A Reinforcement Learning-Based Facilitator for Simulated Group Motivational Interviewing

Alafate Abulimiti[1], Vladislav Maraev[1,3], Agnès Helme-Guizon[4], Catherine Pelachaud[1,2]

[1]ISIR – Institut des Systèmes Intelligents et de Robotique, Sorbonne University
[2]Centre National de la Recherche Scientifique (CNRS)
[3]Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg
[4]CERAG, Grenoble IAE – INP, UGA, Université Grenoble Alpes

Emails: alafate.abulimiti@isir.upmc.fr, vladislav.maraev@gu.se
agnes.helme-guizon@univ-grenoble-alpes.fr, catherine.pelachaud@sorbonne-universite.fr

Abstract

Motivational Interviewing (MI) is a widely validated approach to behavior change, but existing virtual MI agents operate only in one-on-one settings, ignoring the cost-effectiveness and peer-support dynamics of *group* MI. We present a simulation environment and reinforcement learning (RL) based dialogue manager for group MI, in which a discrete Soft-Actor-Critic (SAC) policy selects facilitator dialogue acts and a large language model generates utterances, with two LLM-prompted participant agents as interlocutors. Our model supports adaptation to different participant profiles. We compared our dialogue manager with four LLM-based ones at the dialogue acts level. We observed that RL yields a significantly different facilitator policy, which showed the tendency to generate more directive acts and adapt to varying group compositions. Participant profile adaptation was the strongest in groups containing an open-to-change participant. We released the code in https://github.com/neuromaancer/rl-gmi-facilitator.

1 Introduction

The demand for behavioral change support services continues to outpace the capacity of healthcare systems worldwide, creating substantial barriers to timely care (Cameron et al., 2017). Motivational Interviewing (MI; Miller and Rose, 2009)—a client-centered therapeutic approach that facilitates behavioural change—is among the most widely validated interventions for substance use, diet, and exercise (Rubak et al., 2005). Virtual agents capable of delivering MI have been proposed as a scalable supplement to human facilitators (Steenstra et al., 2024; Meyer and Elsweiler, 2025), but existing systems operate exclusively in one-on-one settings. Delivering *Group Motivational Interviewing (GMI)*· offers distinct advantages: it is more cost-effective, allowing facilitators to reach more participants simultaneously, and it enables peer support dynamics—such as mutual influence on change talk and group cohesion—that serve as additional therapeutic factors absent in dyadic interaction (Velasquez et al., 2006; Yalom and Leszcz, 2020; D'Amico et al., 2015). Group cohesion is the therapeutic bond that encompasses both each member's relationship with the group as a whole and the relationships among members, and meta-analytic evidence identifies it as one of the strongest predictors of treatment outcomes in group therapy (Burlingame et al., 2011).

Recent large language models (LLMs) have enabled virtual MI agents with impressive linguistic capabilities, yet they lack built-in mechanisms for controlling session progression through therapeutic stages or adapting strategy to individual participant profiles. A promising alternative is the hybrid approach of separating dialogue act selection using reinforcement learning (RL) for the former and an LLM for the latter (Galland et al., 2025). However, this architecture has only been explored in dyadic MI with a single participant.

In this paper, we present a fully simulated group MI environment in which one facilitator agent and two participant agents interact over a structured session. The facilitator uses a hybrid architecture: a Discrete Soft-Actor-Critic (SAC) policy (Haarnoja et al., 2018; Christodoulou, 2019) selects from 14 dialogue acts, and an LLM generates utterances conditioned on the selected act. The two simulated participant agents are LLMs prompted with distinct behavioral profiles that shape their stance toward change.

Our contributions are as follows: 1) a simulation environment and RL-based dialogue manager for group MI that extends the hybrid RL+LLM architecture of Galland et al. (2025) from dyadic to triadic sessions; and 2) an empirical finding that RL yields a significantly different facilitator policy that adapts to group composition.

This paper is organized as follows. Section 2 re-

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 871–890
August 2–5, 2026.

views related work. Section 3 describes the framework, including the simulation environment, agent designs, reward function, and training procedure. Section 4 presents experimental results. Section 5 outlines conclusions and future work.

2 Related Work

Our work sits at the intersection of three research threads: AI-assisted motivational interviewing, reinforcement learning for dialogue management, and LLM-based participant simulation. We review each in turn and position our contributions relative to the state of the art. Also, We distinguish human and clinical MI/GMI studies, which motivate our therapeutic constructs, from simulation and RL studies, which motivate our training framework.

2.1 AI-Assisted Motivational Interviewing

Motivational Interviewing (MI) is a client-centered therapeutic approach that facilitates behavioral change by eliciting a client's own motivations through techniques such as open questions, affirmations, reflections, and summaries (Miller and Rose, 2009; Hettema et al., 2005). A growing body of evidence supports its effectiveness across a range of health-related domains, including substance use, smoking cessation, weight management, and physical activity (Rubak et al., 2005; Apodaca and Longabaugh, 2009).

Early computational approaches to MI relied on rule-based architectures (Cameron et al., 2017; Park et al., 2019; Prochaska et al., 2021). Large language models have since broadened the scope of MI dialogue systems, enabling virtual agents that replicate MI techniques for alcohol counseling (Steenstra et al., 2024), generate reflections that approach but do not yet match human facilitator quality (Basar et al., 2025), infer client motivational states to guide topic selection (Yang et al., 2025), and increase readiness to change in a randomized controlled trial (Meyer and Elsweiler, 2025). At the same time, Kong and Moon (2025) demonstrated that current LLMs produce and fail to detect MI-inconsistent responses, underscoring the need for structured safeguards. All of these systems operate exclusively in dyadic settings.

While these systems operate in one-on-one settings, clinical practice often delivers MI in group formats (Velasquez et al., 2006; Wagner and Ingersoll, 2012). Group MI introduces unique dynamics: peer influence on change talk (D'Amico et al., 2015), inter-member cohesion as a therapeutic factor (Yalom and Leszcz, 2020), and the facilitator's need to manage multiple participants simultaneously. The prior work on multi-agent group MI is Geng et al. (2025), who developed a multi-chatbot system (one facilitator and two peer bots) for premenstrual syndrome management and found that the group condition increased engagement and social learning compared to one-on-one chatbot interaction. However, their system relies entirely on LLM prompting without structured dialogue management or reinforcement learning for creating adaptive facilitator strategies in response to participants.

Our work addresses this gap by combining group MI facilitation with a structured RL-based dialogue manager, enabling principled stage progression and strategic adaptation across participant profiles (see section 3.1.1)—capabilities that pure prompting approaches do not provide.

2.2 Reinforcement Learning for Dialogue Management

Reinforcement learning has a long history in dialogue policy optimization, from early work on spoken dialogue systems (Su et al., 2017) to more recent applications in counseling and therapy. Srivastava et al. (2023) proposed response-act guided RL for mental health counseling, training a dialogue agent to select appropriate response acts using both content and act-level rewards. He et al. (2024) introduced a dual-process framework combining instinctive policy models with Monte Carlo tree search for goal-directed dialogue planning.

Most directly related to our work, Galland et al. (2025) proposed an RL-based dialogue manager that governs an LLM for one-on-one MI dialogues. Their system uses hierarchical reinforcement learning (HRL) with a master policy selecting MI phases and sub-policies managing phase-specific actions, combined with Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017) for adaptation across participant profiles. They demonstrated that the RL-conditioned LLM outperforms a standalone LLM baseline in accumulated reward.

Our approach shares the core insight of separating dialogue act selection (via RL) from utterance generation (via LLM), but differs in several respects. First, we address *group* MI rather than one-on-one sessions, requiring the agent acting as a facilitator to manage two participants with potentially different profiles within a single episode. Second,

we use a flat Discrete SAC architecture (Haarnoja et al., 2018; Christodoulou, 2019) rather than the hierarchical RL of Galland et al. (2025). Hierarchical RL suffers from training non-stationarity as task complexity grows (Hu et al., 2023; Pateria et al., 2021), and the multi-participant state space of group MI would amplify this risk.

2.3 Simulated Environments for Dialogue Training

Training dialogue agents through simulation is essential when real data is unavailable or ethically constrained. In mental health and therapeutic domains, Qiu and Lan (2024) used LLM-to-LLM role-playing to simulate therapeutic interactions, and De Duro et al. (2025) released a dataset of simulated mental health dialogues comparing multiple LLMs against human conversational patterns. For MI specifically, simulation work remains limited. Galland et al. (2024b) demonstrated that LLM-conditioned participant simulators produce natural and coherent utterances for MI training, Yosef et al. (2024) built an LLM-based digital patient platform for MI session assessment, and Steenstra et al. (2025) developed an LLM-powered simulated patient for MI counselor training with turn-level feedback. All address dyadic MI only. No prior work simulates group MI sessions, which is the gap our simulation environment fills.

3 Method

We present a framework for simulating group motivational interviewing sessions with one facilitator agent and two participant agents. In the absence of publicly available group MI dialogue data, simulation is the only practical route to training and evaluating such a system. The facilitator uses a hybrid architecture: a reinforcement learning policy selects dialogue acts, and an LLM generates the corresponding utterances.

3.1 Agent Design

3.1.1 Participant Agents

Each participant is an LLM prompted with a *participant type* profile that shapes its stance toward change.

Participant types: Following Galland et al. (2024a), we define three participant profiles that differ in their initial stance toward change. Each profile is realized as a natural-language persona prompt given to the LLM, which generates the participant's utterances conditioned on the persona and the conversation so far: 1) **Open-to-Change**: willing to change and open to new ideas, though change still takes effort; builds on suggestions that feel practical or personally meaningful and can voice tentative commitment earlier than the other profiles. 2) **Resistant-to-Change**: values autonomy, defends current behaviors as coping mechanisms, and reacts negatively to perceived pressure. 3) **Receptive**: begins skeptical and gradually shifts through ambivalence toward cautious openness as the session unfolds. The full participant prompts for each profile are provided in Appendix A.1.1.

Dialogue act taxonomy: participants select from 12 dialogue acts (Malhotra et al., 2021). Following the MI participant taxonomy of Malhotra et al. (2021), we additionally introduce three social acts—normalizing, encouraging, and empathizing with the other participant—adapted from the facilitator's supportive acts so that participants can perform peer support behavior without assuming the facilitator's role. The task-oriented acts carry the core MI dynamics (change talk vs. sustain talk, ambivalence, perspective shift), while the peer-supportive acts carry the peer-support dynamics that distinguish group from individual therapy. Complete definitions for each act are given in Appendix A.1.2.

