Tailored Conversations beyond LLMs: A RL-Based Dialogue Manager

Anonymous ACL submission

Abstract

In this work, we propose a novel framework that integrates large language models (LLMs) with an RL-based dialogue manager for Motivational Interviews (MI). MI is a therapeutic approach that emphasizes collaboration and supports behavioral change by guiding patients to explore the reason and motivation behind their unhealthy behaviors. By leveraging hierarchical reinforcement learning to model the structured phases of MI and employ meta-learning to enhance adaptability across diverse user types, our approach enhances adaptability and efficiency, enabling the system to learn from limited data, transition fluidly between MI phases, and personalize responses to heterogeneous patient needs. Our findings demonstrate that the proposed dialogue manager outperforms an LLM baseline in terms of reward, effectively structuring MI interactions while maintaining adaptability.

1 Introduction

In recent years, the demand for mental health services has surged, outpacing the availability of resources and creating a substantial gap in access to care (Cameron et al., 2017). As a result, many patients face extended waiting periods before receiving therapy (Cameron et al., 2017; Denecke et al., 2020). To address this challenge, virtual agents capable of simulating Motivational Interviewing (MI) have been proposed as a potential interim support system for individuals awaiting treatment. These agents can provide immediate assistance, particularly beneficial in therapeutic approaches requiring multiple sessions (Fiske et al., 2019). However, their role is not to replace human therapists but to serve as a supplementary tool that enhances existing therapeutic interventions.

Motivational Interviewing (MI) poses a particularly complex challenge for dialogue systems, traditionally addressed through intricate rule-based frameworks (Prochaska et al., 2021; Olafsson et al., 2020). However, recent advances in natural language processing (NLP) have paved the way for leveraging large language models (LLMs) such as GPT-like architectures (Baktash and Dawodi, 2023) in such applications (Steenstra et al., 2024), significantly expanding the scope of dialogue systems across various domains.

While these models exhibit remarkable language generation capabilities, they also present significant limitations — many of which can be addressed through insights from "traditional" dialogue research. In particular, LLMs often lack the controllability and structured decision-making of conventional rule-based systems, that are more predictable and interpretable (Shidara et al., 2020).

Ruled-based domain-specific dialogue systems (Hadi et al., 2024) offer notable advantages, including improved controllability, explainability, and the ability to integrate expert knowledge. However, they are typically less adaptable and more resource-intensive to develop. In contrast, LLMs demonstrate strong adaptability across domains but pose challenges in achieving control, transparency, and efficiency. Additionally, incorporating expert knowledge into LLMs often requires extensive domain-specific data (Hadi et al., 2024). Notably, reinforcement learning (RL)-based dialogue managers (Pecune and Marsella, 2020) have shown promise in enhancing control and coherence in dialogue systems. Hence, a promising approach involves hybrid models that combine the strengths of both paradigms—leveraging the adaptability and generative capabilities of LLMs while integrating a domain-specific dialogue manager to regulate interactions (Abu-Rasheed et al., 2024; Galland et al., 2024).

In this work, we investigate a hybrid approach in which an RL-based dialogue manager governs an LLM to simulate MI dialogues, aiming to balance adaptability and control for more effective virtual

therapy support.

2 Theoretical background : Motivational Interviewing (MI)

Motivational Interviewing (MI) is a therapeutic approach that emphasizes collaboration and supports behavioral change by guiding patients to explore the reason and motivation behind their unhealthy behaviors.

