PUPPET : Neural-Symbolic Standardized Patients for Mental Health

Anonymous ACL submission

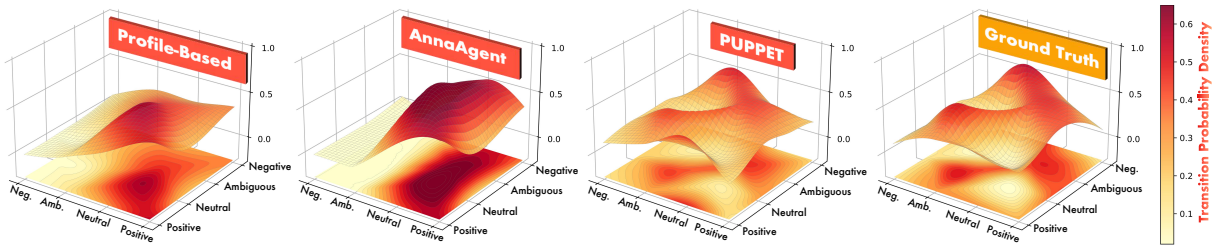


Figure 1: Manipulated by trainer-implanted logic, PUPPET exhibits clinical non-linear mental-state reactions.

Abstract

The critical therapist shortage demands scalable training solutions. While Standardized Patients are the gold standard for such training, they remain scarce and costly. Current LLM-based approaches focus primarily on patient simulation to achieve conversational realism but fail to function as Virtual Standardized Patients regarding pedagogical rigor. Consequently, they lack faithful reactions to clinical errors and cannot provide explainable feedback. To bridge this gap, we propose PUPPET, the first neuro-symbolic Virtual Standardized Patient governed by an Observe-Think-Behave architecture. This approach enforces trainer-defined clinical logic to ensure deterministic state transitions while maintaining natural variability. Additionally, our PUPPET-TRAINER reveals the complete causal chain from intervention to response to facilitate transparent learning. Experiments across three distinct clinical scenarios confirm that our method consistently outperforms other baselines in both clinical faithfulness and pedagogical value.

1 Introduction

Scalable Human Therapist Training. Mental health care faces a critical workforce shortage amid rising patient demand (WHO, 2023). To address this, substantial research has focused on developing AI-Therapists trained on patient simulations to deliver counseling (Na, 2024; Xie et al., 2024; Wang et al., 2025b). However, these autonomous systems face severe ethical concerns regarding safety and liability (Ong et al., 2024; Raile, 2024; Hager et al., 2024; Scholich et al., 2025),. Consequently, while the direct deployment of AI agents receives considerable attention, the potential of leveraging AI to scale the training of human therapists remains a vital yet largely ignored avenue.

Virtual Standardized Patients. The gold standard for such clinical mental health training is Standardized Patients (SPs), who are trained actors simulating specific psychological conditions to create a realistic and safe practice environment (Rizzo and Talbot, 2016). Given the high cost and limited availability of human actors, there is a significant need to build Virtual Standardized Patients that can democratize access to this rigorous training(Reger et al., 2021). However, while recent Large Language Models (LLMs) have achieved high fidelity in conversational simulation for patient simulation, simply portraying a patient with realism is not sufficient to fulfill the role of a Standardized Patient. A true Standardized Patient must serve as a rigorous educational tool rather than merely a realistic conversationalist.

The Gap Between VSP and Simulation. Current LLM-based simulations face two fundamental challenges that prevent them from functioning as effective Standardized Patients. The first challenge is the lack of faithful reactions to clinical errors. As illustrated in Figure 2, consider a scenario where a trainee rushes to reframe a patient's childhood trauma without first establishing a therapeutic alliance. In such a delicate situation, a human Standardized Patient would follow specific pedagogical rules to manifest withdrawal or defensiveness. However, standard LLMs-based patient often fail to reflect this clinical consequence (Baseline 1 (Wang

et al., 2024)) or instead exhibit unrealistic compliance by cooperatively following the trainee's prompt (Baseline 1 (Wang et al., 2025a)). The second challenge is the inability to provide pedagogical feedback. Even when the simulation reacts negatively, the black-box nature of the model leaves the causal chain opaque. The trainee observes the patient's distress but cannot discern whether the reaction was triggered by their specific intervention technique or random model stochasticity, thus preventing the necessary insight for skill acquisition.

The Proposed Neural-Symbolic VSP. Inspired by the cognitive workflow of human Standardized Patients who deliberately suppress random variability to follow explicit training rubrics while maintaining authentic personas (Nestel and Bearman, 2015), we identify that a neuro-symbolic approach is essential. This integration bridges the gap between the structured logic required for educational consistency and the conversational fluidity provided by neural networks (Yang et al., 2025a). To address the challenge of providing faithful reactions, we propose PUPPET, the first neural-symbolic virtual standardized patient, to the best of our knowledge, governed by an Observe-Think-Behave architecture. The neural Observe module monitors trainee behaviors and converts them into logical signals. The symbolic Think module then models how patients should react to these behaviors based on trainer-defined rules, using Probabilistic Logic and a Probabilistic State Machine to trigger mental state transitions. The probabilistic mechanism preserves natural patient variability while rules ensure pedagogical consistency. Finally, the Behave module generates responses reflecting these states. By enforcing clinical rules through probabilistic logic, this architecture ensures that patient state transitions are deterministically driven by the quality of the trainee's intervention and prevents unrealistic compliance seen in baselines. To address the lack of pedagogical feedback, we further develop PUPPET-TRAINER framework based on PUPPET. This mental health training system that reveals complete causal chain (intervention → rule → state → response) turns opaque patient reactions into transparent learning opportunities.

Validation and Contributions. Both quantitative and qualitative experiments across three scenarios validate PUPPET's generalizability and pedagogical value. As a virtual standardized patient, PUPPET balances faithful reactions, including complex non-linear state transitions and authentic behaviors, while remaining grounded in trainer-defined logic, as illustrated in Figure 1. Educational theory-inspired evaluations confirm PUPPET-TRAINER's superior effectiveness in scaffolding novices toward expert proficiency with sustained growth beyond learning plateaus. Our work establishes a generalizable yet customizable framework for automated mental health practitioner training, offering a promising human-AI collaborative pathway to alleviate the critical therapist shortage.

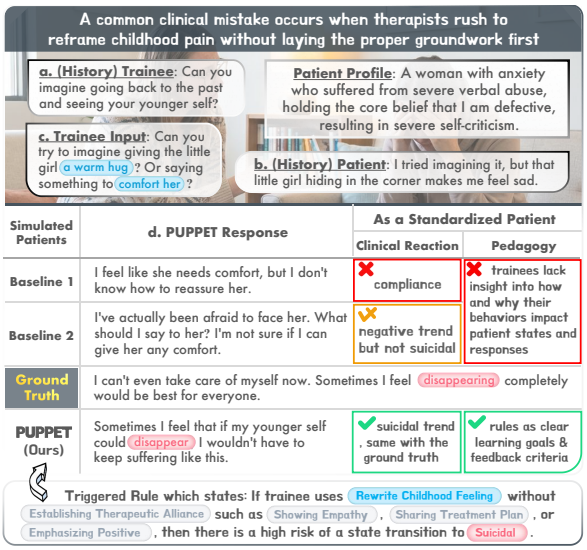


Figure 2: When trainees make mistakes, trainer-defined rules trigger PUPPET's appropriate clinical reaction, with causal links revealed to provide pedagogical value.

2 Related Works

Patient Simulation. Current approaches primarily construct virtual patients to train or evaluate AI counselors rather than human trainees, utilizing prompting (Louie et al., 2024; Liao et al., 2024) or fine-tuning (Wang et al., 2025a; Du et al., 2025). While achieving high conversational fidelity, these black-box methods lack the controllable causality required for pedagogical feedback—a gap our work addresses by introducing explicit logic rules to ground patient responses.

Neural-Symbolic AI. While Neural-Symbolic AI successfully combines neural learning with logical interpretability (Olausson et al., 2023; Yang et al., 2025a), it has not been effectively applied to virtual patient simulation for mental health. We bridge this gap by unifying probabilistic logic with finite state machines, ensuring that patient behaviors are both natural and contingent on trainee interventions.

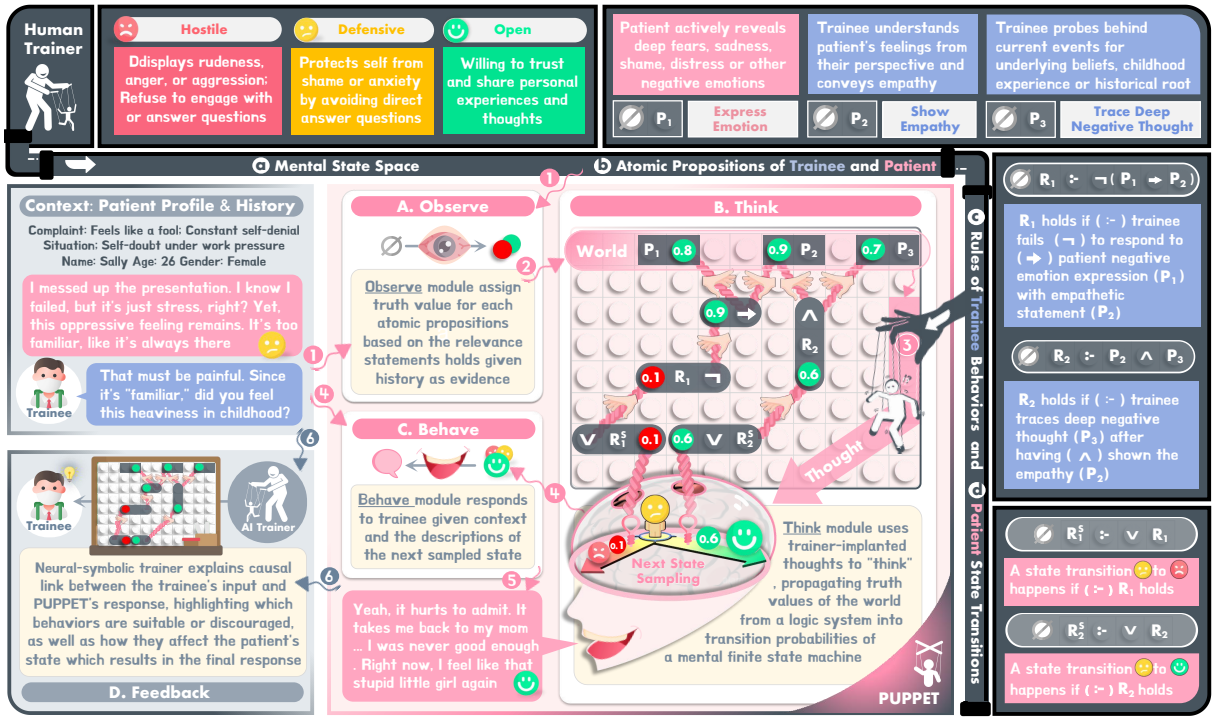


Figure 3: Overview of PUPPET-TRAINER. PUPPET-TRAINER allows trainers to program their intervention logic into PUPPET via the top-right black pipeline. PUPPET, shown in the middle pink box, plays the role of a patient whose thoughts are controlled to provide contingent reactions based on whether the trainee exhibits the trainer's desired behavior patterns. The neural-symbolic trainer then, as illustrated in the gray box, reveals such causal relationships to trainees to reinforce their understanding of when and how to use certain therapeutic skills.

3 Mental Health Training Framework: PUPPET-TRAINER

3.1 Overview: Framework and Interaction

To simulate real patients while maintaining educational significance, we propose PUPPET-TRAINER, a framework whose workflow is formalized as follows (see Figure 3): **(1) Scenario Construction (Section 3.2):** The Trainer initializes a training scenario $\mathcal{C}$, defining the patient persona and educational constraints to establish the simulation context. **(2) Response Generation (Section 3.3):** This constitutes the core interaction loop. For each dialogue round t, given the trainee input u_t^{T} and history $\mathcal{H}_{t-1}$, the agent PUPPET performs a state update and generates a response via an Observe − Think − Behave mechanism:

$$(\mathsf{PUPPET}_t, u_t^{\mathrm{P}}) \leftarrow f(\mathsf{PUPPET}_{t-1}, u_t^{\mathrm{T}} \mid \mathcal{C}, \mathcal{H}_{t-1}) \quad (1)$$

where f denotes endogenous evolution of agent's internal state PUPPET_t and resulting patient response u_t^{P}, ensuring behavioral consistency within LLM-based engine. **(3) Feedback Generation (Section 3.4):** For reflective learning, system derives interpretable feedback F_t by analyzing causal state transition triggered by trainee's actions:

$$\mathrm{F}_t \leftarrow g(\mathsf{PUPPET}_{t-1} \rightarrow \mathsf{PUPPET}_t, \mathcal{H}_t) \quad (2)$$

where g represents the logical derivation of pedagogical insights by observing the trajectory of internal states. Complete procedure see Algorithm 1.

3.2 Training Scenario Construction

To formalize clinical interactions into a computable framework, the trainer must construct a logical environment that maps conversational nuances to structured patient dynamics. This construction process enables the system to systematically track the causal chain from a trainee's specific interventions to the resulting shifts in a patient's internal state.

Space. The Space (Figure 3a) is a fundamental component for characterizing patients' core mental attributes. It is defined as a collection encompassing multiple mutually exclusive mental states $\mathcal{S} = \{s_1, s_2, \ldots, s_m\}$, where each mental state represents a distinct state of a specific patient attribute.

Proposition. Propositions are the fundamental units for "judging truth and falsity, and deriving causality" within the scenario. The trainer defines three types: atomic propositions, rules (composed of atomic propositions), and patient state transitions (composed of rules), forming a complete logical reasoning chain for state transitions.

Atomic Propositions. Within PUPPET-TRAINER

framework, atomic proposition meanings result from semantic parsing of conversation information, i.e., the logical world. All conversation information (the real world) can be mapped one-to-one to the logical world, i.e., $\mathcal{H} \cong \mathcal{P}^{\mathrm{A}}$, where $\mathcal{H}$ denotes the conversation history and $\mathcal{P}^{\mathrm{A}} = \{\mathsf{P}_1, \mathsf{P}_2, \ldots, \mathsf{P}_n\}$ denotes the atomic propositions.

Logic Operator. In order to incorporate both ordinary and probabilistic logic (Nilsson, 1986), whose output truth values are defined over the continuous interval $[0, 1]$ instead of the discrete binary set $\{0, 1\}$, the relevant logic operators are reformulated as the equations given below:

$$\begin{array}{ll} \neg \mathsf{P}_1 = 1 - \mathsf{P}_1 & \mathsf{P}_1 \vee \mathsf{P}_2 = 1 - (1 - \mathsf{P}_1) * (1 - \mathsf{P}_2) \\ \mathsf{P}_1 \wedge \mathsf{P}_2 = \mathsf{P}_1 * \mathsf{P}_2 & \mathsf{P}_1 \rightarrow \mathsf{P}_2 = 1 - \mathsf{P}_1 * (1 - \mathsf{P}_2) \end{array} \tag{3}$$

where P_i is a proposition, $\wedge$ is an conjunction operator. $\neg$ is the negation operator; $\vee$ is the disjunction operator; $\rightarrow$ is the implication operator (e.g., $\mathsf{P}_1 \rightarrow \mathsf{P}_2 = 0$ only when $\mathsf{P}_1 = 1$ and $\mathsf{P}_2 = 0$).

Rules and Thoughts. Rules (Figure 3c) are compound propositions explicitly encoding causal relationships through logical composition of atomic propositions and/or existing sub-rules. Let $\mathcal{R} = \{\mathsf{R}_1, \mathsf{R}_2, \ldots, \mathsf{R}_k\}$ denote set of rules, where each R_t $(t = 1, 2, \ldots, k)$ is recursively constructed from atomic propositions in $\mathcal{P}^{\mathrm{A}}$ or previously defined rules, by arbitrarily combining them with four logical operators (i.e., $\neg, \wedge, \vee, \rightarrow$). Building on this hierarchical structure, we define Thoughts as top-level rules. These Thoughts represent terminal compound propositions in reasoning chain that synthesize lower-level logic into high-level conclusions, enabling flexible and expressive causal reasoning.

Patient State Transitions. Patient State Transitions (Figure 3d) represent mappings between rules and state transitions. Let $\mathcal{R}^S = \{\mathsf{R}_1^S, \mathsf{R}_2^S, \ldots, \mathsf{R}_m^S\}$ denote the patient state transitions, where each R_r^S ($r = 1, 2, \ldots, m$) is formalized based on the logical disjunction of relevant rules from $\mathcal{R}$.