3.1.2 Facilitator Agent

The facilitator operates as a two-stage pipeline: a dialogue policy selects an act based on the current observation, and an LLM generates the corresponding utterance conditioned on the selected act and the dialogue history.

Dialogue act space: The RL policy outputs one of $|\mathcal{A}_T| = 14$ the facilitator's dialogue acts (Malhotra et al., 2021). A NoAct action additionally allows the facilitator to remain silent. Complete definitions for each act are given in Appendix A.1.3.

Observation space: The observation $\mathbf{o}_t$ at time step t is a fixed-dimensional vector that captures recent dialogue history from all three speakers. It contains four components: per-speaker act histories (normalized dialogue-act indices for Participant A, Participant B, and the facilitator), per-speaker utterance embeddings computed using *all-mpnet-base-v2* (Reimers and Gurevych, 2019), a normalized turn pointer indicating session progress, and a one-hot encoding of the current GMI stage. Thus, the

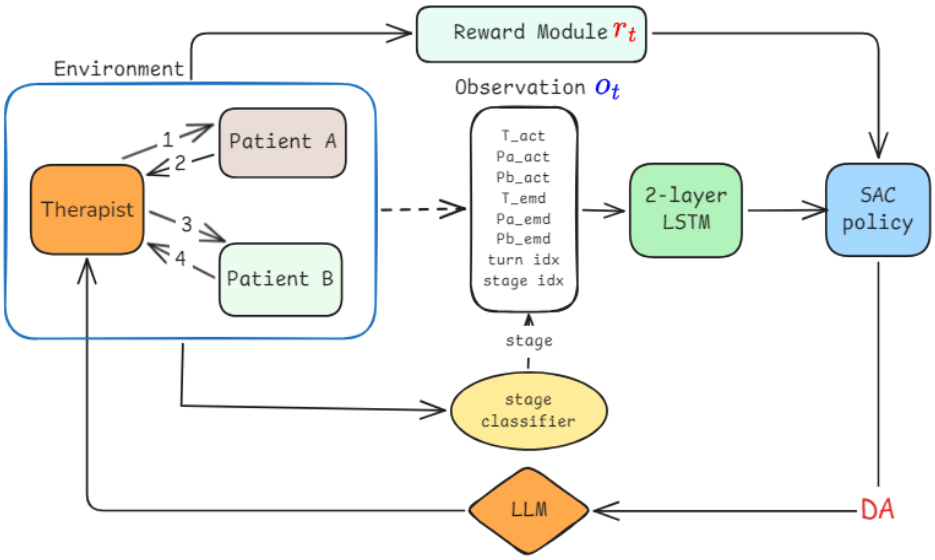

Figure 1: The facilitator and two participant agents (A, B) interact within the environment, alternating in a strict schedule ($T \to A \to T \to B \to T$..., arrows 1–4). At each step, the environment produces an observation $\mathbf{o}_t$—act histories and sentence embeddings for all three speakers, plus a turn pointer and a stage one-hot from an LLM-based stage classifier—together with a five-component reward r_t. The observation feeds a 2-layer LSTM and a Discrete SAC policy that selects one of 14 facilitator dialogue acts (DA), which an LLM-based NLG module renders into an utterance.

group setting is represented not only by an additional participant, but by separate participant histories and pair-level profile composition.

Utterance generation: Given the selected dialogue act a_t and the dialogue history, the NLG module prompts the base LLM to produce a facilitator utterance. The prompt includes the act catalogue, act-specific few-shot examples when available, and instructions to produce natural spoken-style language.

3.2 Simulation Environment

We simulate a group MI session as a sequential decision-making process in which a facilitator agent interacts with two simulated participants (participant A and participant B) over a fixed number of T=60 turns (30 for the facilitator, 15 per participant), a pragmatic pilot-based choice, balancing sufficient interaction length for simulated stage progression while remaining trackable for RL training. All sessions are themed around *dietary change*. At the start of each episode, two participant profiles are sampled from the three types described in §3.1.1, yielding six possible pairings.

Turn structure: Each environment step comprises one facilitator turn followed by one participant response, alternating between participant A and participant B across successive steps. We adopt this fixed schedule because realistic floor management in multi-party dialogue (Traum, 2004) would introduce substantial variance into RL training; a deterministic turn order keeps the learning signal focused on dialogue act selection.

Session stage tracking: We track session progression through the four phases of group MI proposed by Wagner and Ingersoll (2012): *Engaging*, *Exploring*, *Broadening*, and *Moving-to-Action*. At each step, an auxiliary LLM classifies the current stage from a window of recent dialogue. Because raw classifications are unstable, a smoothing level requires at least two subsequent predictions of the stage to commit to it (details in Appendix A.2.2).

3.3 Reward Function

The reward at each step is the sum of five components, each targeting a different aspect of effective group MI facilitation.

Participant outcome (r_{out}): Scores the participant's dialogue act on a valence scale derived from the participant taxonomy (Appendix A.1.2), ranging from strongly positive for concrete change talk to strongly negative for sustain talk (see table 6 in Appendix). Change talk and new perspectives receive additional weight, reflecting MI's central

premise that change talk is the primary pathway to behavior change (D'Amico et al., 2015).

Stage transition (r_{stage}): Wagner and Ingersoll (2012) recommend a generally sequential four-phase model for group MI, arguing that later phases flow from the successful completion of earlier ones; at the same time, they note that the progression is not fixed and that facilitators should return to earlier phases when needed. We encode both principles: the later phases are worth more, while regression incurs no penalty—accommodating both the imperfect stage estimator and the practical reality that sessions often revisit earlier phases when new concerns surface.

Strategic act (r_{strat}): Provides early-exploration guidance so the agent does not waste training on acts far from the current therapeutic phase: $+\beta$ for acts in the current stage's repertoire (Appendix A.2.1), zero for adjacent-stage acts, and $-\beta$ for all others. The magnitude is kept small so that participant outcomes remain the dominant signal.

Cohesion reward (r_{coh}): A lexicon-based signal rewards participant utterances containing cohesion-related language (e.g., agreement, or shared experience) computed as a weighted combination of categories from the SEANCE tool (Crossley et al., 2017) and scaled by γ_{coh}. This component encourages the facilitator to foster group cohesion—a therapeutic factor unique to group settings (Yalom and Leszcz, 2020).

Repetition penalty (r_{repeat}): Guards against reward hacking: consecutive uses of the same facilitator act incur an exponentially doubling penalty, making single-act exploitation rapidly unprofitable.

Total reward: Writing a_t^T, a_t^P for the facilitator and participant acts, u_t^P for the participant utterance, s_t for the committed stage index (0 for ENGAGING, increasing through MOVING-TO-ACTION), and k_t for the current facilitator-act repeat streak, the components are:

$$r_{\text{out}} = v(a_t^P), \qquad r_{\text{coh}} = \gamma_{\text{coh}} \cdot c(u_t^P)$$

$$r_{\text{stage}} = \begin{cases} \alpha \cdot \min(s_t, 3) & s_t > s_{t-1} \\ 0 & \text{else} \end{cases}$$

$$r_{\text{strat}} = \begin{cases} +\beta & a_t^T \in \mathcal{A}(s_t) \\ 0 & a_t^T \in \mathcal{A}(s_t \pm 1) \\ -\beta & \text{else} \end{cases}$$

$$r_{\text{repeat}} = \begin{cases} -\delta \cdot 2^{k_t - 2} & k_t \geq 2 \\ 0 & k_t \leq 1 \end{cases}$$

Where $v(\cdot)$ is the participant-act valence map, $\mathcal{A}(s)$ the stage-s act repertoire (Appendix A.2.1), and $c(\cdot)$ the normalized SEANCE cohesion score. The total is the unweighted sum $r_t = r_{\text{out}} + r_{\text{stage}} + r_{\text{strat}} + r_{\text{coh}} + r_{\text{repeat}}$, with relative influence set by component magnitudes rather than explicit weights. Hyperparameters are in Appendix A.2.8.

3.4 RL Training

We train the facilitator policy with Discrete SAC (Haarnoja et al., 2018; Christodoulou, 2019), whose maximum-entropy objective encourages exploration across the discrete act space, keeping all hyperparameters fixed across the four specific runs (Appendix A.2.8). The structured observation is encoded by per-channel encoders whose outputs are concatenated and passed through an LSTM, so that the policy can condition on how the participant pair's behavior evolves across turns, providing implicit profile adaptation without the meta-learning machinery. The LSTM's output is the shared representation for the discrete actor and twin Q-value critics. At each facilitator turn, the policy selects a dialogue act, the LLM realizes the act as an utterance, the two participant simulators respond, and the transition is stored in the replay buffer for off-policy SAC updates.

3.5 LLM-Only Baseline

We compare the RL facilitator against an LLM-only baseline that shares the full pipeline but replaces the trained policy with an LLM prompt for dialogue act selection; all other components are identical, so performance differences isolate the effect of RL. We evaluate four open-weight LLMs in the 7–12B parameter range—Llama 3.1 8B (Grattafiori et al., 2024), Mistral 7B (Jiang et al., 2023), Mistral Nemo 12B (Mistral AI Team, 2024), and Qwen 3.5 9B (Qwen Team, 2026)—chosen because our hybrid architecture offloads strategic act selection to the RL policy, reducing the LLM's role to conditioned utterance generation for which models in this range are adequate, and because RL training requires hundreds of LLM calls per episode, making larger models impractical. Full prompts can be found in Appendices A.2.3 and A.2.4.

3.6 Evaluation Framework

We evaluate using a three-tier framework designed to capture both surface-level quality and MI-specific effectiveness.

Automatic metrics: We report standard generation-quality and diversity scores: BARTScore (Yuan et al., 2021) as a reference-free fluency measure, and distinct-2 lexical diversity scores of Li et al. (2016), which penalize degenerate repetition in dialogue generation. These metrics characterize *how* the facilitator speaks but not *what* it accomplishes, motivating the additional measures below.