2.1 Dialogue with Multiple Phases

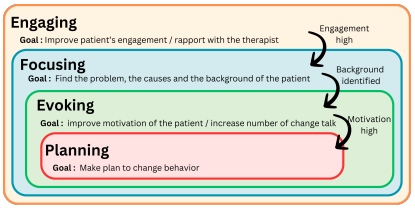

Figure 1: Phases of Motivational Interviewing

Complex dialogues, such as those in Motivational Interviewing (MI), evolve through distinct phases, each guided by unique long-term strategies (Miller and Rollnick, 2012) (see Figure 1). The dialogue usually begins with an *engaging* phase, where rapport is established, and patient engagement with the therapist is fostered. This is followed by a *focusing* phase, where core issues, their underlying causes, and the patient's background are identified to set a clear focus for the conversation. The third phase is the *evoking* phase, which involves encouraging the patient's motivation for change by eliciting and amplifying "change talk". Finally, *planning* involves developing a specific, actionable plan for behavior change based on the patient's motivation and goals.

Therapists must ensure that specific objectives, such as achieving high levels of engagement, clarifying core issues, and cultivating sufficient motivation, are met before transitioning between phases. Furthermore, the process is not strictly linear, as therapists may need to revisit earlier phases depending on the patients' evolving motivation and engagement. Individual variability in engagement and motivation necessitates a flexible approach.

For a virtual therapist employing MI, effectively navigating across these phases is crucial. This requires discerning when to progress to the next phase, when to revisit earlier phases, and how to adapt the interaction to align with each patient's unique needs and circumstances.

2.2 Patients types in MI

Patients participating in motivational interviewing (MI) exhibit varying levels of readiness to change their behaviors. As proposed by (anonymous, 2024a), these patients can be classified into three categories: *Open-to-Change*, *Resistant-to-Change*, and *Receptive*. *Open-to-Change* individuals demonstrate a strong willingness to modify unhealthy behaviors. *Resistant-to-Change* patients are generally reluctant to alter their current behaviors, showing a preference for maintaining the status quo. *Receptive* patients, while initially exhibiting low motivation to change, gradually develop a higher motivation to adopt healthier behaviors as the conversation progresses.

These classifications capture variations in patients' responses and therapists' strategies, as discussed in (anonymous, 2024a). The ability to adapt the flow of dialogue to these three patient types can significantly enhance the efficiency of the therapist dialogue model.

A dialogue system for MI should be able to take into account the particular challenges that such dialogs arises. Such a dialogue system should be able to navigate across phases while being able to adapt to different types of users.

3 Related Work

Motivational Interviewing (MI) presents significant challenges for dialogue systems, as it necessitates both a structured progression through its four distinct phases —engagement, focusing, evoking, and planning (Miller and Rollnick, 2012) — and adaptability to diverse patients profiles (anonymous, 2024a). While existing systems such as Woebot (Prochaska et al., 2021) have demonstrated the feasibility of MI-based chatbots by incorporating therapeutic frameworks like cognitive behavioral therapy (CBT) and mindfulness, they predominantly rely on static or rule-based architectures. Steenstral et al. (Steenstra et al., 2024) identified the limitations of rule-based approaches in maintaining adherence to therapeutic protocols and proposed leveraging LLMs for this application, demonstrating promising results.

The advent of large language models (LLMs) has transformed dialogue generation, offering new possibilities for MI-based interactions (Steenstra et al., 2024). Models such as GPT-like systems (Baktash and Dawodi, 2023) exhibit strong generative capabilities and adaptability across diverse applications,

which can be leveraged to enhance the social dimensions of interaction. This aspect is particularly relevant, as Kanaoka et al. (Kanaoka and Mutlu, 2015) emphasized the critical role of social engagement and rapport in facilitating behavioral change. Recent research has underscored the importance of cognitive modeling and adaptability in dialogue systems to more accurately account for the mental states of both the agent and the user. For instance, He et al. (He et al., 2024) introduced dual reasoning mechanisms that enable LLMs to incorporate contextual nuances, while Zhang et al. (Zhang et al., 2020) explored interactive agent representations to improve dialogue coherence. Despite these advancements, LLMs remain constrained in terms of controllability and domain specificity (Shidara et al., 2020). Recent efforts have sought to enhance control and applicability, particularly within task-oriented dialogue systems. Yao et al. (Yao et al., 2023) employed reinforcement learning (RL) to optimize LLM prompting strategies; however, black-box LLMs continue to exhibit limitations in controllability and often generate repetitive responses. To address these shortcomings, Xu et al. (Xu et al., 2023) demonstrated the advantages of integrating fine-tuned, smaller language models with larger LLMs, while Yu et al. (Yu et al., 2023) employed Monte Carlo tree search for optimal action selection, improving both coherence and practical utility. Although these studies mark significant progress in hybrid model design, they often fall short of accommodating a broad spectrum of user types. Integrating the structured controllability of classical dialogue models with the generative flexibility of LLMs within an adaptive hybrid framework could facilitate more dynamic, personalized, and effective MI interactions.

