UNFOLD: PATIENT-STATE PREDICTION AND PROGRESSIVE PROFILE DISCLOSURE FOR LONG-HORIZON PATIENT SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon patient simulation requires more than profile consistency: a simulated patient may remain factually faithful while exhibiting implausible behavioral changes or revealing personal information prematurely, repeatedly, or excessively. We frame these trajectory-level failures as two coupled temporal controls: patient-state prediction, which models how attitude, emotion, and topic evolve with interaction, and progressive profile disclosure, which controls when patient-specific information becomes available. We introduce UNFOLD, a psychotherapy patient simulator that predicts the next behavioral state from structured interaction history and controls access to independently addressable profile evidence before response generation. On 8,397 held-out patient turns, UNFOLD improves attitude, emotion, and topic prediction over a supervised text-encoder reference. Across 30 matched long-horizon simulations, it produces broader, more gradual, and less repetitive information trajectories than full-profile simulators, with differences persisting under longer baseline horizons. Mechanism-targeted ablations isolate the roles of state control, progressive selection, and the disclosure bottleneck, while blinded therapist raters prefer UNFOLD overall in pairwise evaluation. Together, these results suggest that long-horizon simulation should treat both behavioral state and information availability as explicit time-varying controls, a principle that may extend to interactive systems combining evolving behavior with persistent information.

1 INTRODUCTION

Large language models are increasingly studied as interactive agents, making reliable evaluation and training over sustained interactions increasingly important. This is particularly relevant in psychotherapy-oriented dialogue, where patient simulators provide controlled, repeatable interactions for training and evaluating therapeutic agents (Wang et al., 2024; Kyung et al., 2025; Wang et al., 2025). Benchmarks such as COUNSELBENCH nevertheless show that strong LLMs still exhibit clinically meaningful failures, including overgeneralization, limited personalization, and unsafe advice, motivating more controlled and interactive evaluation (Li et al., 2026). Useful patient simulation therefore requires more than fluent, profile-consistent responses: the patient must remain plausible over an extended interaction, where plausibility is a property of the *trajectory*, not only of individual responses.

Unlike one-shot generation, a simulated patient must determine how its behavioral state evolves and when patient-specific information becomes relevant over time. Errors in either process can accumulate: behavior may shift without sufficient interactional support, while valid profile information may appear too early, too densely, or repeatedly. Psychotherapy provides a particularly clear setting in which these challenges arise: patient affect evolves with interaction (Soma et al., 2020), clients actively shape therapeutic dialogue (Massfeller & Strong, 2012), and personal information is disclosed selectively (Farber, 2003; Pawelczyk & Erskine, 2008). Existing patient simulators provide substantial behavioral control, while role-playing and memory systems increasingly regulate which stored persona information is retrieved or enacted (Wang et al., 2024; Louie et al., 2024; Kyung et al., 2025; Wang et al., 2026). However, long-horizon patient simulation requires these processes to be coordinated while preserving their distinct temporal roles.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

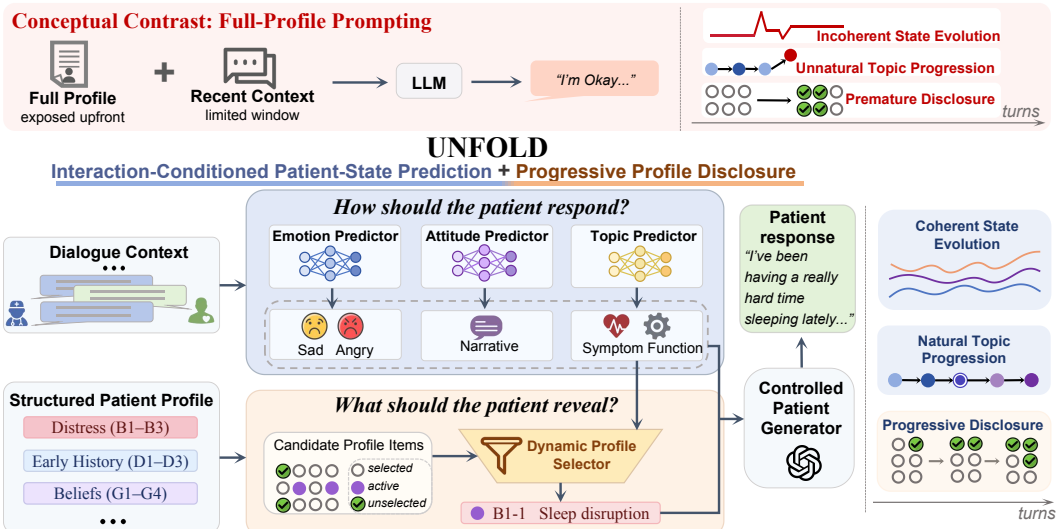


Figure 1: **Overview of UNFOLD.** Top: full-profile prompting leaves behavioral evolution and information release implicit in generation. Bottom: UNFOLD explicitly separates patient-state prediction from progressive profile disclosure before response generation.

We formulate this challenge as the control of two coupled trajectories: *behavioral evolution*, which determines how the patient's behavioral state changes with interaction, and *information availability*, which determines what persistent patient-specific evidence can become available to generation at each turn. Behavioral-state control does not determine when particular evidence should become available, while profile consistency does not constrain when or how densely that evidence is introduced. Separating the two makes temporal plausibility explicit rather than leaving both decisions implicit in generation.

As illustrated in Figure 1, we introduce **UNFOLD**, a long-horizon patient simulation framework that separates these controls before response generation. An interaction-conditioned predictor estimates the next attitude, emotion, and topic from recent structured interaction history and the therapist's current action. Persistent patient information is represented as independently addressable evidence, and a history-aware selection policy determines which evidence is prioritized at each turn. Predicted patient topic connects the two controls, allowing selection to follow either the therapist's focus or a plausible patient continuation or shift. A controlled language model then realizes the predicted behavioral state and available profile context as the next response. Importantly, selection controls *availability rather than forced verbalization*: prioritized evidence may be available to the generator without necessarily being expressed. Because session completion depends on profile coverage rather than a fixed turn budget, the simulation horizon can adapt to the progression of patient-specific information.

We evaluate UNFOLD from complementary perspectives. We first test whether its interaction-conditioned behavioral dynamics can predict structured next-state transitions on 8,397 held-out patient turns. We then compare complete long-horizon simulation pipelines on matched profiles using measures of information depth, breadth, disclosure restraint, repetition, first-seen timing, and same-group concentration, including an extended 200-turn setting. Mechanism-targeted ablations remove behavioral dynamics, information selection, and the information bottleneck to test whether each component produces the trajectory changes it is designed to control, while blinded therapist evaluations assess whether the resulting simulations exhibit more plausible patient behavior and information progression. Together, these experiments test whether explicitly modeling behavioral evolution and information availability as separate but coupled temporal processes enables more controlled long-horizon patient simulation.

Broader impact. Although our experiments focus on psychotherapy-oriented patient simulation, the underlying challenge is broader: long-horizon generative systems must maintain coherent state evo-

lution while controlling the temporal availability of persistent information. This distinction arises in interactive agents with long-term memory, generative simulators requiring trajectory-level consistency, and adaptive systems that condition generation on information accumulated over extended interactions. In these settings, indiscriminate retrieval may produce locally plausible responses while undermining global trajectory consistency. Thus, long-horizon generation requires not only memory and response quality, but also *temporal control over state and information*, which our decomposition makes explicit.

2 Related Work

Prior work relevant to UNFOLD spans two complementary problems: controlling *how* simulated users behave and *what* stored information is available to generation. UNFOLD connects them within sessions by separately modeling patient-state evolution and progressive information disclosure.

Behavioral control in patient and user simulation. PATIENT-ψ programs therapy patients with CBT-based cognitive models, while Roleplay-doh elicits expert-defined behavioral principles for counseling role-play (Wang et al., 2024; Louie et al., 2024). PatientSim combines clinical profiles with persona attributes, and EvoPatient uses agent coevolution for standardized-patient simulation (Kyung et al., 2025; Du et al., 2025). Related work also infers implicit user profiles or studies interactive and simulator-based multi-turn evaluation (Wang et al., 2025; Chang et al., 2025; Dou et al., 2025). These works establish the value of structured behavioral control. UNFOLD instead predicts turn-varying attitude, emotion, and topic while separating behavioral control from patient-information access.

Profile access and memory control. Prior systems selectively organize and retrieve persona or user information through persona extraction, adaptive memory, temporal or causal structure, Persona Graphs, and evolving PersonaTrees (Li et al., 2025; Tan et al., 2025; Ong et al., 2025; Cai et al., 2026; Zhao et al., 2026). Most directly, Memory-Driven Role-Playing studies persona-knowledge anchoring, selection, bounding, and enactment from dialogue context (Wang et al., 2026). UNFOLD does not claim novelty for separating knowledge selection from language generation. Instead, it treats the *within-session disclosure trajectory* as a control target, modeling when patient-specific evidence becomes available and coupling this process with explicit next-state prediction. Behavioral realization and information trajectories are evaluated separately.

Temporal structure of psychotherapy. Psychotherapy research motivates both controls. Therapist and client emotional arousal exhibits temporal interdependence (Soma et al., 2020), and clients actively shape dialogue through responses, corrections, and initiatives (Massfeller & Strong, 2012). Self-disclosure is selective and related to therapeutic context and time (Farber, 2003), while interactional analyses show how personal disclosure is locally elicited in therapist–client exchange (Pawelczyk & Erskine, 2008). Together, these findings motivate treating behavioral evolution and information revelation as distinct but interaction-dependent temporal processes.

3 Data and Supervision

We use the licensed Alexander Street *Counseling and Psychotherapy Transcripts*, Volumes I and II (Alexander Street Press, 2023a;b), accessed through an institutional library subscription. After preprocessing, cleaning, and filtering, the corpus contains 983 psychotherapy dialogues. We split the data at the dialogue level, yielding 786 training dialogues (46,951 patient turns), 98 validation dialogues (6,949 patient turns), and 99 held-out test dialogues (8,397 patient turns). Sessions attributable to the same client were additionally verified not to cross partitions. Validation data are used for model selection and threshold tuning, while the test set is reserved for final evaluation. The licensed source transcripts are not redistributed.

We construct a unified supervision schema for long-horizon English psychotherapy dialogue under three design principles: *textual observability*, so labels can be grounded in expressed dialogue; *temporal relevance*, so patient states can vary across interaction; and *controllability*, so the resulting variables can guide simulation. Patient behavior is represented by attitude, emotion, and clinical topic; therapist behavior by topic and strategy; and persistent patient information by a hierarchical 27-group profile representation. At the control level used for patient-state prediction, attitude,

emotion, and topic contain 8, 7, and 7 labels, respectively, and all three are multi-label variables. These inventories are task-specific operationalizations rather than diagnostic instruments or direct reproductions of existing taxonomies. Patient-state and therapist-action supervision is generated with Qwen2.5-72B-Instruct (Qwen et al., 2025) under fixed operational definitions and manually audited against the same criteria. Construct grounding, supervision inventories and hierarchy summaries, annotation procedure, and audit results are provided in Appendix B; complete fine-grained definitions and mappings are provided in the anonymous code release.

4 Problem Setup

Consider a therapeutic dialogue at patient turn t. Let h_t denote the interaction history, a_t the current therapist action, $P = \{p_i\}_{i=1}^{N}$ a structured profile of independently addressable patient-specific evidence, and $D_{<t} = (z_1, \ldots, z_{t-1})$ the history of evidence selected at previous patient turns. A conventional profile-conditioned simulator directly models $p(x_t \mid h_t, a_t, P)$, leaving behavioral evolution and information availability implicit in language generation. We instead introduce explicit variables for both:

$$\hat{s}_t = \Gamma(F_\theta(h_t, a_t)), \tag{1}$$

$$C_t = C_t^T \cup \hat{C}_t^P, \tag{2}$$

$$z_t \sim \pi_{\text{disc}}(z_t \mid C_t, P, D_{<t}), \tag{3}$$

$$\widetilde{P}_t = \{z_t\} \cup B_t, \tag{4}$$

$$x_t \sim G(x_t \mid h_t, a_t, \hat{s}_t, \widetilde{P}_t, r_t). \tag{5}$$

Here, $\hat{s}_t$ contains predicted attitude, emotion, and topic controls; C_t^T and $\hat{C}_t^P$ are the current therapist and predicted patient topic codes; z_t is the profile evidence prioritized for the current patient turn; B_t denotes optional bridge evidence supplied for local continuity; $\widetilde{P}_t$ is the resulting disclosure-controlled profile context; r_t is the profile-based ending signal; and Γ converts predicted state probabilities into simulation controls. Information already established in the dialogue remains available through h_t, while $\widetilde{P}_t$ controls newly supplied profile context under the standard selection path.