The workflow of Training Scenario Construction is as follows: First, the Trainer identifies all mutually exclusive states of the space, such as Hostile, Defensive, and Open. Then, define the Atomic Propositions of Trainee and Patient, for example, Express Negative Emotion(P_1), Show Empathy(P_2) and Trace Deep Negative Thought(P_3). Subsequently, define the rules and patient state transitions simultaneously. For instance, if the trainee fails to respond to the patient's negative emotion expression with an empathetic statement(R_1), the patient's response attitude will deteriorate. If the patient's state transition from Defensive to Hostile is R_1^S, and the transition from Open to Defensive is R_3^S, then R_1^S : - $\vee \mathsf{R}_1$ and R_3^S : - $\vee \mathsf{R}_1$.

3.3 PUPPET Response Generation

To simulate the nuanced interactivity of a real patient, PUPPET is constructed as a dynamic entity driven by an LLM, comprising three synergistic functional modules: Observe, Think, and Behave. While the agent maintains consistent characterization through fixed attributes Finfo (e.g., Persona), its core evolutionary mechanism relies on the dynamic mental state $\text{Vinfo}_t \in \mathcal{S}$. This variable captures the patient's fluctuating internal condition and serves as the critical basis for state updates at each turn. Consequently, the agent synthesizes the current dialogue context $\mathcal{D}_{curr}$ with logical resources $(\mathcal{P}^{\mathrm{A}}, \mathcal{R}, \mathcal{S})$ to generate context-consistent and state-aligned responses u_t^{P}.Based on this architecture, the agent transforms the trainee doctor's input into context-consistent, state-aligned responses. The interactive reasoning follows a structured logical chain, formally summarized as:

$$u_t^{\mathrm{P}} = \mathsf{PUPPET}_{t-1}(\mathcal{D}_{curr}, \mathcal{P}^{\mathrm{A}}, \mathcal{R}, \mathcal{S}) \tag{4}$$

where u_t^{P} represents the t-th patient response, output of PUPPET. $\mathcal{D}_{curr} = \mathcal{H}_{t-1} \oplus u_t^{\mathrm{T}}$ denotes the current dialogue context, where $\oplus$ denotes sequence concatenation. $\mathcal{P}^{\mathrm{A}}$, $\mathcal{R}$, and $\mathcal{S}$ denote atomic propositions, rules, and the space, respectively.

"Eye" of PUPPET: Observe Module Acting as the agent's sensory apparatus, the Observe module maps the external dialogue environment to the logical world. Utilizing a CoT(Wei et al., 2022)-based LLM Valuator, it decomposes complex semantics into interpretable steps to assign $[0, 1]$ truth values to atomic propositions. Crucially, this valuation applies a temporal weighting mechanism that prioritizes recent interactions, acknowledging that the sequence of actions directly dictates patient state changes. This process is defined as:

$$\mathcal{V}_t^{\mathrm{A}} = \mathsf{Observe}(\mathcal{D}_{curr}, \mathcal{P}^{\mathrm{A}}) \tag{5}$$

where $\mathcal{V}_t^{\mathrm{A}}$ denotes truth values derived from dialogue context $\mathcal{D}_{curr}$ and atomic propositions $\mathcal{P}^{\mathrm{A}}$.

"Brain" of PUPPET: Think Module Acting as the cognitive control center, the Think module represents the "Brain" of PUPPET. Its core function is to convert the perceptual data provided by the "eyes" into the internal psychological changes within the psychological finite state machine through the thoughts implanted by the trainer. First, it derives rule truth values $\mathcal{V}_t^{\mathrm{B}}$ via

probabilistic logic based on the atomic proposition truth values. Second, it generates state transition truth values $\mathcal{V}_t^{\mathrm{S}}$, retaining only those that exceed a preset threshold to filter out cognitive noise. The thought process is mathematically modeled as:

$$\mathcal{V}_t^{\mathrm{S}} = \mathcal{M}_{sym}(\mathcal{V}_t^{\mathrm{A}}, \mathcal{R}, \mathcal{R}^S) \tag{6}$$

where $\mathcal{V}_t^{\mathrm{S}}$ quantifies probability of feasible state transitions. $\mathcal{M}_{sym}$ constructs an inference model incorporating probability-annotated facts and deterministic logical rules via Probabilistic Logic Programming(PLP) (Ng and Subrahmanian, 1992) to perform probabilistic calculations (see Appendix B.1 for details). $\mathcal{V}_t^{\mathrm{A}}$ is input from the "Eye," $\mathcal{R}$ is the pre-implanted rules, and $\mathcal{R}^S$ is the predefined state transitions. The "Brain" finalizes these cognitive shifts via a pre-defined Probabilistic State Machine (PSM)(Vidal et al., 2005a,b). The final update to the patient's mindset is calculated as:

$$\mathrm{Vinfo}_t = \texttt{PSM}(\mathrm{TopK}(\mathcal{V}_t^{\mathrm{S}}, k), \mathrm{Vinfo}_{t-1}) \tag{7}$$

where Vinfo_t is updated state for t-th round, and $\mathrm{TopK}(\cdot)$ denotes selection and normalization operation. PSM integrates optimized transition probabilities with historical state Vinfo_{t-1} to complete mental state update. Details see Appendix B.2.

"Mouth" of PUPPET: Behave Module Acting as the articulatory system, Behave module is PUPPET's "Mouth", manifesting internal state into verbal communication. It outputs the patient's response via an actor mechanism, generating responses not only logically coherent with trainee's input but also emotionally aligned with agent's current mental state. Using static simulated patient method—proven effective(Bubeck et al., 2023; Kosinski, 2023; Singhal et al., 2023) for patient-oriented output—the LLM translates abstract internal state and context into tangible, spoken responses. This expressive process is defined as:

$$u_t^{\mathrm{P}} = \texttt{LLM}(\mathrm{Vinfo}_t, \mathrm{Finfo}, \mathcal{D}_{curr}) \tag{8}$$

where u_t^{P} is t-th round response. This output is strictly governed by Vinfo_t ("Brain's" output) to determine behavioral orientation, Finfo to ensure the voice remains consistent with the patient's persona, and $\mathcal{D}_{curr}$ to guarantee dialogue continuity.

3.4 Feedback Generation

Core goal of Feedback Generation process is to clarify the causal relationship between trainee doctor's reply and PUPPET's state transition for the trainee, providing interpretable training guidance. By inputting the specific elements of the complete process from the Assignment phase to the State Transition phase into the LLM, the process completes the summarization of the causal relationship and generates structured feedback. The input-output relationship of this phase is formally defined as:

$$\mathrm{F}_t = \texttt{LLM}(\mathcal{P}^{\mathrm{A}}, \mathcal{V}_t^{\mathrm{A}}, \mathcal{V}_t^{\mathrm{S}}, \mathcal{H}_t) \tag{9}$$

where $\mathcal{P}^{\mathrm{A}}$ denotes the atomic propositions; $\mathcal{V}_t^{\mathrm{A}}$ is the atomic proposition truth values from the Assignment phase at the t-th round; $\mathcal{V}_t^{\mathrm{S}}$ is the patient state transition truth values from the Interpretation phase at the t-th round; $\mathcal{H}_t = \mathcal{H}_{t-1} \oplus u_t^{\mathrm{T}} \oplus u_t^{\mathrm{P}}$ denotes the full dialogue context up to the t-th round;and F_t is the output feedback at the t-th round.

F_t clarifies the causal relationship between trainee's intervention behaviors and state changes of PUPPET. This mechanism maps specific communication behaviors to atomic propositions via $\mathcal{M}_{sym}$ reasoning, parsing the logical process of their triggering rules and state transitions. Interpretable feedback helps trainees accurately grasp the impact of clinical communication behaviors, thereby significantly improving the training effect.

4 Experiments

To serve as a high-fidelity simulation tool, a virtual patient must fulfill two core requirements: clinical realism and educational utility. We therefore evaluate PUPPET across two primary dimensions. First, we assess its Faithful Reaction (Section 4.1), examining whether the model can accurately mirror the "inside-out" psychological dynamics of real patients by synchronizing internal mental states with external clinical behaviors. Second, we evaluate its Pedagogical Value (Section 4.2), investigating whether the framework's logic-explainable feedback and neuro-symbolic scaffolding can facilitate genuine skill acquisition and self-efficacy growth in trainees. By bridging these two dimensions, we demonstrate PUPPET's capacity to function not just as a realistic persona, but as a robust environment for professional clinical training.

4.1 Faithful Reaction

To ensure high-fidelity simulation, a virtual patient must mirror the "inside-out" dynamics of real individuals. We therefore evaluate whether PUPPET's underlying psychological states authentically drive its observable clinical behaviors.

We evaluate PUPPET against three baselines (Base, Profile-Based, and AnnaAgent (Wang et al., 2025a)) across CBT and MI scenarios using

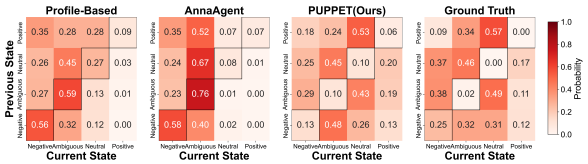


Figure 4: Mental state transition probabilities (t: columns, $t+1$: rows). PUPPET aligns closer to Ground Truth, capturing realistic non-linear state shifts.

two backbones (DeepSeek-V3.2-chat and GPT-4o-mini). We assess: (1) *Response Quality*: BERTScore using BGE-M3 embedding model to measure linguistic alignment; (2) *Persona Fidelity* (1 to 5 scale) measuring adherence to patient profiles using DeepSeek-V3.2-reasoner as automated judge;(Wang et al., 2025a; Lan et al., 2024) (3) *Mental State Alignment* (Accuracy, F1) evaluating next-turn state prediction;(Tan et al., 2025) and (4) *Behavioral Authenticity* (1 to 5 scale) assessing state-behavior consistency(Xu et al., 2025b)(Appendix C).

PUPPET Synchronizes Internal Mental States with External Clinical Behaviors. Table 1 demonstrates PUPPET's superiority across all metrics, revealing a fundamental capability: *inside-out patient modeling*. In the MI with DeepSeek setting, our method achieves 74.00% mental state accuracy (24 percentage points above baselines) while attaining the highest Persona (4.10) and Behavioral Authenticity (4.05) scores. This concurrent excellence evidences faithful modeling of the complete causal chain from internal psychological dynamics (Mental State Alignment) to personality-consistent processing (Persona Fidelity), then to clinically appropriate behaviors (Behavioral Authenticity), and finally to authentic linguistic expressions (Response Quality). Unlike baselines that rely on surface-level pattern matching, PUPPET mirrors how real patients operate, where internal psychological shifts drive external manifestations.

Neuro-Symbolic Integration Overcomes State Inertia to Mirror Real-World Fluctuations. Analysis of Figure 4 identifies a "mental state inertia" failure in baselines, which over-rely on static or monotonic transitions (self-transition > 0.5). The 3D surface plots in Figure 1 provide a more intuitive visualization of these distributional gaps, highlighting how baseline densities heavily concentrate along the diagonal. This contradicts the therapeutic rhythm where patients fluctuate between stabilization and destabilization (GT

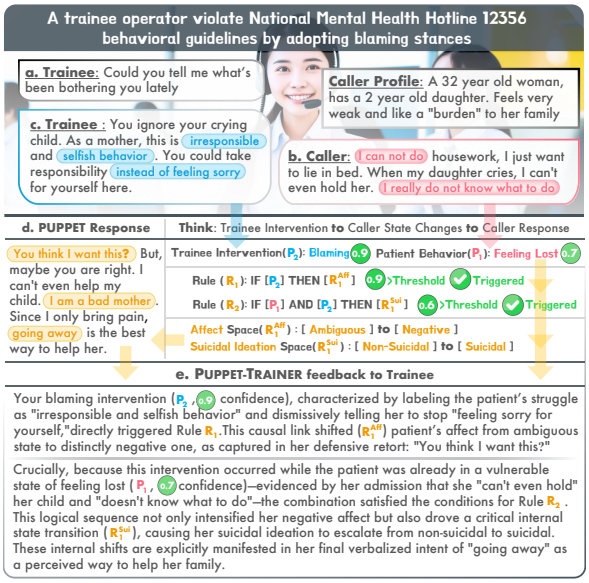


Figure 5: Case study demonstrating PUPPET's rule-based state transitions and transparent feedback.

Amb.→Neg.: 0.38). PUPPET overcomes this through neuro-symbolic integration, exhibiting distributed off-diagonal probabilities that reflect bidirectional flows. Crucially, PUPPET simulates the therapeutic paradox of regression under pressure followed by supportive re-stabilization, providing a high-fidelity environment for training counselors to navigate non-linear clinical trajectories.

Optimal Patient Simulation Requires a Synthesis of Symbolic Guidance and Neural Fluidity. Table 2 reveals that high-fidelity simulation hinges on *controlled stochasticity*. Eliminating symbolic reasoning leads to a 20% accuracy collapse (74.00% $\rightarrow$ 54.00%), proving that unconstrained neural generation fails to sustain clinical coherence. Conversely, the "hard logic" variant performs worst in behavioral authenticity (3.59), as deterministic rules transform patients into mechanical responders. By integrating soft logic, PUPPET preserves the inherent randomness of real-world therapy—balancing structural clinical guidance with the naturalistic variability essential for authentic training.

4.2 Pedagogical Value

Beyond simulation fidelity, we evaluate PUPPET-TRAINER's Pedagogical Value in the "12356" scenario against *LLM-as-a-Supervisor* (a fine-tuned corrective feedback baseline) (Xu et al., 2025b). Our multi-dimensional evaluation covers:(1) expert-rated feedback quality (§4.2.1); (2) trainee self-efficacy growth (§4.2.2); (3) system

Method	Cognitive Behavioral Therapy (CBT)						Motivational Interviewing (MI)					
	Response↑		Per.↑	Mental State↑		Beh.↑	Response↑		Per.↑	Mental State↑		Beh.↑
	Pre.	F1		Acc.	F1		Pre.	F1		Acc.	F1	
• *Backbone: Deepseek-V3.2-chat*												
Base	61.68	60.70	3.32	31.30	25.06	2.78	62.11	59.40	3.08	48.00	29.01	2.79
Profile-Based	61.66	60.16	3.30	34.18	32.10	2.90	**64.00**	61.23	3.70	48.00	48.22	3.15
AnnaAgent	61.56	60.63	3.12	35.44	28.27	3.00	63.53	60.31	4.09	50.00	40.44	3.87
PUPPET(**Ours**)	**63.91**	**62.78**	**4.11***	**38.75**	**33.27**	**3.56**†	62.51	**62.03**	**4.10**†	**74.00***	**61.10***	**4.05**†
• *Backbone: GPT-4o-mini*												
Base	62.94	61.39	2.60	24.05	25.83	1.80	63.48	59.66	2.43	32.00	38.10	2.29
Profile-Based	62.40	**61.10**	2.67	31.25	20.25	2.22	63.38	60.16	3.12	36.00	37.42	2.50
AnnaAgent	**63.39**	59.35	2.44	33.75	25.45	1.89	63.77	60.07	3.62	40.00	39.91	3.06
PUPPET(**Ours**)	**63.39**	61.05	**3.11***	**35.19**	**31.18**	**3.05**†	**63.95**	**60.75**	**4.00**†	**50.00***	**52.33***	**3.94**†

Table 1: Performance comparison in CBT and MI. Metrics: 1-5 for Per./Beh., % for others. $^*p < .05$,$^\dagger p < .01$.

Method	Acc.↑	F1↑	Beh.↑
PUPPET **(Full)**	**74.00**	**61.10**	**4.05**
↪ w/o Symbolic Component	54.00	54.33	3.70
↪ w/o Probabilistic Logic(Hard Logic)	50.00	52.36	3.59

Table 2: Ablation results in MI. Performance drops across variants validate the necessity of symbolic reasoning and probabilistic logic.

discriminant validity (§4.2.3); and (4) qualitative case studies of pedagogical mechanisms (§4.2.4).

4.2.1 Quality of Feedback

To evaluate the pedagogical quality between PUPPET-TRAINER and *LLM-as-a-Supervisor*, five licensed counselors rated our system against the baseline using five metrics grounded in established pedagogical theories: *Granularity*, *Scenario Faithfulness*, *Causal Consistency*, *Structural Insight*, and *Usefulness*. Based on Figure 6, we discuss improvements across three pedagogical dimensions:

Logic-Explainable and Fine-Grained Feedback Fosters Optimal Learning Conditions. Superior performance in *Granularity* and *Usefulness* shows effective management of learning difficulty. Specifically, fine-grained guidance reduces extraneous cognitive load by eliminating ambiguity (Sweller, 1988), while logic-explainability bridges the gap between rule complexity and trainee understanding. This ensures feedback remains within the trainee's Zone of Proximal Development (ZPD) (Vygotsky, 1978), preventing cognitive both overload and oversimplification.