Reinforcement learning has been instrumental in optimizing dialogue policies and refining system behavior. Traditional RL-based approaches have primarily focused on enhancing user engagement and task success rates. Walker et al. (Walker, 2000) and Li et al. (Li et al., 2016) demonstrated the efficacy of RL in training conversational agents for goal-oriented tasks, while Weber et al. (Weber et al., 2018) illustrated its effectiveness in selecting contextually appropriate actions, such as humor or sound effects, to enrich user experiences. More advanced frameworks have incorporated both social and task-oriented rewards (Pecune and Marsella, 2020) or jointly trained user and dialogue policies (Takanobu et al., 2020). Although these methods enable adaptation to different user types, they remain largely confined to task-oriented systems that rely on predefined natural language templates.

In this work, we propose a novel framework that integrates LLMs with an RL-based dialogue manager to structure MI dialogues across different phases while dynamically adapting to diverse patient profiles. By synergizing structured control with generative flexibility, our approach enhances adaptability and efficiency, enabling the system to learn from limited data, transit fluidly between MI phases, and personalize responses to heterogeneous patient needs.

The remainder of this paper is organized as follows: Section 4 details our proposed methodology, while Section 5 describes our evaluation environment. Section 6 presents the experimental results, and Section 7 provides an analysis and interpretation of our findings.

4 Method

This section outlines the methodology for developing a dialogue manager capable of navigating the distinct phases of Motivational Interviewing (MI) while adapting to diverse patients profile. The complete architecture is depicted in Figure 2.

4.1 Problem description

The objective of the proposed model is to predict the optimal action a_t at each time step t, given the dialogue context c. Each action corresponds to a dialogue act representing the virtual therapist's strategic behavior. The agent sentence is then generated by a conditioned large language model (LLM) that produces an utterance coherent with the context and realizes the selected dialogue act, as validated in (anonymous, 2024b).

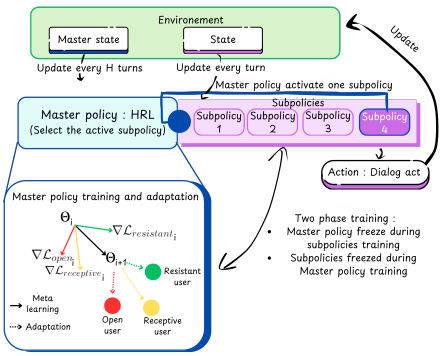

Figure 2: Hierarchical architecture of the dialogue manager

4.2 Hierarchical RL for Managing Dialogue Phases

The dialogue manager employs a hierarchical reinforcement learning (HRL) framework to manage dialogue phases comprising a master policy and N sub-policies. Each sub-policy is dedicated to a specific phase of the Motivational Interviewing (MI) process, while the master policy governs transitions between these phases over a fixed horizon H (see Figure 2). The master policy orchestrates the dialogue by selecting the appropriate phase for the next H dialogue turns based on the current master state. Meanwhile, sub-policies execute phase-specific strategies and actions, leveraging their respective sub-policy states. Notably, the master state and the sub-policy state differ in composition, as long-term planning requires distinct information from short-term decision-making. This hierarchical structure ensures dynamic and context-sensitive dialogue management, allowing real-time adjustments to both the patient's evolving needs and the interaction context. It balances global objectives, such as increasing motivation for behavior change, with more localized goals, such as answering patient inquiries. The fixed decision interval of H turns reduces the training horizon, simplifying the learning process and enhancing the model's adaptability to diverse users by focusing on shorter-term adjustments. By leveraging hierarchical reinforcement learning (HRL), the model effectively manages the different phases of MI dialogue while minimizing adaptation complexity. The sub-policies handle local objectives, while the master policy adjusts to user-specific global goals. Once the sub-policies are trained, adapting to a new user requires fine-tuning only the master policy, which operates on a smaller horizon and action space. This approach enables efficient transitions between MI phases, ensuring that interactions remain tailored to the needs of each individual user.

