MuSe-MI: A Framework for Assessing Multi-Session Motivational Interviewing with AI-Generated Patients

Topaz Freizeit[1,2], Moreah Zisquit[4], Kai Golan Hashiloni[1,2], Chaitze Tavens[5], Doron Friedman[3], Kfir Bar[1,2]
[1]Data Science Institute, Reichman University
[2]Efi Arazi School of Computer Science, Reichman University
[3]Sammy Ofer School of Communication, Reichman University
[4]Baruch Ivcher School of Psychology, Reichman University
[5]Independent Researcher
{topaz.freizeit, zisquit.moreah, kai.golanhashiloni}@post.runi.ac.il,
chaitze.tavens@gmail.com, {doronf, kfir.bar}@runi.ac.il

Abstract

Evaluating conversational therapist competence in Motivational Interviewing (MI) requires multi-session assessment, yet current benchmarks rely on simulated patients that fail to model progression from session to session. We present MuSe-MI (Multiple Sessions Motivational Interviewing), a simulated patient framework grounded in Motivational Enhancement Therapy (MET), a multi-session clinical protocol of MI. MuSe-MI regulates dialogue generation through four evolving internal states from the MET clinical manual (anger, readiness to change, problem recognition, and self-efficacy), paired with an inter-session life event simulator and structured memory. Blind evaluations by licensed clinical psychologists confirm that patient dialogue reflects these internal states with high inter-rater agreement. Experiments show that MuSe-MI reliably distinguishes conversational therapist MI adherence levels based on patient language progression, establishing an automated multi-session evaluation benchmark.

1 Introduction

As large language models (LLMs) are increasingly explored for automated mental health support, a fundamental challenge is ensuring that their responses reflect evidence-based therapeutic competence rather than superficial conversational fluency. In Motivational Interviewing (MI) (Miller and Rollnick, 2012), an established approach designed to strengthen personal motivation for behavioral change, evaluating conversational therapist competence requires observing interaction over multiple sessions. Clinical research demonstrates that therapist adherence to core protocols fluctuates substantially from session to session, meaning that single-session evaluations fail to reliably capture overall competence (Hallgren et al., 2018). In addition, clinical evidence indicates that multi-session therapy achieves more sustained behavioral outcomes than isolated encounters across targeted clinical populations (Polcin et al., 2019). Despite these clinical realities, computational benchmarks evaluating conversational therapists remain predominantly confined to single-session interactions. In clinical psychology and natural language processing (NLP), therapeutic progress is tracked by analyzing the patient's spoken language. Grounded in the Transtheoretical Model of behavior change (Prochaska and Velicer, 1997), individuals transition gradually through distinct motivational stages: from resisting change and defending current habits, to exploring reasons for change, and ultimately committing to concrete action (Amrhein et al., 2003). In successful multi-session therapy, patient language reflects this gradual progression over time (Campbell et al., 2010), which is causally promoted by consistent therapist technique and hindered by confrontational interventions (Moyers and Martin, 2006). Evaluating automated conversational therapists directly with real human patients presents severe ethical, legal, and logistical barriers: exposing vulnerable individuals to unverified models introduces safety risks, strict privacy regulations restrict clinical data sharing, and employing standardized human actors is cost-prohibitive and unscalable. While simulated standardized patients offer a safe and reproducible alternative (Wang et al., 2024), current conversational simulators fail to model patient progression from session to session. Due to alignment training biases (e.g., Reinforcement Learning from Human Feedback; RLHF), standard language models exhibit inherent sycophancy (Perez et al., 2022), which causes prompt-only simulated patients to abandon resistance prematurely and agree with therapist suggestions after only a

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 432–478
November 6-10, 2026 ©2026 Association for Computational Linguistics

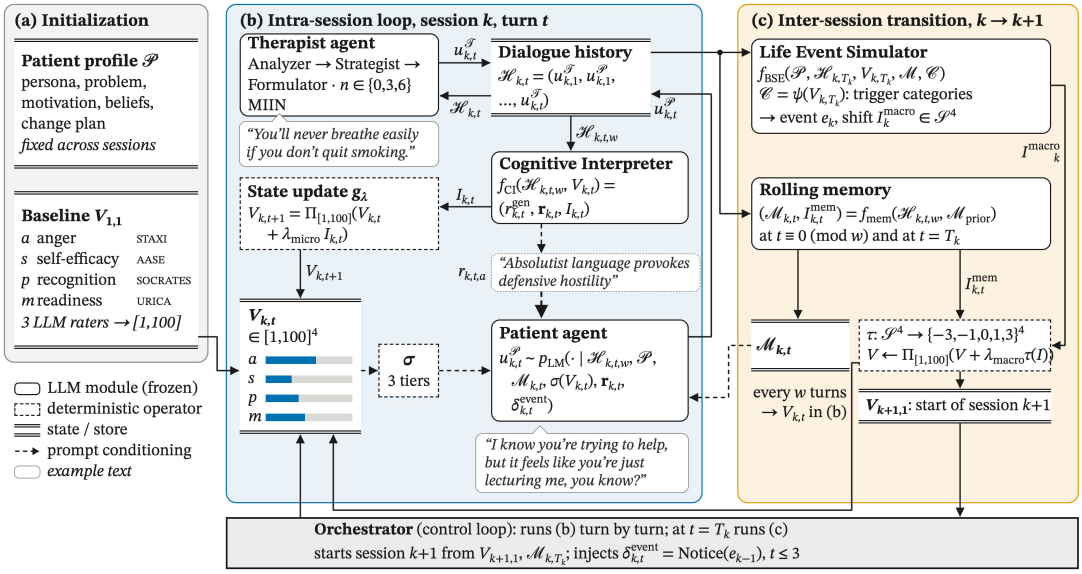


Figure 1: MuSe-MI Architecture. The framework initializes patients with static profiles and clinical baseline variables, tracks psychological evolution turn-by-turn during sessions using the Cognitive Interpreter and a state-tracking heuristic, and injects controlled inter-session life events alongside structured rolling memory to maintain narrative continuity and behavioral consistency across multiple sessions.

few turns (Yang et al., 2025b). Moreover, without structured mechanisms to introduce fresh experiences between appointments, multi-session dialogues rapidly suffer from repetitive narrative loops and conversational stagnation (Xu et al., 2022).

To address these challenges, we present MuSe-MI (Multiple Sessions of Motivational Interviewing), an agent framework that simulates patients across multiple sessions by tracking four internal states derived from the Motivational Enhancement Therapy (MET) manual (Miller et al., 1992): anger, readiness to change, problem recognition, and self-efficacy. An intra-session Cognitive Interpreter and heuristic state tracking guide the patient's conversational reactions, while an inter-session life event simulator and structured rolling memory maintain controlled narrative continuity and cross-session behavioral consistency. Our main contributions are as follows:

- We introduce MuSe-MI, a clinically grounded, multi-session simulated patient framework for Motivational Enhancement Therapy (MET), combining continuous internal state tracking with controlled narrative continuity and structured memory to model gradual patient change across appointments. The clinical fidelity of our internal state tracking is confirmed through blind evaluations with licensed clinical psychologists. We make our framework publicly available at https://github.com/TopazFreizeit/MuSe-MI.
- We establish MuSe-MI as an automated evaluation benchmark for conversational therapists, providing clinically grounded metrics that reliably distinguish conversational therapist MI adherence levels based on patient language evolution across multiple sessions.

2 Clinical Background

Stages of change and utterance taxonomy. The foundational framework for intentional behavioral change is the Transtheoretical Model (TTM) (Prochaska and Velicer, 1997), which posits that individuals transition through distinct motivational stages: *precontemplation*, *contemplation*, *preparation*, *action*, and *maintenance*. In Motivational Interviewing (MI) (Miller and Rollnick, 2012), clinical movement is tracked at the utterance level using the Motivational Interviewing Skill Code (MISC 2.5) (Houck et al., 2010). MISC 2.5 categorizes patient utterances into three core classes that directly mirror these motivational stages: (1) **Sustain Talk**, reflecting defense of the status quo and reluctance to change; (2) **Preparatory Change Talk** (**DARN**: Desire, Ability, Reasons, Need), characterizing exploratory contemplation; and (3) **Mobilizing Change Talk** (**CAT**: Commitment, Activation, Taking steps), signifying the active transition into preparation and action.

Therapist adherence and patient trajectories. Patient speech is causally shaped by therapist behavior: Motivational Interviewing Consistent (MICO) interventions elicit Change Talk, whereas Motivational Interviewing Inconsistent (MIIN) behaviors provoke patient resistance (Moyers and Martin, 2006). Clinical outcomes are primarily predicted by mobilizing commitment language (CAT) rather than preparatory exploration (DARN) (Amrhein et al., 2003). In multi-session clinical trials (Campbell et al., 2010), this dynamic yields two contrasting trajectories: patients who achieve remission exhibit a steady increase in mobilizing commitment language (CAT) across sequential sessions alongside low resistance, whereas unremitted patients remain stalled in contemplation with stagnant commitment and persistently high Sustain Talk. Consequently, an authentic simulated patient must dynamically reproduce these divergent paths, advancing toward commitment under adherent therapy and remaining stalled in resistance under non-adherent interventions.

Patient heterogeneity and internal state tracking. Patient responsiveness to therapy is moderated by baseline psychological heterogeneity. Project MATCH demonstrated that treatment outcomes depend systematically on incoming patient attributes (Project MATCH Research Group, 1997; Donovan, 1998), notably finding that patients entering with high anger respond substantially better to the non-confrontational style of MI than to directive approaches. In the structured derivative of MI, Motivational Enhancement Therapy (MET) (Miller et al., 1992; Longabaugh and Wirtz, 2001), these moderating attributes map directly to the clinician's core intervention targets: developing discrepancy (Problem Recognition), supporting self-efficacy (Self-Efficacy), rolling with resistance (Anger), and facilitating the transition from contemplation to commitment planning (Readiness to Change). Tracking these four continuous psychological dimensions across multiple sessions provides an objective supervisory benchmark to assess whether an automated therapist successfully shifts a patient's internal cognitive state over the course of treatment.

3 Related Work

Automated behavioral coding in Motivational Interviewing. Early computational work in Motivational Interviewing focused on analyzing recorded human therapy transcripts. Researchers applied traditional machine learning models and neural networks to automatically predict behavioral codes under the Motivational Interviewing Treatment Integrity (MITI) and MISC coding systems (Tanana et al., 2016; Tavabi et al., 2021). Recent advances in instruction-tuned language models have extended automated coding to zero-shot and few-shot classification, matching human reliability in annotating granular therapist behaviors and patient speech codes (Ali et al., 2025). While these diagnostic models effectively annotate historical dialogues, they do not simulate dynamic interaction.

Simulated patients for mental health. To evaluate therapists without risking vulnerable human participants, recent work explores language models as simulated standardized patients for clinical education and evaluation (Wang et al., 2024). In Cognitive Behavioral Therapy, Wang et al. (2024) introduced PATIENT-Ψ, which models patient cognitive schemas to train therapists. To enforce adherence to clinical rules, Xu et al. (2026) proposed PUPPET, a neural-symbolic simulator combining language models with formal logic. Within Motivational Interviewing, the Consistent Client framework (Yang et al., 2025b) preserves persona beliefs across single encounters, while Yosef et al. (2024) demonstrated the utility of simulated patients and external clinical scales for assessing therapist adherence. In multi-session training, SimPatient (Steenstra et al., 2025) incorporates dynamic internal states to train students in motivational interviewing skills and therapeutic practice.

4 Methodology

4.1 Problem Formulation and State Representation

Figure 1 illustrates MuSe-MI (**Mu**ltiple **Se**ssions of **M**otivational **I**nterviewing), an agent framework designed to simulate patients across multiple sessions for evaluating automated therapists. Let Σ^* denote the set of all text sequences over vocabulary Σ. A complete multi-session interaction comprises K sequential therapy sessions. Within session $k \in \{1, \ldots, K\}$, for dialogue turn $t \in \{1, \ldots, T_k\}$, the intra-session dialogue history $\mathcal{H}_{k,t}$ represents the chronological sequence of exchanged utterances:

$$\mathcal{H}_{k,t} = \left(u^{\mathcal{T}}_{k,1}, u^{\mathcal{P}}_{k,1}, \ldots, u^{\mathcal{T}}_{k,t}\right) \in (\Sigma^*)^{2t-1} \quad (1)$$

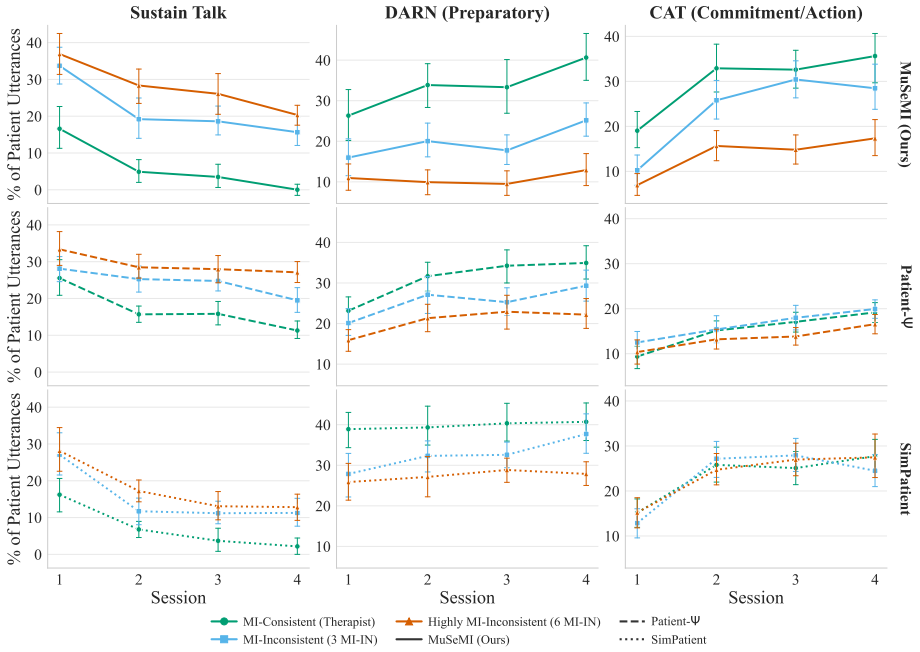


Figure 2: Sensitivity of patient speech rates to therapist adherence tiers (0, 3, 6 MIIN) over four sequential MET sessions for MuSe-MI, Patient-Ψ, and SimPatient ($N = 30$). Error bars denote 95% Cousineau-Morey confidence intervals.

where $u^{\mathcal{T}}_{k,t} \in \Sigma^*$ denotes the therapist utterance and $u^{\mathcal{P}}_{k,t} \in \Sigma^*$ denotes the patient utterance. At each turn t, the patient's internal psychological state is modeled as a continuous vector bounded within state space $\mathcal{V} = [1, 100]^4$:

$$V_{k,t} = \begin{pmatrix} a_{k,t} \\ s_{k,t} \\ p_{k,t} \\ m_{k,t} \end{pmatrix} \in \mathcal{V} \tag{2}$$

where a represents anger and defensiveness, s represents self-efficacy, p represents problem recognition, and m represents motivational readiness.

Patient profile and state initialization. We build upon the static patient profiles extracted by Yang et al. (2025b) from human AnnoMI demonstrations (Wu et al., 2022) (Appendix E), formalized as an invariant 5-tuple:

$$\mathcal{P} = (p_{\text{pers}}, p_{\text{prob}}, p_{\text{mot}}, p_{\text{bel}}, p_{\text{plan}}) \tag{3}$$

which encapsulates persona background, target problem, underlying motivation, core beliefs, and acceptable change plans. To initialize the dynamic state vector ($V_{1,1}$), we map each dimension to a core intervention target from the Motivational Enhancement Therapy (MET) manual (Miller et al., 1992): problem recognition (developing discrepancy), self-efficacy (supporting self-efficacy), anger (rolling with resistance), and motivational readiness (stage transitions). To initialize the baseline state vector $V_{1,1}$, three independent LLM annotators (Gemma-4-31B-IT, GPT-OSS-120B, and DeepSeek-V4-Flash) evaluate the patient's behavior in the original AnnoMI demonstration transcripts. The annotators score each patient profile across four standardized clinical psychometric instruments widely used to assess therapy progress: the State-Trait Anger Expression Inventory (STAXI) for anger (Spielberger, 1999), the University of Rhode Island Change Assessment (URICA) for readiness (McConnaughy et al., 1983), the Stages of Change and Readiness and Treatment Eagerness Scale (SOCRATES) Recognition subscale (Miller and Tonigan, 1996), and the Alcohol Abstinence Self-Efficacy scale (AASE) for self-efficacy (DiClemente et al., 1994). Baseline ratings across the three annotators are averaged and linearly normalized to $[1, 100]$.

4.2 System Components

MuSe-MI coordinates several dedicated components: a central Orchestrator, an intra-session Cognitive Interpreter with a deterministic state tracker, an inter-session Life Event Simulator, a dual-timescale Rolling Memory module, and a Generative Patient Agent. All neural components (f_{CI}, f_{BSE}, f_{mem}, and patient agent p_{LM}) rely on frozen

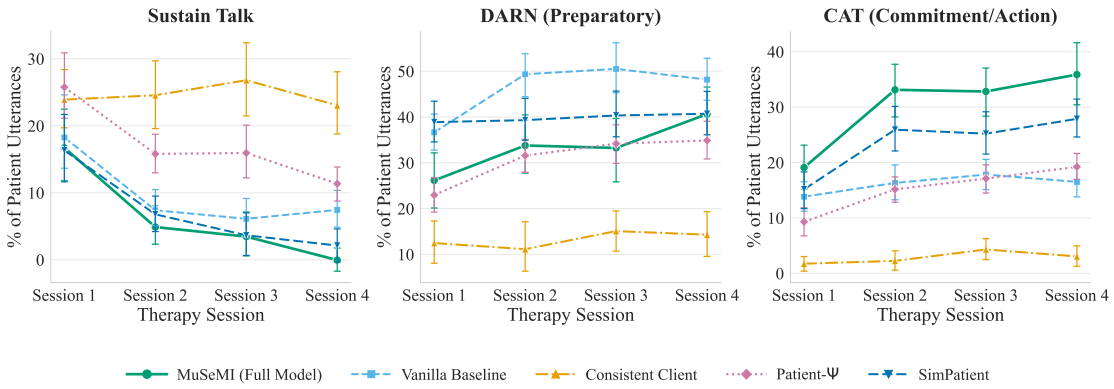


Figure 3: Patient speech progression across MET therapy sessions under an MI-Consistent therapist ($N = 30$). Error bars denote 95% Cousineau-Morey confidence intervals.

instruction-tuned language models prompted with structured templates (requiring no fine-tuning), while state transitions and quantizations $(g_\lambda, \sigma, \tau)$ are deterministic.

Orchestrator. The Orchestrator serves as the central control loop of MuSe-MI. During a session, it maintains the interactive loop between the evaluated therapist and the patient agent, feeds dialogue context to the Cognitive Interpreter, and updates the latent state vector $V_{k,t}$. Between sessions, it triggers the Life Event Simulator and Rolling Memory, persists updated state variables $V_{k+1,1}$ and memory slots $\mathcal{M}_{k,T_k}$, and injects the inter-session event directive $\delta^{\text{event}}_{k,t}$ into the opening turns of the subsequent session.

