INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

A Reinforcement Learning-Based Facilitator for Simulated Group Motivational Interviewing

Alafate Abulimiti | Vladislav Maraev | Agnès Helme-Guizon | Catherine Pelachaud

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

Core idea: RL controls the facilitator strategy at the dialogue-act level; the LLM realizes fluent utterances in a simulated triadic group MI session.

1 How the hybrid facilitator works

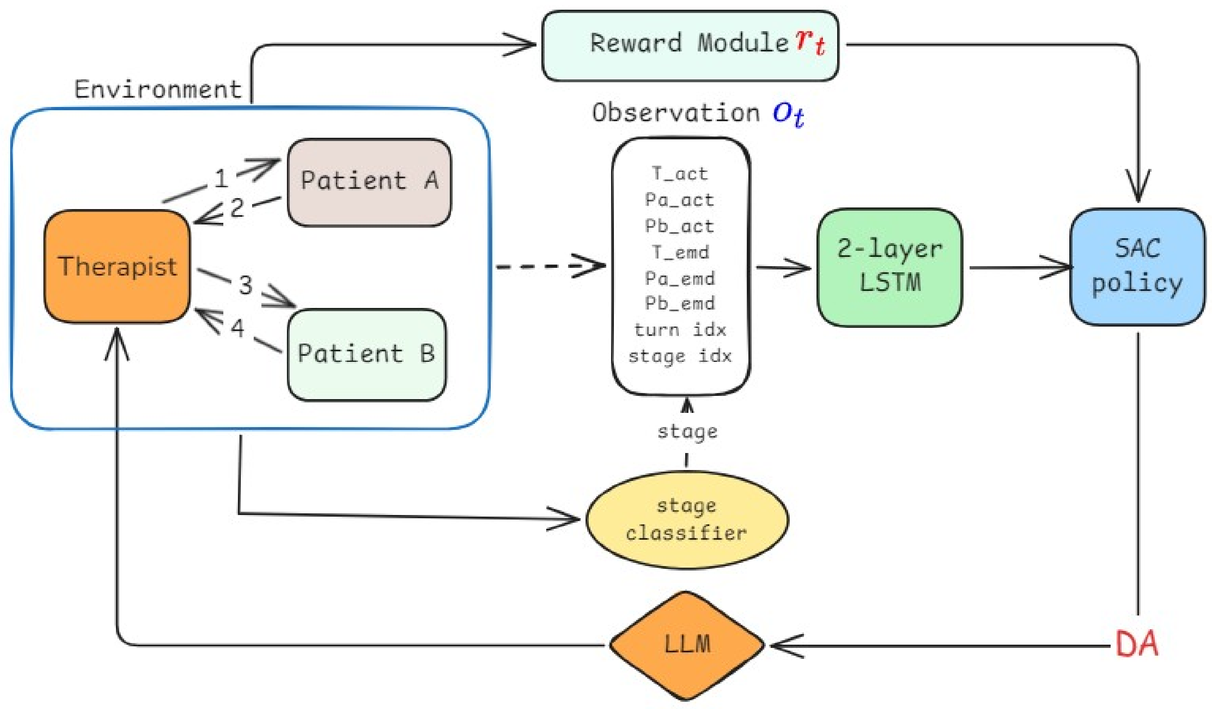


Figure 1: group MI environment, observation, reward, LSTM, SAC policy, and LLM realization loop.

- **Observe the triad and context**
  Recent dialogue acts (DAs) from both participants, sentence embeddings, turn progress, and the current group MI stage.
- **Decide the next dialogue act**
  A 2-layer LSTM feeds a Discrete SAC policy that selects the facilitator's next act.
- **Realize utterance**
  The selected act is passed to an LLM, which generates a natural, context-appropriate utterance.
- **Learn from outcomes**
  A reward module evaluates outcomes and feeds the signal back to the policy.