4.3 Meta-Learning for User Adaptation

To facilitate rapid adaptation to new users, the master policy is trained using meta-learning techniques, specifically the Model-Agnostic Meta-Learning (MAML) algorithm (Finn et al., 2017). This approach enables the dialogue manager to generalize efficiently across diverse user types while retaining the flexibility to quickly adapt to novel ones. By leveraging MAML, the master policy not only learns strategies shared among various patients' profiles but also fine-tunes its behavior in response to individual patients needs, ensuring both robustness and personalization.

4.4 Algorithm and Training Framework

In this subsection we present formally our algorithm and training framework.

4.4.1 Dialogue Management Algorithm

The model aims to predict the optimal dialogue act a_t to maximize a reward function $\mathcal{R}(s_t, a_t)$. The system comprises a master policy θ and a set of N sub-policies $\psi_0, \ldots, \psi_N$. The master policy, with a discrete action space of size N, determines the appropriate sub-policy to use for the next H turns, while each sub-policy ψ_i manages the dialogue acts of the corresponding phase within an action space of size $N_{da} = 13$. At each time step t, the algorithm operates as follows. If $t \mod H = 0$, the master policy selects the next sub-policy: $A_t = \theta(s_t^{master})$. Otherwise, the previous master action is reused: $A_t = A_{t-1}$. The sub-policy corresponding to A_t then generates the next dialogue act: $a_t = \psi_{A_t}(s_t)$. This action a_t influences the environment, producing a user response and updating the state to s_{t+1} and s_{t+1}^{master}. The detailed algorithm is presented in Algorithm 1

4.4.2 Training Framework

The training framework leverages a model-based reinforcement learning (RL) approach. A model-based approach enables efficient reuse of dialogue turns across multiple iterations as policies evolve. Specifically, we utilize the Soft Actor-Critic (SAC) algorithm (Haarnoja et al., 2018), which enhances the system's adaptability to new human users in online interactions. This approach allows for policy updates at each turn, maintaining the information from previous turns. Each training epoch targets a specific user type and begins with cloning the master policy θ. The optimization process occurs in two phases. In the first phase, the master policy θ is fixed, and the sub-policies $\psi_0, \ldots, \psi_n$ are optimized using SAC. In the second phase, the sub-policies remain fixed while the cloned master policy θ_{clone} is optimized using SAC. After these optimizations, the updated policies are evaluated. Finally, the master policy θ is updated using the MAML algorithm. The complete training process is detailed in Algorithm 2.

Algorithm 1 Hierarchical Dialogue Management Algorithm

Input: State s_t and s_t^{master}, Master policy θ, Sub-policies $\psi_0, \ldots, \psi_n$, Time horizon H
Initialize: Master policy θ, Sub-policies $\psi_0, \ldots, \psi_n$
for each time step t **do**
    if $t \mod H = 0$ **then**
        Compute master action $A_t = \theta(s_t^{master})$
    else
        Reuse previous master action $A_t = A_{t-1}$
    end if
    Select sub-policy ψ_{A_t} based on A_t
    Compute dialogue act $a_t = \psi_{A_t}(s_t)$
    Apply action a_t to the environment
    Observe user response and new state s_{t+1} and s_{t+1}^{master}
end for