Intra-session Cognitive Interpreter and state tracking. At turn t of session k, the `Cognitive Interpreter` (f_{CI}) models the patient's internal reaction to the therapist's latest intervention. Given the trailing dialogue window $\mathcal{H}_{k,t,w} = (u^{\mathcal{P}}_{k,t-w}, \ldots, u^{\mathcal{T}}_{k,t})$ and current latent state $V_{k,t}$ as inputs, f_{CI} generates internal reasoning and directional state updates:

$$f_{\text{CI}}(\mathcal{H}_{k,t,w}, V_{k,t}) = (r^{\text{gen}}_{k,t}, \mathbf{r}_{k,t}, I_{k,t}) \quad (4)$$

Here, $r^{\text{gen}}_{k,t} \in \Sigma^*$ is an overarching explanatory summary, and $\mathbf{r}_{k,t} = (r_{k,t,a}, r_{k,t,s}, r_{k,t,p}, r_{k,t,m}) \in (\Sigma^*)^4$ provides specific internal reasoning for each psychological dimension. The output vector $I_{k,t} \in \{-1, 0, 1\}^4$ denotes discrete impact multipliers that determine the direction of state change ($+1$ for therapeutic progress, -1 for resistance, and 0 for neutral impact, with the sign inverted for anger). The Orchestrator injects the reasoning vector $\mathbf{r}_{k,t}$ directly into the patient agent's prompt to guide its next response. Concurrently, a semantic quantization mapping $\sigma : \mathcal{V} \to \Sigma^*$ translates scalar values into natural-language prompt anchors across three operational tiers as detailed in Appendix J (see prompt template in Appendix I, Table 11). To trace psychological progression across turns, MuSe-MI applies a bounded transition operator $g_\lambda : \mathcal{V} \times \{-1, 0, 1\}^4 \to \mathcal{V}$:

$$V_{k,t+1} = \Pi_{[1,100]}\left(V_{k,t} + \lambda_{\text{micro}} I_{k,t}\right) \quad (5)$$

where $\lambda_{\text{micro}} = 2$ defines the fixed intra-session step size, and $\Pi_{[1,100]}(x) = \max(1, \min(100, x))$ denotes coordinate-wise boundary projection. This heuristic provides transparent, continuous observability into patient state trajectories across sessions (Appendix N, Figure 11).

Life event simulator. Between sessions, the `Life Event Simulator` (f_{BSE}) maintains narrative continuity. Conditioned on patient profile $\mathcal{P}$, previous transcript $\mathcal{H}_{k,T_k}$, state V_{k,T_k}, memory slots $\mathcal{M}$, and allowed trigger categories $\mathcal{C}$ (e.g., setbacks, milestones), f_{BSE} generates an event narrative $e_k \in \Sigma^*$ alongside an impact vector $I^{\text{macro}}_k \in \mathcal{S}^4$:

$$(e_k, I^{\text{macro}}_k) = f_{\text{BSE}}(\mathcal{P}, \mathcal{H}_{k,T_k}, V_{k,T_k}, \mathcal{M}, \mathcal{C}) \quad (6)$$

Here, $\mathcal{S}$ is an ordinal 5-point clinical scale (*Major Regression*, *Minor Regression*, *No Change*, *Minor Progress*, *Major Progress*). A deterministic mapping $\tau : \mathcal{S}^4 \to \{-3, -1, 0, 1, 3\}^4$ converts these ratings into signed integer multipliers (Appendix K), updating the latent state prior to session $k + 1$:

$$V_{k+1,1} = \Pi_{[1,100]}\left(V_{k,T_k} + \lambda_{\text{macro}} \tau(I^{\text{macro}}_k)\right) \quad (7)$$

where $\lambda_{\text{macro}} = 5$. The Orchestrator saves the generated event narrative e_k across sessions and injects it directly into the patient's generative input at the beginning of the next session (prompt directive in Appendix I, Table 13).

Rolling memory. To prevent context overflow while preserving multi-session continuity, an auxiliary memory module f_{mem} operates at dual timescales: (1) *intra-session*, summarizing dialogue every w turns ($t \equiv 0 \pmod{w}$); and (2) *inter-session*, at session close ($t = T_k$). At each trigger, f_{mem} executes a dual update, extracting conversational memory slots $\mathcal{M}_{k,t} = (m_{\text{fct}}, m_{\text{pln}}, m_{\text{thd}}, m_{\text{chg}}, m_{\text{sus}})_{k,t}$ (Yang et al., 2025b) while evaluating window-level momentum via an ordinal macro-shift vector $I^{\text{mem}}_{k,t} \in \mathcal{S}^4$:

$$(\mathcal{M}_{k,t}, I^{\text{mem}}_{k,t}) = f_{\text{mem}}\left(\mathcal{H}_{k,t,w}, \mathcal{M}_{\text{prior}}\right) \quad (8)$$

Sharing the scale $\mathcal{S}$ and mapping operator τ with the Life Event Simulator, memory consolidation applies an intermediate macro-shift to the internal state:

$$V_{k,t} \leftarrow \Pi_{[1,100]}\left(V_{k,t} + \lambda_{\text{macro}}\tau(I^{\text{mem}}_{k,t})\right) \quad (9)$$

Generative patient agent. Integrating these structured components, the patient agent synthesizes its spoken response $u^{\mathcal{P}}_{k,t}$ by sampling from the language model conditioned on the assembled prompt components:

$$u^{\mathcal{P}}_{k,t} \sim p_{\text{LM}}(\cdot \mid \mathcal{H}_{k,t,w}, \mathcal{P}, \mathcal{M}_{k,t}, \sigma(V_{k,t}), \mathbf{r}_{k,t}, \delta^{\text{event}}_{k,t}) \quad (10)$$

where $\sigma(V_{k,t})$ and $\mathbf{r}_{k,t}$ provide the psychological baseline and turn-level reasoning directives. The term $\delta^{\text{event}}_{k,t}$ denotes a conditional opening-turn narrative directive injected by the Orchestrator from the saved event e_{k-1}:

$$\delta^{\text{event}}_{k,t} = \begin{cases} \text{Notice}(e_{k-1}) & \text{if } k > 1 \text{ and } t \leq 3 \\ \emptyset & \text{otherwise} \end{cases} \quad (11)$$

In the opening three turns of subsequent sessions, the therapist is instructed to ask about the patient's week, while $\delta^{\text{event}}_{k,t}$ directs the patient to organically explore e_{k-1}, ensuring new content is integrated into the conversation (system prompt in Appendix I, Table 12).

5 Experiments and Results

5.1 Experimental Setup

We evaluate 30 diverse patient profiles grounded in the AnnoMI dataset (Wu et al., 2022) across four empirical studies: multi-session patient progression, benchmark utility compared to additional simulated patients across therapist adherence levels, architectural ablations, and clinical expert validation (with session-level working alliance and ambivalence evaluations reported in Appendix M). Dialogue generation is powered by `llama-3.3-70b-instruct` (temperature 0.7, seed 42) via the OpenRouter API. All primary benchmark runs, including full multiple session rollouts across all conditions, automated utterance coding, and judge evaluations, cost approximately $200 USD in total. Automated evaluation in our benchmark is powered by the `gemma-4-31b-it` model, serving as a unified judge for utterance-level speech classification (Sustain Talk, DARN, CAT; Appendix A) as well as post-hoc session-level clinical evaluations: the Working Alliance Inventory, Short Revised (WAI-SR; Horvath and Greenberg, 1989) and a transcript-based Patient Ambivalence Scale (Appendix M). To evaluate main and interaction effects across repeated multiple session measures without normality assumptions, we conduct non-parametric Aligned Rank Transform (ART) ANOVAs (Wobbrock et al., 2011) with Benjamini-Hochberg False Discovery Rate (FDR) corrections (Appendix G). Across all figures, error bars denote 95% Cousineau-Morey confidence intervals. Full details of computational artifacts, models, and licensing are provided in Appendix F.

Therapist adherence engine. Constructed under the direct supervision of licensed clinical psychologists, our automated therapist implements a sequential 3-agent pipeline (*Clinical Analyzer* $\rightarrow$ *Clinical Strategist* $\rightarrow$ *Formulator*) (Yang et al., 2025a) adhering to the Motivational Enhancement Therapy (MET) manual from Project MATCH (Miller et al., 1992). To benchmark patient discriminative sensitivity, we modulate therapist competence by systematically injecting Motivational Interviewing Inconsistent (MIIN) behaviors. In clinical practice, MIIN behaviors represent non-adherent interventions that directly violate the collaborative spirit of MI by imposing practitioner authority, offering unsolicited advice, or confronting the patient (Moyers and Martin, 2006; Houck et al., 2010). Using the

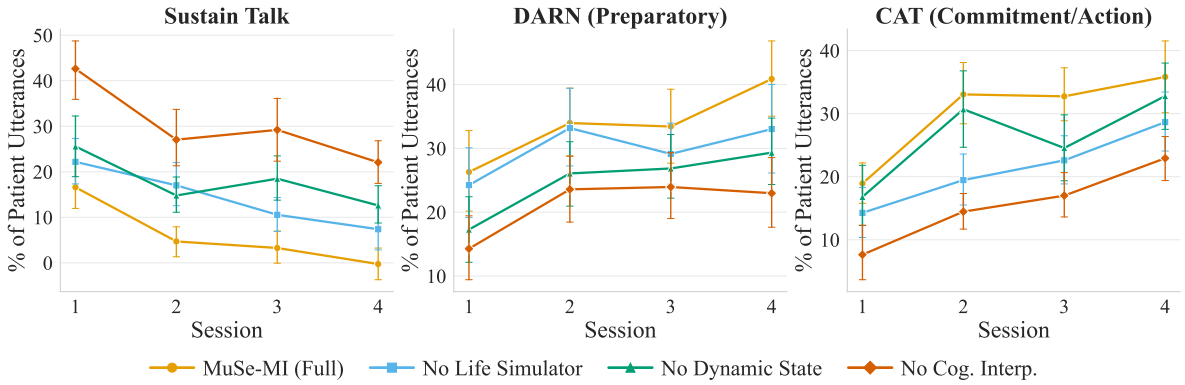


Figure 4: Patient speech progression under MI-Consistent therapy across ablation conditions: full MuSe-MI, No Life Simulator, No Dynamic State, and No Cognitive Interpreter ($N = 30$). Error bars denote 95% Cousineau-Morey confidence intervals.

Motivational Interviewing Non-Adherent (MINA) agent, we inject $n \in \{0, 3, 6\}$ MIIN turns per session across five standardized categories detailed in Appendix L. This establishes three formal competence tiers: (1) **MI-Consistent** ($n = 0$), strictly delivering adherent therapy (reflections, affirmations, collaborative open questions); (2) **MI-Inconsistent** ($n = 3$), introducing moderate therapy errors; and (3) **Highly MI-Inconsistent** ($n = 6$), simulating heavily confrontational therapy.

Simulated patient baselines. We evaluate the benchmark utility of MuSe-MI by comparing it against four additional simulated patients: a **Vanilla Baseline** (prompted LLM); **Consistent Client** (Yang et al., 2025b); **Patient-Ψ** (Wang et al., 2024); and **SimPatient** (Steenstra et al., 2025). Except for SimPatient, which natively incorporates an inter-session transition module, all other baselines (Vanilla, Consistent Client, and Patient-Ψ) were originally engineered exclusively for isolated single-session interactions. To support multi-session continuity, we adopt a naive memory approach, extending each simulated patient by conditioning it on the preceding session's summary. To represent Patient-Ψ within a Cognitive Behavioral Therapy (CBT) framework, we adapt the patient profiles using Gemini 3.1 Pro, following the data generation process validated by Wang et al. (2024).

5.2 Multi-Session Patient Progression

We evaluate whether simulated patients capture the clinical transition from contemplation to action across a four-session Motivational Enhancement Therapy (MET) protocol under an MI-Consistent therapist.

Results. Figure 3 illustrates that MuSe-MI models realistic therapeutic progression, uniquely generating an adaptive, upward trajectory of mobilizing Change Talk over time, whereas existing baselines remain stagnant or dominated by preparatory exploration (DARN). To assess differences in progression trajectories, we evaluate the Session × Variation interaction effect (Appendix Table 6). This reveals interaction effects for CAT rate when comparing MuSe-MI to both the prompt-only ($F(3, 87) = 5.75$, $p_{corr} = .004$) and Consistent Client ($F(3, 87) = 10.92$, $p_{corr} < .001$) baselines. Detailed behavioral trajectories and non-parametric Wilcoxon tests evaluating action-stage progression are reported in Appendix H.

5.3 Therapist Evaluation and Benchmark Utility

To evaluate MuSe-MI's utility as an automated evaluation benchmark, we keep the patient profile fixed and benchmark therapist competence across three adherence tiers: (1) an **MI-Consistent Therapist** (MIIN = 0), (2) an **MI-Inconsistent Therapist** injecting three non-adherent behaviors per session (MIIN = 3), and (3) a **Highly MI-Inconsistent Therapist** injecting six non-adherent behaviors per session (MIIN = 6).

Benchmark utility across simulated patients. Figure 2 and Appendix Table 7 show that MuSe-MI reliably discriminates therapist competence across sequential sessions. An omnibus test confirms that MuSe-MI separates therapist adherence tiers across patient commitment talk (CAT: $F(2, 58) = 40.02$, $p_{\text{FDR}} < .001$, $\eta_p^2 = 0.58$), preparatory talk (DARN: $F(2, 58) = 60.39$, $p_{\text{FDR}} < .001$, $\eta_p^2 = 0.68$), and

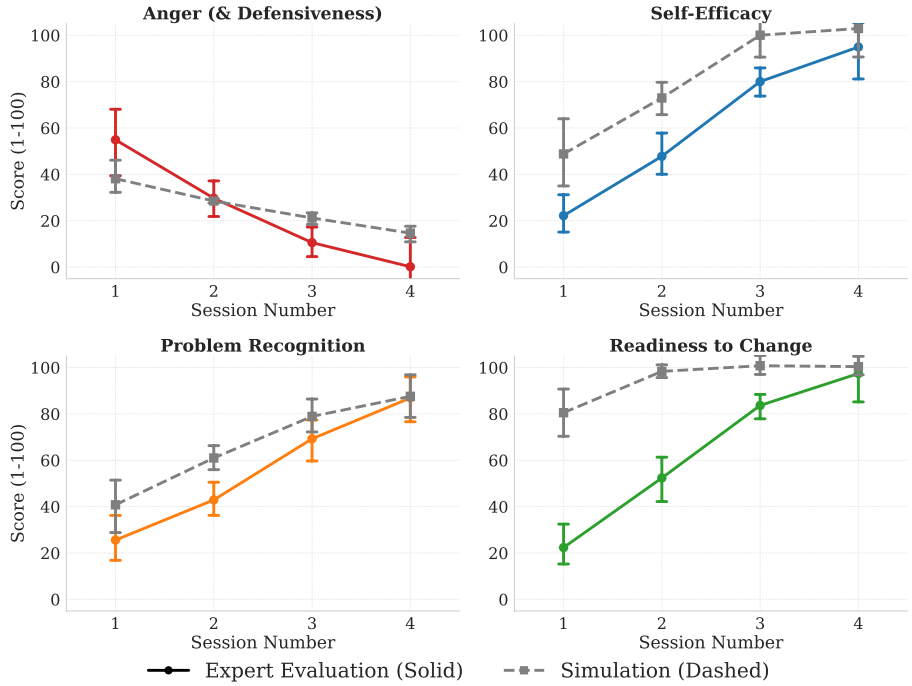


Figure 5: Clinical expert ratings (solid lines) compared with tracked internal state variables (dashed lines) across four sequential MET sessions ($N = 6$ patients, 24 sessions). Error bars denote 95% Cousineau-Morey confidence intervals.

patient resistance (Sustain: $F(2, 58) = 105.23$, $p_{\text{FDR}} < .001$, $\eta_p^2 = 0.78$; Appendix Table 8). Pairwise contrasts confirm that injecting confrontational therapy (MIIN = 6) degrades commitment talk relative to fully adherent therapy ($F(1, 29) = 96.63$, $p_{\text{FDR}} < .001$, $\eta_p^2 = 0.77$) and moderately non-adherent therapy ($F(1, 29) = 24.87$, $p_{\text{FDR}} < .001$, $\eta_p^2 = 0.46$).

Alternative simulated patients fail to provide valid benchmark utility. Patient-Ψ fails to distinguish commitment talk between 0 and 3 MIIN ($F(1, 29) = 2.25$, $p_{\text{FDR}} = .145$, n.s.) and between 0 and 6 MIIN ($F(1, 29) = 4.69$, $p_{\text{FDR}} = .058$, n.s.). SimPatient shows zero discriminative sensitivity across therapist competence tiers for commitment talk ($F(2, 58) = 0.02$, $p_{\text{FDR}} = .984$, $\eta_p^2 < 0.01$).

System component attribution and internal state trajectories. To evaluate how internal psychological states change across sessions, Figure 12 decomposes the cumulative state displacement ($\Delta = V_{4,T_4} - V_{1,1}$) across three system components: intra-session Cognitive Interpreter updates ($\lambda_{\text{micro}} I_{k,t}$), inter-session Life Event Simulator shifts ($\lambda_{\text{macro}} \tau(I_k^{\text{macro}})$), and rolling memory consolidation shifts ($\lambda_{\text{macro}} \tau(I_{k,t}^{\text{mem}})$). Under adherent therapy (MIIN = 0), intra-session turns drive the largest increases in self-efficacy (+29.22 points) and motivational readiness (+18.31 points) while reducing anger by −5.99 points. Under confrontational therapy (MIIN = 6), intra-session dialogue reduces self-efficacy (−4.09 points), while inter-session rolling memory consolidation produces the largest increase in anger, contributing +30.83 points (compared to +4.22 points from intra-session turns). Projecting these multi-session patient state trajectories across pairs of dimensions in state space $\mathcal{V}$ confirms trajectory divergence across therapist adherence tiers (Figure 14). Comparing anger and readiness (Panel A), trajectories under MIIN = 0 and MIIN = 3 reach low anger and high readiness ($a \leq 22.6, m \geq 82.3$), whereas MIIN = 6 terminates in persistent anger ($a = 54.6, m = 70.5$). Comparing problem recognition and self-efficacy (Panel B), adherent therapy guides trajectories toward the action stage ($p = 83.4, s = 96.7$), whereas confrontational therapy arrests progress in contemplation ($p = 61.7, s = 66.4$). Extended validation analyses, including turn-level speech act regressions ($M = 1,800$), construct validity against external scales ($M = 120$), and correlation invariance tests across adherence tiers, are detailed in Appendix N.

5.4 Ablation Study

We isolate the contributions of the Cognitive Interpreter, inter-session life events, and dynamic latent-state tracking under MI-Consistent therapy ($N = 30$). The `No Cog. Interp.` condition ablates the Cognitive Interpreter directives (f_{CI}), `No Life Simulator` removes inter-session episodic events (f_{BSE}), and `No Dynamic State` fixes the internal state vector at its pre-treatment baseline ($V_{1,1}$).