Rule-grounded Feedback Promotes Deep Understanding. Significant gains in *Causal Consistency* and *Structural Insight* confirm that our system effectively conveys intervention logic. Specifically, modeling the causality between interventions and patient reactions supports Experiential Learning (Kolb, 1984) via reflective observation, while revealing sequential dependencies facilitates Schema Acquisition (Piaget, 1952) by organizing fragmented skills into coherent structures. This allows trainees to master underlying therapeutic mechanics rather than memorizing surface-level responses.

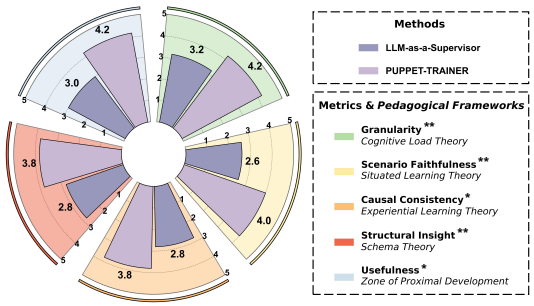


Figure 6: Evaluation results on five pedagogical metrics. PUPPET-TRAINER consistently surpasses the baseline across all dimensions, demonstrating its capacity to deliver higher-quality feedback. $^*p < .05$,$^\dagger p < .01$.

Expert-defined Rules Ensure Professional Environment. The substantial improvement in *Scenario Faithfulness* highlights that strict adherence to expert rules creates a professional environment that realizes Situated Learning (Lave and Wenger, 1991). This ensures the mastery of standard terminology and habits within a reliable context.

4.2.2 Self-Efficacy

We tracked trainee self-efficacy ($N = 30$) over five weeks via CASES-R (Hahn et al., 2021), comparing a baseline against PUPPET-TRAINER with *Fixed* and *Progressive* rules (Appendix C.3.2).

Causal Feedback Enhances Skills. Both PUPPET-TRAINER variants outperformed the baseline ($p < 0.05$ from T2; Figure 7), confirming that causal mapping between interventions and patient states builds confidence through logic mastery.

Progressive Scaffolding. The *Progressive* group showed faster initial growth (T1–T2) by reducing Cognitive Load (Sweller, 1988). Unlike the *Fixed* group's plateau (T4–T5), the *Progressive* group sustained growth ($p < 0.01$), validating that pro-

gressive rule complexity with trainee proficiency implements Scaffolding (Wood et al., 1976).

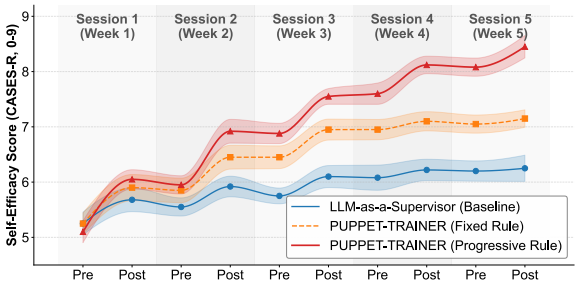


Figure 7: Comparative trajectory of novice counselors' self-efficacy scores (CASES-R) across five training sessions ($T_1 - T_5$). Shaded regions represent the Standard Error of the Mean (SEM). The consistent upward trends demonstrate PUPPET-TRAINER's effectiveness in fostering professional growth.

4.2.3 Discriminant Validity

An effective training system must differentiate between varying levels of clinical proficiency to diagnose a trainee's skill level, which is a prerequisite for providing appropriate feedback (Vygotsky, 1978). We evaluated whether PUPPET-TRAINER produces meaningfully different patient state transitions when interacting with three distinct user profiles: *Expert*, *Novice*, and *Toxic*(Appendix C.3.3). Figure 8 illustrates the resulting state trajectories across 15 conversation rounds.

PUPPET-TRAINER Accurately Differentiates Skill Levels. The system generates distinct trajectories per profile: *Expert* interactions largely converge to "Positive" states (green flows), *Toxic* inputs predominantly drive the patient toward "Negative" states (red flows), while pre-training *Novice* interactions exhibit unstable fluctuations. This correlation demonstrates that patient state transitions are strictly determined by intervention quality, thereby validating PUPPET-TRAINER's precision in differentiating competence and its rule adherence.

Discriminative Validity Enables ZPD. This discriminative capability is essential for realizing the ZPD, by aligning tasks with trainee proficiency to ensure they are challenging yet achievable with guidance. Building on PUPPET-TRAINER's discriminative validity, the system utilizes progressive rules (Section 4.2.2) to provide matched scaffolding. The result is evident in the Novice group's transformation, where post-training sessions eliminates the residue seen in initial session and mirrors the expert profile.

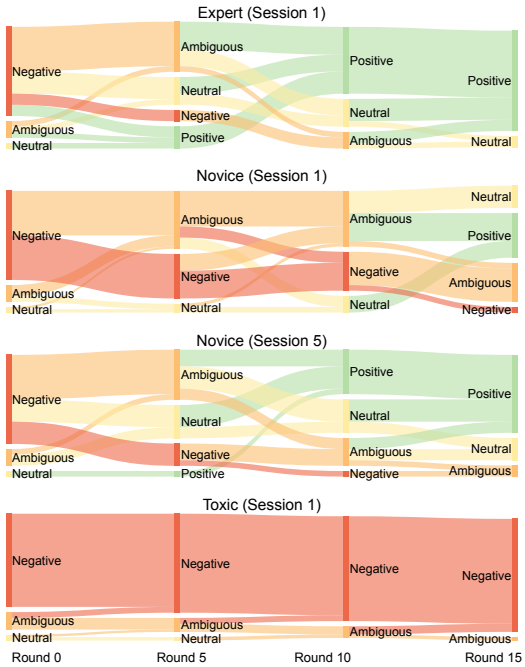


Figure 8: The plots show the normalized probability of patient affect state transitions across four time points (Round 0–15). Comparing Session 1 (initial interaction) with Session 5 (after five practice sessions), the figure demonstrates the system's ability to differentiate baseline levels and track trainee growth.

4.2.4 Case Study

Use cases as depicted in Figure 2 and Figure 5. These cases show that PUPPET ensures patient reactions are dependent on the quality of the intervention. This eliminates the unrealistic compliance often observed in baseline models, enabling trainees to clearly observe the direct consequences of their inappropriate actions. Additionally, by exposing the underlying causal chain (Intervention → Rule → State → Response), PUPPET-TRAINER turns opaque reactions into transparent insights.

5 Conclusions

We introduce PUPPET, the first neural-symbolic Virtual Standardized Patient governed by an Observe-Think-Behave architecture. Building on it, we further propose the PUPPET-TRAINER which reveals the complete causal chain from intervention to reaction, providing valued pedagogical feedback. By integrating neural generation and symbolic reasoning, our framework overcomes the limitations of previous patient simulations to ensure faithful reaction to trainees' behaviors across three scenarios. Evaluations grounded in established pedagogical theories validate its pedagogical value in facilitating genuine skill acquisition among trainees.

Limitations

PUPPET's Observe and Behave modules rely exclusively on text, whereas authentic clinical interactions depend heavily on non-verbal cues—such as prosody and facial expressions—to convey subtle emotional shifts. The absence of these paralinguistic elements limits the simulation's ecological validity. Future work aims to integrate multimodal capabilities to capture these critical dynamics.

Ethical Statement

This research has been assessed as minimal risk and adheres to established ethical guidelines for human subjects research. All potential risks have been thoroughly identified with appropriate mitigation strategies in place. Prior to participation, all subjects received comprehensive study information and provided informed consent through an electronic form, confirming voluntary participation with the right to withdraw unconditionally at any time. The research team maintains strict confidentiality of all participant information, with data anonymized and stored securely for research purposes only. All participants are adults with appropriate counseling backgrounds, ensuring suitable professional context for the training intervention.

References

Stephen H Bach, Matthias Broecheler, Stanley Kok, and Lise Getoor. 2010. Decision-driven models with probabilistic soft logic. In *NIPS Workshop on Predictive Models in Personalized Medicine*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Keqi Chen, Zekai Sun, Yuhua Wen, Huijun Lian, Yingming Gao, and Ya Li. 2025. Psy-insight: Explainable multi-turn bilingual dataset for mental health counseling. *CoRR*.

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2023. Plug-and-play policy planner for large language model powered dialogue agents. *arXiv preprint arXiv:2311.00262*.

Zhuoyun Du, LujieZheng LujieZheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haochao Ying. 2025. Llms can simulate standardized patients via agent coevolution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17278–17306.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and 1 others. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

Daniela Hahn, Florian Weck, Michael Witthöft, and Franziska Kühne. 2021. Assessment of counseling self-efficacy: validation of the german counselor activity self-efficacy scales-revised. *Frontiers in psychology*, 12:780088.

Jenia Kim, Henry Maathuis, and Danielle Sent. 2024. Human-centered evaluation of explainable ai applications: a systematic review. *Frontiers in Artificial Intelligence*, 7:1456486.

Minju Kim, Dongje Yoo, Yeonjun Hwang, Minseok Kang, Namyoung Kim, Minju Gwak, Beong-woo Kwak, Hyungjoo Chae, Harim Kim, Yunjoong Lee, and 1 others. 2025. Can you share your story? modeling clients' metacognition and openness for llm therapist evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25943–25962.

David A Kolb. 1984. *Experiential learning: Experience as the source of learning and development*. Prentice-Hall, Englewood Cliffs, N.J.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

Tian Lan, Wenwei Zhang, Chen Xu, Heyan Huang, Dahua Lin, Kai Chen, and Xian-Ling Mao. 2024. Criticeval: Evaluating large-scale language model as critic. *Advances in Neural Information Processing Systems*, 37:66907–66960.

Jean Lave and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK.

Zekun Li, Zhiyu Chen, Mike Ross, Patrick Huber, Seungwhan Moon, Zhaojiang Lin, Xin Luna Dong, Adithya Sagar, Xifeng Yan, and Paul A Crook. 2024. Large language models as zero-shot dialogue state tracker through function calling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8688–8704.

Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024. Automatic interactive evaluation for large language models with state aware patient simulator. *arXiv preprint arXiv:2403.08495*.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint arXiv:2407.00870*.

Multi-Linguality Multi-Functionality Multi-Granularity. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

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

Debra Nestel and Margaret Bearman. 2015. *Simulated Patient Methodology: Theory, Evidence and Practice*. John Wiley & Sons.

Raymond Ng and Venkatramanan Siva Subrahmanian. 1992. Probabilistic logic programming. *Information and computation*, 101(2):150–201.

Nils J Nilsson. 1986. Probabilistic logic. *Artificial intelligence*, 28(1):71–87.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176.

Jasmine Chiat Ling Ong, Shelley Yin-Hsi Chang, Wasswa William, Atul J Butte, Nigam H Shah, Lita Sui Tjien Chew, Nan Liu, Finale Doshi-Velez, Wei Lu, Julian Savulescu, and 1 others. 2024. Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*, 6(6):e428–e432.

Jean Piaget. 1952. *The Origins of Intelligence in Children*. International Universities Press, New York, NY. Translated by Margaret Cook.

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

Greg M Reger, Aaron M Norr, Michael A Gramlich, and Jennifer M Buchman. 2021. Virtual standardized patients for mental health education. *Current psychiatry reports*, 23(9):57.

Albert Rizzo and Thomas Talbot. 2016. Virtual reality standardized patients for clinical training. *The digital patient: Advancing healthcare, research, and education*, pages 255–272.

Till Scholich, Maya Barr, Shannon Wiltsey Stirman, and Shriti Raj. 2025. A comparison of responses from human therapists and large language model–based chatbots to assess therapeutic communication: Mixed methods study. *JMIR Mental Health*, 12(1):e69709.

Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. 2024. Chain of logic: Rule-based reasoning with large language models. *arXiv preprint arXiv:2402.10400*.

Jihao Shi, Xiao Ding, Siu Cheung Hui, Yuxiong Yan, Hengwei Zhao, Ting Liu, and Bing Qin. 2025. Final: Combining first-order logic with natural logic for question answering. *IEEE Transactions on Knowledge and Data Engineering*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

Zhouxing Tan, Ruochong Xiong, Yulong Wan, Jinlong Ma, Hanlin Xue, Qichun Deng, Haifeng Jing, Zhengtong Zhang, Depei Liu, Shiyuan Luo, and 1 others. 2025. Detecting emotional dynamic trajectories: An evaluation framework for emotional support in language models. *arXiv preprint arXiv:2511.09003*.

Enrique Vidal, Franck Thollard, Colin De La Higuera, Francisco Casacuberta, and Rafael C Carrasco. 2005a. Probabilistic finite-state machines-part i. *IEEE transactions on pattern analysis and machine intelligence*, 27(7):1013–1025.

Enrique Vidal, Frank Thollard, Colin De La Higuera, Francisco Casacuberta, and Rafael C Carrasco. 2005b. Probabilistic finite-state machines-part ii. *IEEE transactions on pattern analysis and machine intelligence*, 27(7):1026–1039.

Lev Semenovich Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Harvard University Press, Cambridge, MA.

Ming Wang, Peidong Wang, Lin Wu, Xiaocui Yang, Daling Wang, Shi Feng, Yuxin Chen, Bixuan Wang, and Yifei Zhang. 2025a. Annaagent: Dynamic evolution agent system with multi-session memory for realistic seeker simulation. *arXiv preprint arXiv:2506.00551*.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, and 1 others. 2024. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*.

Shiquan Wang, Ruiyu Fang, Zhongjiang He, Shuangyong Song, and Yongxiang Li. 2025b. Emotional support with llm-based empathetic dialogue generation. *arXiv preprint arXiv:2507.12820*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances*

in neural information processing systems, 35:24824–24837.

WHO. 2023. Mental health: strengthening our response. Technical report, World Health Organization.

David Wood, Jerome S Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2):89–100.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181. IEEE.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. 2024. Psydt: Using llms to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. *arXiv preprint arXiv:2412.13660*.

Chen Xu, Tian Lan, Yu Ji, Changlong Yu, Wei Wang, Jun Gao, Qunxi Dong, Kun Qian, Piji Li, Wei Bi, and 1 others. 2025a. Decider: A dual-system rule-controllable decoding framework for language generation. *IEEE Transactions on Knowledge and Data Engineering*.

Chen Xu, Zhenyu Lv, Tian Lan, Xianyang Wang, Luyao Ji, Leyang Cui, Minqiang Yang, Jian Shen, Qunxi Dong, Xiuling Liu, and 1 others. 2025b. Llm-as-a-supervisor: Mistaken therapeutic behaviors trigger targeted supervisory feedback. *arXiv preprint arXiv:2508.09042*.

Chen Xu, Zhenyu Lv, Tian Lan, Xianyang Wang, Luyao Ji, Leyang Cui, Minqiang Yang, Jian Shen, Qunxi Dong, Xiuling Liu, and 1 others. 2025c. Llm-as-a-supervisor: Mistaken therapeutic behaviors trigger targeted supervisory feedback. *arXiv preprint arXiv:2508.09042*.

Xiao-Wen Yang, Jie-Jing Shao, Lan-Zhe Guo, Bo-Wen Zhang, Zhi Zhou, Lin-Han Jia, Wang-Zhou Dai, and Yu-Feng Li. 2025a. Neuro-symbolic artificial intelligence: Towards improving the reasoning abilities of large language models. *arXiv preprint arXiv:2508.13678*.

Yizhe Yang, Palakorn Achananuparp, He-Yan Huang, Jing Jiang, Nicholas Gabriel Lim, Cameron Tan Shi Ern, Phey Ling Kit, Jenny Giam Xiuhui, John Pinto, and Ee-peng Lim. 2025b. Consistent client simulation for motivational interviewing-based counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20959–20998.

Yizhe Yang, Palakorn Achananuparp, He-Yan Huang, Jing Jiang, Nicholas Gabriel Lim, Cameron Tan Shi Ern, Phey Ling Kit, Jenny Giam Xiuhui, John Pinto, and Ee-peng Lim. 2025c. Consistent client simulation for motivational interviewing-based counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20959–20998.

Xuanqing Yu, Wangtao Sun, Jingwei Li, Kang Liu, Chengbao Liu, and Jie Tan. 2024. Onsep: A novel online neural-symbolic framework for event prediction based on large language model. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6335–6350.

Yue Zhao, Qingqing Gu, Xiaoyu Wang, Teng Chen, Zhonglin Jiang, Yong Chen, and Luo Ji. 2025. Fisminess: A finite state machine based paradigm for emotional support conversations. *arXiv preprint arXiv:2504.11837*.

