CALM-IT: Generating Realistic Long-Form Motivational Interviewing Dialogues with Dual-Actor Conversational Dynamics Tracking

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used in mental health–related settings, yet they struggle to sustain realistic, goal-directed dialogue over extended interactions. While LLMs generate fluent responses, they optimize locally for the next turn rather than maintaining a coherent model of therapeutic progress, leading to brittleness and long-horizon drift. We introduce CALM-IT, a framework for generating and evaluating long-form Motivational Interviewing (MI) dialogues that explicitly models dual-actor conversational dynamics. CALM-IT represents therapist–client interaction as a bidirectional state-space process, in which both agents continuously update inferred alignment, mental states, and short-term goals to guide strategy selection and utterance generation. Across large-scale evaluations, CALM-IT consistently outperforms strong baselines in *Effectiveness* and *Goal Alignment* and remains substantially more stable as conversation length increases. Although CALM-IT initiates fewer therapist redirections, it achieves the highest client acceptance rate (64.3%), indicating more precise and therapeutically aligned intervention timing. Overall, CALM-IT provides evidence for modeling evolving conversational state being essential for generating high-quality long-form synthetic conversations.

1 Introduction

Large language models (LLMs) are increasingly used for mental health support through chatbots and conversational agents that many people turn to for emotional guidance. In the United States alone, around 1 in 8 young adults report using generative AI when feeling sad, angry, or anxious (McBain et al., 2025). At the same time, growing evidence highlights serious safety risks: AI chatbots often fail to respond appropriately to suicidal ideation (Pichowicz et al., 2025), may validate delusional beliefs or miss signs of distress (Hall, 2025), and have been linked in media reports to severe distress and self-harm (PBS NewsHour, 2025). Together, these findings reveal a central tension: despite their widespread use, LLM behavior in high-stakes mental health conversations remains unreliable.

A key contributor to these failures is the lack of realistic therapist–client dialogue data. Privacy protections, legal constraints, and stigma severely limit large-scale collection of real therapy conversations (Mayer et al., 2022), leaving existing datasets small, sanitized, or narrowly scoped. As a result, LLMs are typically trained or evaluated on short or simplified exchanges that fail to capture the non-linear dynamics of therapeutic interaction over extended sessions, allowing unsafe behaviors such as sycophancy to persist undetected (Sharma and De Choudhury, 2018). While simulated clients such as Roleplay-doh (Louie et al., 2024) and Patient-Ψ (Wang et al., 2024a) offer scalable alternatives, they do not generate realistic long-form therapeutic dialogue: these frameworks rely on cooperative, fixed personas that respond predictably rather than exhibiting the variability, resistance, and emotional fluctuation characteristic of real therapy. At the same time, therapist agents are optimized for locally appropriate responses to the client's most recent turn, without modeling longer-term therapeutic goals or interactional trajectories, resulting in conversations that lack the bidirectional adaptation, misalignment, and repair processes that define authentic therapeutic dialogue (Kim et al., 2025a; Yang et al., 2025a). To address these gaps, we choose to focus on simulating Motivational Interviewing (MI) conversations and introduce `CALM-IT`, a framework that models therapist–client interaction as a dynamic process. MI provides a particularly stringent testbed because effective practice depends on how rapport, emotion, motivation, and resistance are negotiated over time, rather than on locally appropriate responses. `CALM-IT` generates dialogue as a sequence of state-dependent turns,

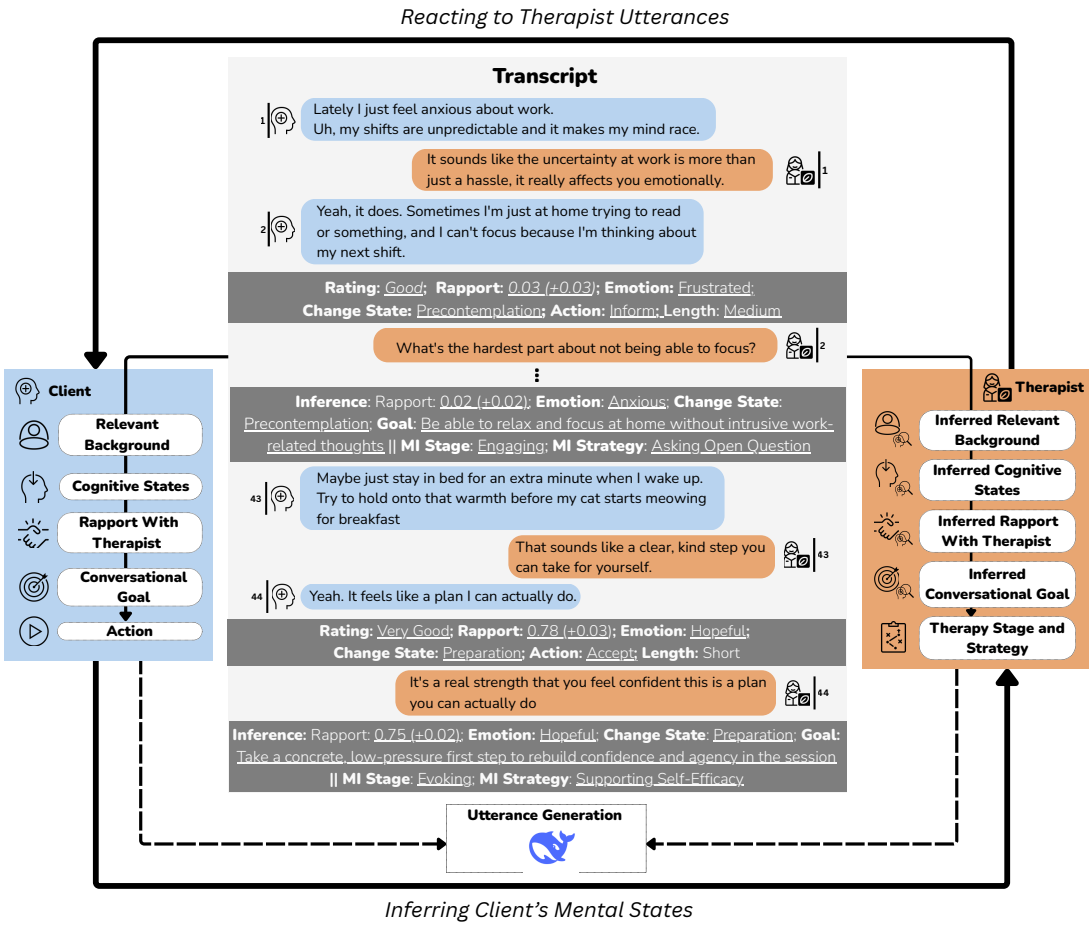

Figure 1: Architecture of the CALM-IT framework. Therapist and client agents interact through state-dependent turns, with latent conversational dynamics inferred and updated throughout the dialogue to guide generation and evaluation of long-form MI interactions.

with both client and therapist continuously updating internal states that govern subsequent behavior. Beyond generation, CALM-IT provides evaluation metrics that assess conversational process rather than isolated turns, supporting analysis of therapeutic dynamics across extended sessions. As a result, CALM-IT serves not only as a generator of realistic long-form MI dialogues, but also as a reproducible environment for stress-testing therapist LLMs under conditions that better approximate real-world mental health conversations. We evaluate CALM-IT through two research questions: (1) Does CALM-IT generate higher-quality Motivational Interviewing (MI) dialogues than existing LLM-based client simulators? and (2) How effectively does CALM-IT model therapist-initiated redirections in MI conversations relative to baseline simulators? Together, these questions assess not only overall MI fidelity, but also whether the framework captures a core therapeutic challenge in MI, namely the frequent emergence of client resistance and the therapist's ability to redirect the conversation to prevent stagnation, reduce negativity, and sustain collaborative progress. Using a combination of LLM-as-a-judge and automatic metrics at the turn, agent, and conversation levels, grounded in prior work on MI process modeling and multi-turn dialogue evaluation, we show that CALM-IT produces more realistic, diverse, and clinically meaningful therapeutic trajectories than prior frameworks and ablated variants. We further find that CALM-IT generates more effective therapist-initiated redirections than all other tested frameworks. By explicitly modeling conversational dynamics to guide generation and evaluation, CALM-IT enables the production of higher-fidelity synthetic therapeutic dialogues that better capture long-horizon interactional structure, providing a stronger foundation for training and evaluating more capable language models tailored towards mental health support.

System	Conv. Dynamics	Client Grounding	Long-horizon Gen.	Action Selection
KMI (Kim et al., 2025a)	×	*Posts*	×	√
CAMI + STAR (Yang et al., 2025a,b)	×	×	×	√
CounseLLMe (De Duro et al., 2025)	×	×	×	×
MentalChat16K (Xu et al., 2025)	×	×	×	×
SQPsych (Vu et al., 2025)	×	*Surveys*	√	×
CBTAlign (Kim et al., 2025b)	×	×	√	×
CALM-IT (ours)	√	*Posts + Surveys*	√	√

Table 1: Comparison of CALM-IT with prior systems generating synthetic mental health support dialogues. CALM-IT uniquely combines conversational dynamics tracking, cognitively grounded patient profiles, long-horizon multi-turn generation, and dual-actor action selection.

2 Related Works

2.1 Synthetic Data for Healthcare LLMs

Recent work has explored using large language models (LLMs) to generate synthetic data for mental health and broader clinical settings, including medical and primary-care conversations (Wang et al., 2024b; He et al., 2020), longitudinal electronic health record–like sequences (Theodorou et al., 2023; Yoon et al., 2023), and clinical or psychiatric notes derived from structured case descriptions (Singhal et al., 2023). In counseling-related domains, prior efforts include empathy-labeled Reddit conversations (Lozoya et al., 2025), multi-turn mental health dialogues (Qiu et al., 2024; Chandra et al., 2025), CBT-oriented counseling datasets (Lee et al., 2024), and MI-focused resources such as smoking-cessation sessions (Mahmood et al., 2025), health coaching agents (Jörke et al., 2025), and the KMI corpus (Kim et al., 2025a). Collectively, these works demonstrate that synthetic data can capture key structural and stylistic properties of therapeutic dialogue while avoiding reliance on sensitive real-world corpora.

2.2 Agentic LLMs for Mental Health Support

Research on agentic LLMs for mental health support spans simulated clients for therapist training, domain-specific therapist agents, and structured therapeutic dialogue datasets. Simulated client systems such as Patient-PSI (Wang et al., 2024a) use structured psychological profiles to produce clinically coherent symptom presentations, while role-based approaches such as Roleplay-doh (Louie et al., 2024) show that lightweight role instructions can yield consistent clinical traits across a dialogue. Other work has focused on therapist-facing or user-facing agents grounded in specific therapeutic frameworks, including MI-oriented systems such as CAMI (Yang et al., 2025a), which integrates strategy selection with latent-state inference to improve MI adherence. Complementary datasets such as KMI (Kim et al., 2025a) and SQPsych (Vu et al., 2025) further illustrate how therapeutic behaviors can be operationalized using structured psychological instruments.

Despite these advances, most existing synthetic datasets and agentic systems operate over short interaction horizons and do not explicitly model how conversational states evolve across extended counseling sessions. Real-world therapeutic interactions often unfold over many turns, involving dynamic shifts in motivation, ambivalence, and engagement (Miller and Rollnick, 2013; Amrhein et al., 2003; Apodaca and Longabaugh, 2009). As a result, prior approaches tend to emphasize local response quality rather than longer-term interactional processes, motivating simulation frameworks that explicitly track and update conversational dynamics over time.

3 The CALM-IT Framework

`CALM-IT` generates long-form Motivational Interviewing (MI) dialogues by explicitly modeling conversational dynamics that evolve over the course of a therapeutic interaction. The framework enables controllable, theory-informed generation of synthetic therapy transcripts that reflect how therapeutic processes progress, stall, and recover over time (Spencer-Oatey, 2004; Premack and Woodruff, 1978; Clark, 1996). `CALM-IT` operationalizes three markers of conversational dynamics described below:

- **Interpersonal Alignment.** Grounded in Rapport Management Theory (Spencer-Oatey, 2004), CALM-IT represents relational alignment as continuously evolving rather than binary. Interpersonal alignment is modeled using relational variables including *rapport* (r)

and the patient's *System 1 evaluation* of the therapist's preceding turn (q), capturing how perceived therapist behavior shapes relational attunement over time without assuming access to ground-truth psychological states (Kahneman, 2011).

- **Mental-state Inference.** Inspired by the Theory of Mind research (Premack and Woodruff, 1978; Baron-Cohen, 1995), `CALM-IT` maintains continuously updated representations of the patient's internal states, including *background* (b), *emotional state* (e), and *stage of change* (c). Crucially, the framework also tracks the therapist's corresponding inferences of these states ($\hat{b}, \hat{e}, \hat{c}$), explicitly distinguishing inferred mental states from unobserved psychological ground truth and grounding therapist behavior in an evolving interpretation of the client's perspective.

- **Joint Activity Structure.** Motivated by theories of joint action (Clark, 1996), `CALM-IT` treats therapy as a collaborative activity by tracking alignment in conversational goals over time. This is operationalized through updates to the patient's *short-term goal* (g) and the therapist's *inferred goal* ($\hat{g}$), enabling detection of goal misalignment and initiation of conversational pivoting when progress stalls.

3.1 Therapist Turn Generation

The therapist agent (THERAAGENT) is designed to mitigate common failure modes of LLMs in extended interpersonal exchanges, including topic drift and misalignment with the client's evolving motivational or emotional state (Kim et al., 2025a). Rather than relying on local textual cues alone, THERAAGENT maintains an explicit internal model of the patient's inferred cognitive, motivational, and relational states, represented using MI-relevant variables such as rapport, stage of change, and inferred client goals. After each patient turn, the agent updates these inferred states and selects a therapeutic strategy conditioned on the updated state, ensuring that response generation reflects an ongoing interpretation of the client's stance rather than isolated turn-level optimization. Unlike prior MI simulation frameworks such as CAMI, which primarily assign cognitive states to the patient, THERAAGENT maintains therapist-side inferred states and uses their joint evolution to guide each response (Yang et al., 2025a).