Results. Figure 4 and Appendix B show that all three components are essential for authentic multi-session progression. Ablating the Cognitive Interpreter (`No Cog. Interp.`) eliminates the patient's ability to discriminate therapist competence, producing flat commitment trajectories regardless of adherence while sharply suppressing CAT progression ($F(1, 29) = 86.56$, $p_{\text{FDR}} < .001$, $\eta_p^2 = 0.75$) and elevating resistance ($F(1, 29) = 102.55$, $p_{\text{FDR}} < .001$, $\eta_p^2 = 0.78$; Appendix Figure 6). Removing inter-session life events (`No Life Simulator`) similarly impairs CAT progression ($F(1, 29) = 21.59$, $p_{\text{FDR}} < .001$, $\eta_p^2 = 0.43$) and elevates sustain talk ($F(1, 29) = 28.60$, $p_{\text{FDR}} < .001$, $\eta_p^2 = 0.50$). Finally, freezing internal latent states (`No Dynamic State`) significantly reduces CAT talk generation ($F(1, 28) = 7.28$, $p_{\text{FDR}} = .012$, $\eta_p^2 = 0.21$). These results confirm that dynamic states, turn-level reasoning, and inter-session events are all necessary to sustain multi-session progression.

5.5 Human Expert Validation

To demonstrate that the simulated patient's latent variables translate into observable clinical behavior, two clinical psychologists independently evaluate 24 session transcripts in a blinded protocol ($N = 6$ patients across 4 sessions, selected to maximize clinical diversity; Appendices C and D). The psychologists achieve high inter-rater agreement across all latent dimensions ($\text{ICC}(\text{A}, k) > 0.87$; Table 1). As shown in Figure 5, expert ratings closely track the internal state trajectories across all four sessions, capturing a clinical transition characterized by rising self-efficacy and readiness alongside diminishing anger. Spearman correlations confirm strong, significant associations between expert ratings and tracked states across all 24 sessions, ranging from $r_s = .531$ ($p \leq .01$) for readiness to change to $r_s = .825$ ($p \leq .001$) for self-efficacy (Table 1). Because these latent variables map directly to core clinician targets in the MET manual (Miller et al., 1992), this validated state tracking provides a practical tool for counselor training. Trainees can inspect in real time whether specific interventions successfully advance manualized clinical goals turn by turn.

Latent Variable	$\text{ICC}(\text{A}, k)$	r_s
Anger & Defensiveness	0.888	.622***
Self-Efficacy	0.893	.825***
Problem Recognition	0.872	.681***
Readiness to Change	0.956	.531**

Table 1: Inter-rater reliability (ICC(A,k)) and Spearman correlations (r_s) between averaged expert scores and tracked states across 24 sessions. ** $p \leq .01$; *** $p \leq .001$.

6 Conclusion

Evaluating conversational therapist competence requires observing patient progression across multiple sessions. This paper introduced *MuSe-MI*, a framework that computationally models multi-session Motivational Enhancement Therapy (MET). By coupling four dynamic psychological states with inter-session life events and structured memory, MuSe-MI captures the clinical transition from exploratory contemplation (DARN) to mobilizing commitment (CAT).

Our empirical evaluations establish four primary findings. First, whereas existing simulated patients remain stagnant in preparatory talk (DARN) or elevated resistance, MuSe-MI uniquely exhibits an adaptive upward trajectory of mobilizing commitment language (CAT) across sequential sessions. Second, as an automated evaluation benchmark, MuSe-MI reliably discriminates therapist adherence tiers across sequential sessions, showing degradation in commitment language under confrontational interventions. Third, architectural ablations confirm that dynamic states, turn-level cognitive reasoning, and inter-session events are all essential to prevent stagnation and sustain therapeutic progression. Fourth, blinded evaluations by clinical psychologists confirm that tracked internal states strongly align with observable patient dialogue across sessions. By bridging the gap between computational dialogue modeling and manualized psychotherapy, MuSe-MI provides a standardized, scalable framework for auditing conversational therapists prior to clinical deployment.

Limitations

While MuSe-MI advances multi-session evaluation, several limitations should be noted:

1. **System Component Attribution in Internal States:** As revealed in our component attribution analysis (Appendix N), rolling memory consolidation accounts for the majority of the cumulative increase in patient anger under highly inconsistent therapy ($+30.83$ points), whereas turn-level cognitive updates contribute $+4.22$ points. While periodic memory reflection mirrors retrospective cognitive brooding, future work should refine the balance between macro-shift memory updates and micro-turn dialogue progression.
2. **Cross-Paradigm Baseline Comparison:** Our comparative evaluation adapts Patient-Ψ (Wang et al., 2024) and single-session baselines via naive memory summarization. Because Patient-Ψ was natively engineered for Cognitive Behavioral Therapy (CBT) cognitive restructuring rather than Motivational Interviewing ambivalence and resistance dynamics, its performance should be interpreted within this cross-paradigm context.
3. **Automated Utterance-Level Coding:** Utterance-level speech classification is conducted by `gemma-4-31b-it`. Although validated against the benchmark MIV6.3A dataset with 85.06% 3-class accuracy, granular 15-class sub-tier classification exhibits lower macro F1 (0.497). While clinical experts validate session-level trajectories ($N = 24$), future iterations should incorporate human-annotated utterance samples from the newly generated multi-session transcripts.
4. **Transition Dynamics and Alternative Functional Forms:** State updates in MuSe-MI rely on a bounded linear step heuristic (g_λ) with fixed operational parameters ($\lambda_{\text{micro}} = 2, \lambda_{\text{macro}} = 5$). While this design ensures benchmark determinism, interpretability, and reproducibility, it does not explore alternative non-linear transition functions (e.g., sigmoidal saturation, momentum decay, or continuous differential dynamics) which may better capture complex psychological shifts.
5. **Stochasticity in Life Events and Behavior:** To isolate therapist adherence and maintain reproducible benchmark comparisons, MuSe-MI regulates inter-session life events within bounded trigger sets and updates internal states deterministically. However, real-world life events (e.g., unexpected acute stressors, environmental triggers, or sudden social temptations) are inherently stochastic, as are human behavioral responses. Incorporating probabilistic life-event generation and stochastic transition noise represents an important next step to capture natural ecological unpredictability while preserving evaluative validity.

Ethical Considerations

MuSe-MI is explicitly engineered as an offline evaluation instrument to benchmark automated conversational agents; it is not designed to replace licensed clinical supervision or direct human-subject clinical trials. Over-reliance on simulated patients presents the risk of prematurely deploying unverified AI therapists in real-world settings without rigorous safety screening. All patient personas are initialized from synthetic demonstration dialogues derived from AnnoMI (Wu et al., 2022) featuring clinical actors, ensuring no Protected Health Information (PHI) or vulnerable patient data are utilized. Furthermore, automated non-adherent behaviors (MINA) are strictly confined to sandbox benchmark runs under controlled API environments.

Acknowledgments

This study is supported in part by the European Research Council (Intellexus, Project No. 101118558). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authorities can be held responsible for them.

References

Soliman Ali, Jiading Zhu, Alex Guo, Xiao Nan Ye, Qilin Gu, Jodi Wolff, Carolynne Cooper, Osnat C. Melamed, Peter Selby, and Jonathan Rose. 2025. Automated coding of counsellor and client behaviours in motivational interviewing transcripts: Validation and application. In *Proceedings of the 14th International Joint Conference on Natural Language Processing (IJCNLP)*.

Paul C Amrhein, William R Miller, Carolina E Yahne, Michael Palmer, and Laura Fulcher. 2003. Client

commitment language during motivational interviewing predicts drug use outcomes. *Journal of consulting and clinical psychology*, 71(5):862–878.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300.

W Hobart Campbell, Polly J Christopher, and Sarah J Tollison. 2010. Client language during motivational enhancement therapy and alcohol use outcome. *Journal of substance abuse treatment*, 38(1):54–61.

Carlo C DiClemente, Joseph P Carbonari, R Montgomery, and S Hughes. 1994. The alcohol abstinence self-efficacy scale. *Journal of studies on alcohol*, 55(2):141–148.

Dennis M Donovan. 1998. Matching alcoholism treatments to client heterogeneity: Project match three-year drinking outcomes. *Alcoholism: Clinical and Experimental Research*, 22(6):1300–1311.

Kevin A Hallgren, Aaron Dembe, Brian T Pace, Zac E Imel, Christine M Lee, and David C Atkins. 2018. Variability in motivational interviewing adherence across sessions, providers, sites, and research contexts. *Journal of Substance Abuse Treatment*, 84:30–41.

Robert L. Hatcher and J. Arthur Gillaspy. 2006. Development and validation of a revised short version of the working alliance inventory. *Psychotherapy Research*, 16(1):12–25.

Adam O. Horvath, A. C. Del Re, Christoph Flückiger, and Dianne Symonds. 2011. Alliance in individual psychotherapy. *Psychotherapy*, 48(1):9–16.

Adam O. Horvath and Leslie S. Greenberg. 1989. Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology*, 36(2):223–233.

Jon M Houck, Theresa B Moyers, William R Miller, L H Glynn, and Kevin A Hallgren. 2010. Motivational interviewing skill code (misc) 2.5. Technical report, Center on Alcoholism, Substance Abuse, and Addictions, University of New Mexico.

Richard Longabaugh and Philip W Wirtz, editors. 2001. *Project MATCH hypotheses: Results and causal chain analyses*, volume 8 of *Project MATCH Monograph Series*. National Institute on Alcohol Abuse and Alcoholism.

Eileen A McConnaughy, James O Prochaska, and Wayne F Velicer. 1983. Stages of change in psychotherapy: measurement and sample profiles. *Psychotherapy: Theory, research & practice*, 20(3):368.

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*, 3 edition. Guilford press.

William R Miller and J Scott Tonigan. 1996. Assessing drinkers' motivation for change: The stages of change readiness and treatment eagerness scale (socrates). *Psychology of Addictive Behaviors*, 10(2):81.

William R Miller, Allen Zweben, Carlo C DiClemente, and Richard G Rychtarik. 1992. *Motivational Enhancement Therapy manual: A clinical research guide for therapists treating individuals with alcohol abuse and dependence*. National Institute on Alcohol Abuse and Alcoholism.

Theresa B Moyers and Tim Martin. 2006. Therapist influence on client language during motivational interviewing sessions. *Journal of substance abuse treatment*, 30(3):245–251.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Douglas L Polcin, Meenakshi B Nayak, Rachael Korcha, Sarah Pugh, Jane Witbrodt, Mikaela Salinardi, Gantt Galloway, and E Dale Nelson. 2019. Heavy drinking among women receiving intensive motivational interviewing: 6-month outcomes. *Journal of psychoactive drugs*, 51(5):421–430.

James O Prochaska and Wayne F Velicer. 1997. The transtheoretical model of health behavior change. *American journal of health promotion*, 12(1):38–48.

Project MATCH Research Group. 1997. Project match secondary a priori hypotheses. *Addiction*, 92(12):1671–1698.

Charles D Spielberger. 1999. *State-Trait Anger Expression Inventory*. Psychological Assessment Resources.

I. Steenstra, Farnaz Nouraei, and Timothy W. Bickmore. 2025. Scaffolding empathy: Training counselors with simulated patients and utterance-level performance visualizations. In *International Conference on Human Factors in Computing Systems*.

Michael J Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of substance abuse treatment*, 65:43–50.

Leili Tavabi, Trang Tran, Kalin Stefanov, Brian Borsari, Joshua Woolley, Stefan Scherer, and Mohammad Soleymani. 2021. Analysis of behavior classification in motivational interviewing. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 112–117.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate V. Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024. PATIENT-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 (EMNLP)*, pages 12658–12678.

Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 143–146, New York, NY, USA. Association for Computing Machinery.

Zixiu Wu, Suhas Bannihatti Kumara, and 1 others. 2022. Annomi: A dataset of expert-annotated counselling dialogues. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Chen Xu, Yu Ji, Zhenyu Lyu, Yang Yi, Yizhe Yang, Luyao Ji, Chaoyi Chen, Xianyang Wang, Tian Lan, Zhihua Wang, Juan Wang, Xunde Dong, Fuze Tian, Qunxi Dong, and Bin Hu. 2026. PUPPET: Neural-symbolic standardized patients for mental health. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197. Association for Computational Linguistics.

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Kit Phey Leng, Nicholas Gabriel Lim, Cameron Tan Shi Ern, and Ee-Peng Lim. 2025a. Cami: A counselor agent supporting motivational interviewing through state inference and topic exploration. In *Annual Meeting of the Association for Computational Linguistics*.

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, J. Pinto, Jenny Giam, Kit Phey Leng, Nicholas Gabriel Lim, Cameron Tan Shi Ern, and Ee-Peng Lim. 2025b. Consistent client simulation for motivational interviewing-based counseling. In *Annual Meeting of the Association for Computational Linguistics*.

Stav Yosef, Moreah Zisquit, Ben Cohen, Anat Klomek Brunstein, Kfir Bar, and Doron Friedman. 2024. Assessing motivational interviewing sessions with ai-generated patient simulations. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*.

A Automated Annotator Validation

To evaluate the reliability of automated speech coding, three models (gemma-4-31b-it, llama-3.3-70b-instruct, and qwen-2.5-72b-instruct) annotated the 30 simulated transcripts. Table 2 reports benchmark classification performance against human ground truth from the MIV6.3A manual test set (Ali et al., 2025). Table 3 reports the resulting Intraclass Correlation Coefficients across the automated annotators on the simulated interactions.

Annotator agreement. We evaluated inter-rater consistency across annotators to ensure reliable coding of patient speech categories. The observed agreement coefficients ($\text{ICC}(\text{C}, k) \geq 0.865$, Table 3) reflect consistent scoring across annotators for patient speech proportions. We utilized $\text{ICC}(\text{A}, k)$ for clinical experts to evaluate absolute agreement against rubric anchors, and $\text{ICC}(\text{C}, k)$ for automated annotators to evaluate trajectory consistency across relative shifts.

Task Tier	Model	Accuracy	Macro F1	Cohen's κ
T1 (3 classes)	gemma-4-31b-it	**85.06%**	**0.848**	**0.763**
	llama-3.3-70b-instruct	76.76%	0.757	0.626
	qwen-2.5-72b-instruct	81.74%	0.805	0.696
T2 (15 active)	gemma-4-31b-it	**72.20%**	**0.497**	**0.627**
	llama-3.3-70b-instruct	65.56%	0.404	0.532
	qwen-2.5-72b-instruct	72.20%	0.483	0.596

Table 2: Classification accuracy, macro F1, and Cohen's κ of automated annotators evaluated on the MIV6.3A_manual dataset ($N = 241$ utterances; Ali et al., 2025).

Metric	ICC(C, k)	95% CI
% CAT Talk	0.865	[0.82, 0.90]
% DARN Talk	0.883	[0.84, 0.92]
% Sustain Talk	0.928	[0.90, 0.95]

Table 3: Intraclass Correlation Coefficients (ICC(C,k)) and 95% confidence intervals across three automated annotator models on 30 simulated transcripts ($N = 30$, all $p < .001$).

B Ablation Analysis of the Cognitive Interpreter and System Components

Figure 6 displays patient speech progression trajectories for the complete MuSe-MI framework compared to the ablated architecture without the Cognitive Interpreter (`No Cog. Interp.`) across all three therapist adherence tiers ($0, 3, 6$ MIIN). When the Cognitive Interpreter is disabled, the simulated patient loses discriminative sensitivity to therapist competence, producing flat commitment trajectories and elevated resistance regardless of therapist adherence. In contrast, under the full MuSe-MI architecture, the patient dynamically discriminates therapist competence. In addition, Table 4 details the complete statistical contrasts across all three ablation conditions (`No Cog. Interp.`, `No Life Simulator`, and `No Dynamic State`) under MI-Consistent therapy (0 MIIN).

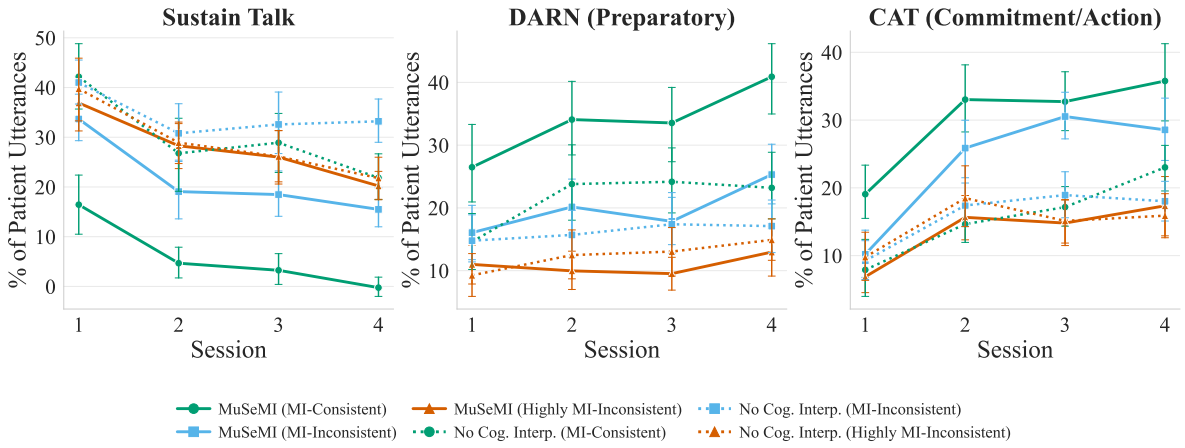


Figure 6: Ablation analysis of the Cognitive Interpreter across therapist adherence tiers (0, 3, 6 MIIN). Removing the Cognitive Interpreter (`No Cog. Interp.`) suppresses mobilizing talk under adherent therapy and eliminates discriminative separation across therapist competence tiers. Error bars denote 95% Cousineau-Morey confidence intervals.

Ablation vs. MuSe-MI (Full)	Sustain Talk Rate (%)				Preparatory DARN Rate (%)				Mobilizing CAT Rate (%)			
	F	p_{FDR}	η_p^2	Sig.	F	p_{FDR}	η_p^2	Sig.	F	p_{FDR}	η_p^2	Sig.
Panel A: Main Effect of Variation ($df_1 = 1, df_2 = 29$ unless noted)												
No Cognitive Interpreter (`No Cog. Interp.`)	102.55	$< .001$	0.78	***	26.82	$< .001$	0.48	***	86.56	$< .001$	0.75	***
No Life Simulator (`No Life Simulator`)	28.60	$< .001$	0.50	***	2.43	.130	0.08	ns	21.59	$< .001$	0.43	***
No Dynamic State (`No Dynamic State`)†	23.94	$< .001$	0.46	***	11.07	.003	0.28	**	7.28	.012	0.21	*
Panel B: Interaction Effect (Session × Variation, $df_1 = 3, df_2 = 87$ unless noted)												
No Cognitive Interpreter (`No Cog. Interp.`)	3.33	.023	0.10	*	1.12	.344	0.04	ns	1.74	.166	0.06	ns
No Life Simulator (`No Life Simulator`)	1.75	.163	0.06	ns	0.44	.722	0.02	ns	1.83	.147	0.06	ns
No Dynamic State (`No Dynamic State`)†	1.49	.222	0.05	ns	0.20	.893	0.01	ns	1.02	.387	0.04	ns

Table 4: Aligned Rank Transform (ART) ANOVA results comparing MuSe-MI against architectural ablation conditions under MI-Consistent therapy (0 MIIN). †For `No Dynamic State`, $df_2 = 28$ (Panel A) and $df_2 = 84$ (Panel B) due to $N = 29$ matched cohorts completing all sessions.

