

StoryMI: Steerable Multi-Agent Therapeutic Dialogue Generation

Qingyu Meng¹, Min Chen¹, Dingming Liu^{1,2}, Yifan Mo¹,
Yue Su¹, Xin Sun³, Koen Hindriks¹, Jiahuan Pei^{1*}

¹Vrije Universiteit Amsterdam, ²Bol.com, ³NII, Tokyo Institute of Technology

q.meng@vu.nl, j.pei2@vu.nl

Abstract

Large language models (LLMs) can generate fluent dialogue, but prior works lack situational grounding, dynamic strategy control, and evaluation aligned with clinical standards in motivational interviewing (MI). We introduce **StoryMI**, a multi-LLM agent framework for controllable MI dialogue generation, where questionnaire-based client profiles are expanded into situational stories that provide narrative context for the dialogue. Therapist and client agents generate MI-coded utterances guided by MI codes selected by the interaction agent, while an interaction agent dynamically coordinates exchanges to control MI strategies during a multi-turn conversation. We propose a two-level evaluation protocol: lexical metrics and MI-specific measures of macro-level counseling strategies, alongside LLM-as-judge and human expert assessments. We construct a dataset of 6K simulated MI dialogues grounded in 1K questionnaire-story pairs, covering 12 MI codes and 13 symptom domains, and benchmark six open- and closed-source LLMs. Our results show that situational grounding and macro-level control can improve MI adherence and clinical plausibility, demonstrating the effectiveness of a structured multi-agent workflow for psychotherapy dialogue generation. We provide code and data for reproducibility.¹

1 Introduction

Motivational interviewing (MI) is a directive, client-centered counseling approach for eliciting clients' motivation for behavioral change (Miller and Rollnick, 2013). Behavioral coding (BC), which categorizes therapist and client behaviors (Tavabi et al., 2021), mitigates client resistance and disengagement associated with confrontational or paternalistic interaction styles. In ad-

*Corresponding author.

¹<https://github.com/Beren-sds/StoryMI>

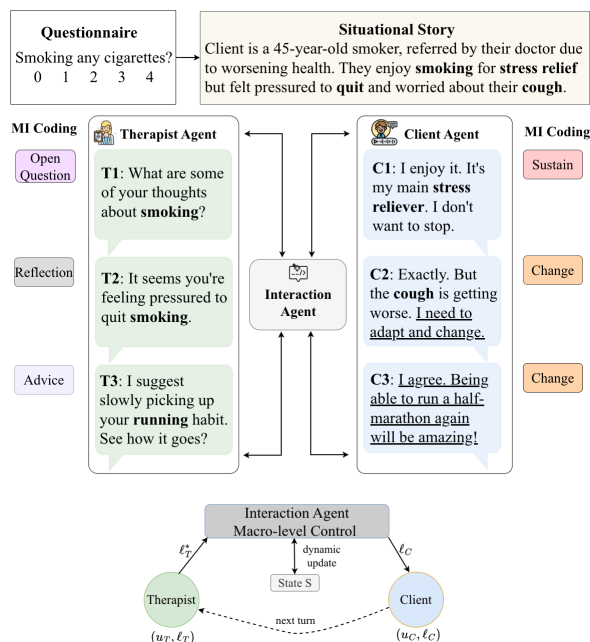


Figure 1: StoryMI workflow. Situational story derived from questionnaire grounds MI-coded therapist–client dialogue. An interaction agent dynamically controls and coordinates MI-coded exchanges between therapist and client agents, with dynamic state updates. Bold and underlined spans indicate story overlap and intent to change.

dition, *macro-level control* of a counseling strategy (a trajectory of behavior codes) is essential to guide the flow of the counseling (Shah et al., 2022). For example, in Motivational Interviewing Skill Code (MISC) (Miller, 2000), reflections should exceed questions by at least two to one; complex paraphrasing should dominate over simple echoing; and open questions should guide clients towards self-initiated change talk.

Unlike traditional methods (Li et al., 2022; Majumder et al., 2020), large language models (LLMs) enable psychotherapy dialogue agents through instruction following (Ouyang et al., 2022), standard adherence (Chiu et al., 2024), multi-party inter-

action simulation (Park et al., 2023). However, three barriers remain: (1) *Situational grounding*. DSM5AgentFlow (Ozgun et al., 2025) generates client profiles from questionnaires and simulates therapist–client agent conversations using an explainable diagnosis agent. But Likert-scale ratings in questionnaires do not capture the client’s situational context and personal intention, which are central to motivational interviewing. (2) *Dynamic controllability*. Recent work includes automated BC (Wu et al., 2022; Sun et al., 2024) and MI-guided dialogue generation (Sun et al., 2025b), but MI codes remain static, script- or turn-level transformations rather than adaptive, situationally grounded responses. Multi-LLM agent role-play frameworks (Li et al., 2023; Qiu and Lan, 2024) produce fluent text but lack the structured control and fail to capture complex MI interaction dynamics. (3) *Evaluation fidelity*. Accessing the therapeutically meaningful dialogue requires more than surface-level coherence metrics, such as automatic lexical metrics (See et al., 2019) and LLM-as-judge rubrics (Zheng et al., 2023; Liu et al., 2023). None of the metrics captures macro-level control of counseling strategies (see Table 1), which is essential.

We introduce StoryMI, a multi-agent framework that grounds MI-coded client-therapist dialogue in situational stories derived from questionnaires, as shown in Figure 1. First, client profiles are generated from questionnaires and expanded into situational stories, providing rich contextual grounding for MI-coded dialogue. Next, therapist and client agents alternately generate utterances along with their corresponding MI behavioral codes. An interaction management agent dynamically coordinates these exchanges, updating agent states to ensure adherence to MI strategies while maintaining naturalistic, situational conversation.

The contributions of this work are threefold:

- We design a multi-LLM agent workflow, which combines *situational story contextualization* with dynamically *controllable MI-coded interactions*, benchmarked across six representative LLMs.
- We propose MI-specific metrics, together with existing automatic lexical metrics and LLM-as-judge rubrics, to quantify both micro and macro-level clinical counseling strategy.
- We collect a dataset of 6,000 conversations grounded on 1,000 questionnaire-story pairs, covering 12 types of MI codes and 13 symptom domains for our evaluation and further study.

Table 1: Comparison of most related work. The symbols ✓, (✓), ✗ indicate full support, partial support, and no support. CP = Client Profiling, TM = Therapist Modeling, MS = Multi-turn Strategy, DA = Dual Agent, MI = Motivational Interview, HE = Human Evaluation.

Related Work	CP	TM	MS	DA	MI	HE
ChatPsych (Chen et al., 2023)	✗	✗	✗	(✓)	✗	✓
PATIENT-Ψ (Wang et al., 2024c)	(✓)	✗	(✓)	(✓)	✗	✓
NoteChat (Wang et al., 2024b)	(✓)	✗	✓	(✓)	✗	✗
ClientCAST (Wang et al., 2024a)	✗	✗	✓	(✓)	✗	✓
InteractAgent (Qiu and Lan, 2024)	✗	✗	✓	✓	✗	✓
CPsyCoun (Zhang et al., 2024)	✗	✗	✓	✗	✗	✓
MDD-5K (Yin et al., 2025)	(✓)	✗	✓	✓	✗	✗
DSM5AgentFlow (Ozgun et al., 2025)	✓	✗	✓	✓	✗	✓
AMIE (Tu et al., 2025)	(✓)	✗	✓	(✓)	✗	✓
BiMISC (Sun et al., 2024)	✗	✓	✗	✗	✓	✓
ScriptAlign (Sun et al., 2025a)	✗	✓	✓	✗	✓	✓
RethinkMI (Sun et al., 2025b)	✗	✓	✓	✗	✓	✓
StoryMI (Ours)	✓	✓	✓	✓	✓	✓

2 Related Work

2.1 MI-Adherent Dialogue Generation

Prior work on MI dialogue generation focuses on utterance-level strategy modeling without grounding responses in client-specific situations. BiMISC (Sun et al., 2024) demonstrates that LLMs can predict MI coding categories and that explicit codes improve alignment at the utterance level, though the setup operates on isolated turns without multi-turn flows or client profiling. ScriptAlign (Sun et al., 2025a) introduces script-strategy aligned generation using expert-crafted scripts paired with strategy labels, raising MI fidelity but limiting scalability due to handcrafted script dependence. RethinkMI (Sun et al., 2025b) proposes predicting MI strategies as intermediate reasoning before conditioning generation, improving controllability and explainability. DIIR (Xie et al., 2024) produces a framework to apply MI conversation strategies in the form of inductive rules from expert demonstrations. VirtualMI (Steenstra et al., 2024) develops a virtual counselor to conduct MI counseling in the context of alcohol use.

StoryMI addresses both gaps by replacing script dependence with standardized questionnaire profiling plus situational stories, enabling broad, clinically plausible coverage without manual script writing. Our controllable MI coding conditions each turn on behavioral strategy, while situational stories grounds responses in concrete client contexts.

2.2 Psycho-Conversation Simulation

Single-Agent Systems. ChatPsych (Chen et al., 2023) demonstrates tuning-free prompting for diag-

nostically competent chatbots but lacks modality-specific scaffolding and systematic multi-turn coding. PATIENT- Ψ (Wang et al., 2024c) develops CBT-style AI patients with cognitive distortions, yielding controllable personas yet remaining therapy-agnostic about MI mechanisms. NoteChat (Wang et al., 2024b) conditions doctor-patient role-play on clinical notes for documentation quality rather than counseling strategy. These approaches model the therapist or client separately, lacking macro-level strategy control.