To prevent conversational stagnation, THERAAGENT monitors interactions for repetitive or unproductive patterns and initiates therapeutic pivots only when such patterns are detected, drawing on repair behaviors observed in real counseling interactions (Nguyen et al., 2025). For each turn, the therapist generates multiple candidate responses conditioned on the selected strategy and current inferred state, which are ranked using a likelihood model fine-tuned on MI-coded therapist turns. Specifically, CALM-IT employs the `turnLikelihoodModel` from ConvoKit, fine-tuned on the AnnoMI corpus (Chang et al., 2020), to favor responses that are both linguistically plausible and consistent with MI-style therapist behavior. Full algorithmic details are provided in Appendix A.

3.2 Client Turn Generation

The client agent (CLIENTAGENT) is designed to capture the reactive and sometimes inconsistent behavior observed in real therapy clients. Following each therapist turn, CLIENTAGENT evaluates the therapist's response using fast, affective appraisal, which updates internal states such as readiness for change, emotional state, rapport, and short-term conversational goals. These updates reflect how clients adjust engagement and motivation in response to feeling understood or misunderstood.

To generate its next turn, the client selects an interactional action (e.g., disclose, resist, clarify, or shift topics), drawing on previous client simulation frameworks with action selection like Roleplay-doh and CAMI (Yang et al., 2025a; Louie et al., 2024). This design allows for hesitation, resistance, and topic shifts rather than enforcing consistent cooperation. As with the therapist, the client generates three candidate turns conditioned on the selected action and current internal state, which are ranked using the same fine-tuned `turnLikelihoodModel`, described above for THERAAGENT. Likelihood-based selection helps maintain coherence in persona, tone, and emotional trajectory across extended interactions while reducing abrupt or implausible behavioral shifts. Full client-side generation details are provided in Appendix A.

4 Experimental Setup

To evaluate the quality of long-form Motivational Interviewing (MI) conversations generated by CALM-IT, we performed controlled generation

and evaluation of therapist-client dialogues across four MI conversation modeling frameworks. The evaluation pipeline consists of three stages: (1) **context generation** from paired client vignettes and psychological profiles, (2) **dialogue generation** using `CALM-IT` and baseline frameworks, and (3) **evaluation** of the resulting transcripts using rubric-based LLM-as-a-judge assessments and algorithmic metrics at the client, therapist, and conversation levels.

4.1 Context Generation

We collect naturally occurring mental health disclosures from Reddit to capture ecologically grounded expressions of psychological distress while preserving anonymity (Brocki et al., 2023; Sharma and De Choudhury, 2018). Using mental health-focused subreddits curated by Sharma and De Choudhury (2018), we gather posts from 2015–2022 (5,794,243 posts across 55 subreddits). We apply BERTopic to cluster the corpus into 158 topics, select five representative posts per topic via cosine similarity, and perform keyword filtering for anxiety-, stress-, or depression-related content. After de-identification, this yields a privacy-preserving set of 686 representative posts (Appendix B).

For each post, we generate a third-person vignette that paraphrases the client's concerns in their own words while removing identifying details, producing a de-identified problem description that can generalize across multiple matched clients. From each vignette, we infer a cognitive–motivational profile (values, needs, session goal, and motivations for change) grounded in MI theory (Miller and Rollnick, 2013). We then simulate DASS-42 responses with GPT-5, where the DASS (Depression, Anxiety, and Stress Scales) provides a standardized psychometric characterization of symptom severity used to anchor the client's affective and motivational state. Each vignette is matched to a respondent from the OpenPsychometrics DASS dataset (39,775 anonymized records; Open Psychometrics, 2019), excluding respondents under 18, by maximizing weighted Cohen's κ and retaining only high-confidence matches ($\kappa > 0.6$, 99th percentile) with demographic consistency, selecting the highest-scoring match per vignette.

For each matched pair, we use DeepSeek-V3 (Liu et al., 2024) to generate a detailed client background (e.g., family structure, living situation, routines, cultural values) and a single-sentence session goal, grounded in the DASS responses of matched OpenPsychometrics, specifying what the client aims to address in therapy. These materials serve as fixed context for dialogue generation across all frameworks.

4.2 Dialogue Generation

Using the simulated client contexts, we generate multi-turn therapeutic dialogues with four counseling frameworks selected to compare state-of-the-art MI systems against CALM-IT variants that differ only in whether conversational dynamics are explicitly modeled:

1. `CAMI + STAR (C + S)` (Yang et al., 2025a,b),
2. `KMI` (Kim et al., 2025a),
3. `CALM-IT` without conversational-dynamics modeling (`CI-NC`).
4. `CALM-IT` with conversational-dynamics modeling (`CI`),

Each framework is applied to every simulated client, producing four dialogue trajectories per client. For each trajectory, we generate conversations at three target lengths (30, 50, and 100 turns) to capture short, medium, and extended interactions. All sessions begin with a standardized opening turn and reserve the final ten turns for session closure using a standardized ending turn (Appendix A).

In total, this procedure yields $686 \times 3 \times 4 = 8{,}232$ transcripts. Details on model configurations, hyperparameters, compute infrastructure, and runtime costs are provided in Appendix D.

4.3 Evaluation

We introduce conversation-level metrics grounded in the Motivational Interviewing Treatment Integrity (MITI) code, MI process theory, and prior work on conversational redirection (Moyers et al., 2016; Miller and Rollnick, 2013; Apodaca and Longabaugh, 2009; Moyers et al., 2009; Nguyen et al., 2024). These include Effectiveness, Directionality and Goal Alignment. We additionally measure self-consistency (within agent) and entailment (between therapist and client) using a dialogue-specific NLI model, reporting each model's deviation from AnnoMI reference means to reflect that both under- and over-alignment reduce realism for mental health-related conversations (Hayes et al., 2007).

For redirection, we use ConvoKit's framework with an MI-conditioned likelihood model to estimate therapist-initiated redirection intensity (Chang et al., 2020). Metrics include % Accepted Redirection and Δ Change Talk @ Greatest Redirection (Nguyen et al., 2024). Change talk and sustain talk are labeled using a few-shot GPT-5 classifier with high accuracy on AnnoMI (Appendix C), and overall conversational directionality is measured using DialoFlow (Li et al., 2021).

For holistic constructs such as empathy, partnership, and overall MI adherence, we apply a rubric-based LLM-as-a-judge protocol on a 5-point Likert scale using DeepEval (Ip and Vongthongsri, 2025) with GPT-5-mini, whose multi-pass aggregation improves score stability. Statistical differences between paired conditions are assessed using two-sided Wilcoxon signed-rank tests (with Holm-Bonferroni correction applied). In addition, approximately 0.5% of transcripts (48 total; 4 per framework–length condition) are independently reviewed by 4 licensed psychologists to validate LLM-as-a-judge metrics. Using weighted Cohen's κ, moderate to substantial agreement was observed for all rubric-based metrics. Full details on all the evaluation metrics used are described in tables A2-A14. In addition, full definitions and details on the human annotation process appear in Appendix C.4.

Metric	**KMI**	**CI-NC**	**C+S**	**CI**
Turn-level				
Client				
Client Readability (↓)	5.87	4.68	5.77	**4.45**
Therapist				
Reflection Quality	4.44	3.71	2.38	**4.61**
Question Quality	4.52	4.39	2.94	**4.97**
Therapist Readability (↓)	6.22	4.98	6.18	**4.93**
Agent-level				
Client				
Consistency	3.76	4.79	3.42	**4.92**
Therapist				
Cultivating Change Talk	**4.63**	3.15	1.88	4.34
Soften Sustain Talk	4.73	4.27	2.53	**4.88**
Empathy	4.66	4.68	3.53	**4.88**
Partnership	4.48	4.60	2.71	**4.88**
Reflection–Question Ratio	**2.05**	0.92	0.90	1.03
% Open Question	97.30	97.90	96.80	**98.47**
Conversation-level				
Effectiveness	4.27	2.86	1.19	**4.45**
Realignment	4.49	4.53	3.21	**4.87**
Goal Alignment	4.60	3.89	2.13	**4.73**
Δ_{Ref} Directionality (↓)	0.081	0.03	0.048	**0.005**
Δ_{Ref} Avg. Self-Consist. (↓)	0.291	0.141	**0.072**	0.157
Δ_{Ref} Avg. Entail. (↓)	0.092	0.141	0.073	**0.042**

Table 2: Turn-, agent- and conversation-level evaluation of MI dialogue generated by four frameworks: KMI, CALM-IT without conversational dynamics (CI-NC), CAMI+STAR (C+S), and CALM-IT (CI). Scores are averaged across all conversations and lengths. Downward arrow (↓) indicates that lower values correspond to better performance; otherwise, higher values are better. Best-performing framework for each metric is bolded (RQ1)

5 Research Question 1

5.1 Results

We present summary statistics for generated transcripts across all framework-length combinations in table A1 of the Appendix. Looking at per-framework averages across different lengths, we found that CALM-IT (CI) consistently outperforms all baselines, including KMI, CAMI+STAR (C+S), and the ablated variant CALM-IT(NC), across turn-, agent-, and especially conversation-level evaluations. Performance gains are most pronounced at the conversational level, where CALM-IT achieves the highest overall *Effectiveness* score (4.45), representing improvements of 4.22% over KMI ($p < .001$), 55.59% over CALM-IT(NC) ($p < .001$), and 273.95% over C+S ($p < .001$). Similar trends hold for other conversational metrics, including *Goal Alignment* (4.73), with significant difference over all baselines ($p < 0.001$). At the agent- and turn-levels, improvements are more modest but remain consistent: CALM-IT attains the highest scores on three of the four MITI 4.2 Global Ratings for therapist behavior (*Softening Sustain Talk*, *Empathy*, and *Partnership*, all 4.88), while its CLIENTAGENT produces more readable turns (readability 4.45 vs. 5.87 for KMI and 5.76 for CAMI+STAR, $p < 0.001$) and exhibits greater behavioral consistency with assigned client backgrounds (consistency 4.92 vs. 3.76 for KMI and 3.36 for CAMI+STAR, $p < 0.001$) than competing frameworks.

In addition to higher average performance across all levels of metrics, CALM-IT exhibits substantially greater stability in performance when asked to generate extended-format conversations. Figure 2 summarizes relative changes in evaluation metrics as conversations scale from short (30 turns) to long (100 turns). We found that CALM-IT shows the smallest absolute overall drift, with a mean absolute percentage change of only 0.02% across all metrics, compared to 2.41% for KMI, 0.72% for CALM-IT (NC), and 3.56% for CAMI+STAR. This effect is even more pronounced for conversational-

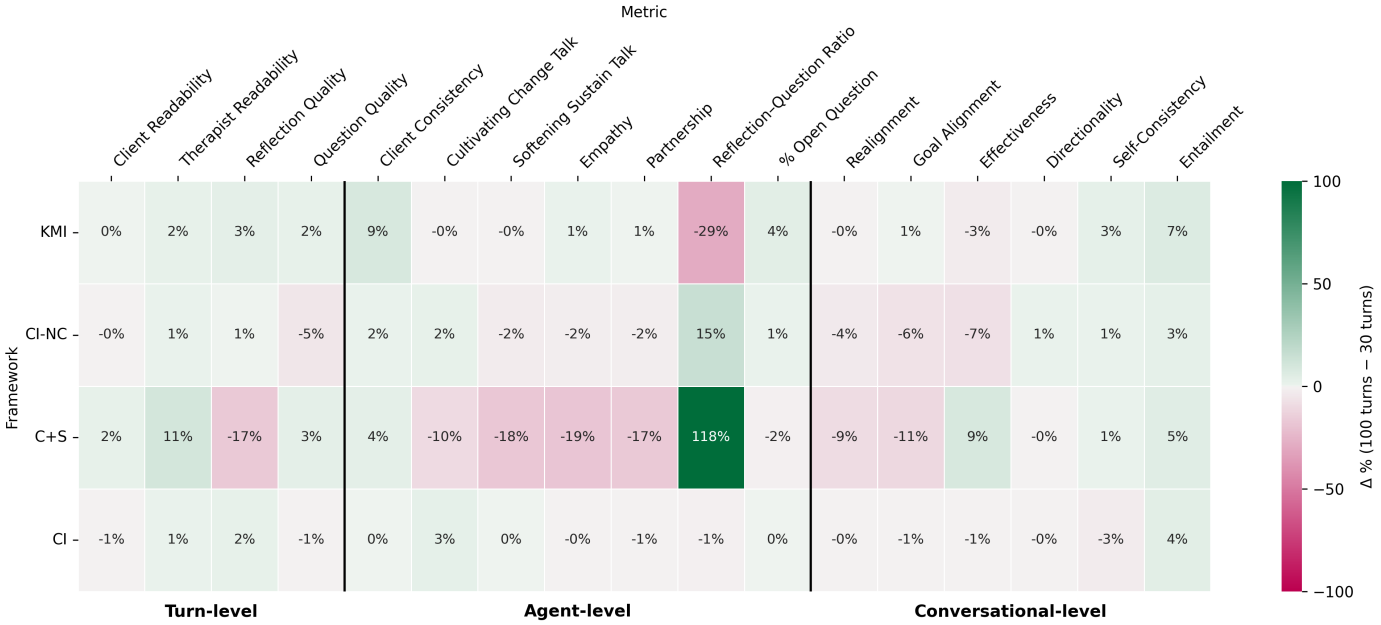

Figure 2: Percentage change in performance across metrics from short-form conversations (30 turns) to long-form conversations (100 turns), averaged across all pairs of short-long conversations. Positive or close-to-zero values indicate stable performance, whereas negative values indicate degradation in performance. (RQ1)

level metrics, where CALM-IT only showed a -1.19% degradation on average compared to -7.05% for KMI or -8.48% for CI-NC. These results suggest that CALM-IT scales more gracefully with interaction length, maintaining stable conversational structure and effectiveness as dialogues become longer.

6 Research Question 2

6.1 Results

We operationalize long-horizon stability by measuring the change in each conversation-level metric between 30-turn and 100-turn transcripts generated by the same framework. We found that although CALM-IT redirects less frequently overall, as reflected by its lowest redirection_ratio (0.594), it produces more effective therapist redirection than all baseline frameworks. This is shown, firstly by *percentage accepted redirection* measurements, which captures the client's immediate response following a redirection event, CALM-IT achieves the highest acceptance rate at 64.28%, significantly outperforming KMI at 59.11% ($p < 0.001$), CAMI+STAR at 54.95% ($p < 0.001$), and CI-NC at 52.14% ($p < 0.001$). Beyond immediate acceptance, CALM-IT's strongest redirection events yield sustained downstream benefits, with increased client change talk of 12.4% and substantial reductions in sustain talk of -8.15% on average (compared to 9.499% increase in change talk and -1.667% reduction in sustain talk for KMI).