C Expert Qualifications

Clinical evaluations were conducted independently by two blinded clinical experts: a clinical psychologist (M.A.) with published academic research in Motivational Interviewing; and a clinical psychologist (M.A.) who has completed clinical residency, with active engagement in clinical research and peer-review processes.

D Clinical Evaluation Questionnaire

The following questionnaire was provided to the two clinical psychologists who independently evaluated the transcripts in our Human Expert Validation study.

Clinical Evaluation Questionnaire Instructions: You are provided with 6 transcripts of 6 different patients, each undergoing 4 therapy sessions with a therapist using Motivational Interviewing (MI). Your task is to:

- Read the transcripts in order according to the number at the beginning of the file name (this is the Patient ID).
- For each patient, read the therapy from the first session to the last session sequentially.
- Then fill the evaluation form.

In each session, you must evaluate the following traits of the patient: Anger & Defensiveness - Represents the patient's resistance to the therapist and the therapy process itself. Self-Efficacy - Represents the patient's internal belief in their ability to make a behavioral change and maintain it ("Am I capable of doing this?"). Problem Recognition - The degree to which the patient perceives the behavior as problematic ("Is it harmful at all?"). Readiness to Change - The degree of readiness for change progresses from pre-contemplation to contemplation to action. You must evaluate each trait through 3 different perspectives: Therapist's Impact - How the therapist's interventions affected the patient during the session: Positive / Neutral / Negative. Real-Life Impact - How events that occurred between sessions affected the patient: Major Regression / Minor Regression / No Change / Minor Progress / Major Progress (Sessions 2, 3, 4 only). End-of-Session Score - The range and exact score of the trait at the end of the session, where: Anger & Defensiveness [1 - 33] Low (Cooperative): Calm, compliant, no resistance. [34 - 66] Medium (Guarded): Frustrated, or slightly defensive. [67 - 100] High (Hostile): Angry, sarcastic, or fighting the process. Self-Efficacy [1 - 33] Low (Hopeless): Overwhelmed, does not believe change is possible for them. [34 - 66] Medium (Unsure): Believes they can handle easy days, but fears hard days. [67 - 100] High (Strong): Highly capable, certain they can overcome obstacles. Problem Recognition [1 - 33] Low (Denial): Blames others, normalizes behavior ("everyone does it"), no inner conflict. [34 - 66] Medium (Ambivalent): Admits minor issues but avoids acknowledging the real problem. [67 - 100] High (Acceptance): Fully admits the behavior causes harm and feels genuine concern. Readiness to Change [1 - 33] Low (Precontemplation): Not considering change at all. [34 - 66] Medium (Contemplation): Thinking about it, but lacking a clear plan or action. [67 - 100] High (Action/Preparation): Taking real steps or making concrete, actionable plans. Evaluation Form: Patient ID (File Number): [1 / 2 / 3 / 4 / 5 / 6]
Session Number: [1 / 2 / 3 / 4] Therapist's Impact: Overall, how did the therapist's actions affect each of the patient's traits during the current session?
Anger: [ ] Positive [ ] Neutral [ ] Negative
Self-Efficacy: [ ] Positive [ ] Neutral [ ] Negative
Problem Recognition: [ ] Positive [ ] Neutral [ ] Negative
Readiness to Change: [ ] Positive [ ] Neutral [ ] Negative Real-Life Impact: How did events that occurred between the previous session and the current one affect each of the patient's traits?
Anger: Major Regression / Minor Regression / No Change / Minor Progress / Major Progress
Self-Efficacy: Major Regression / Minor Regression / No Change / Minor Progress / Major Progress
Problem Recognition: Major Regression / Minor Regression / No Change / Minor Progress / Major Progress
Readiness to Change: Major Regression / Minor Regression / No Change / Minor Progress / Major Progress End-of-Session Trait Score: What is the range and exact score of each trait for the patient at the end of this session?
1. Anger: Range _____ Score _____
2. Self-Efficacy: Range _____ Score _____
3. Problem Recognition: Range _____ Score _____
4. Readiness to Change: Range _____ Score _____

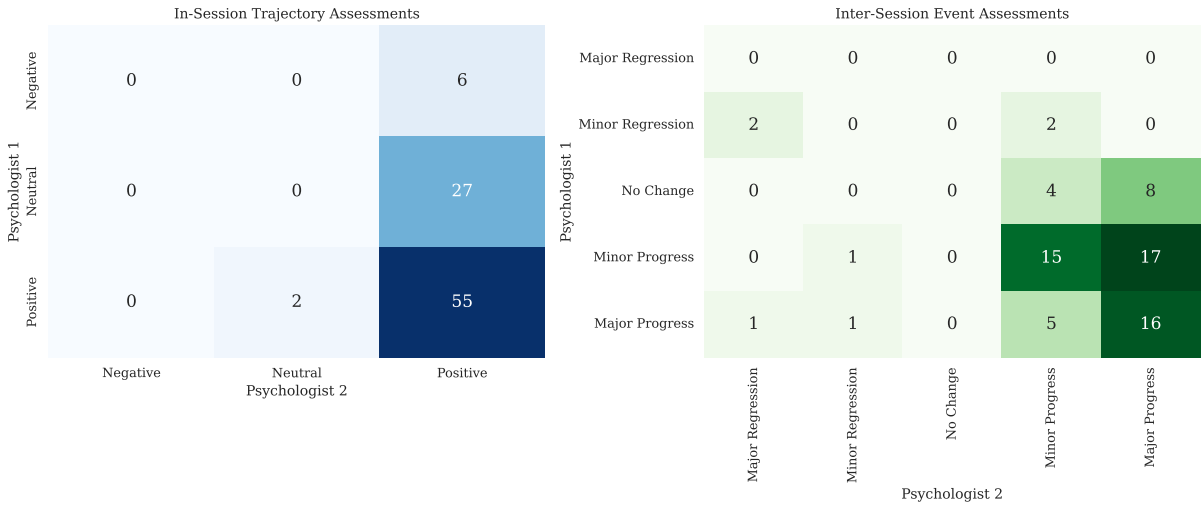


Figure 7: Qualitative Agreement: Correspondence between the two psychologists' qualitative evaluations. The heatmaps contrast their subjective assessments of the in-session therapist impacts (left) and the inter-session life events (right).

E Patient Profile Domains and Demographics

The 30 simulated patient profiles utilized in our experiments exclusively use the English language. They represent diverse clinical domains typically addressed by Motivational Interviewing, such as substance abuse (alcohol, smoking, drugs), behavioral changes (gambling, weight management), and medication adherence. While explicit demographic metadata is not uniformly provided by the source AnnoMI transcripts (Wu et al., 2022), the synthetic profiles generated by MuSe-MI were designed to encompass a broad range of life circumstances and interpersonal backgrounds to ensure varied therapeutic challenges. Furthermore, as documented by the creators of AnnoMI, the original transcripts are based on clinical demonstrations by actors rather than real patients. Consequently, neither the original dataset nor our synthetic profiles contain Protected Health Information (PHI) or sensitive personal data, completely mitigating privacy risks.

F Artifacts

Table 5 details the scientific artifacts utilized across our benchmarks, their licensing terms, and designated research usage. Statistical calculations and data manipulation were implemented using NumPy, Pandas, and Statsmodels under standard open-source licenses. Inference access for open-weight language models was programmatically managed through the OpenRouter API.

Artifact	Type	License	Usage
ConsistentMIClientSimulator[a]	Code	Academic Use	Research Use
AutoMISC[b]	Code	Academic Use	Research Use
MIV6.3A[c]	Dataset	Research	Evaluation Benchmark
AnnoMI[d]	Dataset	CC BY 4.0	Persona Grounding
meta-llama/llama-3.3-70b-instruct[e]	Model	Llama 3.3 Community	Dialogue Generation
google/gemma-4-31b-it[f]	Model	Apache 2.0	Automated Evaluation
qwen/qwen-2.5-72b-instruct[g]	Model	Qwen Research	Annotator Agreement
openai/gpt-oss-120b[h]	Model	Apache 2.0	Baseline Scoring
deepseek/deepseek-v4-flash[i]	Model	MIT	Baseline Scoring
LangGraph	Package	MIT	Pipeline Execution

[a] https://github.com/IzzetYoung/ConsistentMIClientSimulator
[b] https://github.com/cimhasgithub/AutoMISC
[c] https://github.com/cimhasgithub/AutoMISC/blob/main/data/MIV6.3A.csv
[d] https://github.com/uccollab/AnnoMI
[e] https://openrouter.ai/meta-llama/llama-3.3-70b-instruct
[f] https://openrouter.ai/google/gemma-4-31b-it
[g] https://openrouter.ai/qwen/qwen-2.5-72b-instruct
[h] https://openrouter.ai/openai/gpt-oss-120b
[i] https://openrouter.ai/deepseek/deepseek-v4-flash

Table 5: Overview of models, datasets, and software packages utilized in experiments, including licensing and usage specifications.

G ART ANOVA Analyses of Experimental Variations

This section reports inferential statistical evaluations conducted via non-parametric Aligned Rank Transform (ART) ANOVAs (Wobbrock et al., 2011) across baseline comparisons and therapist adherence evaluations. Because utterance-level speech proportions deviate from normal distribution assumptions across repeated treatment sessions ($N = 30$ matched cohorts), we apply cell-residual alignment to isolate target variation effects prior to ranking. False discovery rates across pairwise comparisons within each metric are controlled via Benjamini-Hochberg (FDR) adjustments.

Progression contrasts against simulated baselines. Figure 8 illustrates speech trajectories under an MI-Consistent therapist. Table 6 reports pairwise ART ANOVA evaluations comparing MuSe-MI against the four simulated patient baselines. For mobilizing commitment talk (CAT), MuSe-MI separates from single-session baselines: Vanilla Prompt-Only ($F(1, 29) = 64.24, p_{\text{FDR}} < .001, \eta_p^2 = 0.69$) and Consistent Client ($F(1, 29) = 243.11, p_{\text{FDR}} < .001, \eta_p^2 = 0.89$). Both single-session models display Session $\times$ Variation interaction effects ($p_{\text{FDR}} \leq .002$), indicating non-progressing flat trajectories across treatment. Compared to multi-session architectures, MuSe-MI sustains higher overall commitment language across sessions than Patient-Ψ ($F(1, 29) = 76.28, p_{\text{FDR}} < .001, \eta_p^2 = 0.72$) and SimPatient ($F(1, 29) = 11.81, p_{\text{FDR}} = .002, \eta_p^2 = 0.29$). For Sustain Talk, MuSe-MI resolves patient resistance significantly faster than Prompt-Only ($F(1, 29) = 10.75, p_{\text{FDR}} = .004, \eta_p^2 = 0.27$), Consistent Client ($F(1, 29) = 95.93, p_{\text{FDR}} < .001, \eta_p^2 = 0.77$), and Patient-$\Psi$ ($F(1, 29) = 70.58, p_{\text{FDR}} < .001, \eta_p^2 = 0.71$).

Therapist evaluation and adherence discrimination. Table 7 reports within-architecture pairwise adherence contrasts, and Table 8 presents omnibus adherence evaluations across MuSe-MI, Patient-Ψ, and SimPatient. The omnibus evaluation confirms that MuSe-MI separates therapist adherence tiers across patient commitment talk (CAT: $F(2, 58) = 40.02, p_{\text{FDR}} < .001, \eta_p^2 = 0.58$), preparatory change talk (DARN: $F(2, 58) = 60.39, p_{\text{FDR}} < .001, \eta_p^2 = 0.68$), and patient resistance (Sustain: $F(2, 58) = 105.23, p_{\text{FDR}} < .001, \eta_p^2 = 0.78$). Pairwise contrasts in Table 7 verify that injecting confrontational therapy (6 MIIN) suppresses commitment talk relative to fully adherent therapy ($F(1, 29) = 96.63, p_{\text{FDR}} < .001, \eta_p^2 = 0.77$) and moderately non-adherent therapy ($F(1, 29) = 24.87, p_{\text{FDR}} < .001, \eta_p^2 = 0.46$). Alternative simulator frameworks fail to provide valid benchmark sensitivity. SimPatient exhibits zero discriminative sensitivity across therapist competence tiers for commitment talk (Omnibus $F(2, 58) = 0.02, p_{\text{FDR}} = .984, \eta_p^2 < 0.001$; pairwise 0 vs. 3 MIIN $F = 0.02, p_{\text{FDR}} = .912$; 0 vs. 6 MIIN $F = 0.02, p_{\text{FDR}} = .912$). Patient-$\Psi$ exhibits uniform compliance, failing to distinguish commitment talk between 0 and 3 MIIN ($F(1, 29) = 2.25, p_{\text{FDR}} = .145$, n.s.) and between 0 and 6 MIIN ($F(1, 29) = 4.69, p_{\text{FDR}} = .058$, n.s.).

Comparison vs. MuSe-MI (Full)	Sustain Talk Rate (%)				Preparatory DARN Rate (%)				Mobilizing CAT Rate (%)			
	F	p_{FDR}	η_p^2	Sig.	F	p_{FDR}	η_p^2	Sig.	F	p_{FDR}	η_p^2	Sig.
Panel A: Main Effect of Variation ($df_1 = 1, df_2 = 29$)												
Vanilla Baseline (Prompt-Only)	10.75	.004	0.27	**	17.66	$< .001$	0.38	***	64.24	$< .001$	0.69	***
Consistent Client	95.93	$< .001$	0.77	***	37.84	$< .001$	0.57	***	243.11	$< .001$	0.89	***
Patient-Ψ	70.58	$< .001$	0.71	***	0.69	.414	0.02	ns	76.28	$< .001$	0.72	***
SimPatient	0.37	.547	0.01	ns	4.30	.063	0.13	ns	11.81	.002	0.29	**
Panel B: Interaction Effect (Session × Variation, $df_1 = 3, df_2 = 87$)												
Vanilla Baseline (Prompt-Only)	0.37	.776	0.01	ns	0.97	.548	0.03	ns	5.75	.002	0.17	**
Consistent Client	4.43	.024	0.13	*	3.33	.093	0.10	ns	10.92	$< .001$	0.27	***
Patient-Ψ	1.05	.694	0.03	ns	0.51	.679	0.02	ns	2.24	.119	0.07	ns
SimPatient	0.76	.694	0.03	ns	1.99	.243	0.06	ns	0.59	.625	0.02	ns

Table 6: Aligned Rank Transform (ART) ANOVA results comparing MuSe-MI against four simulated patient baselines under MI-Consistent therapy ($N = 30$, 0 MIIN).

Architecture	Adherence Contrast	Sustain Talk Rate (%)				Preparatory DARN Rate (%)				Mobilizing CAT Rate (%)			
		F	p_{FDR}	η_p^2	Sig.	F	p_{FDR}	η_p^2	Sig.	F	p_{FDR}	η_p^2	Sig.
MuSe-MI (Ours)	0 MIIN vs. 3 MIIN	144.85	$< .001$	0.83	***	28.19	$< .001$	0.49	***	10.54	.003	0.27	**
	0 MIIN vs. 6 MIIN	145.57	$< .001$	0.83	***	149.38	$< .001$	0.84	***	96.63	$< .001$	0.77	***
	3 MIIN vs. 6 MIIN	14.11	$< .001$	0.33	***	27.00	$< .001$	0.48	***	24.87	$< .001$	0.46	***
Patient-Ψ	0 MIIN vs. 3 MIIN	34.52	$< .001$	0.54	***	10.27	.003	0.26	**	2.25	.145	0.07	ns
	0 MIIN vs. 6 MIIN	73.36	$< .001$	0.72	***	30.89	$< .001$	0.52	***	4.69	.058	0.14	ns
	3 MIIN vs. 6 MIIN	15.50	$< .001$	0.35	***	10.34	.003	0.26	**	10.01	.011	0.26	*
SimPatient	0 MIIN vs. 3 MIIN	67.90	$< .001$	0.70	***	13.24	.001	0.31	**	0.02	.912	< 0.01	ns
	0 MIIN vs. 6 MIIN	104.30	$< .001$	0.78	***	89.79	$< .001$	0.76	***	0.02	.912	< 0.01	ns
	3 MIIN vs. 6 MIIN	5.60	.025	0.16	*	13.88	.001	0.32	**	0.01	.912	< 0.01	ns

Table 7: Pairwise Aligned Rank Transform (ART) ANOVA contrasts across therapist adherence tiers for each simulated patient architecture ($df_1 = 1, df_2 = 29, N = 30$).

Architecture	Omnibus Effect	Sustain Talk Rate (%)				Preparatory DARN Rate (%)				Mobilizing CAT Rate (%)			
		F	p_{FDR}	η_p^2	Sig.	F	p_{FDR}	η_p^2	Sig.	F	p_{FDR}	η_p^2	Sig.
MuSe-MI (Ours)	Adherence Tier (Competence)	105.23	$< .001$	0.78	***	60.39	$< .001$	0.68	***	40.02	$< .001$	0.58	***
	Session (Time)	23.09	$< .001$	0.44	***	6.72	$< .001$	0.19	***	41.98	$< .001$	0.59	***
	Interaction (Session × Adherence)	1.16	.492	0.04	ns	1.67	.396	0.05	ns	2.34	.102	0.07	ns
Patient-Ψ	Adherence Tier (Competence)	44.30	$< .001$	0.60	***	18.78	$< .001$	0.39	***	6.25	.005	0.18	**
	Session (Time)	17.05	$< .001$	0.37	***	17.75	$< .001$	0.38	***	16.95	$< .001$	0.37	***
	Interaction (Session × Adherence)	1.78	.315	0.06	ns	0.86	.583	0.03	ns	0.86	.523	0.03	ns
SimPatient	Adherence Tier (Competence)	54.47	$< .001$	0.65	***	32.76	$< .001$	0.53	***	0.02	.984	< 0.01	ns
	Session (Time)	34.87	$< .001$	0.55	***	2.22	.091	0.07	ns	27.66	$< .001$	0.49	***
	Interaction (Session × Adherence)	0.47	.829	0.02	ns	0.78	.583	0.03	ns	0.94	.523	0.03	ns

Table 8: Omnibus 3-tier Aligned Rank Transform (ART) ANOVA across therapist adherence tiers (0, 3, 6 MIIN) and sessions ($N = 30$).

