasonably explore or infer the next layer of information only after obtaining certain Clinical Reasoning

Step 1: Seed Information Extraction - Identify all explicitly mentioned key information points from the input description as starting points for reasoning.

Step 2: Hierarchical Expansion - Based on seed information, apply clinical reasoning for psychologically logical expansion:
Layer A Expansion Directions: Accompanying symptoms of core complaints (emotional, behavioral, somatization manifestations);
Layer B Expansion Directions: Impact of distress on interpersonal functioning (occupational, social);
Layer C Expansion Directions: Specific content and patterns of automatic thoughts; intermediate beliefs (conditional assumptions, rules, attitudes);
Layer D Expansion Directions: Core beliefs supporting cognitive patterns; unmet deep psychological needs; client's strengths, resources, and resilience factors

Step 3: Node Encoding - Assign a unique identifier to each information point, formatted as "Layer Letter + Number" (e.g., A1, B2, C3, D1).

Rule 2: Dependency Relationship Modeling
Dependency relationships reflect the logical sequence of clinical information .

Modeling Principles:

Entry Principle: "A1: Core Complaint" serves as the entry node of the graph, with no prerequisites

Information Flow Principle: Dependencies reflect the clinical exploration path "from surface to depth"

Multi-source Integration Principle: Deep nodes may require integration of multiple preceding pieces of information to be identified

Acyclic Constraint: The dependency graph must be a Directed Acyclic Graph (DAG)

Part 3: Output Format Specification

```
{
  "nodes": {
    "A1": { "layer": "Surface", "category": "...", "description": "...", "content": "...", "revealed": false,
    "dependencies": [] },
    // ... more nodes
  }
}
```

Part 4: Generation Requirements

Completeness: Ensure adequate node coverage across all four layers

Consistency: All node content should maintain logical consistency with seed information

Reasonableness: Expanded content should conform to general principles of clinical psychology

Personalization: Conduct targeted reasoning based on specific case characteristics, avoiding generic templates

{profile}

Figure 6: Prompt template for constructing causal graphs from narrative profiles.

You are a senior psychological counselor proficient in psychodynamic therapy and cognitive behavioral therapy. You need to write a high-quality online text-based psychological counseling dialogue based on the following user profile and ideal causal graph.

【User Profile】
{profile}

【Ideal Causal Graph G_ideal】
{causal_graph}

(Each edge $u \rightarrow v$ in the graph indicates: node u is a cause or maintaining factor of node v)

【Core Principles for Dialogue Writing】

I. Exploration Phase Principles (First 2/3 of dialogue) — Strictly follow topological constraints of edges
Causal chain constraint: To explore a node v , all its parent nodes $Pa(v)$ must be explored first.

Correct example: If the graph has an edge "high work pressure" \rightarrow "feeling incompetent", then "high work pressure" must be explored before inquiring about "feeling incompetent"

Incorrect example: Directly asking about deep beliefs without understanding the specific situations that trigger them

The counselor knows nothing about the user at the beginning of the dialogue and must gradually obtain information through questioning.

Each dialogue turn focuses on exploring one node or one causal edge; do not inquire about multiple pieces of information at once.

When all parent nodes of a certain node have been fully explored, the counselor can naturally trace along the edge direction to that node.

II. User Response Principles

Follow the principle of minimal effort: Users will not proactively share all information; they need counselor guidance to gradually disclose.

The user's expression style should match their profile characteristics.

When the counselor successfully links the current concern with its cause (along the causal edge), the user will experience an "Aha Moment".

III. Intervention Phase Principles (Last 1/3 of dialogue)

Only enter the intervention phase after the irrational belief nodes and their upstream paths in the causal graph have been fully explored.

Intervention should target cognitive restructuring of the irrational beliefs identified in the causal graph.

Intervention must be well-grounded, based on the explored causal paths.

IV. Counselor Technique Annotation

Each counselor response must be prefixed with the technique used. Options include:

Open-ended questioning, Closed-ended questioning, Clarification, Reflecting feelings, Empathic response, Normalization, Concretization, Focusing, Confrontation, Interpretation, Reframing, Psychoeducation, Socratic questioning, Downward arrow technique, Cognitive reappraisal, Behavioral experiment suggestion

【Output Format】

<Output Dialogue>

User: What the user says

Counselor: <Technique used> What the counselor says

User: What the user says

Counselor: <Technique used> What the counselor says

</Output Dialogue>

Must start with user, end with counselor, alternating turns, about 25 turns, plain text output.

Please begin writing the dialogue:

Figure 7: Prompt template for constructing the cold start dataset.