Algorithm 2 Training Process for Hierarchical Dialogue Manager

Input: Master policy θ, Sub-policies $\psi_0, \ldots, \psi_n$, User simulator, Replay buffer D_{sub}, Replay buffer D_{master}
Initialize: Master policy θ, Sub-policies $\psi_0, \ldots, \psi_n$, Replay buffer D_{sub}, Replay buffer D_{master}
for each training epoch **do**
    Sample user type t and apply to the simulator
    Phase 1: Sub-policy Optimization
    Fix θ and for N_{sub} dialogues:
    for each dialogue **do**
        Generate transition $(s_t, a_t, r_t, s_{t+1}, A_t)$ and store in replay buffer D_{sub}
        Sample a batch B from D_{sub}
        for each sub-policy ψ_i **do**
            Optimize ψ_i: $\psi_i \leftarrow \text{SAC}(\psi_i, B_{A_t=i})$
        end for
    end for
    Phase 2: Master Policy Optimization
    Clone master policy: $\theta_{\text{clone}} \leftarrow \theta$
    Fix $\psi_0, \ldots, \psi_n$ and for N_{master} dialogues:
    for each dialogue **do**
        Generate transition (s_t, a_t, r_t, s_{t+1}) and store in replay buffer D_{master}
        Optimize θ_{clone}: $\theta_{\text{clone}} \leftarrow \text{SAC}(\theta_{\text{clone}})$
    end for
    Evaluation
    Evaluate the updated policies θ_{clone} and $\psi_0, \ldots, \psi_n$ on the task
    Master Policy Update
    Update the master policy θ using the MAML algorithm: $\theta \leftarrow \text{MAML}(\theta_{\text{clone}})$
    Empty Replay buffer D_{sub} and Replay buffer D_{master}
end for

5 Evaluation Environment

In this section we present the evaluation env of our framework on a MI dialogue environment and compare our model with a state of the art LLM baseline. We also perform ablations to demonstrate the efficiency of each of the model components showing that incorporating knowledge on the flow of the dialogue in the dialogue manager development improves the resulting conversations.

5.1 Baseline

We use as a baseline a Nemo Instruct LLM[1] prompted as validated in (Steenstra et al., 2024). The prompt incorporates information on MI strategies, as well as techniques for addressing specific problems, such as *Drinking*, *Smoking*, and *Sedentary Lifestyle*. This approach was validated in (Steenstra et al., 2024) through experiments with human participants.

5.2 Evaluation Environment

The evaluation environment includes a simulated user described in Section 5.2.1 with a specific type T and a problem P, where $P \in \{\text{Smoking, Alcohol, Sedentary Lifestyle}\}$. Additionally, the environment incorporates a Mistral LLM, specifically a Nemo Instruct LLM[1], which is prompted to generate both therapist and patient utterances based on the context, discussion theme, and dialogue act. This generation was validated in (anonymous, 2024c).

5.2.1 Simulate patients in MI

Simulating patients in MI has been explored in prior research. For instance, (anonymous, 2024c) proposed a prompt to simulate a user with a large language model (LLM) (anonymous, 2024b). This approach has been validated to produce contextually relevant, natural dialogue acts and utterances (anonymous, 2024b,c), although the differences between user types have not been tested and might be a limitation of this simulator. We use this user simulator to train and evaluate our dialogue manager. In the following of this article, the term user refers to this user simulator.

5.2.2 Action Space

The agent operates in a discrete action space consisting of 13 possible dialogue acts, which are categorized into task-oriented dialogue acts and socially oriented dialogue acts. Task-oriented dialogue acts include *Asking for Consent or Validation, Providing Medical Education and Guidance, Planning with the Patient, Giving a Solution, Asking about Current Emotions, Inviting a Shift in Outlook, Asking for Information*, and *Reflection*. Socially-oriented dialogue acts include *Empathic reactions, Acknowledging Progress and Encouraging, Backchanneling, Greeting or Closing, and Normalizing Experiences while Providing Reassurance*. This taxonomy was introduced in (anonymous, 2024a).

