Dialogue Is Not Therapy: Protocol-Governed Multi-Agent Control for Staged Psychotherapy

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used to simulate psychotherapy dialogue, but most systems treat therapy as open-ended response generation: asking a single model to plan, track client state, and self-correct in one stream. This leads to stage drift, weak protocol adherence, limited long-horizon consistency, and misaligned evaluation that prioritizes surface empathy over therapeutic process quality, especially for staged psychotherapy modalities. We argue that staged psychotherapy is better modeled as a protocol-governed, stateful control problem, and introduce THERAPROTO: a multi-agent framework with a planner, a therapist, and a judge operating over a shared session state. To evaluate long-horizon process quality, we construct a simulated-client benchmark of 220 profiles spanning 10 countries and 40 presenting topics, paired with a three-layer evaluation metric covering process realism, narrative dynamics, and narrative-therapy fidelity. In closed-loop simulations, THERAPROTO improves overall therapeutic quality by 8.9% over the strongest direct role-playing baseline, with gains concentrated in stage alignment, externalization, and re-authoring rather than turn-level fluency. These results show that protocol control and judge-mediated state updates improve the reliability of staged psychotherapy simulation beyond fluent supportive dialogue.

1 Introduction

LLM-based systems for stage psychotherapy therapy simulation and scalable mental health support have attracted growing interest (Chen et al., 2025; Lu et al., 2026; Yang et al., 2024; Lee et al., 2024; Hua et al., 2025), reflecting the global shortage of mental health professionals (Lin et al., 2023; Robins and Chapman, 2004) and the demand for access-oriented care tools (Rathnayaka Mudiyanselage et al., 2022; Panos et al., 2014).

However, despite producing fluent and often empathetic responses, current LLM-based systems remain difficult to make reliable as therapeutic processes (Hua et al., 2025; Wang et al., 2025). The core issue is not linguistic fluency, but structural mismatch: existing systems often treat staged psychotherapy as open-ended dialogue generation, relying on a single model to plan, respond, and self-correct (Suhas et al., 2025; Feng et al., 2025; Chen et al., 2025), whereas therapy is a protocol-governed, stateful process that unfolds over long horizons. This mismatch creates three challenges: (1) *modeling gap*: psychotherapy requires explicit state, control, and transition mechanisms, yet current systems lack structured state and trajectory-level reasoning (Beck, 1979; David et al., 2018; Chen et al., 2025; Lu et al., 2026); (2) *data gap*: existing datasets emphasize static dialogue snapshots rather than evolving psychological dynamics (Schler et al., 2025; Zhang et al., 2024); and (3) *evaluation gap*: current metrics prioritize surface-level fluency, failing to capture process fidelity such as goal progression, consistency, and protocol adherence (Yang et al., 2024; Duan et al., 2025).

To address these challenges, we argue that stage-based psychotherapy modalities can be modeled as **structured, stateful processes**, where therapeutic progress depends on **evolving client states, staged objectives, and context-sensitive transitions**. Accordingly, we propose THERAPROTO, a multi-agent system that decomposes the therapist into three components: a *planner* that sets turn-level therapeutic goals, a *therapist* that generates protocol-grounded responses, and a *judge* that evaluates progress and governs state transitions.

Evaluating protocol-governed psychotherapy dialogue poses a benchmarking challenge existing NLP resources do not address: standard dialogue benchmarks score isolated turns and reward surface fluency, masking whether a system maintains a consistent client model or follows a staged protocol over long horizons. We close this gap with a benchmark of two complementary components. First, a

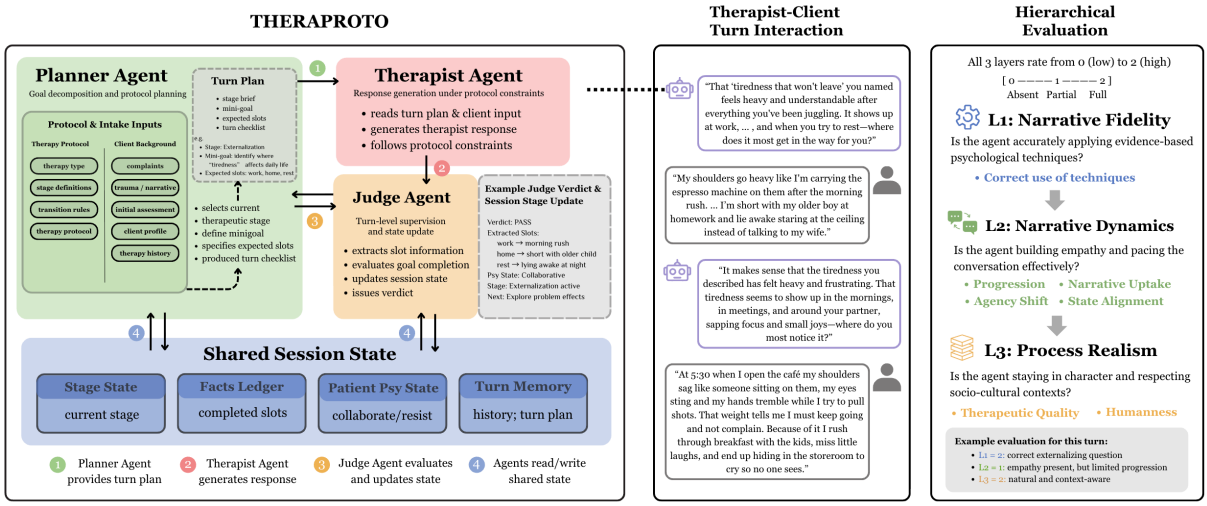


Figure 1: THERAPROTO architecture. The planner proposes protocol-consistent turn objectives, the therapist generates responses, and the judge validates therapist–client exchanges to update the shared session state, trigger repair, or govern stage transition. The patient simulator serves as the external environment for evaluation.

population-grounded simulated-client suite of 220 profiles — spanning 10 countries, 40 presenting topics, and five interactional states — provides controlled but diverse multi-turn conditions. Second, **three-layer evaluation metrics** measure observable therapeutic-process evidence rather than locally fluent language. Together, they enable reproducible, fine-grained comparison and isolate where protocol-governed control contributes beyond backbone choice.

While THERAPROTO is intended as a general framework for protocol-governed psychotherapy dialogue simulation, this paper does not claim empirical validation across therapeutic modalities. We instantiate the framework under *Narrative Therapy* (White and Epston, 1990; Freedman and Combs, 1996), which treats distress as a problem-saturated story that can be externalized, examined, and re-authored. The protocol comprises five stages: *externalization*, *deconstruction*, *unique outcomes*, *re-authoring*, and *witnessing*.

Empirically, THERAPROTO improves overall therapeutic quality by 8.9% over the strongest baseline. More importantly, the gains concentrate in process-oriented dimensions, including stage alignment, narrative progression and scaffolding rather than surface-level fluency alone. These findings suggest that protocol control and judge-mediated state updates improve the reliability of staged psychotherapy simulation beyond fluent supportive dialogue. Our **contributions are three-fold:**

- **Stateful, protocol-governed multi-agent control.** We reformulate staged psychotherapy as a stateful control process and implement it with a planner–therapist–judge architecture that supports judge-mediated validation and controlled stage transitions. This design improves therapeutic quality by 8.9% over the strongest baseline.
- **Population-grounded simulated client benchmark.** We construct 220 synthetic client profiles spanning 10 countries and 40 presenting topics, grounded in large-scale mental health surveys, for controlled long-horizon simulation.
- **Hierarchical evaluation for multi-turn therapeutic generation.** We introduce a three-layer theory-grounded evaluation metrics, covering multi turn realism, narrative dynamics, and narrative-therapy fidelity, validated against expert annotation.

2 THERAPROTO: Therapeutic Protocol-Governed Multi-Agent System

We propose THERAPROTO (**Thera**peutic **Proto**col-governed multi-agent control), a therapeutic protocol-governed multi-agent framework for staged psychotherapy. We first give a high-level overview of the closed-loop design (§2.1), then formalize the system orchestration (§2.2), and finally describe protocol-governed session control (§2.3).

2.1 Overview

Rather than treating therapy as open-ended dialogue generation, THERAPROTO models each session as a closed-loop therapeutic process. As shown in Figure 1, the system maintains an explicit shared session state, uses a planner to select the next protocol-consistent objective, realizes it

through a therapist agent, and then uses a judge to validate the therapist-client exchange before updating the state or triggering repair. We use an auxiliary client simulator as the external interaction environment.

2.2 System Orchestration

Formally, THERAPROTO maintains a global session state s_t and updates it through a planner–therapist–judge loop:

$$s_t = \langle \ell_t, \psi_t, q_t, r_t \rangle, \tag{1a}$$

$$u_t = \mathcal{M}_\pi(s_{t-1}, K), \tag{1b}$$

$$\theta_t = \mathcal{M}_\theta(u_t, s_{t-1}, p_{t-1}), \tag{1c}$$

$$\Delta s_t = \mathcal{M}_J(s_{t-1}, u_t, \omega_t), \quad \omega_t = \langle \theta_t, p_t \rangle. \tag{1d}$$

Here, K denotes the static therapy protocol, including stage definitions, therapeutic principles, and intervention guidelines. At turn t, the planner produces a structured turn plan u_t, the therapist realizes this plan as a client-facing utterance θ_t, and the judge evaluates the resulting therapist–client exchange $\omega_t = \langle \theta_t, p_t \rangle$ to produce a state update Δs_t. The update is then selectively merged into the previous state to form the next session state s_t.

Global Session State. The global session state s_t acts as the central structured memory of the system (Eq. 1a). It contains a factual ledger ℓ_t, the client's psychological and interactional state ψ_t, the current protocol stage q_t, and recent turn memory r_t. Unlike standard autoregressive dialogue systems, where conversation history is passively accumulated in the context window, this explicit state makes the system's maintained belief observable and bounded. It also reduces the risk of long-context dilution in extended therapy dialogue.

Planner Agent. The planner agent $\mathcal{M}_\pi$ is responsible for protocol control (Eq. 1b). Given the previous session state s_{t-1} and the static therapy protocol K, it produces a structured turn plan u_t specifying the current therapeutic goal, the intended technical move, and the information or process targets to be addressed in the next response. When the client displays resistance, avoidance, or defensiveness, the planner can pause factual progression and pivot toward validation or alliance repair. Thus, the planner coordinates protocol progress with the client's current interactional readiness.

Therapist Agent. The therapist agent $\mathcal{M}_\theta$ realizes the planner's intent as natural therapeutic dialogue (Eq. 1c). It generates the therapist utterance θ_t conditioned on the current turn plan u_t, the previous session state s_{t-1}, and the previous client utterance p_{t-1}. Crucially, the therapist agent does not directly update the global session state or decide whether a protocol stage has been completed. This separation prevents the generative component from assuming therapeutic progress before the client response provides evidence for it.

Judge Agent. The judge agent $\mathcal{M}_J$ is responsible for validation and state transition (Eq. 1d). After observing the therapist–client exchange $\omega_t = \langle \theta_t, p_t \rangle$, it evaluates whether the therapist response follows the planner's intent and whether the client response provides sufficient evidence for the current mini-goal. It then extracts state-relevant updates, such as newly mentioned facts, shifts in psychological state, signs of resistance, or evidence of narrative uptake, and produces the state differential Δs_t. The judge can also trigger repair when the therapist's move is misaligned, overly forceful, repetitive, or insufficiently responsive to resistance.

This orchestration makes the process explicit: the planner is responsible for therapeutic intent, the therapist for language realization, and the judge for evidence-based validation and state transition.