Multi-Agent Systems. InteractAgent (Qiu and Lan, 2024) role-plays counselor-client interactions and fine-tunes on synthetic data, while CPsyCoun (Zhang et al., 2024) reconstructs multi-turn dialogues from clinical reports. MDD-5K (Yin et al., 2025) synthesizes psychiatrist-patient dialogues via neuro-symbolic diagnosis trees. DSM5AgentFlow (Ozgun et al., 2025) autonomously generates Likert-scale questionnaires and conducts simulations emphasizing diagnostic trustworthiness. These systems support multi-turn interaction but are neither controllable through explicit behavior coding nor situationally grounded in narrative client contexts.

Evaluation Fidelity. Almost all prior work relies on generic metrics (fluency, coherence) without alignment between automatic evaluation and clinical constructs. AMIE (Tu et al., 2025) optimizes diagnostic medical dialogue through self-play but targets medical diagnosis rather than psychotherapy skill expression. ClientCAST (Wang et al., 2024a) benchmarks therapist performance through simulated client interactions but does not supply a generative recipe for MI-structured dialogues. StoryMI introduces specific metrics measuring macro-level strategies and systematically compares automatic, LLM-based, and human evaluation to quantify alignment on therapeutic dimensions.

3 Workflow

StoryMI consists of three key components: (1) *Questionnaire-Based Profiling* (§3.1) establishes psychologically consistent client representations with standard questionnaire assessment; (2) *Situational Story Contextualization* (§3.2) bridges abstract symptom scores to narrative grounding, enabling the situational specificity that MI requires; and (3) *Controllable MI Dialogue Simulation* (§3.3) produces therapeutically coherent in-

teractions through multi-agent coordination. We simulate dialogue with three agents: client, therapist, and interaction manager. The interaction manager takes a special role of coordinating turns, selecting MI strategies, and enforcing therapeutic constraints via shared dialogue states.

3.1 Questionnaire-Based Profiling

This module constructs structured client profiles from standardized clinical instruments to ensure consistent and controllable symptom presentation across dialogues.

We model client heterogeneity using the DSM-5 questionnaire (First, 2024; Narrow et al., 2013), which contains 23 items spanning 13 symptom domains (e.g., depression, anger, sleep problems). For each item in the questionnaire, we use LLM generate (1) a severity score $s_i \in \{0, \dots, 4\}$ (from not at all to nearly every day), and (2) a brief first-person rationale describing the client’s subjective experience. The role-based system prompts and constrained JSON schemas are provided in Appendix F. The resulting structured profile \mathcal{P} serves as an explicit, interpretable control signal that grounds downstream dialogue generation in stable psychological characteristics while supporting explainability and reproducibility (Kim and Wang, 2025). A heterogeneity analysis of 1,000 client profiles is in Appendix C.

3.2 Situational Story Contextualization

This component transforms abstract questionnaire responses into situationally grounded narratives that enable naturalistic therapeutic exchange.

Raw questionnaire responses lack the contextual richness needed for naturalistic dialogue (Gao et al., 2023; Qiu and Lan, 2024). For instance, a depression score of “3” indicates frequency but not circumstances, triggers, or personal meaning. We address this gap by generating a situational story \mathcal{N} of approximately 200 words that transforms the client profile \mathcal{P} into concrete life experiences. This narrative grounding enables persona consistency across multi-turn interactions, provides emotional texture for authentic therapeutic exchange, and supports scalable generation of psychologically plausible profiles without manual curation (Kim et al., 2023). The story generation employs conditional prompting to select a primary symptom and construct a coherent first-person narrative around a specific scene, using constrained generation for symptom-behavior alignment.

Algorithm 1 MI-Coded Dialogue Generation

Require: Large language model \mathcal{M} , turn $\in [T_{\min}, T_{\max}]$, context window k

Initialize Dialogue State \mathcal{S} :

- 1: $\mathcal{S}.\mathcal{H} \leftarrow []$ \triangleright Dialogue history: list of (u, ℓ) pairs
- 2: $\mathcal{S}.\mathcal{C} \leftarrow []$ \triangleright MI code trajectory
- 3: $\mathcal{S}.t \leftarrow 0$ \triangleright Turn counter
- 4: $\mathcal{S}.\psi \leftarrow \text{False}$ \triangleright Completion flag
- 5: $\mathcal{S}.\text{context} \leftarrow (\mathcal{H}, \mathcal{C}, t, \psi)$

Define Nodes \mathcal{N} :

- 6: \mathcal{N}_C : ClientNode produces (u_C, ℓ_C) given $(\mathcal{S}, \mathcal{P}, \mathcal{N})$
- 7: \mathcal{N}_T : TherapistNode produces (u_T, ℓ_T) given (\mathcal{S}, ℓ_T^*)
- 8: \mathcal{N}_I : InteractionNode selects ℓ_T^* , updates \mathcal{S} , checks termination $\mathcal{S}.\psi$

Questionnaire-Based Profiling & Situational Story Contextualization

- 9: $\mathcal{P} \leftarrow \mathcal{N}_C.\text{FillQuestionnaire}(\mathcal{M})$ \triangleright Client profile
- 10: $\mathcal{N} \leftarrow \mathcal{N}_C.\text{GenerateStory}(\mathcal{P}, \mathcal{M})$ \triangleright Situational story

Controllable MI Dialogue Simulation

- 11: $(u_0, \ell_0) \leftarrow \mathcal{N}_T.\text{Greet}(\mathcal{M})$ \triangleright Initial greeting
- 12: $\mathcal{S} \leftarrow \mathcal{N}_I.\text{Update}(\mathcal{S}, u_0, \ell_0)$
- 13: **while** $\mathcal{S}.\psi \neq \text{True}$ **and** $\mathcal{S}.t < T_{\max}$ **do**
- 14: $(u_C, \ell_C) \leftarrow \mathcal{N}_C.\text{Generate}(\mathcal{S}, \mathcal{M})$
- 15: $\mathcal{S} \leftarrow \mathcal{N}_I.\text{Update}(\mathcal{S}, u_C, \ell_C)$
- 16: $\ell_T^* \leftarrow \mathcal{N}_I.\text{SelectStrategy}(\mathcal{S}.\mathcal{C}[-k:], \ell_C, \mathcal{M})$ \triangleright
- Macro-level control
- 17: $(u_T, \ell_T) \leftarrow \mathcal{N}_T.\text{Generate}(\mathcal{S}, \ell_T^*, \mathcal{M})$
- 18: $\mathcal{S} \leftarrow \mathcal{N}_I.\text{SynchronizeState}(\mathcal{S}, u_T, \ell_T)$
- 19: $\mathcal{S}.t \leftarrow \mathcal{S}.t + 1$ \triangleright Increment turn counter
- 20: **if** $\mathcal{S}.t \geq T_{\min}$ **then**
- 21: $\mathcal{S}.\psi \leftarrow \mathcal{N}_I.\text{CheckTermination}(\mathcal{S}.\mathcal{H}, \mathcal{M})$
- 22: **end if**
- 23: **end while**
- 24: **return** $\mathcal{S}.\mathcal{H}$

3.3 Controllable MI Dialogue Simulation

In this component, the interaction manager coordinates the client-therapist interactions with macro-level control and dynamic state updates. Algorithm 1 formalizes controllable MI-coded dialogue generation using a multi-agent workflow. It first initializes the dialogue state and generates client profiles from questionnaires, which are then expanded into situational stories providing narrative grounding. Therapist and client agents alternately produce utterances paired with MI behavioral codes, while the interaction management agent dynamically selects strategies, updates states, and monitors termination. This loop continues until the dialogue reaches the maximum turn limit or meets the completion criteria, resulting in a fully simulated, MI-consistent conversation.

3.3.1 MI Coding Scheme

We follow MISC/MITI (Miller et al., 2003; Moyers et al., 2016) scheme. Therapist behaviors partition into three categories: (1) *Reflection* (simple vs. complex) captures empathic echoing or summarization, (2) *Question* (open vs. closed) encompasses

open-ended exploration and closed inquiries, and (3) *Input* includes information-giving, advice, affirmations, or goal-setting initiated by therapist.

Client utterances receive tri-partite motivational classification: (1) *Change* expresses desire, ability, or commitment toward behavioral change, (2) *Sustain* articulates resistance or preference for status quo, and (3) *Neutral* constitutes non-motivational content discourse.

3.3.2 MI-Conditioned Generation

Each agent produces a dual output (u, ℓ) comprising both utterance and MI code. Generation proceeds in two stages. First, the Interaction Agent classifies the client’s motivational orientation and selects an appropriate therapist strategy $\ell_T^* \in \mathcal{L}_T$ according to MI strategies. For instance, change talk typically receives reflective responses that reinforce motivation, while sustain talk calls for open questions exploring ambivalence. Second, the selected code conditions the therapist’s response generation. This decoupling enables macro-level therapeutic control while preserving generation flexibility. The context is limited to the five most recent turns (Yen et al., 2024).

3.3.3 Dynamic Interaction Management

The interaction manager coordinates the client-therapist interactions with macro-level control and dynamic state updates. Figure 1 depicts the turn-by-turn coordination: the client agent produces an utterance with its MI code, the interaction agent classifies the client’s MI state and selects an appropriate therapist strategy, and the therapist agent generates a response conditioned on this strategy.