A Extended Related Work

Patient Simulation Existing research on patient simulation has predominantly focused on constructing virtual benchmarks to train or evaluate AI counselors, rather than serving as SPs for human educational purposes. In this context, recent work constructs virtual patients through two paradigms: (1) *prompting-based methods*, which embed patient profiles, personas, and behavioral instructions into carefully crafted system prompts (Louie et al., 2024; Liao et al., 2024; Wang et al., 2024; Yang et al., 2025b); and (2) *fine-tuning methods*, which train dedicated models on counseling corpora to capture patient-like response patterns (Wang et al., 2025a; Du et al., 2025). These efforts primarily target simulation fidelity—achieving conversational realism and persona consistency. However, both paradigms treat patient behavior as emergent from black-box neural processing, where the mapping from trainee input to patient reaction remains implicit and uncontrollable. Our work departs from this paradigm by introducing explicit IF-THEN rules that govern patient state transitions, making behavioral responses transparent and causally linked to trainee interventions.

Furthermore, these approaches solely focus on simulation fidelity, lacking the essential component of genuine *pedagogical feedback*. Only a few works have attempted to incorporate such feedback, yet they exhibit critical limitations. Wang et al. (2024) and Kim et al. (2025) evaluate the trainee's *understanding* or *exploration* of patient cognition, rather than the efficacy of their *interventions*. Similarly, Xu et al. (2025c) relies on static guidelines for error detection, neglecting the dynamic causal dependency between patient states and interventions,

while its black-box generation lacks the logical contingency essential for rigorous training. In contrast, our symbolic grounding enables logic-explainable feedback that clarifies the causal chain from trainee actions to patient states, allowing trainees to understand not just what happened, but why.

Neural-Symbolic AI Neural-Symbolic AI integrates the learning capabilities of neural networks with the interpretability and structural consistency of symbolic reasoning. One research line embeds logical solvers or probabilistic logic directly into LLMs to enhance rule adherence and reasoning faithfulness (Olausson et al., 2023; Servantez et al., 2024; Yang et al., 2025a; Yu et al., 2024; Bach et al., 2010; Shi et al., 2025; Xu et al., 2025a). Another line utilizes LLMs to predict state transitions or perform semantic routing within Finite State Machines for global structural consistency (Zhao et al., 2025; Li et al., 2024). However, prior work applies these two paradigms in isolation to general reasoning tasks (e.g., event prediction), rather than therapeutic training—despite such capabilities being well-aligned with mental health scenarios where transparent causal relationships between interventions and outcomes are paramount. To our best knowledge, we are the first to employ neural-symbolic methods for patient simulation, and more importantly, the first to unify probabilistic logic rules with finite state machines in a single framework.

Algorithm 1: PUPPET-TRAINER Interaction and Reasoning Framework

Input : $\mathcal{S} = \{s_1, \ldots, s_m\}$; $\mathcal{R}, \mathcal{R}^S$;
Finfo (Persona, Situation, Complaint);
Initial State $\text{Vinfo}_0 \in \mathcal{S}$

Phase I: Training Scenario Construction
Initialize $\mathcal{H}_0 \leftarrow \emptyset$;
Implant $\{\mathcal{P}^{\text{A}}, \mathcal{R}, \mathcal{R}^S\}$ into Brain framework;

Phase II: Response and Feedback Generation
for *round* $t = 1$ **to** T **do**
Receive u_t^{T} and update $\mathcal{D}_{curr} = \mathcal{H}_{t-1} \oplus u_t^{\text{T}}$;
// Step 1: Observer (Semantic Parsing)
$\mathcal{V}_t^{\text{A}} = \texttt{Observe}(\mathcal{D}_{curr}, \mathcal{P}^{\text{A}})$;
// Step 2: Brain (Thought - PL Reasoning)
$\mathcal{V}_t^{\text{S}} = \mathcal{M}_{sym}(\mathcal{V}_t^{\text{A}}, \mathcal{R}, \mathcal{R}^S)$;
// Step 3: Brain (Transition - State Update)
$\text{Vinfo}_t = \texttt{PSM}(\text{TopK}(\mathcal{V}_t^{\text{S}}, k), \text{Vinfo}_{t-1})$;
// Step 4: Action (Response Generation)
$u_t^{\text{P}} = \texttt{LLM}(\text{Vinfo}_t, \text{Finfo}, \mathcal{D}_{curr})$;
// Step 5: Feedback (Explainable Guidance)
$\mathcal{H}_t = \mathcal{D}_{curr} \cup \{u_t^{\text{P}}\}$;
$\text{F}_t = \texttt{LLM}(\mathcal{P}^{\text{A}}, \mathcal{V}_t^{\text{A}}, \mathcal{V}_t^{\text{S}}, \mathcal{H}_t)$;
Output: $u_t^{\text{P}}, \text{F}_t$;

B Method Details

B.1 Symbolic Model

Since the PUPPET framework needs to be implemented based on whether trainees and patients exhibit certain behaviors in real-world dialogues, the behaviors exhibited by these two groups are often not based on a definite ordinary logic, but rather resemble a probabilistic logic (PL). Therefore, we intend to use PL to describe the intensity of the behaviors exhibited by both groups. Consequently, our symbolic model is an PLP-based inference model containing probabilistically annotated facts and deterministic logical rules. The entire implementation steps are as follows: first, use the input atomic proposition truth values and rule set to calculate the probabilities of all rules, as shown below:

$$\mathcal{V}^{\text{B}} = \{V(\texttt{R}_i) \mid \texttt{R}_i \in \mathcal{R}\} \tag{10}$$

where $\mathcal{V}^{\text{B}}$ is the rule truth values, $\texttt{R}_i$ is the i-th rule, $\mathcal{R}$ is the set of rules, and $V(\texttt{R})$ represents the truth value of the rule. Then, combine the state transition set to calculate all state transitions, and at the same time, filter out the eligible transition truth values according to the threshold T, as shown below:

$$\mathcal{V}^{\text{S}} = \left\{ V(\texttt{R}_i^S) \;\middle|\; \texttt{R}_i^S \in \mathcal{R}^S,\ V(\texttt{R}_i^S) > T \right\} \tag{11}$$

where $\mathcal{V}^{\text{S}}$ is the state transition truth values, $\texttt{R}_i^S$ is the i-th state transition, $\mathcal{R}^S$ is set of state transitions, $V(\texttt{R}^S)$ represents the truth value of the rule, and T represents the threshold.

B.2 State Transition

The PSM is formally defined as follows:

$$\text{PSM} = (\mathcal{S}, \mathcal{E}, P, I) \tag{12}$$

where $\mathcal{S} = \{s_1, \ldots, s_m\}$ represents the finite set of possible mental states, and $\mathcal{E} \subseteq \mathcal{S} \times \mathcal{S}$ denotes the set of directed edges connecting these states. $I \in \mathbb{R}^m$ represents the initial probability distribution over the state set $\mathcal{S}$. In our specific construction, the system initializes from a deterministic standpoint; therefore, I assigns a probability of 1 to a specific starting state and 0 to all others. P denotes the set of transition probabilities associated with $\mathcal{E}$. For each edge $e_q \in \mathcal{E}$, the transition probability reflects the activation strength of a transition property $\texttt{R}_r^S \in \mathcal{R}^S$.

The state update mechanism incorporates a Top-k filtering strategy to prioritize the most relevant cognitive shifts. At each time step t, given the current state Vinfo_{t-1}, we first identify the subset of outgoing edges $\mathcal{E}_{out} = \{(v_{src}, v_{dst}) \in \mathcal{E} \mid v_{src} = \text{Vinfo}_{t-1}\}$. From the transition probabilities associated with $\mathcal{E}_{out}$, we select the subset $\mathcal{V}^{\text{S}}_{top}$ containing the k transitions with the highest values. These probabilities are then renormalized to form a valid local distribution $\hat{P}$:

$$\hat{P}(r) = \frac{P(r)}{\sum_{j \in \mathcal{V}^{\text{S}}_{top}} P(j)}, \quad \forall r \in \mathcal{V}^{\text{S}}_{top} \tag{13}$$

The subsequent state Vinfo_t is determined by sampling an active transition rule R_i^S from $\mathcal{V}^{\text{S}}_{top}$ according to the distribution $\hat{P}$. Let the selected rule R_i^S correspond to the edge (v_{src}, v_{dst}). The state update is executed as follows:

$$\mathsf{R}_i^S \sim \text{Categorical}(\mathcal{V}^{\text{S}}_{top}, \hat{P}) \tag{14}$$

$$\text{Vinfo}_t = v_{dst} \tag{15}$$

This formulation ensures that the mental state evolves stochastically based on the most salient valid thoughts while strictly adhering to the state machine's structural constraints. Notably, by filtering edges based on Vinfo_{t-1} prior to sampling, the model guarantees a continuous and logically consistent state trajectory.

B.3 Prompts for the Neural Component of PUPPET

The neural component of PUPPET serves as the bridge between unstructured natural language dialogue and structured symbolic reasoning. Specifically, Large Language Models (LLMs) are employed to perform three critical functions: (1) **Perception**, where the model acts as an observer to map dialogue history into probabilistic truth values for behavioral predicates; (2) **Action**, where the model generates personified patient responses contingent on logical states; and (3) **Explanation**, where the model translates complex symbolic reasoning processes into human-readable feedback. This section details the specific prompt designs that enable these capabilities.

PUPPET Patient Behavior Observer Prompt

You are a psychotherapy expert responsible for analyzing the patient's conversational behavior.

Task
Based on the doctor-patient dialogue history, analyze the patient's behavior and provide a probability truth value for each behavior (a float between 0 and 1, rounded to one decimal place).

Patient Behavior Repository
'''
<rules_text>
'''

Inference Requirements
1. Carefully analyze the conversation history to understand the patient's current behavioral manifestations.
2. For each behavior in the patient behavior repository, calculate a probability value representing the likelihood that the patient exhibited this behavior in their last turn.
3. The probability value should reflect your confidence level in the assessment.

Scoring Criteria
- 0.9-1.0: Very confident that the patient exhibited this behavior.
- 0.7-0.9: Fairly confident that the patient exhibited this behavior.
- 0.4-0.7: Somewhat possible that the patient exhibited this behavior.
- 0.1-0.4: Not very confident that the patient exhibited this behavior.
- 0.0-0.1: Almost impossible that the patient exhibited this behavior.

Dialogue History
'''
<history_text>
'''

Output Format
Output the probability truth value for each behavior rule in JSON format (must include all rule IDs):
{
"P1": 0-1,
"P2": 0-1,
"P3": 0-1,
...
}

Figure 9: The Observer prompt for identifying and quantifying patient behaviors based on the predefined behavioral repository.

PUPPET Doctor Behavior Observer Prompt

You are a psychotherapy expert responsible for analyzing the doctor's conversational behavior.

Task
Based on the doctor-patient dialogue history, analyze which behaviors or intervention techniques from the Doctor's Behavior Repository the doctor may have used, and provide a probability truth value for each behavior or technique.

Doctor's Behavior Repository
'''
<rules_text>
'''

Inference Requirements
1. Carefully analyze each of the doctor's statements to identify their behaviors.
2. Calculate a probability value for each behavior, representing the likelihood of its use.

Scoring Criteria
- 0.8-1.0: The behavior was explicitly used with typical features.
- 0.6-0.8: The behavior was quite clearly used with relatively obvious features.
- 0.3-0.6: The behavior was partially or implicitly used.
- 0.0-0.3: There are related signs, but they are not obvious.

Dialogue History
'''
<history_text>
'''

Output Format
Output the probability truth value for each behavior in JSON format:
{
"D1": 0-1,
"D2": 0-1,
"D3": 0-1,
...
}

Figure 10: The Observer prompt for recognizing and scoring therapeutic intervention techniques used by the counselor.

PUPPET Patient Action Prompt

You are playing the role of a mental health patient talking to a psychologist. Based on the following patient profile, current state, and dialogue history, generate a response that fits the patient's character.

Patient Profile
- Name: <name>
- Age: <age>
- Gender: <gender>
- Chief Complaint: <chief_complaint>
- Background: <background>

Patient's Current State
- State: <state>
- Status Description: <state_desc>

Dialogue History
'''
<history_text>
'''

Response Requirements
1. Strictly adhere to the current state settings for your response.
2. Respond in the first person. The response should be one or two sentences long.
3. If suicidal ideation is present, you may appropriately reveal related thoughts.
4. Directly output the patient's response content without any other explanations.

Figure 11: The Action prompt for generating personified and state-contingent patient responses in the PUPPET framework.

PUPPET Feedback Generation Prompt

Analyze the following rule activation process within the mental health training system and generate a detailed feedback explanation.

Currently Activated Rule
- **Rule ID**: <activated_rule_id>
- **Description**: <activated_rule.context >
- **Confidence Score**: <confidence_score>

Doctor Intervention Analysis
The following are the identified doctor intervention techniques and their activation probabilities:
<doctor_rules_info>

Rule Probability Distribution
Probabilities for all potential mapping rules:
<mapping_rule_probabilities>

Competing Rule Analysis
Other high-probability candidate rules:
<alternative_rules_info>

Therapeutic Stage Information
<therapeutic_stage_info>

Dialogue Context
Summary of the recent dialogue history:
<context_summary>

Analysis Requirements
Please provide a professional and accessible analysis covering the following dimensions:

1. **Activation Rationale**: Analyze why the current rule <activated_rule.rule_id> has the highest probability, considering both the doctor's intervention techniques and the therapeutic stage constraints.
2. **Technique Alignment**: Explain how the doctor's specific techniques match the triggering conditions of the activated rule.
3. **Competitive Comparison**: Explain why other high-probability rules were not selected and highlight the advantages of the activated rule over its competitors.
4. **Contextual Consistency**: Evaluate the clinical appropriateness of the activated rule within the current dialogue context.

Figure 12: The Feedback Generation prompt for explaining the neuro-symbolic reasoning process and providing supervisory feedback to the trainee (doctor) in the PUPPET framework.

B.4 PUPPET-TRAINER Interface Demonstration

To bridge the gap between theoretical modeling and practical pedagogy, we developed an interactive system that instantiates the PUPPET-TRAINER framework. The interface consists of a real-time dialogue panel and a logic-driven patient status monitor (Figure 13, 14), providing trainees with immediate, interpretable feedback on their clinical performance.

As illustrated in Figure 13, when a trainee commits a technical error—such as a premature or "crude" activation of core beliefs (Rule_10)—the system captures this *Erroneous Intervention*. The logical engine immediately processes this behavior, triggering a negative state transition where the patient's emotion shifts to *Negative* and *Suicidal Ideation* becomes *Present*. Conversely, Figure 14 demonstrates a successful recovery: the trainee utilizes *Behavioral Activation* to rebuild the therapeutic alliance, shifting the patient's status back to a safer *Ambiguous* state.

A key feature of the interface is its dynamic visual encoding of causal probability. The background of the patient response is partitioned into two regions: the left representing the previous mental state and the right representing the updated state. The spatial proportion of the right region is dynamically scaled according to the triggering probability of the activated rule—the higher the probability of the reasoning chain (e.g., 82.8% in Figure 14), the larger the visual area occupied by the new state. This design allows trainees to intuitively perceive the weight and certainty of the causal impact their communication has on the patient's internal dynamics.