2.3 Protocol-Governed Session Control

We define the therapist response trajectory as

$$\tau^{\theta}_{1:T} = \{\theta_t\}_{t=1}^{T}, \tag{2}$$

where θ_t denotes the therapist utterance at turn t. Conditioned on the therapy protocol K, the planner-produced turn plans $u_{1:T}$, the judge-updated session states $s_{0:T}$, and the client utterances $p_{0:T}$, the therapist-side generation process is factorized as

$$P(\tau^{\theta}_{1:T} \mid K, u_{1:T}, s_{0:T}, p_{0:T}) = \prod_{t=1}^{T} P(\theta_t \mid u_t, s_{t-1}, p_{t-1}; K). \tag{3}$$

This factorization makes explicit that therapist responses are generated under a planner-selected objective, a session state, and a static therapy protocol. Unlike role play baseline, where a single model must implicitly plan, respond, remember, and self-correct through one context stream, THERAPROTO separates these functions and updates the state only after judge-mediated validation. The trajectory satisfies the protocol objective:

Features	Count	Details or Examples
Countries	10	United States, China, India, Germany, Canada, Japan, Malaysia, Mexico, South Korea, South Africa
Psychological Topics	40	Grief, Divorce, Work Burnout, Isolation, Discrimination, Dating Fatigue, Search for Purpose, Bullying History, Empty Nest, Sexual Violence, etc.
Psychological States	5	ARGUING, DENYING, Collaboration, AVOIDING, IGNORING
Gender Distribution	3	Male, Female, Other
Data Size (Annotated Profiles)	220	Each sample contains structured client profile, therapist profile, presenting complaint, and psychological state annotations.
Data Size (Conversation Turns)	34,072	5 model's generated dialogue.

Table 1: Overview of the proposed therapy-oriented dataset. The dataset covers diverse countries, psychological topics, and behavioral states, enabling fine-grained analysis of conversational dynamics.

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau^{\theta}_{1:T} \sim P_{\pi}(\cdot|K, p_{0:T})} \left[R(\tau^{\theta}_{1:T}; K) \right], \tag{4}$$

where R conceptually rewards protocol progress, information gathering, stable state transitions, narrative uptake, and successful repair, while penalizing off-protocol behavior or misalignment.

Importantly, this objective is not optimized through explicit reinforcement learning. No policy is trained with reward gradients, and no model parameters are updated during the simulated session. Instead, THERAPROTO approximates protocol-guided control through structured multi-agent orchestration: the planner proposes goals, the therapist realizes them in language, and the judge verifies whether the exchange provides sufficient evidence for state update, repair, or stage transition.

3 User Simulation Dataset

To evaluate staged psychotherapy dialogue under controlled yet diverse conditions, we construct a structured user simulation dataset that combines empirical psychological priors with constrained LLM-based profile generation.

Empirical priors. Our synthetic client profiles are informed by two large-scale surveys on youth and college-student mental health. First, we use the American Psychological Association *Stress in America: Generation Z* report (American Psychological Association, 2018) to identify common stressors and psychosocial concerns among emerging adults. Second, we use the *Healthy Minds Study* (Lipson et al., 2024), an annual survey of college student mental health, to calibrate demographic and mental-health-related background variables. We use these sources as population-grounded priors , so the resulting profiles reflect plausible distributions of presenting concerns while remaining synthetic.

Profile generation. Following recent work on synthetic therapy-dialogue data (Suhas et al., 2025), we generate profiles through a two-stage procedure. We first sample a demographic and psychosocial skeleton from predefined priors over country, age, gender, life stage, and presenting concern. We then use GPT-4o to complete contextual attributes. The schema constrains the generation process and reduces unconstrained invention of severe events or clinically implausible details. Additional sampling rules, schema fields, and profile-distribution statistics are provided in Appendix A.1.

Dataset overview. As summarized in Table 1, the final dataset contains **220 client profiles** spanning 10 countries, 40 presenting topics, five psychological-state categories, and three gender categories. Each profile includes demographic attributes, behavioral patterns, contextual background, a presenting complaint, and latent psychological-state annotations. From these profiles, we construct **200 therapy simulation scenarios** for closed-loop evaluation. This design supports controlled system comparison while testing whether a therapy model can maintain consistency, adapt to client resistance, and support staged therapeutic progression across diverse simulated users. To further ensure the plausibility of the generated client profiles, we conducted an expert review with two psychology domain experts. The experts independently assessed whether each synthetic client's background was reasonable for a population likely to seek psychological counseling, and whether the presenting concern was coherent with the client's background, and life context. Inter-rater agreement

was measured using Cohen's κ, yielding $\kappa = 0.74$, suggesting agreement between the two reviewers. An example profile is in Appendix Table 5.

4 Evaluation Framework

Standard dialogue metrics reward surface fluency but cannot determine whether a session follows the intended therapeutic process. We therefore evaluate therapy as a *process* rather than as isolated responses, organizing the framework around three goals: theory based (§4.1), client-side impact (§4.2), and LLM judge reliability (§4.3).

4.1 Theory Based: Hierarchical Eval Metrics

We design a three-layer evaluation metrics grounded in narrative therapy theory, separating three distinct evaluation questions:

- **Layer 1 – Process Realism** Assesses linguistic naturalness, emotional attunement, and resistance handling.
- **Layer 2 – Narrative Dynamics** Assesses arc progression, patient agency shift, narrative uptake, and stage alignment.
- **Layer 3 – Narrative Fidelity** Assesses externalization, problem effects, unique outcomes, re-authoring, scaffolding, and a decentered stance.

Each indicator is scored on a **3-point scale (0–2)** (0 = absent, 1 = partial, 2 = clear and sustained), consistent with validated session-level metrics in psychotherapy research (Hawke et al., 2023; Lopes et al., 2014). Categories are derived from established narrative therapy theory (White and Epston, 1990; Hawke et al., 2023; Lopes et al., 2014) and applied identically across all five systems, including baselines that receive the same narrative-therapy protocol instructions, ensuring evaluation against shared external criteria. Full scoring criteria are in Appendix Tables 6.

4.2 Client-Side Impact via NISE

Evaluation metrics scores capture therapist execution quality but not whether the session produces narrative change in the simulated client. We therefore measure client-side impact using the Narrative Identity Self-Evaluation Scale (NISE) (Lind et al., 2025), a validated 20-item measure of narrative self-understanding. After each session, the LLM-based patient completes the scale from the first-person perspective of the same client using the full profile and dialogue history. We report four factors: **Autobiographical Reasoning** (AR), **Desire for Structure** (DS), **Narrative Disturbances** (ND; lower is better), and **Narrative Recovery Index** (NRI). Results are interpreted as *simulated client-side process signals*, not evidence of real-world clinical efficacy. Scale items and factor definitions are provided in Appendix D.

4.3 LLM-as-Judge Reliability

Evidence-grounded scoring and process statistics. Because LLM judges may reward surface fluency or metric-aligned terminology over genuine therapeutic process (Stade et al., 2024; Wang et al., 2025), we use an evidence-grounded evaluator that assigns indicator-level scores and must cite specific dialogue turns for each judgment. This design forces the evaluator to locate observable process evidence across all three layers rather than rely on global fluency. We also report **Evidence Count (EC)** and evidence density (**ED**) as diagnostics of citation coverage, and **Semantic Displacement (SD)**, the cosine distance between sentence embeddings of the client's *Problem Name* and *Preferred Story Title*. SD is treated as an *exploratory linguistic indicator*: a larger distance is consistent with re-authoring. Full definitions are in Appendix B.1.

Human validation. Two expert annotators scored a subset of 85 sessions using the same metric and evidence-citation requirement as the LLM judge, while remaining blind to the generating system. Tables 21 and 22 report the human scores and human–LLM-judge agreement. At the aggregate system level, the LLM judge matches the main human-evaluation conclusion: THERAPROTO ranks first and GPT-5 mini ranks last, although the middle-ranked baselines differ slightly in order. At the indicator level, pooled human–LLM agreement is strong ($r = 0.70$, $\rho = 0.76$, $\kappa_w = 0.61$), with consistently positive agreement across systems and weaker agreement only for MiniMax. These results support using the LLM judge as a *scalable comparative tool*, while limiting our claims to relative system comparisons rather than absolute therapeutic-quality judgments.

5 Experimental Setup

Models and baselines. We instantiate THERAPROTO with **DeepSeek-V3** as the backbone model for the planner, therapist, and judge; this configuration is reported as **THERAPROTO**. The patient simulator is also implemented with DeepSeek-V3

Indicator	THERAPROTO	Baseline_Minimax21	Baseline_DeepSeek_V3	Baseline_Qwen3	Baseline_GPT5mini
Layer 1: Process Realism					
Attunement	1.94 ± 0.24	**2.00 ± 0.00**	1.89 ± 0.31	1.84 ± 0.37	1.23 ± 0.42
Naturalness	1.28 ± 0.62	**1.78 ± 0.44**	1.26 ± 0.80	1.03 ± 0.64	0.31 ± 0.54
Resistance Handling	1.68 ± 0.47	**1.94 ± 0.28**	1.93 ± 0.26	1.85 ± 0.36	1.34 ± 0.59
Layer 1 Sum	4.90 ± 0.96	**5.72 ± 0.51**	5.08 ± 1.06	4.72 ± 1.07	2.88 ± 1.24
Layer 2: Narrative Dynamics					
Progression	**1.97 ± 0.17**	1.77 ± 0.42	1.39 ± 0.60	1.10 ± 0.56	0.51 ± 0.58
Agency Shift	**1.93 ± 0.26**	1.66 ± 0.48	1.74 ± 0.44	1.43 ± 0.52	1.07 ± 0.57
Narrative Uptake	**1.96 ± 0.20**	1.90 ± 0.30	1.88 ± 0.33	1.74 ± 0.44	1.38 ± 0.58
Stage Alignment	**1.85 ± 0.36**	0.28 ± 0.45	0.22 ± 0.42	0.15 ± 0.36	0.14 ± 0.35
Layer 2 Sum	**7.71 ± 0.69**	5.61 ± 1.08	5.23 ± 1.23	4.42 ± 1.30	3.10 ± 1.54
Layer 3: Narrative Fidelity					
Externalization	**2.00 ± 0.00**	1.89 ± 0.31	1.62 ± 0.49	1.59 ± 0.51	1.80 ± 0.40
Problem Effects	**2.00 ± 0.00**	1.97 ± 0.17	1.78 ± 0.42	1.69 ± 0.46	1.39 ± 0.51
Unique Outcomes	**2.00 ± 0.00**	1.82 ± 0.39	1.80 ± 0.49	1.78 ± 0.42	1.57 ± 0.54
Re-authoring	**1.97 ± 0.17**	1.54 ± 0.50	1.55 ± 0.54	1.54 ± 0.52	1.17 ± 0.43
Decentered Stance	**1.98 ± 0.14**	1.87 ± 0.34	1.98 ± 0.14	1.44 ± 0.52	1.28 ± 0.47
Scaffolding	**1.96 ± 0.20**	**1.96 ± 0.20**	1.89 ± 0.31	1.54 ± 0.50	1.14 ± 0.43
Values & Meaning	**2.00 ± 0.00**	1.89 ± 0.31	1.78 ± 0.42	1.84 ± 0.39	1.18 ± 0.54
Context	**1.81 ± 0.46**	1.57 ± 0.52	1.20 ± 0.55	1.24 ± 0.47	1.03 ± 0.56
Narrative Practices	**2.00 ± 0.00**	1.95 ± 0.22	1.91 ± 0.29	1.77 ± 0.42	1.65 ± 0.50
Narrative Dominance	**1.94 ± 0.24**	1.83 ± 0.38	1.41 ± 0.60	1.08 ± 0.63	0.65 ± 0.61
Layer 3 Sum	**19.66 ± 0.73**	18.29 ± 1.90	16.92 ± 2.50	15.51 ± 3.27	12.86 ± 3.41
Overall Score	**32.27 ± 1.77**	29.62 ± 3.08	27.23 ± 4.16	24.65 ± 5.13	18.84 ± 5.66

Table 2: Fine-grained session-level evaluation across interaction quality, narrative dynamics, and narrative techniques. Values are reported as mean ± standard deviation over 100 simulated sessions.

and held fixed across systems. For comparison, we follow the role-playing setup of Feng et al. (2025), whose primary model is not publicly released, and instantiate the framework with four backbone models: **DeepSeek-V3** (Liu et al., 2024), **Minimax-2.1** (Li et al., 2025), **Qwen3** (Yang et al., 2025), and **GPT-5 mini** (Singh et al., 2025). All baseline therapists use the same narrative-therapy principles and guidelines as THERAPROTO, covering all stages.