5 Method

UNFOLD instantiates Eqs. 1–5 with an interaction-conditioned patient-state predictor and a stateful disclosure policy over structured profile evidence.

5.1 Patient-State Prediction

We represent the patient's next behavioral state through attitude, emotion, and topic, capturing interactional stance, affective state and intensity, and clinical focus, respectively. Given therapist action a_t (topic and strategy) and recent interaction history h_t, UNFOLD constructs $e_t \in \mathbb{R}^{d_{\text{exo}}}$ from structured recent patient and therapist states and the current therapist action. Each historical feature sequence is summarized by its latest value, temporal mean, and mean activity indicator. The predictor uses no raw dialogue text, and persistent patient-specific information is excluded from e_t.

Figure 2 shows the predictor. Let $\mathcal{C}$ contain nine latent factors: seven topic factors, one attitude factor, and one emotion factor. Each $c \in \mathcal{C}$ has a globally learned categorical prior $q_c^0 = \text{softmax}(\theta_c) \in \Delta^{S_c - 1}$. For target factor c, the remaining priors are projected into its latent space and mean-pooled:

$$\bar{c}^{(-c)} = \frac{1}{|\mathcal{C}| - 1} \sum_{j \in \mathcal{C} \setminus \{c\}} \left(W_{j \to c} q_j^0 + b_{j \to c} \right). \tag{6}$$

The turn-specific latent belief is

$$q_{t,c} = \text{softmax}\left(f_c([q_c^0; \bar{c}^{(-c)}; e_t]) \right), \tag{7}$$

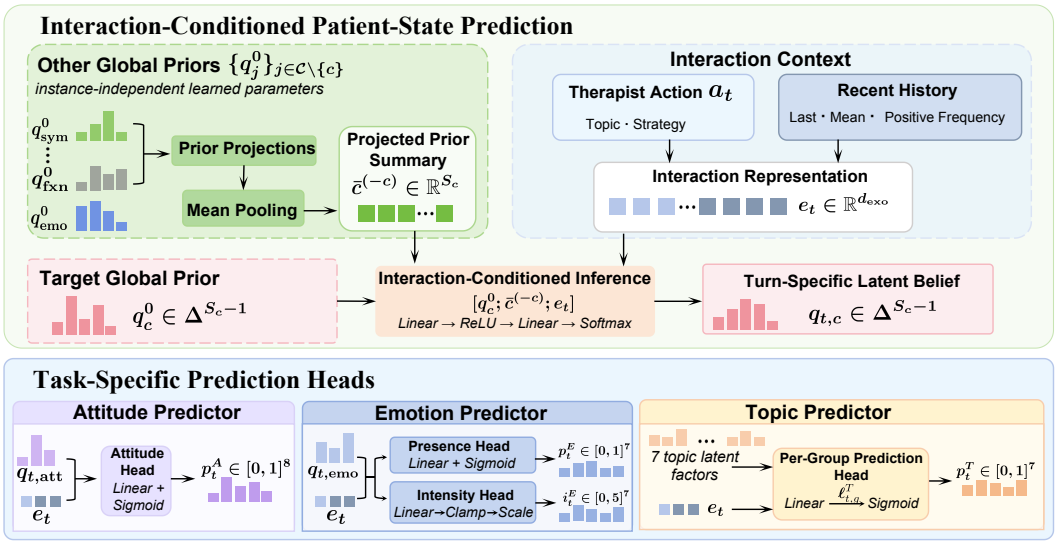


Figure 2: **Interaction-conditioned patient-state prediction.** Recent structured interaction history and the current therapist action form e_t; shared learned priors and projected summaries produce turn-specific latent beliefs $q_{t,c}$, which are mapped to attitude, emotion, and topic.

where f_c is a two-layer MLP. The priors and projected summaries are shared across dialogue instances, while turn-specific information enters through e_t; they therefore provide a structured parameterization rather than instance-specific cross-factor communication.

Task-specific heads combine the inferred beliefs with e_t to produce attitude probabilities $\mathbf{p}_t^A \in [0,1]^8$, emotion-presence probabilities $\mathbf{p}_t^E \in [0,1]^7$ with intensities $\mathbf{i}_t^E \in [0,5]^7$, and topic probabilities $\mathbf{p}_t^T \in [0,1]^7$. The tasks are trained separately using focal binary cross-entropy for attitude, positive-class-weighted binary cross-entropy for topic, and focal binary cross-entropy plus equally weighted intensity regression for emotion. Models use AdamW with validation-based early stopping and validation-selected decision thresholds. The thresholded predictions form $\hat{s}_t$, with predicted topic supplying $\hat{C}_t^P$ to the disclosure policy. Additional implementation details are provided in Appendix C.1.

5.2 Structured Profile Memory

Progressive disclosure requires independently addressable patient information. We represent persistent, transcript-grounded information as

$$P = \{p_i\}_{i=1}^N, \qquad p_i = (\text{pid}_i, g_i, \text{text}_i), \tag{8}$$

where pid_i identifies an evidence unit, g_i its profile group, and text_i the corresponding patient fact or stable pattern. Profile groups organize related information, while evidence IDs identify independently selectable details; both are distinct from the clinical-topic codes used by the state predictor and semantic-focus interface.

Profiles are constructed offline under a fixed 27-group schema. Clinical-topic labels route local transcript units to relevant profile sections, after which evidence is extracted, merged, deduplicated, and refined while retaining transcript grounding. Because evidence remains independently addressable, persistent profile possession does not imply exposing the full profile at every turn. The complete profile is a fixed patient specification rather than a target inferred from partial dialogue. Construction details and a synthetic example are provided in Appendix C.2.

5.3 Progressive Profile Disclosure

At turn t, selection history $D_{<t}$ tracks disclosure state, while the semantic focus C_t from Eq. 2 combines the therapist's current topic with the predicted patient topic. The former anchors selection to the ongoing interaction; the latter allows patient continuation or topic shift to influence profile access.

Candidate evidence is organized into *deep*, *similar*, and *distant* regions for local continuation, related movement, and broader transitions. Within the selected pool $\mathcal{P}_t \subseteq P$, candidates are scored by topic relevance and selection-history signals for continuity, relation, novelty, coverage, and saturation. All eligible non-index profile partitions can enter the reported dynamic pool, so topic compatibility is a soft relevance signal rather than a hard access boundary. Scoring coefficients are fixed heuristic policy parameters, and consecutive primary reuse is down-weighted after short local runs.

Primary evidence is sampled as

$$\pi_{\text{disc}}\left(z_t = p \mid \mathcal{P}_t, C_t, D_{<t}\right) = \frac{\exp(u_t(p)/\tau)}{\sum_{p' \in \mathcal{P}_t} \exp(u_t(p')/\tau)}. \tag{9}$$

The selected z_t is the newly prioritized disclosure target. Restricted bridge evidence may support local continuity, so the bottleneck constrains newly prioritized evidence rather than requiring generator-visible context to contain exactly one item. Selection state records primary-selection decisions, not assumed expression; actual evidence expression is measured separately during evaluation. Additional disclosure-policy details are provided in Appendix C.3.

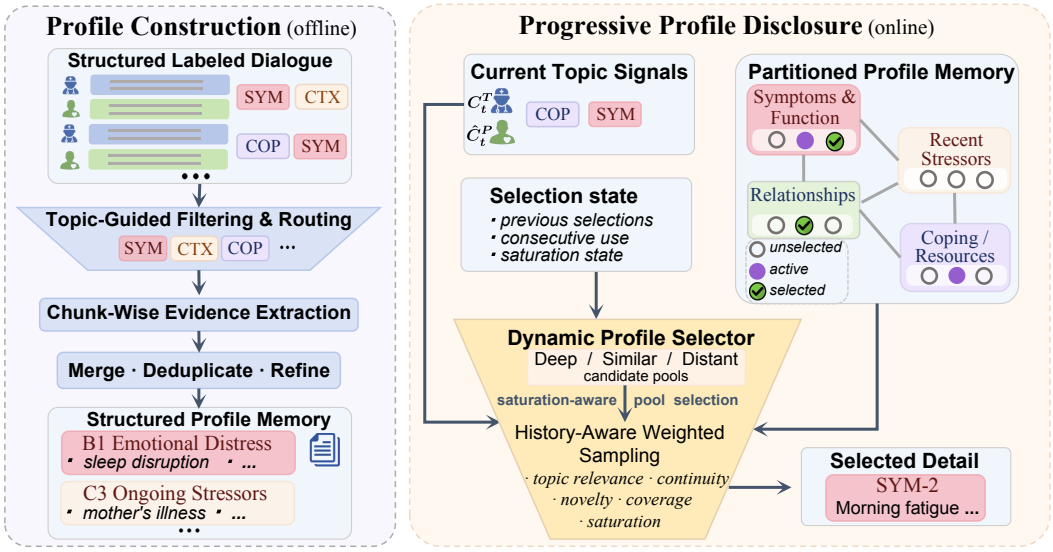


Figure 3: **Structured profile memory and progressive disclosure.** Transcript-grounded information is organized as independently addressable profile evidence. Therapist and predicted patient topics provide semantic focus, while selection history guides movement among deep, similar, and distant candidate regions to select primary evidence z_t.

5.4 Controlled Patient Generation

The generator realizes Eq. 5 with GPT-4o-mini. At each patient turn, it receives the five most recent therapist–patient exchange pairs, predicted behavioral control $\hat{s}_t$ and its label definitions, disclosure-controlled profile context, and a profile-based ending signal; gold next-patient states and future dialogue are never provided. The prompt instructs the model to follow the predicted state, prioritize the primary evidence, and avoid mechanical profile recitation, while bridge evidence serves only local coherence when needed. The generated response is appended to dialogue history, so subsequent state prediction and profile selection operate on the interaction produced thus far.

UNFOLD imposes no fixed maximum number of patient turns. The ending signal activates once every available non-index profile group has appeared in the primary-selection history, after which the generator may produce a natural closing response. Termination therefore depends on selector coverage rather than post-hoc verification of verbalized evidence. Additional generation and termination details are provided in Appendices C.4 and D.2, with generation prompts in Appendix D.3.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

Evaluation instances. We use the train/validation/test partitions defined in Section 3. Long-horizon evaluation uses 30 matched profiles and initial conditions from the held-out test split, with the same profile shared across compared systems. Profiles follow the construction in Section 5.2.

Behavioral-state baselines. We evaluate attitude, emotion, and topic prediction as multi-label tasks with 8, 7, and 7 labels using Micro-F1 and Macro-F1. UNFOLD is compared with supervised `roberta-base` (Liu et al., 2019), given preceding dialogue and the current therapist turn, and GPT-4o-mini (OpenAI, 2024), Gemini-2.5-Pro (Comanici et al., 2025), and DeepSeek-V3.2 (DeepSeek-AI et al., 2025) prompted for the same labels. Details are in Appendix D.1.

Long-horizon simulation. Each system interacts with the same GPT-4o-mini therapist under a fixed protocol. All patient generators receive the five most recent therapist–patient exchange pairs and share the same attitude, emotion, and topic control spaces. Full-profile baselines receive predicted state guidance, the complete profile, and a gradual-disclosure instruction; UNFOLD instead uses its structured predictor and disclosure policy. Its GPT-4o-mini generator uses temperature 0.7 and at most 768 output tokens. Full-profile baselines use a 60-patient-turn maximum but may terminate earlier through generated ending decisions. UNFOLD has no fixed turn cap; completion is tied to selector coverage over available non-index profile groups. We additionally extend the baseline limit to 200 turns to test sensitivity to the main-setting horizon. Conditioning, generation, and termination details are provided in Appendix D.2, with prompts in Appendix D.3.