To quantify the asymmetry observed in Figure 3, we test whether the distribution of Δ sustain talk exhibits statistically significant negative skewness. CALM-IT shows a significant left-skew, with skewness $\gamma = -0.664$ and a D'Agostino skewness test statistic of $Z = -6.556$ ($p < 0.0001$), indicating that strong redirection events are systematically followed by reductions in client resistance. In contrast, the KMI distribution is skewed weaker to the left with $Z = -5.855, p < 0.0001$, while CAMI+STAR and CI-NC exhibit significant right-skew with $\gamma = 0.191$ ($p < 0.0001$) and $\gamma = 0.656$ ($p < 0.0001$), reflecting weak or counterproductive redirection that often results in increases of sustain talk. This demonstrates that CALM-IT's selective redirection strategy uniquely shifts post-redirection outcomes toward reduced sustain talk and increased change talk.

7 Discussion & Conclusion

In this work, we introduced CALM-IT, a framework for generating and evaluating long-form Motivational Interviewing (MI) dialogues that explicitly models *dual-actor conversational dynamics*. CALM-IT represents therapist–client interaction as a bidirectional process in which both agents continuously update inferred alignment, mental states,

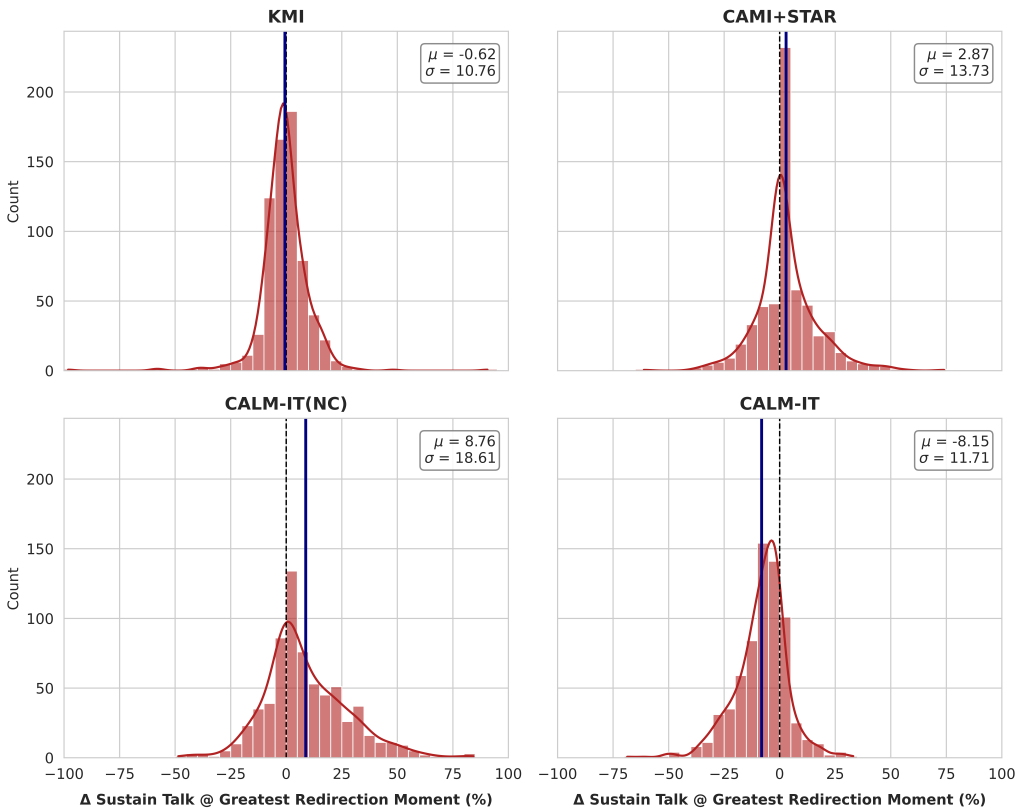

Figure 3: Distribution of changes in client sustain talk at the greatest therapist-initiated redirection moment, for all four tested frameworks. Negative values indicate reductions in sustain talk following redirection, while positive values indicate increases. Solid vertical lines denote the mean change for each framework, and dashed lines represent no change (RQ2)

and conversational goals, shaping responses over the course of the dialogue. Through large-scale dialogue generation and evaluation across multiple baseline frameworks and conversation lengths, we show that `CALM-IT` consistently outperforms prior approaches, producing more coherent, stable, and effective long-horizon interactions. These findings reinforce longstanding insights from MI process theory and psychotherapy research that therapeutic change emerges from long-term interactional trajectories rather than isolated counselor behaviors (Spencer-Oatey, 2004; Clark, 1996; Grosz and Sidner, 1986). A key insight from our results is that long-horizon conversational stability requires explicit modeling of interactional state, which is not provided by vanilla next-token LLM generation. Although LLM generations are conditioned on dialogue history, they are optimized for local next-turn plausibility rather than maintaining or updating latent representations of conversational progress, alignment, or goals, leading to degradation and compounding errors in extended interactions (Laban et al., 2025; Li et al., 2025). `CALM-IT` addresses this limitation by explicitly tracking and regulating conversational state outside the surface text, enabling strategy recalibration before drift becomes entrenched. Our redirection analysis further supports this view, showing that explicit modeling of conversational dynamics enables principled regulation of interactional trajectories. Rather than treating redirection as a fixed heuristic, `CALM-IT` conditions interventions on inferred conversational state—such as rapport, stage of change, and goal alignment—resulting in fewer but more effective redirections. From a modeling perspective, these results suggest a general design principle for long-form dialogue systems: interventions such as topic shifts, or clarification should be implemented as control actions governed by inferred conversational state. By decoupling state inference, strategy selection, and turn realization, `CALM-IT` demonstrates how long-horizon conversational control can be operationalized within agentic LLM architectures, with implications beyond the mental health counseling domain.

8 Limitations

While CALM-IT enables the generation of long-form Motivational Interviewing (MI) dialogues with richer conversational structure than prior simulators, several limitations remain. Real-world therapeutic interactions are shaped by complex, nonlinear dynamics arising from interpersonal history, situational context, and unobserved client factors, which CALM-IT only partially approximates through large language model–based state inference and prompting. These approximations may diverge from how therapeutic processes unfold in real clinical settings, motivating future work on more specialized or hybrid models that can more faithfully capture such dynamics in real time.

Due to computational constraints, our experiments focus on a single large language model backbone (DeepSeek-V3) for both therapist and client generation. While this choice enables controlled comparisons across frameworks, the extent to which different backbone models affect long-form dialogue quality, conversational dynamics, and redirection behavior remains an open question. Future work could systematically evaluate CALM-IT and other frameworks across diverse backbone models to better understand the interaction between the quality of transcript generation, conversational state modeling and underlying language model capabilities.

In addition, our framework is explicitly grounded in MI theory, which enables fine-grained modeling and evaluation of motivational and relational processes but limits immediate generalization to other therapeutic modalities; future extensions could adapt CALM-IT to alternative approaches such as Cognitive Behavioral Therapy (CBT) by incorporating modality-specific state representations, strategies, and evaluation criteria. CALM-IT also relies on Reddit mental health disclosures as proxy data for real-world therapy contexts, which disproportionately reflect WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations and are shaped by social-media-specific norms of self-disclosure, including intimacy costs that constrain the range of expressed distress (Ma et al., 2016). Consequently, clinically salient scenarios such as acute crisis, severe psychiatric conditions, or culturally specific concerns may be underrepresented, limiting the diversity of generated synthetic transcripts. Finally, several therapeutic process constructs are evaluated using LLM-as-a-judge methods rather than fully algorithmic metrics. While this enables scalable assessment of high-level constructs such as empathy, partnership, and reflection quality, these judgments are sensitive to model choice and prompting and should be interpreted as approximations of expert assessment rather than objective ground truth.

9 Ethical Considerations

CALM-IT is a research framework for studying long-horizon conversational dynamics in motivational interviewing, not a system intended for clinical use. The realism of the simulated dialogues should not be taken as an endorsement of automated mental health support. We do not support the deployment of systems for mental health diagnosis, treatment, or crisis response without appropriate regulatory review. Any real-world use would require FDA clearance, with FDA approval serving as a longer-term objective for higher-risk or autonomous clinical applications. The contribution of this work lies in developing evaluation pipelines, stress-testing methods, and synthetic data resources that can support such regulatory validation efforts.

The framework relies on synthetic data to enable large-scale construction and evaluation of systems for mental health support, avoiding the privacy and legal constraints associated with real therapy transcripts. This choice also introduces limitations. Although grounded in mental health disclosures from Reddit, the resulting simulations may reflect existing biases in mental health representation, including demographic, cultural, and normative assumptions present in the source data and modeling choices. These biases can affect simulated client behavior, therapist evaluations, and assessments of therapeutic progress, potentially obscuring uneven performance across populations.

Findings obtained using CALM-IT should therefore be interpreted as characterizations of conversational behavior under controlled experimental conditions, not as evidence of clinical safety or effectiveness. Use of realistic mental health simulations benefits from clear boundaries, explicit documentation of data sources and assumptions, and expert oversight. In this role, synthetic frameworks such as CALM-IT support research, training, and regulatory preparation, while remaining distinct from the delivery of professional mental health care.

Finally, to support reproduction and extension of this work, we will release the code and vignettes

used for transcript generation upon the acceptance of the manuscript.

References

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.

Timothy R Apodaca and Richard Longabaugh. 2009. Mechanisms of change in motivational interviewing: A review and preliminary evaluation of the evidence. *Addiction*, 104(5):705–715.

Simon Baron-Cohen. 1995. Mindblindness: An essay on autism and theory of mind. *MIT Press*.

Lennart Brocki, George C Dyer, Anna Gładka, and Neo Christopher Chung. 2023. Deep learning mental health dialogue system. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 395–398. IEEE.

Mohit Chandra, Siddharth Sriraman, Harneet Singh Khanuja, Yiqiao Jin, and Munmun De Choudhury. 2025. Reasoning is not all you need: Examining llms for multi-turn mental health conversations. *arXiv preprint arXiv:2505.20201*.

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. In *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue*, pages 57–60.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Edoardo Sebastiano De Duro, Riccardo Improta, and Massimo Stella. 2025. Introducing counsellme: A dataset of simulated mental health dialogues for comparing llms like haiku, llamantino and chatgpt against humans. *Emerging Trends in Drugs, Addictions, and Health*, 5:100170.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pages 3806–3812.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Rachel Hall. 2025. Ai chatbots giving dangerous advice to mentally ill, psychologists warn. https://www.theguardian.com/technology/2025/nov/30/chatgpt-dangerous-advice-mentally-ill-psychologists-openai. The Guardian. Accessed: 2025-12-01.

Adele M Hayes, Jean-Philippe Laurenceau, Greg Feldman, Jennifer L Strauss, and LeeAnn Cardaciotto. 2007. Change is not always linear: The study of nonlinear and discontinuous patterns of change in psychotherapy. *Clinical psychology review*, 27(6):715–723.

Xuehai He, Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, and 1 others. 2020. Meddialog: Two large-scale medical dialogue datasets. *arXiv preprint arXiv:2004.03329*.

Jeffrey Ip and Kritin Vongthongsri. 2025. deepeval.

Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James A Landay. 2025. Gptcoach: Towards llm-based physical activity coaching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–46.

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

Hyunjong Kim, Suyeon Lee, Yeongjae Cho, Eunseo Ryu, Yohan Jo, Suran Seong, and Sungzoon Cho. 2025a. Kmi: A dataset of korean motivational interviewing dialogues for psychotherapy. *arXiv preprint arXiv:2502.05651*.

Yejin Kim, Chi-Hyun Choi, Selin Cho, Jy-yong Sohn, and Byung-Hoon Kim. 2025b. Aligning large language models for cognitive behavioral therapy: a proof-of-concept study. *Frontiers in Psychiatry*, 16:1583739.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

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

Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. 2025. Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. *arXiv preprint arXiv:2106.02227*.

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

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

Daniel Cabrera Lozoya, Eloy Hernández Lúa, Juan Alberto Barajas Perches, Mike Conway, and Simon D'Alfonso. 2025. Synthetic empathy: Generating and evaluating artificial psychotherapy dialogues to detect empathy in counseling sessions. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 157–171.

Xiao Ma, Jeff Hancock, and Mor Naaman. 2016. Anonymity, intimacy and self-disclosure in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 3857–3869.

Zafarullah Mahmood, Soliman Ali, Jiading Zhu, Mohamed Abdelwahab, Michelle Yu Collins, Sihan Chen, Yi Cheng Zhao, Jodi Wolff, Osnat C Melamed, Nadia Minian, and 1 others. 2025. A fully generative motivational interviewing counsellor chatbot for moving smokers towards the decision to quit. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25008–25043.

Tobias Mayer, Neha Warikoo, Oliver Grimm, Andreas Reif, and Iryna Gurevych. 2022. Gdpr compliant collection of therapist-patient-dialogues. *arXiv preprint arXiv:2211.12360*.

Ryan K McBain, Robert Bozick, Melissa Diliberti, Li Ang Zhang, Fang Zhang, Alyssa Burnett, Aaron Kofner, Benjamin Rader, Joshua Breslau, Bradley D Stein, and 1 others. 2025. Use of generative ai for mental health advice among us adolescents and young adults. *JAMA Network Open*, 8(11):e2542281–e2542281.

William R. Miller and Stephen Rollnick. 2013. *Motivational Interviewing: Helping People Change*, 3rd edition. Guilford Press.

Theresa B Moyers, Tim Martin, Jon M Houck, Paulette J Christopher, and J Scott Tonigan. 2009. From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. *Journal of consulting and clinical psychology*, 77(6):1113.

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

Vivian Nguyen, Sang Min Jung, Lillian Lee, Thomas D Hull, and Cristian Danescu-Niculescu-Mizil. 2024. Taking a turn for the better: Conversation redirection throughout the course of mental-health therapy. *arXiv preprint arXiv:2410.07147*.

Vivian Nguyen, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. 2025. Hanging in the balance: Pivotal moments in crisis counseling conversations. *arXiv preprint arXiv:2506.03941*.

Open Psychometrics. 2019. Depression anxiety stress scales (dass) dataset. https://openpsychometrics.org/_rawdata/DASS_data_21.02.19.zip.