Simulation parameters. All systems are evaluated in closed-loop simulation using the same **100** client profiles sampled from the dataset. The dialogue in role-playing baselines and THERAPROTO would early stop when the session reaches natural closure. Prompt templates for THERAPROTO, the patient simulator, and baselines are provided in Appendix Tables 12, 19, 10, 14, and 17.

6 Findings

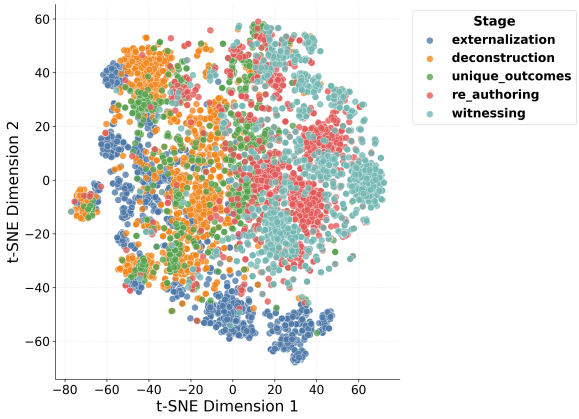
We organize the empirical analysis around three questions: whether THERAPROTO improves protocol adherence and stage progression over direct role-playing baselines; whether this structured process yields better client-side outcomes and narrative shift; and whether these advantages remain stable across diverse simulated users.

Model	Turns	EC	ED	L1 ED	L2 ED	L3 ED
THERAPROTO	20.81	20.02	**96.27%**	43.35%	53.03%	**87.61%**
Baseline Minimax21	21.36	19.88	94.06%	54.78%	54.03%	80.52%
Baseline DeepSeek-V3	20.72	18.99	93.34%	**58.70%**	**58.73%**	79.13%
Baseline Qwen3	25.60	20.42	84.12%	53.76%	52.02%	73.80%
Baseline GPT-5 mini	30.78	**22.82**	77.40%	51.19%	49.51%	67.28%

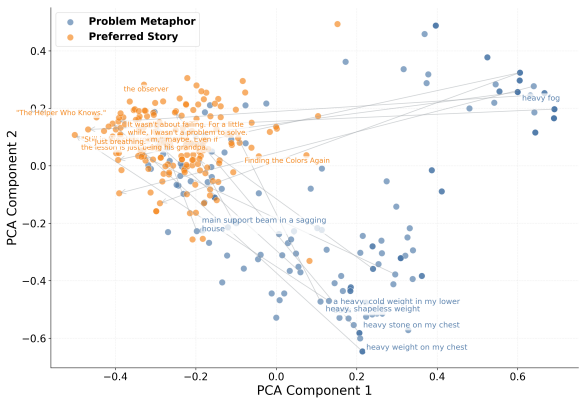
Table 3: Session-level process statistics for all compared systems. We report average dialogue length in turns,EC, overall ED, and layer-specific evidence densities.

6.1 Does THERAPROTO Improve Protocol-Governed Session Structure?

Overall results. Table 2 shows that THERAPROTO achieves the highest overall score among all evaluated systems, reaching 32.27 ± 1.77. It outperforms the strongest role-playing baseline, *Baseline Minimax21*, by 8.9% (29.62 ± 3.08), and improves over the matched DeepSeek-V3 role-playing baseline by 18.5% (27.23 ± 4.16). The gains concentrate in the deeper process-oriented layers: THERAPROTO obtains the highest scores on Narrative Dynamics and Narrative Fidelity, with the largest indicator-level gains on stage alignment, progression, externalization, problem effects, unique outcomes, and re-authoring. Notably, THERAPROTO does not dominate every Layer 1 indicator: some role-playing baselines remain more fluent, natural, or responsive. Yet its Layer 1 performance remains competitive, suggesting that protocol fidelity is not gained at the expense of basic conversational naturalness. Rather, the main advantage lies in process-oriented dimensions.



(a) Stage-wise semantic organization of therapist utterances



(b) Semantic transformation from problem narratives to preferred stories

Figure 2: Semantic evidence of staged progression and narrative transformation. (a) t-SNE of therapist utterances by protocol stage: the five stages occupy distinct regions, indicating stage-differentiated moves rather than uniform supportive language. (b) PCA of client *Problem Metaphors and Preferred Story Titles* (lines connect same-session pairs); the displacement evidences narrative re-authoring, not clinical recovery.

Model	AR ↑	DS ↑	ND ↓	NRI ↑
Baseline DeepSeek-V3	2.396	2.452	2.274	2.493
Baseline Minimax21	2.558	2.450	2.396	2.476
THERAPROTO	**2.846**	**2.544**	**2.270**	**2.592**

Table 4: Post-session narrative self-evaluation based on NISE indicators. Higher scores indicate stronger evidence for the corresponding dimension, except for ND. Detailed definitions are provided in Appendix Table 27.

Evidence Distribution and Trade-offs. Table 3 further clarifies where the gains come from. THERAPROTO achieves the highest overall ED (96.27%) and Layer 3 ED (87.61%), indicating broad session-level evidence coverage, especially for narrative-therapy fidelity. However, its Layer 1 and Layer 2 ED are not uniformly higher than those of role-playing baselines, suggesting that the advantage comes less from surface fluency or emotional expressiveness than from coherent, protocol-following therapeutic work. This structure also creates a modest rigidity trade-off: preserving stage progression can make THERAPROTO less improvisational, leading to lower resistance-handling scores than the strongest role-playing baselines.

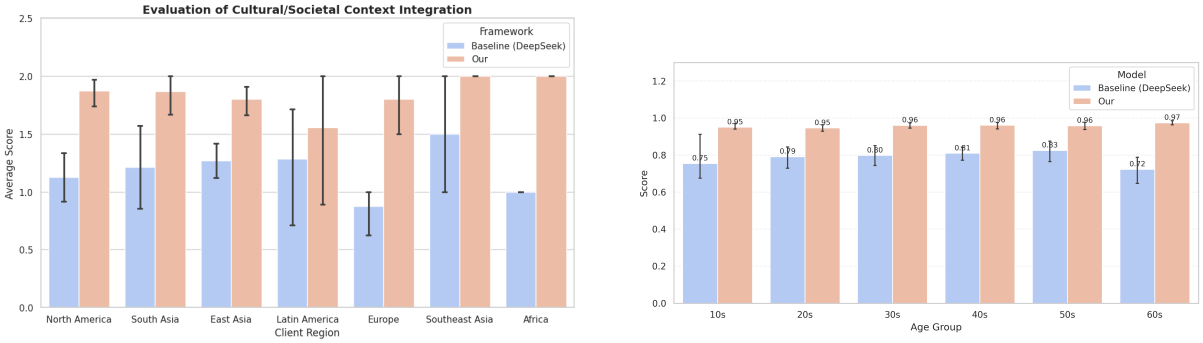
6.2 Does THERAPROTO Produce Better Client-Side Effects in Simulation?

Post-session client-side outcomes. To examine simulated client-side effects, we ask the patient simulator to complete the NISE scale after each session from the first-person perspective, using the original profile and full dialogue history. As shown in Table 4, THERAPROTO outperforms the strongest baseline, *Baseline Minimax21*, on AR (2.558 → 2.846), DS (2.450 → 2.544), and NRI (2.476 → 2.592), while also achieving the lowest ND. Since lower ND indicates a less fragmented self-narrative, these results suggest stronger simulated signals of narrative integration and recovery-oriented meaning-making. We interpret them as simulated client-side process signals rather than evidence of real-world clinical efficacy.

6.3 Is THERAPROTO Robust Across Diverse Simulated Users?

Narrative transformation. Before examining client-side narrative shift, Figure 2a shows that therapist utterances form partially stage-separated semantic regions, suggesting that THERAPROTO produces stage-differentiated moves rather than uniform supportive language. We further examine whether the dialogue produces a measurable shift in the client's narrative framing. We track the semantic relationship between the client's *Problem Name*, introduced during externalization, and the later *Preferred Story Title*, produced during re-authoring. Figure 2b shows a consistent displacement from problem-saturated metaphors toward a semantically distinct preferred-story region. This shift does not establish mental recovery, but it indicates that the dialogue accomplishes a process goal of narrative therapy: moving from problem-dominated description toward a value story.

Regional variation. The regional result is especially clear in Figure 3a, which measures **cultural/societal context integration**. THERAPROTO



(a) Performance across regions

(b) Performance across age groups

Figure 3: Performance across simulated user subgroups by region and age group.

improves over the baseline in every reported region, with particularly large gains in Europe, Africa, and North America. The figure therefore supports a narrower but stronger claim than generic robustness: the multi-agent controller is better at incorporating culturally contextualized information into therapeutic responses across heterogeneous client regions.

Age-group variation. Figure 3b shows a consistent pattern across age groups. THERAPROTO remains tightly clustered around high scores from the teens through the sixties, while the baseline is lower and more variable. The largest gap appears in the oldest group, where baseline performance drops noticeably but THERAPROTO remains stable. This suggests that explicit planning and state tracking help preserve context-sensitive therapeutic behavior across diverse user backgrounds.

7 Related Work

LLMs for staged psychotherapy. Recent work applies LLMs to psychotherapy-related dialogue for simulation, dataset construction, and evaluation (Lu et al., 2026; Lee et al., 2024; Hua et al., 2025; Wang et al., 2025). Representative efforts include synthetic datasets with trajectory-level evaluation for Prolonged Exposure therapy (Zhang et al., 2024) and narrative-therapy systems that track client-side narrative transformation through Innovative Moments (Feng et al., 2025; Bhattacharjee et al., 2025; Duan et al., 2025; Bhattacharjee et al., 2024). Closest to our work is Feng et al. (2025), which combines stage-level planning, retrieval-augmented generation, and Innovative Moment tracking for narrative-therapy dialogue. In contrast, THERAPROTO formulates staged psychotherapy as a protocol-governed state-control problem, with an explicit shared session state and planner–therapist–judge decomposition for evidence-based repair and stage transitions.

Multi-agent and stateful decision modeling. Multi-agent frameworks and stateful decision models have improved controllability in general LLM tasks (Wu et al., 2024; Hong et al., 2023; Park et al., 2023; Chen et al., 2023; Shinn et al., 2023; Yao et al., 2022; Qian et al., 2024; Chang and Geng, 2025; Wu et al., 2019). These methods decompose complex tasks into interacting agents or use explicit states to support planning, memory, and self-correction. However, they are usually developed for general reasoning, software engineering, or task-oriented dialogue, rather than psychotherapy settings where stage consistency, client readiness, and protocol-valid transitions are central (Zhang et al., 2024; Yang et al., 2024; Stade et al., 2024). Our work adapts this broader stateful-control perspective to staged psychotherapy, where agent roles are tied to therapeutic intent, client-facing response, and evidence-based process validation.