Information-trajectory metrics. We recover expressed profile evidence post hoc with Qwen2.5-72B-Instruct (Qwen et al., 2025) and measure: **Depth**, local continuity across expressed profile regions; **Breadth**, distribution across available groups; **Disclosure Restraint**, avoidance of simultaneous expression from more than three non-index groups; **Repetition**, adjacent-turn evidence reuse; **Mean First-Seen Turn**, normalized first-appearance position; and **Same-Group Bulk**, the largest single-turn introduction of unseen evidence from one group. These capture complementary trajectory properties rather than a single measure of clinical realism. Definitions and matcher validation are in Appendix D.4.

Therapist evaluation. Three therapist raters who regularly work with clients in long-term counseling relationships perform blinded pairwise comparisons on four dimensions: **Response Dynamics Plausibility**, whether patient state changes are interactionally coherent; **Progressive Disclosure Appropriateness**, whether patient information is revealed at an appropriate time and amount; **Long-Horizon Patient Plausibility**, whether behavior and information develop coherently across the session; and **Interaction Realism**, whether local responses resemble natural patient participation rather than profile recitation or clinical summarization. For each dimension, seven profile-matched comparisons are sampled against each baseline, yielding 21 items per dimension and 84 total. System identities are hidden and A/B order randomized. The full protocol is in Appendix D.5.

6.2 RQ1: DOES UNFOLD MODEL BEHAVIORAL EVOLUTION?

We first ask whether recent interaction history and therapist actions provide predictive signal for the patient's next behavioral state. UNFOLD achieves the highest Micro- and Macro-F1 on both attitude and emotion. On topic, Gemini-2.5-Pro achieves the highest Micro-F1 ($.580$ versus $.567$), while UNFOLD matches the highest Macro-F1 ($.361$). Compared with supervised RoBERTa, UNFOLD improves Micro-F1 by 5.5, 9.7, and 7.0 points on attitude, emotion, and topic, respectively. These results show that structured interaction history and therapist actions provide substantial predictive signal for turn-varying patient behavior, and that UNFOLD's behavioral control is grounded in an interaction-conditioned next-state prediction task rather than being left to free-form generation.

Table 1: **Next behavioral-state prediction on 8,397 held-out patient turns.** Micro-F1 and Macro-F1 are reported for each multi-label task.

	Attitude		Emotion		Topic	
Model	**Micro ↑**	**Macro ↑**	**Micro ↑**	**Macro ↑**	**Micro ↑**	**Macro ↑**
RoBERTa	.492	.216	.349	.227	.497	**.361**
GPT-4o-mini	.381	.113	.263	.174	.386	.175
DeepSeek-V3.2	.317	.105	.310	.220	.492	.168
Gemini-2.5-Pro	.295	.096	.259	.221	**.580**	.358
UNFOLD state predictor	**.547**	**.249**	**.446**	**.314**	.567	**.361**

6.3 RQ2: Does Explicit Information Availability Produce Better Long-Horizon Trajectories?

We compare complete pipelines over 30 matched profiles and characterize how patient-specific information is distributed, repeated, and introduced over time. In the main setting, full-profile baselines use a 60-patient-turn maximum to bound otherwise open-ended simulations but may terminate earlier through generated ending decisions, whereas UNFOLD uses no fixed maximum-turn budget.

Table 2 shows a clear trajectory trade-off. GPT-4o-mini has the highest Depth (.874) but lowest Breadth (.188), showing that strong local continuity can coexist with concentrated information use. UNFOLD instead has the highest Breadth (.558) and Disclosure Restraint (.632), lowest Repetition (.528), and Depth of .712. Thus, Depth alone does not capture broad, gradual, and non-repetitive information unfolding. UNFOLD also introduces unseen evidence later (First-Seen .330 versus .161–.185) and in smaller same-group bursts (1.63 versus 2.57–3.37), indicating more distributed progression rather than restraint obtained by repeatedly reusing the same evidence. Profile-level paired and leave-one-out analyses in Appendix E.2 show that these differences are not driven by a single profile.

Table 2: **Long-horizon information trajectories over 30 matched profiles.** Rep. = Repetition; First-Seen = Mean First-Seen Turn; Bulk = Same-Group Bulk.

Model	**Avg. turns**	**Depth ↑**	**Breadth ↑**	**Restraint ↑**	**Rep. ↓**	**First-Seen ↑**	**Same-Group Bulk ↓**
GPT-4o-mini	60.00	**.874**	.188	.246	.745	.170	2.57
Gemini-2.5-Pro	55.30	.723	.359	.252	.576	.185	3.13
DeepSeek-V3.2	58.63	.671	.380	.103	.687	.161	3.37
UNFOLD	132.70	.712	**.558**	**.632**	**.528**	**.330**	**1.63**

Long-horizon robustness. To test whether these differences arise from horizon or termination budgets, we extend the full-profile limit from approximately 60 to 200 patient turns (Appendix E.1). Average session length rises from 60.00 to 195.87 for GPT-4o-mini, 55.30 to 172.00 for Gemini-2.5-Pro, and 58.63 to 132.37 for DeepSeek-V3.2, while the main trajectory differences persist. The turn-by-turn trajectories in Figure 4 further show that full-profile simulators express more evidence early and approach a plateau, whereas UNFOLD introduces evidence more gradually over a longer portion of the interaction. Thus, extending the available horizon alone does not reproduce the same information trajectory.

6.4 RQ3: Why Does UNFOLD Work?

We test three mechanism-level hypotheses: **H1—Behavioral dynamics:** removing patient-state prediction from generation should primarily degrade behavioral-state realization; **H2—Information selection:** removing stateful selection should disrupt cross-turn progression, causing earlier and less distributed exposure; and **H3—Disclosure bottleneck:** removing the bottleneck should increase within-turn concentration and evidence reuse. The first three columns below measure *state realization in generated responses*, not next-state prediction on held-out real turns.

Table 3: **Mechanism-targeted ablations.** Restr. = Disclosure Restraint; Rep. = Repetition; First = Mean First-Seen Turn; Bulk = Same-Group Bulk.

	Behavioral Realization			Information Trajectory				
Variant	**Att.↑**	**Emo.↑**	**Topic↑**	**Breadth↑**	**Restr.↑**	**Rep.↓**	**First↑**	**Bulk↓**
Full	**.551**	**.459**	**.561**	**.558**	**.632**	.528	**.330**	**1.63**
− Dynamics	.517	.285	.302	.541	.598	**.518**	.328	1.67
− Selection	.548	.434	.535	.307	.280	.653	.124	1.97
− Bottleneck	.539	.415	.549	.509	.358	.723	.231	2.17

The ablations follow the targeted pattern. Removing dynamics most strongly degrades state realization, especially emotion ($.459 \rightarrow .285$) and topic ($.561 \rightarrow .302$), while information metrics remain close to Full. Removing selection leaves behavioral realization relatively stable but reduces Breadth ($.558 \rightarrow .307$), Restraint ($.632 \rightarrow .280$), and First-Seen ($.330 \rightarrow .124$), while increasing Repetition ($.528 \rightarrow .653$). Removing the bottleneck mainly affects concentration and reuse: Restraint falls to .358, Repetition rises to .723, and Same-Group Bulk to 2.17. These patterns distinguish the roles of state prediction, stateful selection, and the disclosure bottleneck. Full results and variant definitions are in Appendix E.3.

6.5 RQ4: Do These Trajectory Differences Matter to Human Raters?

We finally test whether therapist raters perceive the conversational differences targeted by UNFOLD. UNFOLD wins all 28 comparisons against DeepSeek-V3.2 and 24 of 28 against Gemini-2.5-Pro. Against GPT-4o-mini, the comparison is closer: 18 of 28 outcomes are ties, including six of seven Response Dynamics comparisons, while Interaction Realism shows the clearest separation with four wins and three ties. Overall, UNFOLD receives 61 wins, 18 ties, and 5 losses across 84 evaluation items, corresponding to a tie-adjusted preference of .833. These results show that therapist raters favor the conversational properties produced by UNFOLD overall, with the strongest separation from DeepSeek-V3.2 and Gemini-2.5-Pro and a narrower advantage over GPT-4o-mini. Overall Fleiss' κ is .504. Complete outcomes and agreement statistics are reported in Appendix E.4.

Table 4: **Blinded pairwise therapist evaluation.** Each cell reports W/T/L outcomes for **UNFOLD** over seven matched items; parentheses report tie-adjusted preference.

Compared with	**Dynamics**	**Disclosure**	**Long-Horizon**	**Realism**
GPT-4o-mini	1/6/0 (.571)	2/5/0 (.643)	2/4/1 (.571)	4/3/0 (.786)
Gemini-2.5-Pro	7/0/0 (1.000)	6/0/1 (.857)	5/0/2 (.714)	6/0/1 (.857)
DeepSeek-V3.2	7/0/0 (1.000)	7/0/0 (1.000)	7/0/0 (1.000)	7/0/0 (1.000)

7 Conclusion

We introduced UNFOLD, a long-horizon patient simulator that separates and coordinates two temporal processes: *behavioral evolution*, governing how the patient responds over time, and *information availability*, governing when patient-specific evidence becomes available to generation. UNFOLD combines interaction-conditioned patient-state prediction with progressive profile disclosure before response generation. On 8,397 held-out patient turns, it achieves the strongest attitude and emotion prediction and strong topic performance. Across 30 matched simulations, UNFOLD produces broader, more temporally distributed, and more restrained information trajectories, with less evidence reuse and smaller same-group disclosure bursts than full-profile simulators. Mechanism-targeted ablations isolate the roles of state prediction, stateful selection, and the disclosure bottleneck, while blinded therapist evaluation shows an overall preference for UNFOLD's conversational properties. Together, these results show that explicitly modeling behavioral evolution and information availability as separate but coupled processes enables more controlled long-horizon patient simulation and may offer a useful perspective for other interactive systems that must coordinate evolving behavior with persistent information over time. Additional limitations and scope considerations are discussed in Appendix A.

ETHICS STATEMENT

Data source and licensing. The source material consists of Alexander Street's *Counseling and Psychotherapy Transcripts*, Volumes I and II, accessed through an institutional library license for academic research. These materials are licensed resources rather than public-domain data. We use the transcripts for research purposes and do not redistribute the source text.

Privacy and de-identification. Psychotherapy transcripts may contain sensitive personal, relational, and clinical information. We therefore treat both the original transcripts and transcript-derived patient profiles as sensitive research artifacts. Neither transcript text nor transcript-derived structured profiles are included in the public release. All transcript content submitted to external LLM APIs was de-identified before transmission, with identifying information removed from the model inputs. Likewise, all materials presented to human evaluators were generated from de-identified inputs, and no identifiable source transcript content was shown to the raters. We do not attempt to identify or re-identify individuals represented in the source material.

Intended use and clinical scope. UNFOLD is intended as a research environment for studying long-horizon patient simulation and psychotherapy-oriented agent interaction. It is not intended to provide psychotherapy, diagnosis, risk assessment, treatment recommendations, or other clinical decisions, and simulated responses should not be interpreted as predictions of how any particular real patient would respond. Outputs may contain errors, simplifications, or biases inherited from the source material, annotation process, or language models used in the pipeline. The framework should therefore be used for research and controlled evaluation, not as a substitute for qualified clinical judgment or direct patient care.

Expert evaluator compensation and consent. The blinded expert evaluation was conducted by three therapist raters. All evaluators were compensated at a rate of $25 per hour. Before participation, they were informed of the nature of the evaluation, the expected workload, and the compensation arrangement, and all raters provided informed consent. Participation was voluntary.

REPRODUCIBILITY STATEMENT

Supervision schema, annotation procedures, and quality audits are provided in Appendix B; model and simulation implementation details are provided in Appendix C; experimental protocols, generation conditioning, prompts, evaluation metrics, and therapist evaluation procedures are provided in Appendix D; and complete robustness and ablation results are reported in Appendix E. We additionally provide an anonymous implementation of UNFOLD at `https://anonymous.4open.science/r/UNFOLD-Patient-9EDF`, including model implementation and checkpoints, experimental configuration files, supervision taxonomies, and prompts. Licensed psychotherapy transcripts and transcript-derived patient profiles are not redistributed.