2 What changes when we train with RL?

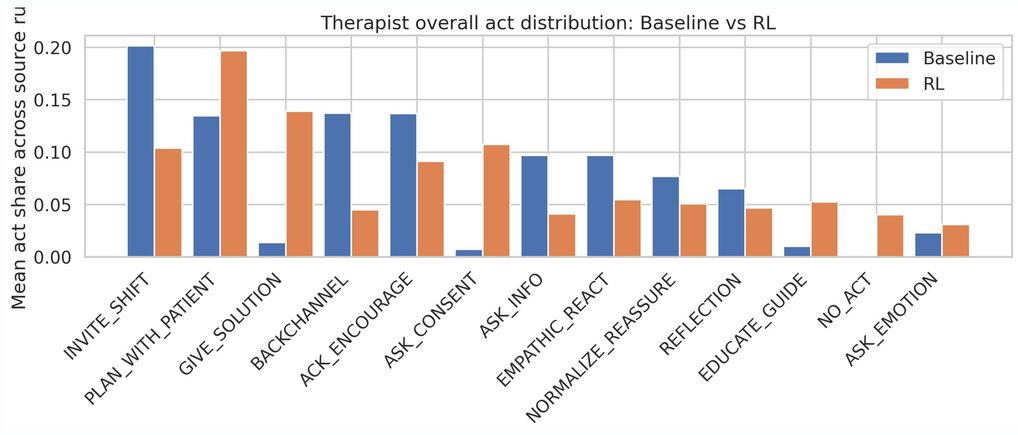
LLM	F: Reflections BL → RL	F: Questions BL → RL	P: Change talk BL → RL	P: Sustain talk BL → RL	JSD
Llama 3.1 8B	0.12 → 3.47***	3.82 → 4.87**	0.83 → 0.63	1.83 → 2.25	**0.246*** **
Mistral 7B	0.05 → 0.67***	0.65 → 5.05***	10.38 → 16.23***	0.80 → 1.13	**0.417*** **
Mistral Nemo 12B	4.07 → 5.12**	5.38 → 2.75↓***	4.37 → 4.90	1.87 → 2.02	**0.095*** **
Qwen 3.5 9B	3.30 → 4.23**	4.90 → 5.68	3.93 → 5.90***	6.67 → 6.23	**0.148*** **

Table 1: -mean MI act counts, simplified as BL → RL; arrows/stars preserve the significance markers. Good practice: more change talk and reflections, less sustain talk and questions (partial results).

- ✓ **RL increases facilitator reflections** across all four LLMs.
- ✓ **Participant change talk shifts** are model-dependent but meaningful.
- ✓ **RL produces clear policy shifts** with significant *Jensen–Shannon divergence* (JSD) for every model.
- ✓ **The learned policy is strategically different** not merely more fluent.

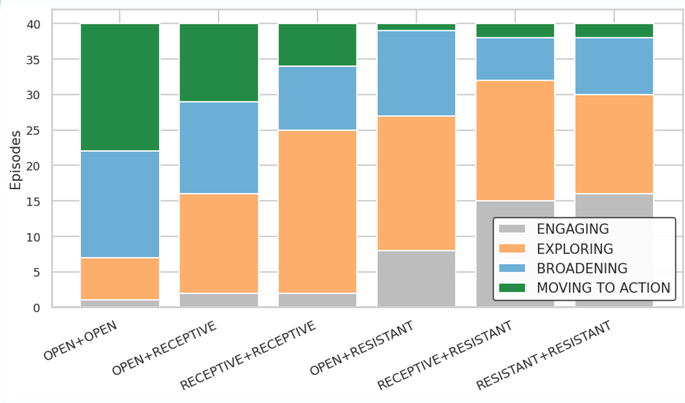
3 What behavior does RL actually learn?

Overall policy shift: Baseline vs. RL



RL shifts the facilitator away from generic invites toward patient-centered planning, solutions, encouragement, and information-seeking compare to pure LLM baseline.

Stage progression under RL



RL policies more often reach later stages, in receptive and resistant cases.

Key findings

JSD = 0.166*** **Global policy shift**

4/4 **LLMs** increase facilitator reflections

OPEN strongest group-composition adaptation

Interpretation: the learned policy is not merely more fluent; it becomes strategically different.

References

- William R. Miller and Gary S. Rose. 2009. Toward a theory of motivational interviewing. American Psychologist, 64(6):527–537.
- 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.
- Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2025. Tailored conversations beyond LLMs: A RL-based dialogue manager. Preprint, arXiv:2506.19652.
- Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. Proceedings of the 35th International Conference on Machine Learning, PMLR 80:1861–1870.
- Petros Christodoulou. 2019. Soft actor-critic for discrete action settings.
- 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.

Action, Cognition, Interaction and Embedded Decisions Team (ACIDE)