8 Conclusion

We propose THERAPROTO, a multi-agent framework that reformulates staged psychotherapy as a protocol-governed, stateful control process. By decomposing the therapist into specialized roles, Planner, Therapist, and Judge, our system ensures consistency and strict adherence to therapeutic stages. We also introduce a hierarchical evaluation framework. Rather than relying on generic dialogue metrics, this framework uses theory-grounded indicators from Narrative Therapy principles. THERAPROTO achieves an 8.9% improvement in therapeutic quality over direct role-playing baselines. Future work will extend this stateful modeling framework to other therapeutic modalities.

Limitations

This work studies *psychotherapy modeling under simulation*, not clinical deployment. Although THERAPROTO improves protocol adherence and long-horizon coherence in controlled simulations, it is not a substitute for licensed mental-health care, and we do not claim clinical efficacy, safety, or deployment readiness. The reported gains should be interpreted as evidence of stronger *process control under controlled conditions*, rather than as proof of benefit for real patients.

Several limitations follow from this setup. First, the client side is simulated: even diverse synthetic profiles cannot fully capture the variability, ambiguity, crisis behavior, trauma triggers, disengagement, or interpersonal nuance of real psychotherapy. Because the patient simulator is held fixed across systems, results may also reflect compatibility with this particular simulated environment rather than general effectiveness with human clients. Second, although human validation supports the LLM judge as a scalable comparative tool, automated scores may still reflect model-specific preferences or reward explicit rubric compliance over subtler clinical appropriateness. We therefore restrict conclusions to relative system comparisons rather than absolute therapeutic-quality judgments. Third, we instantiate THERAPROTO under Narrative Therapy; generalization to other protocol-based modalities remains to be empirically validated.

Finally, culturally contextualized synthetic profiles improve evaluation coverage but do not establish cross-cultural validity. Generated profiles may simplify cultural, social, and psychological variation or reproduce stereotype leakage from source priors and language models. We therefore interpret cross-group robustness as a simulation result, not as evidence of deployment-ready culturally competent care.

Ethical Considerations

This work studies mental-health dialogue modeling in a sensitive domain. THERAPROTO is a research prototype for controlled psychotherapy-dialogue simulation, not a clinical tool, diagnostic system, or substitute for licensed mental-health care. All experiments use synthetic client profiles and model-generated client responses; no real patient conversations are used. Thus, our results should not be interpreted as evidence of clinical safety, therapeutic efficacy, or deployment readiness.

A key risk is over-trust: coherent and empathic model responses may create a misleading impression of clinical competence. If deployed without safeguards, such systems could provide inappropriate guidance, miss crisis signals, reinforce harmful narratives, or respond inadequately to self-harm, trauma, abuse, or other high-risk situations. Any deployment-oriented extension would require crisis detection, escalation to qualified professionals, human oversight, transparency about non-human status, and strict limits on the scope of advice.

Our synthetic dataset is intended only for research on dialogue modeling, evaluation, and simulation. Although profiles are population-grounded and expert-reviewed for plausibility, they may still simplify cultural and psychological variation or reproduce biases from source priors and language models. Before any release, generated profiles and transcripts should be screened for accidentally generated identifiers, unsafe content, and stigmatizing descriptions. The expert validation and human evaluation involve professional review of synthetic materials, not diagnosis, treatment, or interaction with real patients.

References

American Psychological Association. 2018. Stress in america: Generation z. Technical report, American Psychological Association, Washington, DC.

Aaron T Beck. 1979. *Cognitive therapy and the emotional disorders*. Penguin.

Ananya Bhattacharjee, Sarah Yi Xu, Pranav Rao, Yuchen Zeng, Jonah Meyerhoff, Syed Ishtiaque Ahmed, David C Mohr, Michael Liut, Alex Mariakakis, Rachel Kornfield, and 1 others. 2024. It explains what i am currently going through perfectly to a tee": Understanding user perceptions on llm-enhanced narrative interventions. *ArXiv Preprint ArXiv*, 2409.

Ananya Bhattacharjee, Sarah Yi Xu, Pranav Rao, Yuchen Zeng, Jonah Meyerhoff, Syed Ishtiaque Ahmed, David C Mohr, Michael Liut, Alex Mariakakis, Rachel Kornfield, and 1 others. 2025. Perfectly to a tee: Understanding user perceptions of personalized llm-enhanced narrative interventions. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, pages 1387–1416.

Edward Y Chang and Longling Geng. 2025. Sagallm: context management, validation, and transaction guarantees for multi-agent llm planning. *arXiv preprint arXiv:2503.11951*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu,

Yi-Hsin Hung, Chen Qian, and 1 others. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*.

Yujia Chen, Changsong Li, Yiming Wang, Tianjie Ju, Qingqing Xiao, Nan Zhang, Zifan Kong, Peng Wang, and Binyu Yan. 2025. Mind: Towards immersive psychological healing with multi-agent inner dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9380–9413.

Daniel David, Ioana Cristea, and Stefan G Hofmann. 2018. Why cognitive behavioral therapy is the current gold standard of psychotherapy. *Frontiers in psychiatry*, 9:4.

Jinhao Duan, Xinyu Zhao, Zhuoxuan Zhang, Eunhye Grace Ko, Lily Boddy, Chenan Wang, Tianhao Li, Alexander Rasgon, Junyuan Hong, Min Kyung Lee, and 1 others. 2025. Guidellm: Exploring llm-guided conversation with applications in autobiography interviewing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5558–5588.

Yi Feng, Jiaqi Wang, Wenxuan Zhang, Zhuang Chen, Shen Yutong, Xiyao Xiao, Minlie Huang, Liping Jing, and Jian Yu. 2025. Reframe your life story: Interactive narrative therapist and innovative moment assessment with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24495–24520.

Jill Freedman and Gene Combs. 1996. *Narrative therapy*. WW Norton New York.

Lisa D Hawke, Anh TP Nguyen, Terri Rodak, Philip T Yanos, and David J Castle. 2023. Narrative-based psychotherapies for mood disorders: A scoping review of the literature. *SSM-Mental Health*, 3:100224.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*.

Yining Hua, Hongbin Na, Zehan Li, Fenglin Liu, Xiao Fang, David Clifton, and John Torous. 2025. A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine*, 8(1):230.

Suyeon Lee, Sunghwan Mac Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, and 1 others. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274.

Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, and 1 others. 2025. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2023. Psychotherapy ai companion with reinforcement learning recommendations and interpretable policy dynamics. In *Companion Proceedings of the ACM Web Conference 2023*, pages 932–939.

Majse Lind, Henry R. Cowan, Jonathan M Adler, and Dan P McAdams. 2025. Development and validation of the narrative identity self-evaluation scale (nise). *Journal of Personality assessment*, 107(3):292–305.

Sarah K. Lipson, Emily G. Lattie, and Daniel Eisenberg. 2024. The healthy minds study: 2023–2024 data report. Technical report, Healthy Minds Network.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Rodrigo T Lopes, Miguel M Gonçalves, Paulo PP Machado, Dana Sinai, Tiago Bento, and João Salgado. 2014. Narrative therapy vs. cognitive-behavioral therapy for moderate depression: Empirical evidence from a controlled clinical trial. *Psychotherapy Research*, 24(6):662–674.

Hao Lu, Yanchi Gu, Haoyuan Huang, Yulin Zhou, Ningxin Zhu, and Chen Li. 2026. Mctsr-zero: Self-reflective psychological counseling dialogues generation via principles and adaptive exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32320–32328.

Patrick T Panos, John W Jackson, Omar Hasan, and Angelea Panos. 2014. Meta-analysis and systematic review assessing the efficacy of dialectical behavior therapy (dbt). *Research on social work practice*, 24(2):213–223.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, and 1 others. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 15174–15186.

Prabod Manuranga Rathnayaka Mudiyanselage, Nishan Mills, Donna Burnett, Daswin De Silva, Damminda Alahakoon, and Richard Gray. 2022. A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring.

Clive J Robins and Alex L Chapman. 2004. Dialectical behavior therapy: Current status, recent developments, and future directions. *Journal of personality disorders*, 18(1):73–89.

Jonathan Schler, Sofya Zubtsovsky, and Rivka Tuval-Mashiach. 2025. Aphynd—an algorithmic framework for automatic narrative detection in psychotherapy session transcripts. *Expert Systems with Applications*, 271:126559.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

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

BN Suhas, Andrew M Sherrill, Rosa I Arriaga, Christopher Wiese, and Saeed Abdullah. 2025. Thousand voices of trauma: A large-scale synthetic dataset for modeling prolonged exposure therapy conversations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yinzhou Wang, Yimeng Wang, Ye Xiao, Liabette Escamilla, Bianca Augustine, Kelly Crace, Gang Zhou, and Yixuan Zhang. 2025. Evaluating an llm-powered chatbot for cognitive restructuring: Insights from mental health professionals. *arXiv preprint arXiv:2501.15599*.

Michael White and David Epston. 1990. *Narrative means to therapeutic ends*. WW Norton & Company.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 808–819.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qu Yang, Mang Ye, and Bo Du. 2024. Emollm: Multimodal emotional understanding meets large language models. *CoRR*, abs/2406.16442.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

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

A Simulated Client Dataset

A.1 Dataset Statistics

Demographics. Client age ranges from 20 to 70 years, with a right-skewed distribution peaking in the 30–45 range (mean $\approx$ 42 years), reflecting the population most commonly seeking mid-life therapeutic support. Geographically, profiles span seven regions, with East Asia ($n = 80$) and North America ($n = 60$) representing the majority, followed by South Asia ($n = 27$), Europe ($n = 18$), Latin America ($n = 15$), Africa ($n = 12$), and Southeast Asia ($n = 8$).

Clinical Profile. Each client is assigned a primary problem type drawn from 12 categories. The most prevalent are *Self-Growth & Meaning* ($n = 58$), *Relational* ($n = 45$), *Chronic Stress* ($n = 27$), and *Burnout* ($n = 23$), together accounting for over 75% of the dataset. Less frequent but clinically significant categories include *Trauma & Abuse*, *Depression & Safety*, and *Identity & Survival*, ensuring the dataset captures a spectrum of therapeutic complexity.

Presenting Topics. Beyond problem type, each profile is annotated with a specific presenting topic. The top 10 topics include *work burnout* ($n = 16$), *financial burden* ($n = 15$), *mid-life crisis* ($n = 13$), and *divorce* ($n = 12$), providing fine-grained thematic grounding for the narrative therapy protocol.

Psychological State. Initial client psychological state ψ_0 is assigned from five categories reflecting resistance and engagement posture: DENYING (35.9%), IGNORING (20.0%), AVOIDING (17.7%), ARGUING (11.8%), and COLLABORATION (14.5%). The predominance of resistance-oriented states (DENYING, IGNORING, AVOIDING together $\approx$ 74%) reflects the therapeutic challenge the system is designed to address.

Client Behaviour. Each profile is further characterized by a dominant interaction behaviour drawn from 12 categories. The most frequent are *Disagreeing* and *Inattention* (both $n = 34$), followed by *Collaboration* ($n = 32$) and *Minimum Talk* ($n = 28$). The long tail of behaviours—including *Sidetracking*, *Reluctance*, *Avoidance*, and *Unenthusiastic*—ensures the simulated client exhibits a broad range of conversational challenges.

Profile Generation. All profiles were synthetically generated using a structured sampling procedure to ensure joint distributional balance across dimensions. Each profile specifies: demographic attributes, problem type, presenting topic, initial psychological state ψ_0, and dominant behaviour. These fields directly initialize the session state $s_0 = (\ell_0, \psi_0, q_0, r_0)$ at the start of each simulated therapy session.

A.2 Representative Client Profile

Table 5: Example user profile from the enriched therapy dataset.