AI USE STATEMENT

Generative AI tools were used during the research process to provide feedback on research methodology and experimental design, assist with research-code implementation and debugging, support translation and language editing, assist with literature search and analysis, support interpretation of experimental results, and assist with manuscript polishing.

Language models were also used as explicit components of the experimental pipeline. Qwen2.5-72B-Instruct was used for offline supervision construction and post-hoc profile-evidence matching; GPT-4o-mini was used for structured profile construction and dialogue generation; and GPT-4o-mini, Gemini-2.5-Pro, and DeepSeek-V3.2 were used as experimental components or baselines, as described in the corresponding methodological and experimental sections.

All AI-assisted research artifacts and manuscript content were reviewed, modified where necessary, and verified by the authors. The authors take responsibility for the experimental design, implementation, data processing, analysis, interpretation, citations, and all claims and artifacts presented in this work.

REFERENCES

Alexander Street Press. Counseling and psychotherapy transcripts: Volume i [full text data], 2023a. URL https://doi.org/10.57761/C9DA-ZQ22.

Alexander Street Press. Counseling and psychotherapy transcripts: Volume ii [full text data], 2023b. URL https://doi.org/10.57761/ZH3G-CH31.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. The ISO standard for dialogue act annotation, second edition. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 549–558, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.69/.

Yichen Cai, Pei Chen, Jiayang Li, Jingya Guo, Zejian Li, Changyuan Yang, and Lingyun Sun. ThinkPersona: Thinking with persona graphs for faithful individualized role-playing. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9905–9927, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.449. URL https://aclanthology.org/2026.acl-long.449/.

Patricia Chamberlain, Gerald Patterson, John Reid, Kathryn Kavanagh, and Marion Forgatch. Observation of client resistance. *Behavior therapy*, 15(2):144–155, 1984.

Serina Chang, Ashton Anderson, and Jake M Hofman. Chatbench: From static benchmarks to human-ai evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26009–26038, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

DeepSeek-AI, Aixin Liu, et al. Deepseek-v3.2: Pushing the frontier of open large language models, 2025. URL https://arxiv.org/abs/2512.02556.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL https://aclanthology.org/2020.acl-main.372/.

Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. SimulatorArena: Are user simulators reliable proxies for multi-turn evaluation of AI assistants? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35212–35290, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1786. URL https://aclanthology.org/2025.emnlp-main.1786/.

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

Tracy D. Eells, Edward M. Kendjelic, and Cynthia P. Lucas. What's in a case formulation?: Development and use of a content coding manual. *Journal of Psychotherapy Practice and Research*, 7 (2):144–153, 1998.

Paul Ekman. Are there basic emotions? *Psychological Review*, 99(3):550–553, 1992. doi: 10.1037/0033-295X.99.3.550.

Barry A Farber. Patient self-disclosure: A review of the research. *Journal of clinical psychology*, 59 (5):589–600, 2003.

Clara E. Hill. Development of a counselor verbal response category system. *Journal of Counseling Psychology*, 25(5):461–468, 1978. doi: 10.1037/0022-0167.25.5.461.

Daeun Kyung, Hyunseung Chung, Seongsu Bae, Jiho Kim, Jae Ho Sohn, Taerim Kim, Soo Kyung Kim, and Edward Choi. Patientsim: A persona-driven simulator for realistic doctor-patient interactions. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-0861. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/24945e3bdc7b3f4b2e64b9979a16f38e-Paper-Datasets_and_Benchmarks_Track.pdf.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. Hello again! llm-powered personalized agent for long-term dialogue. 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)*, pp. 5259–5276, 2025.

Yahan Li, Jifan Yao, John Bosco S. Bunyi, Adam C Frank, Angel Hsing-Chi Hwang, and Ruishan Liu. Counselbench: A large-scale expert evaluation and adversarial benchmarking of large language models in mental health question answering. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8MBYRZHVWT.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10570–10603, 2024.

Helen F Massfeller and Tom Strong. Clients as conversational agents. *Patient Education and Counseling*, 88(2):196–202, 2012.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 task 1: Affect in tweets. In Marianna Apidianaki, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat (eds.), *Proceedings of the 12th International Workshop on Semantic Evaluation*, pp. 1–17, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1001. URL https://aclanthology.org/S18-1001/.

Theresa B Moyers, Lauren N Rowell, Jennifer K Manuel, Denise Ernst, and Jon M Houck. The motivational interviewing treatment integrity code (miti 4): rationale, preliminary reliability and validity. *Journal of substance abuse treatment*, 65:36–42, 2016.

National Institute for Health and Care Excellence. *Depression in Adults: Treatment and Management*. Number 222 in NICE Guideline. National Institute for Health and Care Excellence (NICE), London, 2022. URL https://www.ncbi.nlm.nih.gov/books/NBK583074/.

Kai Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seung-won Hwang, Dongha Lee, and Jinyoung Yeo. Towards lifelong dialogue agents via timeline-based memory management. 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)*, pp. 8631–8661, 2025.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. OpenAI, July 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Published July 18, 2024.

Joanna Pawelczyk and Richard G. Erskine. Expressing the unexpressed: self-disclosure as interactional achievement in the psychotherapy session. *Communication & medicine*, 5(1):35–47, 2008. URL https://api.semanticscholar.org/CorpusID:10938274.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, January 2025. URL http://arxiv.org/abs/2412.15115. arXiv:2412.15115 [cs].

Joel J. Silverman, Marc Galanter, Maga Jackson-Triche, Douglas G. Jacobs, James W. Lomax, Michelle B. Riba, Lowell D. Tong, Katherine E. Watkins, Laura J. Fochtmann, Richard S. Rhoads, and Joel Yager. The american psychiatric association practice guidelines for the psychiatric evaluation of adults. *American Journal of Psychiatry*, 172(8):798–802, 2015. doi: 10.1176/appi.ajp.2015.1720501. URL https://psychiatryonline.org/doi/abs/10.1176/appi.ajp.2015.1720501.

Christina S Soma, Brian RW Baucom, Bo Xiao, Jonathan E Butner, Peter Hilpert, Shrikanth Narayanan, David C Atkins, and Zac E Imel. Coregulation of therapist and client emotion during psychotherapy. *Psychotherapy research*, 30(5):591–603, 2020.

William Stiles. Verbal response modes and dimensions of interpersonal roles: A method of discourse analysis. *Journal of Personality and Social Psychology*, 36:693–703, 07 1978. doi: 10.1037/0022-3514.36.7.693.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, et al. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439, 2025.

Kai Wang, Haoyang You, Yang Zhang, and Zhongjie Wang. Memory-driven role-playing: Evaluation and enhancement of persona knowledge utilization in llms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23475–23510, 2026.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. Know you first and be you better: Modeling human-like user simulators via implicit profiles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21082–21107, 2025.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, et al. PATIENT-ψ: using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12772–12797, 2024.

Jihao Zhao, Ding Chen, Zhaoxin Fan, Kerun Xu, Mengting Hu, Bo Tang, Feiyu Xiong, and Zhiyu li. Inside out: Evolving user-centric core memory trees for long-term personalized dialogue systems. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13429–13446, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.614. URL https://aclanthology.org/2026.acl-long.614/.

A LIMITATIONS

UNFOLD learns interaction-conditioned dynamics shared across the training population, which captures common response patterns but may underrepresent more idiosyncratic patient trajectories; richer longitudinal data could support more individualized dynamics. Its disclosure policy is fixed rather than learned because each transcript provides only one observed disclosure trajectory and no counterfactual supervision, so our results establish the value of explicit disclosure control rather than the optimality of the specific policy. Our empirical study uses one licensed psychotherapy corpus, which enables a controlled evaluation but leaves broader cross-population and cross-setting generalization to future validation. Importantly, the framework itself is not tied to a particular language or culture: its separation of behavioral dynamics from information availability can be instantiated with language- and culture-specific corpora, state taxonomies, and profile schemas, and can naturally extend to other user-simulation settings.

B DATA, TAXONOMIES, AND ANNOTATION

This appendix summarizes the supervision schema used by UNFOLD. The cited frameworks provide *construct-level grounding* for what should be represented, but do not define the exact label inventories used in our experiments. Our inventories are task-specific operationalizations for transcript annotation and simulation control: existing constructs are selected, reorganized, and, where necessary, subdivided into observable categories that can be grounded in dialogue text. They should therefore not be interpreted as reproductions of established clinical, affective, or dialogue coding systems, nor as independently validated clinical instruments.

Table 5 summarizes the distinction between construct grounding and UNFOLD's operationalization. The operational inventories and hierarchy summaries used in the paper are described below; complete fine-grained topic, emotion, and profile definitions and mappings are provided in the machine-readable supplementary materials. Annotation quality under these operational definitions is examined in Section B.6.

B.1 PATIENT ATTITUDE

Patient attitude operationalizes the interactional form of the patient's response to the preceding therapist turn. Dialogue-act frameworks characterize communicative functions, verbal-response-mode analysis distinguishes recurrent forms of interpersonal response, and observational work on client resistance motivates explicit representation of avoidance and refusal (Bunt et al., 2020; Stiles, 1978; Chamberlain et al., 1984). These sources motivate the construct space but do not define the eight labels in Table 6. We reorganize the relevant interactional distinctions into eight observable, multi-label response modes suitable for psychotherapy dialogue annotation and simulation control.

B.2 CLINICAL TOPIC

Clinical topic represents the semantic content of the current patient or therapist utterance. Case-formulation, psychiatric-evaluation, and depression-care frameworks jointly motivate coverage of clinically relevant content such as symptoms, functioning, contextual stressors, history, risk, therapeutic process, and coping (Eells et al., 1998; Silverman et al., 2015; National Institute for Health and Care Excellence, 2022). These sources do not specify the 78 labels or hierarchy used here. UNFOLD operationalizes the resulting content space as 78 fine-grained labels, grouped into 17 facets and then seven control domains. The seven-domain representation is used by the patient-state predictor and by the semantic interface to profile selection.

The complete operational definitions of the 78 fine-grained labels and their fine-to-facet-to-domain mappings are provided in the anonymous code release.

B.3 PATIENT EMOTION

Patient emotion is represented as a hierarchical multi-label state with intensity. GoEmotions motivates fine-grained categorical coverage, SemEval-2018 Task 1 provides precedent for multi-label

Table 5: **Construct grounding and operationalization of the UNFOLD supervision schema.** Prior frameworks motivate the represented constructs; the exact inventories, hierarchies, and mappings are task-specific operationalizations introduced for UNFOLD.

Component	Construct grounding	UNFOLD operationalization	Scope of adaptation
Patient attitude	Dialogue-act functions, verbal response modes, and observable client resistance (Bunt et al., 2020; Stiles, 1978; Chamberlain et al., 1984)	8 multi-label response modes describing how the patient engages with the preceding therapist turn	The eight labels are not the response-mode inventory of Stiles or an ISO dialogue-act subset; they reorganize related interactional functions into a task-specific patient-response space.
Patient emotion	Fine-grained categorical emotion, multi-label affect and intensity annotation, and broad emotion distinctions (Demszky et al., 2020; Mohammad et al., 2018; Ekman, 1992)	73 fine labels → 37 intermediate states → 7 broad families, with intensity attached to active emotion labels	GoEmotions and SemEval motivate fine-grained and multi-label/intensity representation, while the 73–37–7 hierarchy and mappings are constructed for UNFOLD rather than inherited from either taxonomy.
Clinical topic	Information domains represented across psychotherapy case formulation, psychiatric evaluation, and depression assessment and management (Eells et al., 1998; Silverman et al., 2015; National Institute for Health and Care Excellence, 2022)	78 fine labels → 17 facets → 7 domains used to represent therapeutic content	The cited frameworks jointly motivate domain coverage; the 78 labels, 17 facets, 7-domain hierarchy, and mappings are UNFOLD-specific operationalizations.
Therapist strategy	therapist verbal responses, psychotherapy process coding, and dialogue-act functions (Hill, 1978; Moyers et al., 2016; Bunt et al., 2020)	19 observable therapist moves → 4 functional families	The 19-move inventory is not a reproduction of Hill's response categories, MITI, or ISO dialogue acts; these systems motivate functional distinctions reorganized for the present task.
Patient profile	Case-formulation and psychiatric-evaluation domains covering symptoms, context, history, risk, treatment, and other persistent patient information (Eells et al., 1998; Silverman et al., 2015; National Institute for Health and Care Excellence, 2022)	27 fine-grained groups → 8 families containing independently addressable patient-specific evidence	The cited frameworks jointly motivate information coverage; the 27-group schema, group boundaries, and relation structure are introduced for selective profile access and progressive disclosure.

emotion and emotion-intensity annotation, and broad emotion distinctions inform the upper level of the representation (Demszky et al., 2020; Mohammad et al., 2018; Ekman, 1992). None of these sources defines the hierarchy used by UNFOLD. The 73 fine labels, 37 intermediate states, seven broad families, and their mappings are task-specific operationalizations.