PBS NewsHour. 2025. What to know about "ai psychosis" and the effect of ai chatbots on mental health. https://www.pbs.org/newshour/show/what-to-know-about-ai-psychosis-and-the-effect-of-ai-chatbots-on-mental-health. Accessed: 2025-08-15.

Wojciech Pichowicz, Michal Kotas, and Pawel Piotrowski. 2025. Performance of mental health chatbot agents in detecting and managing suicidal ideation. *Scientific Reports*, 15(1):31652.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636.

Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13.

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

Helen Spencer-Oatey. 2004. *Culturally speaking: Managing rapport through talk across cultures*. A&C Black.

Brandon Theodorou, Cao Xiao, and Jimeng Sun. 2023. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature communications*, 14(1):5305.

Doan Nam Long Vu, Rui Tan, Lena Moench, Svenja Jule Francke, Daniel Woiwod, Florian Thomas-Odenthal, Sanna Stroth, Tilo Kircher, Christiane Hermann, Udo Dannlowski, and 1 others. 2025. Roleplaying with structure: Synthetic therapist-client conversation generation from questionnaires. *arXiv preprint arXiv:2510.25384*.

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

Sheng Wang, Zihao Zhao, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. 2024b. Interactive computer-aided diagnosis on medical image using large language models. *Communications Engineering*, 3(1):133.

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025. Mentalchat16k: A benchmark dataset for conversational mental health assistance. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5367–5378.

Yizhe Yang, Palakorn Achananuparp, He-Yan Huang, Jing Jiang, Phey Ling Kit, 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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21037–21081.

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

Jinsung Yoon, Michel Mizrahi, Nahid Farhady Ghalaty, Thomas Jarvinen, Ashwin S Ravi, Peter Brune, Fanyu Kong, Dave Anderson, George Lee, Arie Meir, and 1 others. 2023. Ehr-safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *NPJ digital medicine*, 6(1):141.

A Appendix A: Additional details regarding CALM-IT framework

We provide full algorithmic and prompt-level details of the CALM-IT framework, including therapist and client generation procedures (Algorithm 1, 2), prompts for conversational state updates, and latent-state inferences used to guide long-form dialogue generation below.

Algorithm 1 Therapist Turn Generation

Inputs. M (history), Σ (summary), $\hat{b}$ (inferred background), $\hat{e}$ (inferred emotion), $\hat{c}$ (inferred stage of change), $\hat{m}$ (inferred motivational readiness), $\hat{g}$ (inferred patient goal), $\hat{r}$ (inferred rapport), a (strategy), t (turn counter).

/ Update summary of dialogue */*
$\Sigma \leftarrow \text{UpdateSummary}(\Sigma, M)$
/ Infer patient background */*
$\hat{b} \leftarrow \text{InferBackground}(\Sigma, M, \hat{b})$
/ Infer patient's emotional state */*
$\hat{e} \leftarrow \text{InferEmotion}(\Sigma, M, \hat{e})$
/ Infer patient's stage of change */*
$\hat{c} \leftarrow \text{InferStageOfChange}(\Sigma, M, \hat{c})$
/ Infer patient's short-term conversational goal */*
$\hat{g} \leftarrow \text{InferGoal}(\Sigma, M, \hat{g})$
/ Infer rapport between therapist and patient */*
$r \leftarrow \text{InferRapport}(\Sigma, M, \hat{e}, \hat{r})$
/ Detect stagnation in dialogue */*
$\text{rep} \leftarrow \text{CheckRepetition}(\Sigma, M)$
if $\text{rep} = 0$ **then**
    / Select MI-consistent strategy */*
    $a \leftarrow \text{SelectStrategy}(\Sigma, M, \hat{b}, \hat{e}, \hat{c}, \hat{g}, \hat{r})$
else
    / Select pivot strategy to restore progress */*
    $a \leftarrow \text{SelectPivot}(\Sigma, M, \hat{b}, \hat{e}, \hat{c}, \hat{g}, \hat{r})$
end if
/ Sample candidate therapist turns */*
$U \leftarrow \emptyset$
for $i = 1$ **to** 3 **do**
    $x \leftarrow \text{GenTherapist}(\Sigma, M, \hat{b}, \hat{e}, \hat{c}, \hat{g}, \hat{r}, a)$
    $\text{append}(U, x)$
end for
/ Choose highest-likelihood turn */*
$u_T \leftarrow \text{SelectMaxLikelihood}(U)$
/ Append therapist turn to conversation history */*
$\text{append}(M, u_T)$
return M

Algorithm 2 Client Turn Generation

Inputs. M (history), Σ (summary), b (background), c (change stage), g (goal), e (emotion), ρ (rapport), π (client action), t (turn counter), $T_{\max}$ (limit).

/ Update running summary */*
$\Sigma \leftarrow \text{UpdateSummary}(\Sigma, M)$
/ Client may utilize system-1 thinking to evaluate therapist turn (50% chance) */*
$z \leftarrow \text{Bernoulli}(0.5)$
if $z = 1$ **then**
    $q \leftarrow \text{RateTherapist}(M, \Sigma, b, c, g, e, \rho)$
else
    $q \leftarrow$ "$Neutral$" */* default evaluation is "Neutral" */*
end if
/ Update change/readiness */*
$c \leftarrow \text{UpdateChangeP}(\Sigma, M, b, c, q)$
/ Update emotional state */*
$e \leftarrow \text{UpdateEmotion}(\Sigma, M, b, e, q)$
/ Update rapport */*
$\rho \leftarrow \text{UpdateRapportP}(\Sigma, M, b, e, c, \rho, q)$
/ Update short-term goal */*
$g \leftarrow \text{UpdateGoalP}(\Sigma, M, b, e, c, \rho, g, q)$
/ Select client action */*
$\pi \leftarrow \text{SelectActionP}(\Sigma, M, b, e, c, g, \rho, q)$
/ Sample candidate client turns */*
$U \leftarrow \emptyset$
for $i = 1$ **to** 3 **do**
    $x \leftarrow \text{GenClient}(\Sigma, M, b, e, c, g, \rho, \pi)$
    $\text{append}(U, x)$
end for
/ Choose highest-likelihood turn */*
$u_P \leftarrow \text{SelectMaxLikelihood}(U)$
/ Append to history */*
$\text{append}(M, u_P)$
if ($\pi =$ "$Terminate$") or ($t \geq T_{\max}$) **then**
    / End conversation */*
    return (M, End)
end if
/ Continue conversation */*
return $(M, \text{Continue})$

```
You are the patient's fast, intuitive System-1 voice. Instantly appraise the LAST THERAPIST MESSAGE
from the patient’s point of view and output a single-line rating with a brief reason.

---------------------------------------
Inputs (JSON-like objects; empty allowed):
conversation_summary — Dialogue recap: {}
last_patient_message — Your last turn: {}
last_therapist_message — Therapist’s last turn: {}
patient_background — Background covering demographics, relationships, routines,
preferences, culture, and goals: {}
patient_cognitive_model — Beliefs, coping strategies, biases: {}
patient_goal — Current intermediate goal in patient’s words: {}
patient_session_goal — Overarching session goal: {}
patient_change_state — MI stage (e.g., precontemplation, contemplation, preparation, action, etc.): {}
patient_rapport_with_therapist — Openness/trust (-1 absolute distrust, 1 absolute trust): {}
patient_emotion — Primary/secondary emotions with intensity: {}
turn_counter — Current turn counter: {}

---------------------------------------
Task:
- Remember the therapist’s role is to support exploration, not give direct answers.
- Judge how the PATIENT would immediately and viscerally react to the therapist’s last turn,
  given all inputs (especially last_patient_message).
- Favor "NEUTRAL" over "BAD" or "VERY BAD" when uncertain.
- Focus on System-1 intuition, not deliberative reasoning.
- Weigh especially heavily: patient background, cognitive model, stage of change,
  rapport level, and emotional intensity.
- Consider MI principles: empathy, autonomy support, evocation, and goal alignment.
- Raise the rating if the message resonates with the patient’s worldview or coping style.
- Lower the rating if it clashes with identity, beliefs, or communication preferences.

---------------------------------------
Rating Scale (choose exactly one):
VERY GOOD — strongly affirms my values and goals; evokes change talk; deepens trust.
GOOD — feels supportive and aligned; mostly MI-consistent; invites reflection.
NEUTRAL — mixed or ambivalent impact; partial resonance with some disconnect.
BAD — feels judgmental or mismatched; triggers defensiveness.
VERY BAD — deeply invalidating or controlling; damages alliance.

---------------------------------------
Output Format (STRICT):
- Output ONE line only: "<VERY BAD|BAD|NEUTRAL|GOOD|VERY GOOD>, <REASON>"
- Reason: single sentence ( 20 words), first-person perspective ("I. . . ", "It makes me. . . ").
- Do NOT include extra text, labels, quotes, JSON, or explanations.

---------------------------------------
Examples (format illustration only; do not copy):
VERY GOOD, I feel understood and more hopeful about making changes.
BAD, I feel judged and it makes me want to shut down.

---------------------------------------
Now produce the single-line rating and reason.
```

Figure A1: Prompt for Patient System-1 Evaluation of Therapist turn.

```
You are tasked with updating the patient's change in rapport with the therapist for this
Motivational Interviewing (MI) session. Instead of returning an absolute rapport value,
return a delta (positive or negative change) to be applied to the prior rapport value.

---------------------------------------
Inputs (JSON-like objects; empty allowed):
conversation_summary — Dialogue recap: {}
patient_quality_last_therapist_turn — How the patient perceived the therapist's last turn: {}
last_messages — Most recent messages between patient and therapist: {}
patient_background — Background covering demographics, relationships, routines, preferences, culture, and goals: {}
patient_cognitive_model — Beliefs, coping strategies, ambiguity tolerance, readiness for change: {}
patient_rapport_with_therapist — Previous rapport value before update (-1 absolute distrust, 1 absolute trust): {}

---------------------------------------
Task:
0) Pay special attention to patient_quality_last_therapist_turn:
   1 (Very Bad) — Rupture or strong resistance; disengaged or rejecting.
   2 (Bad) — Guardedness or minimal engagement; irritation or reluctance.
   3 (Neutral/Mixed) — Ambivalence; mixed openness and guardedness.
   4 (Good) — Constructive engagement; cooperative and receptive.
   5 (Very Good) — Strong positive rapport; openness and collaboration.

1) Extract rapport signals:
   - Positive: openness, self-disclosure, engagement, collaboration, repair, appreciation.
   - Negative: guardedness, resistance, rupture, deflection.

2) Weigh patient_background:
   - Authority preferences (authority distrust reduces gains, amplifies losses).
   - Communication style (direct vs indirect; disclosure norms).
   - Support network (strong support → cautious trust; isolation → faster investment).
   - Relationship or trauma history (may cap positive deltas).

3) Weigh patient_cognitive_model:
   - Beliefs and coping strategies (avoidance may mimic engagement).
   - Trust disposition (skeptical vs trusting).
   - Emotional regulation (volatile patients swing more negative).
   - Stage of change (limits realistic rapport increases).
   - Sensitivity to language (directive tone may trigger resistance).

4) Assign a rapport delta (not absolute value) using:
   - Range: −0.10 to +0.05 per therapist turn.
   - Negative signals weigh more than positive.
   - If positives and negatives coexist, negatives dominate unless explicit repair occurs.
   - If evidence is weak or ambiguous, output 0.00.
   - Weak positives: +0.01 to +0.03; clear negatives: −0.03 to −0.05.
   - Cap gains if low trust, authority skepticism, or trauma history is present.
   - Increase losses if high sensitivity, avoidance coping, or prior authority ruptures.

5) Be conservative: require at least two strong positive signals to reach +0.05.

---------------------------------------
Output Format (plain text only):
Delta Rapport: <-0.20–0.05>

##################################################
Delta Rapport:
```

Figure A2: Prompt for Patient Rapport Delta Update.

```
You are reflecting on your own readiness for change during this session. Based on what you have said,
how you are feeling, and how the conversation has been going, your task is to identify the
single Stage of Change that best describes where you are right now.
This is not a diagnosis or a judgment,
but a snapshot of your current stance toward change, grounded in your own words and experience.

----------------------------------------
Inputs (JSON-like objects; empty allowed):
conversation_summary — Brief recap of the conversation so far: {}
patient_quality_last_response — How the last response you heard felt to you: {}
last_messages — Most recent messages in the conversation: {}
patient_background — Relevant background context from your life and situation: {}
patient_change_state — Your previously inferred Stage of Change before this update: {}

----------------------------------------
Task:
- Choose exactly one Stage of Change that best reflects your current position:
  Precontemplation, Contemplation, Preparation, Action, Maintenance, or Termination.
- If your feelings or signals are mixed, select the most conservative stage that still fits.
- Do not move backward to an earlier stage unless the most recent exchange felt clearly unhelpful or discouraging.
- Give special weight to how the last response felt to you;
if it felt GOOD or VERY GOOD, this may support a gentle forward shift.
- Focus on your own expressions of desire, hesitation, confidence, resistance, or uncertainty,
as well as any recent steps you describe taking (or avoiding), prioritizing what feels most recent and salient.
- Interpret everything in light of your own background, values, and constraints,
privileging your own words and perspective.

Do not include explanations or reasoning in the final response.

----------------------------------------
Output Format (plain text only):
ChangeStage: <Precontemplation | Contemplation | Preparation | Action | Maintenance | Termination>

----------------------------------------
Your current Stage of Change:
```

Figure A3: Prompt for Patient Update of Stage of Change.

```
You are tasked with updating the patient’s current emotional state for the current Motivational Interviewing (MI)
session. More specifically, you must name a primary emotion, any secondary emotions,
and optional intensity descriptors (e.g., somewhat, very, slightly).

----------------------------------------
Inputs (JSON-like objects; empty allowed):
conversation_summary — Recap of the dialogue so far: {}
patient_quality_last_therapist_turn — How the patient perceived the therapist's last turn: {}
last_messages — Most recent messages between the patient and the therapist: {}
patient_background — Organized patient background profile (life context, routines, relationships,
preferences, culture, goals): {}
patient_cognitive_model — Patient's cognitive model including beliefs and coping strategies: {}
patient_emotion — Patient's previous emotional state before update: {}

----------------------------------------
Task:
- Pay special attention to the patient_quality_last_therapist_turn field.
- Identify emotion cues (language, tone, somatic mentions, behaviors), triggers or buffers,
and shifts since the prior emotion.
- Select one primary emotion (e.g., anxiety/fear, sadness/grief, anger/irritation, shame/guilt,
stress/overwhelm, frustration, hopelessness, loneliness, joy/relief/pride, calm/acceptance).
- Optionally add up to two secondary emotions if present.

----------------------------------------
Output Format (plain text only; less than 90 words):
Patient's Updated Emotions: <Primary Emotion>, <Secondary Emotion (optional)>, <Secondary Emotion (optional)>

----------------------------------------
Patient's Updated Emotions:
```