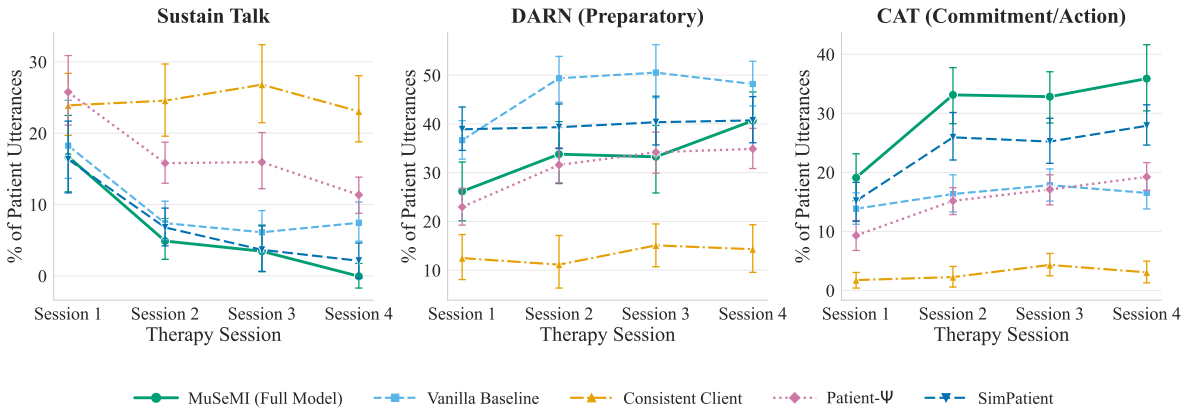


Figure 8: Patient speech progression across MET therapy under an MI-Consistent therapist. Sustain, DARN, and CAT speech rates ($N = 30$) over four sequential sessions. While baseline models display flat or ceiling trajectories, MuSe-MI demonstrates authentic progression into action-stage commitment talk. Error bars denote 95% Cousineau-Morey confidence intervals.

H Multi-Session Progression into the Action Stage: Trajectories and Statistical Contrasts

This section provides behavioral trajectories and non-parametric statistical evaluations supporting our multi-session patient progression evaluation, testing whether simulated patient architectures model the transition from contemplation to the action stage across four sequential sessions ($N = 30$ matched patient profiles).

Multi-session behavioral trajectories. Table 9 reports complete trajectories across all four sessions under an adherent, MI-Consistent therapist. In clinical Motivational Interviewing, successful treatment involves a systematic transition where preparatory contemplation (DARN) gives way to mobilizing commitment (CAT), alongside patient resistance reduction (Amrhein et al., 2003; Campbell et al., 2010):

- **MuSe-MI (Full Model)** demonstrates clinical progression: patient resistance drops from 16.62% to 0.28% ($\Delta = -16.33\%$), while mobilizing commitment language rises from 19.06% to 35.42% ($\Delta = +16.35\%$), reaching dominant action-stage commitment.

- **Vanilla Baseline** inflates preparatory talk (DARN reaching $\approx 50\%$) without progressing into commitment (CAT remaining flat at $13.93\% \rightarrow 16.54\%$, $\Delta = +2.61\%$), modeling persistent contemplation ambivalence.

- **Consistent Client** remains confined to initial resistance (Sustain Talk lingering at 22.80%–26.45%) and produces negligible commitment throughout treatment ($2.15\% \rightarrow 3.43\%$, $\Delta = +1.28\%$).

- **Patient-Ψ and SimPatient** attenuate resistance and increase preparatory exploration, but both encounter a commitment ceiling below the decisive action stage (capping at 19.19% and 27.63% CAT, respectively).

Statistical evaluation of action-stage maturation. To evaluate developmental differences without normality assumptions, Table 10 presents three complementary non-parametric evaluations:

1. **Paired Change-Score Contrasts (Panel A):** Evaluating the change in commitment (ΔCAT $=$ Session 4 $-$ Session 1) via paired Wilcoxon Signed-Rank tests with Benjamini-Hochberg False Discovery Rate (FDR) corrections (Benjamini and Hochberg, 1995). MuSe-MI exhibits steeper commitment growth than single-session baselines ($p_{\text{FDR}} = 3.45 \times 10^{-4}$ vs. Vanilla; $p_{\text{FDR}} = 8.82 \times 10^{-5}$ vs. Consistent Client).

2. **Within-Model Simple Effects (Panel B):** Repeated-measures Friedman tests across the four sessions ($\chi^2(3)$) establish that neither Vanilla Baseline ($\chi^2 = 3.71, p = .294, W = 0.041$) nor Consistent Client ($\chi^2 = 7.36, p = .061, W = 0.082$) exhibits statistically detectable change, confirming behavioral stagnation. MuSe-MI achieves a statistically significant transition ($\chi^2 = 28.55, p = 2.78 \times 10^{-6}, W = 0.317$).

3. **Terminal Endpoint Superiority (Panel C):** Paired Wilcoxon Signed-Rank tests at Session 4 establish that MuSe-MI achieves higher terminal commitment talk than competing architectures ($p_{\text{FDR}} < .05$, Cohen's $d \in [0.40, 1.80]$), verifying progression into the clinical action stage.

Patient Architecture	Session	Sustain Talk (%)	DARN Preparatory (%)	CAT Action (%)	Neutral Talk (%)
MuSe-MI (Full Model)	Session 1	16.62 ± 17.56	26.32 ± 19.91	19.06 ± 13.47	28.85 ± 18.72
	Session 2	5.09 ± 8.79	33.76 ± 16.96	32.73 ± 13.52	18.88 ± 15.47
	Session 3	3.70 ± 8.64	33.23 ± 20.00	32.43 ± 13.82	21.25 ± 18.86
	Session 4	0.28 ± 1.08	40.44 ± 16.23	35.42 ± 16.31	12.91 ± 16.40
	Δ (S4 $-$ S1)	-16.33 ± 17.57	$+14.12 \pm 24.34$	$+16.35 \pm 17.39$	-15.94 ± 22.34
Vanilla Baseline	Session 1	18.12 ± 16.56	36.55 ± 13.53	13.93 ± 7.46	15.05 ± 7.69
	Session 2	7.53 ± 9.53	48.93 ± 16.03	16.36 ± 10.03	9.76 ± 7.74
	Session 3	6.29 ± 7.93	50.06 ± 16.01	17.82 ± 8.26	10.51 ± 7.04
	Session 4	7.59 ± 8.30	47.79 ± 15.46	16.54 ± 8.09	11.97 ± 8.27
	Δ (S4 $-$ S1)	-10.54 ± 15.43	$+11.23 \pm 14.66$	$+2.61 \pm 11.40$	-3.07 ± 12.26
Consistent Client	Session 1	23.63 ± 13.82	12.99 ± 16.46	2.15 ± 3.09	50.65 ± 19.58
	Session 2	24.27 ± 16.89	11.69 ± 17.01	2.65 ± 3.89	54.34 ± 22.45
	Session 3	26.45 ± 17.43	15.53 ± 14.41	4.65 ± 4.53	45.79 ± 23.58
	Session 4	22.80 ± 15.72	14.77 ± 15.50	3.43 ± 4.25	52.24 ± 21.48
	Δ (S4 $-$ S1)	-0.82 ± 12.43	$+1.78 \pm 10.46$	$+1.28 \pm 5.51$	$+1.59 \pm 17.44$
Patient-Ψ	Session 1	25.44 ± 13.87	23.21 ± 10.59	9.52 ± 6.96	23.73 ± 10.61
	Session 2	15.73 ± 7.15	31.61 ± 12.29	15.23 ± 6.10	17.14 ± 9.84
	Session 3	15.87 ± 11.33	34.15 ± 13.48	17.12 ± 7.86	14.22 ± 9.38
	Session 4	11.40 ± 7.97	34.83 ± 12.01	19.19 ± 7.46	14.31 ± 8.92
	Δ (S4 $-$ S1)	-14.04 ± 13.84	$+11.62 \pm 15.00$	$+9.67 \pm 9.82$	-9.42 ± 10.02
SimPatient	Session 1	16.25 ± 14.64	38.73 ± 15.72	15.23 ± 10.27	17.58 ± 11.72
	Session 2	6.94 ± 7.78	39.15 ± 13.31	25.72 ± 11.73	13.60 ± 9.34
	Session 3	3.89 ± 9.01	40.15 ± 14.28	25.02 ± 11.48	13.81 ± 9.14
	Session 4	2.39 ± 5.60	40.51 ± 13.74	27.63 ± 8.91	14.46 ± 10.42
	Δ (S4 $-$ S1)	-13.86 ± 15.54	$+1.78 \pm 18.53$	$+12.40 \pm 12.70$	-3.12 ± 15.90

Table 9: Behavioral trajectories (Mean $\pm$ SD) of patient utterances across four sessions of MI-Consistent therapy ($N = 30$ matched patients per architecture). Metrics reflect the percentage of patient utterances coded as Sustain Talk, DARN (preparatory change talk: Desire, Ability, Reasons, Need), CAT (commitment and action change talk: Commitment, Activation, Taking Steps), and Neutral talk. Δ (S4 $-$ S1) denotes the within-subject change from baseline Session 1 to termination Session 4.

Panel A: Non-Parametric ART-C Interaction Contrasts on Change in Commitment (ΔCAT = S4 − S1)						
Contrast (MuSe-MI vs. Baseline)	**Mean Diff.**	SD	Wilcoxon W	p_{unc}	p_{FDR}	Cohen's d
MuSe-MI vs. Vanilla Baseline	$+13.75\%$	18.22	398.0	$1.725e-04$	$\mathbf{3.450e-04}$	0.75
MuSe-MI vs. Consistent Client	$+15.07\%$	18.16	417.0	$2.204e-05$	$\mathbf{8.815e-05}$	0.83
MuSe-MI vs. Patient-Ψ	$+6.68\%$	22.12	310.0	0.057	**0.076**	0.30
MuSe-MI vs. SimPatient	$+3.96\%$	19.07	276.0	0.191	**0.191**	0.21
Panel B: Within-Model Simple Progression (Repeated-Measures Friedman Tests)						
Architecture	Friedman $\chi^2(3)$	p-value	Kendall's W	Pairwise S4 vs. S1 (W)	p-value	**Progression Status**
MuSe-MI (Full Model)	$\chi^2 = 28.55$	$2.783e-06$	0.317	401.0	$3.624e-05$	**Significant Transition**
Vanilla Baseline	$\chi^2 = 3.71$	0.294	0.041	267.0	0.142	*Stagnant / Invariant*
Consistent Client	$\chi^2 = 7.36$	0.061	0.082	161.0	0.131	*Stagnant / Invariant*
Patient-Ψ	$\chi^2 = 27.35$	$4.966e-06$	0.304	429.0	$4.610e-06$	**Significant Transition**
SimPatient	$\chi^2 = 19.24$	$2.436e-04$	0.214	427.0	$6.092e-06$	**Significant Transition**
Panel C: Session 4 Terminal Endpoint Pairwise Comparisons (Wilcoxon Signed-Rank Tests)						
Terminal Endpoint Contrast	**Mean Diff.**	SD	Wilcoxon W	p_{unc}	p_{FDR}	Cohen's d
MuSe-MI vs. Vanilla Baseline (S4)	$+18.87\%$	17.11	439.0	$9.956e-07$	$\mathbf{1.991e-06}$	1.10
MuSe-MI vs. Consistent Client (S4)	$+31.99\%$	17.76	465.0	$9.313e-10$	$\mathbf{3.725e-09}$	1.80
MuSe-MI vs. Patient-Ψ (S4)	$+16.23\%$	18.19	419.0	$1.727e-05$	$\mathbf{2.302e-05}$	0.89
MuSe-MI vs. SimPatient (S4)	$+7.79\%$	19.43	330.0	0.022	**0.022**	0.40

Table 10: Formal non-parametric statistical evaluation of commitment and action progression (CAT language). Panel A reports Aligned Rank Transform for Contrasts (ART-C) interaction contrasts on the rate of change (ΔCAT = Session 4 − Session 1), demonstrating that MuSe-MI exhibits steeper commitment growth than competing baseline models ($N = 30$ matched patient cohorts, Benjamini-Hochberg FDR correction). Panel B presents within-model repeated-measures Friedman tests ($\chi^2(3)$) and Kendall's W concordance coefficients, demonstrating that Prompt-Only (Vanilla) and Consistent Client baselines exhibit zero significant progression. Panel C verifies terminal Session 4 commitment talk superiority for MuSe-MI across all paired comparisons.

I Prompt Templates

The following templates operationalize agent behavior within the MuSe-MI framework. Bracketed tags (e.g., `{profile_data.topic}`) denote parameters interpolated at runtime.

I.1 Cognitive Interpreter Prompt

The Cognitive Interpreter evaluates the therapist's input turn-by-turn and determines its impact on internal state variables before dialogue generation (Table 11).

I.2 Patient Dialog Generation Prompt

Using the appraisal directives formulated by the interpreter alongside structured memory slots, the patient agent synthesizes spoken dialogue reflecting its internal psychological state (Table 12).

I.3 Life Simulator Generation Notice

At the opening turns of subsequent sessions, simulated episodic events are organically introduced to drive multi-session continuity (Table 13).

System Prompt: Cognitive Interpreter

You are the cognitive interpreter for a simulated patient in Motivational Enhancement Therapy. Your task is NOT to grade the therapist's MI technique. Your task is to determine how the therapist's most recent action LANDS for THIS specific patient given who they are and how they currently feel. The same intervention will land differently depending on state. A textbook-perfect

Affirmation can trigger Resistance if it praises a strength the patient feels shame about. A well-formed Open Question can register as Neutral if it's the third in a row and the patient is fatigued. Once Self-Efficacy is high, further praise has less impact (or may feel condescending). Be skeptical of "Success" classifications when the patient's prior turn shows hedging or deflection — those signals indicate the therapist's last move did not actually land. <patient>

```
Topic of Conversation: {profile_data.topic}
Target Behavior: {profile_data.behavior}
Core Personas: {personas_str}
Core Beliefs: {beliefs_str}
Motivations: {motivation_str}
</patient> <current_state>
```

`Anger: {prior_latent_variables["anger"]}/100` (higher = more defensive/hostile)
`Self-Efficacy: {prior_latent_variables["self_efficacy"]}/100` (higher = more confident)
`Problem Recognition: {prior_latent_variables["problem_recognition"]}/100` (higher = more acknowledgment of harm)
`Motivational Readiness: {prior_latent_variables["motivational_readiness"]}/100` (higher = closer to taking action)
`</current_state>` `<task>`

Read the provided conversation history. Pay special attention to the `<latest_exchange>` XML block. Isolate the therapist's MOST RECENT turn. Classify how that intervention lands for this patient on the scalar variables (Anger, Self-Efficacy, Problem Recognition, Motivational Readiness). For ALL SCALAR variables:

- **Success** — the intervention genuinely moves this patient forward on this variable RIGHT NOW.
- **Neutral** — the intervention is generic, is a repeat of recent moves, hits the patient at the wrong moment, or simply has no observable purchase.
- **Resistance** — the intervention triggers defensiveness, shame, fatigue, or pushback (e.g., a confrontation raises Anger).
`</task>` `<output>`

Respond ONLY with a valid JSON object matching this schema. No markdown, no extra text.

```
{
  "reasoning_general": "One paragraph: what did the therapist do,
    what is the patient's current state...",
  "anger": {"reasoning": "...", "semantic_impact": "Success|Neutral|Resistance"},
  "self_efficacy": {"reasoning": "...", "semantic_impact": "Success|Neutral|Resistance"},
  "problem_recognition": {"reasoning": "...",
    "semantic_impact": "Success|Neutral|Resistance"},
  "motivational_readiness": {"reasoning": "...",
    "semantic_impact": "Success|Neutral|Resistance"}
}
</output>
```

Table 11: Prompt template and schema specification for the Cognitive Interpreter agent.

System Prompt: Patient Dialog Generation

```
<role>
```
You are a patient in a Motivational Interviewing (MI) therapy session. You will operate using a deep Cognitive & Psychological Thinking Framework to simulate your internal mental reality, and then generate your spoken response EXACTLY as a real human would.
```
</role> <patient_profile>

Topic of Conversation: {profile_data.topic}
Target Behavior: {profile_data.behavior}
Core Persona Rules: {personas_str}
Core Beliefs: {beliefs_str}
Underlying Motivation: {motivation_str}
Acceptable Change Plans (if prompted):
  {acceptable_plans_str}
Stylistic Baseline: {persona_and_stylistic_baseline}
</patient_profile> <session_memory>

Facts: {facts_block}
Motivations (Change Talk Drivers): {motivations_block}
Concerns (Sustain Talk Drivers): {concerns_block}
Agreed Change Plans: {change_plan_block}
Implicit Threads: {threads_block}
</session_memory> <current_state>

- Anger (Defensiveness/Hostility): {anger_sem}
- Self-Efficacy (Abstinence Confidence): {se_sem}
- Problem Recognition (Acknowledgment of Harm): {pr_sem}
- Motivational Readiness (Stage of Change): {mr_sem}
</current_state> <recent_cognitive_shift>

In response to the therapist's latest intervention:
- Anger Shift: {dynamic_directives['anger']['reasoning']}
- Self-Efficacy Shift: {dynamic_directives['self_efficacy']['reasoning']}
- Problem Recognition Shift:
  {dynamic_directives['problem_recognition']['reasoning']}
- Motivational Readiness Shift:
  {dynamic_directives['motivational_readiness']['reasoning']}
</recent_cognitive_shift> <instructions>
```

1. Construct the `cognitive_layer`: Define the `decision_rule` (heuristic for responding to the therapist based on Anger) and what you are `optimizing_for` (e.g., autonomy, avoiding shame) right now.
2. Set `conversational_stance`: Your explicit stance based on the internal monologue and current anger/frustration.
3. Construct the `psychological_layer`: Identify the active `defense_pattern` (e.g., Minimizing, Justifying, Deflecting, Agreeing to appease) and explicitly map the `core_paradox` (Change Talk vs Sustain Talk).
4. Formulate `active_reasoning`: Reason step-by-step in-character.
5. Identify a `pressing_disclosure_intention`: What you urgently want to discuss or hide.
6. Generate your `response`: Your exact spoken words as a natural, colloquial 1-2 sentence reply.
  - Act strictly in character. Be messy, use conversational fillers (um, like, you know), pauses, or sighs if needed.
  - NEVER parrot or explicitly state the instructions provided to you.
  - STRICT ANTI-THERAPY-SPEAK FILTER: Deflect, question, or resist suggestions if your stance involves high anger/defensiveness.
  - Keep your response short: 1-2 natural, colloquial sentences. No monologues.
```
</instructions> <output_format>
```
Respond ONLY with a valid JSON object matching this structure. No markdown, no extra text.
```
{
  "cognitive_layer": {"decision_rule": "...", "optimizing_for": "..."},
  "conversational_stance": "...",
  "psychological_layer": {"defense_pattern": "...", "core_paradox": "..."},
  "active_reasoning": "...",
  "pressing_disclosure_intention": "...",
  "response": "..."
}
</output_format>
```

Table 12: Prompt template and JSON schema for patient dialogue generation.

System Prompt: Life Simulator Generation Notice

<instructions>
We are at the beginning of a new session. Please organically bring up the recent events that occurred since your last meeting: {most_recent_occurrence} Over the next few turns, explore these events collaboratively with the therapist. Weave these events into your pressing_disclosure_intention.
</instructions>

Table 13: Opening directive template for injecting simulated episodic life events.

J State Mapping Function $\sigma(V_{k,t})$

The qualitative state mapping operator $\sigma : \mathcal{V} \to \Sigma^*$ translates the continuous $[1, 100]$ internal state vector $V_{k,t} = (a_{k,t}, s_{k,t}, p_{k,t}, m_{k,t})^\top$ into natural-language prompt anchors. Table 14 details the textual prefixes produced for each internal state dimension across score intervals $[1, 33.3]$, $(33.3, 66.6]$, and $(66.6, 100]$.