This is implemented by three key functions in Algorithm 1: *SelectStrategy* examines recent client codes $\mathcal{C}[-k:]$ and selects a therapist strategy following MI strategies. *SynchronizeState* updates the shared history \mathcal{H} and code trajectory \mathcal{C} after each turn, ensuring consistent context for generation. *CheckTermination* monitors for natural closure indicators (gratitude, farewell) via LLM-based detection after $\tau \geq T_{\min}$ turns, with a hard bound at T_{\max} to balance fidelity with computational feasibility (Pérez-Rosas et al., 2016).

4 Evaluation Protocol

We propose a two-level evaluation protocol: lexical metrics and MI-strategy metrics, alongside LLM-as-judge and human expert assessments.

4.1 Lexical Metrics

We employ four standard metrics capturing lexical fluency and diversity (See et al., 2019). (1) *Perplexity* measures fluency via language-model likelihood. Higher perplexity indicates less predictable, less templatic phrasing. (2) *Self-BLEU* quantifies intra-session redundancy by computing the average BLEU scores between each utterance and the rest of the dialogue. Lower values indicate less redundancy. (3) *Distinct-2* measures the ratio of unique bi-grams to total bi-grams, capturing lexical diversity. (4) *Entropy* quantifies token distribution uniformity across the dialogue.

4.2 Strategy Metrics

We introduce six behavioral metrics grounded in established MI quality indicators (See details in Appendix A). Here, the threshold δ follow the recommended macro-level strategies threshold defined in MISC coder’s manual (Miller, 2000).

- *Code Entropy* measures the diversity of therapeutic strategy usage via normalized Shannon entropy over the distribution of MI code categories. Higher values indicate balanced technique application across the defined strategies.
- *Strategy Adherence* quantifies alignment with the golden MI strategy distribution, computed as the negative exponential of KL divergence from MISC best practices (50% reflections, 25% questions, 20% input, 5% other). Scores near 1.0 indicate closer adherence to recommended practice patterns.
- *Reflection Depth* quantifies how much semantic content a reflection adds beyond the client’s original statement, computed as a weighted combination of sentence-level similarity (ensuring topical relevance) and token-level information gain (measuring novel therapeutic insight).
- *Complex Reflection Ratio* ($\delta > 0.5$) measures the proportion of reflections classified as paraphrase or summarization versus simple repetition, operationalizing reflection quality beyond surface-level echoing.
- *Open Question Ratio* ($\delta > 0.7$) captures question type distribution, favoring exploratory open-ended inquiries over closed yes/no questions that limit client exploration.
- *Reflection-to-Question Ratio* ($\delta > 2.0$) operationalizes the core MISC indicator that MI-adherent sessions should emphasize reflective responses over questioning, promoting client-

driven exploration.

4.3 LLM-as-Judge and Human Alignment

We define six rubrics motivated by MI principles using 5-point Likert scales, to evaluate therapeutic dialogue quality for both LLM and human judges: *Coherence* (logical dialogue flow) (Dziri et al., 2019), *Depth* (psychological insight and understanding), *Progress* (therapeutic advancement towards resolution), *Naturalness* (conversational authenticity), *Empathy* (affective attunement to client state) (Sharma et al., 2020; Cabrera Lozoya et al., 2025), and *Adherence* (adherence to MI strategies).

Two expert annotators with clinical psychology backgrounds independently rated 120 dialogues (20 per model) using identical rubrics as the LLM-as-a-Judge evaluation. The stratified sampling ensured balanced representation across models. Inter-annotator reliability reached moderate agreement using quadratic-weighted Cohen’s κ (Artstein and Poesio, 2008) ($\kappa = 0.51$ across all dimensions and models, with the highest reliability observed for Adherence as $\kappa = 0.79$).

5 Experimental Setup

5.1 Research Question

We seek to answer three research questions in our evaluation: **(RQ1)** Can we simulate client-therapist conversations adhering to MI strategies with multiple LLM-based agents? **(RQ2)** How does situational story and MI codes influence the mental health conversation quality? **(RQ3)** How well does LLM-as-a-Judge align with human annotators with respect to metric and model evaluation levels?

5.2 Dataset and LLMs

We construct a synthetic dataset of 6,000 multi-turn dialogues (113K+ utterances) grounded in 1,000 questionnaire–story pairs. Each questionnaire covers 23 DSM-5 items across 13 symptom domains. Each story averages around 200 words. Average dialogue lengths range from 13.3 to 25.6 turns, with utterance lengths from 27 words to 133 words. Detailed statistics are provided in Table 8. We use seven LLMs, including GPT-5-Nano² and five open-source LLMs provided by Ollama,³ i.e., LLaMA 3.1-8B, Qwen 2.5-7B, Gemma-7B, OpenChat-7B, and Phi-4-14B (hereafter, GPT,

²<https://platform.openai.com/docs/models/gpt-5-nano>

³<https://ollama.com/>

Table 2: Overall performance across two level metrics. **Bold** and underline indicate the best and second-best results. ✓ and ✗ indicate whether a metric meets the recommended threshold. ↑ higher is better; ↓ lower is better.

Metric	GPT-5-Nano	LLaMA 3.1-8B	Phi-4-14B	OpenChat-7B	Gemma-7B	Qwen 2.5-7B	Overall
<i>Lexical Metrics</i>							
Entropy (%) ↑	90.7	87.5	88.6	89.4	<u>89.3</u>	89.1	89.1
Distinct-2 (%) ↑	84.3	75.7	<u>76.8</u>	69.8	64.0	73.0	73.9
Perplexity ↑	<u>12.2</u>	10.9	17.1	7.7	7.3	10.0	10.9
Self-BLEU (%) ↓	16.7	27.0	<u>25.5</u>	39.3	49.3	34.2	32.0
<i>Strategy Metrics</i>							
Code Entropy (%) ↑	<u>80.0</u>	83.2	87.9	86.6	79.7	72.3	81.6
Strategy Adherence (%) ↑	<u>80.9</u>	53.0	53.0	81.3	80.1	81.0	71.6
Reflection Depth (%) ↑	<u>72.5</u>	45.7	42.5	67.0	72.1	75.9	62.6
Complex Reflection Ratio ($\delta > 50\%$)	✓ 98.3	✓ 62.2	✓ 58.0	✓ 82.6	✓ 97.4	✓ 99.0	✓ 83.0
Open Question Ratio ($\delta > 70\%$)	✗ 6.9	✓ 91.8	✓ 95.2	✗ 11.7	✗ 29.0	✗ 19.8	✗ 42.4
Reflection/Question Ratio ($\delta > 2.0$)	✓ 3.6	✗ 0.1	✗ 0.5	✗ 1.9	✓ 4.4	✓ 6.3	✓ 2.8

LLaMA, Phi, OpenChat, Gemma, Qwen), and GLM-5 (short for GLM) as an independent evaluator to rule out evaluator bias.

5.3 Implementation details

We use LangGraph⁴ to develop StoryMI (Appendix E). All 1,000 questionnaires and background stories were generated once using the same base model LLaMA with fixed decoding parameters and reused for all six generation models. In the ablation studies (without story, without MI code, both removed), all other decoding and inference settings for GPT stay unchanged compared with the full condition. We set the temperature to 0.7 and top-p to 0.9 for open-source LLMs. LLaMA served as the reference model for generating all the questionnaire profiles and situational stories. We use GPT as the closed-source LLM, which also serves as the LLM-as-judge evaluator. We run open-source LLMs on the region-based national supercomputer cluster using a single NVIDIA A100 GPU (40GB).

6 Results

6.1 Overall Performance (RQ1)

6.1.1 Lexical and Strategy Evaluation

Table 2 reports overall performance across lexical diversity and MI-specific strategy metrics.

First, StoryMI enables surface-level language quality across diverse LLMs. GPT attains the highest Entropy (90.7%) and Distinct-2 (84.3%), alongside the lowest Self-BLEU (16.7%), indicating both high diversity and low repetition. Open-source models also demonstrate competitive fluency (e.g., Phi achieves 88.6% Entropy and 17.1 Perplexity),

indicating that they enables fluent and diverse dialogue generation across model families.

Second, StoryMI consistently enables macro-level MI strategy control. Strategy Adherence is high (>80%) for most models (GPT, OpenChat, Gemma, and Qwen), indicating close alignment with the target MI strategy distribution. The Complex Reflection Ratio further demonstrates adherence to MI strategies, with all models exceeding the recommended threshold of 50%. These results directly validate StoryMI’s interaction agent as an effective mechanism for enforcing MI-adherent macro-strategies beyond turn-level prompting.

Third, lexical metrics alone fail to explain therapeutic quality differences. For instance, Phi shows the highest Perplexity (17.1) and Code Entropy (87.9%) but low Strategy Adherence (53.0%) and shallow Reflection Depth (42.5%). In contrast, Qwen achieves the highest Reflection Depth (75.9%) and Complex Reflection Ratio (99.0%) despite moderate lexical scores. This divergence demonstrates that MI-specific metrics reveal clinically meaningful distinctions that surface-level metrics cannot capture, underscoring the necessity of StoryMI’s evaluation framework.

6.1.2 LLM-as-Judge and Human Evaluation

Figure 2 presents rubric-based evaluation from both LLM judges and human experts.

First, Depth shows the strongest cross-evaluator consistency. Both LLM judges and human experts assign high Depth scores to dialogues with substantive psychological exploration, with LLaMA-generated dialogues receiving the highest ratings from both evaluator types (LLM: 3.84; Human: 4.96). This agreement suggests that Depth captures observable dialogue properties that both automated and human evaluation can reliably assess.