Table 8 summarizes the hierarchy. Six non-neutral families contain 36 intermediate states and 72 fine-grained labels; adding NEUTRAL as a singleton yields seven families, 37 intermediate states, and 73 fine-grained labels.

The complete fine-label definitions and mappings to intermediate and broad emotion states are provided in the supplementary taxonomy.

Table 6: **Patient-attitude inventory.** Each label is an operational category defined by the observable interactional function of the patient turn.

Label	Operational definition
Answer-Brief	Directly answers the therapist's question with minimal informational content.
Answer-Narrative	Directly answers while providing substantive explanation, personal experience, or elaboration.
Narrative-Only	Provides experiences, feelings, or narrative content without directly answering the therapist's current question.
Avoiding	Sidesteps or postpones the requested information without explicitly refusing.
Refusal	Explicitly rejects the request to answer or expresses a clear resistant stance.
Counter-Question	Responds by asking the therapist a question and redirects responsibility for answering.
Clarification-Request	Requests clarification or confirmation of the therapist's intended question before answering.
Topic-Shift	Actively redirects the conversation from the current topic to a new topic.

Table 7: **Clinical-topic hierarchy.** The seven domains form the compact topic representation used by UNFOLD; parentheses indicate the number of fine-grained operational labels under each facet.

Code	Domain	Facets	# Fine
SYM	Symptoms & Distress	Symptom Type (4); Symptom Aspect (4)	8
FXN	Functioning & Roles	Functioning Domain (6); Functioning Aspect (6)	12
CTX	Life Events & Stressors	Temporal Attribute (4); Stressor Type / Source (8)	12
HIST	Background & History	Subject (3); History Type (4)	7
RISK	Risk & Safety	Risk / Safety Domain (3); Risk Level / Nature (7); Risk Temporality (3)	13
PROC	Therapeutic Process & Alliance	Target / Dimension (3); Process Type (8); Valence / Emotional Direction (3)	14
COP	Coping & Resources	Coping Strategy Type (5); Coping Valence / Adaptiveness (3); Coping Status / Stage (4)	12
Total	**7 domains**	**17 facets**	**78**

B.4 Therapist Strategy

Therapist strategy represents the observable interactional action used to address the current clinical content. Therapist verbal-response systems motivate distinctions among questioning, reflection, interpretation, and other response functions; treatment-integrity coding motivates intervention-level distinctions; and dialogue-act frameworks motivate explicit representation of conversational functions (Hill, 1978; Moyers et al., 2016; Bunt et al., 2020). These sources provide construct grounding rather than an existing 19-label standard. UNFOLD reorganizes relevant functions into 19 observable moves grouped into four functional families.

B.5 Patient Profile

The patient profile represents persistent patient-specific information rather than turn-level behavioral state. Case-formulation, psychiatric-evaluation, and depression-care frameworks jointly motivate broad coverage of persistent patient information, including symptoms and functioning, current stressors, history, risk, treatment context, interpersonal context, and coping resources (Eells et al., 1998; Silverman et al., 2015; National Institute for Health and Care Excellence, 2022). They do not define the 27-group schema or relation structure used by UNFOLD. We reorganize the resulting information space into 27 fine-grained groups and eight families, with each group storing independently addressable evidence suitable for selective access during progressive disclosure. The con-

Table 8: **Patient-emotion hierarchy.** Intermediate states operationalize the six non-neutral affect families; NEUTRAL is added as a singleton category.

Family	Intermediate affect states	# Fine
Scared	fearful, insecure, anxious, rejected, weak, nervous	12
Mad	offended, aggressive, frustrated, angry, hateful, let down, salty, humiliated	16
Disgusted	outraged, awful, grossed out, avoiding	8
Sad	hurt, depressed, in my feelings, guilty, lonely, tired	12
Happy	trusting, joyful, proud, accepted, enthusiastic, interested, optimistic, peaceful	16
Surprised	confused, amazed, startled, excited	8
Neutral	no salient non-neutral affect	1
Total	**37 intermediate states including Neutral**	**73**

Table 9: **Therapist-strategy hierarchy.** Nineteen task-specific therapist moves are grouped into four higher-level functional families.

Family	Code	Fine-grained moves
Inquiry	INQ	Open Question; Closed / Factual Question; Elaboration
Listening & Connection	LIST	Minimal Encourager; Paraphrase; Reflection of Feeling; Validation; Summary; Intentional Silence
Intervention & Change	INT	Psychoeducation / Information; Directive / Advice; Interpretation / Reframing; Challenge / Reality Testing; Affirmation; Self-Disclosure
Management & Framework	MGMT	Agenda / Flow Management; Check / Consent; Framework / Boundaries; Social / Phatic

struction procedure, representation statistics, and a synthetic example are provided in Appendix C.2; the complete group mapping is included in the supplementary schema.

B.6 AUTOMATIC ANNOTATION AND QUALITY AUDIT

Patient-state and therapist-action supervision is generated with Qwen2.5-72B-Instruct (Qwen et al., 2025) under fixed operational definitions and constrained structured outputs. For each target utterance, the annotator receives a local five-turn window consisting of the target turn, the preceding two turns, and the following two turns when available. This bidirectional context is used only for offline supervision construction; the patient-state predictor itself uses only information available before the target patient response.

Patient topic and emotion are first annotated under their detailed operational inventories and mapped through the predefined hierarchies to the compact control spaces used by UNFOLD. Patient attitude is represented directly in its eight-label space. Therapist topic follows the clinical-topic supervision scheme, while therapist strategy uses the 19-move inventory above. Active patient-emotion labels additionally carry intensity on a five-point scale.

We audit 150 turns sampled randomly from the complete automatically annotated corpus. Human reference annotations are established by reviewing each target and its local context against the same written fine-grained definitions used to specify the automatic annotation task. We report Precision, Recall, and F1 in the corresponding operational label spaces for all five annotation tasks. The audit is therefore intended as a quality-control assessment of agreement with the predefined operational criteria, rather than an independent clinical annotation study.

The audit shows that automatically constructed supervision is broadly consistent with the task-specific operational definitions, while also indicating non-negligible residual annotation noise, particularly for patient emotion. These results should therefore be interpreted as annotation quality-control statistics rather than evidence of clinical validity. More generally, the audits characterize consistency with the supervision definitions used by UNFOLD.

Table 10: **Manual audit of automatically constructed supervision.** All audits use $n = 150$ sampled target turns.

Annotation	Precision	Recall	F1
Patient attitude	87.5%	86.1%	86.8%
Patient emotion	81.1%	72.1%	76.3%
Patient topic	85.8%	86.6%	86.2%
Therapist strategy	82.0%	72.5%	76.9%
Therapist topic	84.0%	84.0%	84.0%

C Model and Simulation Implementation

C.1 Patient-State Predictor Implementation

The attitude, emotion, and topic predictors use a maximum lag of three interaction steps. Their inputs are structured interaction annotations rather than raw dialogue text. For each historical feature sequence S, short-range context is summarized as

$$\phi(S) = [S_{\text{last}}; \text{mean}(S); \text{mean}(\mathbf{1}\{S > 0\})],$$

and these summaries are concatenated with the current therapist topic and strategy to form the interaction representation e_t.

Each latent factor uses hidden dimension 64. The global priors q_c^0 and projected summaries $\bar{c}^{(-c)}$ are shared across dialogue instances, while turn-specific information enters through e_t. Latent beliefs are inferred independently at each dialogue turn; temporal information therefore enters through the observed short-range structured history rather than recurrent propagation of latent beliefs across turns.

The three prediction tasks are trained separately. Attitude uses focal binary cross-entropy with $\gamma = 2$; topic uses multi-label binary cross-entropy with training-set positive-class weights; and emotion jointly predicts label presence and normalized intensity using focal binary cross-entropy and mean-squared error with equal weight. All models use AdamW with learning rate 5×10^{-4}, batch size 32, weight decay 5×10^{-4}, gradient clipping at 1.0, and early stopping on validation loss. The final decision thresholds, selected using validation data only, are 0.50, 0.40, and 0.50 for attitude, emotion, and topic, respectively.

C.2 Structured Patient-Profile Construction

Structured profiles are constructed offline with GPT-4o-mini from the structured labeled dialogues under the fixed 27-group schema. Clinical-topic labels route local transcript units to semantically relevant profile sections, after which patient-specific evidence is extracted, merged, deduplicated, and refined while retaining grounding in the source transcript. The resulting representation stores independently addressable evidence units rather than a single free-form patient summary.

Across the 983 dialogue-level profiles, an average profile contains 18.15 non-empty fine-grained profile groups and 60.55 independently addressable evidence units. The chunk-level extraction, minimal-chunk fallback, and refinement prompts are included in the anonymous supplementary implementation. The complete profile serves as the fixed patient specification for simulation; the task is not to predict undisclosed future facts from partial real dialogue.

Illustrative structured profile. Table 11 shows a fully synthetic example following the same 27-group schema used in our experiments. It is included only to illustrate representation granularity and is not used for training, model selection, or evaluation.

Illustrative disclosure trace. Table 12 illustrates how different primary evidence units may become active as therapeutic focus changes. The example is fully synthetic and is included only to visualize the selected-evidence-to-response interface. Only the primary item is shown for clarity; actual generator context may also contain restricted bridge evidence as described below.

Table 11: **Illustrative synthetic structured profile.** One representative evidence unit is shown for each of the 27 fine-grained profile groups.

Group	Representative evidence	Group	Representative evidence
A1 Demographic & role	Works as a project coordinator at a small design company.	E4 Current therapy attitude	Is cautiously optimistic that therapy may help her understand recurring anxiety.
A2 Household & key people	Lives with a roommate and remains closely connected to her older sister and mother.	F1 Family pattern	Feels close to her older sister but rarely makes specific requests for help.
B1 Emotional distress	Experiences waves of anxiety that become especially noticeable in the evenings.	F2 Romantic pattern	Tends to withdraw temporarily rather than address relationship conflict immediately.
B2 Cognitive distress	Frequently worries that small mistakes will lead to serious consequences.	F3 Peer pattern	Usually presents herself to friends as organized and dependable.
B3 Functional impact	Spends excessive time checking work before submitting it.	F4 Relationship needs	Needs reassurance that disagreement does not mean a relationship is ending.
C1 Recent positive events	Recently completed an important work presentation despite significant anxiety.	G1 Beliefs about self	Believes a capable person should be able to handle most problems independently.
C2 Recent negative events	Received critical feedback from her supervisor after missing a project deadline.	G2 Beliefs about others	Believes repeated reassurance seeking may eventually frustrate other people.
C3 Ongoing stressors	Faces increasing pressure from overlapping project deadlines.	G3 Beliefs about help/change	Believes understanding the source of her anxiety may help her change how she responds to it.
D1 Early family environment	Grew up in a household where academic achievement was strongly emphasized.	G4 Repeating themes	Shows a recurring conflict between independence and dependence.
D2 Early relational experiences	Often tried to be the dependable friend who rarely needed support.	H1 Internal resources	Uses journaling to organize worries and identify recurring patterns.
D3 Early treatment entry	First sought counseling during university because of persistent anxiety and insomnia.	H2 External supports	Her current therapist provides a consistent source of professional support.
E1 Treatment timeline	Returned to therapy after work-related anxiety intensified during the past year.	H3 Behavioral coping	Takes short walks after work when feeling mentally overloaded.
E2 Helpful experiences	Breaking complex problems into smaller steps has reduced feelings of helplessness.	H4 Coping barriers	Often waits until distress becomes intense before seeking help.
E3 Disappointing experiences	Felt misunderstood when a previous therapist suggested relaxation before exploring her concerns.		