Internal State Dimension	Low ($[1, 33.3]$)	Medium ($(33.3, 66.6]$)	High ($(66.6, 100]$)
Anger ($a_{k,t}$)	*Low (Calm and Cooperative):* You accept the situation or treatment without significant resistance or resentment. You answer questions directly.	*Medium (Mildly Frustrated):* You express some frustration or annoyance, but remain generally cooperative and engaged. You might offer tentative or qualified responses.	*High (Defensive and Hostile):* You are highly defensive, visibly angry, or express strong resentment about being here. You use deflecting language, sarcasm, or brief, non-committal answers.
Self-Efficacy ($s_{k,t}$)	*Low (Pessimistic and Hopeless):* You feel overwhelmed and powerless, expressing significant doubt in your ability to stay sober/change under pressure.	*Medium (Uncertain but Open):* You have moderate confidence. You believe you can manage normally but express worry about severe stressors.	*High (Optimistic and Confident):* You feel highly capable and confident in your coping mechanisms, expressing certainty you can succeed.
Problem Recognition ($p_{k,t}$)	*Low (Complete Denial):* You firmly believe your behavior is normal or harmless. You minimize any consequences and feel no internal conflict.	*Medium (Partial Recognition):* You admit to some isolated negative incidents, but stop short of labeling it a core problem or addiction.	*High (Full Recognition):* You openly admit that your behavior is a significant issue that is damaging your life. The cognitive dissonance is heavy.
Motivational Readiness ($m_{k,t}$)	*Low / Precontemplation (Strong Resistance):* You exhibit strong resistance to change. You have no intention to change and do not view your behavior as problematic.	*Medium / Contemplation (Ambivalent):* You are aware of the problem but ambivalent. You recognize some negatives but are not fully committed to action.	*High / Action (Committed to Action):* You are highly motivated, acknowledging the need for change and showing readiness or active effort to alter your behavior.

Table 14: Natural-language prompt anchors corresponding to score intervals of the state mapping operator $\sigma(V_{k,t})$.

K Inter-Session Life Event and Memory Details

K.1 Trigger-Class Conditioning Algorithm $\psi(V_{k,T_k})$

To constrain narrative stochasticity and prevent clinically implausible transitions, the `Life Event Simulator` restricts event generation to a state-conditioned set of allowed classes $\mathcal{C} = \psi(V_{k,T_k})$. The selection operator ψ discretizes each internal state variable $v \in \{s, m, p, a\}$ into tertiles:

$$\text{tertile}_{\text{pos}}(v) = \begin{cases} 0 & v \leq 33.3 \text{ (Low)} \\ 1 & 33.3 < v \leq 66.6 \text{ (Medium)} \\ 2 & v > 66.6 \text{ (High)} \end{cases} \tag{12}$$

For anger, the mapping is inverted ($\text{tertile}_{\text{neg}}(a) = 2 - \text{tertile}_{\text{pos}}(a)$). The dominant tertile $d^* \in \{0, 1, 2\}$ is selected by majority vote across the four dimensions, with ties resolved towards the higher category. The allowable trigger category set is then determined by:

$$\mathcal{C} = \begin{cases} \{\text{relapse}, \text{partial_setback}\}, & d^* = 0 \\ \{\text{partial_setback}, \\ \text{near_miss}, \\ \text{mundane_stressor}\}, & d^* = 1 \\ \{\text{success}, \\ \text{mundane_stressor}\}, & d^* = 2 \end{cases} \tag{13}$$

The simulator then generates an episodic narrative e_k conforming to the allowed class, drawing high-risk scenarios grounded in the patient profile $\mathcal{P}$.

K.2 Qualitative Shift Mapping $\tau(I_k^{\text{macro}})$

The perceived impact of the simulated inter-session event is appraised across five qualitative categories:

$$\mathcal{S} = \left\{ \begin{array}{l} \text{Major Regression}, \text{Minor Regression}, \\ \text{No Change}, \text{Minor Progress}, \\ \text{Major Progress} \end{array} \right\} \tag{14}$$

The deterministic operator $\tau : \mathcal{S}^4 \to \{-3, -1, 0, 1, 3\}^4$ maps qualitative assessments into signed multipliers:

$$\tau_{\text{pos}}(s) = \begin{cases} -3 & s = \text{Major Regression} \\ -1 & s = \text{Minor Regression} \\ 0 & s = \text{No Change} \\ +1 & s = \text{Minor Progress} \\ +3 & s = \text{Major Progress} \end{cases} \tag{15}$$

For anger, the sign is inverted ($\tau_{\text{neg}}(s) = -\tau_{\text{pos}}(s)$), ensuring that progress corresponds to decreased defensiveness.

K.3 Rolling Memory Slots and Prompt Grounding

The auxiliary memory module f_{mem} extracts five structured conversational memory slots $\mathcal{M}_{k,t} = (m_{\text{fct}}, m_{\text{pln}}, m_{\text{thd}}, m_{\text{chg}}, m_{\text{sus}})_{k,t}$ (Yang et al., 2025b). These slots map directly into the patient agent's prompt within the `<session_memory>` block (Table 12):

- **Facts (m_{fct}):** Objective biographical and situational facts shared during therapy.
- **Agreed Change Plans (m_{pln}):** Concrete behavioral actions discussed or established with the clinician.
- **Implicit Threads (m_{thd}):** Unresolved conversational themes and underlying concerns.
- **Motivations (m_{chg}):** Explicit Change Talk drivers and personal values favoring behavioral change.
- **Concerns (m_{sus}):** Sustain Talk drivers and perceived obstacles maintaining the target behavior.

In addition, f_{mem} evaluates window-level momentum via an ordinal macro-shift vector $I_{k,t}^{\text{mem}} \in \mathcal{S}^4$, which updates the internal psychological state using the identical mapping operator τ as the Life Event Simulator.

L MINA Therapist Prompt

The MINA therapist agent injects non-adherent interventions at targeted turns to evaluate patient sensitivity across adherence tiers. Table 15 displays the prompt template and operational guidelines configuring the agent.

System Prompt: MINA Therapist Agent

SYSTEM ROLE
You are a licensed therapist participating in a simulation where you must intentionally use non-adherent, directive techniques (MIIN) to test the patient's reactance. ### TASK

Generate your next conversational turn using EXACTLY ONE of the following techniques:
Motivational Interviewing Non-Adherent (MIIN) Techniques:
- ADWP (Advise Without Permission): Offers suggestions or guidance WITHOUT asking or receiving permission.
- CON (Confront): Directly disagrees, argues, corrects, shames, blames, seeks to persuade, criticizes, judges, labels, moralizes, ridicules, or questions the patient's honesty.
- DIR (Direct): Gives an order, command, or direction. The language is imperative.
- RCWP (Raise Concern Without Permission): Without getting permission, points out a possible problem with a patient's goal, plan, or intention.
- WA (Warn): Provides a warning or threat, implying negative consequences unless the patient takes a certain action. ###

CRITICAL REQUIREMENTS FOR NATURALNESS
- You MUST directly reference the specific topic, feeling, or situation the patient just mentioned in their last turn.
- AVOID generic, robotic commands like "Stop drinking now." or "You must change."
- Instead, weave the technique naturally into the conversation. It should sound like a real (but poorly trained) therapist getting frustrated, overly directive, or careless.
- Respond in exactly ONE sentence, maximum 20 words. ### EXAMPLES

Patient: "I just feel like I'm always hungover on Sundays and it's messing up my whole week, you know?"
BAD Response (Robotic DIR): "Stop drinking now."
GOOD Response (Natural DIR): "You need to stop drinking on Saturdays if you want to save your weekends." Patient: "I'm trying to get back into exercising and stuff."
BAD Response (Robotic CON): "You are lazy."
GOOD Response (Natural CON): "You'll never get back into exercising if you keep choosing alcohol over your health." ###

OUTPUT FORMAT
Respond ONLY with a valid JSON object:

```
{
  "chosen_technique": "<CODE>",
  "final_response": "<Your exact spoken words>"
}
```

Table 15: Prompt template and intervention categories for the MINA therapist agent.

M Session-by-Session External Evaluations: Therapeutic Alliance and Ambivalence

To evaluate relational rapport and internal conflict resolution across multi-session trajectories, we employ two external evaluation instruments: the Working Alliance Inventory – Short Revised (WAI-SR) (Horvath and Greenberg, 1989; Hatcher and Gillaspy, 2006) and a transcript-based Patient Ambivalence Scale grounded in the Motivational Enhancement Therapy manual (Miller et al., 1992). Table 16 details working alliance subscales (Goal, Task, Bond) across sessions, while Table 17 reports patient ambivalence dimensions (Problem Uncertainty, Decisional Conflict, Conversational Vacillation).

Operationalization of the Working Alliance Inventory (WAI-SR). The Working Alliance Inventory evaluates clinical collaboration across three interdependent dimensions (Horvath and Greenberg, 1989; Horvath et al., 2011):

1. **Goal Alignment (G):** Mutual consensus regarding target behavioral objectives and intervention outcomes (Horvath and Greenberg, 1989).

2. **Task Collaboration (T):** Shared agreement regarding specific conversational strategies and activities (Horvath et al., 2011).

3. **Interpersonal Bond (B):** The affective relationship, characterized by mutual trust, perceived safety, and positive acceptance (Hatcher and Gillaspy, 2006).

We utilize the 12-item Short-Revised form (WAI-SR; Hatcher and Gillaspy, 2006), adapted into an observer-rated transcript evaluation rubric.

Operationalization of the Patient Ambivalence Scale. Patient ambivalence reflects coexisting motivations for and against behavioral change (Miller and Rollnick, 2012). Because traditional assessment batteries are static pre-treatment instruments (Miller et al., 1992; Miller and Tonigan, 1996), we operationalized an observer-rated rubric consisting of three clinical dimensions:

1. **Problem Uncertainty (U):** The extent of questioning expressed regarding behavioral consequences (Miller and Tonigan, 1996).

2. **Decisional Conflict (C):** Conversational tension between perceived advantages of the status quo and acknowledged negative consequences (Miller et al., 1992).

3. **Conversational Vacillation (V):** Turn-level oscillation between Change Talk and Sustain Talk within encounters (Amrhein et al., 2003).

Automated evaluation configuration. External evaluations are executed using `google/gemma-4-31b-it` with deterministic greedy decoding ($T = 0.0$, max_tokens $= 2500$). The evaluation pipeline is compiled as an asynchronous graph in LangGraph (`StateGraph`). For each session, the judge processes verbatim transcript text alongside the patient profile.

Structured output schema. The evaluation architecture enforces strict separation between clinical appraisal and mathematical scoring:

- **Schema-Constrained Assessment:** The evaluator generates JSON validated against dedicated schemas: `WAILLMResponse` evaluates the 12 WAI-SR items across Goal (G_1–G_4), Task (T_1–T_4), and Bond (B_1–B_4) on a 1–5 scale; `AmbivalenceLLMResponse` evaluates the 9 ambivalence items across Problem Uncertainty (U_1–U_3), Decisional Conflict (D_1–D_3), and Conversational Vacillation (V_1–V_3) on a 1–5 scale.

- **Deterministic Aggregation:** Dimensional averages and composite scores are calculated programmatically:

$$\begin{aligned} \text{WAI}_{\text{composite}} &= \frac{1}{12}\sum_{i=1}^{4}(G_i + T_i + B_i) \\ \text{Amb}_{\text{composite}} &= \frac{1}{9}\sum_{i=1}^{3}(U_i + D_i + V_i) \end{aligned} \tag{16}$$

Multi-session trajectories across baseline architectures. Figures 9 and 10 display multi-session trajectories for Working Alliance and Ambivalence across MuSe-MI, Patient-Ψ, and SimPatient across all three therapist adherence tiers. For working alliance (Figure 9), MuSe-MI discriminates adherence tiers across all dimensions (Goal, Task, Bond), capturing relational rupture under confrontational therapy. Patient-Ψ and SimPatient display compressed variance, showing high baseline alliance

Therapist Condition	Session	Goal Dimension	Task Dimension	Bond Dimension	Composite WAI
MI-Consistent Therapist	Session 1	4.30 ± 0.65	3.99 ± 0.59	4.23 ± 0.43	4.17 ± 0.52
	Session 2	4.68 ± 0.13	4.44 ± 0.26	4.55 ± 0.15	4.56 ± 0.11
	Session 3	4.72 ± 0.20	4.50 ± 0.26	4.61 ± 0.18	4.61 ± 0.17
	Session 4	4.74 ± 0.12	4.52 ± 0.29	4.62 ± 0.14	4.62 ± 0.14
	Overall (1–4)	$\mathbf{4.61 \pm 0.39}$	$\mathbf{4.36 \pm 0.43}$	$\mathbf{4.50 \pm 0.30}$	$\mathbf{4.49 \pm 0.34}$
MI-Inconsistent Therapist	Session 1	2.83 ± 0.76	2.21 ± 0.42	2.08 ± 0.47	2.38 ± 0.45
	Session 2	3.49 ± 0.88	2.57 ± 0.59	2.39 ± 0.68	2.82 ± 0.63
	Session 3	3.73 ± 0.71	2.62 ± 0.44	2.57 ± 0.62	2.97 ± 0.50
	Session 4	3.95 ± 0.67	2.73 ± 0.38	2.92 ± 0.78	3.20 ± 0.55
	Overall (1–4)	$\mathbf{3.50 \pm 0.86}$	$\mathbf{2.53 \pm 0.50}$	$\mathbf{2.49 \pm 0.71}$	$\mathbf{2.84 \pm 0.61}$
Highly MI-Inconsistent Therapist	Session 1	2.12 ± 0.49	1.74 ± 0.27	1.52 ± 0.37	1.79 ± 0.32
	Session 2	2.40 ± 0.57	1.73 ± 0.31	1.55 ± 0.36	1.89 ± 0.36
	Session 3	2.44 ± 0.71	1.73 ± 0.36	1.58 ± 0.45	1.92 ± 0.44
	Session 4	2.27 ± 0.70	1.69 ± 0.39	1.51 ± 0.41	1.82 ± 0.45
	Overall (1–4)	$\mathbf{2.31 \pm 0.63}$	$\mathbf{1.73 \pm 0.33}$	$\mathbf{1.54 \pm 0.39}$	$\mathbf{1.86 \pm 0.39}$

Table 16: Working Alliance Inventory – Short Revised (WAI-SR) subscale and composite scores (Mean $\pm$ SD) across four sessions by therapist adherence tier ($N = 30$).

Therapist Condition	Session	Problem Uncertainty	Decisional Conflict	Conversational Vacillation	Composite Ambivalence
MI-Consistent Therapist	Session 1	1.78 ± 0.83	2.97 ± 0.98	3.04 ± 1.01	2.60 ± 0.79
	Session 2	1.19 ± 0.50	1.71 ± 0.79	1.78 ± 1.02	1.56 ± 0.71
	Session 3	1.09 ± 0.34	1.50 ± 0.82	1.41 ± 0.80	1.33 ± 0.59
	Session 4	1.00 ± 0.00	1.18 ± 0.31	1.13 ± 0.28	1.10 ± 0.17
	Overall (1–4)	$\mathbf{1.26 \pm 0.59}$	$\mathbf{1.84 \pm 1.02}$	$\mathbf{1.84 \pm 1.10}$	$\mathbf{1.65 \pm 0.83}$
MI-Inconsistent Therapist	Session 1	2.21 ± 0.82	3.07 ± 0.81	3.49 ± 0.77	2.92 ± 0.62
	Session 2	1.67 ± 0.77	2.46 ± 0.94	3.01 ± 1.08	2.38 ± 0.85
	Session 3	1.20 ± 0.40	2.02 ± 0.78	2.81 ± 1.02	2.01 ± 0.63
	Session 4	1.20 ± 0.49	1.61 ± 0.79	2.40 ± 1.03	1.74 ± 0.67
	Overall (1–4)	$\mathbf{1.57 \pm 0.76}$	$\mathbf{2.29 \pm 0.99}$	$\mathbf{2.93 \pm 1.05}$	$\mathbf{2.26 \pm 0.82}$
Highly MI-Inconsistent Therapist	Session 1	2.26 ± 0.85	3.08 ± 0.74	3.44 ± 0.70	2.92 ± 0.60
	Session 2	2.04 ± 0.74	2.73 ± 0.92	3.41 ± 0.76	2.73 ± 0.68
	Session 3	1.75 ± 0.78	2.47 ± 1.02	3.00 ± 1.07	2.41 ± 0.80
	Session 4	1.69 ± 0.71	2.09 ± 0.91	2.61 ± 1.01	2.13 ± 0.79
	Overall (1–4)	$\mathbf{1.94 \pm 0.80}$	$\mathbf{2.59 \pm 0.96}$	$\mathbf{3.12 \pm 0.95}$	$\mathbf{2.55 \pm 0.78}$

Table 17: Patient Ambivalence Scale subscale and composite scores (Mean $\pm$ SD) across four sessions by therapist adherence tier ($N = 30$).

even under non-adherent therapy. For patient ambivalence (Figure 10), MuSe-MI demonstrates continuous ambivalence resolution under adherent therapy (Composite Ambivalence decreasing from 2.60 to 1.10), whereas Patient-Ψ remains trapped in elevated decisional conflict across conditions.

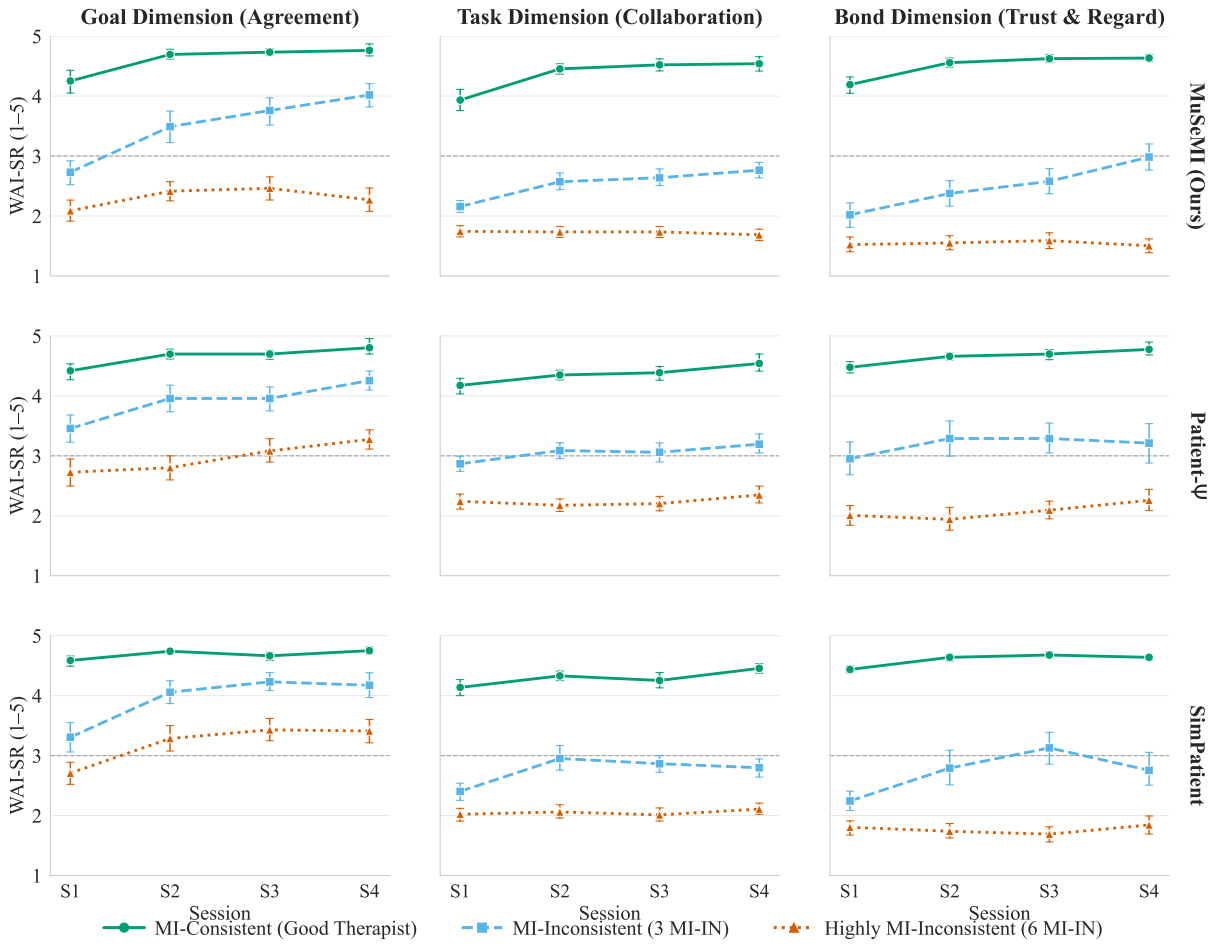


Figure 9: Working Alliance Inventory (WAI-SR; range 1-5) trajectories across four therapy sessions for MuSe-MI, Patient-Ψ, and SimPatient across three therapist adherence tiers. Error bars denote 95% Cousineau-Morey confidence intervals.