⁴<https://www.langchain.com/langgraph>

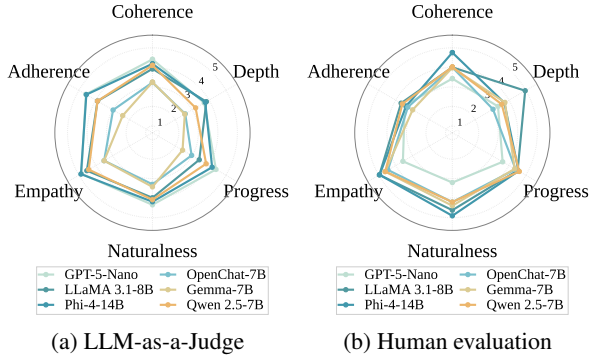


Figure 2: Therapeutic dialogue quality evaluation in Likert scale (1 to 5 points) across six dimensions.

Second, Naturalness and Adherence show the largest cross-evaluator divergence. LLM judges produce uniformly high scores across models (e.g., GPT Adherence: 4.59), while human experts assign more varied ratings with notably lower scores on certain dimensions (e.g., GPT Naturalness: 3.20). This pattern indicates that Naturalness and Adherence require assessment of subtle qualities, such as conversational authenticity and technique fidelity, that LLM judges have difficulty capturing.

Third, the Open Question Ratio correlates with higher human Empathy ratings. Models exceeding the 70% Open Question Ratio threshold (LLaMA: 91.8%; Phi: 95.2%) receive the highest human Empathy scores (4.97 and 4.95, respectively). Models with lower open question ratios (GPT: 6.9%; OpenChat: 11.7%) receive correspondingly lower Empathy ratings. This correlation supports the MI principle that open questions embody an accepting therapeutic stance, and demonstrates that StoryMI’s controllable generation enables systematic investigation of technique-outcome relationships.

6.2 Ablation Study (RQ2)

We conduct an ablation study of the key components of StoryMI, as shown in Table 3.

First, MI coding drives strategy adherence. Removing MI codes drops Strategy Adherence by 16.7 points (80.9% to 64.2%) while lexical metrics remain highly stable. This confirms that interaction-level coding enables macro-level strategy control. Second, situational stories improve behavioral diversity. Removing stories reduces Code Entropy by 5.0 points while maintaining Strategy Adherence, indicating that narrative grounding enriches therapeutic exchanges between client and therapist. Third, combined removal produces the largest degradation. Strategy Adherence drops from 80.9%

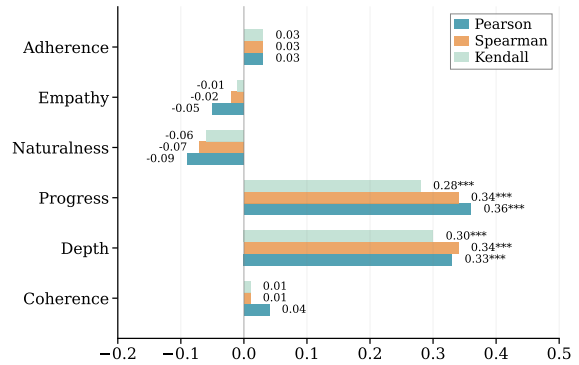
Table 3: Ablation study on GPT-5-Nano. **Bold** = largest degradation for each metric.

Metric	Full	w/o Story	w/o MI	w/o Both
<i>Lexical Metrics</i>				
Entropy (%) ↑	90.7	90.9	90.4	90.8
Distinct-2 (%) ↑	84.3	84.7	84.5	85.5
Perplexity ↑	12.2	12.8	12.1	12.8
Self-BLEU (%) ↓	16.7	15.4	17.6	15.0
<i>Strategy Metrics</i>				
Code Entropy (%) ↑	80.0	75.0	69.9	70.1
Strategy Adherence (%) ↑	80.9	78.8	64.2	61.4
Reflection Depth (%) ↑	72.5	73.8	68.5	63.4
Complex Reflection Ratio ($\delta > 50\%$)	98.3	98.6	90.5	82.3
Open Question Ratio ($\delta > 70\%$)	6.9	4.7	6.6	23.2
Reflection/Question Ratio ($\delta > 2.0$)	3.6	4.8	5.3	3.9

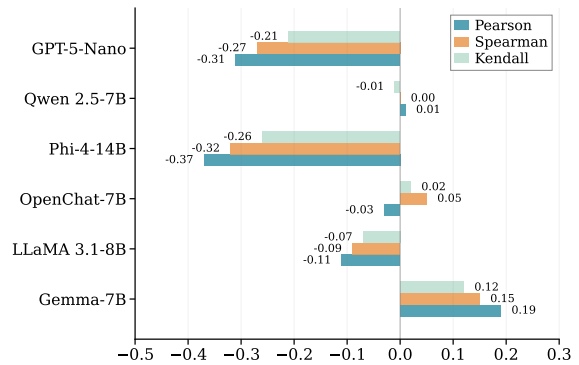
to 61.4%, demonstrating that both components contribute complementary benefits (qualitative examples in subsection D.1).

6.3 Human Alignment Analysis (RQ3)

Figure 3 presents correlation analysis between LLM-based and human evaluations.



(a) Dimension-level Correlation



(b) Model-level Correlation

Figure 3: Correlation between LLM-based and human evaluations. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

First, Depth and Progress show significant positive correlations. Depth ($r=0.33$) and Progress ($r=0.36$) indicate that LLM judges can reliably

identify dialogues that explore issues meaningfully and advance logically. These dimensions assess observable dialogue properties: whether conversations move beyond surface remarks and progress without repetition.

Second, Coherence, Naturalness, Empathy, and Adherence show near-zero correlations. Despite high human inter-rater reliability on Adherence ($\kappa=0.79$), in general LLM-human correlations remain weak ($|r|<0.10$). These dimensions require assessment of subtler qualities: conversational flow, authenticity, affective attunement, and therapeutic technique fidelity.

Third, no model achieves strong cross-evaluator consistency. All models show weak correlations ($|r|<0.38$), indicating that LLM-human agreement on relative dimension rankings is limited regardless of overall dialogue quality. Overall, LLM-as-Judge can efficiently screen for dialogue development quality at scale, while human experts remain necessary for validating therapeutic fidelity and conversational authenticity.

6.4 Self- vs Cross-Evaluator Bias Analysis

Bias Verification. To rule out evaluator bias, we re-evaluated all GPT-generated dialogues with GLM using the same rubric and conducted a paired t -test against the self-evaluation scores, as shown in Table 4.

We find that GPT rates its own dialogues *lower* than GLM on 5 of 6 dimensions ($p < 0.001$), indicating no inflated scoring tendency. Depth is the sole exception (+0.19), where GPT gives itself a slightly higher score.

Cross-Model Ranking Consistency. Table 5 compares the two evaluators’ results across six models, including their ranking positions. The top-4 and bottom-2 model clusters are consistent; notably, GLM ranks Phi first and GPT second, confirming that the original ranking holds under an independent evaluator. Furthermore, the top-4 and bottom-2 model ranking clusters remain consistent across two evaluators.

Human Alignment Consistency. Table 6 compares the alignment of two evaluators’ judgments with human ratings. GLM achieves higher human correlation on 4 of 6 dimensions, averaging across three correlation metrics. Progress is particularly notable ($r = 0.54$ vs. 0.43). On Naturalness, GPT’s correlation is near zero ($r = 0.01$), while GLM reaches significance ($\rho = 0.22$, $p < .05$),

Table 4: Comparison of self- and cross-evaluator results across different rubrics, including a paired t -test of their differences. Diff. = GPT - GLM).

	LLM Evaluators			Paired t -test	
	GPT-5-Nano	GLM-5	Diff.	t	p
Coherence	4.48	4.93	-0.45	-8.35	<.001
Depth	3.74	3.55	+0.19	3.27	.001
Progress	4.46	4.94	-0.48	-9.20	<.001
Naturalness	4.39	4.81	-0.42	-6.75	<.001
Empathy	4.90	4.98	-0.08	-2.36	.020
MI Alignment	4.71	4.94	-0.23	-4.35	<.001
Overall	4.45	4.69	-0.24	-6.03	<.001

Table 5: Comparison of the two evaluators’ alignment with human annotations, measured using Pearson’s r , Spearman’s ρ , and Kendall’s τ correlation coefficients. Diff. = GPT-GLM).

Model	LLM Evaluators			Rank Position	
	GPT-5-Nano	GLM-5	Diff.	GPT-5-Nano	GLM-5
GPT-5-Nano	4.45	4.69	-0.24	1	2
Phi-4-14B	4.31	4.78	-0.47	2	1
Qwen-2.5-7B	3.94	4.17	-0.23	3	4
LLaMA-3.1-8B	3.93	4.49	-0.57	4	3
OpenChat-7B	3.20	2.83	+0.37	5	5
Gemma-7B	2.82	2.59	+0.23	6	6

consistent with our finding that Naturalness shows the largest human-LLM gap (§6.3). Depth is the only dimension where GPT correlates more strongly, possibly reflecting familiarity with its own generation patterns.

Table 6: Comparison of the two evaluators’ judgment alignment with human annotations, measured by Pearson’s r , Spearman’s ρ , and Kendall’s τ correlation coefficients. Significance: * $p < .05$, † $p < .01$, ‡ $p < .001$.