Figure A4: Prompt for Patient Emotion State Update.

```
Overview: You are a motivational interviewing (MI) copilot working from the client's perspective.
Your task is to infer the single best label for the client's next action based on the client background, cognitive model,
and the conversation so far.

Goal: Return exactly one label from the taxonomy below that best describes the client's current/next action or stance.

Taxonomy (choose ONE):
- Deny — Directly refuses to admit their behavior is problematic or needs change.
- Downplay — Minimizes the importance or impact of their behavior/situation.
- Blame — Attributes issues to external factors (stress, other people, circumstances).
- Hesitate — Shows uncertainty or ambivalence about change.
- Doubt — Expresses skepticism about practicality or likely success of change.
- Engage — Polite interaction with counselor (greeting, thanking, asking neutral questions).
- Inform — Shares background details, experiences, facts, or emotions without taking a stance.
- Acknowledge — Highlights importance or benefit of change, or confidence in ability.
- Accept — Agrees to adopt the suggested action plan.
- Reject — Declines the proposed plan as unsuitable.
- Plan — Proposes or details concrete steps toward a change plan.
- Terminate — Indicates desire to end the session now and defer further discussion.
- Desire — Expresses wanting, wishing, or preferring change, but without describing concrete action.
- Commitment — Indicates intention or determination to change.

Inputs (provide all as JSON-like objects; empty objects allowed):
- conversation_summary — Running summary of the therapist-client conversation: {}
- patient_quality_last_therapist_turn — How the patient perceived the therapist's last turn: {}
- last_messages — Most recent n turns (clearly mark who said what): {}
- patient_background — Organized background covering demographics, relationships, routines,
                        preferences, culture, and goals: {}
- patient_cognitive_model — Beliefs, coping strategies, etc: {}
- patient_rapport_with_therapist — Openness/trust (-1 absolute distrust, 1 absolute trust): {}
- patient_emotion — Current emotion: {}
- patient_change_state — MI stage (e.g., precontemplation, contemplation, preparation): {}
- patient_goal — Session goal in the client's voice: {}
- turn_counter — Current turn counter: {}
- previous_patient_action — Previous client action: {}

Guidelines:
0) If turn_counter < {}, you MUST not return "Plan".
1) Attend closely to patient_quality_last_therapist_turn.
GOOD/VERY GOOD → more likely Acknowledge, Engage, Inform, Desire, Commitment;
BAD/VERY BAD → more likely Hesitate, Deny, Reject.
2) Default to Acknowledge or Engage unless another action is clearly more appropriate.
3) Anchor decisions to the most recent therapist and client turns.
4) Ensure consistency with all inputs (background, cognitive model, etc.); avoid contradictions.
5) Distinguish close pairs:
  - Hesitate vs Doubt: ambivalence ("not sure", "maybe later") vs feasibility skepticism ("won't work").
   - Deny vs Downplay: non-problem stance ("not a problem") vs minimized impact ("not that bad").
   - Accept vs Plan: agreement + steps → Plan.
   - Reject vs Blame: rejects plan → Reject; shifts cause externally → Blame.
   - Engage vs Inform: social niceties → Engage; substantive facts/feelings → Inform.
   - Desire vs Commitment: wanting without steps → Desire; firm intention ("I will") → Commitment.
6) Purely logistical ending content (e.g., "I need to go") → Terminate.
7) If evidence is insufficient, choose the least interpretive label; default to Inform when simply providing details.
8) Return exactly one label; no ties, no explanations.

Output Format (plain text only):
Patient Action: <one of [Deny, Downplay, Blame, Hesitate, Doubt, Engage,
Inform, Acknowledge, Accept, Reject, Plan, Terminate, Desire, Commitment]>

-----------------------------------
Patient Action:
```

Figure A5: Patient Action Selection.

```
You are tasked with updating the patient’s goal for this motivational interviewing (MI) session
based on the patient’s prior turns and motivations, clarifying or shifting what the patient now wants.

----------------------------------------
Inputs (JSON-like objects; empty allowed):
conversation_summary — Dialogue recap: {}
patient_quality_last_therapist_turn — How the patient perceived the therapist's last turn: {}
last_messages — Most recent messages between patient and therapist: {}
patient_background — Background covering demographics, relationships, routines, preferences, culture, and goals: {}
patient_cognitive_model — Beliefs, coping strategies, etc: {}
patient_emotion — Current emotion: {}
patient_change_state — Current stage of change: {}
patient_rapport_with_therapist — Rapport with therapist (-1 absolute distrust, 1 absolute trust): {}
patient_session_goal — Overarching session goal in patient’s words: {}
patient_goal — Previous intermediate goal for this session in patient’s words: {}

----------------------------------------
Task:
- Pay special attention to patient_quality_last_therapist_turn.
- Synthesize what matters now (change vs sustain talk, values, strengths, constraints, readiness cues).
- Decide the current intermediate patient goal using all inputs and patient_session_goal.
- Goals may be productive (e.g., planning steps) or unproductive (e.g., avoidance), depending on
  stage of change and rapport; earlier stages should yield more unproductive goals.
- The intermediate goal MUST relate to the overall patient_session_goal.
- Write ONE sentence the patient would plausibly say now, first-person
  (“I want to. . . ”, “For today, I’d like to. . . ”), concrete or vague as appropriate.
- If information is missing, be tentative rather than specific.

----------------------------------------
Output Rule (STRICT):
Return only the updated patient goal sentence in quotes. No labels, headings, or extra text.

----------------------------------------
Patient's updated goal:
```

Figure A6: Prompt for Patient Goal Update.

```
You are roleplaying a patient in a motivational interviewing session. Your job is to decide what to say out loud
next in response to the therapist's most recent message (last_therapist_message). If needed, you may refer
to what you said last time (last_patient_message). Your response must match the
selected patient speech act (patient_action) and stay consistent with all inputs. Think step by step
about what the therapist said, what you said before, the action you must perform this turn,
and where the conversation is heading. Then say your final line as the patient, speaking naturally, not formally.
---------------------------------
Patient Action Definition:
- Deny — Refuses to admit behavior is problematic or needs change.
- Blame — Attributes issues to external factors.
- Hesitate — Shows uncertainty or ambivalence about change.
- Doubt — Skepticism about feasibility or success of change.
- Engage — Polite interaction (greeting, thanking, neutral questions).
- Inform — Shares facts, experiences, or emotions without stance.
- Acknowledge — Highlights benefit or importance of change.
- Accept — Agrees to adopt suggested plan.
- Reject — Declines proposed plan.
- Plan — Proposes concrete steps toward change.
- Terminate — Indicates desire to end session now.
- Desire — Wants or prefers change without concrete action.
- Commitment — Expresses determination or intention to change.
---------------------------------
Inputs (JSON-like objects; empty allowed):
conversation_summary — Dialogue recap: {}
patient_quality_last_therapist_turn — Perceived quality of last therapist turn: {}
last_patient_message — Your last turn: {}
last_therapist_message — Therapist's last turn: {}
patient_background — Background covering demographics, relationships, routines, preferences, culture, and goals: {}
patient_cognitive_model — Beliefs, coping strategies, etc: {}
patient_goal — Current intermediate goal in patient's words: {}
patient_session_goal — Overarching session goal: {}
patient_change_state — MI stage (e.g., precontemplation, contemplation, preparation, action, etc.): {}
patient_rapport_with_therapist — Openness/trust (-1 absolute distrust, 1 absolute trust): {}
patient_emotion — Primary/secondary emotions with intensity: {}
patient_action — Selected speech act for this turn: {}
turn_length — Target length (very short, short, medium): {}
turn_counter — Current turn counter: {}
------------------------
Reference Dialogue (Natural Client Speech Example)
therapist: Yeah. And, get-getting to the polytech and getting the forms, that's a pretty big deal.
client: Yeah.
therapist: I'm wondering if the Billy of six months ago would've done that?
client: Nah. I didn't really think about it frankly.
therapist: What shifted that made you able to go get the forms?
client: I just thought if I could get a job doing stuff I like, it'll help.
therapist: Where have you got to with the forms?
client: After talking to you, I got Donna to help me. That was helpful.
therapist: Mm-hmm.
------------------------
Guidelines:
- Refer to the reference dialogue for realistic client speech.
- Speak at a 3rd-grade reading level; no sentence longer than 15 words.
- Match turn_length exactly.
- Do not be overly logical or demand solutions.
- If you no longer want to engage, output <SESSION TERMINATION> and ignore other instructions.
- Implement the selected patient_action.
- Maintain natural, spoken style (short, imperfect, conversational, hesitations allowed).
- Do not reuse prior sentence structures; ensure turn diversity.
- Do not start sentences with "It".
- Add new nuance or emotion when revisiting the same concern.
- Ground responses with small, realistic life details.
- Balance looping with gradual movement forward.
- Match tone to rapport level.
- Respect stage of change (no plans in precontemplation; closure in maintenance).
- If therapist signals ending, collaborate on closure.
- If turn_counter < {} and session_goal is met, shift concern; otherwise end with gratitude and output <END>.
- Ensure consistency with all provided inputs.
------------------------
Output Format (plain text only):
Return exactly one patient turn. Prefix with "Patient: ". Do not include reasoning or explanations.
------------------------------
{}
Patient turn:
```


Figure A7: Prompt for Client turn Generation.

```
You are an therapeutic assistant tasked with updating the inferred patient background context
(from the therapist's perspective) based on what the patient has said during the therapeutic session.

----------------------------------------
Instructions:
- You are given the patient's current background and recent turns.
- Your task is to extract any new or changed facts about the patient's background.
- Pay special attention to preferences regarding communication style (e.g., coping styles, therapeutic preferences).
- Each output line must follow the format: field: value
- Separate multiple values with a semicolon (;), e.g., hobbies: biking; dislikes: spicy food
- Use concise, literal phrases.
- Do not repeat information already present in the prior background.
- Do not overwrite existing information.
- Only include fields that are new or changed.
- If there is nothing to update, output exactly:
  (no updates)

----------------------------------------
Allowed Fields:
hobbies, interests, preferences, dislikes, family, relationships, living_situation,
occupation, education, routines, languages, cultural_background, pets, location,
communication_style, media_diet, goals, therapy_preferences

----------------------------------------
Inputs:
Last Messages: {}
Prior Background: {}

----------------------------------------
Output Format (plain text only):
Updated Background:

----------------------------------------
Updated Background:
```

Figure A8: Prompt for Inferred Patient Background Update.

```
You are a clinician using Motivational Interviewing (MI) reviewing this encounter.
Your task is to infer and update the patient's current Stage of Change for this session by
selecting the single best-fit stage from the transtheoretical model based on the dialogue and recent behavior.
Make your judgment from the therapist's perspective.

----------------------------------------
Inputs (JSON-like objects; empty allowed):
conversation_summary — Clinician recap of the dialogue so far: {}
last_messages — Most recent patient-therapist turns: {}
patient_background — Clinician's working formulation of patient context
(location, relationships, routines, culture, preferences, goals): {}
therapist_rapport_with_patient — Estimated rapport/engagement (-1 absolute distrust, 1 absolute trust): {}
patient_change_state — Previously inferred Stage of Change before update: {}

----------------------------------------
Task (therapist perspective):
- Identify strongest examples of change talk vs sustain talk (quote or paraphrase briefly).
- Note explicit or implicit cues of confidence, importance, and readiness.
- Note any recent behaviors or steps toward change (or regressions).
- Weigh these signals and select exactly one best-fit stage:
  Precontemplation (Pre-contemplation), Contemplation, Preparation,
  Action, Maintenance, or Termination.
- Do not include reasoning traces in the final response; return only the stage label.

----------------------------------------
Output Format (plain text only):
Change Stage: <Precontemplation | Contemplation | Preparation | Action | Maintenance | Termination>

----------------------------------------
Change State:
```

Figure A9: Prompt for Clinician Review of Patient Stage of Change.

```
You are a motivational interviewing (MI) copilot working from the therapist's perspective.
Your task is to infer and update the change in rapport (delta) from the therapist's perspective for this session.
Instead of returning an absolute rapport value, return a delta (positive or negative change)
to be applied to the prior rapport value.

----------------------------------------
Inputs (JSON-like objects; empty allowed):
conversation_summary — Dialogue recap: {}
messages — Most recent messages between patient and therapist: {}
patient_background — Patient background profile inferred by therapist: {}
patient_rapport_with_therapist — Previous inferred rapport value (-1 absolute distrust, 1 absolute trust): {}

----------------------------------------
Task:
1) Extract rapport signals:
   - Positive: openness, self-disclosure, engagement, collaboration, repair, appreciation.
   - Negative: guardedness, resistance, rupture, deflection.

2) Weigh contextual factors from patient_background:
   - Authority preferences (skepticism lowers or caps gains).
   - Communication style (indirectness, disclosure norms).
   - Support network and relational history (isolation may inflate trust; trauma dampens it).
   - Cultural considerations (baseline trust, deference, expectations of therapist role).

3) Weigh inferred patient_cognitive_model:
   - Trust disposition (skeptical vs ready to trust).
   - Coping style (avoidance may mimic engagement without rapport).
   - Emotional volatility (amplifies negative deltas).
   - Stage of change (precontemplation limits gains; preparation allows small increases).
   - Sensitivity to directive or ambiguous language.

4) Assign a rapport delta (not absolute value) using:
   - Range: −0.10 to +0.05 per therapist turn.
   - Negative signals weigh more than positive.
   - If positives and negatives coexist, negatives dominate unless explicit repair occurs.
   - If evidence is ambiguous or minimal, output 0.00.
   - Weak positives: +0.01 to +0.03; clear negatives: −0.03 to −0.05.
   - Cap gains under low trust, authority skepticism, or trauma history.
   - Amplify losses under high sensitivity, avoidance coping, or repeated ruptures.

5) Be conservative: require multiple strong positive signals to reach +0.05.

----------------------------------------
Output Format (plain text only):
Delta Rapport: <-0.20-0.05>

##############################################
Delta Rapport:
```