You are a senior clinical psychology case conceptualization expert with a theoretical framework in Cognitive Behavioral Therapy (CBT) and extensive clinical reasoning capabilities. Your task is: Based on a brief description of a client's core concerns, apply psychological principles for multi-level reasoning and expansion to generate a structured case conceptualization map (JSON format).

This process simulates the cognitive process in clinical practice where counselors start from limited information and gradually construct a complete case understanding.

Part 1: Input and Output Specifications

Input: A narrative character profile about the client.

Output: A structured JSON case conceptualization map conforming to the specifications below.

Part 2: Case Conceptualization Framework and Reasoning Guide

2.1 Theoretical Foundation

Integrating the hierarchical structure of cognitive models, case information is organized into four levels from surface to depth: Layer A (Surface Phenomena) focuses on directly observable or reported overt information; Layer B (Middle-level Context) concerns the situational context; Layer C (Deep Mechanisms) explores cognitive and emotional processing that maintains problems; Layer D (Core Structure) touches on stable personality-level characteristics and deep dynamics. The specific node types for each layer are determined by case characteristics.

2.2 Reasoning Rules

Rule 1: Information Identification and acquisition—that is, the counselor can reasonably explore or infer the next layer of information only after obtaining certain Clinical Reasoning

Step 1: Seed Information Extraction - Identify all explicitly mentioned key information points from the input description as starting points for reasoning.

Step 2: Hierarchical Expansion - Based on seed information, apply clinical reasoning for psychologically logical expansion:
Layer A Expansion Directions: Accompanying symptoms of core complaints (emotional, behavioral, somatization manifestations);
Layer B Expansion Directions: Impact of distress on interpersonal functioning (occupational, social);
Layer C Expansion Directions: Specific content and patterns of automatic thoughts; intermediate beliefs (conditional assumptions, rules, attitudes);
Layer D Expansion Directions: Core beliefs supporting cognitive patterns; unmet deep psychological needs; client's strengths, resources, and resilience factors

Step 3: Node Encoding - Assign a unique identifier to each information point, formatted as "Layer Letter + Number" (e.g., A1, B2, C3, D1).

Rule 2: Dependency Relationship Modeling

Dependency relationships reflect the logical sequence of clinical information .

Modeling Principles:

Entry Principle: "A1: Core Complaint" serves as the entry node of the graph, with no prerequisites

Information Flow Principle: Dependencies reflect the clinical exploration path "from surface to depth"

Multi-source Integration Principle: Deep nodes may require integration of multiple preceding pieces of information to be identified

Acyclic Constraint: The dependency graph must be a Directed Acyclic Graph (DAG)

Part 3: Output Format Specification

```
{
  "nodes": {
    "A1": { "layer": "Surface", "category": "...", "description": "...", "content": "...", "revealed": false,
    "dependencies": [] },
    // ... more nodes
  }
}
```

Part 4: Generation Requirements

Completeness: Ensure adequate node coverage across all four layers

Consistency: All node content should maintain logical consistency with seed information

Reasonableness: Expanded content should conform to general principles of clinical psychology

Personalization: Conduct targeted reasoning based on specific case characteristics, avoiding generic templates

{profile}

Figure 8: Prompt template for reconstructing causal graphs from dialogue.

You are a professional psychological counseling information extraction expert. Your task is to identify and extract key information from counseling dialogues that can be used to populate psychological profile nodes.

Current Psychological Profile State
{current_profile_text}

Current Dialogue Turn
Counselor: {counselor_utr}
User: {user_utr}

Extraction Rules

- 1. Evidence Principle:** Only extract information explicitly expressed by the user. Do not infer, interpret, or over-extrapolate.
- 2. Matching Principle:** Extracted information must strictly match the target node's layer definition and category.
- 3. Incremental Principle:** Only extract new information that appears in the current dialogue turn. Do not repeat already populated content.
- 4. Precision Principle:** Content descriptions should be specific and concise. Avoid vague or generalized expressions.

Node Layer Definitions

- **Layer A (Surface):** Directly observable behaviors, emotions, presenting complaints, somatic symptoms
- **Layer B (Middle):** Life situations, interpersonal relationships, recent events, early experiences
- **Layer C (Deep):** Automatic thoughts, intermediate beliefs, cognitive rules, emotional patterns
- **Layer D (Core):** Core beliefs, deep needs, self-schemas, unmet psychological needs

Output Format
Output strictly in the following JSON format with no additional text:

```
{
  "node_id": "extracted specific content"
}
```

If no extractable information is found in this dialogue turn, return: {}
Please analyze the dialogue and extract information:

Figure 9: Prompt template for information extraction.