Dimension	GLM-5			GPT-5-Nano		
	r	ρ	τ	r	ρ	τ
Coherence	.128	.157	.134	.098	.100	.087
Depth	.313 [†]	.314 [†]	.267 [†]	.368 [‡]	.384 [‡]	.336 [‡]
Progress	.538 [‡]	.504 [‡]	.408 [‡]	.432 [‡]	.417 [‡]	.351 [‡]
Naturalness	.166	.219*	.182*	.013	.039	.036
Empathy	.034	.049	.047	.012	.082	.075
MI Alignment	.169	.172	.148	.111	.156	.137

7 Discussion and Implication

Complementary Evaluation Paradigms. Our correlation analysis shows that LLM judges and human experts capture different aspects of therapeutic dialogue quality. LLM evaluators reliably assess observable properties such as topical exploration and logical progression, while human experts prove necessary for subtler qualities including conversational authenticity and affective attunement. This divergence suggests a practical division of labor: LLM-based methods can efficiently screen dialogue development at scale (Zheng et al., 2023; Liu et al., 2023), while human judgment remains

necessary for validating therapeutic fidelity (Basar et al., 2025). We advocate for hybrid protocols that use LLM efficiency for initial filtering while reserving expert assessment for dimensions where automated evaluation falls short, such as empathy and naturalness as observed in our study.

Open Questions and Therapeutic Stance. Our findings show a consistent relationship between open question usage and human-perceived empathy: models with higher open question ratios receive higher empathy ratings from human experts. This aligns with MI theory, which positions therapeutic empathy as emerging from an accepting, exploratory stance (Miller and Rollnick, 2013). Open questions create space for client elaboration and signal genuine interest in the client’s perspective, which human raters appear to recognize more readily than LLM judges. The controllable generation in StoryMI allows investigation of how MI strategies relate to human perceived quality, further extending prior work on strategy-aligned generation (Sun et al., 2025a).

Multiple Paths to Therapeutic Modeling. Our evaluation shows that models achieve therapeutic plausibility through distinct strategic profiles: some excel through exploratory questioning while others emphasize reflective techniques. Both paths can lead to favorable human ratings, suggesting that therapeutic quality is not tied to a single behavioral pattern. Specifically, strategy metrics ensure technique fidelity to MI principles (Moyers et al., 2016), while human evaluation validates perceived therapeutic effectiveness. This complementarity supports multi-dimensional assessment that considers both behavioral adherence and subjective quality (See et al., 2019).

Client Profiling and Persona Diversity. Relying on standardized questionnaire instruments inherently constrains the persona space: the resulting client profiles are shaped by the fixed item set and Likert-scale format of the DSM-5 cross-cutting symptom measures, which may not fully capture idiosyncratic personal narratives, cultural expressions of distress, or contextual factors beyond symptom frequency. This profiling choice propagates through the pipeline, influencing both the diversity of situational stories and the range of therapeutic dynamics represented in the generated dataset. Future work could explore complementary profiling strategies, such as open-ended intake

interviews, to broaden persona coverage.

Implications. MI-style therapeutic quality can be shaped by controlling dialogue strategies rather than relying on model scale alone. Different strategy profiles achieve comparable human-rated quality, suggesting MI systems should adapt to context rather than follow rigid recipes. Evaluation frameworks should jointly report strategy adherence and human perception, since either alone misses key aspects of therapeutic stance.

8 Conclusion

This work shows that structured multi-agent coordination enables controllable and situationally grounded MI dialogue generation. StoryMI integrates questionnaire-based client profiling, narrative contextualization, and macro-level MI behavioral coding to produce therapeutically coherent dialogues. Our evaluation demonstrates that MI-specific metrics capture clinically meaningful distinctions beyond lexical measures alone. The workflow, dataset, and evaluation protocol together provide a foundation for future research, with open directions including extension to validation in real-world training and clinical settings.

Limitations

We acknowledge several limitations. First, our framework focuses on motivational interviewing as a single therapeutic modality, whereas real clinical practice often integrates multiple approaches; extending controllable generation to multi-modality therapy would require additional coding schemes beyond MISC. Second, the dialogues are validated through expert annotation rather than interaction with actual clients or deployment in clinical training settings, so user studies with trainees or clinicians remain necessary to establish real-world applicability. Third, the questionnaire-based profiling reflects Western diagnostic norms embedded in DSM-5 and may require adaptation for cross-cultural contexts where MI is practiced with different populations. Finally, while we demonstrate LLM–human alignment patterns, establishing causal relationships between specific MI strategies and therapeutic outcomes would require longitudinal studies with actual client populations, which is beyond the scope of this work.

References

- Ron Artstein and Massimo Poesio. 2008. [Inter-coder agreement for computational linguistics](#). *Computational Linguistics*, pages 555–596.
- Erkan Basar, Xin Sun, Iris Hendrickx, Jan de Wit, Tibor Bosse, Gert-Jan De Bruijn, Jos A. Bosch, and Emiel Kraemer. 2025. [How well can large language models reflect? A human evaluation of LLM-generated reflections for motivational interviewing dialogues](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1964–1982. Association for Computational Linguistics.
- Daniel Cabrera Lozoya, Eloy Hernandez Lua, 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*, pages 157–171. Association for Computational Linguistics.
- Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. [LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation](#). *Preprint*, arXiv:2305.13614.
- Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. [A computational framework for behavioral assessment of LLM therapists](#). *Preprint*, arXiv:2401.00820.
- Nouha Dziri, Ehsan Kamaloo, Kory Mathewson, and Osmar 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*, pages 3806–3812. Association for Computational Linguistics.
- Michael B First. 2024. *DSM-5-TR Handbook of Differential Diagnosis*. American Psychiatric Association Publishing.
- Silin Gao, Beatriz Borges, Soyoung Oh, Deniz Bayazit, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2023. [PeaCoK: Persona commonsense knowledge for consistent and engaging narratives](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 6569–6591. Association for Computational Linguistics.
- Brian Hyeonseok Kim and Chao Wang. 2025. [Large language models for interpretable mental health diagnosis](#). *Preprint*, arXiv:2501.07653.
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. [SODA: Million-scale dialogue distillation with social commonsense contextualization](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949. Association for Computational Linguistics.
- Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. [CAMEL: Communicative agents for "mind" exploration of large language model society](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 51991–52008. Curran Associates, Inc.
- Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. [Knowledge bridging for empathetic dialogue generation](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10993–11001. AAAI Press.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.
- Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. [MIME: MIMicking emotions for empathetic response generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8968–8979. Association for Computational Linguistics.
- William R. Miller. 2000. *Motivational Interviewing Skill Code (MISC) Coder’s Manual*. Unpublished manuscript. Revised 11/29/2000. University of New Mexico.
- William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. [Manual for the motivational interviewing skill code \(MISC\)](#). *Unpublished manuscript*. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico.
- William R Miller and Stephen Rollnick. 2013. *Motivational Interviewing: Helping People Change*, 3rd edition. Guilford Press.
- Theresa B. Moyers, Lauren N. Rowell, Jennifer K Manuel, Denise B. 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.
- William E. Narrow, Diana E. Clarke, S. Janet Kuramoto, Helena C. Kraemer, David J. Kupfer, Lisa Greiner, and Darrel A. Regier. 2013. [DSM-5 field trials in the United States and Canada, Part III: Development and reliability testing of a cross-cutting symptom assessment for DSM-5](#). *American Journal of Psychiatry*, 170(1):71–82.

- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Mithat Can Ozgun, Jiahuan Pei, Koen Hindriks, Lucia Donatelli, Qingzhi Liu, and Junxiao Wang. 2025. [Trustworthy AI psychotherapy: Multi-agent LLM workflow for counseling and explainable mental disorder diagnosis](#). In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, page 2263–2272. Association for Computing Machinery.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22. Association for Computing Machinery.
- Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. [Building a motivational interviewing dataset](#). In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51. Association for Computational Linguistics.
- Huachuan Qiu and Zhenzhong Lan. 2024. [Interactive agents: Simulating counselor-client psychological counseling via role-playing LLM-to-LLM interactions](#). *Preprint*, arXiv:2408.15787.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. [What makes a good conversation? How controllable attributes affect human judgments](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1702–1723. Association for Computational Linguistics.
- Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. 2022. [Modeling motivational interviewing strategies on an online peer-to-peer counseling platform](#). *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–24.
- Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. [A computational approach to understanding empathy expressed in text-based mental health support](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5263–5276. Association for Computational Linguistics.
- Ian Steenstra, Farnaz Nouraei, Mehdi Arjmand, and Timothy Bickmore. 2024. [Virtual agents for alcohol use counseling: Exploring LLM-powered motivational interviewing](#). In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, pages 1–10. Association for Computing Machinery.
- Xin Sun, Jan de Wit, Zhuying Li, Jiahuan Pei, Abdallah El Ali, and Jos A. Bosch. 2025a. [Script-strategy aligned generation: Aligning LLMs with expert-crafted dialogue scripts and therapeutic strategies for psychotherapy](#). *Proceedings of the ACM on Human-Computer Interaction*, 9(7).
- Xin Sun, Jiahuan Pei, Jan de Wit, Mohammad Aliannejadi, Emiel Kraemer, Jos T.P. Dobber, and Jos A. Bosch. 2024. [Eliciting motivational interviewing skill codes in psychotherapy with LLMs: A bilingual dataset and analytical study](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 5609–5621. ELRA and ICCL.
- Xin Sun, Xiao Tang, Abdallah El Ali, Zhuying Li, Pengjie Ren, Jan de Wit, Jiahuan Pei, and Jos A Bosch. 2025b. [Rethinking the alignment of psychotherapy dialogue generation with motivational interviewing strategies](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1983–2002. Association for Computational Linguistics.
- Leili Tavabi, Trang Tran, Kalin Stefanov, Brian Borsari, Joshua Woolley, Stefan Scherer, and Mohammad Soleymani. 2021. [Analysis of behavior classification in motivational interviewing](#). In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 110–115. Association for Computational Linguistics.
- Tao Tu, Mike Schaekermann, Anil Palepu, and 1 others. 2025. [Towards conversational diagnostic artificial intelligence](#). *Nature*, 642(8067):442–450.
- Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024a. [Towards a client-centered assessment of LLM therapists by client simulation](#). *Preprint*, arXiv:2406.12266.
- Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024b. [NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes](#). In *Findings of the Association for Computational Linguistics*, pages 15183–15201. Association for Computational Linguistics.
- Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024c. [PATIENT- \$\psi\$: Using large language models to simulate patients for training mental health professionals](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural*

Language Processing, pages 12772–12797. Association for Computational Linguistics.