MI-specific metrics: We report four episode-level counts derived from dialogue act annotations: facilitator *reflections* (R) and *questions* (Q), and participant *change-talk* (CT) and *sustain-talk* (ST) acts. Raw counts are more stable than ratios when denominators are small. The counts are grounded: MI coding frameworks emphasize reflective responding over excessive questioning (Moyers et al., 2016), and participant change talk is more strongly associated with behavior-change outcomes than sustain talk (Apodaca and Longabaugh, 2009; Magill et al., 2014). We additionally report *facilitator act entropy* (Shannon entropy of the session's act distribution) as a measure of strategic diversity.

LLM-as-judge: An LLM judge (Llama 3.3 70B Instruct[1]) evaluates facilitator performance at two granularities. At the *turn level*, the judge scores each facilitator utterance on four criteria—appropriateness, specificity, engagement, and naturalness—using a single-pass evaluation. At the *dialogue level*, the judge evaluates complete episodes using a chain-of-thought critique-first format (Liu et al., 2023), identifying strengths and weaknesses before scoring the same four criteria. All criteria use anchored 1–5 rubrics with detailed descriptors at scores 1, 3, and 5, adapted from the MI evaluation rubric of Basar et al. (2025) and calibrated to discourage positivity bias (Zheng et al., 2023) (score 3 = "adequate baseline"). Full judge prompts and rubric descriptors are provided in Appendix A.2.5.

We rely on automatic metrics and LLM-as-judge scoring rather than human annotation. Recent work has shown that strong LLM judges can match inter-human agreement levels on open-ended dialogue evaluation when equipped with detailed criteria and anchored rubrics (Zheng et al., 2023; Bavaresco et al., 2025). We acknowledge that LLM judges can show degraded agreement in expert-knowledge domains (Szymanski et al., 2025); we mitigate this with the rubric design detailed in Appendix A.2.5 with detailed score descriptors, chain-of-thought critique-first prompting at the dialogue level, and explicit anti-positivity-bias calibration. Because our judge is not validated against human expert annotations for group MI, human expert evaluation of group MI sessions remains an important direction for future validation (see Section 5).

[1]Ranked 1st on *Judge Arena*. https://huggingface.co/spaces/AtlaAI/judge-arena

4 Results

4.1 Baseline BL and RL Comparison

Across the four paired BL–RL comparisons (Table 1), RL largely preserves response quality but shifts facilitator strategy differently across LLMs, with the largest changes in the MI-specific act counts.

RL largely preserves response quality, but not uniformly: For Llama, Mistral 7B, and Qwen, no turn-level judge criterion reaches significance, indicating that RL training does not materially degrade turn-level stylistic quality on these LLMs. The dialogue-level judge tells a more mixed story: Mistral 7B shows a significant decrease in appropriateness ($p<.01$) but a significant increase in specificity ($p<.001$), and Mistral Nemo shows significant decreases in dialogue-level appropriateness and naturalness (both $p<.001$). Llama and Qwen are unaffected at both granularities. BARTScore and distinct-2 remain numerically similar across all four LLMs, suggesting that the main quality differences are localized to judge-rated criteria rather than to surface fluency.

RL shifts MI-specific behavior, differently per LLM: The bottom block of Table 1 shows where RL has its main effect. All four LLMs increase facilitator's *reflective* acts under RL, with Llama showing the largest absolute gain (R^F: $0.12 \rightarrow 3.47$, $p<.001$) from a near-zero baseline, and Mistral Nemo reaching the highest RL reflection count (5.12, $p<.01$). Mistral Nemo also reduces its question count (Q^F: $5.38 \rightarrow 2.75$, $p<.001$), yielding the most balanced act-level tradeoff in the table. Qwen modestly increases both reflections and change talk without a significant question-count shift. Mistral 7B is the clearest outlier: RL drives change-talk counts to 16.23 ($p<.001$) while reflections, though technically rising ($0.05 \rightarrow 0.67$, $p<.001$), remain negligible. MI coding guidelines emphasize reflective listening as the core counsel-

LLM	Method	Auto		Turn-level judge				Dialogue-level judge			
		BART	D-2	App	Spe	Eng	Nat	App	Spe	Eng	Nat
Llama 3.1 8B	BL	$-3.40_{\pm 0.57}$	0.31	$4.67_{\pm 0.54}$	$4.28_{\pm 0.93}$	$4.73_{\pm 0.51}$	$4.93_{\pm 0.25}$	$3.92_{\pm 0.38}$	$3.53_{\pm 0.64}$	$4.35_{\pm 0.65}$	$3.90_{\pm 0.44}$
	RL	$-3.40_{\pm 0.66}$	0.33	$4.58_{\pm 0.78}$	$4.33_{\pm 1.11}$	$4.67_{\pm 0.87}$	$4.83_{\pm 0.73}$	$3.83_{\pm 0.45}$	$3.50_{\pm 0.56}$	$4.43_{\pm 0.62}$	$3.82_{\pm 0.39}$
Mistral 7B	BL	$-3.49_{\pm 0.69}$	0.35	$3.78_{\pm 0.75}$	$3.02_{\pm 1.32}$	$3.82_{\pm 0.97}$	$4.27_{\pm 0.96}$	$3.98_{\pm 0.13}$	$3.10_{\pm 0.44}$	$4.47_{\pm 0.50}$	$3.72_{\pm 0.45}$
	RL	$-3.45_{\pm 0.63}$	0.36	$3.95_{\pm 0.83}$	$3.45_{\pm 1.20}$	$3.95_{\pm 1.06}$	$4.53_{\pm 0.78}$	$3.72_{\pm 0.61}^{\downarrow **}$	$3.47_{\pm 0.53}^{\uparrow ***}$	$4.50_{\pm 0.50}$	$3.55_{\pm 0.53}$
Mistral Nemo 12B	BL	$-3.83_{\pm 0.84}$	0.42	$4.02_{\pm 0.83}$	$3.53_{\pm 1.35}$	$3.97_{\pm 1.17}$	$4.70_{\pm 0.64}$	$3.98_{\pm 0.13}$	$3.58_{\pm 0.49}$	$4.53_{\pm 0.50}$	$3.95_{\pm 0.22}$
	RL	$-3.92_{\pm 0.98}$	0.44	$3.90_{\pm 1.03}$	$3.57_{\pm 1.46}$	$3.83_{\pm 1.33}$	$4.50_{\pm 0.97}$	$3.77_{\pm 0.46}^{\downarrow ***}$	$3.52_{\pm 0.53}$	$4.42_{\pm 0.53}$	$3.67_{\pm 0.47}^{\downarrow ***}$
Qwen 3.5 9B	BL	$-3.96_{\pm 0.56}$	0.38	$4.65_{\pm 0.57}$	$4.65_{\pm 0.75}$	$4.73_{\pm 0.60}$	$4.95_{\pm 0.22}$	$4.00_{\pm 0.00}$	$3.68_{\pm 0.47}$	$4.15_{\pm 0.60}$	$3.83_{\pm 0.37}$
	RL	$-3.89_{\pm 0.70}$	0.39	$4.43_{\pm 0.82}$	$4.50_{\pm 0.89}$	$4.68_{\pm 0.76}$	$4.82_{\pm 0.62}$	$3.98_{\pm 0.13}$	$3.57_{\pm 0.53}$	$4.33_{\pm 0.51}$	$3.82_{\pm 0.39}$

LLM	Method	Facilitator			Participant		JSD
		R^F	Q^F	Ent.	CT^P	ST^P	
Llama 3.1 8B	BL	$0.12_{\pm 0.32}$	$3.82_{\pm 1.07}$	2.82	$0.83_{\pm 1.38}$	$1.83_{\pm 2.65}$	0.246^{***}
	RL	$3.47_{\pm 1.85}^{\uparrow ***}$	$4.87_{\pm 2.38}^{\uparrow **}$	3.58	$0.63_{\pm 1.06}$	$2.25_{\pm 3.23}$	
Mistral 7B	BL	$0.05_{\pm 0.22}$	$0.65_{\pm 0.66}$	2.32	$10.38_{\pm 5.10}$	$0.80_{\pm 1.22}$	0.417^{***}
	RL	$0.67_{\pm 0.82}^{\uparrow ***}$	$5.05_{\pm 2.13}^{\uparrow ***}$	2.73	$16.23_{\pm 5.22}^{\uparrow ***}$	$1.13_{\pm 1.56}$	
Mistral Nemo 12B	BL	$4.07_{\pm 0.92}$	$5.38_{\pm 1.44}$	3.35	$4.37_{\pm 4.99}$	$1.87_{\pm 2.04}$	0.095^{***}
	RL	$5.12_{\pm 2.53}^{\uparrow **}$	$2.75_{\pm 2.08}^{\downarrow ***}$	3.51	$4.90_{\pm 5.90}$	$2.02_{\pm 2.14}$	
Qwen 3.5 9B	BL	$3.30_{\pm 1.49}$	$4.90_{\pm 0.60}$	3.16	$3.93_{\pm 3.58}$	$6.67_{\pm 6.77}$	0.148^{***}
	RL	$4.23_{\pm 2.42}^{\uparrow **}$	$5.68_{\pm 4.14}$	3.50	$5.90_{\pm 5.97}^{\uparrow ***}$	$6.23_{\pm 6.40}$	

Table 1: BL vs. RL facilitator profiles across four LLMs (Llama 3.1 8B, Mistral 7B, Mistral Nemo 12B, Qwen 3.5 9B). **Top**: automatic metrics (BART, distinct-2) and LLM-as-judge scores (1–5) — App (appropriateness), Spe (specificity), Eng (engagement), Nat (naturalness). **Bottom**: episode-mean act counts and policy divergence. *Facilitator*: R^F (reflections), Q^F (questions), Ent. (act entropy). *Participant*: CT^P (change talk), ST^P (sustain talk). JSD: Jensen–Shannon divergence between BL and RL act distributions per LLM. Significance from pair-blocked permutation tests over 60 episodes; ↑/↓ mark RL>/<BL. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. D-2, Ent. descriptive.

ing skill distinguishing MI from more directive approaches (Moyers et al., 2016), so a policy that elicits heavy change talk with almost no reflections is difficult to interpret as MI-consistent; whether this extreme CT count reflects genuine elicitation or a narrowing artifact of the RL policy warrants further inspection. Act entropy increases descriptively across all four LLMs, most substantially for Llama ($2.82 \rightarrow 3.58$), consistent with broader use of the act repertoire under RL.