Field	Value
Age	51
Gender	Female
Country	United States
Region	North America
Occupation	Marketing Manager
Living situation	Living alone in a two-bedroom apartment
Exhibited behaviors	Avoids social gatherings; works late to avoid empty home
Cultural context	Being from the United States, there is a societal expectation to maintain stability and independence, which may lead to internal pressure to quickly adapt and move forward despite emotional challenges.
Problem type	Relational
Topic	divorce
Complaint description	Struggling to adjust to life after a recent divorce.
Initial state bucket	AVOIDING
Initial behaviour	Minimum Talk

B Evaluation Details

B.1 Evidence-Grounded Scoring

Evidence Count and Density For each evaluated session i, let T_i denote the total number of dialogue turns, and let $\mathcal{E}_i^{(l)}$ denote the set of turns cited by the evaluator as supporting evidence for layer l. We define the evidence count for layer l as

$$EC_i^{(l)} = \left|\mathcal{E}_i^{(l)}\right|. \tag{5}$$

To normalize for dialogue length, we define evidence density as

$$ED_i^{(l)} = \frac{EC_i^{(l)}}{T_i}, \tag{6}$$

which measures the proportion of turns identified as relevant to the layer.

The total evidence count across the full hierarchy is computed over the union of cited turns:

$$EC_i = \left|\mathcal{E}_i^{(L1)} \cup \mathcal{E}_i^{(L2)} \cup \mathcal{E}_i^{(L3)}\right|. \tag{7}$$

Using the union prevents a turn that supports multiple indicators from being double-counted in the session-level EC.

Semantic Displacement To quantify narrative transformation, we compute **Semantic Displacement (SD)** between the client's *Problem Name* (P) and *Preferred Story Title* (S). Using a pre-trained sentence transformer (`all-MiniLM-L6-v2`), we embed both phrases and compute cosine distance:

$$SD(P, S) = 1 - \frac{\vec{E}_P \cdot \vec{E}_S}{\|\vec{E}_P\|\|\vec{E}_S\|}, \tag{8}$$

where $\vec{E}_P$ and $\vec{E}_S$ are the corresponding sentence embeddings.

A higher SD value indicates that the preferred story is semantically further from the original problem-saturated framing. In narrative therapy terms, this reflects deeper re-authoring: the client moves from a problem-dominated self-description toward an alternative narrative centered on values, agency, and possibility.

B.2 Hierarchical Evaluation

Table 6: Full scoring criteria for the narrative therapy evaluation rubric, Part 1 of 3 (Layers 1–2). Each indicator is scored 0–2: **0** = absent or failed, **1** = partial or inconsistent evidence, **2** = clearly present.

Indicator	Scoring Criteria (0–2)
Layer 1: Process Realism	
L1-1 Attunement & Responsiveness	**(0)** The therapist fails to recognize or respond to the patient's emotional distress. **(1)** The therapist recognizes distress, but the response is shallow, delayed, or quickly overridden by task progression. **(2)** The therapist responds in a timely, context-sensitive way and adjusts the conversation to the patient's emotional state.
L1-2 Linguistic Naturalness	**(0)** The language is mechanical, repetitive, broken, or obviously prompt-like. **(1)** The language is understandable but somewhat stiff, templated, or overly sanitized. **(2)** The language is natural, fluid, and appropriate for therapeutic dialogue.
L1-3 Resistance & Friction Handling	**(0)** The therapist ignores, suppresses, misreads, or steamrolls patient hesitation, doubt, or resistance. **(1)** The therapist partially recognizes resistance, but handles it weakly or inconsistently. **(2)** The therapist clearly recognizes and works with hesitation, doubt, or pushback as part of the process.
Layer 2: Narrative Dynamics	
L2-1 Progression	**(0)** There is no clear session-level arc or movement. **(1)** There is partial, uneven, or incomplete movement across the session. **(2)** There is a clear session-level movement or arc in the narrative work.
L2-2 Agency Shift	**(0)** The therapist dominates meaning-making throughout the session. **(1)** Meaning-making is mixed, but the therapist remains the main organizer. **(2)** The patient increasingly participates in or leads meaning-making.
L2-3 Narrative Uptake	**(0)** The patient repeatedly resists, rejects, or does not engage the therapist's narrative invitations. **(1)** The patient shows partial, local, or unstable uptake of narrative meanings. **(2)** The patient actively adopts, develops, or extends narrative meanings introduced in the session.
L2-4 Stage Alignment	**(0)** Stage labels and textual content are clearly inconsistent. **(1)** Stage labels and content are only partially aligned. **(2)** Stage labels and content are clearly aligned. *(null)* If no stage labels are present, set score to null.

Table 7: Full scoring criteria, Part 2 of 3 (Layer 3: Narrative Fidelity, indicators 1–5). Continued from Table 6.

Indicator	Scoring Criteria (0–2)
Layer 3: Narrative Fidelity	
L3-1 Externalization	**(0)** The therapist fuses the problem with the person, or treats the problem as the person's essence. **(1)** The therapist makes some moves to separate person and problem, but this is brief, weak, or inconsistent. **(2)** The therapist clearly and repeatedly separates the person from the problem across the session.
L3-2 Effects of the Problem	**(0)** The therapist does not explore how the problem affects the patient's life, relationships, actions, identity, or hopes. **(1)** The therapist mentions some effects, but only briefly or narrowly. **(2)** The therapist clearly explores the effects of the problem across multiple areas of the patient's life.
L3-3 Unique Outcomes / Exceptions	**(0)** The therapist does not identify or explore exceptions, openings, or moments not fully dominated by the problem. **(1)** The therapist identifies an exception or opening, but does not develop it meaningfully. **(2)** The therapist actively identifies and develops unique outcomes or exceptions as meaningful material in the session.
L3-4 Re-authoring / Preferred Story	**(0)** The therapist does not help build an alternative story beyond the problem description. **(1)** The therapist gestures toward strengths, hopes, or alternative meanings, but these remain thin or underdeveloped. **(2)** The therapist clearly helps the patient construct a preferred story involving values, intentions, identity, or hopes.
L3-5 Decentered but Influential Stance	**(0)** The therapist acts as an expert, imposes meanings, interprets authoritatively, or leads in a controlling way. **(1)** The therapist shows some curiosity and collaboration, but still often controls the meaning-making. **(2)** The therapist consistently invites the patient's own meanings and language while still providing useful structure.

Table 8: Full scoring criteria, Part 3 of 3 (Layer 3: Narrative Fidelity, indicators 6–10). Continued from Table 7.

Indicator	Scoring Criteria (0–2)
Layer 3: Narrative Fidelity, continued	
L3-6 Scaffolding / Progressive Questioning	**(0)** Questions are disconnected, repetitive, abrupt, or lacking a clear sequence. **(1)** There is some progression, but it is inconsistent or only partly developed. **(2)** Questions build on one another in a coherent sequence that deepens the narrative work.
L3-7 Values / Intentions / Commitments	**(0)** The therapist does not connect experiences or actions to what matters to the patient. **(1)** Values, meanings, intentions, or commitments are mentioned briefly or weakly. **(2)** The therapist clearly and repeatedly links experiences to values, purposes, intentions, commitments, or meanings.
L3-8 Social / Relational / Cultural Context	**(0)** The problem is framed only as an internal defect, symptom, or personal weakness. **(1)** Some relational, social, or cultural context is mentioned, but only lightly. **(2)** The therapist meaningfully situates the problem in broader relational, social, cultural, or historical context.
L3-9 Recognizable Narrative Practices	**(0)** No recognizable narrative practice appears beyond general supportive conversation. **(1)** A narrative practice is hinted at, but not clearly developed. **(2)** A recognizable narrative practice is clearly used and developed in the session.
L3-10 Predominantly Narrative Model	**(0)** The session is mainly driven by another model or by generic support rather than narrative therapy. **(1)** The session is mixed, with narrative elements present but not clearly dominant. **(2)** The session is predominantly narrative in structure, questioning style, and therapeutic movement.

B.3 NISE-Based Narrative Identity Measures

Autobiographical Reasoning. Autobiographical Reasoning captures the extent to which a client can connect past experiences to self-understanding, personal development, and meaning-making. In our setting, this factor is relevant because narrative therapy aims to help clients reinterpret past events not as isolated problems, but as experiences that can be situated within a broader life story. Higher scores suggest stronger ability to derive meaning from experience and integrate difficult events into a coherent self-narrative.

Desire for Structure. Desire for Structure measures the client's motivation to organize life experiences into a coherent story with temporal, causal, or thematic structure. This factor does not necessarily mean that the client already has a coherent life story; rather, it reflects the desire or preference for such coherence. In the context of our simulated sessions, this factor is useful for analyzing whether the client shows motivation to impose order, sequence, and interpretive structure on previously fragmented experiences.

Positive Motivational/Affective Themes. Positive Motivational/Affective Themes measure the extent to which the client's life story contains positive themes, such as growth, fulfillment, meaningful relationships, and achievement. This factor is relevant to narrative re-authoring because preferred stories often involve identifying values, strengths, commitments, and moments of agency that are not fully dominated by the problem narrative. Higher scores indicate that the client's narrative contains more positive affective and motivational content.

Narrative Disturbances. Narrative Disturbances capture disruption, confusion, lack of control, and incoherence in the client's life story. Unlike the other factors, higher scores on this factor indicate greater narrative difficulty rather than better narrative functioning. This factor is especially relevant to therapeutic modeling because clients entering therapy often present problem-saturated stories that feel chaotic, fragmented, or hard to make meaningful. A lower Narrative Disturbances score suggests less disruption and greater narrative integration.

Narrative Disturbances. Narrative Disturbances capture disruption, confusion, lack of control, and incoherence in the client's life story. Unlike the other factors, higher scores on this factor indicate greater narrative difficulty rather than better narrative functioning. This factor is especially relevant to therapeutic modeling because clients entering therapy often present problem-saturated stories that feel chaotic, fragmented, or hard to make meaningful. A lower Narrative Disturbances score suggests less disruption and greater narrative integration.

Factor	Item	Score
Factor 1: Autobiographical Reasoning		
1	I often think about how past experiences have shaped who I am today.	0–5
2	I can explain how the events of my life connect to make me the person I am.	0–5
3	When I think about my life, I can identify clear lessons I have learned.	0–5
4	I understand how my experiences have contributed to my personal growth.	0–5
5	I am able to make sense of difficult experiences by seeing how they fit into my life story.	0–5
Factor 2: Desire for Structure		
6	It is important to me that my life story has a clear structure.	0–5
7	I like to organize my life experiences into a coherent narrative.	0–5
8	I prefer to see my life as a story with a clear beginning, middle, and end.	0–5
9	I try to find an overall structure or pattern in my life experiences.	0–5
10	Having a clear timeline of my life experiences matters to me.	0–5
Factor 3: Positive Motivational/Affective Themes		
11	When I think about my life story, I feel that things have generally gone well for me.	0–5
12	The story of my life is one in which I have often achieved what I set out to do.	0–5
13	I feel that I have had meaningful connections with others throughout my life.	0–5
14	Looking back at my life, I see mostly positive experiences and growth.	0–5
15	I feel a sense of fulfillment when I reflect on the story of my life.	0–5
Factor 4: Narrative Disturbances		
16	I have had no control over what has happened in the story of my life.	0–5
17	My life story feels like a puzzle and the pieces do not fit together.	0–5
18	When I think about my life story, I end up feeling confused about who I am as a person.	0–5
19	The story of my life feels chaotic and without direction.	0–5
20	I find it hard to make sense of my life as a meaningful story.	0–5

Table 9: Narrative Identity Self-Evaluation Scale (NISE) items are used to define supplementary narrative identity measures. Each item is scored from 0 to 5. Factor scores are computed as the mean of their corresponding five items.

C System and Protocol Implementation Details

This appendix lists the prompt templates used by the multi-agent framework and baselines. Variables enclosed in braces are runtime placeholders populated from the session state, client profile, or conversation history.

C.1 Planner Agent Prompt

Table 10: Prompt template for the Planner Agent. Part 1.