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

Zhouhang Xie, Bodhisattwa Prasad Majumder, Mengjie Zhao, Yoshinori Maeda, Keiichi Yamada, Hiromi Wakaki, and Julian McAuley. 2024. [Few-shot dialogue strategy learning for motivational interviewing via inductive reasoning](#). In *Findings of the Association for Computational Linguistics*, pages 13207–13219. Association for Computational Linguistics.

Howard Yen, Tianyu Gao, and Danqi Chen. 2024. [Long-context language modeling with parallel context encoding](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 2588–2610. Association for Computational Linguistics.

Congchi Yin, Feng Li, Shu Zhang, Zike Wang, Jun Shao, Piji Li, Jianhua Chen, and Xun Jiang. 2025. [MDD-5k: A new diagnostic conversation dataset for mental disorders synthesized via neuro-symbolic LLM agents](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25715–25723. AAAI Press.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. [Judging LLM-as-a-judge with MT-bench and Chatbot Arena](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 46595–46623. Curran Associates Inc.

A MI Strategy Metric Definition

This appendix provides complete mathematical definitions, computational details, and illustrative examples for the proposed MI-specific automatic metrics introduced in §4.

A.1 MI Code Distribution Entropy

MI Code Entropy measures the diversity of therapeutic strategies employed throughout a dialogue session. Higher entropy indicates balanced strategy usage across MI techniques; lower entropy suggests over-reliance on specific strategies.

Definition. Given a sequence of therapist MI codes $\mathbf{c} = (c_1, \dots, c_n)$ where each $c_i \in \mathcal{C} = \{\text{reflection, question, input, other}\}$, we compute:

$$H(\mathbf{c}) = -\frac{1}{\log_2 |\mathcal{C}_{\text{obs}}|} \sum_{c \in \mathcal{C}_{\text{obs}}} p(c) \log_2 p(c) \quad (1)$$

where $p(c) = \text{count}(c)/n$ and $\mathcal{C}_{\text{obs}} \subseteq \mathcal{C}$ is the set of observed categories. We normalize by $\log_2 |\mathcal{C}_{\text{obs}}|$ to ensure $H \in [0, 1]$.

Example. Consider a 10-turn dialogue with the following counts: reflection (6), question (3), therapist_input (1), and other (0).

The observed set is $\mathcal{C}_{\text{obs}} = \{\text{reflection, question, input}\}$, so $|\mathcal{C}_{\text{obs}}| = 3$. The probabilities are $p(\text{refl}) = 0.6$, $p(\text{ques}) = 0.3$, and $p(\text{input}) = 0.1$.

The raw entropy calculation is:

$$\begin{aligned} H_{\text{raw}} &= -(0.6 \log_2 0.6 + 0.3 \log_2 0.3 \\ &\quad + 0.1 \log_2 0.1) \\ &\approx 1.295 \end{aligned}$$

The final normalized score is $H = 1.295 / \log_2(3) = 0.817$.

A.2 Strategy Adherence

The Strategy Adherence (\mathcal{S}_{adh}) quantifies how closely the observed distribution of MI codes match best-practice recommendations.

Ideal Distribution. Based on MISC guidelines we define:

$$P_{\text{ideal}} = \{\text{reflection} : 0.50, \text{question} : 0.25, \text{input} : 0.20, \text{other} : 0.05\}$$

Definition. Given observed distribution P_{obs} , we compute:

$$\mathcal{S}_{adh} = \exp(-D_{\text{KL}}(P_{\text{obs}} \| P_{\text{ideal}})) \quad (2)$$

We apply additive smoothing ($\epsilon = 10^{-6}$) to handle zero counts.

Example. Using the previous counts, $P_{\text{obs}} = \{0.6, 0.3, 0.1, \epsilon\}$. The KL divergence is:

$$\begin{aligned} D_{\text{KL}} &\approx 0.6 \ln(0.6/0.5) + 0.3 \ln(0.3/0.25) \\ &\quad + 0.1 \ln(0.1/0.2) \\ &\approx 0.109 + 0.055 - 0.069 = 0.095 \end{aligned}$$

The final score is $\mathcal{S}_{adh} = \exp(-0.095) = 0.909$.

A.3 Reflection Depth

Reflection Depth (\mathcal{R}_d) computes a continuous score representing the semantic elaboration of a reflection beyond the client’s original statement. It relies on two components: Semantic Similarity and Information Gain.

Semantic Similarity (Sim). We encode the reflection r and client utterance u into embeddings e_r, e_u using the all-MiniLM-L6-v2 model. The raw cosine similarity is normalized to $[0, 1]$:

$$\text{Sim}(r, u) = \frac{1}{2} \left(\frac{e_r \cdot e_u}{\|e_r\| \|e_u\|} + 1 \right) \quad (3)$$

Information Gain (Info). We define T_r and T_u as the sets of lemmatized content tokens (excluding stopwords/punctuation) for r and u . A token $t \in T_r$ is considered *novel* if it is semantically distinct from all tokens in T_u :

$$\text{IsNovel}(t) = \mathbb{I} \left(\max_{t' \in T_u} \text{cossim}(t, t') < \theta \right) \quad (4)$$

where $\text{cossim}(t, t')$ is the cosine similarity between token embeddings and the threshold $\theta = 0.8$. We compute:

$$\text{Info}(r, u) = \frac{\sum_{t \in T_r} \text{IsNovel}(t)}{|T_r|} \quad (5)$$

Definition. The final score is a weighted sum:

$$\mathcal{R}_d = \frac{1}{|R|} \sum_{(r,u) \in R} (0.4 \cdot \text{Sim}(r, u) + 0.6 \cdot \text{Info}(r, u)) \quad (6)$$

Rationale. We weight Information Gain higher (0.6) to reward reflections that add therapeutic meaning (paraphrasing) rather than merely repeating content.

Example. *Client (u):* “I feel so tired.” ($T_u = \{\text{feel, tired}\}$) *Reflection (r):* “It sounds like you are exhausted.” ($T_r = \{\text{sound, exhausted}\}$)

- **Sim:** Embeddings are close but not identical; assume normalized $\text{Sim} \approx 0.70$.
- **Info:** “Exhausted” is similar to “tired” ($\text{cossim} > 0.8$), so it is *not* novel. “Sound” is novel. $\text{Info} = 1/2 = 0.5$.
- **Score:** $0.4(0.7) + 0.6(0.5) = 0.28 + 0.30 = 0.58$.

A.4 Complex Reflection Ratio

The Complex Reflection Ratio (\mathcal{R}_c) measures the proportion of reflections demonstrating therapeutic depth beyond simple repetition.

Metrics Calculation. For a reflection r and client utterance u , we compute: (1) **Similarity (Sim):** Cosine similarity of sentence embeddings (all-MiniLM-L6-v2 model, mapping to 384 dimensional dense vector space). (2) **Info Gain (Info):** The ratio of semantically novel tokens in r not present in u (token similarity threshold < 0.8).

Reflection Classification.

- **Repeat:** $\text{Sim} > 0.9$ and $\text{Info} < 0.15$.
- **Rephrase:** $\text{Sim} > 0.75$ and $\text{Info} < 0.35$.
- **Paraphrase (Complex):** $\text{Sim} > 0.5$ and $\text{Info} < 0.6$.
- **Summarize (Complex):** Otherwise.

Definition.

$$\mathcal{R}_c = \frac{|\{r \in R : \text{class}(r) \in \{\text{para, summ}\}\}|}{|R|} \quad (7)$$

Example. *Client:* “I feel so tired.” *Reflection A (Simple):* “You are very tired.” ($\text{Sim} \approx 0.95, \text{Info} \approx 0.0$). *Reflection B (Complex):* “It sounds like you are exhausted by the pressure.” ($\text{Sim} \approx 0.60, \text{Info} \approx 0.5$). If a session has only these two reflections, $\mathcal{R}_c = 1/2 = 0.5$.

A.5 Open Question Ratio

The Open Question Ratio (\mathcal{Q}_o) captures the distribution of question types, favoring exploratory inquiries that invite elaboration.

Definition.

$$\mathcal{Q}_o = \frac{|\text{OpenQuest.}|}{|\text{OpenQuest.}| + |\text{ClosedQuest.}|} \quad (8)$$

Example. *Q1 (Open):* “What about your thoughts on the materials that you mentioned last time? Could you tell me more about it?”, *Q2 (Closed):* “Did you sleep well?”, *Q3 (Closed):* “Did you take medicine last week on time?” Calculation: $\mathcal{Q}_o = 1/(1 + 2) = 0.33$. (Target > 0.70).