LLM heterogeneity persists under RL: The four LLMs do not converge to a common RL policy; rather, each learns a different redistribution of facilitator strategy shaped by its baseline priors. This LLM-level variation motivates the finer-grained facilitator behavior analysis that follows.

4.2 Facilitator's Simulated Behavior Analysis

To assess whether RL changes facilitator adaptation, we analyze facilitator behavior as a distribution over dialogue acts. We compare baseline and RL policies at three levels: global, pair-conditioned, and recipient-conditioned. At each level, we compared facilitator's dialogue act distribution across different LLMs for different participants' profile; that is, each of the four baseline LLMs contributes one profile and each of the four RL LLMs contributes one profile. We then use Jensen-Shannon divergence (JSD) to quantify the difference between the baseline and RL act distributions. Statistical significance is assessed with exact label permutations over these eight run profiles (four baselines, four RLs), rather than over individual turns, which would overstate significance by ignoring the dependence structure within each model run. Supporting plots are provided in Appendix A.3.

4.2.1 Global and Pair-Conditioned Policy Shift

Baseline and RL facilitator act distributions differ significantly (JSD $= 0.166$, $p < .001$), and the divergence holds for every LLM ($p < .001$; Table 1), ranging from 0.095 (Mistral Nemo) to 0.417 (Mistral 7B). RL decreases mean value for exploratory acts — INVITE_SHIFT (-0.098), BACKCHANNEL (-0.093), ASK_INFO (-0.056) — and toward directive ones: GIVE_SOLUTION ($+0.125$), ASK_CONSENT ($+0.100$), PLAN_WITH_PARTICIPANT ($+0.062$) (Figure 3 in Appendix). The policy is systemati-

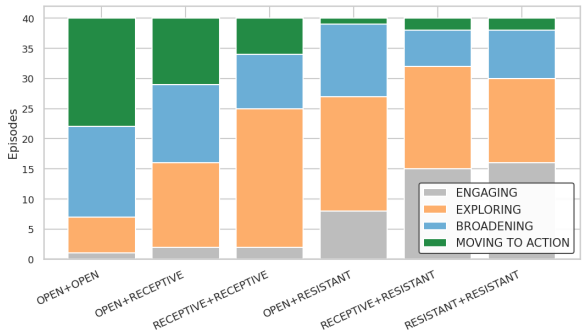

Figure 2: Pair-level final-stage distribution for the RL runs across four LLMs (40 episodes per participant-pair type, 240 episodes total). Pairs containing more ready participants more often end in later stages than resistant-heavy pairs.

cally different from the baseline, and the size of that difference varies by LLM.

The pair-conditioned difference is strongest in groups containing at least one OPEN participant: OPEN+OPEN (JSD $= 0.193$, $p < .05$), OPEN+RESISTANT (JSD $= 0.159$, $p < .05$), and OPEN+RECEPTIVE (JSD $= 0.145$, $p < .05$). Pairs without an Open participant show smaller, non-significant divergences: RECEPTIVE+RECEPTIVE (0.105, $p = 0.114$), RECEPTIVE+RESISTANT (0.117, $p = 0.114$), and RESISTANT+RESISTANT (0.119, $p = 0.200$). Across all the pairs, RL increases GIVE_SOLUTION ($\Delta = +0.054$ to $+0.146$) and ASK_CONSENT ($\Delta = +0.066$ to $+0.079$), while reducing INVITE_SHIFT ($\Delta = -0.088$ to -0.133) and BACKCHANNEL ($\Delta = -0.068$ to -0.108)[2] (Table 10 in Appendix A.3).

4.2.2 Stage Progression

These behavioral shifts co-occur with systematic differences in pair-level final-stage outcomes under RL. As shown in Figure 2, dialogues involving more open-to-change participants tend to end in later stages than those dominated by resistant participants. OPEN+OPEN pairs show the deepest final-stage distribution, with the largest share of episodes ending in MOVING_TO_ACTION. By contrast, RESISTANT+RESISTANT pairs tend to end in earlier stages. Mixed pairs occupy the expected middle ground: OPEN+RECEPTIVE often ends in BROADENING and sometimes in MOVING_TO_ACTION, whereas OPEN+RESISTANT and RECEPTIVE+RESISTANT more often end in EXPLORING or BROADENING. These results suggest that RL learns a pair-sensitive facilitator policy whose behavioral adjustments are associated with deeper progression in groups that contain greater readiness to engage.

[2]Unlike the global act-shift summary, which reports a single pooled RL-minus-BL difference per act, the values here are pair-conditioned ranges: for each act, we compute the RL-minus-BL difference separately within each of the six participant-pair types and report the minimum-to-maximum span across those pair-specific deltas.

5 Conclusion & Future Work

We presented a simulation environment and RL-based dialogue manager that extends the hybrid RL+LLM architecture from dyadic to triadic MI sessions with one facilitator and two participants. Across four open-weight LLMs, RL produces a facilitator policy that is measurably different from the LLM-only baseline, shifting from exploratory acts toward directive ones while preserving turn-level response quality. The difference is strongest in groups containing at least one open-to-change participant and observed with deeper stage progression in those groups. The learned adaptation operates at the group-composition level. All participants in this work are simulated LLMs, so validating the framework with human participants is the most pressing next step. Group MI derives much of its therapeutic value from peer influence on change talk (D'Amico et al., 2015) and inter-member cohesion (Yalom and Leszcz, 2020); while our reward function includes a lexicon-based cohesion signal, directly measuring and optimizing for these group-specific dynamics remains an open task. Other directions include: replacing the fixed turn schedule with natural floor management—a core facilitator skill when managing multiple participants simultaneously, scaling to larger groups and additional health domains.

Limitations & Ethical Considerations

Our framework is evaluated entirely in simulation. Both participants are LLM-prompted agents whose behavior may be more predictable and cooperative than that of real people — disengagement, emotional escalation, and resistance patterns are difficult to elicit reliably from prompted LLMs, while the dialogue is also text-based rather than spoken. Our evaluation uses an LLM judge rather than human annotation. We mitigate known LLM-judge biases (Szymanski et al., 2025) through anchored 1–5 rubrics, explicit anti-positivity-bias calibration, and chain-of-thought critique-first scoring at the dialogue level (Appendix A.2.5). Human expert

validation remains an important next step. The deterministic turn schedule prevents the facilitator from learning floor-management skills such as redirecting attention, handling cross-talk or interruptions, or yielding the floor strategically, though the NOACT action partially compensates by allowing intentional silence. Behavior change is a sensitive domain where poorly timed interventions can increase resistance or cause distress (Kong and Moon, 2025). Although our system operates only in simulation, the long-term goal of deploying such agents with real participants demands caution: the gap between simulated and real therapeutic interaction is not yet understood, and RL-trained policies optimized against simulated participants may not transfer safely to real clinical settings.

Acknowledgments

We thank SIGDIAL anonymous reviewers for their valuable feedback. The authors are supported by a French government grant managed by the Agence Nationale de la Recherche as part of the France 2030 program, reference ANR-22-EXEN-0004 (PEPR eNSEMBLE / MATCHING). Vladislav Maraev was supported by Swedish Research Council (VR) grant 2023-00358 – Social laughter for virtual agents (SocLaVA).

References

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.

Erkan Basar, Xin Sun, Iris Hendrickx, Jan de Wit, Tibor Bosse, Gert-Jan De Bruijn, Jos A. Bosch, and Emiel Krahmer. 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, Abu Dhabi, UAE. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, and 1 others. 2025. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255.

Gary M. Burlingame, Debra Theobald McClendon, and Jennifer Alonso. 2011. Cohesion in group therapy. *Psychotherapy*, 48(1):34–42.

Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2017. Towards a chatbot for digital counselling. In *Proceedings of the 31st International BCS Human Computer Interaction Conference (HCI 2017)*.

Petros Christodoulou. 2019. Soft actor-critic for discrete action settings.

Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2017. Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, 49(3):803–821.

Elizabeth J. D'Amico, Jon M. Houck, Sarah B. Hunter, Jeremy N. V. Miles, Karen Chan Osilla, and Brett A. Ewing. 2015. Group motivational interviewing for adolescents: Change talk and alcohol and marijuana outcomes. *Journal of Consulting and Clinical Psychology*, 83(1):68–80.

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.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1126–1135. PMLR.

Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2024a. EMMI – empathic multimodal motivational interviews dataset: Analyses and annotations. *Preprint*, arXiv:2406.16478.

Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2024b. Simulating patient oral dialogues: A study on naturalness and coherence of conditioned large language models. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents*, pages 1–4, GLASGOW United Kingdom. ACM.

Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2025. Tailored conversations beyond LLMs: A RL-based dialogue manager. *Preprint*, arXiv:2506.19652.

Shixian Geng, Remi Inayoshi, Chi-Lan Yang, Zefan Sramek, Yuya Umeda, Chiaki Kasahara, Arissa J. Sato, Simo Hosio, and Koji Yatani. 2025. Beyond the dialogue: Multi-chatbot group motivational interviewing for premenstrual syndrome (PMS) management. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–18, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR.

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen, and Bing Qin. 2024. Planning like human: A dual-process framework for dialogue planning. *Preprint*, arXiv:2406.05374.

Jennifer Hettema, Julie Steele, and William R. Miller. 2005. Motivational interviewing. *Annual Review of Clinical Psychology*, 1(1):91–111.

Wenning Hu, Hongbin Wang, Ming He, and Nianbin Wang. 2023. Uncertainty-aware hierarchical reinforcement learning for long-horizon tasks. *Applied Intelligence*, 53(23):28555–28569.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego De Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Haein Kong and Seonghyeon Moon. 2025. When LLM therapists become salespeople: Evaluating large language models for ethical motivational interviewing. *Preprint*, arXiv:2503.23566.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 110–119.

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. *Preprint*, arXiv:2303.16634.

Molly Magill, Jacques Gaume, Timothy R Apodaca, Justin Walthers, Nadine R Mastroleo, Brian Borsari, and Richard Longabaugh. 2014. The technical hypothesis of motivational interviewing: A meta-analysis of mi's key causal model. *Journal of consulting and clinical psychology*, 82(6):973.

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. *Preprint*, arXiv:2111.06647.