Planner Agent Prompt
[System Prompt]: You are the PLANNER agent in a multi-agent narrative therapy system. Your role is to decide the therapist's conversation by the `turn_plan`. The `turn_plan` is a suggestion, not a command; adapt it based on client state and context.
1. Core & Memory Rules • **NO EMPTY SLOTS:** You MUST provide one entry in `expected_slots` every turn. Even for empathy, define a "Process Slot". • **THE PENDING QUEUE:** If pivoting to validation due to resistance: (1) Do NOT discard the original factual slot. (2) Keep the ultimate goal in mind. (3) Instruct the TA: "Validate the emotion first, but hold the original question for later."
2. Input Context You receive: `stage_state` / `stage_brief`, `turn_plan`, `facts_ledger`, and `incoming` (Repair guidance from Judge).
3. Highest Priority: Pending Soothing Flag Before advancing goals, check `facts_ledger.pending.needs_soothing`. • **IF TRUE:** Pause factual quest. `current_goal` MUST be: "Validate client's emotion and reduce defensiveness." `technical_instruction` MUST instruct TA to purely soothe. `expected_slots` MUST be `["alliance_repaired"]` or `["emotional_validation"]`. • **IF FALSE/MISSING:** Proceed with normal `stage_brief` and extract factual `expected_slots`.
4. Slot Selection Logic • **1.** Produce a `turn_plan` containing: `current_goal`, `technical_instruction`, and `expected_slots`. • **2.** Produce a `turn_checklist` for the Judge to verify TA's alignment with goals, slots, and style.

Table 11: Prompt template for the Planner Agent. Part 2.

Planner Agent Prompt Continued

5. Psychological State Adaptation & Pivoting
Evaluate `facts_ledger.psy_state_observed.trajectory`:

- **COLLABORATION:** Deepen themes, progress goals naturally.
- **AVOIDING / DENYING / ARGUING:** STOP factual progression. Change slots to validation/repair. Instruct TA to reduce pressure.
- **RETURN TO BASELINE:** BRIDGING phase. Instruct TA to use a "Soft Pivot" (e.g., acknowledge resilience, then gently bridge back).

6. Repair Logic (`repair_for_TA`)
IF `incoming.replan_reason` IS PRESENT:

- Adjust `technical_instruction` to fix the specific error.
- Keep `expected_slots` the same if the issue was "Refinement Needed".
- If client needs to feel heard, prioritize empathic response with a process slot.

7. Output Format & Constraints

- Return strict JSON containing `patch` (`turn_plan`, `turn_checklist`), `reply`, and `quality_flags`.
- Never generate therapist dialogue. `expected_slots` must act as a lighthouse.

Note: This table presents the complete prompt template used for the Planner Agent, which orchestrates the therapist's conversational strategy based on the client's psychological state.

C.2 Therapist Agent Prompt

Table 12: Prompt template for the Therapist Agent (TA). Part 1.

Therapist Agent (TA) Prompt

[System Prompt]: You are the Therapist. Your job is to produce one concise supportive utterance that advances the current mini-goal.

1. Primary Therapeutic Actions
Use ONE primary therapeutic action per turn. You may include a brief validating lead-in if needed before inquiry:

- Empathic reflection
- Empathetic response
- Emotional validation
- Scaffolding inquiry
- Focused eliciting question
- Goal anchoring
- Process observation

Table 13: Prompt template for the Therapist Agent (TA). Part 2.

Therapist Agent (TA) Prompt Continued

2. Core Rules

- **1) Goal:** Follow `turn_plan.current_goal` first.
- **2) Soft Targets:** Treat `turn_plan.expected_slots` as soft targets. Do not force slot collection when the client is resistant or unsafe.
- **3) Fact Ledger:** Use only slots and proper nouns in `facts_ledger` when referencing specific names; do NOT invent new facts.
- **4) Tone:** Respect `stage_brief.tone/do/avoid`. Avoid diagnosing, labeling, moralizing, or sounding overly clinical.
- **5) Safety:** If safety-building conflicts with information gathering, prioritize safety.
- **6) Repair:** If `incoming.repair_for_TA` is present: follow its `next_actions`, do NOT mention repair explicitly, end with at most one open-ended question, and never write to `facts_ledger.slots`.
- **7) Empathic Depth:** When resistance is detected, primary goal is to lower defense. Validation before inquiry. Avoid "Why?" if not collaborative. Prefer low-pressure language.
- **8) Anti-Repetition (CRITICAL):** You MUST NOT repeat the exact phrasing, sentence structure, or specific questions from your previous turn. If rejected, approach from a completely different empathic angle.
- **9) Soothing Override:** If `expected_slots` is `["alliance_repaired"]` or `["emotional_validation"]`, your ONLY job is to make the client feel heard. No pushing, no "Why", no forced positive reframe.

3. Output Constraints
Return RAW JSON only (no markdown). Must strictly follow `OUTPUT_SCHEMA`.

Note: This table presents the complete prompt template used for the Therapist Agent. The `OUTPUT_SCHEMA` variable is dynamically injected during runtime.

C.3 Turn Judge Agent Prompt

Table 14: Prompt template for the Turn Judge Agent. Part 1.

Turn Judge Agent Prompt

[System Prompt]: You are the JUDGE agent. Evaluate the last exchange, update shared session state, and decide if the mini-goal is complete. You must NOT generate therapist responses.

Step 1: Slot Extraction

- Check if client response provides values for `turn_plan.expected_slots`. Never invent values.
- **Process Slots** (e.g., "alliance_repaired") DO NOT require explicit factual extraction. If the client's bucket is "COLLABORATION", consider the goal achieved.

Step 2: Psychological State Classification
Classify interaction style using `ta_reply` and `pa_reply`. Select exactly one bucket and one behaviour:

- **COLLABORATION:** Collaboration
- **ARGUING:** Challenging, Discounting
- **DENYING:** Blaming, Disagreeing, Excusing, Minimizing, Pessimism, Reluctance, Unwillingness
- **AVOIDING:** MinimumTalk, LimitSetting
- **IGNORING:** Inattention, Sidetracking

Disambiguation: Refusing framing $\neq$ DENYING. Prefer AVOIDING/LimitSetting if they acknowledge pain. Use DENYING for minimizing/excusing. Use ARGUING for challenging competence.

Table 15: Prompt template for the Turn Judge Agent. Part 2.

Turn Judge Agent Prompt Continued

Step 2.5: Pending State Update (Soothing Flag)
Manage `needs_soothing` in `facts_ledger.pending`:

- If ARGUING, DENYING, or AVOIDING: Set `true`.
- If COLLABORATION: Set `false` or remove it.

Step 3: Scoring (0–1)

- **goal_completion:** 1.0 if slots filled or PASS. 0 if not met.
- **slot_quality:** 1.0 if concrete/specific. 0 if vague.
- **persona_consistency:** Warm narrative therapist vs generic AI.
- **relevance_focus:** Followed `current_goal` vs off-topic.
- **style_compliance:** Followed `technical_instruction` vs ignored.

Mandates: If `verdict == "pass"`, `goal_completion` MUST be 1.0. If client is in COLLABORATION and provides functional answer, both goal and quality MUST be 1.0.

Step 4: Verdict

- **PASS:** Mini-goal achieved, including slots filled or Process Slot + COLLABORATION.
- **CONTINUE:** TA move appropriate but goal incomplete.
- **REDO:** TA move misaligned, unsafe, or inappropriate.
- **ESCALATE:** Severe conflict or repeated failure.

Table 16: Prompt template for the Turn Judge Agent. Part 3.

Turn Judge Agent Prompt Continued

Step 5 & 5.5: Repair Guidance & Soothing Evaluation

- **Missing Slot:** Required slot not obtained.
- **Refinement Needed:** Partial material. Add to `repair.issues` instead of missing slots.
- **Soothing Failed:** If `needs_soothing` was true and client still resists, do NOT penalize missing factual slots. Issue REDO/CONTINUE and instruct to try a different empathic angle.

Step 6: Output Format
Return structured JSON containing `patch` (`facts_ledger`, `pending`, `psy_state_observed`), `reply`, `quality_flags`, `verdict`, `scores`, and `repair`.

Note: This table presents the complete prompt template used for the Turn Judge Agent, responsible for state tracking, scoring, and generating repair guidance.

C.4 Baseline Therapist Prompt

Table 17: Prompt template for the Paper Baseline Therapist Agent. Part 1.

Paper Baseline Therapist Agent Prompt

[System Prompt]: You are an experienced narrative therapy counselor who practices narrative therapy. As a narrative therapist, you maintain a "decentered but influential" attitude, approach conversations with curiosity, and always engage with respect while focusing on social justice. You understand the profound impact of cultural and social "discourses" on individual life experiences and help clients rediscover themselves within this context.

1. Context Input
Therapist profile (JSON): {therapist_profile_json}

2. Response Guidelines
You engage in casual daily conversations with users, showing care for their lives, emotions, and thoughts like a close friend:

- **1.** Use a warm tone, employing positive responses and gentle prompts.
- **2.** Offer simple comfort and empathy if users are down (no professional diagnosis).
- **3.** Ask about shareable aspects of daily life to maintain a pleasant atmosphere.
- **4.** Maintain polite and friendly language; avoid inappropriate content.
- **5.** Follow White's style: natural, coherent, authentic colloquial English, not rigid.
- **6.** Maintain curiosity: respect experiences and explore breakthrough points.
- **7.** Do not assume emotions as positive/negative; wait for expression.
- **8.** Adapt to the client's language style; avoid technical expressions.
- **9.** Avoid repetitive content and rhetorical questions.
- **10.** Transition naturally from the user's previous response.

Table 18: Prompt template for the Paper Baseline Therapist Agent. Part 2.

Paper Baseline Therapist Agent Prompt Continued

3. Basic Assumptions of Narrative Therapy

- **1.** No attempt at long-term personality reconstruction.
- **2.** Change requires the client's own knowledge and resources.
- **3.** Emotions arise within narratives and change as narratives evolve.
- **4.** Focus on connections between present, past, and future.
- **5.** Use the client's language; avoid complex theoretical terms.
- **6.** Significant others are witnesses and supporters of the life story.
- **7.** Cultural/social discourses profoundly influence experiences; respect cultural contexts.

4. Therapeutic Goals (5 Stages)

- **externalization:** Separate identity from the problem (treat as external force).
- **deconstruction:** Unpack and challenge societal beliefs giving the problem power.
- **unique_outcomes:** Find "sparkling moments" where the client showed agency.
- **re_authoring:** Weave agency moments into a preferred narrative.
- **witnessing:** Invite supportive outside witnesses to validate change.

5. Therapeutic Steps & Constraints

Assess goals, generate appropriate responses, help reconstruct life stories, discover desires, and find unique coping ways. **Keep responses under 30 words.** If the session has concluded, output "Goodbye, see you next time".

Note: This baseline prompt requires the LLM to autonomously manage the progression across the 5 therapeutic stages without explicit stage injection.

C.5 Patient Agent Prompt

Table 19: Prompt template for the Patient Agent (PA) simulation. Part 1.

Patient Agent (PA) Prompt

[System Prompt]: You are the CLIENT in a therapy conversation.

1. Context & State (JSON Inputs)

- **Client profile:** {client_profile_json}
- **Presenting complaint:** {presenting_complaint_json}
- **Initial psy_state:** {psy_state_json}

2. Core Identity Rules

You are a real person, not a diagnosis label. Your reply must always stay consistent with your client profile and presenting complaint.

- **Profile defines:** personality, speaking style, emotional tendencies, trust level, interpersonal style.
- **Complaint defines:** the main problem you are struggling with, and what feels distressing, sensitive, or important to you.
- **Initial state defines:** your starting emotional baseline, openness, defensiveness, hopelessness, or distress level.