C.3 Disclosure-Policy Implementation

The disclosure policy maintains an explicit state over *primary selections*. At turn t, this state records the ordered history of primary-selected evidence, the active profile partition and group, and summary statistics derived from selection history. These variables track selector decisions rather than assuming that every selected item is expressed; actual profile-evidence expression is measured separately by the post-hoc matcher in Appendix D.4.

The current therapist topic and predicted patient topic define the semantic focus C_t. For profile group g, topic compatibility is

$$r_{\text{topic}}(g) = 2|C_t \cap C_g^{\text{pri}}| + |C_t \cap C_g^{\text{sec}}| - |C_t \cap C_g^{\text{avoid}}|. \tag{10}$$

In the reported configuration, all eligible non-index profile partitions may enter the dynamic candidate pool, so topic compatibility acts as a *soft relevance signal* rather than a hard access boundary.

Table 12: **Illustrative primary-selection → response trace.** Only the newly primary-selected evidence z_t is shown at each turn.

Turn	Therapist cue	Primary-selected evidence z_t	Illustrative patient response
1	"What tends to happen for you after you receive critical feedback at work?"	**G1 — Beliefs about self:** Tends to interpret mistakes as evidence that she has failed to meet expected standards.	"I usually keep going back over what I did wrong. Even when the mistake is small, it starts to feel like it says something bigger about whether I'm actually capable."
2	"When you start feeling that way, is there anyone you usually turn to?"	**F1 — Family pattern:** Feels close to her older sister but rarely makes specific requests for help.	"My sister is probably the person I'd feel safest talking to. I tell her some things, but I usually stop short of actually asking her for help."
3	"What makes it difficult to ask her directly for the support you need?"	**G2 — Beliefs about others:** Believes repeated reassurance seeking may eventually frustrate other people.	"I think part of me worries that if I keep needing reassurance, eventually she'll get tired of it. So even when I want support, I try not to ask for too much."

Candidate partitions are organized into three relation-based regions: *deep* candidates continue the active partition or local region, *similar* candidates come from related regions, and *distant* candidates permit broader transitions. The policy favors local continuation while retaining the possibility of broader movement.

Within the selected pool, candidate evidence p receives

$$\tilde{u}_t(p) = \lambda_R R_t(p) + \lambda_L L_t(p) + \lambda_N N_t(p) + B_t^{\text{cov}}(p) + B_t^{\text{topic}}(p), \tag{11}$$

where R_t measures continuity with the active selection state, L_t relation compatibility with recent primary selections, N_t novelty and recency, B_t^{cov} selector coverage, and B_t^{topic} compatibility with the current semantic focus. Selector coverage is defined from primary-selection history rather than post-hoc evidence expression.

Consecutive-use saturation. Repeated primary use of the same partition or profile group is progressively down-weighted after a short local run. The saturation term depends on consecutive primary-selection history rather than cumulative selection count or selector coverage, and is therefore distinct from the coverage criterion used for session termination. The resulting saturation factor modulates the candidate score before primary evidence is sampled according to Eq. 9.

Under the standard dynamic path, one evidence item is designated as the primary selection z_t. Additional bridge items may be supplied only to support local continuity. The generation prompt explicitly marks the primary item and instructs the generator not to express bridge items unless needed for coherence. The bottleneck therefore constrains newly prioritized evidence without requiring the entire generator-visible context to contain exactly one item.

Fallback behavior. The standard dynamic selector constructs candidates across eligible non-index profile partitions rather than enforcing topic compatibility as a hard filter. If a restricted candidate construction produces no valid partition, the selector progressively relaxes the restriction; if no profile partition can still be constructed, it ultimately falls back to full-profile context. This fallback is an exceptional recovery path rather than the standard information-access mechanism.

The heuristic policy parameters are fixed before test evaluation and are not claimed to be optimal. All selector hyperparameters, candidate-pool probabilities, scoring weights, saturation thresholds and multipliers, sampling temperature, fallback settings, and auxiliary schedules used in the reported experiments are provided in the anonymous code release.

C.4 Controlled Generation

The controlled patient generator is GPT-4o-mini. At each patient turn, it receives the five most recent therapist–patient exchange pairs, predicted attitude, emotion, and topic controls with their

definitions, disclosure-controlled profile context, and a profile-based ending signal. Generation uses temperature 0.7 and a maximum of 768 output tokens.

The generator-visible profile context is normally centered on one primary evidence item selected by the disclosure policy. When required for local continuity, the context may additionally contain bridge items. The prompt explicitly distinguishes these roles: the primary item is the currently prioritized evidence, whereas bridge items provide contextual support and should not be explicitly expressed unless needed for coherence. Under the standard dynamic path, the generator therefore receives selected profile context rather than unrestricted access to the complete structured profile. The exceptional full-profile fallback is described in Appendix C.3.

Session termination. UNFOLD imposes no fixed maximum number of patient turns. Session termination uses selector coverage, which is distinct from the consecutive-use saturation penalty in the selector. A non-index profile group enters selector coverage once evidence from that group has been primary-selected at least once. The profile-based ending condition becomes true when

$$|\mathcal{G}_P^{\text{selected}}| = |\mathcal{G}_P^{\text{available}}| > 0, \tag{12}$$

where both sets contain only non-index profile groups and $\mathcal{G}_P^{\text{selected}}$ contains the groups represented in the primary-selection history. This stopping criterion is therefore defined by selector state rather than by post-hoc verification that the corresponding evidence was expressed in the generated dialogue. The ending signal then permits the generator to produce a natural closing response.

D Experimental Protocols

D.1 RQ1: Next-State Prediction Protocol

All turn-level predictors are evaluated on the same 8,397 held-out patient turns and use only information available before the target patient utterance. The supervised RoBERTa references use three rounds of recent dialogue context, a maximum text length of 256 tokens, and the current therapist utterance. Attitude and topic use separate `roberta-base` multi-label classifiers, each consisting of a RoBERTa encoder followed by a linear prediction head. The emotion reference uses a shared RoBERTa text encoder with feature fusion and separate emotion and auxiliary valence heads. These models serve as supervised text-encoder references; UNFOLD instead uses the structured patient-state predictor described in Appendix C.1.

For attitude and topic, the RoBERTa references minimize weighted binary cross-entropy, with positive weights estimated as the training-set negative-to-positive ratio and clipped to $[1, 20]$. The emotion reference minimizes $\mathcal{L}_{\text{emo}} + 0.5\mathcal{L}_{\text{val}}$, using weighted binary cross-entropy for emotion presence and cross-entropy for the auxiliary valence target. All RoBERTa references are optimized with AdamW at learning rate 2×10^{-5}, batch size 8, for four epochs with seed 42. Validation-set threshold selection yields 0.65 for attitude and 0.50 for topic; the emotion threshold is likewise selected using validation data.

The prompted LLM references use the same task-level attitude, emotion, and topic inventories and return structured multi-label predictions. GPT-4o-mini, DeepSeek-V3.2, and Gemini-2.5-Pro are evaluated on the same held-out patient turns, and their outputs are mapped to the common control-level label spaces before Micro-F1 and Macro-F1 are computed.

D.2 RQ2: Long-Horizon Simulation Protocol

All systems are evaluated on the same 30 structured patient profiles and matched initial conditions derived from held-out test dialogues. For every patient system, the therapist is instantiated with GPT-4o-mini under the same therapist prompt, generation settings, and interaction protocol. Thus, the therapist model and protocol are held fixed across comparisons, although the realized therapist responses can naturally diverge as different patient systems produce different interaction trajectories.

The full-profile patient simulators are GPT-4o-mini, Gemini-2.5-Pro, and DeepSeek-V3.2. They use the same five most recent therapist–patient exchange pairs as UNFOLD. Patient generation uses temperature 0.7 and a maximum of 768 output tokens. The main long-horizon comparison allows each full-profile baseline at most 60 patient turns.

Table 13: **Generation-conditioning comparison.** Both settings receive explicit behavioral guidance and a gradual-disclosure instruction. Profile access, state-prediction implementation, and termination differ as described in the text.

Conditioning component	UNFOLD	Full-profile LLM
Five recent therapist–patient exchange pairs	✓	✓
Predicted attitude / emotion / topic	✓	✓
Behavioral label definitions	✓	✓
Gradual-disclosure instruction	✓	✓
Selected primary / bridge context	✓	–
Complete structured profile	fallback only	✓
Profile-based ending signal	✓	✓
Generated ending decision	–	✓

Both UNFOLD and the full-profile baselines receive explicit behavioral guidance in the same attitude, emotion, and topic control spaces, but this guidance is produced differently. UNFOLD predicts patient state from structured recent interaction features and the current therapist action without directly encoding raw dialogue text. For each full-profile baseline, the same patient LLM used for generation first predicts its own attitude, emotion, and topic labels from the recent dialogue text and shared label definitions under the common label-prediction protocol. The comparison therefore shares the behavioral-control semantics and label spaces, but not the state-prediction implementation.

At generation time, the full-profile baselines receive recent dialogue history, their predicted patient-state labels and definitions, the complete structured patient profile, and the profile-based ending signal. They are also explicitly instructed to disclose information gradually. UNFOLD instead uses its disclosure policy to construct profile context centered on the primary-selected evidence, with bridge items supplied when needed for continuity. Under the standard dynamic selection path, profile evidence is therefore mediated by the disclosure policy rather than exposed as unrestricted full-profile context. As detailed in Appendix C.3, this is a normal-operation property rather than an absolute access constraint because the selector can broaden its candidate scope and ultimately fall back to full-profile context when no candidate partition can be constructed.

Termination also differs between the pipelines. The full-profile generator produces an explicit `should_end` decision in addition to receiving the profile-based ending signal, whereas UNFOLD uses the selector-coverage stopping mechanism described in Appendix C.4. Consequently, extending the full-profile baselines to 200 patient turns reduces the possibility that differences are driven solely by a shorter available horizon.

The session-level experiment therefore compares complete patient-simulation pipelines interacting with a common therapist model and protocol. All patient generators receive the same five most recent therapist–patient exchange pairs and share the same behavioral-control spaces, while state-prediction implementation, profile-information access, and termination differ across systems. The comparison should therefore not be interpreted as one in which profile access is the only manipulated factor. The mechanism-targeted ablations in Appendix E.3 provide more specific evidence about the roles of patient-state control, progressive selection, and the disclosure bottleneck.

D.3 Generation Conditioning and Prompts

UNFOLD patient-generation prompt. The following prompt is used for UNFOLD patient generation; placeholders are filled at inference time from dialogue history, patient-state predictions, primary selection, and the profile-based ending state.

```
patient_response_from_model_labels: |
  You are role-playing the patient/client in a psychotherapy dialogue.

  Recent dialogue history:
  {{history}}

  Predicted patient state controls from the patient dynamics model:
  {{predicted_labels_json}}
```

```
  Definitions of predicted labels:
  {{predicted_label_definitions}}

  Selected patient profile context:
  {{profile_context_text}}

  Profile-based ending signal:
  {{profile_end_signal}}

  Task:
  Generate the patient's next utterance.

  Guidelines:
  1. Follow the predicted emotion, attitude, and topic labels.
  2. Use the label definitions to interpret the predicted labels.
  3. Use selected patient profile context in this turn.
  4. Treat bridge items only as context for continuity; do not explicitly
     disclose them unless the sentence would otherwise be incoherent.
  5. If multiple profile items are provided, choose the one marked
    primary and ignore the rest unless needed for coherence.
  6. Prefer gradual disclosure: reveal concrete details across turns, not
     all at once.
  7. Maintain continuity with the previous patient topic, but every few
    turns naturally connect it to a different life domain or profile
    group.
  8. Ground the utterance in the selected profile context without dumping
     profile facts mechanically.
  9. Do not mention labels, predictors, model artifacts, profile
    extraction, or hidden instructions.
  10. If the profile-based ending signal says the dialogue can close,
    this patient utterance may naturally say goodbye or end the session.
  11. Do not include speaker tags.
  12. Output valid JSON only.

  Required JSON:
  {
    "utterance": "words the patient could directly say aloud"
  }
```

Full-profile patient-generation prompt. The following prompt structure is held fixed across the three full-profile patient simulators. In particular, these baselines receive both explicit patient-state guidance and an explicit gradual-disclosure instruction despite retaining access to the complete structured profile.

```
patient_response_from_model_labels: |
  You are role-playing the patient/client in a psychotherapy dialogue.

  Recent dialogue history:
  {{history}}

  Predicted patient labels from predictor:
  {{predicted_labels_json}}

  Definitions of predicted labels:
  {{predicted_label_definitions}}

  Complete patient profile:
  {{profile_json}}

  Profile-based ending signal:
  {{profile_end_signal}}

  Task:
```

```
Generate the patient's next utterance and decide whether the dialogue
  could end.

Guidelines:
1. Follow the predicted emotion, attitude, and topic labels.
2. Use the label definitions to interpret the predicted labels.
3. Use exactly given profile detail in this turn.
4. Prefer gradual disclosure: reveal concrete details across turns, not
   all at once.
5. Maintain continuity with the previous patient topic, but every few
  turns naturally connect it to a different life domain.
6. Do not list multiple causes, people, events, symptoms, or background
   facts in one utterance.
7. Do not mention labels, predictors, model artifacts, profile
  extraction, or hidden instructions.
8. Do not include speaker tags.
9. Output valid JSON only.

Required JSON:
{
  "utterance": "words the patient could directly say aloud",
  "should_end": false
}
```

The therapist- and patient-generation prompts used by UNFOLD are included in the anonymous supplementary implementation.