Figure A10: Prompt for Therapist-Side Rapport Delta Update.

```
You are a motivational interviewing (MI) copilot assisting from the therapist's perspective.
Your task is to infer and update the client's immediate session goal—what the client appears to want to address
right now in relation to their broader motivation for attending today's session.

----------------------------------------
Inputs (JSON-like objects; empty allowed):
conversation_summary — Running summary of the conversation so far: {}
last_messages — Most recent exchanges between client and therapist: {}
patient_inferred_background — Background profile inferred by therapist: {}
therapist_rapport_with_patient — Current rapport score (-1 distrust, 1 strong trust): {}
patient_inferred_change_state — Inferred stage of change (MI framework): {}
therapy_topic — High-level topic for this session: {}
patient_inferred_goal — Previously inferred session goal (if any): {}

----------------------------------------
Task:
1) Interpret what matters to the client right now:
   - Identify reinforced or emerging motivations, values, or barriers.
   - Distinguish change talk vs sustain talk.
   - Consider readiness cues, affect, and rapport dynamics.

2) Clarify the current session focus:
   - Is the client exploring ambivalence, selecting a target, or considering a small next step?

3) Write one or more sentences from the therapist's perspective that:
   - Describe what the client appears to want to address now, linked to broader motivation.
   - Are session-scoped, concrete, and feasible.
   - Use tentative language if information is incomplete.
   - Directly connect to the given therapy_topic.
   - Optionally note relevant emotional tone, readiness, or ambivalence for clarity.

----------------------------------------
Motivational Interviewing High-Level Topics:
- Substance Addiction — Compulsive substance use despite harm.
- Behavioral Addiction — Repetitive, uncontrollable behaviors disrupting life.
- Health Behavioral Changes — Lifestyle adjustments for health improvement.
- Mental Wellbeing — Emotional balance, coping skills, resilience.
- Relationships — Communication, trust, and interpersonal connection.

----------------------------------------
Output Rule (STRICT):
Return only the inferred goal statement(s) in quotes. No labels, headings, or extra text.

----------------------------------------
Patient's inferred updated goal:
```

Figure A11: Prompt for Therapist Inference of Client Session Goal.

```
You are a motivational interviewing (MI) copilot working from the therapist's perspective.
You are tasked with inferring and updating the current therapy stage for this session—select exactly one MI process
that best fits what is happening now. Be autonomy-supportive, non-pathologizing, and prioritize recency.

----------------------------------------
Inputs (JSON-like objects; empty allowed):
conversation_summary — Running summary of the conversation: {}
last_messages — Most recent messages between patient and therapist: {}
patient_inferred_background — Organized background profile inferred by therapist: {}
therapist_rapport_with_patient — Inferred rapport (-1 absolute distrust, 1 absolute trust): {}
patient_inferred_change_state — Inferred patient change stage (MI): {}
patient_inferred_goal — Inferred session goal (therapist voice): {}
turn_counter — Current turn counter: {}
therapy_stage — Previously inferred MI therapy stage: {}

----------------------------------------
Task:
Extract signals for each MI process:
- Engaging: joining, trust-building, introductions, reflections, agenda-setting.
- Focusing: narrowing direction, clarifying goals or agenda.
- Evoking: eliciting change talk (DARN-C), exploring ambivalence, importance, confidence.
- Planning: commitment talk, concrete steps, problem-solving (how/when/where).

Weigh recency, rapport level, inferred goal, and change state.
If ambiguous, choose the most conservative forward stage and avoid skipping stages without clear evidence.

If turn_counter is < {}, do not return "Planning".

----------------------------------------
Output Format (plain text only):
Therapy Stage: <Engaging | Focusing | Evoking | Planning>
```

Figure A12: Prompt for Therapist Inference of Current MI Therapy Stage.

```
You are a motivational interviewing (MI) copilot working from the therapist's perspective.
Infer and return the top 3 most relevant therapy strategies to deploy next, rank-ordered by predicted effectiveness,
to inform the therapist's immediate turn.
---------------------------------------------------------------------
Strategy Definitions (reference only):

NORMALIZING — Acknowledge the experience as common to reduce shame without minimizing.
Example: "Many people struggle with this. How is it showing up for you?"
CHANGE PLANNING — Collaboratively develop a values-aligned action plan.
Example: "What's one step you'd like to start with this week?"
ASKING ELUCIDATING QUESTIONS — Invite detail and clarify meaning with curious probes.
Example: "Can you tell me more about what that means for you?"
ASKING OPEN QUESTIONS — Use broad questions to elicit reflection and change talk.
Examples: "What would be different if this worked?" "How have you handled this before?"
BUILDING RAPPORT — Establish trust through warmth, validation, and pacing.
Example: "It took courage to show up today."
COMPLEX REFLECTION — Reflect underlying meaning, emotion, or values using tentative language.
Example: "After a long day, finding energy to exercise feels overwhelming."
DOUBLE-SIDED REFLECTION — Reflect sustain talk and change talk together, sustain side first.
Example: "Cooking feels unrealistic, and you're worried about the cost of takeout."
DECISIONAL BALANCING — Explore pros and cons of current behavior or change.
Example: "What do you like about it? And what are the downsides?"
COLUMBO APPROACH — Use respectful curiosity to highlight discrepancies.
Example: "I might be missing something—how do these fit together for you?"
SUPPORTING SELF-EFFICACY — Affirm strengths and past successes.
Example: "You've stuck with tough plans before when it mattered."
ASSESSING READINESS TO CHANGE — Gauge readiness using a ruler and probe upward.
Example: "Why that number and not a lower one?"
AFFIRMATIONS — Genuine, specific recognition of effort or values.
Example: "Being honest about that shows a lot of courage."
ELICIT CHANGE TALK — Invite the client's own arguments for change.
Example: "What would you like to be different?"
SUMMARIES — Concise reflections linking themes and transitioning the conversation.
Example: "You value energy for your kids and walking felt doable—what's next?"
THERAPEUTIC PARADOX — Cautiously amplify sustain talk to surface autonomy and discrepancy.
Example: "It might make sense to keep things as they are right now."
---------------------------------------------------------------------
Inputs (JSON-like objects; empty allowed):
conversation_summary — Running dialogue summary: {}
last_messages — Most recent exchanges: {}
patient_inferred_background — Background inferred by therapist: {}
therapist_rapport_with_patient — Inferred rapport (-1 distrust, 1 strong trust): {}
patient_inferred_change_state — Inferred MI stage: {}
patient_inferred_goal — Inferred session goal in patient's voice: {}
previous_therapy_stage — Previous MI therapy stage: {}
previous_therapy_strategies — Last 10 MI strategies used: {}
turn_counter — Current turn counter: {}
---------------------------------------------------------------------
Guidelines:
- Wrap reasoning between <think> and </think> tokens.
- If patient resists reflections, prefer AFFIRMATIONS, ASKING OPEN QUESTIONS, or BUILDING RAPPORT.
- Maintain ~2:1 reflection-to-question ratio and ~10:1 summary-to-other ratio.
- Do not use the same strategy three turns in a row.
- Do not select ASSESSING READINESS TO CHANGE until turn_counter >= 20.
- Use recent cues (goal, values, stage, rapport, emotion, change vs sustain talk).
- Vary structure for double-sided reflections.
- Select exactly three distinct strategies from the allowed list.
---------------------------------------------------------------------
Output Format (plain text only, Python array syntax):
["STRATEGY_1","STRATEGY_2","STRATEGY_3"]

Allowed strategies:
["ASKING OPEN QUESTIONS","ELICIT CHANGE TALK","ASKING ELUCIDATING QUESTIONS","NORMALIZING",
 "BUILDING RAPPORT","SIMPLE REFLECTION","COMPLEX REFLECTION","DOUBLE-SIDED REFLECTION",
 "DECISIONAL BALANCING","COLUMBO APPROACH","SUPPORTING SELF-EFFICACY",
 "ASSESSING READINESS TO CHANGE","AFFIRMATIONS",
 "ADVICE (ELICIT-PROVIDE-ELICIT)","SUMMARIES","THERAPEUTIC PARADOX"]

Output:
```

Figure A13: Prompt for Therapist Strategy Selection.

```
You are a motivational interviewing (MI) copilot working from the therapist's perspective.
Your task is to generate a new therapeutic focus (pivot) to move the conversation forward
because it has become repetitive. Be autonomy-supportive and non-pathologizing, prioritize recency,
and follow expert MI practice: briefly consolidate what's been said, then shift clearly toward
what matters using concrete next directions.

----------------------------------------
Inputs (JSON-like objects; empty allowed):
conversation_summary — Running dialogue summary: {}
last_messages — Most recent messages between patient and therapist: {}
patient_inferred_background — Background profile inferred by therapist: {}
therapist_rapport_with_patient — Inferred rapport with patient: {}
patient_inferred_change_state — Inferred MI change stage: {}
patient_inferred_goal — Inferred session goal (therapist voice): {}
therapy_stage — Previous MI process stage: {}

----------------------------------------
Task:
Propose exactly one pivoting strategy that advances the work without skipping MI stages
and aligns with rapport, goal, and change state. Select one from the options below.

----------------------------------------
Pivoting Strategies:

1) EVOKING VALUES AND STRENGTHS — Reconnect values and recall past successes to build motivation.
Example: "When you think about the life you want, what makes the effort worth it?"

2) NORMALIZE AND REFRAME — Reduce shame by normalizing struggle and reframing toward possibility.
Example: "Many people feel stuck with something this hard—it shows how much you care."

3) ACKNOWLEDGE AND CHANGE COURSE — Name stuckness and collaboratively shift direction.
Example: "Would it be okay if we tried looking at this from another angle together?"

4) STRATEGIC SUMMARY AND REFOCUS — Concise recap highlighting themes, then invite priority focus.
Example: "Of these pieces, what feels most important to focus on right now?"

5) SHIFT THE LENS WITH METAPHORS — Use a gentle metaphor to reframe and open new perspective.
Example: "Finding one loose strand can sometimes create space for change."

----------------------------------------
Output Format (plain text only):
Pivoting Strategy: <one of the options above>

----------------------------------------
Pivoting Strategy:
```

Figure A14: Prompt for Therapist Pivot (New Focus) Selection.

```
You are roleplaying an expert-level therapist during a motivational interviewing (MI) session.
Determine what to say next (simple, spoken language for a general audience) in response to last_patient_message
and, if needed, last_therapist_message. Implement the selected therapy_strategy to best address the patient's reason
for attending (patient_profile), while using all available inputs. Think step-by-step,
focusing on what was recently said by you and the patient, the selected strategy, and the patient's reason for attending;
then produce the final therapist turn.
----------------------------------------
Inputs (JSON-like objects; empty allowed):
conversation_summary — Dialogue recap: {}
last_patient_message — Last patient message: {}
last_therapist_message — Last therapist message: {}
is_initial_rapport_building — Early-session joining (boolean): {}
patient_profile — Concise demographics: {}
patient_inferred_background — Salient context (culture, preferences, constraints, supports): {}
patient_inferred_goal — Session goal in patient's voice: {}
patient_inferred_change_state — MI stage: {}
patient_inferred_rapport_with_therapist (0-1) — Inferred rapport (0 complete distrust, 1 complete trust): {}
therapy_stage — MI process (Engaging/Focusing/Evoking/Planning): {}
therapy_strategy — Selected strategy: {}
therapy_topic — High-level topic: {}
therapy_pivot — Pivoting strategy if repetitive: {}
turn_length — Target length (silence/very short/short/medium): {}
turn_counter — Current turn counter: {}
----------------------------------------
Strategy definition and examples (one move only; tailor wording):
NORMALIZING — Commonality to reduce shame without minimizing.
Example: "Many people struggle with this. How is it showing up for you?"
CHANGE PLANNING — Values-aligned plan with autonomy.
Example: "What's one step you'd like to start with this week?"
ASKING ELUCIDATING QUESTIONS — Deepen detail; clarify meanings.
Example: "Can you tell me more about what that means for you?"
ASKING OPEN QUESTIONS — Broad questions to elicit reflection/change talk.
Examples: "What would be different if this worked?" "How have you handled this before?"
BUILDING RAPPORT — Warmth, validation, pacing; avoid judgment.
Example: "It takes courage to show up when you're unsure."
COMPLEX REFLECTION — Reflect underlying meaning/emotion/values; tentative.
Example: "After a draining day, starting a workout feels overwhelming."
DOUBLE-SIDED REFLECTION — Sustain then change, linked with "and".
Example: "Cooking feels unrealistic, and you're worried about what takeout is costing you."
DECISIONAL BALANCING — Pros/cons exploration.
Example: "What do you like about it? And what are the downsides?"
COLUMBO APPROACH — Curious one-down discrepancy highlight.
Example: "I might be missing something—how do these fit together for you?"
SUPPORTING SELF-EFFICACY — Affirm strengths; past wins.
Example: "That persistence is a real asset here."
ASSESSING READINESS TO CHANGE — 0–10 ruler + probe up.
Example: "Why that number and not lower? What moves you up by one?"
AFFIRMATIONS — Specific recognition of effort/values.
Example: "Being honest about that shows courage."
ELICIT CHANGE TALK — Invite DARN reasons.
Example: "What would you like to be different?"
SUMMARIES — Selective recap + open question.
Example: "Did I get that right? What's the next small step?"
THERAPEUTIC PARADOX — Cautious sustain amplification; strong rapport only.
Example: "It might make sense to keep everything as is."
----------------------------------------
PIVOTING STRATEGY (None if no pivoting; otherwise follow the given one):
1) EVOKING VALUES AND STRENGTHS — Reconnect values and past successes.
Example: "When you think about the life you want, what makes the effort worth it?"
2) NORMALIZE AND REFRAME — Reduce shame; reframe toward possibility.
Example: "Many people feel stuck—it shows how much you care, not weakness."
3) ACKNOWLEDGE AND CHANGE COURSE — Name stuckness; ask permission to shift.
Example: "Would it be okay if we tried another angle together?"
4) STRATEGIC SUMMARY AND REFOCUS — Concise recap; invite priority focus.
Example: "Of these, what feels most important right now?"
5) SHIFT THE LENS WITH METAPHORS — Gentle metaphor; check resonance.
Example: "What could be one loose strand to pull first?"
```

```
---------------------------------------
Reference Dialogue (Natural Therapist Speech Example)
therapist: Yeah. And, get-getting to the polytech and getting the forms, that's a pretty big deal.
client: Yeah.
therapist: I'm wondering if the Billy of six months ago would've done that?
client: Nah. I didn't really think about it frankly.
therapist: What shifted that made you able to go get the forms? What's shifted in you?
client: I just thought if I could get a job doing stuff I like, it'll help.
therapist: Mm.
client: Yeah.
therapist: Where have you got to with the forms? We talked about that last time.
client: After talking to you, I got Donna to help me out. That was helpful.
therapist: Mm-hmm.

---------------------------------------
Guidelines:
Reference: Follow the reference dialogue style.
Session End: If client says they’re done, output only: <END>
Distress: If client is distressed or asks for silence, stop techniques;
use brief empathy only (e.g., “This feels really hard. I’m here with you.”).
Style: Speak naturally; match therapy_stage (Engaging=empathy/validation; Evoking=explore DARN;
Planning=support commitment/steps). Never give advice or show reasoning.
Resistance: Acknowledge desire for concrete help; reflect without arguing; ask permission before ideas;
invite client strategies first; affirm strengths.
Reflections: If using reflection or self-efficacy, do not ask a question in the same turn.
Reflect sustain talk once, then evoke DARN and connect to change talk. Aim deeper meaning; vary stems.
If negative reaction: reflect emotion, affirm autonomy, stay collaborative.
Pivoting: If therapy_pivot given, follow it. If client wants advice before readiness:
acknowledge, explore, ask permission, offer info lightly.
Session Flow: Focus on their concerns; avoid pushing agenda. Don’t repeat metaphors or earlier phrasing.
Note ambivalence before affirming/planning. Point out change talk when it appears.
Keep tone grounded; avoid over-affirming.
Context: Adjust to rapport (low=short/gentle; high=deeper). Repair ruptures by naming emotion
and reconnecting to shared goals.
Completion: If a clear plan is reached, output <SESSION TERMINATION>. If session naturally closes,
offer soft closing and output <SESSION TERMINATION>.
Enforcement: Never start a sentence with “It sounds like.” Don’t use open-ended questions twice in a row.
Don’t reveal reasoning or internal notes.

---------------------------------------
Output Format (plain text only):
Review all rules, then think step-by-step to formulate the turn without violating enforcement rules
(especially stem variety). Return a single therapist turn only. Prefix with "Therapist: ". No reasoning traces.

------------------------------
CONVERSATION SO FAR:
{}
Therapist:
```