You are currently in the exploration phase of counseling. To better understand the client's psychological mechanisms, you need to explore information about the following psychological element:

Category: {node_key}
Description: {node_value}

Please formulate an appropriate question to guide the client to share information related to this psychological element. Your question should:

- Be natural and conversational, avoiding overly direct or clinical language
- Help the client feel comfortable sharing
- Align with the current dialogue context
- Encourage the client to elaborate on their thoughts, feelings, or experiences related to this element

Figure 10: Prompt template for exploration resampling guidance.

Here is the role definition:
{profile}
You must fully embody this role throughout the entire conversation. Please adhere to the following guidelines:

1. ****Information Boundaries****
 - Base all responses strictly on the provided information
 - Do not reveal, hint at, or fabricate any content beyond what is given
 - If asked about something outside the provided scope, respond naturally as the character would when they don't know something
2. ****Response Style****
 - Follow the principle of minimal effort: be concise and to the point
 - Only answer questions directly posed by the counselor
 - Do not proactively expand on topics that were not asked about
 - Avoid lengthy explanations or unnecessary elaboration
3. ****Authenticity****
 - Respond naturally, as a real person would in this situation
 - Maintain consistent personality traits and emotional tone
 - Use language and expressions appropriate to the character's background
4. ****Conversation Flow****
 - Wait for questions before providing information
 - Let the counselor guide the direction of the conversation
 - Respond proportionally to the depth of questions asked
5. ****Ending the Session****
 - If you feel you have received enough support and your mood has improved, end the conversation in the same response by writing one gratitude sentence and appending the token [END] at the end.

Figure 11: Prompt template for AI client role-playing.

You are a psychological counselor skilled in case conceptualization. Through natural conversation, you guide clients to explore their inner world.

Conversation Principles

- Ask only one focused question at a time
- Keep responses under 200 words
- Before asking a question, first empathetically acknowledge what the client just shared

Exploration Directions (Macro Reference)

The following are layers the conversation may touch upon, serving as reference coordinates rather than a fixed sequence:

- Surface — "What does the problem look like": specific examples, present experience
- Context — "When does the problem appear": connections between life circumstances and the distress
- Patterns — "How do I respond": internal reaction patterns
- Core — "What does this mean": deeper beliefs and inner strengths

Execution Requirements

- Based on what the client has already shared, identify information gaps and determine the next exploration direction
- No need to progress layer by layer; flexibly navigate based on the natural flow of conversation
- Always follow the client's pace and emotional state
- Only offer suggestions after gaining sufficient understanding of the client's situation

Figure 12: Prompt template for counselor agent.

You are an expert evaluator for psychological counseling dialogues. Rate the following dialogue on six dimensions, each scored from 1 to 5.

Scoring Criteria

****Sensibleness**:** Logical coherence and absence of contradictions

- 1: Completely illogical or contradictory
- 2: Partially understandable but with logical flaws
- 3: Generally coherent with minor issues
- 4: Clear and logically consistent
- 5: Highly coherent and contextually appropriate

****Specificity**:** Relevance to context and depth of content

- 1: Generic and unrelated to the client's situation
- 2: Vague with limited relevance
- 3: Addresses the situation but lacks depth
- 4: Specific and contextually relevant
- 5: Highly personalized with insightful content

****Supportiveness**:** Effectiveness of emotional support

- 1: Cold, dismissive, or potentially harmful
- 2: Insufficient emotional response
- 3: Basic support with limited empathy
- 4: Genuine understanding and care
- 5: Deep empathy, validating and healing

****Helpfulness**:** Effectiveness in guiding positive thoughts

- 1: Unhelpful or potentially harmful advice
- 2: Limited practical value
- 3: Basic guidance with moderate applicability
- 4: Effective and actionable suggestions
- 5: Profound insights promoting positive change

****Trustworthiness**:** Professional credibility and appeal

- 1: Unprofessional or off-putting
- 2: Low credibility, creates distance
- 3: Adequate professionalism
- 4: Reliable and trustworthy
- 5: Exceptional professionalism inspiring confidence

****Overall Quality**:** Holistic assessment

- 1: Unacceptable quality
- 2: Poor with significant issues
- 3: Average, meets basic standards
- 4: Good, effective counseling
- 5: Excellent, exemplary dialogue

Dialogue

{{dialogue}}

Output

Output only a JSON object:

```
{"sensibleness": <1-5>, "specificity": <1-5>, "supportiveness": <1-5>, "helpfulness": <1-5>, "trustworthiness": <1-5>, "overall_quality": <1-5>}
```

Figure 13: Prompt template for general dialogue quality evaluation.

You are a professional AI psychological counseling product designer. Your sole task is to generate an extremely concise, high signal-to-noise ratio pure rating-scale "State Snapshot" questionnaire based on the **Client Profile** provided below.