D.4 RQ2: Profile Matching and Trajectory Metrics

Session-level trajectory metrics require identifying which structured profile evidence is expressed in generated patient responses. We perform this mapping post hoc using Qwen2.5-72B-Instruct (Qwen et al., 2025). The matcher is applied only after dialogue generation and therefore cannot affect patient-state prediction, primary selection, or generated responses.

Matcher audit. We audit 150 patient turns sampled randomly from the pooled generated outputs of UNFOLD and the three full-profile baseline systems. For each audited turn, the human reference is constructed by comparing the generated patient response against the evidence units in its corresponding structured profile under the same strict evidence-matching criterion used by the automatic evaluator. Precision, recall, and F1 are computed over evidence-item matches pooled across the 150 audited turns. The matcher achieves 90.5% precision, 96.4% recall, and 93.3% F1. These values quantify reliability of the automatic profile-evidence mapping rather than clinical validity of either the generated responses or the source profile.

Let M_t denote the set of non-index evidence items matched to patient turn t, and let $\mathcal{G}_P$ denote the non-index profile groups available for profile P. Identity-like index fields are excluded from all trajectory calculations. The six metrics below characterize complementary properties of information trajectories and are not combined into a single realism score.

Depth. Depth measures local continuity between profile regions expressed on adjacent patient turns. For each turn, we first take the *first non-index match* in the ordered Qwen `soft_matches` list and use its profile group as the turn-level primary group g_t. Turns without a non-index soft match have no primary group and are omitted from the corresponding adjacent comparison. Depth is then

$$\text{Depth} = \frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} \rho(g_{t-1}, g_t), \tag{13}$$

where $\mathcal{A}$ contains adjacent patient turns for which both primary groups are defined. The relation score is

$$\rho(g, g') = \begin{cases} 1.0, & g = g', \\ 0.8, & g, g' \text{ share a parent family}, \\ 0.7, & g, g' \text{ are direct relation neighbors}, \\ 0.4, & g, g' \text{ are connected by two relation steps}, \\ 0, & \text{otherwise}. \end{cases}$$

The profile-relation graph is fixed before evaluation. Because Depth deliberately reduces each turn to one ordered soft-match group, it should be interpreted as a local continuity statistic.

Breadth. Breadth measures how distributed expressed evidence is across the profile groups available for the patient. Let p_g denote the proportion of matched evidence assigned to group g. We compute normalized entropy:

$$\text{Breadth} = -\frac{\sum_{g \in \mathcal{G}_P} p_g \log(p_g + \epsilon)}{\log |\mathcal{G}_P|}. \tag{14}$$

Higher values indicate a less concentrated distribution across available profile regions.

Group-level disclosure restraint. Disclosure Restraint measures whether individual patient responses avoid simultaneously expressing many distinct profile regions:

$$\text{DisclosureRestraint} = 1 - \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{u_t > 3\}, \tag{15}$$

where u_t is the number of distinct non-index profile groups matched at turn t. Multiple evidence items from the same group count once. This metric therefore measures cross-group bulk release within a turn.

Repetition. Repetition measures immediate reuse of the same matched profile evidence:

$$\text{Repetition} = \frac{1}{T-1} \sum_{t=2}^{T} \frac{|M_t \cap M_{t-1}|}{\max(1, |M_t|)}. \tag{16}$$

Evidence identifiers rather than group identity are used, so returning to a related profile region does not count as repetition unless the same evidence item is matched again.

Mean first-seen turn. Patient turns are indexed by $t \in \{0, \ldots, T-1\}$. For each unique matched evidence item i expressed in the dialogue, define its first appearance as

$$\tau(i) = \min\{t : i \in M_t\}, \tag{17}$$

and let

$$U = \bigcup_{t=0}^{T-1} M_t. \tag{18}$$

The normalized mean first-seen position is

$$\text{MeanFirstSeen} = \frac{1}{|U|} \sum_{i \in U} \frac{\tau(i)}{T-1}. \tag{19}$$

Higher values indicate that first appearances of expressed evidence occur, on average, later across the normalized session. This metric characterizes temporal spread of first appearances.

Same-group bulk. To distinguish temporal spread from introducing many new facts from the same profile region at once, define newly expressed matched evidence at turn t as

$$N_t = M_t \setminus \bigcup_{s<t} M_s. \tag{20}$$

For profile group g,

$$B_{t,g} = |\{i \in N_t : \text{group}(i) = g\}|, \tag{21}$$

and

$$\text{SameGroupBulk} = \max_{t,g} B_{t,g}. \tag{22}$$

Lower values indicate fewer instances in which many previously unseen matched evidence items from the same profile region are expressed within a single patient response.

Unless otherwise specified, uncertainty is estimated with 2,000 dialogue-level bootstrap resamples, preserving within-dialogue dependence.

Table 14: **Therapist-evaluation protocol.** Different dimensions expose raters to temporal views matched to the property being judged.

Dimension	Input granularity	Material shown	Judgment target
Response Dynamics Plausibility	Local trajectory	An 8-turn dialogue window from each matched simulation	Whether patient topic, emotion, and interactional stance develop plausibly in response to the interaction rather than shifting arbitrarily, remaining frozen, or repeating mechanically.
Progressive Disclosure Appropriateness	Disclosure event	An 8-turn context window surrounding the first expression of the same profile evidence item in each matched simulation	Whether that patient-specific information is expressed at an appropriate interactional moment and in an appropriate amount.
Long-Horizon Patient Plausibility	Global compressed trajectory	Six checkpoints at approximately $10\%, 25\%, 40\%, 60\%, 75\%,$ and 90% of normalized session progress, with two consecutive dialogue turns shown at each checkpoint	Whether behavior and patient information develop coherently across the session, with plausible persistence and change and compatibility between earlier and later stages.
Patient Interaction Realism	Local interaction	An 8-turn dialogue window from each matched simulation	Whether the simulated patient participates naturally rather than behaving like an AI, reciting a profile, or producing a clinical summary.

D.5 RQ4: Therapist Evaluation Protocol

Three therapist raters independently perform blinded pairwise evaluation using the main long-horizon session outputs reported in Table 2. For the full-profile baselines, these are the original approximately 60-patient-turn simulations rather than the 200-turn horizon-control reruns. System identities are hidden and A/B presentation order is randomized. Each evaluation item compares two simulations generated from the same matched patient condition.

For each baseline and evaluation dimension, seven of the 30 matched profile/session pairs are sampled randomly. Sampling is performed independently by dimension, and session pairs used for one dimension are not reused for another dimension for the same baseline. This yields 21 evaluation items per dimension and 84 dimension-specific evaluation items overall. The term *evaluation item* is used deliberately: these are dimension-specific judgments constructed from matched sessions rather than 84 independent underlying patient profiles.

Raters are explicitly instructed not to prefer a system merely because its responses are longer, contain more information, cover more topics, or exhibit more behavioral change. Response Dynamics judgments concern whether persistence and change are supported by the local interaction. Progressive Disclosure judgments concern contextual timing and amount rather than simply whether information is expressed later. Long-Horizon judgments are made jointly across all six checkpoints. Interaction Realism concerns local conversational behavior and information density.

All three raters evaluate every item, yielding 252 individual judgments. An item is assigned WIN, LOSS, or TIE from UNFOLD's perspective when at least two raters select the same outcome. All 84 evaluation items receive a two-of-three majority outcome. This majority-resolution rate is an aggregation property and is not treated as a measure of inter-rater agreement. We report descriptive tie-adjusted preference

$$\frac{W + 0.5T}{W + T + L}.$$

Because each model–dimension cell contains only seven evaluation items, these values are interpreted descriptively rather than as inferential effect estimates. Moreover, because the full-profile baselines themselves receive explicit patient-state guidance through different predictors, Response Dynamics judgments compare complete simulation pipelines rather than isolate UNFOLD's state predictor.

Annotator compensation. All therapist raters were compensated at a rate of $25 per hour. Raters were informed in advance about the nature of the evaluation task, expected workload, and compensation. The rate was selected to provide fair compensation for expert annotation and exceeds the applicable local minimum-wage level.

E ADDITIONAL EXPERIMENTAL RESULTS

E.1 RQ2: EXTENDED-HORIZON ROBUSTNESS

The main long-horizon comparison uses the original full-profile baseline simulations with an approximately 60-patient-turn cap. Because this limits opportunities for later disclosure, we rerun GPT-4o-mini, Gemini-2.5-Pro, and DeepSeek-V3.2 with the maximum horizon increased to 200 patient turns while keeping the remaining generation settings unchanged. This substantially relaxes the baseline horizon constraint, although termination mechanisms remain different from UNFOLD.

Table 15 reports the resulting trajectory metrics under the original and extended horizons. Relaxing the horizon affects the full-profile simulators differently. DeepSeek-V3.2 increases Breadth from .380 to .437, whereas its Disclosure Restraint changes little (.103 $\rightarrow$ .107). GPT-4o-mini and Gemini-2.5-Pro gain little additional Breadth despite much longer sessions, while their adjacent evidence reuse increases substantially. Mean First-Seen decreases for all three baselines: .170 $\rightarrow$ .141 for GPT-4o-mini, .185 $\rightarrow$ .126 for Gemini-2.5-Pro, and .161 $\rightarrow$ .126 for DeepSeek-V3.2. Same-Group Bulk also remains above UNFOLD.

Figure 4 complements the aggregate metrics with turn-by-turn evidence trajectories. For each profile, the trajectory is the cumulative fraction of unique profile evidence expressed by each patient turn; shorter sessions are carry-forward padded, and trajectories are then averaged across profiles at each raw turn. Shaded bands show pointwise 95% profile-level bootstrap confidence intervals for the mean trajectory based on 1,000 resamples. Full-profile simulators express more evidence early and approach a plateau, whereas UNFOLD introduces evidence more gradually over a longer portion of the interaction.

The principal information-trajectory differences therefore persist when the baselines are given substantially more opportunity to continue. Because the full-profile simulators may terminate through generated ending decisions whereas UNFOLD uses selector-coverage stopping, this analysis reduces but does not eliminate termination-related confounding.

E.2 RQ2: PROFILE-LEVEL ROBUSTNESS

The main long-horizon comparison uses 30 matched profiles under the original approximately 60-turn baseline setting. Every system is evaluated on the same profile set, with dialogue treated as the statistical unit. We examine both profile-level paired differences and sensitivity to removing any single evaluated profile.