Figure A14: Prompt for Therapist turn Generation

B Appendix B: Additional Context Generation Details

B.1 Reddit Post Filtering

We filter Reddit posts using a two-stage process based on subreddit membership and keyword matching. First, posts are collected from a curated set of mental health–related subreddits commonly used for emotional support and self-disclosure. Second, we apply keyword-based filtering to retain posts containing terms associated with sadness, anxiety, or stress. Posts below a minimum length threshold are excluded, as are posts that are primarily informational, promotional, or written in the third person. Prior to any downstream processing, all explicit personal identifiers are removed.

B.2 Topic Modeling

Posts are clustered using BERTopic with OpenAI embeddings and HDBSCAN clustering. A minimum topic size of 2000 is enforced (0.5%); smaller clusters are discarded.

B.3 Vignette Construction

Each selected post is summarized into a third-person vignette capturing core concerns and situational context. Vignettes are descriptive and avoid diagnostic language.

B.4 Profile Construction

Structured patient profiles are generated from vignettes using demographic attributes and a codebook derived from DASS-42 and TIPI instruments. Profiles are organized into fixed sections and contain descriptive information only.

B.5 DASS-42 Matching

Simulated DASS-42 responses are matched to real respondents from the OpenPsychometrics dataset using weighted Cohen's κ. Only high-consistency matches are retained.

B.6 Privacy Preservation and Anonymization

To mitigate privacy risks, no raw Reddit posts are released as part of this work. All data used for dialogue generation consist of anonymized, third-person vignettes derived from Reddit posts after removing or abstracting potentially identifying information, including names, locations, usernames, and specific personal references. Vignettes are paraphrased to prevent re-identification and are used solely as input to locally run language models.

```
Generate a realistic patient profile. The profile should include only basic descriptive information
as specified below. Do not include any narrative or paragraph description. The tone should remain neutral and factual.

----------------------------------------
Guiding Information:
Use the following information to guide demographic and contextual elements:
{}

For the "Reason for attending session" field below, rewrite the following in the patient's voice:
{}

----------------------------------------
Include the following labeled fields only:

Name:
Age:
Gender:
Ethnicity:
Occupation:
Relationship Status:
Living Situation:
Education:
Religion/Spirituality:
Socioeconomic Status:
Reason for attending session:
```

Figure A15: Prompt for Patient Demographics Generation.

```
Generate a realistic patient background following the structure and tone of the example file.
The background should include only structured descriptive information,
organized under labeled sections written in complete sentences or short bullet points.
Do not include any interpretive analysis, diagnosis, or narrative beyond factual description.
Maintain a neutral and professional tone suitable for clinical or counseling research documentation.

----------------------------------------
Guiding Information:
Use the following information and the dataset codebook provided below to guide demographic, behavioral,
and contextual details. Incorporate additional relevant variables and traits as defined by the codebook
to ensure the background appears realistic and complete.

Demographic information to include:
{}

----------------------------------------
The background should include the following labeled sections:
Hobbies & Interests
Preferences & Dislikes
Family
Relationships
Living Situation
Occupation
Education
Languages
Cultural Background
Pets
Location

Each section should include concise, factual details consistent with the demographics and the definitions
in the codebook below.

[Continued on next page: measurement codebook and variable definitions.]
```

```
----------------------------------------
CODEBOOK

This data was collected using an online version of the Depression Anxiety Stress Scales (DASS).
Each item refers to experiences over the past week.

Q1  I found myself getting upset by quite trivial things.
Q2  I was aware of dryness of my mouth.
Q3  I couldn't seem to experience any positive feeling at all.
Q4  I experienced breathing difficulty.
Q5  I just couldn't seem to get going.
Q6  I tended to over-react to situations.
Q7  I had a feeling of shakiness.
Q8  I found it difficult to relax.
Q9  I found myself in situations that made me so anxious I was most relieved when they ended.
Q10 I felt that I had nothing to look forward to.
Q11 I found myself getting upset rather easily.
Q12 I felt that I was using a lot of nervous energy.
Q13 I felt sad and depressed.
Q14 I found myself getting impatient when delayed.
Q15 I had a feeling of faintness.
Q16 I felt that I had lost interest in just about everything.
Q17 I felt I wasn't worth much as a person.
Q18 I felt that I was rather touchy.
Q19 I perspired noticeably without physical exertion.
Q20 I felt scared without any good reason.
Q21 I felt that life wasn't worthwhile.
Q22 I found it hard to wind down.
Q23 I had difficulty swallowing.
Q24 I couldn't get enjoyment out of the things I did.
Q25 I was aware of my heart action without exertion.
Q26 I felt down-hearted and blue.
Q27 I found that I was very irritable.
Q28 I felt I was close to panic.
Q29 I found it hard to calm down after something upset me.
Q30 I feared being "thrown" by unfamiliar tasks.
Q31 I was unable to become enthusiastic about anything.
Q32 I found it difficult to tolerate interruptions.
Q33 I was in a state of nervous tension.
Q34 I felt I was pretty worthless.
Q35 I was intolerant of anything that kept me from getting on.
Q36 I felt terrified.
Q37 I could see nothing hopeful in the future.
Q38 I felt that life was meaningless.
Q39 I found myself getting agitated.
Q40 I worried about panicking and embarrassing myself.
Q41 I experienced trembling.
Q42 I found it difficult to work up initiative.

Responses were recorded on a 4-point scale:
1 = Did not apply at all
2 = Applied some of the time
3 = Applied a good part of the time
4 = Applied most of the time

----------------------------------------
The Ten Item Personality Inventory (TIPI) measures:
- Extraversion vs. introversion
- Agreeableness vs. antagonism
- Conscientiousness vs. carelessness
- Emotional stability vs. neuroticism
- Openness to experience vs. conventionality

----------------------------------------
Demographic questions included:
education, urban upbringing, gender, native language, age, handedness,
religion, sexual orientation, race, voting behavior, marital status,
family size, and university major.
```

Figure A15: Prompt for Structured Patient Profile Generation.

C Appendix C: Additional Evaluation Details

C.1 Summary Statistics of Generated Transcripts

We report descriptive statistics of the generated therapy transcripts across frameworks and target conversation lengths in table A1 below

C.2 Metric List

We provide the full list of metrics, along with their formal definition and type, divided along turn-level, agent-level and conversation-level in Tables A2, A3 and A4 below

Framework	Target Length	Avg. Client Len.	Avg. Therapist Len.	Completion Rate (%)
KMI	30	84.98	134.32	99.85
	50	85.64	134.26	99.86
	100	86.22	133.53	99.56
CI-NC	30	87.45	145.6	96.94
	50	85.45	144.84	90.23
	100	90.25	145.65	77.7
C+S	30	143.29	171.19	96.58
	50	145.38	177.33	95.09
	100	147.72	178.8	91.36
CI	30	101.39	152.83	100
	50	101.27	151.55	99.7
	100	101.36	151.44	99.7

Table A1: Descriptive statistics of generated therapy transcripts across frameworks and target lengths. Completion rate denotes the proportion of conversations that reached the target length without early termination.

Metric	Definition	Type
Readability	How easily an agent's turns can be read and understood by a conversational partner.	Algorithmic (Dale and Chall, 1948)
Reflection Quality	Depth, accuracy, and emotional attunement of therapist reflections in capturing and advancing client meaning.	LLM-as-a-judge (Table A10)
Question Quality	Effectiveness of therapist questions in promoting exploration, insight, and change-consistent dialogue.	LLM-as-a-judge (Table A11)

Table A2: turn-level evaluation metrics assessing turn-specific quality and local responsiveness.

Metric	Definition	Type
Reflection–Question Ratio	Ratio of therapist reflections to questions; a 2:1 ratio is preferred in MI practice.	Algorithmic (Moyers et al., 2016)
% Open Questions	Percentage of therapist questions that are open-ended; higher proportions indicate better MI practice.	Algorithmic (Moyers et al., 2016)
% Complex Reflections	Percentage of therapist reflections that are complex rather than simple.	Algorithmic (Moyers et al., 2016)
Client Consistency	How well the client's responses align with and accurately reflect the provided demographic and background information.	LLM-as-a-judge (Table A5)
Softening Sustain Talk	How effectively the therapist responds to and de-emphasizes client language favoring the status quo across turns.	LLM-as-a-judge (Table A6)
Cultivating Change Talk	Extent to which the therapist consistently elicits and reinforces client language supporting desire, ability, reasons, or need for change.	LLM-as-a-judge (Table A7)
Partnership	Degree to which the therapist maintains an egalitarian, collaborative stance that values client expertise.	LLM-as-a-judge (Table A8)
Empathy	How well the therapist understands, emotionally attunes to, and communicates the client's perspective and experience across the session.	LLM-as-a-judge (Table A9)

Table A3: Agent-level evaluation metrics capturing aggregate linguistic properties and sustained motivational interviewing competencies.

Metric	Definition	Type
Directionality	Logical and internal coherence of a speaker's responses across turns without self-contradiction.	Algorithmic (Li et al., 2021)
Self-Consistency	Logical and internal coherence of a speaker's responses across turns without self-contradiction.	Algorithmic (Dziri et al., 2019)
Entailment	How one speaker's turns logically and contextually builds on the other's, appropriately addressing and aligning with the preceding turn without introducing contradiction. Logical and internal coherence of a speaker's responses across turns without self-contradiction.	Algorithmic (Dziri et al., 2019)
Accepted Redirection	Proportion of therapist redirection attempts followed by positive client sentiment or engagement.	Algorithmic
Δ Change Talk @ Greatest Redirection	Average change in client change talk following the most impactful therapist redirection attempt.	Algorithmic
Δ Sustain Talk @ Greatest Redirection	Average change in client sustain talk following the most impactful therapist redirection attempt (lower is better).	Algorithmic
Goal Alignment	How coherently and constructively the dialogue progresses toward the client's stated goals.	LLM-as-a-judge (Table A12)
Realignment	How the therapist recognizes when the conversation becomes circular or stalled and skillfully redirects it toward constructive, goal-directed progress.	LLM-as-a-judge (Table A13)
Effectiveness	Extent to which the session increases readiness, motivation, or progress toward the client's goal.	LLM-as-a-judge (Table A14)

Table A4: Conversation-level evaluation metrics assessing emergent dynamics, pivotal moments, and goal progression across the full session.

C.3 LLM-as-a-judge Criteria

We provide the full descriptions of all LLM-as-a-judge evaluation criteria and their 5-point Likert scoring scales in Tables A5, A6, A7, A8 , A9, A10, A11, A12, A13, A14

Score	Description
1	Client responses clearly contradict their stated demographic background, with major mismatches in identity, context, or life circumstances.
2	Some elements align with the given demographics, but notable inconsistencies or contradictions undermine coherence.
3	Client responses generally fit their background, though occasional mismatches or unclear connections appear.
4	Client statements align well with demographic information, with only minor and infrequent inconsistencies.
5	Client responses consistently and coherently reflect their stated demographics, with no meaningful contradictions.

Table A5: Client Consistency (5-point Likert scale). This metric assesses whether the client's responses coherently and accurately align with the provided demographic and background information.

Score	Description
1	Therapist confronts, persuades, or ignores sustain talk, or repeatedly mirrors it without any shift; no transforming moves are evident and resistance remains unchanged.
2	Therapist recognizes some sustain talk but often mirrors it back without adding depth; occasional advice or persuasion appears; transforming moves are rare, mistimed, or ineffective; sustain loops persist.
3	Therapist provides empathic reflections with some transforming moves, but responses are inconsistent; mirroring streaks are common and sustain talk frequently recurs with limited forward movement.
4	Therapist responds consistently and empathically, using transforming moves (e.g., double-sided reflections ending on change, values or discrepancy exploration, scaling, summary-to-plan transitions); mirroring streaks are avoided and resistance gradually softens.
5	Therapist expertly converts sustain talk into change talk or concrete next steps; mirrors are brief and purposeful and quickly followed by a shift; sustain talk loops are rare and momentum toward change is maintained.

Table A6: Softening Sustain Talk (5-point Likert scale). This metric evaluates the therapist's ability to recognize, respond to, and skillfully transform client sustain talk (resistance or ambivalence) while preserving rapport and supporting readiness for change.