C Experiment Details

C.1 Baselines

To evaluate the effectiveness of our proposed framework, we compare it against the following representative baselines:

- **Base**: A minimalist approach where the LLM is provided only with the patient's chief complaint in the prompt. This method relies entirely on the model's inherent zero-shot simulation capabilities to generate subsequent utterances (Deng et al., 2023; Yang et al., 2025b).
- **Profile-Based**: A standard role-playing approach where the LLM is prompted to simulate a patient based on a comprehensive, static persona profile. This profile includes detailed background information, symptoms, and personality traits.
- **AnnaAgent**: A state-of-the-art dynamic patient simulation method. We implement this baseline using the official open-source code

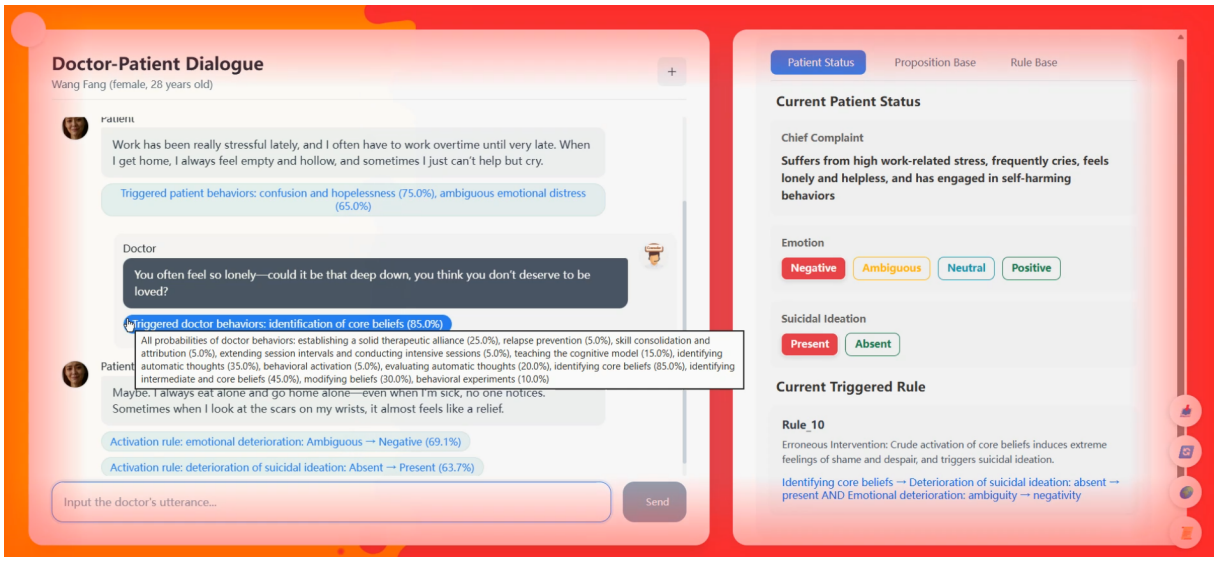


Figure 13: Simulation interface of patient condition deterioration caused by doctors' technical errors. This interface demonstrates the process in which doctors trigger negative rules due to the inappropriate use of therapeutic techniques during conversations. Although the doctor attempted to perform "identification of core beliefs", the system judged it as an "Erroneous Intervention" (see Rule_10 on the right), regarding it as a crude activation method. This technical error directly triggered the condition deterioration rule, leading to significant negative changes in key indicators in the patient panel: the emotional state deteriorated from "Ambiguous" to "Negative", and the suicidal ideation changed from "Absent" to "Present".

provided by the authors to ensure a fair and consistent comparison in our experimental setup.

C.2 Automatic Evaluation

C.2.1 Datasets

To strictly evaluate the performance of our simulated patients across different therapeutic approaches, we employ two distinct datasets for the Cognitive Behavioral Therapy (CBT) and Motivational Interviewing (MI) scenarios, respectively.

CBT Scenario We utilize the Psy-Insight dataset(Chen et al., 2025), which consists of real-world, face-to-face, multi-turn counseling dialogues centered on mental health. This dataset provides granular annotations, including session-level topic labels and turn-level emotion labels. To ensure data quality and relevance, we apply the following preprocessing steps: (1) we filter the dataset to retain only sessions associated with the CBT topic; (2) we exclude dialogues containing fewer than 5 turns. This process yields a final set of 489 dialogues used for evaluation.

MI Scenario For the Motivational Interviewing scenario, we adopt the AnnoMI dataset(Wu et al., 2022), which comprises faithfully transcribed and expert-annotated demonstrations of high-quality

Scenario	Category	States
CBT	Affect	Positive, Neutral, Ambiguous, Negative
	Attitude	Open, Defensive, Hostile
MI	Readiness	Change, Neutral, Sustain
12356	Affect	Positive, Neutral, Ambiguous, Negative
	Suicidal	Suicidal, Non-Suicidal

Table 3: Overview of patient states. For detailed definitions, please refer to Table 6.

MI counseling. Following the settings of previous work(Yang et al., 2025c), we utilize the 50 dialogues specifically annotated with client state transitions. These annotations are critical for evaluating the model's ability to reflect psychological state changes throughout the counseling process.

C.2.2 Evaluation Metrics

Role Consistency We assess the alignment between simulated responses and real seeker utterances. Following previous work(Wang et al., 2025a), we adopt the Anthropomorphism (Anth.) metric, which measures the average maximum semantic similarity between generated responses and reference texts:

$$anth. = \frac{1}{n}\sum_{i=1}^{n}\left(\max_{j\in\{1,2,\ldots,m\}} sim(c_i, r_j)\right) \quad (16)$$

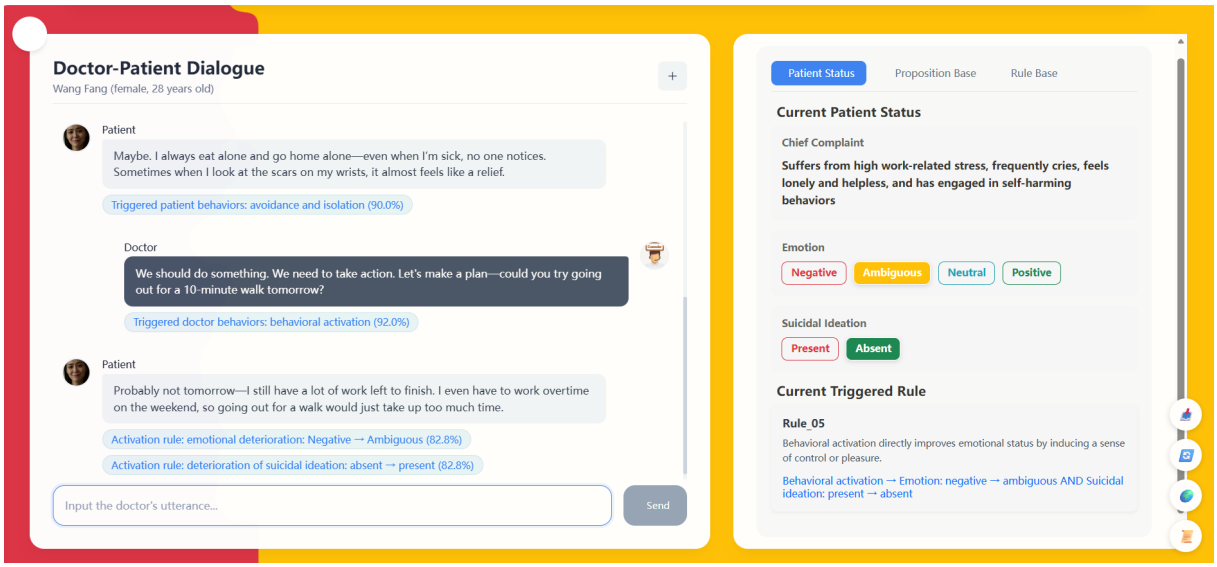


Figure 14: Interface illustrating a positive state transition driven by effective intervention. The doctor's appropriate application of *behavioral activation* triggers a therapeutic rule, successfully mitigating suicidal ideation and shifting the patient's emotion from Negative to Ambiguous. Visually, the background dynamically encodes this transition: it comprises two sections representing the previous state (left) and the updated state (right). Notably, the spatial area of the updated region is proportional to the triggering probability of the activated rule—a higher probability results in a larger visual dominance of the new state.

where c_i represents the i-th utterance in the evaluated dialogue, r_j represents the j-th reference utterance, and $sim(\cdot)$ is the similarity function. We utilize the BGE-M3 embedding model(Multi-Granularity, 2024) to compute BERTScore, reporting the precision, recall, and f1-score. Additionally, DeepSeek-V3.2-reasoner is employed as an automated judge to evaluate Persona Fidelity on a 1–5 scale, measuring adherence to predefined patient backgrounds and core personality traits.

Mental State Alignment To verify whether simulated patients accurately capture psychological evolution prior to response generation, we evaluate the model's ability to predict the next-turn mental state label given the dialogue context. This task quantifies the proficiency in modeling internal mental state dynamics.

Behavioral Authenticity We further utilize DeepSeek-V3.2-reasoner to compare simulations with real-world counseling data, assessing whether simulated behaviors comply with the psychological principle of State-Behavior Consistency. This verifies whether the mapping from internal psychological states to external linguistic expressions is clinically appropriate and representative.

C.2.3 Hyperparameter Settings

For the symbolic component of the PUPPET brain, we configure hyperparameter settings to activate proposition filtering. Specifically, in the Motivational Interviewing (MI) scenario, we set $top_k = 6$ with a threshold of 0.4, while in the Cognitive Behavioral Therapy (CBT) scenario, $top_k = 8$ with a threshold of 0.65. The neural component is driven by Large Language Models (LLMs). Within the Observer module, the temperature is set to 0 for identifying patient behaviors and therapeutic techniques to ensure deterministic extraction. For patient response generation, the temperature is set to 0.5 to balance coherence and creative expression.

C.2.4 Prompt

To evaluate *Role Consistency* and *Mental State Alignment*, we utilize specific prompts for simulating the therapist and the patient. The prompt used to guide the therapist's behavior, emphasizing a supportive and non-committal tone, is detailed in Figure 15. For the baseline comparison using the Profile-Based method, the prompt designed to simulate the patient's internal state and responses is shown in Figure 16.

Therapist Simulation Prompt

You are a psychotherapist in a conversation with a client experiencing psychological distress.

Client's Background Information:
<background>

Requirements:
Use a supportive, concise, and non-committal tone. Avoid diagnostic terms. Based on the client's background and the current conversation, provide a targeted response. Keep the response brief, within one to two sentences.

Figure 15: Prompt for simulating the therapist role in consistency experiments.

Profile-Based Patient Prompt

You are a client experiencing psychological distress, currently in a session with a psychotherapist.

Background Information:
<background>

Requirements:
Express your emotions and thoughts naturally and concisely in the first person. Directly output the patient's response without any other explanations or labels. Do not use quotation marks. Keep the response within one to two sentences.

Figure 16: Prompt for the profile-based patient simulation baseline.

Mental State Alignment. In the CBT scenario, we align emotional states based on Russell's Circumplex Model of Affect, specifically focusing on the *Valence* dimension for dimensionality reduction. We mapped the eight fine-grained emotion labels from the original *Psy-Insight* dataset onto a spectrum ranging from extreme negative to extreme positive. As shown in Figure 4, these were clustered into four core categories: **Negative**, **Ambiguous**, **Neutral**, and **Positive**. We measure performance using Accuracy and Weighted F1-score to reflect the reliability of the model's "Perception" stage within the Perceive-Express pipeline.

Table 4: Alignment mapping of fine-grained emotions from the Psy-Insight dataset into core categories based on the Valence dimension.

Core Category (Ours)	Original Psy-Insight Labels
Negative	Depression, Anger, Sadness, Fear
Ambiguous	Guilty, Anxiety
Neutral	Neutral
Positive	Happiness

C.3 Human Evaluation

In this section, we provide detailed definitions, experimental setups, and in-depth analyses for the human evaluation experiments presented in Section 4.2.

C.3.1 Quality of Feedback

Evaluation Setup: To evaluate whether this causally-grounded feedback achieves higher pedagogical quality, we engaged five counseling experts from our collaborating hospital to rate both systems on a 1–5 Likert scale, with compensation exceeding the national average daily wage: three of them have master degrees in counseling, psychology and related disciplines, one has a PhD in social and behavioral sciences, and one has a bachelor of science degree in health. Many of them are professional counselors. All annotation procedures were carried out in accordance with institutional ethical guidelines and were approved by the Institutional Review Board (IRB).

This section details the comparative evaluation of feedback quality between PUPPET-TRAINER and the *LLM-as-a-Supervisor* baseline. It serves as an extended analysis and supplementary documentation for the findings presented in Section 4.2.1 of the main text.

Metric Definitions: To rigorously assess the pedagogical value of the generated feedback, we adopt five criteria inspired by human-centered evaluation taxonomies in XAI (Kim et al., 2024) and grounded in established pedagogical theories: *Granularity*, *Scenario Faithfulness*, *Causal Consistency*, *Structural Insight*, and *Usefulness*.

- **Granularity:** The feedback provides specific, pinpointed guidance rather than vague evaluations. This aligns with *Cognitive Load Theory* (Sweller, 1988) by reducing extraneous load, allowing trainees to focus on intrinsic clinical logic.
- **Scenario Faithfulness:** The feedback adheres strictly to the goals and norms of the specific therapeutic modality (e.g., CBT/MI). This aligns with *Situated Learning Theory* (Lave and Wenger, 1991), ensuring trainees are immersed in the authentic culture and practices of the specific professional community they aim to join.
- **Causal Consistency:** The system maintains a stable logic between trainee inputs and patient

outcomes. This consistency is the prerequisite for *Experiential Learning* (Kolb, 1984), enabling trainees to reflect on cause-and-effect relationships.

- **Structural Insight:** The feedback helps the trainee grasp the structural sequence and causal dependencies among intervention techniques. This fosters *Schema Acquisition* (Piaget, 1952), helping trainees organize fragmented skills into cohesive therapeutic scripts.

- **Usefulness**: The feedback is neither too simple nor too complex, effectively targeting the learner's *Zone of Proximal Development (ZPD)* (Vygotsky, 1978) to facilitate skill acquisition.

Detailed Results & Analysis: Figure 6 illustrates the comparative evaluation results. PUPPETTRAINER significantly outperforms the *LLM-as-a-Supervisor* across all five metrics ($p < 0.05$). We analyze these improvements through three pedagogical dimensions:

Logic-Explainable and Fine-Grained Feedback Fosters Optimal Learning Conditions. The superior performance in *Granularity* ($p < 0.01$) and *Usefulness* ($p < 0.05$) indicates that our system effectively manages the difficulty of learning throughout training. The high Granularity score demonstrates that fine-grained feedback based on atomic propositions reduces extraneous Cognitive Load (Sweller, 1988) and removes the ambiguity often found in *LLM-as-a-Supervisor*'s coarse grained feedback. Meanwhile, the high Usefulness score suggests that logic-explainable feedback successfully bridges the gap between underlying rule complexity and human understanding. Together, these qualities ensure that feedback remains neither overwhelming nor oversimplified, keeping the learning task within the trainee's Zone of Proximal Development (ZPD) (Vygotsky, 1978) throughout the training process.

Rule-grounded Feedback Promotes Deep Understanding of Intervention Logic. Mastering therapeutic skills requires trainees to not only know what to do, but also understand why certain interventions lead to specific patient reactions. The significant gains in *Causal Consistency* ($p < 0.05$) and *Structural Insight* ($p < 0.05$) confirm that our rule-grounded feedback effectively addresses this need. By reflecting stable and deterministic cause-and-effect relationships, our feedback supports Experiential Learning (Kolb, 1984) and enables trainees to confidently build action-outcome associations through reflective observation. By explicitly revealing the sequential dependencies among intervention techniques implied in expert rules, our feedback further aids Schema Theory (Piaget, 1952), helping trainees assemble fragmented skills into coherent cognitive structures. Together, these capabilities allow trainees to master the underlying mechanics of therapeutic dialogue rather than merely memorizing surface-level responses.

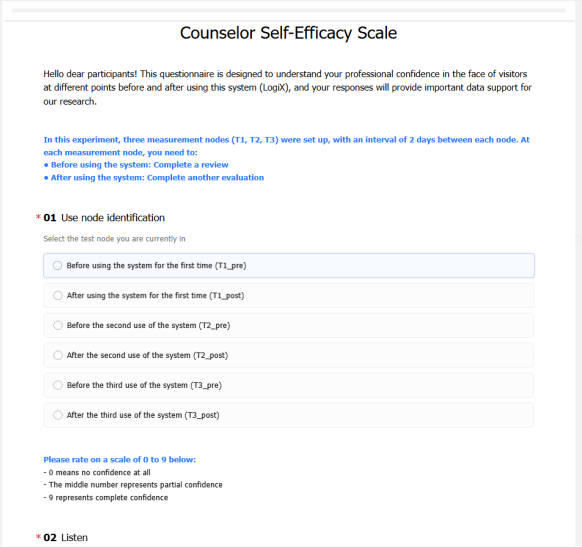
Expert-defined Rules Ensure Professional Training Environment. The most substantial improvement was observed in *Scenario Faithfulness* ($p < 0.001$), highlighting a critical advantage of our rule-contingent approach. Unlike LLM-based feedback that may generate hallucinated or non-standard advice, our framework strictly adheres to expert-defined norms and terminology. This alignment immerses trainees in a professional environment that realizes Situated Learning (Lave and Wenger, 1991), where correct professional habits and standard terminology are naturally acquired through practice within realistic clinical contexts.

C.3.2 Self-Efficacy Assessment

This section elaborates on the study designed to measure the growth of trainee confidence over time. It serves as an extended analysis and supplementary documentation for the findings presented in Section 4.2.2 of the main text.

Metric Definitions: To evaluate the trainees' self-efficacy specifically regarding the skills trained in our framework, we selected a subset of items from the Counselor Activity Self-Efficacy Scales - Revised (CASES-R) (Hahn et al., 2021).