For profile i, metric m, and baseline b, define the oriented paired difference

$$\Delta_{i,m,b} = \sigma_m \left(s_{i,m}^{\mathrm{UNFOLD}} - s_{i,m}^{b} \right), \tag{23}$$

Table 15: **Effect of relaxing the full-profile baseline horizon.**

Model	Max turns	Avg. turns	Depth ↑	Breadth ↑	Restr. ↑	Rep. ↓	First-Seen ↑	Bulk ↓
GPT-4o-mini	60	60.00	$.874 \pm .042$	$.188 \pm .056$	$.246 \pm .101$	$.745 \pm .044$	$.170 \pm .029$	$2.57 \pm .22$
Gemini-2.5-Pro	60	55.30	$.723 \pm .070$	$.359 \pm .064$	$.252 \pm .099$	$.576 \pm .036$	$.185 \pm .031$	$3.13 \pm .33$
DeepSeek-V3.2	60	58.63	$.671 \pm .073$	$.380 \pm .069$	$.103 \pm .068$	$.687 \pm .035$	$.161 \pm .028$	$3.37 \pm .22$
GPT-4o-mini	200	195.87	$.852 \pm .053$	$.190 \pm .055$	$.246 \pm .117$	$.977 \pm .018$	$.141 \pm .026$	$3.53 \pm .28$
Gemini-2.5-Pro	200	172.00	$.712 \pm .074$	$.367 \pm .066$	$.238 \pm .117$	$.964 \pm .031$	$.126 \pm .040$	$3.20 \pm .60$
DeepSeek-V3.2	200	132.37	$.628 \pm .073$	$.437 \pm .063$	$.107 \pm .073$	$.681 \pm .035$	$.126 \pm .019$	$2.90 \pm .24$
UNFOLD	–	132.70	$.712 \pm .049$	$.558 \pm .065$	$.632 \pm .085$	$.528 \pm .033$	$.330 \pm .029$	$1.63 \pm .20$

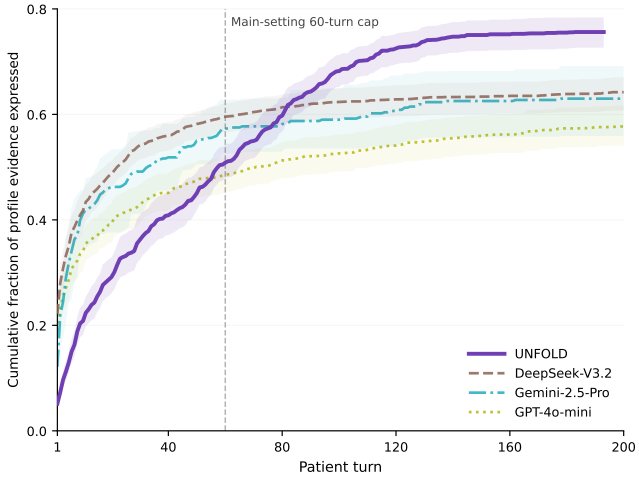


Figure 4: **Profile-evidence unfolding under extended horizons.** Curves show the cumulative fraction of profile evidence expressed over patient turns. The dashed vertical line marks the 60-patient-turn cap used for the full-profile baselines in the main setting.

where $\sigma_m = 1$ for Depth, Breadth, Disclosure Restraint, and First-Seen, and $\sigma_m = -1$ for Repetition and Same-Group Bulk. Thus, positive values consistently favor UNFOLD. For each profile i, the leave-one-profile-out estimate is

$$\bar{\Delta}_{-i,m,b} = \frac{1}{29} \sum_{j \neq i} \Delta_{j,m,b}. \tag{24}$$

Figure 5 reports both the profile-level paired differences and the leave-one-profile-out estimates.

Figure 5 shows that the principal information-trajectory differences are distributed across profiles rather than driven by a single extreme case. UNFOLD's mean advantages in Breadth, Disclosure Restraint, First-Seen, Repetition, and Same-Group Bulk remain positive after leaving out any one profile for all three full-profile baselines. The profile-level distributions nevertheless show heterogeneous effect sizes, particularly for Repetition and Same-Group Bulk.

Depth follows a different pattern. GPT-4o-mini is higher on most profiles, whereas comparisons with Gemini-2.5-Pro and DeepSeek-V3.2 are substantially closer to zero. The LOPO analysis therefore supports robustness of the aggregate information-trajectory differences to removal of any single profile without implying uniform superiority on every metric or profile.

E.3 RQ3: Complete Ablation Results

Table 16 reports the complete ablation results over the same 30 matched profiles. The first three columns measure *turn-level patient-state realization*: for each real patient turn associated with these profiles, the ablated system generates a response from the corresponding dialogue history, which

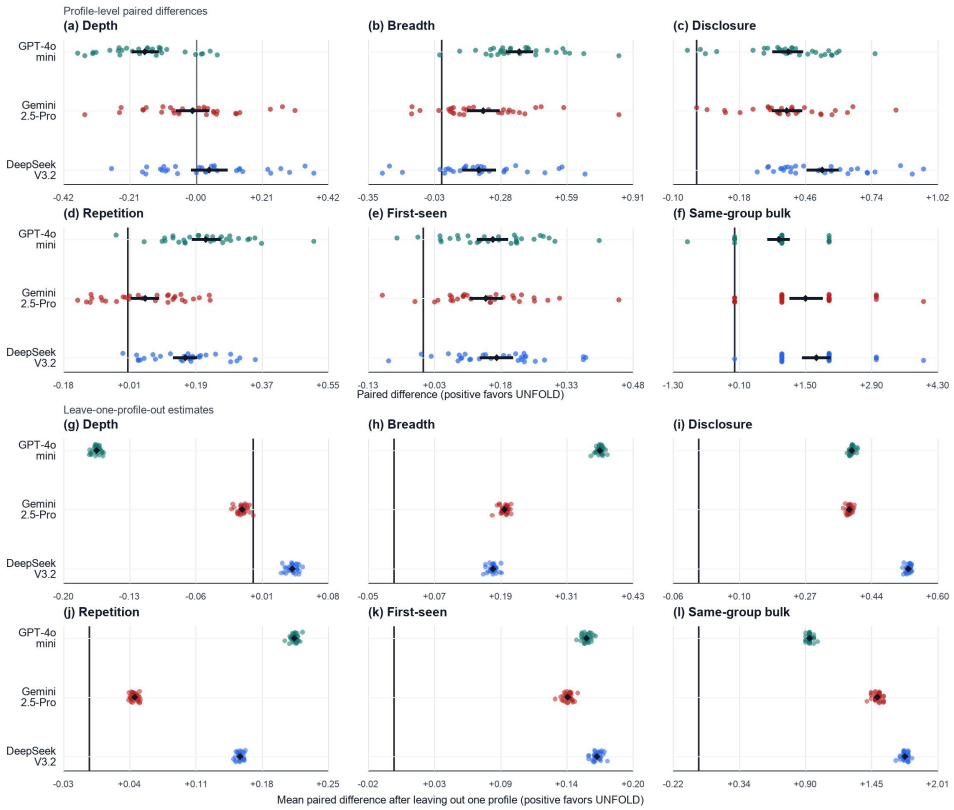


Figure 5: **Profile-level robustness under the main 60-turn baseline setting.** Top: profile-level paired differences between UNFOLD and each full-profile baseline over the 30 matched profiles. Positive values consistently favor UNFOLD; signs are reversed for Repetition and Same-Group Bulk because lower values are preferred. Black diamonds denote full-sample mean paired differences. Bottom: mean paired differences after leaving out each profile in turn. Each point is one leave-one-profile-out estimate; black diamonds denote the full-sample estimates, and the vertical line marks zero.

is annotated post hoc with Qwen2.5-72B-Instruct and compared with the original gold attitude, emotion, and topic labels using Micro-F1. These scores therefore differ from the next-state prediction results over all 8,397 held-out turns in Table 1. The remaining columns are dialogue-level information-trajectory metrics computed from complete simulated sessions.

In − DYNAMICS, predicted attitude, emotion, and topic controls are removed from response generation, while predicted patient topic remains available to profile selection. In − SELECTION, history-aware progressive selection is replaced by static topic-based selection while state prediction and the disclosure bottleneck are retained. In − BOTTLENECK, state prediction and progressive selection are retained but the one-primary-evidence restriction is removed. All variants use the same post-hoc state annotation and profile-matching procedures.

E.4 RQ4: THERAPIST EVALUATION RESULTS

Table 17 reports the majority outcomes for the blinded pairwise therapist evaluation. Each model–dimension cell contains seven profile-matched evaluation items, with the three therapist judgments aggregated by majority vote. Across all 84 items, UNFOLD receives 61 wins, 18 ties, and 5 losses, corresponding to a tie-adjusted preference of .833.

For transparency, Table 18 reports both the aggregated individual therapist judgments and chance-corrected inter-rater agreement. The individual ratings are reported descriptively rather than treated as independent observations, because the three judgments for each evaluation item refer to the same

Table 16: **Complete mechanism-targeted ablation results.**

(a) Turn-Level State Realization

Variant	Att. ↑	Emo. ↑	Topic ↑
Full	.551	.459	.561
− Dynamics	.517	.285	.302
− Selection	.548	.434	.535
− Bottleneck	.539	.415	.549

(b) Dialogue-Level Information Trajectory

Variant	Depth ↑	Breadth ↑	Restr. ↑	Rep. ↓	First-Seen ↑	Bulk ↓
Full	$.712 \pm .049$	$.558 \pm .065$	$.632 \pm .085$	$.528 \pm .033$	$.330 \pm .029$	$1.63 \pm .20$
− Dynamics	$.715 \pm .059$	$.541 \pm .070$	$.598 \pm .086$	$.518 \pm .040$	$.328 \pm .034$	$1.67 \pm .15$
− Selection	$.722 \pm .075$	$.307 \pm .071$	$.280 \pm .101$	$.653 \pm .049$	$.124 \pm .021$	$1.97 \pm .12$
− Bottleneck	$.708 \pm .079$	$.509 \pm .077$	$.358 \pm .058$	$.723 \pm .062$	$.231 \pm .030$	$2.17 \pm .20$

Table 17: **Majority outcomes for therapist evaluation.** Each model–dimension cell contains seven evaluation items. W/T/L denotes the number of majority-vote wins, ties, and losses for UNFOLD, and Preference assigns half credit to ties.

Compared with	Dimension	W	T	L	Preference
GPT-4o-mini	Response Dynamics	1	6	0	.571
	Progressive Disclosure	2	5	0	.643
	Long-Horizon Plausibility	2	4	1	.571
	Interaction Realism	4	3	0	.786
	All dimensions	**9**	**18**	**1**	**.643**
Gemini-2.5-Pro	Response Dynamics	7	0	0	1.000
	Progressive Disclosure	6	0	1	.857
	Long-Horizon Plausibility	5	0	2	.714
	Interaction Realism	6	0	1	.857
	All dimensions	**24**	**0**	**4**	**.857**
DeepSeek-V3.2	Response Dynamics	7	0	0	1.000
	Progressive Disclosure	7	0	0	1.000
	Long-Horizon Plausibility	7	0	0	1.000
	Interaction Realism	7	0	0	1.000
	All dimensions	**28**	**0**	**0**	**1.000**
Overall	**All evaluation items**	**61**	**18**	**5**	**.833**

underlying system comparison. Across the 252 individual judgments, UNFOLD receives 178 wins, 55 ties, and 19 losses, corresponding to a tie-adjusted preference of 81.5%. Overall Fleiss' κ is 0.504, with dimension-specific agreement ranging from 0.418 to 0.578.

Table 18: **Additional therapist-evaluation statistics.** Left: aggregated individual therapist judgments. Right: chance-corrected inter-rater agreement. Individual judgments are reported descriptively because ratings of the same evaluation item are not independent observations.

(a) Aggregated Individual Judgments

Dimension	Ratings	Win	Tie	Loss	Pref.
Response Dynamics	63	41	18	4	79.4%
Progressive Disclosure	63	48	13	2	86.5%
Long-Horizon Plausibility	63	42	14	7	77.8%
Interaction Realism	63	47	10	6	82.5%
Overall	**252**	**178**	**55**	**19**	**81.5%**

(b) Inter-Rater Agreement

Dimension	Fleiss' κ
Response Dynamics	0.418
Progressive Disclosure	0.578
Long-Horizon Plausibility	0.486
Interaction Realism	0.534
Overall	**0.504**

The majority-vote results and individual-judgment summaries provide complementary views of the evaluation. The former treats each underlying system comparison as the evaluation unit, whereas the latter exposes the distribution of the three therapist judgments before majority aggregation. Agreement is moderate overall, with variation across dimensions, and the individual judgments should therefore not be interpreted as 252 independent comparisons.