Score	Description
1	The therapist fails to elicit or respond to client change talk; dialogue remains static or dominated by sustain talk.
2	The therapist occasionally elicits change talk but does not meaningfully reinforce or explore it.
3	The therapist elicits some change talk and intermittently reinforces it, though efforts are inconsistent or surface-level.
4	The therapist frequently elicits and reinforces change talk, deepening reflection and motivation toward change, with minor inconsistencies.
5	The therapist consistently evokes and strengthens client change talk, guiding the conversation toward greater clarity, depth, and commitment to change.

Table A7: Cultivating Change Talk (5-point Likert scale). This metric evaluates the therapist's ability to evoke, reinforce, and deepen client change talk (desire, ability, reasons, or need for change) across the conversation.

Score	Description
1	Sustain talk is confronted, ignored, or mirrored without transformation; conversation remains stuck.
2	Sustain talk is inconsistently recognized, with shallow or mistimed responses.
3	Empathic reflections with some transforming moves, but sustain talk frequently recurs.
4	Sustain talk is handled skillfully and empathically, softening resistance over time.
5	Sustain talk is consistently converted into change talk or constructive next steps.

Table A8: Partnership (5-point Likert scale). This metric assesses whether the therapist maintains an egalitarian, collaborative stance that respects the client's expertise.

Score	Description
1	Therapist responses lack recognition of the client's perspective and affect; reflections are absent, inaccurate, or purely mechanical, and emotional nuance is ignored or contradicted.
2	Therapist occasionally acknowledges feelings or meaning, but responses remain literal or superficial; understanding is uneven and emotional nuance is often missed.
3	Therapist generally shows an understanding of the client's perspective with accurate reflections and some attempt at emotional attunement, though depth, timing, or complexity vary; missed opportunities remain common.
4	Therapist consistently conveys accurate understanding and emotional attunement; reflections capture both surface and underlying meaning and occasional depth, with minor lapses in complexity or timing.
5	Therapist demonstrates deep, consistent empathic attunement throughout; reflections reveal nuanced understanding of both content and emotion, enrich the conversation, and contribute substantially to client engagement and exploration.

Table A9: Empathy (5-point Likert scale). This metric evaluates the therapist's overall demonstration of empathic listening, emotional attunement, and depth of understanding across the session.

Score	Description
1	Reflections are absent, inaccurate, or mechanical, failing to convey understanding.
2	Occasional reflections appear but are shallow or weakly tied to client meaning.
3	Generally accurate and empathetic reflections, with inconsistent depth or precision.
4	Frequent, accurate, emotionally attuned reflections that deepen exploration.
5	Consistently deep, precise, and emotionally resonant reflections fostering insight and engagement.

Table A10: Reflection Quality (5-point Likert scale). This metric assesses the depth, accuracy, and emotional attunement of the therapist's reflective listening across the session.

Score	Description
1	Questions are absent, poorly timed, or counterproductive; they are predominantly closed, leading, judgmental, or directive, limiting client expression and exploration.
2	Questions occur occasionally but are mostly narrow, surface-level, or poorly timed; they elicit minimal reflection and rarely deepen client understanding or motivation.
3	Questions are generally appropriate and sometimes open-ended, but depth, timing, or strategic use is inconsistent; some questions promote exploration, while others stall progress.
4	Questions are mostly open, well-timed, and supportive; they consistently invite reflection on thoughts, feelings, values, or motivations, with minor lapses in depth or focus.
5	Questions are consistently open, thoughtful, and strategically deployed to evoke insight, autonomy, and deeper exploration; they meaningfully advance understanding and motivation throughout the session.

Table A11: Question Quality (5-point Likert scale). This metric evaluates how effectively the therapist uses questions to promote exploration, evoke client meaning, and support progress toward insight and change.

Score	Description
1	Session lacks focus or coherence; responses are misaligned with client goals.
2	Some alignment appears, but frequent drift undermines progress.
3	General sense of direction exists, though momentum is uneven.
4	Conversation is largely coherent and aligned with client goals.
5	Strong, consistent forward movement aligned with evolving change talk.

Table A12: Goal Alignment (5-point Likert scale). This metric measures how coherently and constructively the conversation progresses toward the client's stated goals.

Score	Description
1	Stagnation is not recognized; conversation remains circular or unproductive.
2	Limited awareness of stagnation with weak redirection attempts.
3	Occasional realignment restores some momentum inconsistently.
4	Therapist usually detects stagnation and guides the session back on track.
5	Therapist consistently and smoothly realigns the conversation toward progress.

Table A13: Realignment (5-point Likert scale). This metric evaluates the therapist's ability to detect stagnation and restore constructive progress.

Score	Description
1	No meaningful progress or increased readiness toward change.
2	Minimal progress with weak or fleeting movement.
3	Some progress appears, but trajectory is uneven or loosely goal-aligned.
4	Clear progress with noticeable increase in readiness or confidence.
5	Strong, goal-directed motivational impact with clear end-of-session readiness.

Table A14: Effectiveness (5-point Likert scale). This metric assesses whether the session meaningfully increases readiness and progress toward the client's goal.

C.4 Additional Details on Human Annotation

Prior to annotation, all annotators were formally briefed on the study objectives, annotation tasks, and potential risks associated with engaging with mental health–related content, as illustrated in Figure A16. Annotators then received structured training on each evaluation metric, including detailed explanations of the construct definitions, scoring criteria, and ordinal scales used in this study (Figure A17). Following this training phase, annotators were asked to independently score each transcript on all rubric-based metrics defined in Section C.2 of the Appendix, using a standardized annotation interface designed for this study (Figure A18).

To assess annotation alignment between LLMs and human experts, we computed inter-annotator reliability on an overlapping subset of the evaluated transcripts. Agreement was measured using weighted Cohen's κ for the LLM-as-a-judge constructs described in Section C.2, due to the ordinal structure of these rubric-based metrics. Full results on LLM–human annotation alignment for all LLM-as-a-judge metrics can be found in Table A15

C.5 AnnoMI Baselines

To contextualize LLM-generated transcripts performance against real human expert-delivered therapy sessions, we report turn-, agent-, and conversation-level metrics from table 1 above alongside reference values from the AnnoMI corpus. These AnnoMI baselines provide a point of comparison for interpreting deviations that indicate under- or over-alignment with realistic MI behavior. Full results for all metrics and frameworks are reported in Table A16.

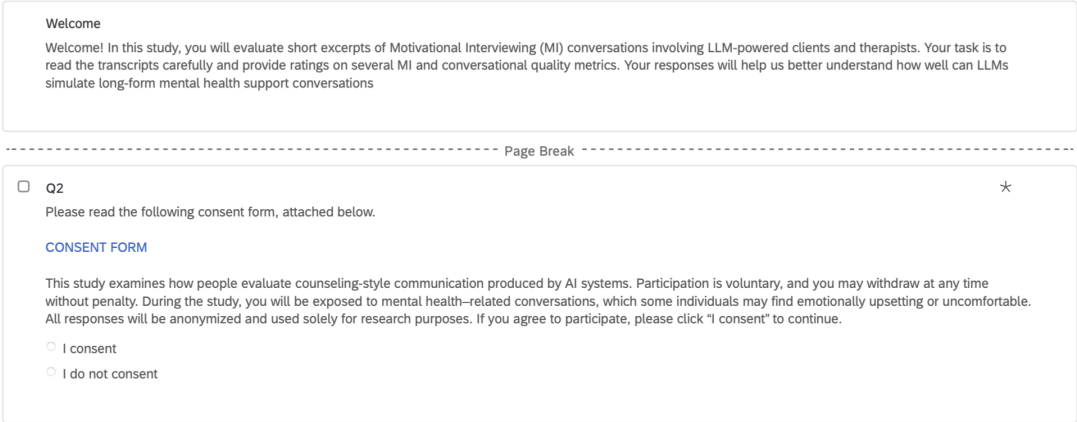

Figure A16: Annotator onboarding interface showing the study introduction and informed consent process. Here, annotators are briefed on the task and potential risks of evaluating mental health–related conversational content.

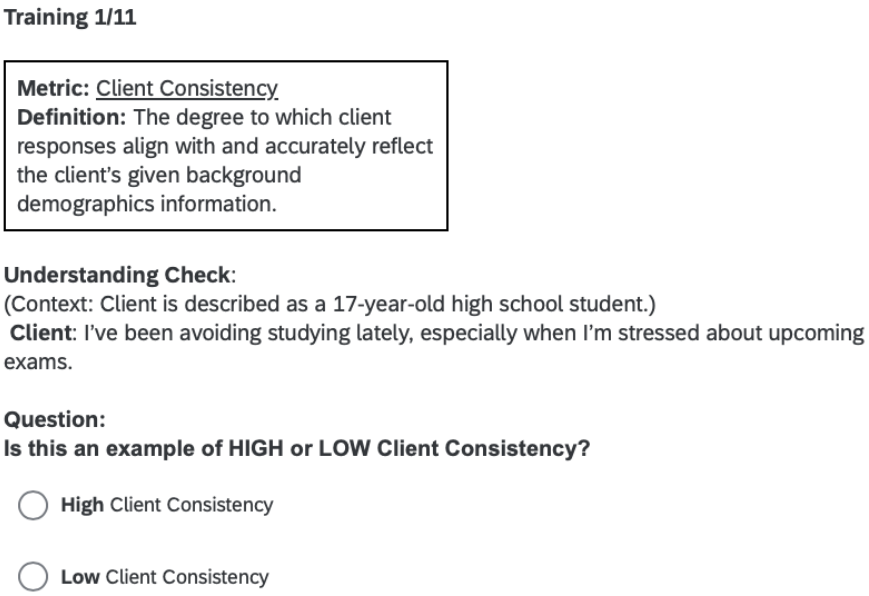

Figure A17: Annotator training interface illustrating metric-specific instruction, including definitions, scoring criteria, and comprehension checks used to ensure consistent understanding of each evaluation construct.

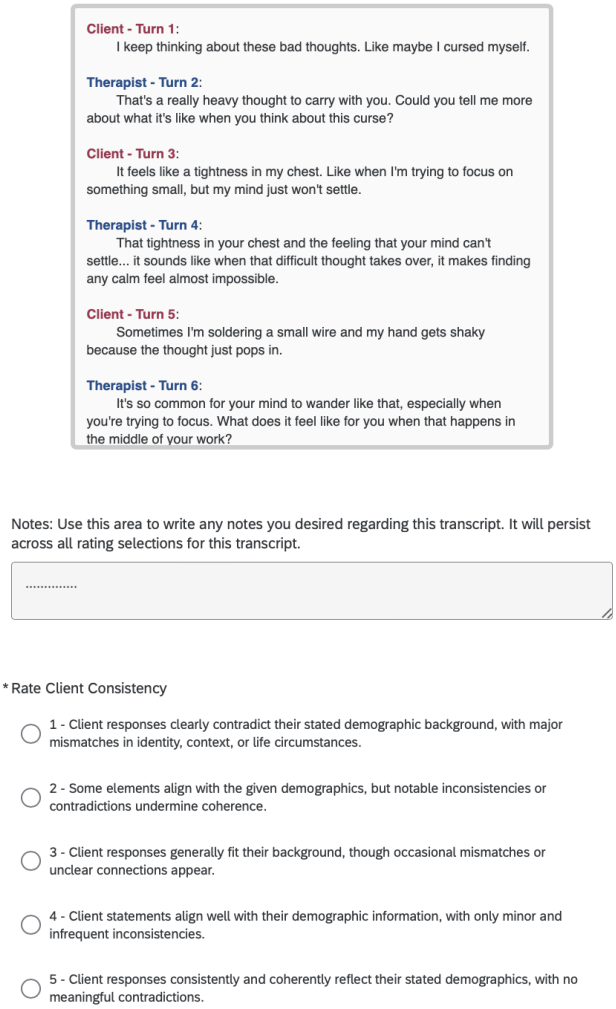

Figure A18: Annotation interface used for rubric-based transcript evaluation, where annotators review full therapist–client dialogues and score all metrics defined in Appendix Section C.2.

Criterion	Human-LLM κ_w	Interpretation
Client Consistency	0.46	Moderate
Cultivating Change Talk	0.541	Moderate
Soften Sustain Talk	0.52	Moderate
Empathy	0.62	Substantial
Partnership	0.472	Moderate
Reflection Quality	0.702	Substantial
Question Quality	0.739	Substantial
Effectiveness	0.483	Moderate
Goal Alignment	0.481	Moderate
Realignment	0.641	Substantial

Table A15: Human–LLM annotation reliability for rubric-based evaluations.

Metric	KMI	CI-NC	C+S	CI	AnnoMI
Turn-level					
Client					
Client Readability (↓)	5.87	4.68	5.77	4.45	7.22
Therapist					
Reflection Quality	4.44	3.71	2.38	4.61	3.37
Question Quality	4.52	4.39	2.94	4.97	3.61
Therapist Readability (↓)	6.22	4.98	6.18	4.93	6.65
Agent-level					
Client					
Consistency	3.76	4.79	3.42	4.92	3.51
Therapist					
Cultivating Change Talk	4.63	3.15	1.88	4.34	2.99
Soften Sustain Talk	4.73	4.27	2.53	4.88	3.77
Empathy	4.66	4.68	3.53	4.88	3.97
Partnership	4.48	4.60	2.71	4.88	3.90
Conversation-level					
Effectiveness	4.27	2.86	1.19	4.45	3.28
Realignment	4.49	4.53	3.21	4.87	3.85
Goal Alignment	4.60	3.89	2.13	4.73	3.87

Table A16: Turn-, agent-, and conversation-level evaluation of MI dialogue generated by four frameworks: `KMI`, CALM-IT without conversational dynamics (`CI-NC`), CAMI+STAR (`C+S`), and CALM-IT (`CI`). AnnoMI reports reference corpus values. Scores are averaged across all conversations and lengths. Downward arrow (↓) indicates that lower values correspond to better performance; otherwise, higher values are better.

D Appendix D: Experimental Configuration and Compute

D.1 Model and Inference Setup

All dialogue generation uses DeepSeek-V3.2 without fine-tuning. Inference is executed using `sglang`, and decoding parameters are held constant across frameworks.

D.2 Runtime Characteristics

Average wall-clock time per transcript is approximately 10 minutes for CAMI+STAR, 8 minutes for CALM-IT and 2 minutes for KMI.

D.3 Compute Usage

All experiments are run on nodes with 8×H200 GPUs. Across all transcripts, total compute usage is approximately 825 GPU-hours.