Selina Meyer and David Elsweiler. 2025. LLM-based conversational agents for behaviour change support: A randomised controlled trial examining efficacy, safety, and the role of user behaviour. *International Journal of Human-Computer Studies*, 200:103514.

William R. Miller and Gary S. Rose. 2009. Toward a theory of motivational interviewing. *American Psychologist*, 64(6):527–537.

Mistral AI Team. 2024. Mistral NeMo.

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.

SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. 2019. Designing a chatbot for a brief motivational interview on stress management: qualitative case study. *Journal of medical Internet research*, 21(4):e12231.

Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. 2021. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys*, 54(5):1–35.

Judith J. Prochaska, Erin A. Vogel, Amy Chieng, Matthew Kendra, Michael Baiocchi, Sarah Pajarito, and Athena Robinson. 2021. A therapeutic relational agent for reducing problematic substance use (Woebot): Development and usability study. *Journal of Medical Internet Research*, 23(3):e24850.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing LLM-to-LLM interactions. *Preprint*, arXiv:2408.15787.

Qwen Team. 2026. Qwen3.5-9B.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sune Rubak, Annelli Sandbæk, Torsten Lauritzen, and Bo Christensen. 2005. Motivational interviewing: a systematic review and meta-analysis. *The British journal of general practice*, 55(513):305.

Aseem Srivastava, Ishan Pandey, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Response-act guided reinforced dialogue generation for mental health counseling. In *Proceedings of the ACM Web Conference 2023*, pages 1118–1129, Austin TX USA. ACM.

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*, IVA '24, pages 1–10, New York, NY, USA. Association for Computing Machinery.

Ian Steenstra, Farnaz Nouraei, and Timothy Bickmore. 2025. Scaffolding empathy: Training counselors with simulated patients and utterance-level performance visualizations. In *Proceedings of the 2025 chi conference on human factors in computing systems*, pages 1–22.

Pei-Hao Su, Pawel Budzianowski, Stefan Ultes, Milica Gasic, and Steve Young. 2017. Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. *Preprint*, arXiv:1707.00130.

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th international conference on intelligent user interfaces*, pages 952–966.

David Traum. 2004. Issues in multiparty dialogues. In Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, and Frank Dignum, editors, *Advances in Agent Communication*, volume 2922, pages 201–211. Springer Berlin Heidelberg, Berlin, Heidelberg.

Mary M. Velasquez, Nanette S. Stephens, and Karen Ingersoll. 2006. Motivational interviewing in groups. *Journal of Groups in Addiction & Recovery*, 1(1):27–50.

Christopher C Wagner and Karen S Ingersoll. 2012. Motivational interviewing in groups.

Irvin D. Yalom and Molyn Leszcz. 2020. *The Theory and Practice of Group Psychotherapy*, 6th edition edition. Basic Books, New York.

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Kit Phey Leng, Nicholas Gabriel Lim, Cameron Tan Shi Ern, and Ee-peng Lim. 2025. CAMI: A counselor agent supporting motivational interviewing through state inference and topic exploration. *Preprint*, arXiv:2502.02807.

Stav Yosef, Moreah Zisquit, Ben Cohen, Anat Klomek Brunstein, Kfir Bar, and Doron Friedman. 2024. Assessing motivational interviewing sessions with ai-generated patient simulations. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 1–11.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Appendix

A.1 Agent Design

A.1.1 Participant Profile Prompts

Each participant agent is prompted with a system message assembled from five components: a shared base instruction, a profile-specific persona (see table 2), the session theme, the current turn index, and the catalogue of allowed dialogue acts (Appendix A.1.2). Only the first two are profile-dependent; they are reproduced verbatim below. The remaining three are set at episode start.

Shared base instruction. Prepended to every profile:

> You are a realistic therapy participant. Do NOT just repeat your stance on change. Share your feelings, personal stories, and struggles. Be conversational, not robotic.

A.1.2 Participant Dialogue Acts

Participant agents select one of twelve dialogue acts on each turn. Table 4 lists each act together with the short definition shown to the LLM in the prompt catalogue. The first nine are standard MI-related participant acts adapted from prior work (Galland et al., 2024a); the last three are peer-supportive acts introduced in our framework to support inter-participant interaction, which is absent in dyadic systems.

A.1.3 Facilitator Dialogue Acts

The facilitator agent selects one of fourteen dialogue acts on each turn. Table 5 lists each act together with the short description shown to the language model when generating utterances. The taxonomy covers the standard MI micro-skills—reflection, open questioning, information-giving, and planning—extended with social and meta-dialogue acts required for group facilitation, including a dedicated silence act (NoAct) that lets the facilitator remain silent.

A.1.4 Participant Act Valence Map

Table 6 shows the valence map used by the participant outcome reward r_{out} (§3.3). Values range from $+1.0$ for concrete change talk to -1.0 for sustain talk, with peer-supportive acts receiving small positive values to encourage group-specific interaction.

A.2 Environment & Training

A.2.1 Stage-Specific Act Repertoires

Table 7 lists the facilitator dialogue acts considered appropriate for each MI stage. These repertoires are used by the strategic act reward r_{strat} (§3.3): at each step, the agent receives one $+\beta$ if its selected act belongs to the current stage's repertoire, zero if it belongs to an adjacent stage, and $-\beta$ otherwise. The mapping is informed by standard MI clinical guidance: engaging and exploring stages emphasize rapport-building and reflective listening, broadening introduces consent-seeking and perspective-shifting, and moving-to-action opens up planning and concrete solutions while retaining reflection as a stabilizing act.

A.2.2 Stage Transition Smoothing

The LLM-based stage classifier can produce noisy predictions, so a smoothing layer mediates between the raw output and the committed stage. The smoother maintains a rolling buffer of the $H{=}5$ most recent raw predictions and enforces the following constraints before committing a transition:

1. **Stepwise projection:** At most one stage change per commit ($|\Delta|_{\max}{=}1$); larger jumps are projected to the nearest single-step neighbor.
2. **Minimum support:** The target must appear at least $s_{\min}{=}3$ times in the buffer ($s_{\min}{=}4$ for multi-stage jumps).
3. **Recency:** The target must also appear in the most recent $r{=}2$ buffer entries.
4. **Minimum dwell:** The agent must have remained in the current stage for at least $d_{\min}{=}2$ turns.
5. **Regression penalty:** Backward transitions require $s_{\min}{+}1$ additional support.

If all conditions are met, the projected target becomes the new committed stage; otherwise the committed stage is unchanged and the target is held as pending.

A.2.3 LLM-Only Baseline Prompt

In the LLM-only baseline (§3.5), the facilitator's dialogue act is selected by an LLM call rather than by a trained policy. The selector is given the system prompt below, and is asked to return a JSON object naming a single act from the catalogue. The chosen act is then passed to the same NLG module used

Profile	Personality	Key Behavior
Open-to-Change	You are 'Open to Change'. You are willing to change your behavior or way of thinking, and you are open to new ideas or suggestions. Change still takes effort, so stay realistic, but do NOT default to resistance. If a suggestion seems practical or personally meaningful, you may build on it rather than immediately pushing it away.	You can express new perspectives and tentative commitment to change earlier than other participant types. Small, concrete next steps are appropriate even before full confidence, as long as they sound plausible for you.
Resistant-to-Change	You are 'Resistant to Change' but NOT hostile without cause. You value your autonomy and hate being told what to do. You believe your behavior helps you cope with stress. You feel annoyed when pushed, but you might open up if the facilitator validates your feelings.	Defend your choices (sustain talk), but also explain *why* (personal information) and express annoyance (negative feeling) if pushed.
Receptive	You are 'Receptive' (evolving). Current state: ⟨*descriptor*⟩†. You fluctuate between wanting to change and staying the same.	Ask questions, express doubt, and slowly warm up to the group.

Table 2: Profile-specific persona and key-behavior strings used to prompt each participant agent, reproduced verbatim. All profiles additionally receive the shared base instruction above. †The Receptive descriptor is turn-gated; see Table 3.

Turn range	Descriptor string
$n/T < 1/3$	Skeptical and hesitant. You are not sure if you belong here.
$1/3 \leq n/T < 2/3$	Curious but conflicted. You see the pros and cons of changing.
$n/T \geq 2/3$	Cautiously optimistic. You are willing to try small steps.

Table 3: Turn-gated descriptor strings for the Receptive profile. The current descriptor is spliced into the Personality field of Table 2 at each turn, letting the Receptive persona shift stance across the session without hidden numerical state.

by the RL condition, so that differences between conditions reflect act selection rather than utterance generation.

System prompt (verbatim).

> You are a Group Motivational Interviewing (GMI) facilitator. Your spirit is a Partnership, built on Acceptance, Compassion, and Evocation.
>
> **YOUR PERSONA**
>
> - You are choosing the NEXT facilitator dialogue act, not writing the full response.
> - Group: The group is small: only "participant A" and "participant B".
> - Focus: Focus on positives, solutions, and the future more than on problems or the past.
>
> **AVAILABLE DIALOGUE ACTS**
> *Inserted at runtime as a bulleted list of* `act name: description` *pairs drawn from the catalogue in Appendix A.1.3; the* NOACT *option is excluded from this list.*
>
> **INSTRUCTIONS**
>
> 1. Read the conversation context.
> 2. Choose the SINGLE most appropriate dialogue act from the list above.
> 3. GOAL: Elicit "Change Talk" (reasons/desire to change) and reduce "Sustain Talk" (reasons to stay the same).
> 4. Prefer acts that move the session forward without becoming formulaic.
> 5. If the session is just starting, select "Greeting or Closing".
> 6. Be aware that the session may progress through Engaging, Exploring, Broadening, and Moving to Action.
>
> You MUST reply with ONLY a JSON object:
>
> ```
> {
>   "act": "<exact act name>"
> }
> ```

A.2.4 Facilitator NLG Prompt

Once a dialogue act a_t has been selected—by the RL policy or by the LLM act selector (Appendix A.2.3) in the baseline—the same NLG module renders it into a natural-language utterance. Sharing a single NLG module across conditions ensures that observed differences reflect act selection rather than utterance generation.