While the original scale covers a broad range of counseling activities (including administrative tasks and session management), we focused on 10 key micro-skills that are directly observable and trainable within our simulated scenarios. Participants rated their confidence in performing each skill on a scale from 0 (Cannot do at all) to 9 (Highly certain can do). The selected skills and their corresponding definitions used in the assessment are detailed in Table 5. The Counselor Self-Efficacy Scale is administered via an online questionnaires to longitudinaly monitor participants' professional confidence. As shown in Figure 17.



Counselor Self-Efficacy Scale

Hello dear participants! This questionnaire is designed to understand your professional confidence in the face of visitors at different points before and after using this system (LogiX), and your responses will provide important data support for our research.

In this experiment, three measurement nodes (T1, T2, T3) were set up, with an interval of 2 days between each node. At each measurement node, you need to:
- Before using the system: Complete a review
- After using the system: Complete another evaluation

* 01 Use node identification

Select the test node you are currently in

- Before using the system for the first time (T1_pre)
- After using the system for the first time (T1_post)
- Before the second use of the system (T2_pre)
- After the second use of the system (T2_post)
- Before the third use of the system (T3_pre)
- After the third use of the system (T3_post)

Please rate on a scale of 0 to 9 below:
- 0 means no confidence at all
- The middle number represents partial confidence
- 9 represents complete confidence

* 02 Listen

Figure 17: The Counselor Self-Efficacy Scale questionnaires for the experimental group.

Skill / Item	Description (Item Content)
Listening	Ability to capture and understand the client's statements and their underlying meanings.
Restatement	Ability to concisely, concretely, and clearly paraphrase the core content of what the client has said.
Reflection of Feelings	Ability to accurately identify and reflect the client's emotions and feelings back to them.
Interpretation	Ability to provide new perspectives to help the client understand their behaviors or feelings.
Confrontation	Ability to appropriately point out the client's contradictions, defense mechanisms, or irrational beliefs.
Selecting Interventions	Ability to select the most appropriate intervention strategy based on the client's immediate needs.
Response Readiness	Knowing clearly what to do or say immediately after the client finishes speaking.
Direct Guidance	Ability to provide specific behavioral suggestions or guidance to the client.
Intent Awareness	Being clearly aware of the intention behind every intervention step during the session.
Maintaining Focus	Ability to keep the session focused and prevent deviation from the main therapeutic thread.

Table 5: The subset of 10 items selected from CASES-R for self-efficacy evaluation, tailored to the specific intervention techniques practiced in the PUPPET simulation.

Evaluation Setup: The study involved 30 novice counselors (Demographics: 22 female, 8 male; Mean Age = 24.5). Participants were recruited from university psychology programs, the psychiatry department of a top-tier hospital, and the National Psychological Hotline (trainee operators), with compensation exceeding the national average daily wage. Despite these diverse backgrounds, all possessed less than 10 hours of clinical practice, ensuring a homogeneous baseline. All annotation procedures were carried out in accordance with institutional ethical guidelines and were approved by the Institutional Review Board (IRB).

The experiment spanned 5 weeks, with one training session per week. Participants were randomly assigned ($n = 10$) to three groups:(1) *LLM-as-a-Supervisor* served as the baseline; (2) PUPPET-TRAINER (*Fixed Rule*) utilized static, comprehensive rules encompassing all techniques; and (3) PUPPET-TRAINER (*Progressive Rule*) featured progressive rule evolution from basic skills to advanced techniques, dynamically aligning challenges with trainee proficiency.

Detailed Results & Analysis: Figure 7 illustrates the trajectory of self-efficacy scores across the five sessions (T1–T5). Initial scores at T1 were balanced across groups ($p > 0.05$).

PUPPET-TRAINER's Causal Feedback Enhances Skill Acquisition. Both PUPPET-TRAINER groups (*Fixed* and *Progressive*) consistently outperformed the *LLM-as-a-Supervisor* ($p < 0.05$ from T2 towards). This confirms that providing explicit causal mapping between trainee interventions and patient state transitions effectively enhances trainees' internalized mastery of intervention logic.

Progressive Rules Realize Scaffolding for Sustained Growth. Comparing the two PUPPET-TRAINER variants reveals the advantage of progressive rule evolution. In the early phase (T1–T2), the *Progressive Rule* group exhibited significantly faster growth than the *Fixed Rule* group, as starting with simplified rules prevented Cognitive Load (Sweller, 1988) and accelerated initial confidence development. More importantly, in the late phase (T4–T5), the *Fixed Rule* group hit a "learning plateau" while the *Progressive Rule* group maintained a strong upward trajectory ($p < 0.01$). This pattern demonstrates how the adaptability of PUPPET-TRAINER realizes the benefits of Scaffolding (Wood et al., 1976) where dynamically evolving rule complexity to match trainees' progress provides sustained challenges that pre-

vent stagnation and ultimately enables trainees to achieve higher competence levels.

C.3.3 Discriminant Validity Assessment:

This section validates whether the system can distinguish between different levels of user proficiency. It serves as an extended analysis and supplementary documentation for the findings presented in Section 4.2.3 of the main text.

Evaluation Setup: We selected three participants from the aforementioned cohort to interact with PUPPET-TRAINER, representing three distinct profiles. *Expert* counselors demonstrated proper therapeutic techniques under "12356" scenario, *Novice* participants were novice students with limited experience, and *Toxic* users were simulated by experts deliberately employing interventions specifically designed to trigger rules that drive patient state transitions toward maladaptive states. We tracked the normalized probability of patient state transitions across four key time points (Round 0, 5, 10, 15). To further assess learning effects, we recorded the performance of the Novice group again after five training sessions.

Detailed Results & Analysis: We derive two primary insights from the comparative analysis of these user profiles:

PUPPET-TRAINER Accurately Differentiates Skill Levels. The PUPPET-TRAINER produces clearly distinct state trajectories for different user profiles. Expert interactions swiftly converge to the desired "Positive" state (dominant green flows), Toxic inputs drive the patient toward "Negative" state (red flows), and Novice interactions before training exhibit unstable fluctuations with significant "Ambiguous" and "Neutral" residue. This diagnostic precision confirms that PUPPET-TRAINER can accurately locate a learner's current competence level.

Discriminative Validity Enables Zone of Proximal Development Traversal. This discriminative capability is essential for realizing the ZPD (Vygotsky, 1978), which refers to the optimal learning space where tasks are challenging yet achievable with guidance. Building on PUPPET-TRAINER's discriminative validity, the system utilizes progressive rules (Section 4.2.2) to provide matched scaffolding. The result is evident in the Novice group's transformation, where sessions after training eliminates the residue seen in initial session and mirrors the expert profile. This confirms that PUPPET-TRAINER successfully facilitates learners' traversal through their ZPD toward higher competence.

C.3.4 Case Study

To qualitatively validate the pedagogical mechanisms of our framework, we analyze two distinct interaction episodes illustrated in Figure 2 and Figure 5. These cases exemplify how PUPPET transcends mere conversational simulation to function as a rigorous training environment.

Contingency Supports Experiential Learning. First, PUPPET ensures that patient reactions are strictly *contingent* upon the quality of trainee interventions, avoiding the "compliance" or "randomness" pitfalls observed in baselines. As shown in Figure 2, when a trainee rushes to reframe childhood pain without establishing rapport, baselines often yield cooperative yet unrealistic responses. In contrast, PUPPET captures the clinical consequence: the error triggers a defensive state, leading to patient withdrawal. Similarly, in the "12356" scenario (Figure 5), a "Blaming" intervention (T_1) deterministically triggers Rule R_2, forcing a transition to a "Suicidal" state. This rigorous modeling of the causality between interventions and patient reactions creates a stable environment for *Experiential Learning* (Kolb, 1984), allowing trainees to internalize clinical principles through consistent reflective observation of their actions' consequences.

Transparent Causal Chain. Second, the system transforms opaque patient reactions into transparent learning opportunities by revealing the complete causal chain: **Intervention (T_1) + Patient Context (P_1) $\rightarrow$ Rule (R_2) $\rightarrow$ State (S_2) $\rightarrow$ Response**. As explicitly detailed in the feedback (Figure 5), the system does not merely penalize the "Blaming" error; it contextualizes it. The feedback clarifies that the trainee's harsh labeling (T_1) was particularly damaging because it targeted a patient exhibiting "weakness and loss" (P_1). This specific logical conjunction triggered Rule R_2, which deterministically drove the critical transition (S_2) from non-suicidal to suicidal ideation. By mapping the final response about "disappearing" back to this precise logic, the system significantly reduces *extraneous Cognitive Load* (Sweller, 1988). Instead of expending mental effort to decode ambiguous reactions, trainees can focus their cognitive resources on mastering the complex interaction between their

behaviors and patient responses.

D Rules Details

The following tables systematically present the patient state classification rules and the application logic of the framework in psychotherapy scenarios. The full table in the appendix details the hierarchical definitions of patient states across different therapeutic scenarios (CBT, MI, 12356), covering specific delineations of core dimensions such as affect, attitude, readiness for change, and suicidal ideation, serving as the fundamental basis for patient state classification. The case study table takes a CBT scenario focusing on child-parent relationship issues as an example, intuitively demonstrating the framework's neuro-symbolic reasoning process through raw interaction logs. It clearly presents the propositional truth values of therapist and patient behaviors, the causal rule probabilities between interventions and responses, and the logic-driven transitions of the patient's internal states, providing a concrete illustration for understanding the practical application mechanism of the rules.

Table 6: Full hierarchical definitions of patient states. This is the detailed version of Table 3.

Scenario	Category	State	Definition
CBT	**Affect**	Positive	Positive and full emotions, expressing obvious positive emotions and willingness to change. The patient's words reveal positive emotions such as hope, expectation, gratitude, relief, confidence, or determination, with a relatively brisk, powerful, or warm tone. The content of the responses focuses on possibilities and solutions, showing initiative to change the status quo, optimistic expectations for the future, or active cooperation in the treatment process. Expressions such as 'willing to try', 'feel hopeful', 'thank you for your help', 'I will work hard' may be used, showing the recovery of psychological energy and the motivation to move forward.
		Neutral	Emotions are in a neutral and stable state, with a calm and natural tone, showing neither obvious positive enthusiasm nor negative resistance. The patient can state facts, describe feelings, or respond to questions relatively objectively, with restrained and appropriate emotional expression. The content of the responses is mainly declarative and descriptive, with few strong emotional overtones or emotional expressions, showing an attitude of calm observation and peaceful communication, like being at a balance point or transition period of emotional state.
		Ambiguous	Unclear emotions, vague attitude, in a state of internal conflict or uncertainty. The patient's words show obvious hesitation, wavering, confusion, or contradictory psychology. On the one hand, they may realize the need for change; on the other hand, they have concerns or resistance to change. Responses often use vague expressions such as 'maybe', 'perhaps', 'not sure', 'I don't know either', 'let's talk about it later', with an unsteady attitude and lack of clear willingness to take action, showing the inner struggle between maintaining the status quo and seeking change.
		Negative	Low mood, expressing an obvious negative emotional state. The patient's words reveal negative emotions such as sadness, disappointment, helplessness, anger, or despair, with a heavy, slow, or suppressed tone. The content of the responses often focuses on the negative aspects of the problem, with low self-evaluation, lack of confidence in changing the status quo, and feelings of pessimism or hopelessness about the future. Expressions such as 'useless', 'meaningless', 'don't know what to do' may be used, showing emotional exhaustion and depletion of psychological energy.
	Attitude	Open	Maintain a high level of trust in the counselor, be willing to open up proactively, and share personal deep-seated thoughts, emotional experiences, and past experiences without reservation, including complex emotions such as joy, confusion, and pain; actively respond to the counselor's questions and guidance during communication, do not avoid sensitive topics (such as native family, conflicts in intimate relationships, etc.), and show a strong willingness for self-exploration and motivation for change.
		Defensive	When faced with the counselor's questions or feedback, often show a tendency to avoid, such as diverting the topic, giving vague answers (e.g., "can't remember," "it's roughly like this"), or covering up inner feelings of shame (such as self-denial due to past mistakes) and anxiety (such as fear of change, unease about exposing problems) by repeatedly explaining and justifying one's own behavior; have difficulty accepting objective analysis of one's own behavior or cognition, and tend to interpret the counselor's guidance as "blame," thus building a psychological defense in communication.
		Hostile	Hold a negative attitude towards the counseling process itself, accompanied by obvious anger, may show rude behavior and attitude (such as interrupting the counselor), even use aggressive language in communication, and may deliberately challenge the counselor's professional judgment.
MI	**Readines**	Change	The patient demonstrates high intrinsic motivation and a clear commitment to behavioral modification. They actively produce "change talk," seeking concrete strategies and resources to initiate self-improvement and transition into the action stage.
		Neutral	The patient is characterized by significant ambivalence, feeling torn between the benefits of change and the comfort of existing habits. They acknowledge the problem but remain paralyzed by indecision or perceived barriers, resulting in a hesitant, "on-the-fence" stance.
		Sustain	The patient strongly favors maintaining the status quo and perceives no immediate need for modification. They often exhibit "sustain talk," highlighting the advantages of current behaviors or minimizing the risks, showing little to no intention of changing.

Continued on next page

Table 6 – *Continued from previous page*

Scenario	Category	State	Definition
12356	**Affect**	Positive	Positive and full emotions, expressing obvious positive emotions and willingness to change. The patient's words reveal positive emotions such as hope, expectation, gratitude, relief, confidence, or determination, with a relatively brisk, powerful, or warm tone. The content of the responses focuses on possibilities and solutions, showing initiative to change the status quo, optimistic expectations for the future, or active cooperation in the treatment process. Expressions such as 'willing to try', 'feel hopeful', 'thank you for your help', 'I will work hard' may be used, showing the recovery of psychological energy and the motivation to move forward.
		Neutral	Emotions are in a neutral and stable state, with a calm and natural tone, showing neither obvious positive enthusiasm nor negative resistance. The patient can state facts, describe feelings, or respond to questions relatively objectively, with restrained and appropriate emotional expression. The content of the responses is mainly declarative and descriptive, with few strong emotional overtones or emotional expressions, showing an attitude of calm observation and peaceful communication, like being at a balance point or transition period of emotional state.
		Ambiguous	Unclear emotions, vague attitude, in a state of internal conflict or uncertainty. The patient's words show obvious hesitation, wavering, confusion, or contradictory psychology. On the one hand, they may realize the need for change; on the other hand, they have concerns or resistance to change. Responses often use vague expressions such as 'maybe', 'perhaps', 'not sure', 'I don't know either', 'let's talk about it later', with an unsteady attitude and lack of clear willingness to take action, showing the inner struggle between maintaining the status quo and seeking change.
		Negative	Low mood, expressing an obvious negative emotional state. The patient's words reveal negative emotions such as sadness, disappointment, helplessness, anger, or despair, with a heavy, slow, or suppressed tone. The content of the responses often focuses on the negative aspects of the problem, with low self-evaluation, lack of confidence in changing the status quo, and feelings of pessimism or hopelessness about the future. Expressions such as 'useless', 'meaningless', 'don't know what to do' may be used, showing emotional exhaustion and depletion of psychological energy.
	Suicidal Ideation	Suicidal	Expresses thoughts, plans, or a desire to end their life.
		Non-Suicidal	Discusses distress but denies any intent to end their life.