A.6 Reflection-to-Question Ratio

The Reflection-to-Question Ratio (\mathcal{R}/\mathcal{Q}) is a core MISC indicator. MI-adherent sessions should emphasize reflective listening over interrogation.

Definition.

$$\mathcal{R}/\mathcal{Q} = \frac{\text{count(reflection)}}{\text{count(question)}} \quad (9)$$

Example. In a session with 12 reflections and 4 questions, $\mathcal{R}/\mathcal{Q} = 12/4 = 3.0$, exceeding the MISC recommended target of > 2.0 .

B MI Code Definitions

Table 7 defines the MI behavioral codes used for turn-level annotation. Therapist codes follow the MITI coding manual (Moyers et al., 2016), and client codes derive from the MISC manual framework (Miller et al., 2003).

Table 7: MI behavioral code taxonomy.

Code	Definition
<i>Therapist Codes</i>	
Reflection	Mirrors back the essence of client statements; includes simple and complex reflections.
Question	Seeks clarity or explores client perspective. <i>Open</i> : invites elaboration. <i>Closed</i> : expects yes/no or brief answer.
Input	Non-reflection, non-question utterances: <i>information-giving, advice, affirmations, or goal-setting.</i>
<i>Client Codes</i>	
Change	Language favoring change: desire, ability, reasons, need, commitment, or taking steps.
Sustain	Language opposing change or favoring status quo.
Neutral	Utterances without directional motivational content.

C Dataset Statistics

Table 8 shows the average turns and length of the generated conversations across models.

Table 8: Average turns and length of the generated conversations across models.

Model	Average Turns	Average Length
GPT-5-Nano	17.19	39.61
LLaMA 3.1-8B	22.58	55.21
Qwen 2.5-7B	25.63	31.06
Gemma-7B	17.89	27.07
OpenChat-7B	16.96	42.02
Phi4-14B	13.29	133.46

Table 9 summarizes demographic and severity score distributions. The wide age range (18–65) and broad severity spread (IQR = 27–44) demonstrate substantial demographic and clinical heterogeneity.

Table 10 reports the symptom severity distribution across all 13 DSM-5 domains. Severity patterns vary widely across domains: depression clus-

Table 9: Demography and score distribution of 1,000 generated client profiles.

Property	Value
Identity types	1 (Adult, 100%)
Age	Mean = 41.1, SD = 13.6, Range = [18, 65] Median = 41.0, IQR = [29.0, 53.0]
Total severity score	Mean = 36.66, SD = 11.4, Range = [11, 73] Q1 = 27.0, Median = 37.0, Q3 = 44.0

ters severe (66.3%), anxiety and sleep concentrate mild-to-moderate, and suicidal ideation is absent in 80.7%, reflecting clinically plausible prevalence. All 13 DSM-5 domains are covered, confirming heterogeneous and representative profiles. The goal of the profiling module is controlled variation for systematic benchmarking, not replication of clinical intake distributions.

Table 10: Symptom severity distribution (% of profiles at each level). Bold indicates the modal category for each domain.

Domain	None	Slight	Mild	Mod.	Severe
Depression	0.3	0.5	1.7	31.2	66.3
Anger	8.6	7.0	39.5	6.2	38.6
Mania	2.0	5.9	26.8	30.4	34.8
Anxiety	20.3	18.5	31.4	21.0	8.7
Somatic symptoms	9.3	23.8	33.3	15.3	18.2
Suicidal ideation	80.7	5.0	4.5	1.8	8.0
Psychosis	42.9	18.9	17.2	13.9	7.0
Sleep problems	12.1	19.5	36.6	18.2	13.5
Memory	22.0	15.0	27.5	23.2	12.2
Repet. thoughts & behav.	10.8	6.9	26.6	27.3	28.3
Dissociation	5.6	3.0	18.8	42.9	29.6
Personality functioning	24.1	13.7	15.5	9.6	37.0
Substance use	54.6	13.1	13.5	7.3	11.5

D Case Study

D.1 Qualitative Ablation Study

We compare how the same client scenario changes across ablation conditions, generated by GPT-5-nano. With the full system, the client produces vivid, situation-specific language and the therapist reflects the emotional content before asking an open question. Without the story, the client opens generically and the therapist’s reflection lacks situational anchoring. Without MI coding, the therapist shifts toward directive problem-solving, bypassing the reflective stance that MI requires. Quantitative ablation results are reported in the main text (Table 3).

Full StoryMI (Turn 2, therapist uses reflection grounded in the situational story):

CLIENT: I'm feeling really on edge after that argument. My heart was racing, and the room felt like it was closing in on me. [...] I'm angry and frustrated, but also a little drained and worried I'll snap again if something small sets me off.

THERAPIST: **That sounds really rough, your body and mind are telling you to slow down, and you're looking for a way to steady yourself when small things push you over the edge. What's one small step you think could help you calm down the next time you notice those hints of tension building?**

w/o Story (Turn 2, same client profile, no situational grounding):

CLIENT: I'm feeling overwhelmed but relieved to be talking about it.

THERAPIST: It sounds like you're carrying a lot right now, but there's some relief in having this space to talk.

w/o MI Coding (Turn 2, story present, no MI code selection):

CLIENT: ...I keep second-guessing myself and wasting time double-checking things that might be fine.

THERAPIST: **That makes sense... Would it help to set a quick, concrete rule for the checks (like one final review after the whole draft) and maybe a timer to keep you moving?**

D.2 High-Quality vs Low-Quality Example

Turn 1 *[sustain → reflection]*

CLIENT: I'm feeling pretty on edge today. The usual numbness is there, but it's like a buzzing under my skin—hard to sit still. [...] The intrusive thoughts about my addiction keep pushing in [...] I'm trying to talk to people about it, but it usually just gets brushed off.

THERAPIST: **That sounds exhausting—like there's a lot buzzing at once and it's hard to sit with it. I hear the numbness, the restlessness, the way, the intrusive thoughts about addiction keep nudging you.**

Turn 2 *[change → question]*

CLIENT: Yeah, it is exhausting. I guess I just want to know I'm not stuck like this forever [...] THERAPIST: **I'm hearing you want a path where these thoughts don't own the day—like a way to ride them rather than be pulled under. What's one small moment today where you felt even a tiny bit of control?**

[Turns 3–14: Client and therapist develop a “one breath” cue, explore barriers, create implementation plan. Varied MI techniques facilitate self-discovery.]

Turn 15 *[reflection]*

CLIENT: I want to remember that pausing, even for one breath, can give me a moment of choice. THERAPIST: That's a grounded takeaway—one breath, one moment of choice. If anything spikes before we meet again, what's the first thing you'll do?

High-Quality (Session 769). Therapist explores multiple topics with empathic engagement, uses varied techniques (reflections, open questions), facilitates natural progression toward change talk.

Turn 1 *[change → reflection]*

CLIENT: I'm feeling really anxious today. When my alarm went off, my thoughts started racing [...] I felt detached [...] I'm worried this is starting to affect how I function at work. THERAPIST: That sounds really heavy to carry, and you're paying attention to how it's hitting your work and focus. It makes sense you'd want to slow the racing thoughts.

[Turns 2–4: Therapist identifies grounding technique. Client describes plan with “I'm here” cue.]

Turns 5–7: Repetitive Pattern *[all reflection]*

CLIENT (T5): I'll sit with my feet flat, take a 4-2-6 cycle, and say “I'm here” [...] THERAPIST: **That sounds like a solid plan you're choosing.** CLIENT (T6): Yeah, I'll start with the 4-2-6, saying “I'm here” [...] THERAPIST: **That sounds like you're taking a clear, doable step.** CLIENT (T7): I think I'm going to try [...] saying “I'm here” [...] THERAPIST: **That sounds like a solid, doable first step.**

[Turns 8–15: Circular pattern continues]

Turn 16 *[reflection] (Abrupt closure)*

CLIENT: I'd say “Back to the breath” [...] THERAPIST: That sounds calm and purposeful [...] Goodbye, take care.

Low-Quality (Session 111). Therapist uses nearly identical phrases (“That sounds like...”) across turns, never explores underlying issues, cycles through same technique without progression.

We compare two GPT-5-Nano sessions to study quality variation and summarize evaluation scores in [Table 11](#).

Table 11: Human and LLM evaluation scores for case study sessions.

Dimension	Sess. 769 (High)		Sess. 111 (Low)	
	Human	LLM	Human	LLM
Coherence	4	4	2	5
Depth	5	4	2	4
Progress	5	4	2	5
Naturalness	5	4	1	5
Empathy	5	5	2	5
Adherence	5	5	2	5

E Implementation and Time Complexity

StoryMI is implemented with LangGraph, a framework for building stateful, multi-agent applications as directed graphs. The workflow defines

a StateGraph with five nodes (TherapistAgent, ClientAgent, CompletionDetector, Identifier, EndSession), where the shared dialogue state flows between nodes via a typed state dictionary and conditional edges implement the branching logic of Algorithm 1. LangGraph is purely an implementation choice and does not affect the method’s generality.

Each turn in StoryMI Full involves: (1) the Interaction Agent classifying the client’s MI code and selecting a therapist strategy (1 call), (2) the Therapist Agent generating a response (1 call), (3) the Client Agent generating a response (1 call), and (4) after turn $T_{\min} = 10$, a completion check (1 call). This totals 3 calls/turn for turns 1–10 and 4 calls/turn thereafter. Table 12 shows that the overhead of macro-level control is one additional LLM call per turn, which is a modest linear cost. Future optimizations include fusing strategy selection and response generation into one call, or using a lightweight completion heuristic to skip the LLM-based check.