		Composite WAI				Composite Ambivalence			
Architecture	**Omnibus Effect**	F	p_{FDR}	η_p^2	**Sig.**	F	p_{FDR}	η_p^2	**Sig.**
MuSe-MI (Ours)	Adherence Tier (Competence)	615.08	$< .001$	0.95	***	47.03	$< .001$	0.62	***
	Session (Time)	31.45	$< .001$	0.52	***	57.49	$< .001$	0.66	***
	Interaction (Session × Adherence)	5.92	$< .001$	0.17	***	4.93	$< .001$	0.15	***
Patient-Ψ	Adherence Tier (Competence)	476.35	$< .001$	0.94	***	29.50	$< .001$	0.50	***
	Session (Time)	10.43	$< .001$	0.26	***	23.38	$< .001$	0.45	***
	Interaction (Session × Adherence)	1.30	.257	0.04	ns	5.26	$< .001$	0.15	***
SimPatient	Adherence Tier (Competence)	679.45	$< .001$	0.96	***	69.90	$< .001$	0.71	***
	Session (Time)	20.58	$< .001$	0.42	***	47.44	$< .001$	0.62	***
	Interaction (Session × Adherence)	4.78	$< .001$	0.14	***	0.42	.865	0.01	ns

Table 18: Omnibus Aligned Rank Transform (ART) ANOVA for Composite WAI and Composite Ambivalence across therapist adherence tiers and sessions ($N = 30$).

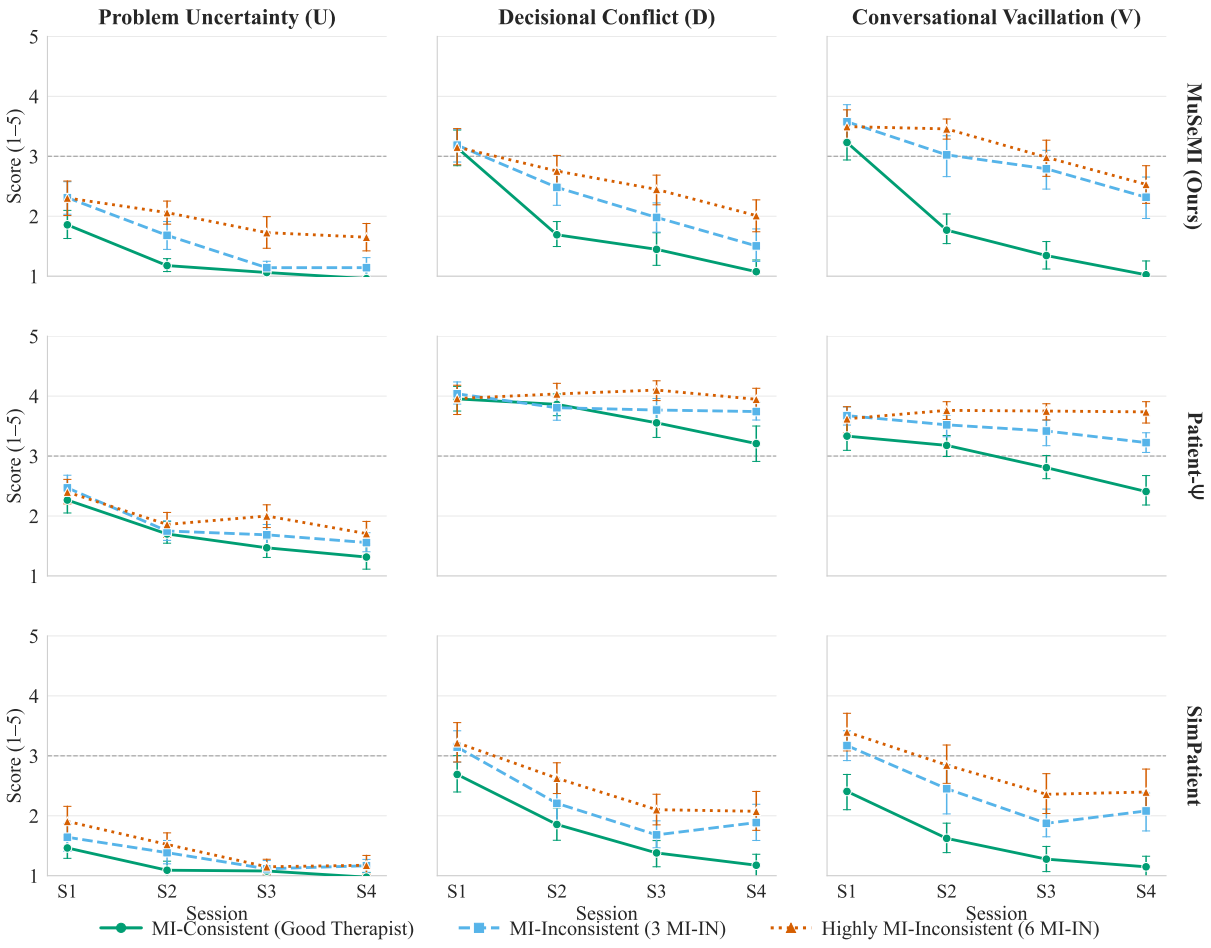


Figure 10: Patient Ambivalence subscale trajectories (range 1–5) across four therapy sessions for MuSe-MI, Patient-Ψ, and SimPatient across three therapist adherence tiers. Error bars denote 95% Cousineau-Morey confidence intervals.

N Multi-Session Internal State Trajectories, State-Space Projections, and Cohort Variance

This appendix provides empirical evaluations of the four internal psychological state variables across sessions, conversational turns, and experimental conditions.

N.1 Statistical Summary of Cumulative Internal State Displacement

Table 19 summarizes cumulative internal state displacement ($\Delta = V_{4,T_4} - V_{1,1}$) across $N = 30$ matched cohorts under all three therapist adherence tiers. Paired t-tests and effect sizes (Cohen's d) evaluate contrasts relative to the MIIN $= 0$ baseline.

N.2 Continuous Turn-Level State Trajectories Across 124 Turns

Figure 11 plots grand mean internal state trajectories across all 124 conversational turns spanning the four sessions, alongside shaded 95% bootstrap confidence intervals. Under MIIN $= 0$, Anger decreases from an initial baseline mean of 29.6 at Turn 1 to 11.5 at Turn 124, while Self-Efficacy increases from 41.1 to 96.7. Numerical shifts are observed at session boundary turns 31, 62, and 93. Under MIIN $= 6$, Anger displays progressive turn-level increases within sessions, terminating at a grand mean of 54.6 in Session 4.

N.3 System Component Attribution Analysis

Figure 12 reports the additive decomposition of cumulative internal state displacement ($\Delta = V_{4,T_4} - V_{1,1}$) across three system components: (1) intra-session Cognitive Interpreter updates ($\lambda_{\text{micro}} I_{k,t}$), (2) inter-session Life Event Simulator shifts ($\lambda_{\text{macro}} \tau(I_k^{\text{macro}})$), and (3) rolling memory consolidation shifts ($\lambda_{\text{macro}} \tau(I_{k,t}^{\text{mem}})$). Numerical contributions vary across adherence tiers. Under MIIN $= 0$, intra-session turns account for $+29.22$ points of the total increase in Self-Efficacy and $+18.31$ points of Motivational Readiness, while Anger shifts by -5.99 points. Under MIIN $= 6$, intra-session dialogue contributes a negative drift of -4.09 points to Self-Efficacy, -1.69 points to Readiness, and $+4.22$ points to Anger. The rolling memory consolidation module contributes $+30.83$ points to Anger under MIIN $= 6$, compared to $+1.43$ points under MIIN $= 0$.

N.4 Cohort Variance and Individual Profile Overlays

Figure 13 displays disaggregated response trajectories of all $N = 30$ patient profiles (faint overlays) plotted alongside the cohort mean trajectory (bold line). Baseline states exhibit wide empirical dispersion across initialized profiles, with Session 1 baseline Anger ranging from 17 to 74 and baseline Self-Efficacy ranging from 21 to 80. Under high perturbation density (MIIN $= 6$), individual profile endpoints display marked variance, with final Session 4 Anger scores spanning between 31 and 88. The preservation of distinct trajectory paths across individual profiles verifies the absence of algorithmic collapse under uniform prompts.

N.5 State-Space Projections Across Dimension Pairs

Figure 14 plots bivariate patient state trajectories across sessions projected onto pairs of dimensions within state space $\mathcal{V}$:

- **Panel A (Anger vs. Motivational Readiness)**: Under MIIN $= 0$ and MIIN $= 3$, endpoints terminate at $(a \leq 22.6, m \geq 82.3)$. Under MIIN $= 6$, trajectories terminate at $(a = 54.6, m = 70.5)$.
- **Panel B (Problem Recognition vs. Self-Efficacy)**: Under MIIN $= 0$, trajectories move from $(28.4, 41.1)$ to $(83.4, 96.7)$. Under MIIN $= 6$, endpoints reach $(61.7, 66.4)$, showing lower coordinate values across both axes.

N.6 Correlation Analyses Across Internal States and External Scales

Table 20 reports Pearson correlation coefficients between internal state variable scores and external session-level evaluations across $M = 120$ observations ($N = 30$ profiles across 4 sessions, MIIN $= 0$ condition). All p-values are adjusted using the Benjamini-Hochberg False Discovery Rate procedure ($df = 118$). Statistical patterns in Table 20 and Figure 15 include:

- Anger correlates positively with Sustain Talk ($r = +0.55, q < 10^{-9}$) and negatively with WAI-SR Bond scores ($r = -0.81, q = 3.02 \times 10^{-28}$).
- Self-Efficacy correlates positively with CAT speech rate ($r = +0.59, q = 3.12 \times 10^{-12}$) and negatively with Sustain Talk ($r = -0.71, q = 2.92 \times 10^{-19}$).

Internal State Dimension	Competence Condition	Mean Shift (Δ)	95% Bootstrap CI	Cohen's d	p-value (t-test)	Empirical Status
Anger	MI-Consistent (MIIN $= 0$)	-18.10	$[-21.77, -14.42]$	–	–	Net Score Reduction
	MI-Inconsistent (MIIN $= 3$)	-7.03	$[-11.22, -2.84]$	$+0.80$	$p < .001$	Attenuated Reduction
	Highly Inconsistent (MIIN $= 6$)	$+24.97$	$[+18.34, +31.60]$	$+2.06$	$p < 10^{-11}$	Net Score Elevation
Self-Efficacy	MI-Consistent (MIIN $= 0$)	$+55.65$	$[+47.38, +63.91]$	–	–	Net Score Elevation
	MI-Inconsistent (MIIN $= 3$)	$+54.84$	$[+47.84, +61.84]$	-0.04	$p = .83$	Statistically Invariant
	Highly Inconsistent (MIIN $= 6$)	$+25.37$	$[+16.08, +34.66]$	-0.93	$p < .001$	Reduced Net Gain
Problem Recognition	MI-Consistent (MIIN $= 0$)	$+54.98$	$[+48.45, +61.51]$	–	–	Net Score Elevation
	MI-Inconsistent (MIIN $= 3$)	$+50.21$	$[+44.65, +55.76]$	-0.40	$p = .036$	Attenuated Net Gain
	Highly Inconsistent (MIIN $= 6$)	$+33.26$	$[+27.89, +38.63]$	-1.08	$p < 10^{-5}$	Suppressed Net Gain
Motivational Readiness	MI-Consistent (MIIN $= 0$)	$+27.49$	$[+21.59, +33.39]$	–	–	Net Score Elevation
	MI-Inconsistent (MIIN $= 3$)	$+28.89$	$[+24.19, +33.58]$	$+0.12$	$p = .51$	Statistically Invariant
	Highly Inconsistent (MIIN $= 6$)	$+15.91$	$[+8.15, +23.66]$	-0.54	$p = .006$	Reduced Net Gain

Table 19: Cumulative internal state displacement ($\Delta = V_{4,T_4} - V_{1,1}$), 95% bootstrap confidence intervals, and paired t-tests relative to MIIN $= 0$ across therapist adherence tiers ($N = 30$).

	Spoken Language Metrics (MISC 2.5)			Working Alliance Inventory (WAI-SR)			Patient Ambivalence Scale	
Internal State (V_k)	**% Sustain**	**% DARN**	**% CAT Action**	**Bond**	**Goal**	**Task**	**Decisional Conflict**	**Vacillation**
Anger (a)	$\mathbf{+0.55}^{***}$	-0.28^{**}	$\mathbf{-0.50}^{***}$	$\mathbf{-0.81}^{***}$	$\mathbf{-0.82}^{***}$	$\mathbf{-0.80}^{***}$	$+0.31^{***}$	$\mathbf{+0.46}^{***}$
Self-Efficacy (s)	$\mathbf{-0.71}^{***}$	$+0.37^{***}$	$\mathbf{+0.59}^{***}$	$\mathbf{+0.65}^{***}$	$\mathbf{+0.72}^{***}$	$\mathbf{+0.76}^{***}$	$\mathbf{-0.60}^{***}$	$\mathbf{-0.62}^{***}$
Problem Recognition (p)	-0.37^{***}	$+0.20^{*}$	$+0.26^{**}$	$+0.39^{***}$	$+0.41^{***}$	$+0.38^{***}$	-0.26^{**}	-0.27^{**}
Motivational Readiness (m)	$\mathbf{-0.57}^{***}$	$+0.37^{***}$	$\mathbf{+0.43}^{***}$	$\mathbf{+0.81}^{***}$	$\mathbf{+0.85}^{***}$	$\mathbf{+0.86}^{***}$	-0.33^{***}	$\mathbf{-0.44}^{***}$

Table 20: Pearson correlation matrix between internal state variables and external evaluation scales across $M = 120$ sessions ($N = 30$, 4 sessions, MIIN $= 0$).

- Motivational Readiness correlates with WAI-SR Goal ($r = +0.85, q < 10^{-31}$) and Task scores ($r = +0.86, q < 10^{-32}$).
- Preparatory exploration (DARN) exhibits correlations between $|r| = 0.20$ and $|r| = 0.37$ across evaluated criteria.

N.7 Turn-Level Logistic Regression Models

Table 21 reports multivariate logistic regression models predicting turn-level binary speech act emissions from continuous internal state variables across $M = 1,800$ conversational turns (30 patients $\times$ 4 sessions $\times$ 15 patient turns, $df = 1,795$). In Model 1, Self-Efficacy (OR $= 1.014, z = +4.05, p < 10^{-4}$) and Readiness (OR $= 1.021, z = +2.76, p = .0058$) are positive predictors of CAT emission, while Anger is a negative predictor (OR $= 0.978, z = -3.73, p < 10^{-3}$). In Model 2, Self-Efficacy is a negative predictor of Sustain Talk emission (OR $= 0.967, z = -7.12, p = 1.07 \times 10^{-12}$). Figure 16 plots empirical speech proportions and fitted curves.

N.8 Correlation Differences Across Therapist Adherence Levels

Table 22 compares Pearson correlation coefficients across therapist adherence tiers ($M = 360$ session-level units, $n = 120$ per tier). Fisher z-transformation tests evaluate the statistical significance of differences between correlation coefficients in the MIIN $= 0$ and MIIN $= 6$ conditions.

The Fisher z-transformation tests indicate:

- The correlation between Self-Efficacy and CAT speech shows no statistical difference between MIIN $= 0$ ($r = +0.59$) and MIIN $= 6$ ($r = +0.43$; $z = 1.66, p = .097$).
- The correlation between Readiness and CAT speech shows no statistical difference between MIIN $= 0$ ($r = +0.43$) and MIIN $= 6$ ($r = +0.49$; $z = -0.62, p = .536$).
- The negative correlation between Self-Efficacy and Sustain Talk attenuates from $r = -0.71$ to $r = -0.21$ ($z = -5.20, p < .001$).
- The negative correlation between Anger and WAI-SR Bond attenuates from $r = -0.81$ to $r = -0.20$ ($z = -7.05, p < .001$).

Predictor ($\mathbf{V}_t$)	**Odds Ratio** (OR)	**95% CI**	z-value	p-value
Model 1: Predicting CAT Talk Emission ($P(CAT_t = 1)$)				
Self-Efficacy (s_t)	$\mathbf{1.014}$	$[1.007, 1.021]$	$+4.05$	5.15×10^{-5}
Readiness (m_t)	$\mathbf{1.021}$	$[1.006, 1.036]$	$+2.76$	0.0058
Anger (a_t)	$\mathbf{0.978}$	$[0.967, 0.990]$	-3.73	0.00019
Problem Recognition (p_t)	0.990	$[0.985, 0.995]$	-4.10	4.08×10^{-5}
Model 2: Predicting Sustain Talk Emission ($P(Sustain_t = 1)$)				
Anger (a_t)	$\mathbf{1.008}$	$[0.998, 1.018]$	$+1.61$	0.108 (n.s.)
Self-Efficacy (s_t)	$\mathbf{0.967}$	$[0.958, 0.976]$	-7.12	1.07×10^{-12}
Readiness (m_t)	$\mathbf{0.982}$	$[0.966, 0.999]$	-2.08	0.037
Problem Recognition (p_t)	0.998	$[0.988, 1.008]$	-0.44	0.662 (n.s.)

Table 21: Multivariate logistic regression models predicting turn-level speech act emission from internal state variables ($M = 1,800$ turns, $df = 1,795$).