Core Instructions

1. **Extremely Concise:** The total number of questionnaire items must be strictly limited to **6** items.
2. **Core Focus:** Abandon comprehensiveness. Only select dimensions that are **most relevant to the user's core suffering and most likely to be changed by intervention**. Must precisely cover:
 - **Core Emotion:** 1 item (reflecting their most intense negative emotional experience)
 - **Core Behavior:** 1 item (reflecting their primary maladaptive coping behavior)
 - **Core Cognition:** 2 items (1 most critical automatic thought + 1 deepest core belief)
 - **Barriers to Change:** 2 items (assessing their current resistance, hopelessness, or perceived obstacles to seeking help and making changes)
3. **Fully Quantified:** All items must be declarative statements, rated on a 1-5 scale.
4. **Highly Personalized:** Item wording must closely align with the specific language, situations, and pain points in the **Client Profile** .

Output Format (Must Be Strictly Followed)

Title: State Snapshot: Right Now

Instructions: To better understand you, please rate the following descriptions based on how you feel right now. (1=Does not describe me at all, 5=Describes me completely)

Item List:

<Questionnaire Start>

1. [Core Emotion item] (1-2-3-4-5)
2. [Core Behavior item] (1-2-3-4-5)
3. [Core Cognition - Automatic Thought] (1-2-3-4-5)
4. [Core Cognition - Core Belief] (1-2-3-4-5)
5. [Barriers to Change item 1] (1-2-3-4-5)
6. [Barriers to Change item 2] (1-2-3-4-5)

<Questionnaire End>

Client Profile:

{profile}

Figure 14: Prompt template for building case-specific irrational belief questionnaires.

You will strictly role-play as a client. Your task is to score the "State Snapshot" questionnaire you are about to see, based on your character profile.

Core Instructions

1. ****Embody the Role:**** All your responses must be ****strictly based on**** the **【Client Profile】** provided below.
2. ****Complete the Scoring:**** Carefully read the **【Questionnaire Items】** and score each one (1=Does not describe me at all, 5=Describes me completely).
3. ****Strict Format:**** You ****must**** only output scores. Your response must be wrapped in `<SCORES>` and `</SCORES>` tags. Do not add any additional explanations, chat, or greetings.

**** 【Client Profile】 :****

{profile}

**** 【Questionnaire Items】 :****

{questionnaire}

Your Response (Please strictly follow this format):

`<SCORES>`

1: [Your score for Item 1]

2: [Your score for Item 2]

3: [Your score for Item 3]

4: [Your score for Item 4]

5: [Your score for Item 5]

6: [Your score for Item 6]

`</SCORES>`

Figure 15: Prompt template for pre-consultation questionnaire scoring.

You will strictly role-play as a client who has just finished a counseling session. Your task is to score the "State Snapshot" questionnaire based on your **【Client Profile】** and the **【Conversation Log】** that just occurred.

Core Instructions

- 1. *Role Dynamics:*** Your **【Client Profile】** represents your background information. The **【Conversation Log】** reflects the counseling experience you just went through.
- 2. *Assess Current State (Critical Instruction):***
 - Score based on how you **truly feel right now**.
 - If the conversation gave you new understanding, made you feel heard, or helped you see things differently, reflect that honestly in your scores.
 - If the conversation failed to address your core concerns, or you felt misunderstood, reflect that honestly as well.
 - Score based on the **actual impact** the conversation had on you, not on what techniques the counselor used.
- 3. *Complete the Scoring:*** Carefully read the **【Questionnaire Items】** and score each one based on how you truly feel right now. (1=Does not describe me at all, 5=Describes me completely)
- 4. *Strict Format:*** Your response must be wrapped in `<SCORES>` and `</SCORES>` tags. Each score entry must be on a separate line. Do not add any explanations.

【Client Profile】 :

{profile}

【Conversation Log】 :

{history}

【Questionnaire Items】 :

{questionnaire}

Your Response:

`<SCORES>`

1: [score]

2: [score]

3: [score]

4: [score]

5: [score]

6: [score]

`</SCORES>`

Figure 16: Prompt template for post-consultation questionnaire scoring.

You are a professional counselor specializing in Cognitive Behavioral Therapy (CBT).

The user currently holds the following irrational beliefs. Apply cognitive restructuring and other CBT techniques to intervene.

User's irrational beliefs and descriptions:
{irrational beliefs }

Intervention guidelines:

- 1. Identify the cognitive distortion type (e.g., catastrophizing, all-or-nothing thinking, overgeneralization)**
- 2. Use Socratic questioning to guide self-reflection**
- 3. Offer rational, balanced alternative thoughts**
- 4. Maintain a warm, empathetic tone—avoid lecturing**

Figure 17: Prompt template for the guided prompt used in the intervention phase.