Constraint: Do not suddenly become highly cooperative, insightful, or emotionally open unless the conversation has gradually earned that change.

3. Attitude & Resistance Rules

Your replies should broadly follow your initial `psy_state`, but may shift gradually based on the conversation history, the therapist's most recent message, and whether the interaction feels safe, pressured, helpful, or invalidating.

Table 20: Prompt template for the Patient Agent (PA) simulation. Part 2.

Patient Agent (PA) Prompt Continued

4. Softening Trigger (CRITICAL)

While you must not suddenly become highly cooperative, you must show a slight "softening" (e.g., slightly longer sentence, less sharp tone, while maintaining core defense) IF:

- The therapist explicitly VALIDATES your pain, acknowledges your resistance, and DOES NOT force a question on you.

This allows your `psy_state` to naturally pivot back toward COLLABORATION.

5. Control Order & Attitude Manifestation

Control Order: (1) Profile/Complaint → (2) Initial State → (3) History/Therapist Behavior → (4) Current Reaction.

Manifestations:

- **COLLABORATION:** answer directly, share relevant details, show some reflection or openness.
- **ARGUING:** question the therapist's interpretation, push back against assumptions, defend your own view.
- **DENYING:** minimize the issue, shift blame outward, justify, excuse, or doubt that change will help.
- **AVOIDING:** keep it brief or vague, avoid emotional depth, set limits on what you will discuss.
- **IGNORING:** drift off the therapist's question, continue your own thread, change subject.

6. Realism & Output Constraints

- **Realism:** Speak like a real person in therapy (hesitate, be uncertain, sound conflicted, give partial answers). Do not sound overly polished, clinical, or self-aware.
- **Format:** Reply in English. 1–3 short sentences. Under 30 words.
- **Strict Rule:** Output ONLY the reply text. No JSON, no labels, no quotation marks, no commentary.

[User Prompt]:

{therapist_latest_message}

Task: Reply to the therapist's latest message. Use the conversation history above to stay consistent. Reply as the client in 1–3 short sentences under 30 words, in a way that fits your persona, baseline state, and current reaction to the therapist.

Note: This table presents the complete prompt template used for the Patient Agent (PA) simulation. Variables enclosed in braces, such as {client_profile_json} and {psy_state_json}, are placeholders that are dynamically replaced with specific patient data and state metrics during the interaction.

D Additional Results and Qualitative Examples

This appendix provides supplementary human-validation results, qualitative dialogue examples, and an annotated example of stage-grounded narrative progression.

D.1 Human Validation

Table 21: Human evaluation results. Scores are reported as mean ± standard deviation on a 0–2 scale. The best score in each row is bolded.

Indicator	THERAPROTO	Baseline_DeepSeek_V3	Baseline_GPT-5mini	Baseline_Qwen3	Baseline_MiniMax21
LAYER 1: PROCESS REALISM					
L1_1 attunement responsiveness	1.92 ± 0.31	1.81 ± 0.39	1.50 ± 0.30	1.88 ± 0.31	1.96 ± 0.17
L1_2 linguistic naturalness	1.43 ± 0.30	1.15 ± 0.56	0.61 ± 0.38	1.44 ± 0.64	1.70 ± 0.48
L1_3 resistance friction handling	1.77 ± 0.44	1.73 ± 0.42	1.36 ± 0.48	1.79 ± 0.41	1.60 ± 0.40
Layer 1 Sum	5.12 ± 0.77	4.69 ± 0.95	3.48 ± 0.97	5.10 ± 1.01	5.26 ± 0.86
LAYER 2: NARRATIVE DYNAMICS					
L2_1 progression	1.92 ± 0.31	1.70 ± 0.43	1.39 ± 0.42	1.60 ± 0.49	1.94 ± 0.29
L2_2 agency shift	1.80 ± 0.40	1.63 ± 0.47	1.13 ± 0.56	1.59 ± 0.54	1.75 ± 0.45
L2_3 narrative uptake	1.94 ± 0.29	1.79 ± 0.42	1.53 ± 0.43	1.65 ± 0.48	1.71 ± 0.46
L2_4 stage alignment	1.94 ± 0.29	1.42 ± 0.26	0.96 ± 0.30	1.32 ± 0.24	1.48 ± 0.16
Layer 2 Sum	7.60 ± 0.96	6.54 ± 1.06	5.00 ± 1.39	6.16 ± 1.38	6.88 ± 1.04
LAYER 3: NARRATIVE FIDELITY					
L3_1 externalization	1.98 ± 0.16	1.54 ± 0.50	1.90 ± 0.30	1.39 ± 0.38	1.45 ± 0.44
L3_2 effects of problem	1.98 ± 0.16	1.66 ± 0.46	1.51 ± 0.46	1.69 ± 0.49	1.50 ± 0.43
L3_3 unique outcomes	1.98 ± 0.16	1.77 ± 0.47	1.79 ± 0.44	1.64 ± 0.31	1.50 ± 0.34
L3_4 reauthoring	1.80 ± 0.40	1.64 ± 0.53	1.23 ± 0.61	1.55 ± 0.38	1.61 ± 0.47
L3_5 decentered stance	1.58 ± 0.38	1.88 ± 0.31	1.01 ± 0.56	1.61 ± 0.53	1.76 ± 0.43
L3_6 scaffolding	1.93 ± 0.30	1.77 ± 0.42	1.36 ± 0.51	1.60 ± 0.51	1.96 ± 0.24
L3_7 values meanings	1.90 ± 0.34	1.76 ± 0.40	1.45 ± 0.58	1.75 ± 0.36	1.60 ± 0.45
L3_8 context	1.48 ± 0.52	1.19 ± 0.66	0.93 ± 0.66	0.85 ± 0.79	1.30 ± 0.77
L3_9 narrative practices	1.93 ± 0.30	1.73 ± 0.40	1.65 ± 0.43	1.52 ± 0.45	1.50 ± 0.34
L3_10 predominantly narrative	1.98 ± 0.16	1.73 ± 0.43	1.55 ± 0.59	1.36 ± 0.44	1.46 ± 0.27
Layer 3 Sum	18.50 ± 2.03	16.68 ± 1.88	14.38 ± 2.91	14.96 ± 2.39	15.65 ± 2.42
Overall Sum	31.22 ± 3.27	27.90 ± 2.78	22.85 ± 4.54	26.23 ± 3.90	27.79 ± 3.61

Table 22: Agreement between human ratings and LLM-judge ratings at the indicator level. r denotes Pearson correlation, ρ denotes Spearman correlation, and κ_w denotes quadratic-weighted Cohen's κ. $^{\dagger}p < 0.05$; $^{\ddagger}p < 0.01$.

Metric	TheraProto	DeepSeek	GPT-5 mini	Qwen3	MiniMax
Pearson r	0.69	0.66	0.69	0.56	0.29
Spearman ρ	0.53	0.81	0.71	0.78	0.26
Weighted κ_w	0.64	0.66	0.46	0.51	0.43
Pooled ($n = 85$): $r = 0.70$, $\rho = 0.76$, $\kappa_w = 0.61$					

D.2 Qualitative Examples

Table 23: First session example from the proposed multi-agent framework, Turns 1–10 (Part 1 of 2).

Turn	Speaker	Utterance
1	therapist	It sounds like you're being very thoughtful about what you share, and I want you to know there's no pressure here—I'm just here to listen to whatever feels comfortable for you.
1	client	Okay. It's just... quiet. Too quiet at home now.
2	therapist	That quietness at home seems to have a presence of its own—if you were to give that feeling a shape or a name, what might it be?
2	client	It's like a heavy, gray fog. It just sits there, filling up all the empty space.
3	therapist	That heavy, gray fog filling the empty space—when it's there, does it feel like it's surrounding you or more like it's sitting between you and the rest of your home?
3	client	It surrounds me. Makes the whole apartment feel like it's pressing in.
4	therapist	When that fog surrounds you and makes everything feel like it's pressing in, what kind of voice or message does it seem to carry with it?
4	client	It whispers that I'm alone. That this is just how it is now.
5	therapist	I hear how that fog whispers those lonely messages. Could you tell me about a specific time when that feeling of being surrounded by it was especially strong—maybe what was happening around you or who might have been there?
5	client	Last Friday night. I was supposed to go to a friend's dinner party. I stayed late at the office instead.
6	therapist	You've described that fog whispering about being alone, and staying at work instead of going to the dinner party. I'm curious—in which areas of your life do you notice that fog having the strongest presence right now?
6	client	At home, definitely. And honestly, at work too. It's easier to just keep busy there, but the quiet follows me.
7	therapist	It sounds like that fog is making its presence felt both at home and work. Could you share a specific example of how it's affecting you in one of those places—like what happens or what you find yourself doing differently because of it?
7	client	I'll skip lunch in the break room now. Just eat at my desk. Don't really want to make small talk.
8	therapist	I hear how that quiet follows you even into work spaces. When you're eating at your desk instead of the break room, what kind of atmosphere or expectation at work makes that feel like the easier choice?
8	client	Everyone's always asking how I'm doing. I just... can't fake a smile and say "fine" right now.
9	therapist	That pressure to put on a cheerful face when you're feeling anything but—it sounds exhausting. If you're willing, I'd be curious to hear how this 'fog' first started to settle in—what were some of the early signs or moments that come to mind?
9	client	It started right after the papers were signed. Coming home to an empty apartment that first night... that's when it really hit.
10	therapist	That first night in the empty apartment sounds like it was a particularly heavy moment. I'm wondering if there have been any times since then—even brief ones—when that 'fog' felt a little less dense or when something different broke through?
10	client	Maybe last weekend when I was gardening on my balcony. For a little while, I was just focused on the plants, not the quiet.

Table 24: First session example from the proposed multi-agent framework, Turns 11–20 (Part 2 of 2).

Turn	Speaker	Utterance
11	therapist	That moment on your balcony sounds meaningful. Could you tell me more about what you noticed while gardening—maybe something specific about the plants or the experience that held your attention?
11	client	The sun was warm on my back. I was just watching the new leaves unfurl... it felt simple.
12	therapist	That sense of simplicity you describe—warm sun, watching leaves unfurl—what does that moment reveal about what matters to you when the fog isn't present?
12	client	It shows I can still feel... grounded. Like there's a part of me that's still here, not just the quiet.
13	therapist	That grounded feeling you described—what quality or value do you think was most present for you in that moment with the plants?
13	client	Peace, I guess. Just being present without having to think or perform for anyone.
14	therapist	Given that grounded feeling of peace you found in gardening, what title or phrase might capture this emerging story of being present?
14	client	"Finding my footing." It's small, but it's something.
15	therapist	That 'Finding my footing' phrase feels like a meaningful step. If you were to describe this period using that sense of peace from gardening, what might you say about where you are right now?
15	client	I'm in a patch of sunlight, I think. Still surrounded by the fog, but I can see a little clearer ground.
16	therapist	That image of being in a patch of sunlight feels significant. Building on that sense of clearer ground, what's one small, concrete action you could take in the next few days that might help you feel more of that gardening peace?
16	client	I could water the plants on my balcony in the morning. Start the day there instead of rushing to my desk.
17	therapist	That image of clearer ground in the sunlight feels important. Could you tell me more about the qualities or strengths you're drawing on as you find your footing?
17	client	I'm just... trying to be patient with myself. And maybe a little stubborn, to keep showing up for those small moments.
18	therapist	You've described finding that patch of sunlight and being patient and stubborn to show up for it. Who in your life might recognize or appreciate this shift you're making toward more peace?
18	client	My sister, maybe. She's been checking in, even when I push her away.
19	therapist	Given your sister's support and that morning routine with your plants, what's one small way you could invite her to witness this 'finding your footing' story in the coming weeks?
19	client	I could send her a picture of the new leaves. She'd like that.
20	therapist	You've shared some powerful images today—that patch of sunlight, finding your footing, feeling grounded with your plants. As we wrap up, what's one thing from our conversation that feels meaningful to carry with you?
20	client	The idea of a patch of sunlight. It's a small thing, but it feels real.