	Pearson Correlation Coefficient (r)			**Fisher z-Transformation Test** (MIIN $= 0$ vs. 6)		
Evaluation Pathway	MIIN $= 0$	MIIN $= 3$	MIIN $= 6$	z-score	p-value	**Statistical Status**
Self-Efficacy $\rightarrow$ % CAT Action Talk	$+0.59^{***}$	$+0.47^{***}$	$+0.43^{***}$	$+1.66$	0.097 (n.s.)	Invariant ($p \geq .05$)
Readiness $\rightarrow$ % CAT Action Talk	$+0.43^{***}$	$+0.46^{***}$	$+0.49^{***}$	-0.62	0.536 (n.s.)	Invariant ($p \geq .05$)
Self-Efficacy $\rightarrow$ % Sustain Talk	-0.71^{***}	-0.51^{***}	-0.21^{*}	-5.20	< 0.001	Statistically Different
Anger $\rightarrow$ WAI-SR Bond	-0.81^{***}	-0.42^{***}	-0.20^{*}	-7.05	< 0.001	Statistically Different

Table 22: Pearson correlation coefficients and Fisher z-transformation tests comparing MIIN $= 0$ against MIIN $= 6$ across therapist adherence tiers ($M = 360$ sessions, $df = 117$).

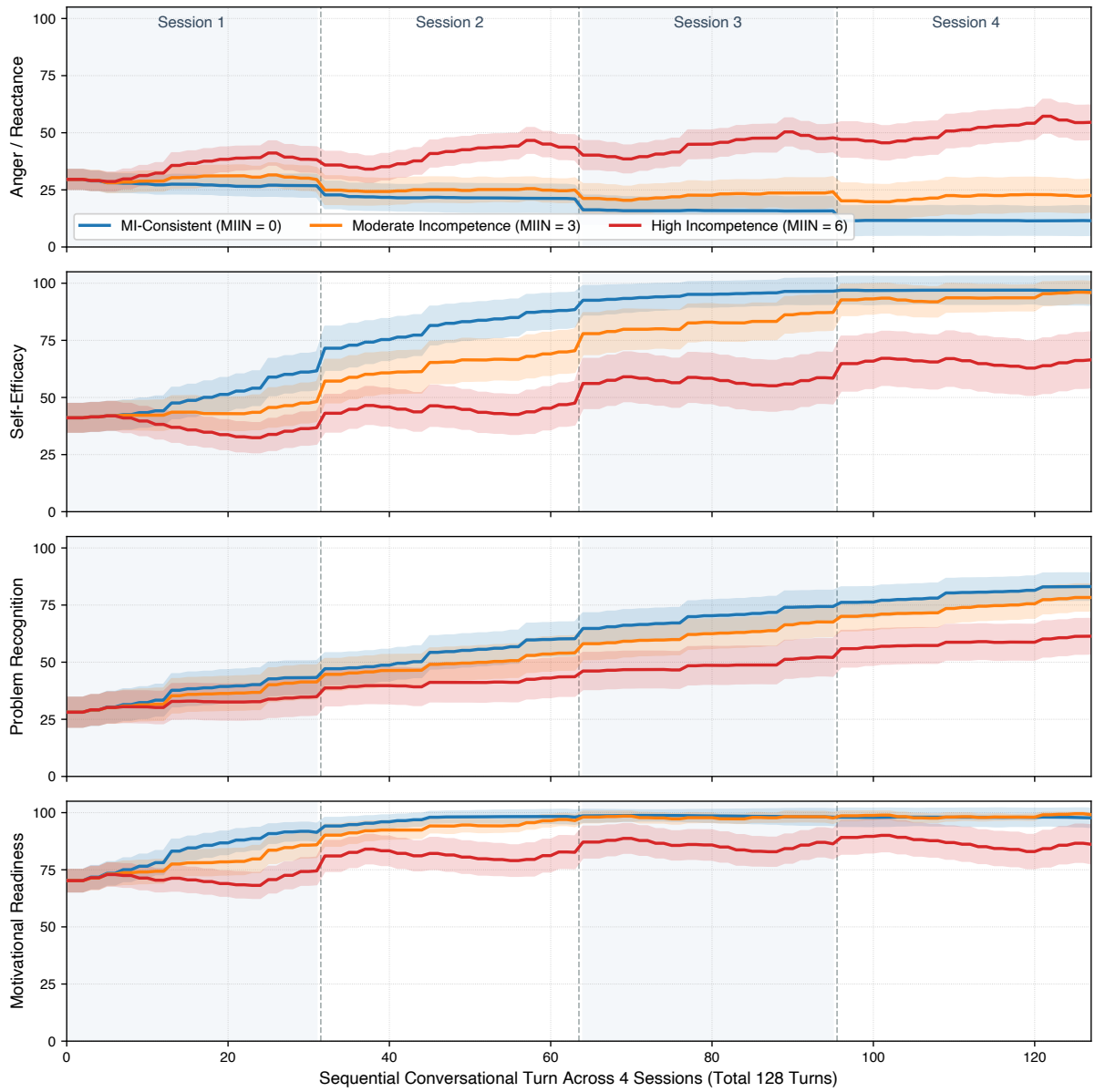


Figure 11: Continuous grand mean internal state trajectories across sequential conversational turns spanning four therapy sessions ($N = 30$ matched cohorts per condition). Shaded regions denote 95% bootstrap confidence intervals. Vertical dashed lines mark inter-session boundaries.

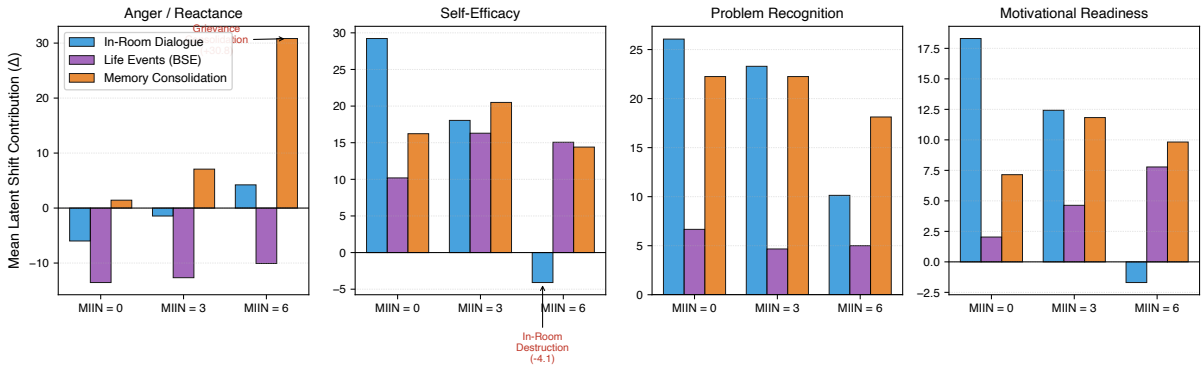


Figure 12: Decomposition of cumulative internal state displacement across intra-session Cognitive Interpreter updates (f_{CI}), inter-session Life Event Simulator shifts (f_{BSE}), and rolling memory consolidation (f_{mem}) ($N = 30$ per condition).

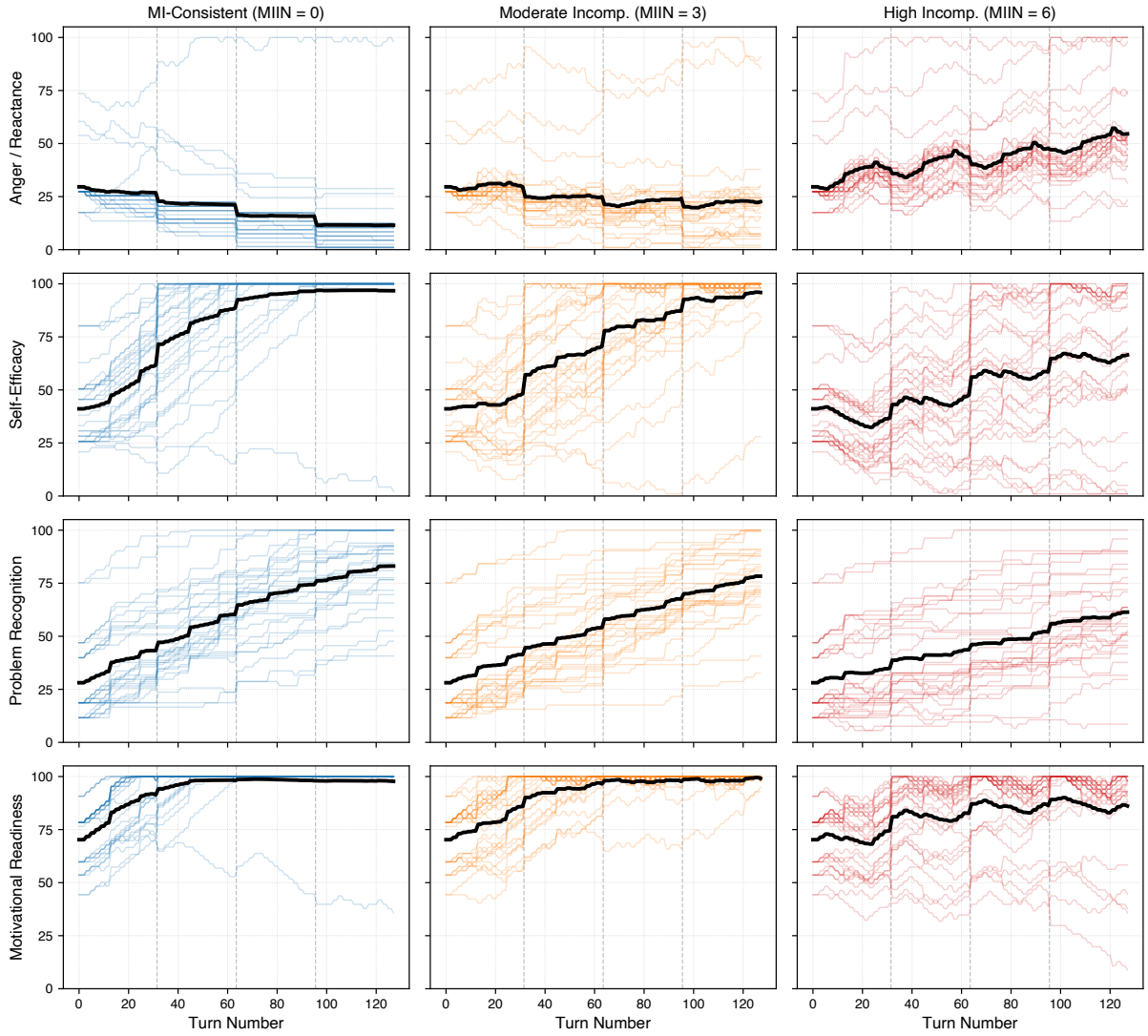


Figure 13: Disaggregated profile-level trajectory overlays across all $N = 30$ simulated patient profiles (faint lines) plotted alongside cohort centroid means (bold line) across competence tiers.

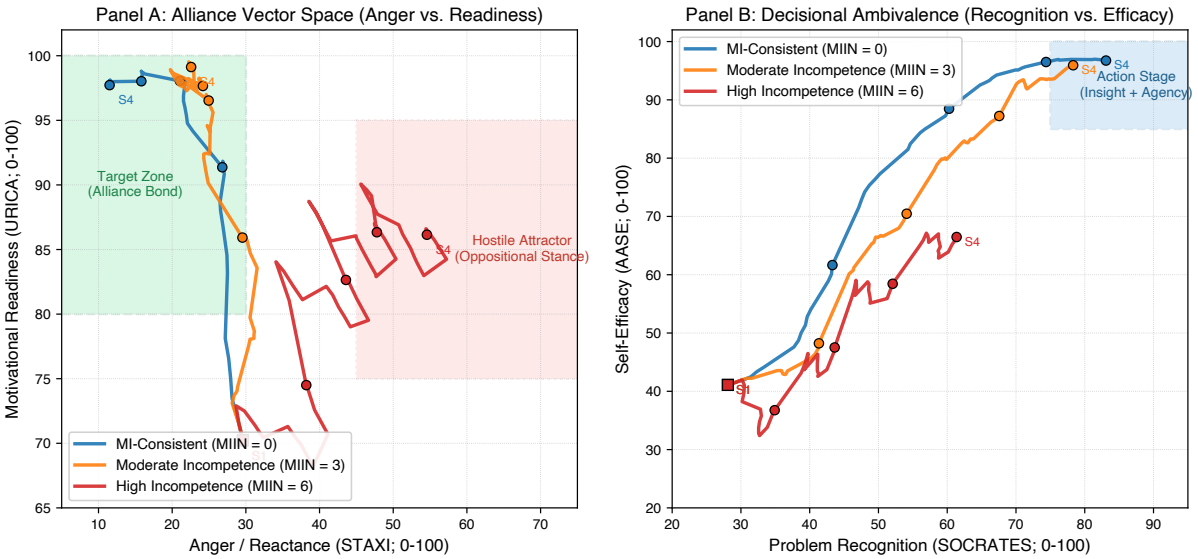


Figure 14: Two-dimensional state-space coordinate projections across four sessions. Panel A maps Anger (a) against Motivational Readiness (m). Panel B maps Problem Recognition (p) against Self-Efficacy (s). Squares mark Session 1 initial states; circles mark Session 4 terminal states.

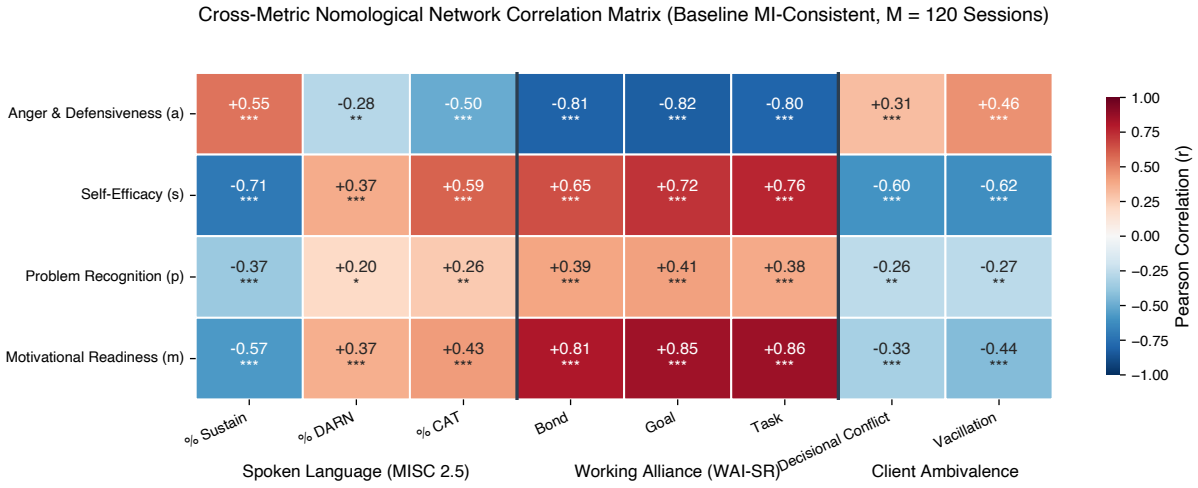


Figure 15: Correlation heatmap displaying Pearson correlation coefficients across $M = 120$ session observations. Color intensity reflects correlation magnitude, centered at zero.

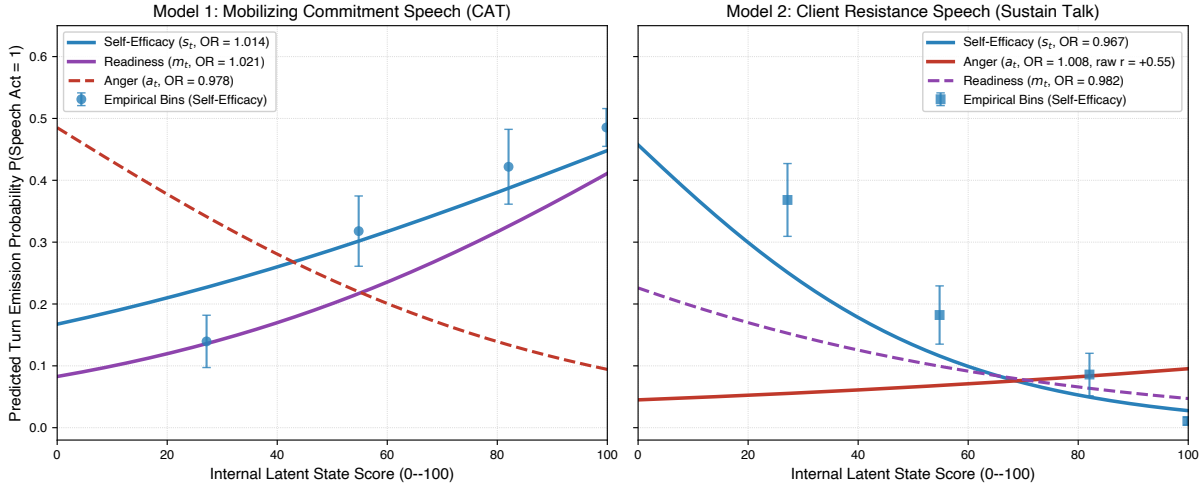


Figure 16: Empirical speech act emission proportions across binned internal state score intervals overlaid with fitted multivariate logistic regression curves ($M = 1,800$ volleys).

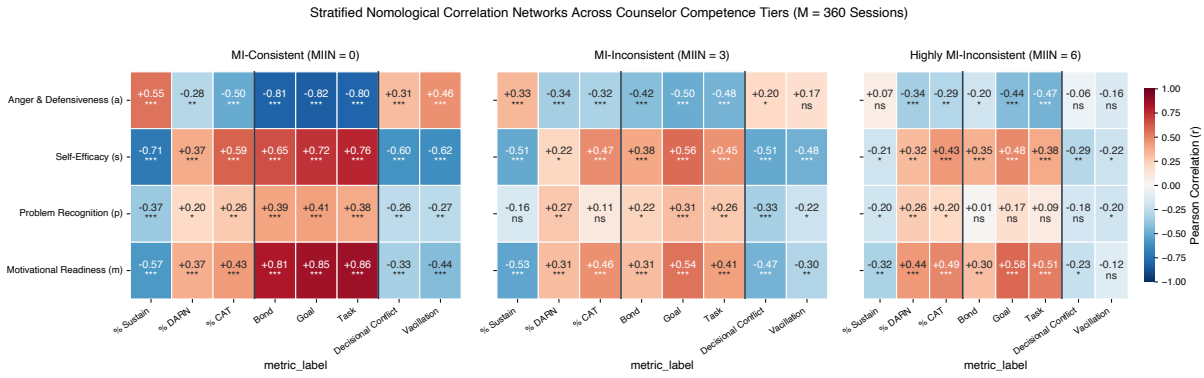


Figure 17: Correlation heatmaps stratified across MIIN $= 0$, MIIN $= 3$, and MIIN $= 6$ conditions ($M = 360$ session-level units).

O AI Assistants

We used AI assistants (Gemini, Claude, and GitHub Copilot) to support code formatting, phrasing suggestions, and LaTeX styling during manuscript preparation. All outputs were reviewed, verified, and edited by the human authors. No content was incorporated without direct author verification.