System prompt template: The system message is assembled at runtime from a base prompt, the session theme, the act catalogue, optional act-specific examples, and a fixed block of style and response rules:

> You are a facilitator utterance generator for a group motivational interviewing session.
>
> Theme: ⟨*theme*⟩.
>
> Allowed acts (for reference):
> *Bulleted list of* `act name: description` *for the 13 acts in Appendix A.1.3 excluding* NOACT.
>
> *Optional act-specific examples:* If a file `⟨short_name⟩.txt` exists for the selected act, its

Act	Description
Standard MI-related acts	
Personal Information	Describes past events or shares personal or medical information.
Change Talk (future)	Expresses a concrete intention to change an unhealthy behavior, including a specific action, a timeframe or trigger, and a personal reason that makes it meaningful. Hollow agreement (e.g. "I'll give it a try") does not qualify.
Sustain Talk (future)	Defends the status quo or expresses intention to keep an unhealthy behaviour.
Positive Feeling	Expresses a positive emotion or sense of relief.
Negative Feeling	Expresses a negative emotion or distress.
New Perspective	Shows a new understanding or perspective shift.
Greeting / Closing	Opens or closes the interaction.
Backchannel	Brief acknowledgment such as "yeah" or "mm-hm".
Ask Medical Information	Asks the facilitator for medical or therapeutic information.
Social acts (group-specific)	
Normalize Experience	Normalizes another participant's experience; reassures through commonality.
Encourage Progress	Acknowledges another participant's effort or progress and offers encouragement.
Empathic Reaction	Explicitly expresses empathy or compassion toward another participant.

Table 4: The 12 participant dialogue acts used by the participant agents. The three social acts at the bottom enable inter-participant interaction, a defining feature of group MI.

Act	Description
Task-oriented acts	
Ask for Consent	Requests the participant's permission or validation before proceeding.
Reflection	Conveys understanding of the participant's statement without judging, interpreting, or advising.
Ask for Information	Seeks further details about events, statements, or background.
Ask about Emotion	Asks the participant to identify or rate a current emotion.
Invite to Shift Outlook	Offers a fresh perspective or invites the participant to imagine alternatives.
Medical Education and Guidance	Provides psychological, medical, or behavioral information.
Give Solution	Offers a direct, concrete suggestion.
Planning with the participant	Co-creates a specific action plan with the participant.
Social and meta-dialogue acts	
Empathic Reaction	Expresses empathy or compassion explicitly.
Greeting / Closing	Opens or closes the session politely.
Backchannel	Brief acknowledgment such as "mm-hm" or "I see".
Normalization/Reassurance	Normalizes the participant's experience and reassures by commonality.
Progress Acknowledgment/Encouragement	Recognizes effort or progress and encourages continued engagement.
NoAct	Intentional silence: the facilitator stays silent.

Table 5: The fourteen facilitator dialogue acts used by the facilitator agent. Acts are split into task-oriented (information-gathering, perspective-shifting, education, and planning) and social/meta-dialogue (affective, relational, and floor-management). The RL policy selects from the full flat 14-way action space; the grouping is for readability only. The NoAct option is specific to group settings, enabling the facilitator to cede the floor to inter-participant interaction when appropriate.

participant Act	Base valence
CHANGE_FUTURE	+1.00
NEW_PERSPECT	+0.50
POS_FEELING	+0.25
participant_ENCOURAGE	+0.25
participant_NORMALIZE	+0.15
participant_EMPATHY	+0.15
PERSONAL_INFO	+0.10
BACKCHANNEL	0.00
GREET_CLOSE	0.00
ASK_MED_INFO	0.00
NEG_FEELING	−0.10
SUSTAIN_FUTURE	−1.00

Table 6: MI-valence assigned to each participant dialogue act. Positive values indicate therapeutic progress; negative values indicate resistance.

> contents are inserted under the header "Examples for ⟨*act*⟩" followed by the instruction "These examples show the FUNCTION of this act. Do NOT copy their wording or sentence openers. Use your own words."
>
> Speak in oral style with brief sentences, natural hesitations (. . . or —) when appropriate, and empathy. Avoid multi-clause paragraphs.
>
> There are exactly two group members. Only say 'participant A' or 'participant B' when clarity truly requires it; otherwise speak naturally to the group or to the person who just spoke.
>
> Your required act now: ⟨*selected act*⟩.
>
> RESPONSE RULES: RESPOND to what the participant just said. Do NOT lecture or monologue. Ask about THEIR experience, feelings, or plans — not about facts or science. Hard limits: one utterance, 4–80 words; no JSON; no meta commentary.

A.2.5 LLM-as-Judge Prompts

We use a two-tier LLM-as-judge strategy: turn-level scoring without chain-of-thought for throughput, and dialogue-level scoring with a mandatory critique-before-scoring step for depth. Both tiers include explicit anti-positivity-bias calibration (score 3 = adequate baseline). The judge is Llama 3.3 70B Instruct (AWQ-INT4 quantized).

A.2.6 Turn-Level Prompt

The turn-level judge scores a single facilitator utterance on four criteria. The system prompt instructs the judge to use the full 1–5 scale and provides anchored descriptors at scores 1, 3, and 5 for each criterion. Table 8 reproduces these anchors.

The user message provides the speaker identity, the preceding conversation context, and the response to evaluate, and asks for a flat JSON object with one score and rationale per criterion.

Stage	**Appropriate facilitator acts**
Engaging	Greeting/Closing, Reflection, Ask for Information, Normalisation/Reassurance, Empathic Reaction, Backchannel
Exploring	Reflection, Ask for Information, Ask about Emotion, Invite to Shift Outlook, Empathic Reaction, Backchannel
Broadening	Reflection, Invite to Shift Outlook, Progress Acknowledgment/Encouragement, Ask for Consent, Empathic Reaction
Moving to Action	Planning with the participant, Give Solution, Ask for Consent, Progress Acknowledgment/Encouragement, Reflection, Greeting/Closing

Table 7: Stage-specific facilitator act repertoires used by the strategic act reward r_{strat}. Acts listed under each stage are considered in-repertoire for that stage; all other acts from the 14-act facilitator taxonomy (Appendix A.1.3) are out-of-repertoire. Acts that appear in multiple stages (e.g., Reflection, Empathic Reaction) are those that remain clinically appropriate across phases of MI. The repertoires were derived from clinical MI guidance and refined during pilot runs.

A.2.7 Dialogue-Level Prompt

The dialogue-level judge evaluates a complete session on five criteria: the same four as the turn level. The system prompt follows a critique-first format: the judge must list three strengths and three weaknesses before scoring, following the chain-of-thought approach of Liu et al. (2023). Table 9 shows the dialogue-level anchors.

The dialogue-level judge returns a JSON object containing a `critique` field (three strengths, three weaknesses) and a `scores` field with one score and rationale per criterion. Both the critique and scores are extracted during evaluation; only the scores are used in the quantitative analysis reported in the main text.

A.2.8 Implementation Details

All RL experiments use a discrete Soft Actor–Critic (SAC) agent implemented with Tianshou[3]. The agent interacts with the group-therapy environment, in which one facilitator responds to two simulated participants. We train a separate RL policy on top of each LLM. For the main pipeline runs, each episode contains 60 turns under the theme *Diet change*.

We use four LLMs for both baseline generation and RL rollouts:

[3] https://github.com/thu-ml/tianshou

Criterion	Score 1	Score 3	Score 5
Appropriateness	Clearly inappropriate: lectures, moralizes, argues, gives unsolicited advice, dismisses feelings, or uses MI-inconsistent techniques.	Generally safe and non-harmful, but generic — could appear in any therapy context without being tailored to MI or to what the participant just said.	Precisely attuned to the participant's emotional state and group dynamic. Applies MI-consistent technique that fits the specific moment and advances the therapeutic goal.
Specificity	Completely generic, could be said to any participant in any session.	References the general topic but misses specific words, emotions, or details.	Directly echoes or builds on the participant's specific words, feelings, or situation.
Engagement	A dead end: flat acknowledgment, closed statement, or monologue that leaves no natural opening.	Keeps the conversation going but in a predictable, formulaic way.	Opens a specific, inviting thread the participant would naturally want to follow. May connect to values, leverage group dynamics, or pose a thought-provoking question.
Naturalness	Clearly machine-generated: stilted phrasing, unnatural sentence structure, or robotic repetition.	Grammatically correct and not obviously robotic, but follows a predictable template.	Genuinely conversational — natural rhythm, appropriate brevity, may include conversational markers.

Table 8: Anchored score descriptors for the four turn-level judge criteria. Scores 2 and 4 are defined as "between" the adjacent anchors. The system prompt instructs the judge that score 3 is the expected baseline for competent but unremarkable responses.

Criterion	Score 1	Score 3	Score 5
Appropriateness	Multiple MI-inconsistent behaviors (confrontation, unsolicited advice, arguing).	Generally MI-consistent but occasionally slips into advice-giving or leading questions.	Consistently MI-consistent; demonstrates the spirit of MI (partnership, evocation, compassion).
Specificity	Responses are generic templates that could apply to any participant.	Some responses reference participant content, but many are interchangeable.	Nearly every response builds on specific participant words, feelings, or experiences.
Engagement	One or both participants disengage; the facilitator fails to re-engage them.	Both participants participate, but engagement is surface-level or one-sided.	Both participants are actively involved; the facilitator skillfully balances attention between them.
Naturalness	Clearly robotic or formulaic — every response follows the same template.	Sounds plausible but somewhat stilted; a trained observer would notice repetitive patterns.	Sounds genuinely conversational; varied language, natural rhythm, appropriate brevity.

Table 9: Anchored score descriptors for the five dialogue-level judge criteria. The score-distribution guidance instructs the judge that score 3 is the expected baseline, score 4 requires genuine MI skill, and score 5 is reserved for exemplary sessions.

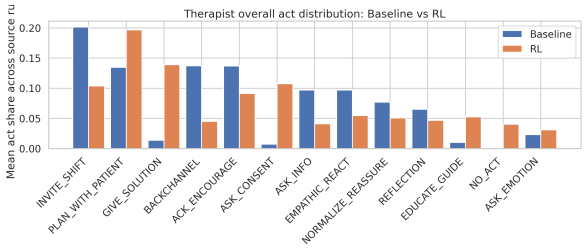

Figure 3: Global facilitator policy shift from the LLM-only baseline to RL. RL increases directive acts such as GIVE_SOLUTION, ASK_CONSENT, and PLAN_WITH_PARTICIPANT, while reducing exploratory acts such as INVITE_SHIFT, BACKCHANNEL, and ASK_INFO.