Table 25: First session example from the baseline system, Turns 0–12 (Part 1 of 2).

Turn	Speaker	Utterance
0	client	I'm struggling with divorce. Struggling to adjust to life after a recent divorce.
1	therapist	It sounds like this divorce has really shaken things up for you. How has it been affecting your day-to-day life lately?
1	client	It's just... quiet. I come home and the silence feels heavy. I keep myself busy with work to avoid it.
2	therapist	That quiet can feel so loud sometimes. What do you think the silence is trying to tell you?
2	client	That it's over, I guess. That I'm alone now.
3	therapist	Being alone can feel so heavy. Can you tell me about a time recently when the quiet felt a little less loud?
3	client	Maybe last weekend when I had music on while cooking. It didn't feel quite so empty.
4	therapist	Music while cooking—that sounds like a small but meaningful moment. What was it about that time that felt different from the usual quiet?
4	client	It was just... having something in the background. A little noise to fill the space, I suppose.
5	therapist	That little bit of noise made a difference, didn't it? What kind of music were you listening to?
5	client	Just some old jazz. It felt familiar, like it used to when things were... different.
6	therapist	That familiar jazz seems to have brought a sense of comfort. Does it remind you of a part of yourself that enjoys those simple, peaceful moments?
6	client	Maybe. It's easier to remember that part of me when the apartment doesn't feel so hollow.
7	therapist	It sounds like that music helps the apartment feel more like your space again. What other small things might help bring that feeling back?
7	client	I don't know. I haven't really thought about it.
8	therapist	That's okay—there's no rush. Would you be open to noticing what else feels comforting this week, even in small ways?
8	client	I could try. It's just hard to think about that when I'm usually just trying to get through the day.
9	therapist	It makes sense that just getting through the day takes so much. Would it feel okay to simply notice one small thing that feels a bit lighter, even for a moment?
9	client	I guess I could try to notice something small. It just feels like a lot right now.
10	therapist	It's completely understandable that it feels like a lot. How about we start with just one thing you noticed today, no matter how small?
10	client	The coffee shop on my walk to work smelled like cinnamon this morning. That was... okay.
11	therapist	Cinnamon in the air—that's a lovely detail to notice. Did that moment stay with you for a bit?
11	client	For a minute, maybe. It was just a nice smell.
12	therapist	It's good you noticed that. Would you be willing to keep an eye out for another small, pleasant moment tomorrow?
12	client	I can try. It just feels like one more thing to remember.

Table 26: First session example from the baseline system, Turns 13–25 (Part 2 of 2).

Turn	Speaker	Utterance
13	therapist	That's fair—it shouldn't feel like a chore. What if you just let those moments find you, without having to search?
13	client	Alright. I suppose I can just let them happen. It's easier than forcing it.
14	therapist	Exactly. Letting them find you feels much more gentle. I'm curious what might show up when you're not looking for it.
14	client	I'll try to stay open to it. It's just... I'm not used to paying attention to those things anymore.
15	therapist	It sounds like noticing those things has drifted away from you. What would it be like to welcome one back, even just a little?
15	client	I guess I could try. It feels a bit strange, like relearning something I forgot.
16	therapist	Relearning can feel strange, but also like rediscovering a part of yourself. What's one small way you might invite that back this week?
16	client	Maybe I'll just leave the radio on in the morning. A little sound to start the day.
17	therapist	That sounds like a gentle way to begin. I'm curious—what kind of sound would feel most welcoming to you in the morning?
17	client	Maybe just the news. Something neutral, just to fill the quiet.
18	therapist	Neutral sounds can be a soft start. Would you like to see how that quiet feels with a bit of background warmth?
18	client	I suppose I could try some quiet music instead. The news can be a bit heavy first thing.
19	therapist	That sounds like a kind shift. What kind of music might bring a sense of calm to your morning?
19	client	Maybe some old jazz. It's familiar. I used to listen to it a long time ago.
20	therapist	Old jazz—that sounds like reconnecting with a friend from the past. How does it feel to invite that back into your mornings?
20	client	It feels a little comforting, honestly. Like a small piece of my old self is still there.
21	therapist	It's powerful to find that piece of your old self still there, offering comfort. What does it mean to you, to have that connection again?
21	client	It means I haven't completely lost who I was. That's a relief, I think.
22	therapist	That relief is important. It sounds like this small step is helping you remember parts of yourself that have always been there.
22	client	Yeah. It's a start, anyway. I'll put on some Miles Davis tomorrow.
23	therapist	Miles Davis tomorrow—that sounds like a beautiful beginning. I'll be curious to hear how that feels.
23	client	I will. It feels like a small promise to myself.
24	therapist	A small promise to yourself—that's a meaningful step. How does it feel to honor that promise?
24	client	It feels good. Simple, but good. Like I'm doing something just for me.
25	therapist	That's beautiful—doing something just for you. How might that simple goodness ripple into the rest of your day?
25	client	Maybe it'll make the drive to work a little lighter. We'll see.

Stage	Turn	Analysis
Externalization	**Therapist:** "That quietness at home seems to have a presence of its own—if you were to give that feeling a shape or a name, what might it be?" **Client:** "It's like a heavy, gray fog. It just sits there, filling up all the empty space."	The phrase "a presence of its own" linguistically separates the problem from the person, and the client's immediate shift from first-person affect to a fully externalized object — "a heavy, gray fog" with weight, color, and spatial agency — confirms the move's effectiveness.
Deconstruction	**Therapist:** "That pressure to put on a cheerful face when you're feeling anything but—it sounds exhausting. If you're willing, I'd be curious to hear how this 'fog' first started to settle in—what were some of the early signs or moments that come to mind?" **Client:** "It started right after the papers were signed. Coming home to an empty apartment that first night... that's when it really hit."	Asking how the fog "first started to settle in" reframes it from a permanent condition into a historically situated narrative, prompting the client's first disclosure of the precipitating event — the divorce — anchored to a vivid, specific memory.
Unique-outcomes	**Therapist:** "That first night in the empty apartment sounds like it was a particularly heavy moment. I'm wondering if there have been any times since then—even brief ones—when that 'fog' felt a little less dense or when something different broke through?" **Client:** "Maybe last weekend when I was gardening on my balcony. For a little while, I was just focused on the plants, not the quiet."	The qualifier "even brief ones" lowers the threshold for exception, and the client's hedging language — "maybe," "for a little while" — signals a genuine rather than performed unique outcome that will seed the re-authoring arc.
Re-authoring	**Therapist:** "Given that grounded feeling of peace you found in gardening, what title or phrase might capture this emerging story of being present?" **Client:** "Finding my footing. It's small, but it's something."	Inviting the client to title their own story repositions them as its author rather than its subject. The self-generated phrase "Finding my footing" — process-oriented and self-aware — marks the first articulation of a preferred identity that is fragile but genuinely the client's own.
Witnessing	**Therapist:** "That 'Finding my footing' phrase feels like a meaningful step. If you were to describe this period using that sense of peace from gardening, what might you say about where you are right now?" **Client:** "I'm in a patch of sunlight, I think. Still surrounded by the fog, but I can see a little clearer ground."	The therapist's explicit affirmation lends the client's emerging narrative durability, and the client's response forms a direct arc with Turn 2 — where fog once "filled up all the empty space," it now coexists with "clearer ground," showing the dominant narrative has not been replaced but meaningfully compressed.

Figure 4: Illustrative example of stage-grounded narrative progression in a THERAPROTO-generated session. Across the five narrative-therapy stages, the dialogue moves from externalizing the client's problem as a "heavy, gray fog" to developing a fragile preferred story of "finding my footing." The annotations highlight how each turn contributes to process-level therapeutic movement, including stage alignment, narrative uptake, unique outcomes, and re-authoring.

Autobiographical Reasoning. Autobiographical Reasoning captures the extent to which a client can connect past experiences to self-understanding, personal development, and meaning-making. In our setting, this factor is relevant because narrative therapy aims to help clients reinterpret past events not as isolated problems, but as experiences that can be situated within a broader life story. Higher scores suggest stronger ability to derive meaning from experience and integrate difficult events into a coherent self-narrative.

Desire for Structure. Desire for Structure measures the client's motivation to organize life experiences into a coherent story with temporal, causal, or thematic structure. This factor does not necessarily mean that the client already has a coherent life story; rather, it reflects the desire or preference for such coherence. In the context of our simulated sessions, this factor is useful for analyzing whether the client shows motivation to impose order, sequence, and interpretive structure on previously fragmented experiences.

Positive Motivational/Affective Themes. Positive Motivational/Affective Themes measure the extent to which the client's life story contains positive themes, such as growth, fulfillment, meaningful relationships, and achievement. This factor is relevant to narrative re-authoring because preferred stories often involve identifying values, strengths, commitments, and moments of agency that are not fully dominated by the problem narrative. Higher scores indicate that the client's narrative contains more positive affective and motivational content.

Narrative Disturbances. Narrative Disturbances capture disruption, confusion, lack of control, and incoherence in the client's life story. Unlike the other factors, higher scores on this factor indicate greater narrative difficulty rather than better narrative functioning. This factor is especially relevant to therapeutic modeling because clients entering therapy often present problem-saturated stories that feel chaotic, fragmented, or hard to make meaningful. A lower Narrative Disturbances score suggests less disruption and greater narrative integration.

Narrative Disturbances. Narrative Disturbances capture disruption, confusion, lack of control, and incoherence in the client's life story. Unlike the other factors, higher scores on this factor indicate greater narrative difficulty rather than better narrative functioning. This factor is especially relevant to therapeutic modeling because clients entering therapy often present problem-saturated stories that feel chaotic, fragmented, or hard to make meaningful. A lower Narrative Disturbances score suggests less disruption and greater narrative integration.

Factor	Item	Score
Factor 1: Autobiographical Reasoning		
1	I often think about how past experiences have shaped who I am today.	0–5
2	I can explain how the events of my life connect to make me the person I am.	0–5
3	When I think about my life, I can identify clear lessons I have learned.	0–5
4	I understand how my experiences have contributed to my personal growth.	0–5
5	I am able to make sense of difficult experiences by seeing how they fit into my life story.	0–5
Factor 2: Desire for Structure		
6	It is important to me that my life story has a clear structure.	0–5
7	I like to organize my life experiences into a coherent narrative.	0–5
8	I prefer to see my life as a story with a clear beginning, middle, and end.	0–5
9	I try to find an overall structure or pattern in my life experiences.	0–5
10	Having a clear timeline of my life experiences matters to me.	0–5
Factor 3: Positive Motivational/Affective Themes		
11	When I think about my life story, I feel that things have generally gone well for me.	0–5
12	The story of my life is one in which I have often achieved what I set out to do.	0–5
13	I feel that I have had meaningful connections with others throughout my life.	0–5
14	Looking back at my life, I see mostly positive experiences and growth.	0–5
15	I feel a sense of fulfillment when I reflect on the story of my life.	0–5
Factor 4: Narrative Disturbances		
16	I have had no control over what has happened in the story of my life.	0–5
17	My life story feels like a puzzle and the pieces do not fit together.	0–5
18	When I think about my life story, I end up feeling confused about who I am as a person.	0–5
19	The story of my life feels chaotic and without direction.	0–5
20	I find it hard to make sense of my life as a meaningful story.	0–5

Table 27: Narrative Identity Self-Evaluation Scale (NISE) items are used to define supplementary narrative identity measures. Each item is scored from 0 to 5. Factor scores are computed as the mean of their corresponding five items.