Table 12: Per-turn LLM call comparison. GPT = GPT-5-Nano, LLaMA = LLaMA 3.1-8B

Config.	$t \leq 10$	$t > 10$	GPT (avg 17t)	LLaMA (avg 23t)
StoryMI Full	3	4	~59	~82
w/o MI code	2	3	~42	~60

F Agent Prompt Templates

Table 13–16 present the prompt templates used for each agent in the StoryMI framework.

Table 13: Questionnaire Profiling & Story Generation

Questionnaire Profiling	
System: You are now a client seeking psychological counseling. Your basic information: {client_info}. Question list: {questions}.	
Task: For every question (exactly 23), you must: (1) Choose one integer score from 0 to 4 (0 = “Not at all”, 4 = “Almost always”) that best fits the client’s feelings. (2) Write one short explanation (1–2 sentences) reflecting the severity, as if the client were speaking.	
Constraints: The arrays must have exactly 23 elements.	
Output:	{“scores”: [s1...s23], “explanations”: [“exp1”...“exp23”]}
Situational Story Generation	
System: Based on the questionnaire screening results, write a first-person narrative (<200 words).	
Requirements: Choose one primary symptom (most severe). Focus on ONE specific scene (work, dinner, morning routine). Describe concrete actions and behaviors. Show how the symptom disrupts normal activity. Use short, direct sentences with minimal adjectives.	
Input:	Questionnaire results: {results}; User explanations: {user_response}.
Output:	Return only the story without additional text.

Table 14: Client Agent

Response Generation
System: You are a client receiving psychological counseling. This is your story/past traumatic experience: {background_story}. MI Code Definitions: {mi_codes}.
Task: Generate a response that naturally embodies the specified client MI code while maintaining consistency with your story and the conversation flow.
Client MI Codes: <i>Change</i> = language favoring behavior change (desire, ability, reasons, commitment); <i>Sustain</i> = language opposing change or favoring status quo; <i>Neutral</i> = no directional motivational content.
Constraints: Use natural, colloquial language; avoid metaphors and dramatic wording. Generate only ONE utterance per turn. Don’t start with “It seems that” or similar phrases.
Input: Conversation history: {messages}; Therapist utterance: {therapist_utterance}; Target client MI code: {client_mi_code}.
Output: Return only the client response content.

Table 15: Therapist Agent

Response Generation
System: You are an experienced psychotherapist skilled in MI techniques. {wrap_up_instruction}.
Task: Generate a response that strictly follows the selected MI code.
Constraints: Generate 1–2 utterances using casual, natural language. Do not use repetitive sentence patterns. Avoid “It seems that”, “It sounds like” phrases.
Input: MI Codes: {mi_codes}; Selected Code: {therapist_mi_code}; History: {messages}.
Output: Return only the therapist response content.

Table 16: Interaction Agent

MI Code Selection
System: You are an expert MI strategy selector.
Task: (1) Classify the MI code of the client’s last utterance (change/sustain/neutral). (2) Select the optimal MI technique the therapist should use next.
Phase-aware selection: Early = more open questions and reflections; Middle = more complex reflections; Later = information giving and advice.
Input: MI Code Definitions: {mi_codes}; Conversation History: {messages}.
Output: {“client_mi_code”: “<change sustain neutral>”, “therapist_mi_code”: “<reflection question therapist_input>”}
Session Monitor
System: You are a session monitor.
Task: Output valid JSON {“result”: “<complete continue>”, “reason”: “...”} to determine if the session should end.
Rules: Return complete if the therapist uses closing cues (e.g., “wrap up”, “goodbye”) without introducing new topics, or if the client explicitly ends. Return continue if new topics emerge, substantive questions need answers, or ending signals are ambiguous.

G Annotation Guideline and Rubrics

Table 17 summarizes the annotation guidelines template provided to human evaluators. Human annotators and the LLM evaluator used identical rubrics for the six evaluation dimensions, as summarized in Table 18.

H Ethical Considerations

This work simulates therapeutic dialogues for research purposes only and is not intended for direct clinical deployment. Real-world use of such systems would require meticulous safety validation via experienced psychotherapy practitioners with professional oversight, and compliance with regional

Table 17: Annotation guidelines template for human evaluators.

Task Overview	
Therapist	Uses MI skills (reflections, open questions) to explore and resolve ambivalence.
Client	Shares thoughts, feelings, and experiences.
Evaluation Criteria (see Table 18)	
MI Codes for Therapist	
Reflection	Mirrors back client’s expressed content.
Question	Open (elaboration) or Closed (yes/no).
Input	Information, Advice, Options, Goal-Setting.
MI Codes for Client	
Change Talk	Supports behavior change.
Sustain Talk	Resists change; favors status quo.
Neutral	No directional content.
Key Steps	
(1) Upload a JSON file to load data samples. (2) Read the MI code definitions. (3) Read the full conversation carefully. (4) Score using all criteria. (5) Click “Download CSV” when done; save as SourceFileName_AnnotatorName.csv.	
Notes	
If unable to finish all annotations at once, record the sample number to resume later. Use the provided rubrics (1–5 scale). Be objective, focus on what is present; avoid assumptions. Leave comments if something stands out.	

healthcare regulations. Potential risks include inappropriate therapeutic responses, reinforcement of harmful patterns, and user’s over-reliance on automated systems for counseling. We emphasize that AI-generated therapeutic content cannot replace qualified mental health professionals and should only be used as supplementary training or research tools under expert supervision.

I AI Usage Disclosure

AI tools were used only in a limited capacity to assist with language editing. Specifically, we use Le Chat by Mistral AI to improve the clarity and readability of the manuscript, and the refinement of the reference formatting. All scientific contributions, including research design, data collection, analysis, results, and conclusions, have been independently conducted and verified by the authors.

Criterion	Rubric (5 = Excellent, 4 = Good, 3 = Average, 2 = Poor, 1 = Very Poor)
Coherence*	<p>Is the conversation logically structured, with smooth transitions between steps?</p> <p>5 = The conversation is well-structured, with smooth transitions and a logical flow from one step to the next.</p> <p>4 = The conversation is mostly coherent, with only minor inconsistencies or abrupt transitions.</p> <p>3 = The conversation has some structural issues, with occasional jumps or awkward transitions.</p> <p>2 = The conversation lacks logical flow, with multiple abrupt or confusing transitions.</p> <p>1 = The conversation is highly disorganized, making it difficult to follow the sequence of steps.</p>
Depth*	<p>To what extent does the dialogue move beyond surface remarks to examine underlying emotions, cognitions, life history, and relational patterns, thereby demonstrating therapeutic depth?</p> <p>5 = Systematically peels back multiple layers including current emotions, bodily sensations, core beliefs, and history.</p> <p>4 = Key emotions and several cognitive or historical aspects are explored with minor gaps.</p> <p>3 = Identifies main emotions or thoughts but lacks consistent follow-up; insights stay shallow.</p> <p>2 = Discussion remains at the level of events or generic emotion labels with little inquiry into causes.</p> <p>1 = Essentially small talk; no meaningful exploration of emotion, cognition, or background.</p>
Progress*	<p>Does the conversation effectively move forward in a logical manner?</p> <p>5 = Progresses efficiently, covering each step logically and without unnecessary repetition.</p> <p>4 = Generally moves forward well, with only minor delays or repetitions.</p> <p>3 = Progresses but sometimes gets stuck or moves inefficiently.</p> <p>2 = Struggles to move forward, often repeating steps or getting sidetracked.</p> <p>1 = Lacks clear progress, frequently revisiting previous steps or failing to complete.</p>
Naturalness*	<p>Does the conversation feel fluid and human-like, avoiding robotic or overly scripted responses?</p> <p>5 = Feels natural and human-like, with engaging and varied responses.</p> <p>4 = Mostly natural, but some responses feel slightly mechanical or repetitive.</p> <p>3 = Has a mix of natural and robotic responses, with some unnatural phrasing.</p> <p>2 = Often feels artificial or scripted, with little variation in responses.</p> <p>1 = Highly robotic or formulaic, making it feel unnatural and disengaging.</p>
Empathy [†]	<p>Does the therapist convey accurate understanding and acceptance of the client's perspective?</p> <p>5 = Precisely identifies and reflects feelings, conveying deep empathy and understanding.</p> <p>4 = Clearly reflects emotions and shows empathy but lacks some nuance.</p> <p>3 = Briefly acknowledges emotion without deeper exploration or meaning expansion.</p> <p>2 = Only surface-level response or misreads the emotion.</p> <p>1 = No empathy shown, or the emotion is dismissed or ignored.</p>
Adherence [†]	<p>Does the therapist align with the predicted Motivational Interviewing (MI) skills?</p> <p>5 = Clearly applies the skill with precision and directly advances the client's change process.</p> <p>4 = Mostly applies the skill with only minor wording issues or missing nuance.</p> <p>3 = Partial demonstration; noticeable gaps or mixed elements.</p> <p>2 = Minimal or unclear use of the skill; some non-MI content present.</p> <p>1 = No MI-consistent skill evident, or the response counters MI strategies.</p>

Table 18: Evaluation rubric for LLM and human annotators. Symbols [†] and * denote therapist-focused and conversation-level criteria, respectively.