`meta-llama/Llama-3.1-8B-Instruct`, `mistralai/Mistral-7B-Instruct-v0.3`, `mistralai/Mistral-Nemo-Instruct-2407`, and `Qwen/Qwen3.5-9B`. During RL training, the same LLM is used for the participant LLM, facilitator-side rollout generation, and stage-estimation LLM through a local vLLM server.

For the main training pipeline, we run 10 epochs with 3200 environment steps per epoch, 160 steps per collection, 0.1 gradient updates per environment step, 160 warm-up steps, and 1 test episode per evaluation phase. We use a replay buffer of size 50,000, batch size 256, discount factor $\gamma = 0.99$, target smoothing coefficient $\tau = 0.005$, actor learning rate 3×10^{-4}, and critic learning rate 3×10^{-4}. We use automatic entropy tuning, with target entropy set internally as $-\log(|\mathcal{A}|) \times 0.98$.

The full job requests 2 `gpu:a6000` devices, 16 CPU cores, and 128 GB RAM. During training, GPU 1 is used for SAC optimization and GPU 2 serves the rollout-time vLLM server. The training vLLM server uses tensor parallel size 1, maximum context length 8192, maximum batched tokens 8192, maximum concurrent sequences 64, and GPU memory utilization 0.85.

A.3 Facilitator Behavior Analysis

A.3.1 Global Policy Shift

Figure 3 shows the aggregate act-share difference between baseline and RL, pooled across all four LLMs and all six participant-pair types.

A.3.2 Pair-Conditioned Shift

Table 10 breaks the global shift down by participant-pair type, showing how the RL-induced redistribution varies across the six pair configurations.

A.3.3 Per-LLM Pair-Specific Tables

Act	OPEN+OPEN			OPEN+RECEPTIVE			RECEPTIVE+RECEPTIVE			OPEN+RESISTANT			RECEPTIVE+RESISTANT			RESISTANT+RESISTANT		
	LLM	RL	Δ	LLM	RL	Δ	LLM	RL	Δ	LLM	RL	Δ	LLM	RL	Δ	LLM	RL	Δ
ACK_ENCOURAGE	0.167	0.115	-0.053	0.149	0.076	-0.073	0.135	0.062	-0.073	0.123	0.085	-0.038	0.109	0.036	-0.073	0.110	0.070	-0.040
ASK_CONSENT	0.006	0.085	+0.079	0.006	0.079	+0.073	0.008	0.076	+0.067	0.006	0.072	+0.066	0.007	0.082	+0.076	0.008	0.074	+0.066
ASK_EMOTION	0.022	0.007	-0.014	0.016	0.003	-0.013	0.017	0.021	+0.004	0.024	0.020	-0.004	0.026	0.037	+0.011	0.029	0.039	+0.010
ASK_INFO	0.086	0.019	-0.066	0.083	0.023	-0.060	0.087	0.043	-0.044	0.104	0.035	-0.069	0.098	0.085	-0.012	0.106	0.087	-0.019
BACKCHANNEL	0.140	0.032	-0.108	0.134	0.038	-0.096	0.131	0.048	-0.082	0.126	0.025	-0.101	0.135	0.067	-0.068	0.131	0.060	-0.071
EDUCATE_GUIDE	0.000	0.031	+0.031	0.017	0.027	+0.010	0.024	0.030	+0.006	0.003	0.020	+0.018	0.013	0.023	+0.009	0.001	0.019	+0.019
EMPATHIC_REACT	0.074	0.097	+0.023	0.079	0.102	+0.022	0.086	0.092	+0.006	0.103	0.088	-0.015	0.104	0.077	-0.028	0.116	0.060	-0.055
GIVE_SOLUTION	0.011	0.157	+0.146	0.019	0.117	+0.098	0.023	0.077	+0.054	0.004	0.103	+0.099	0.013	0.087	+0.074	0.009	0.076	+0.067
GREET_CLOSE	0.033	0.105	+0.072	0.033	0.125	+0.092	0.033	0.135	+0.101	0.033	0.131	+0.098	0.033	0.123	+0.090	0.033	0.112	+0.079
INVITE_SHIFT	0.179	0.072	-0.107	0.181	0.084	-0.097	0.189	0.101	-0.088	0.207	0.088	-0.120	0.208	0.074	-0.133	0.205	0.082	-0.123
NORMALIZE_REASSURE	0.052	0.014	-0.039	0.062	0.024	-0.037	0.073	0.027	-0.046	0.082	0.048	-0.034	0.083	0.050	-0.033	0.092	0.042	-0.051
NO_ACT	0.000	0.011	+0.011	0.000	0.019	+0.019	0.000	0.019	+0.019	0.000	0.021	+0.021	0.000	0.006	+0.006	0.000	0.013	+0.013
PLAN_WITH_PATIENT	0.182	0.198	+0.015	0.168	0.186	+0.019	0.142	0.156	+0.014	0.115	0.156	+0.041	0.094	0.121	+0.027	0.080	0.121	+0.041
REFLECTION	0.047	0.058	+0.011	0.053	0.098	+0.044	0.052	0.114	+0.062	0.069	0.108	+0.039	0.077	0.131	+0.054	0.079	0.144	+0.065
TOTAL	1.000	1.000	-0.000	1.000	1.000	-0.000	1.000	1.000	+0.000	1.000	1.000	+0.000	1.000	1.000	-0.000	1.000	1.000	-0.000

Table 10: Pair-conditioned therapist act proportions for baseline (LLM) and RL, averaged across four models. This table includes the full 14-act therapist inventory. Within each participant-pair type, the LLM and RL columns each sum to 1. Delta is RL minus LLM.

Llama 3.1 8B						
		Facilitator			Participant	
Pair	Method	R	Q	Ent.	CT	ST
OPEN + OPEN	BL	$0.00_{\pm.00}$	$3.20_{\pm.92}$	$2.70_{\pm.05}$	$2.30_{\pm2.21}$	$0.00_{\pm.00}$
	RL	$2.80_{\pm1.62}^{\uparrow***}$	$4.70_{\pm2.58}$	$3.19_{\pm.20}^{\uparrow***}$	$2.00_{\pm1.41}$	$0.00_{\pm.00}$
OPEN + RECEPTIVE	BL	$0.00_{\pm.00}$	$2.90_{\pm.88}$	$2.71_{\pm.07}$	$2.00_{\pm.94}$	$0.00_{\pm.00}$
	RL	$3.00_{\pm1.33}^{\uparrow***}$	$4.10_{\pm2.02}$	$3.22_{\pm.16}^{\uparrow***}$	$1.00_{\pm1.05}$	$0.00_{\pm.00}$
RECEPTIVE + RECEPTIVE	BL	$0.00_{\pm.00}$	$3.40_{\pm1.17}$	$2.65_{\pm.11}$	$0.10_{\pm.32}$	$0.00_{\pm.00}$
	RL	$3.20_{\pm1.32}^{\uparrow***}$	$3.90_{\pm2.13}$	$3.25_{\pm.20}^{\uparrow***}$	$0.00_{\pm.00}$	$0.00_{\pm.00}$
OPEN + RESISTANT	BL	$0.00_{\pm.00}$	$4.40_{\pm.52}$	$2.63_{\pm.11}$	$0.50_{\pm.71}$	$3.30_{\pm1.95}$
	RL	$2.80_{\pm1.32}^{\uparrow***}$	$4.60_{\pm2.41}$	$3.31_{\pm.12}^{\uparrow***}$	$0.70_{\pm.82}$	$3.40_{\pm1.26}$
RECEPTIVE + RESISTANT	BL	$0.10_{\pm.32}$	$4.20_{\pm.79}$	$2.69_{\pm.12}$	$0.10_{\pm.32}$	$2.30_{\pm1.42}$
	RL	$4.60_{\pm2.59}^{\uparrow***}$	$6.00_{\pm2.62}$	$3.31_{\pm.17}^{\uparrow***}$	$0.10_{\pm.32}$	$2.10_{\pm1.29}$
RESISTANT + RESISTANT	BL	$0.60_{\pm.52}$	$4.80_{\pm.63}$	$2.70_{\pm.16}$	$0.00_{\pm.00}$	$5.40_{\pm3.50}$
	RL	$4.40_{\pm2.07}^{\uparrow***}$	$5.90_{\pm2.18}$	$3.24_{\pm.12}^{\uparrow***}$	$0.00_{\pm.00}$	$8.00_{\pm3.20}$
Mean (60 eps)	BL	$0.12_{\pm.32}$	$3.82_{\pm1.07}$	$2.68_{\pm.11}$	$0.83_{\pm1.38}$	$1.83_{\pm2.65}$
	RL	$3.47_{\pm1.85}$	$4.87_{\pm2.38}$	$3.25_{\pm.16}$	$0.63_{\pm1.06}$	$2.25_{\pm3.23}$
Total (60 eps)	BL	7	229	—	50	110
	RL	208	292	—	38	135