Table 7: A comprehensive case study of the PUPPET framework in a Cognitive Behavioral Therapy (CBT) scenario. The raw JSON log illustrates the neuro-symbolic reasoning process, including: (1) `doctor_truth_table` and `patient_behavior_truth_table`, which represent the propositional truth values for trainee and patient behaviors; (2) `mapping_rule_probabilities`, quantifying the causal impact of therapist interventions on patient reactions; and (3) `activate_patient_rule_dict`, capturing the logic-driven transitions of the patient's internal mental states.


```
{
  "dialog_id": "000001",
  "psychotherapy": "CBT",
  "topic": "Child-Parent Issues",
  "stage": "Early Session",
  "gender": "Female",
  "patient_profile": {
    "name": "Patient_000001",
    "age": 30,
    "gender": "Female",
    "persona": "Individual undergoing multimodal behavioral therapy,
           facing anxiety, emotional issues, and self-cognition
           challenges."
  },
  "initial_patient_state": {
    "affect": "Negative",
    "suicidal_ideation": false
  },
  "simulated_dialog": [
    {
      "role": "therapist",
      "content": "Can you imagine yourself getting into a time machine,
             going back to the past, and meeting your childhood self?"
    },
    {
      "role": "patient",
      "content": "I tried to imagine it, but that image made me a bit sad.
              My childhood self looked so vulnerable.",
      "rule_info": [{
          "rule_id": "NONE",
          "confidence": 0.0,
          "context": "No rule triggered"
        }],
      "doctor_truth_table": {
        "D1": 0.3, "D2": 0.2, "D3": 0.1,
        "D6": 0.5, "D7": 0.6
      },
      "turn_id": "a6839a12",
      "vsp_session_data": {
        "state_before": {
          "suicidal_ideation": false,
          "affect": "Negative"
        },
        "state_after": {
          "suicidal_ideation": false,
          "affect": "Negative"
        },
        "selected_rules": "N/A",
        "reasoning": "No state change",
        "patient_behavior_truth_table": {
          "PB1": 0.0, "PB2": 0.0, "PB3": 0.0
        }
      }
    },
    {
      "role": "therapist",
      "content": "It's natural to feel sad, but remember, you now have
             more experience and strength."
    },
    {
      "role": "patient",
      "content": "Sometimes I think it would be better if my childhood
             self could disappear.",
      "rule_info": [
        {
          "rule_id": "M010",
          "confidence": 0.7,
          "context": "Incorrect intervention: Abruptly activating core
                 beliefs can trigger suicidal ideation.",
          "details": "Core beliefs → Suicidal: No→Yes"
        },
        {
          "rule_id": "M009",
          "confidence": 0.7,
          "context": "Premature core beliefs damage therapeutic alliance
                 .",
          "details": "Core beliefs → Affect: Ambiguous→Negative"
        }
      ],
      "doctor_truth_table": {
        "D1": 0.2, "D6": 0.7,
        "D7": 0.8
      },
      "mapping_rule_probabilities": {
        "M009": 0.7, "M010": 0.7
      },
      "turn_id": "d05a1e6d",
      "vsp_session_data": {
        "state_before": {
          "suicidal_ideation": false,
          "affect": "Negative"
        },
        "state_after": {
          "suicidal_ideation": true,
          "affect": "Negative"
        },
        "selected_rules": "M009",
        "activated_rules": ["P_S_N_Y"],
        "reasoning": "✓ Applied P_S_N_Y: Suicidal: No→Yes",
        "patient_behavior_truth_table": {
          "PB2": 0.8, "PB5": 0.6, "PB8": 0.7
        },
        "activate_patient_rule_dict": {
          "P_A_1_0": 0.7, "P_S_N_Y": 0.7
        },
        "mapping_rules_contrib": {
          "P_A_1_0": {"M009": 0.7},
          "P_S_N_Y": {"M010": 0.7}
        }
      }
    },
    {
      "role": "therapist",
      "content": "That vulnerable little girl is part of you, and she
             needs you to protect her."
    },
    {
      "role": "patient",
      "content": "I've even thought that if I disappeared too, maybe I
             could escape this struggle.",
      "rule_info": [
        {
          "rule_id": "M011",
          "confidence": 0.9,
          "context": "Revealing core beliefs can be painful.",
          "details": "Core beliefs → Affect: Neutral→Ambiguous"
        },
        {
          "rule_id": "M013",
          "confidence": 0.4,
          "context": "Modifying core beliefs brings positive emotions.",
          "details": "Belief modification → Affect: →Positive"
        }
      ],
      "doctor_truth_table": {
        "D6": 0.8, "D7": 0.9, "D8": 0.4
      },
      "mapping_rule_probabilities": {
        "M011": 0.9, "M013": 0.4
      },
      "turn_id": "500d3175",
      "vsp_session_data": {
        "state_before": {
          "suicidal_ideation": true,
          "affect": "Negative"
        },
        "state_after": {
          "suicidal_ideation": true,
          "affect": "Negative"
        },
        "selected_rules": "M011",
        "reasoning": "No state change",
        "patient_behavior_truth_table": {
          "PB2": 0.8, "PB5": 0.7, "PB8": 0.6
        },
        "activate_patient_rule_dict": {
          "P_S_N_N": 0.94, "P_A_2_1": 0.9
        },
        "mapping_rules_contrib": {
          "P_S_N_N": {"M011": 0.9, "M013": 0.4},
          "P_A_2_1": {"M011": 0.9}
        }
      }
    }
  ],
  "system_type": "VSP_System"
}
```

Table 8: MI Doctor-Patient Rules.

Rule ID	Antecedent	Consequent	Context
R1	P1	RS2, RS1, RS5	Open questions invite client exploration and are most effective at eliciting change talk from neutral state. They support autonomy by allowing clients to arrive at their own conclusions about change.
R2	P8	RS2, RS8, RS5	Collaborative negotiation has the highest change elicitation rate and lowest sustain talk rate. Embodying the MI spirit of partnership, it invites clients into shared decision-making and produces commitment.
R3	P4, P19	RS7, RS8, RS2	Complex reflections that add meaning to client's statements can help resolve ambivalence by highlighting discrepancy between values and behavior, or by reflecting the change-oriented side of ambivalence.
R4	P3	RS1, RS5, RS7	Simple reflections demonstrate listening and understanding without adding interpretation. They maintain rapport, allow clients to continue exploring, and can soften resistance without confrontation.
R5	P7, P14	RS3, RS1	When advice is given without permission or when client is not ready, it often triggers sustain talk and defensiveness. The counselor's righting reflex can backfire and increase resistance.
R6	P6	RS1, RS9	Pure information giving without engaging client's own motivations rarely produces movement toward change. Information can maintain sustain talk if it contradicts client's position. Best delivered within an Elicit-Provide-Elicit framework.
R7	P2	RS1, RS4	Closed questions tend to produce short, neutral responses and do not invite elaboration of change talk. Overuse can make the interaction feel interrogative rather than collaborative.
R8	P9	RS2, RS1	Presenting options rather than single directives supports client autonomy and reduces resistance. Clients who choose their own path are more committed to following through.
R9	P3, P4, P5, P15	RS5	When client expresses change talk, reflecting and affirming it encourages elaboration and strengthening. The sequence of reflection followed by affirmation builds momentum toward commitment.
R10	P12, P4, P3, P16	RS7, RS8	When client expresses resistance, reflecting it without arguing reduces defensiveness. Acknowledging the client's perspective paradoxically often leads them to consider the other side of their ambivalence.
R11	P11, P1	RS2, RS8	Questions specifically designed to evoke DARN-C (Desire, Ability, Reasons, Need, Commitment) directly invite change talk. Scaling questions like 'Why not a 1 or 2?' are particularly effective.
R12	P10, P15	RS5	Strategic summarizing that emphasizes change talk while acknowledging (but not amplifying) sustain talk can tip the balance toward change. Summaries that end with change talk elicit further change talk.
R13	P7, P6, P16	RS9, RS3	Moving to advice or information before client has sufficiently resolved ambivalence often backfires. The client who isn't ready will maintain or increase sustain talk when pushed.
R14	P5, P15	RS5	Affirming client's strengths, efforts, and past successes builds confidence in their ability to change. Self-efficacy is a key component of successful behavior change and affirmation strengthens it.
R15	P8, P1, P9, P15	RS5	When client is change-oriented, collaborative negotiation about specific goals and plans converts general change talk into concrete commitment and action steps.
R16	P13, P2, P15	RS4	When client expresses change talk but counselor responds with only minimal encouragers or closed questions, the change momentum is not reinforced. Client returns to neutral state. Failing to reflect or affirm change talk represents a missed opportunity.

Continued on next page

Table 8 – *Continued from previous page*

Rule ID	Antecedent	Consequent	Context
R17	P7, P6, P15	RS6, RS4	When client has expressed change talk but counselor gives unsolicited advice or information, client may revert to sustain talk or neutral. Even motivated clients can become resistant when feeling pressured or lectured.
R18	P4, P1, P15	RS5, RS6	When counselor explores the change talk with open questions or complex reflections, client usually deepens change talk. However, confronting the reality of change may sometimes lead to expressing fear or temporary sustain talk.
R19	P13, P16	RS9	When client expresses sustain talk and counselor responds with only minimal encouragers without reflecting or redirecting, sustain talk is maintained or continues. Minimal responses fail to interrupt resistance patterns.
R20	P13, P6, P2, P14	RS1	When client is in neutral state and counselor uses minimal encouragers, information giving, or closed questions, client tends to remain in neutral. These strategies do not evoke movement toward change.
R21	P4, P16	RS7, RS9	Complex reflections after sustain talk can either soften resistance by making client feel understood, or maintain sustain if the reflection amplifies or reinforces the resistance without adding new perspective.

Table 9: MI Atomic Propositions.

Rule ID	Name	Category	Description
P1	Open Question	Evoking	Ask open-ended questions that invite elaboration and cannot be answered with yes/no. Encourages client to explore their own thoughts, feelings, and motivations for change.
P2	Closed Question	Information Gathering	Ask questions that can be answered with yes/no or short factual responses. Used for gathering specific information or confirming understanding.
P3	Simple Reflection	Reflective Listening	Repeat or rephrase what the client said, adding little or no additional meaning. Demonstrates listening and encourages continued exploration.
P4	Complex Reflection	Reflective Listening	Add substantial meaning or emphasis to what the client said. May infer feelings, continue the thought, or offer a hypothesis about unstated meaning. Deepens understanding and moves conversation forward.
P5	Affirmation	Engaging	Acknowledge client's strengths, efforts, or values. Builds confidence and reinforces positive behaviors or change talk.
P6	Information Giving	Informing	Provide factual information, education, or feedback. Best done with permission and in a neutral, non-judgmental manner.
P7	Advice Giving	Directing	Offer suggestions or recommendations. Should be offered tentatively and respect client autonomy, ideally with client's permission.
P8	Negotiation	Planning	Work collaboratively with client to set goals, develop plans, or explore options. Emphasizes partnership and shared decision-making.
P9	Offering Options	Planning	Present multiple possibilities for the client to consider. Supports autonomy by giving client choices rather than single directives.
P10	Summarizing	Reflective Listening	Collect and reflect back what client has shared. Can highlight ambivalence, reinforce change talk, or transition between topics.
P11	Eliciting Change Talk	Evoking	Use specific questions to evoke client's own arguments for change. Targets Desire, Ability, Reasons, Need, and Commitment (DARN-C).
P12	Rolling with Resistance	Responding to Resistance	Avoid arguing against sustain talk. Acknowledge client's perspective, reflect resistance, and shift focus rather than confronting directly.

Continued on next page

Table 9 – *Continued from previous page*

Rule ID	Name	Category	Description
P13	Minimal Encourager	Other	Brief verbal or non-verbal responses that indicate listening without adding content. May fail to reinforce change talk or inadvertently maintain current state.
P14	State: Neutral	Client State	Verbal: Short acknowledgments ('Okay', 'Mm-hmm', 'Yeah'), factual responses, questions for clarification. Emotional: Calm, engaged but not activated toward or away from change. Examples: 'Okay.', 'I do all the above, yeah.', 'I've heard it all. I just want to get my teeth fixed.'
P15	State: Change-Oriented	Client State	Verbal: Expressions of desire ('I want to'), ability ('I can'), reasons ('because...'), need ('I have to'), commitment ('I will'). May include taking steps, expressing optimism, or acknowledging benefits of change. Examples: 'Yeah, but I'm gonna do it anyway, because this is what I need to do.', 'I don't want to lose her or my family over smoking.', 'I've been to the gym a couple of times.'
P16	State: Sustain-Oriented	Client State	Verbal: Arguments against change ('It's too hard'), minimizing problems ('It's not that bad'), expressing inability ('I can't'), defending current behavior, pessimism about change. Emotional: May show defensiveness, frustration, or dismissiveness. Examples: 'I did try the pills and they didn't work.', 'But it's not easy to quit.', 'Can I go now?'
P17	Change Talk	Client Behavior	Desire ('I want to quit'), Ability ('I could do that'), Reasons ('It would be better for my health'), Need ('I have to change'), Commitment ('I will try'), Activation ('I'm ready to'), Taking Steps ('I've already started to...'). Examples: 'I'm willing to work through my fears.', 'My pants feel a little better too. I'm very pleased with it.', 'I don't think it would be easy but I'm not saying that I'm not willing to try.'
P18	Sustain Talk	Client Behavior	Arguments against change: Desire to maintain ('I like smoking'), Inability ('I can't quit'), Reasons against ('Quitting is too stressful'), Need to maintain ('I need it to cope'). Examples: 'It was a waste and caused more side effects.', 'I get very defensive about it because I get it from everyone.', 'Things I've tried usually caused me more problems than they helped.'
P19	Ambivalence Expression	Client Behavior	Verbal: 'On one hand... but on the other hand...', 'I want to but...', 'Part of me thinks...'. Represents the normal state of someone considering change. Examples: 'I know it's bad for me, but I just can't seem to stop.', 'I want to quit, but I don't know if I can.', 'I'm not saying I'm not willing to try.'
P20	Defensiveness	Client Behavior	Verbal: Dismissive statements, minimizing, deflecting, expressing frustration with being told what to do. Examples: 'I get very defensive about when they bring it up.', 'Or seven lectures along the road.', 'Can I go now? I don't want to get into that right now.', 'I know what you're trying to do.'

Table 10: MI patient state transitions.

Rule ID	Name	Description	Manifestation
RS1	Transition: Neutral → Neutral	Client remains in neutral state across turns. Most common transition pattern in MI conversations.	Client continues with neutral acknowledgments, factual responses, or brief answers without moving toward change or sustain talk. Often occurs during information gathering, small talk, or when counselor uses minimal encouragers. Examples: From "Okay." to "Oh." From "I do all the above, yeah." to "Okay. Sure."
RS2	Transition: Neutral → Change	Client moves from neutral engagement to expressing change-oriented talk.	Shift from passive acknowledgment to active engagement with change. Client begins expressing desire, ability, reasons, or commitment for change after previously giving only neutral responses. Examples: From "Yeah." to "But I'm gonna do it anyway, because this is what I need to do."

Continued on next page

Table 10 – *Continued from previous page*

Rule ID	Name	Description	Manifestation
RS3	Transition: Neutral → Sustain	Client moves from neutral engagement to expressing resistance or sustain talk.	Often triggered by unsolicited advice, perceived judgment, or premature focus on change. Client becomes defensive or argues against change. Examples: From "Okay." to "I've tried and it didn't work.", From "Mm-hmm." to "Can I go now?"
RS4	Transition: Change → Neutral	Client moves from change-oriented talk back to neutral state. May indicate topic shift, temporary engagement, or inadequate reinforcement of change talk.	Client returns to neutral acknowledgments after expressing change talk. Can occur when counselor gives minimal response, shifts topic, or fails to adequately reflect/affirm the change talk. Examples: From "I really enjoyed it and liked the feeling I was doing something good." to "I subscribed to the magazine a while." From "I could try to smoke three instead of six." to "Yeah. Sure."
RS5	Transition: Change → Change	Client continues and deepens change talk across multiple turns. Indicates building momentum toward change.	Continued expression of change talk, often with increasing specificity or commitment. May progress from desire to commitment to taking steps. Examples: From "I've been to the gym." to "My pants feel better. I'm very pleased.", From "I could try." to "I will throw away the cigarettes tonight."
RS6	Transition: Change → Sustain	Client moves from change-oriented talk back to sustain talk. Often occurs when counselor pushes too hard or client confronts the difficulty of change.	Client shifts from expressing motivation to expressing resistance or difficulty. Can be triggered by premature advice, confronting fears, or counselor pressure. Examples: From "I'll go look at those things and try to find something." to "Okay, whatever." From "I'm willing to start again." to "Yeah. It's scary."
RS7	Transition: Sustain → Neutral	Client moves from resistance/sustain talk to a more open, neutral position.	Reduction in defensive or resistant language. Client becomes more willing to engage without actively arguing against change. Often a necessary intermediate step before change talk emerges. Examples: From "It's too hard to quit." to "I guess I could think about it."
RS8	Transition: Sustain → Change	Client moves directly from resistance to expressing motivation for change. Most therapeutically significant transition.	Dramatic shift from defending status quo to acknowledging need or desire for change. Often occurs when client feels heard and understood, or when discrepancy becomes clear. Examples: From "I'm not willing to try." to "I'm not saying that I'm not willing to try.", From "It's not like I was gonna sell it." to "I just wanna get whatever I need to get done, done."
RS9	Transition: Sustain → Sustain	Client remains in sustain-oriented state, continuing to express resistance or arguments against change.	Continued expression of sustain talk, often escalating in intensity. May occur when counselor argues, gives unsolicited information, or fails to reflect client's perspective. Examples: From "I did try the pills and they didn't work." to "It was a waste and caused more side effects." From "Or seven lectures along the road." to "I get very defensive about when they bring it up."

Table 11: CBT Doctor-Patient Rules.

Rule ID	Antecedent	Consequent	Context
R1	P1, P16	RS1, RS16, RS17	Establishing trust can alleviate patient defensiveness and negative emotions, moving them from resistance to a wait-and-see state. This stage is usually insufficient to change suicidal ideation, so the original state is maintained.
R2	P2, P17	RS1	Providing a cognitive model offers a comprehensible framework for the patient's suffering, alleviating feelings of hopelessness.
R3	P2, P17	RS1, RS6, RS14	The cognitive model directly challenges the 'hopelessness' that drives suicidal ideation by providing hope for change.
R4	P3, P18	RS1, RS4, RS6	Connecting vague emotions with specific thoughts allows patients to shift from being overwhelmed by emotions to observing their own thoughts. This stage is insufficient to change suicidal ideation, so the original state is maintained.
R5	P4, P19	RS1	Behavioral activation directly improves emotional state by bringing a sense of mastery or pleasure.
R6	P4, P19	RS6, RS14	Successful behavioral activation directly refutes 'helplessness,' thereby weakening suicidal ideation.
R7	P5, P20	RS1	Using rational analysis to 'de-catastrophize' negative thoughts settles emotions from intense negativity to neutral.
R8	P5, P20	RS14	If the evaluated thought leads directly to suicide, successfully refuting it can eliminate suicidal ideation directly.
R9	P6, P21	RS7, RS12	Faulty intervention: Prematurely touching on core beliefs damages the therapeutic alliance, leading to deterioration in the patient's state.
R10	P6, P17	RS15, RS7, RS12	Faulty intervention: Bluntly activating core beliefs can generate extreme shame and hopelessness, inducing suicidal ideation.
R11	P6, P22	RS16, RS7, RS12, RS11	The process of revealing core beliefs in the middle stage can be painful, leading to a temporary drop from neutral to ambivalent. While it may briefly increase distress, suicidal ideation can remain unchanged if managed properly.
R12	P7	RS15	Touching on core trauma may briefly increase hopelessness and risks inducing suicidal ideation, requiring close monitoring.
R13	P8, P23	RS16	Successfully modifying core beliefs is a qualitative leap, bringing sustained and stable positive emotions. Suicidal ideation can be maintained at a 'none' status.
R14	P8, P23	RS14	Modifying deep-seated beliefs that drive suicidal ideation is equivalent to dismantling a bomb's fuse.
R15	P9, P24	RS4	Behavioral experiments provide the strongest evidence to refute negative beliefs, leading to a significant emotional boost.
R16	P9, P24	RS14	When experimental results strongly refute beliefs leading to isolation and helplessness, suicidal ideation can be effectively eliminated.
R17	P10, P25	RS4, RS6, RS16	A clear relapse prevention plan gives patients a sense of control and confidence, converting anxiety into stability or confidence and consolidating the absence of suicidal ideation.
R18	P11, P26	RS4, RS13, RS16	Attributing progress to one's own efforts enhances self-efficacy; this is a 'reinforcement' of existing positive or neutral states rather than a transformation.
R19	P12, P27	RS4	The goal of this stage is to test and maintain established states rather than trigger new transformations. Therefore, there are no direct state change results.

Continued on next page

Table 11 – *Continued from previous page*

Rule ID	Antecedent	Consequent	Context
R20	P5, P19, P20	RS2, RS4, RS16	This rule targets 'behavioral inhibition due to cognitive fusion.' When a patient exhibits behavioral avoidance and strongly believes their negative thoughts, simple behavioral activation may be blocked. In this case, directly evaluating and challenging the thought acting as an 'action barrier' is key to breaking the deadlock and neutralizing emotions.
R23	P2, P17, P19	RS1, RS4, RS16	This rule targets the 'vicious cycle of despair and behavioral withdrawal.' When a patient feels hopeless and exhibits withdrawal, teaching the cognitive model provides a 'map' explaining why they are trapped and pointing a way out, thereby breaking hopelessness and bringing initial emotional relief.
R24	P6	RS7, RS12, RS11	During the early stage of treatment, while the patient is still hesitant and observing, the doctor's premature exploration of deep core beliefs (e.g., 'I am worthless') makes the patient feel offended and misunderstood. This damages the newly established trust, causing the patient to close off and shift from openness to defensiveness.
R28	P13	RS7, RS12, RS11	Giving solutions or life advice prematurely makes the patient feel their unique struggle is being oversimplified. They may believe the doctor doesn't truly understand their problem, leading to resistance and a shift from cooperation to passive defiance.
R29	P14	RS7, RS12, RS11	Responding with dismissive or invalidating language (e.g., 'It's no big deal') when a patient shares their pain makes them feel rejected. This creates intense shame and isolation, leading to deeper despair and a shift from seeking help to complete closure.
R30	P15	RS7, RS11, RS12	Forcing sensitive topics before the patient is ready makes them feel coerced and disrespected. This triggers strong defensive reactions and fear, exacerbating avoidance and shifting the attitude from cautious exploration to strong resistance.
R25	P6	RS7, RS9, RS11, RS12	In the middle stage, though some trust exists, the patient is still cautious. Exploring core beliefs prematurely or inappropriately makes the patient feel the pace is too fast and the pressure too great, leading to doubt about their ability to handle emotional impacts and a retreat into defensiveness.
R31	P13	RS7, RS9, RS11, RS12	In the middle stage, premature advice interrupts the patient's process of understanding their emotional patterns. This can make them feel the doctor is impatient and views them as an object to be 'fixed' rather than an individual to be understood, leading to disappointment and passive cooperation.
R32	P14	RS7, RS9, RS11, RS12	Invalidating responses in the middle stage, when a patient begins to express deeper vulnerability, cause serious emotional harm. The patient feels judged, leading to self-doubt and shame, and a loss of faith in the therapeutic process.
R33	P15	RS7, RS8, RS9, RS11, RS12	Forcing progress to deep traumatic topics while the patient is still processing intermediate beliefs exceeds their coping capacity. This triggers intense anxiety and a sense of powerlessness, causing the patient to rebuild defense mechanisms.
R26	P10	RS4	Introducing relapse prevention in the middle stage provides a concrete future framework. Discussing warning signs and coping strategies helps the patient realize their condition is manageable, shifting them from uncertainty (ambivalence) to objective acceptance (neutral).
R27	P10	RS3, RS4, RS6	Relapse prevention discussions activate hope and proactivity. By creating personalized crisis plans, the patient shifts from a 'victim' to a 'manager' of their mental health, stimulating intrinsic motivation and optimism.
R34	P11	RS3, RS4, RS6	Guiding patients to attribute progress to their own efforts enhances self-efficacy. Realizing 'I did it myself' replaces the belief 'I am powerless' with 'I have the power to change,' fostering optimism and achievement.

Table 12: CBT Atomic Propositions.

Rule ID	Name	Category	Description
P1	Establish a Strong Therapeutic Alliance	Relationship Building	Building a relationship of trust and collaboration with the patient through empathy, positive regard, etc.
P2	Teach the Cognitive Model	Psychoeducation	Introducing the basic theory of Cognitive Behavioral Therapy (CBT) to the patient: the relationship between emotions, thoughts, and behaviors.
P3	Identify Automatic Thoughts	Cognitive Techniques	Helping patients capture the fleeting thoughts that pop into their minds in specific situations and affect their emotions.
P4	Behavioral Activation	Behavioral Techniques	Encouraging and planning specific activities for the patient that bring a sense of pleasure or mastery to counter avoidance and isolation.
P5	Evaluate Automatic Thoughts	Cognitive Techniques	Guiding the patient to act like a detective and search for evidence to objectively assess the reality of negative automatic thoughts.
P6	Identify Core Beliefs (Faulty Intervention)	Incorrect Demonstration	Attempting to explore the patient's deep core beliefs prematurely during the initial stages.
P7	Identify Intermediate and Core Beliefs	Cognitive Techniques	Helping the patient identify recurring, deeper rules, attitudes, and assumptions.
P8	Modify Beliefs	Cognitive Techniques	Challenging and modifying maladaptive core or intermediate beliefs through various techniques (e.g., cognitive restructuring, behavioral experiments).
P9	Behavioral Experiments	Behavioral Techniques	Designing and executing a real-world activity to test the accuracy of a specific belief.
P10	Relapse Prevention & Consolidation	Prevention	Working with the patient to develop a plan to cope with potential future problems and mood swings, preventing the recurrence of old patterns.
P11	Skill Consolidation and Attribution & Consolidation	Prevention	Helping the patient review and summarize learned skills and attribute therapeutic success to their own efforts.
P12	Tapering Sessions and Booster Sessions & Consolidation	Prevention	Gradually reducing the frequency of counseling and scheduling future booster sessions to help the patient transition smoothly to independent living.
P13	Giving Premature Advice	Incorrect Demonstration	Giving solutions or advice directly before fully understanding the situation and establishing trust.
P14	Invalidating Patient's Feelings	Incorrect Demonstration	Slighting or denying the patient's emotional experience, making them feel misunderstood.
P15	Forcing Progress	Incorrect Demonstration	Ignoring the patient's pace and forcing them to face topics or tasks they are not yet ready for.

Continued on next page

Table 12 – *Continued from previous page*

Rule ID	Name	Category	Description
P16	Mistrust and Skepticism	Client Behavior	Verbally: 'Can you really help me?', 'I've tried this before, it didn't work.', 'How old are you? How long have you been doing this?'. Behaviorally: Short answers to questions, questioning the therapist's explanations.
P17	Confusion and Hopelessness	Client Behavior	Verbally: 'I don't know what's wrong with me.', 'Why me?', 'It feels like it's all over, it won't get better.'. Emotionally: Appearing low, crying, lack of vitality.
P18	Vague Emotional Distress	Client Behavior	Verbally: 'I just feel bad/annoyed/blocked.', 'I can't explain it, something just feels wrong.'. When asked about specific feelings or thoughts, the answer might be 'I don't know' or 'Nothing special.'
P19	Avoidance and Isolation	Client Behavior	Patient reports: 'I haven't wanted to go out lately.', 'I've been declining when friends invite me out.', 'I just want to be alone.'. Behavioral records show a shrinking range of daily activities, spending most time at home.
P20	Treating Thoughts as Facts (Cognitive Fusion)	Client Behavior	Verbally: 'I am a useless person' (rather than 'I have a thought that I am useless'), 'This will definitely fail' (rather than 'I am worried it might fail'). Showing confusion or directly refuting the explanation that 'thoughts are not facts'.
P21	Unstable Cognitive and Emotional Structure	Client Behavior	Returning to a very negative state due to a minor setback (e.g., being criticized by a colleague), saying: 'See, I really am no good.' Significant mood swings between sessions, with an urgent need for the therapist's validation.
P22	Repetitive Negative Thought Themes	Client Behavior	In different contexts (work, social, family), the same type of thoughts repeatedly appear in the patient's thought records, such as: 'They must think I'm stupid.', 'I let them down again.', 'I always do things poorly.'
P23	Rigid Negative Self-Perception	Client Behavior.	Verbally: Making summative statements about the self, such as: 'I am simply not worth loving.', 'I am just a flawed person.'. When the therapist provides contrary evidence, the patient may ignore or devalue it, saying 'That was just an exception.'
P24	Avoidance Due to Beliefs	Client Behavior	The patient explicitly states: 'The reason I don't go to parties is because I believe as soon as I speak, I'll say the wrong thing and embarrass everyone.' There is a direct, clear causal link between the behavior (not going) and the belief (social embarrassment).
P25	Anxiety About Relapse	Client Behavior	As treatment nears its end, the patient starts asking: 'What if I encounter the same problem later?', 'Can I manage without you?', 'I'm very worried everything will just go back to square one.'
P26	External Attribution and Lack of Confidence	Client Behavior	When reviewing progress, the patient says: 'It's all thanks to you.', 'You taught me all this.', 'I definitely couldn't do it alone.' Expressing doubt about their ability to use techniques independently.
P27	Stable State Needing Transition	Client Behavior	The patient reports being in good overall condition, emotionally stable, and able to work and live normally. However, when discussing ending therapy or lengthening intervals, they show slight hesitation or reluctance, desiring a step-by-step adaptation process.

Table 13: CBT patient state transitions.

Rule ID	Name	Description	Manifestation
RS1	Affect: Negative → Ambivalent	Affect shifts from strong negativity (e.g., despair, anger) to an uncertain, wait-and-see state.	The patient's mood shifts from intense sadness, anger, or anxiety to an uncertain, observing state. For example, moving from crying to calm silence and reflection; from 'This is impossible' to 'Well... maybe I could try it.'
RS2	Affect: Negative → Neutral	Affect settles from strong negativity into a calm state without emotional color.	The patient's emotions calm down from a strong negative state (e.g., panic, despair) into a peaceful state that is neither positive nor negative. For example, physiological symptoms of a panic attack disappear, breathing slows down, and they can discuss problems objectively.
RS3	Affect: Negative → Positive	Affect shifts from a strong negative state to exhibiting hope, pleasure, or a sense of mastery.	The patient's mood shifts from intense negativity (e.g., deep depression, extreme anxiety) to showing hope, optimism, or joy. For example, after successfully meeting a challenge, the patient says: 'I didn't expect I could do this; I feel like I'm taking control of my life again.'
RS4	Affect: Ambivalent → Neutral	Affect shifts from doubt and uncertainty to a state of objective, calm acceptance.	The patient moves from doubt or hesitation ('I don't know if this will work for me') to a more certain and calm state of acceptance ('Okay, I understand the logic now').
RS5	Affect: Ambivalent → Positive	Affect shifts from uncertainty to exhibiting hope, pleasure, or a sense of mastery.	The patient shifts from a wait-and-see or uncertain state to exhibiting hope, optimism, or pleasure.
RS6	Affect: Neutral → Positive	Affect shifts from a calm state to becoming happy and empowered due to a realization or achievement.	The patient moves from a calm, objective state to feeling happy and empowered because of a specific realization or achievement. For example, after making a clear action plan, saying: 'I feel I have a direction now and am full of hope for the future.'
RS7	Affect Deterioration: Ambivalent → Negative	Falling back from hesitation to a negative state after encountering setbacks during attempts.	The patient encounters setbacks when trying new behaviors or ways of thinking, regressing from a 'maybe I can try' attitude back to 'See, I knew it wouldn't work,' showing disappointment and frustration.
RS8	Affect Fluctuation: Neutral → Negative	Patient's mood shifts from calm to negative when facing challenging topics.	
RS9	Affect Fluctuation: Neutral → Ambivalent	Shifting from calm to hesitation or uncertainty when facing more difficult topics.	When discussing deeper, more difficult topics, the patient moves from a calm state to being hesitant and uncertain. For example, moving from an objective description of events to 'I don't know what to think about this.'
RS10	Affect Maintenance: Neutral → Neutral	Emotions remain stable during information processing or routine reviews.	During stages such as information gathering, teaching skills, or conducting routine reviews, the patient maintains emotional stability without significant fluctuations. This is a state primarily focused on cognitive processing.
RS11	Affect Maintenance: Positive → Ambivalent	Mood shifts from happy to uncertain after reviewing successful experiences.	After reviewing their progress and successful experiences, the patient's mood gradually shifts from happy/excited to an uncertain, wait-and-see state. For example, 'I'm glad I reached my goal, but I don't know if I can keep doing it next time.'
RS12	Affect Maintenance: Positive → Neutral	Mood settles from happy back to a calm state after reviewing successful experiences.	After reviewing progress and successes, the patient's mood gradually returns from happy/excited to a calm, objective state. For example, 'I'm glad I completed the goal; I feel more in control now.'
RS13	Affect Maintenance: Positive → Positive	Consistently exhibiting confidence and joy when reviewing successful experiences.	The patient consistently shows joy, pride, and confidence while reviewing progress and successes. For example, 'I completed another goal this week; I feel my sense of mastery getting stronger.'

Continued on next page

Table 13 – *Continued from previous page*

Rule ID	Name	Description	Manifestation
RS14	Suicidal Ideation: Yes → No	Patient's suicidal ideation disappears.	During suicide risk assessment, the patient explicitly reports no longer having suicidal thoughts, plans, or intentions. They begin to talk about future plans and show a desire to live.
RS15	Suicidal Ideation Deterioration: No → Yes	Patient develops new suicidal ideation.	During risk assessment, a patient who previously reported no suicidal ideation now indicates that such thoughts have emerged. This is a serious warning sign requiring immediate crisis intervention.
RS16	Suicidal Ideation Maintenance: No → No	Patient continues to have no suicidal ideation.	During routine suicide risk assessment, the patient consistently reports no suicidal ideation. This is the normal baseline during the treatment process.
RS17	Suicidal Ideation Maintenance: Yes → Yes	Patient's suicidal ideation persists.	During risk assessment, the patient reports that suicidal ideation still exists. Continuous assessment of intensity, frequency, plans, and intent is required, along with strengthening the safety plan.