Table 11: Pair-specific BL vs. RL comparisons for **Llama 3.1 8B**. Episode-mean act counts ($n = 10$ per pair). *Facilitator*: R (reflections), Q (questions), Ent. (act entropy). *Participant*: CT (change talk), ST (sustain talk). Significance markers on RL rows from exact two-sided permutation tests within each pair; ↑/↓ mark RL>BL or RL<BL. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Mistral 7B						
		Facilitator			Participant	
Pair	Method	R	Q	Ent.	CT	ST
OPEN + OPEN	BL	$0.10_{\pm.32}$	$0.80_{\pm.79}$	$2.28_{\pm.28}$	$15.30_{\pm2.75}$	$0.00_{\pm.00}$
	RL	$0.30_{\pm.48}$	$6.10_{\pm2.42}^{\uparrow***}$	$2.48_{\pm.22}$	$21.40_{\pm2.46}^{\uparrow***}$	$0.00_{\pm.00}$
OPEN + RECEPTIVE	BL	$0.10_{\pm.32}$	$0.70_{\pm.67}$	$2.27_{\pm.25}$	$15.40_{\pm3.44}$	$0.20_{\pm.63}$
	RL	$0.30_{\pm.67}$	$4.90_{\pm2.13}^{\uparrow***}$	$2.49_{\pm.31}$	$18.00_{\pm4.47}$	$0.50_{\pm.71}$
RECEPTIVE + RECEPTIVE	BL	$0.00_{\pm.00}$	$0.90_{\pm.57}$	$2.26_{\pm.28}$	$11.60_{\pm2.41}$	$0.10_{\pm.32}$
	RL	$0.80_{\pm.79}^{\uparrow*}$	$5.70_{\pm2.67}^{\uparrow***}$	$2.42_{\pm.31}$	$19.10_{\pm3.87}^{\uparrow***}$	$0.20_{\pm.42}$
OPEN + RESISTANT	BL	$0.00_{\pm.00}$	$0.30_{\pm.48}$	$2.16_{\pm.15}$	$10.00_{\pm3.77}$	$0.90_{\pm.88}$
	RL	$0.90_{\pm.74}^{\uparrow**}$	$4.30_{\pm1.64}^{\uparrow***}$	$2.41_{\pm.31}^{\uparrow*}$	$12.90_{\pm5.09}$	$1.60_{\pm1.43}$
RECEPTIVE + RESISTANT	BL	$0.00_{\pm.00}$	$0.60_{\pm.84}$	$2.13_{\pm.13}$	$5.90_{\pm1.91}$	$1.90_{\pm1.20}$
	RL	$0.70_{\pm.67}^{\uparrow*}$	$5.20_{\pm2.04}^{\uparrow***}$	$2.50_{\pm.22}^{\uparrow***}$	$14.90_{\pm3.73}^{\uparrow***}$	$1.60_{\pm1.51}$
RESISTANT + RESISTANT	BL	$0.10_{\pm.32}$	$0.60_{\pm.52}$	$2.12_{\pm.41}$	$4.10_{\pm2.13}$	$1.70_{\pm1.77}$
	RL	$1.00_{\pm1.25}^{\uparrow*}$	$4.10_{\pm1.45}^{\uparrow***}$	$2.40_{\pm.30}$	$11.10_{\pm3.48}^{\uparrow***}$	$2.90_{\pm2.02}$
Mean (60 eps)	BL	$0.05_{\pm.22}$	$0.65_{\pm.66}$	$2.20_{\pm.27}$	$10.38_{\pm5.10}$	$0.80_{\pm1.22}$
	RL	$0.67_{\pm.82}$	$5.05_{\pm2.13}$	$2.45_{\pm.27}$	$16.23_{\pm5.22}$	$1.13_{\pm1.56}$
Total (60 eps)	BL	3	39	—	623	48
	RL	40	303	—	974	68

Table 12: Pair-specific BL vs. RL comparisons for **Mistral 7B**. Episode-mean act counts ($n = 10$ per pair). *Facilitator*: R (reflections), Q (questions), Ent. (act entropy). *Participant*: CT (change talk), ST (sustain talk). Significance markers on RL rows from exact two-sided permutation tests within each pair; ↑/↓ mark RL>BL or RL<BL. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Mistral Nemo 12B						
		Facilitator			Participant	
Pair	Method	R	Q	Ent.	CT	ST
OPEN + OPEN	BL	$3.80_{\pm.92}$	$4.90_{\pm.99}$	$3.10_{\pm.12}$	$13.80_{\pm2.82}$	$0.00_{\pm.00}$
	RL	$2.20_{\pm1.23}^{\downarrow**}$	$0.90_{\pm.99}^{\downarrow***}$	$2.95_{\pm.06}^{\downarrow**}$	$16.20_{\pm4.16}$	$0.00_{\pm.00}$
OPEN + RECEPTIVE	BL	$3.60_{\pm1.17}$	$4.20_{\pm1.32}$	$3.27_{\pm.13}$	$5.50_{\pm3.06}$	$0.50_{\pm.85}$
	RL	$4.90_{\pm1.97}$	$1.40_{\pm1.17}^{\downarrow***}$	$3.15_{\pm.13}$	$6.40_{\pm2.67}$	$0.90_{\pm1.20}$
RECEPTIVE + RECEPTIVE	BL	$3.90_{\pm.99}$	$4.40_{\pm1.71}$	$3.27_{\pm.15}$	$2.80_{\pm2.15}$	$1.00_{\pm1.15}$
	RL	$5.30_{\pm2.06}$	$2.30_{\pm1.06}^{\downarrow**}$	$3.15_{\pm.22}$	$3.00_{\pm1.15}$	$1.00_{\pm1.25}$
OPEN + RESISTANT	BL	$4.70_{\pm.48}$	$6.40_{\pm.84}$	$3.00_{\pm.22}$	$2.90_{\pm1.60}$	$2.50_{\pm1.35}$
	RL	$5.40_{\pm1.43}$	$2.80_{\pm1.62}^{\downarrow***}$	$3.22_{\pm.11}^{\uparrow*}$	$3.20_{\pm1.62}$	$3.20_{\pm1.14}$
RECEPTIVE + RESISTANT	BL	$4.30_{\pm.67}$	$5.80_{\pm.79}$	$3.12_{\pm.15}$	$1.10_{\pm1.20}$	$2.10_{\pm1.45}$
	RL	$5.80_{\pm2.97}$	$5.10_{\pm1.20}$	$3.20_{\pm.19}$	$0.50_{\pm.85}$	$2.00_{\pm.47}$
RESISTANT + RESISTANT	BL	$4.10_{\pm.88}$	$6.60_{\pm.84}$	$3.08_{\pm.14}$	$0.10_{\pm.32}$	$5.10_{\pm1.52}$
	RL	$7.10_{\pm2.64}^{\uparrow**}$	$4.00_{\pm2.62}^{\downarrow*}$	$3.06_{\pm.20}$	$0.10_{\pm.32}$	$5.00_{\pm2.62}$
Mean (60 eps)	BL	$4.07_{\pm.92}$	$5.38_{\pm1.44}$	$3.14_{\pm.18}$	$4.37_{\pm4.99}$	$1.87_{\pm2.04}$
	RL	$5.12_{\pm2.53}$	$2.75_{\pm2.08}$	$3.12_{\pm.18}$	$4.90_{\pm5.90}$	$2.02_{\pm2.14}$
Total (60 eps)	BL	244	323	—	262	112
	RL	307	165	—	294	121

Table 13: Pair-specific BL vs. RL comparisons for **Mistral Nemo 12B**. Episode-mean act counts ($n = 10$ per pair). *Facilitator*: R (reflections), Q (questions), Ent. (act entropy). *Participant*: CT (change talk), ST (sustain talk). Significance markers on RL rows from exact two-sided permutation tests within each pair; ↑/↓ mark RL>BL or RL<BL. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Qwen 3.5 9B						
		Facilitator			Participant	
Pair	Method	R	Q	Ent.	CT	ST
OPEN + OPEN	BL	$1.70_{\pm.48}$	$4.70_{\pm.48}$	$3.02_{\pm.08}$	$10.00_{\pm1.63}$	$0.00_{\pm.00}$
	RL	$1.90_{\pm1.29}$	$2.10_{\pm1.37}^{\downarrow***}$	$2.56_{\pm.26}^{\downarrow***}$	$17.20_{\pm1.62}^{\uparrow***}$	$0.00_{\pm.00}$
OPEN + RECEPTIVE	BL	$2.70_{\pm.82}$	$4.80_{\pm.63}$	$3.02_{\pm.09}$	$6.20_{\pm1.23}$	$0.90_{\pm.88}$
	RL	$3.90_{\pm2.08}$	$2.70_{\pm1.06}^{\downarrow***}$	$3.13_{\pm.23}$	$8.40_{\pm1.26}^{\uparrow**}$	$0.70_{\pm.82}$
RECEPTIVE + RECEPTIVE	BL	$2.30_{\pm1.49}$	$4.70_{\pm.95}$	$3.05_{\pm.09}$	$2.60_{\pm1.43}$	$1.00_{\pm.67}$
	RL	$4.80_{\pm3.16}^{\uparrow*}$	$5.40_{\pm3.78}$	$3.15_{\pm.25}$	$3.20_{\pm1.32}$	$0.70_{\pm1.06}$
OPEN + RESISTANT	BL	$3.60_{\pm1.17}$	$5.00_{\pm.67}$	$2.94_{\pm.09}$	$4.00_{\pm.82}$	$9.60_{\pm1.17}$
	RL	$4.30_{\pm1.49}$	$4.00_{\pm2.16}$	$3.14_{\pm.18}^{\uparrow**}$	$5.80_{\pm1.55}^{\uparrow**}$	$9.90_{\pm1.60}$
RECEPTIVE + RESISTANT	BL	$4.80_{\pm.42}$	$5.00_{\pm.00}$	$3.02_{\pm.06}$	$0.80_{\pm.63}$	$10.40_{\pm.84}$
	RL	$5.10_{\pm2.51}$	$9.10_{\pm3.54}^{\uparrow**}$	$2.87_{\pm.27}$	$0.80_{\pm.79}$	$9.60_{\pm1.78}$
RESISTANT + RESISTANT	BL	$4.70_{\pm.82}$	$5.20_{\pm.42}$	$2.92_{\pm.07}$	$0.00_{\pm.00}$	$18.10_{\pm2.13}$
	RL	$5.40_{\pm2.27}$	$10.80_{\pm2.86}^{\uparrow***}$	$2.85_{\pm.27}$	$0.00_{\pm.00}$	$16.50_{\pm2.22}$
Mean (60 eps)	BL	$3.30_{\pm1.49}$	$4.90_{\pm.60}$	$2.99_{\pm.09}$	$3.93_{\pm3.58}$	$6.67_{\pm6.77}$
	RL	$4.23_{\pm2.42}$	$5.68_{\pm4.14}$	$2.95_{\pm.32}$	$5.90_{\pm5.97}$	$6.23_{\pm6.40}$
Total (60 eps)	BL	198	294	—	236	400
	RL	254	341	—	354	374

Table 14: Pair-specific BL vs. RL comparisons for **Qwen 3.5 9B**. Episode-mean act counts ($n = 10$ per pair). *Facilitator*: R (reflections), Q (questions), Ent. (act entropy). *Participant*: CT (change talk), ST (sustain talk). Significance markers on RL rows from exact two-sided permutation tests within each pair; ↑/↓ mark RL>BL or RL<BL. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.