5.2.3 State Space

The agent's state space includes information from the most recent agent's and user's dialogue acts. User can use 9 different dialogue acts: *Changing Unhealthy Behavior, Sustaining Unhealthy Behavior, Sharing Negative/Positive Feelings or Emotions, Sharing Personal Information, Realization or Understanding, Greeting or Closing, Backchanneling,* and *Asking for Medical Information*. Additionally, the state space incorporates the current timestamp and an encoded representation of the dialogue context, which comprises the last three utterances.

5.2.4 Master State Space

The master policy's state space is composed of an approximation of COntext knowledge, engagement approximation and Evocation approximation. Context knowledge approximation is measured by the number of times the user employs the *Sharing Personal Information* dialogue act. Engagement approximation is determined by the number of times the user utilizes the *Sharing Positive/Negative Feelings* dialogue act. Evocation approximation is quantified by the number of uses of the *Understanding or New Perspective* dialogue act.

5.2.5 Reward Function

The reward function is designed to predict therapy outcomes by assigning specific values to different user dialogue acts. Prior research underscores the critical role of user responses, such as *sustain talk*, which is linked to poorer treatment outcomes (Magill et al., 2014), and *change talk*, which is associated with reduced risk behaviors during follow-up assessments (Magill et al., 2018). Additionally, the reward function incentivizes structured progression through the MI phases. A reward of $+5$ is assigned for *Changing Unhealthy Behavior*, as this represents the desired outcome, whereas a penalty of -5

[1] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

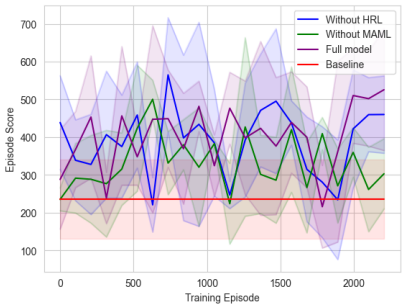

Figure 3: Evolution of the reward function during training

is given for *Sustaining Unhealthy Behavior*, which should be discouraged. In the Engagement phase, a reward of $+50$ is granted for expressing feelings. Once at least two emotions have been expressed by the user, a reward of $+100$ is assigned for providing information in the Focusing phase. After at least two pieces of information have been shared, a reward of $+150$ is given for evoking-related dialogue acts, culminating in a reward of $+200$ for planning-related dialogue acts.

5.2.6 Episode Termination

An episode concludes after 40 turns or when the agent performs a closing action, marking the end of the dialogue interaction.

5.3 Hyperparameters

The model is trained for 22 epochs, with 105 conversations conducted per epoch. Of these, 30 are used to train the master policy, while 60 are dedicated to training the sub-policies and 5 to evaluation. To speed up the training process, 5 conversations are performed in parallel. There are 6 sub-policies in total, and the master policy is executed every 3 turns. The user's type is fixed randomly at the beginning of each epoch. The sub-policies are trained with a learning rate of 10^{-7} and batch size of 10, the master policy uses a learning rate of 10^{-6} and batch size of 10, and the MAML (Model-Agnostic Meta-Learning) algorithm operates with a learning rate of $4 * 10^{-7}$. Each network is composed of 2 linear layers intercalated with Leaky ReLU activation functions and hiddensize of 32. The model is trained over 16 hours using a 42GB GPU.

6 Results

In this section, we present our results and ablation studies. Figure 3 illustrates the mean reward evolution across all three user types—Open to Change, Resistant to Change, and Hesitant—throughout the training process. At each evaluation epoch, five conversations are conducted with each user type. Additionally, Table 1 presents the final experimental results. Our model's reward performance significantly surpasses that of the baseline, demonstrating that conditioning an LLM with our dialogue manager enhances the proportion of desirable dialogue acts. However, the reward fluctuates significantly during training due to the high variability in the environment. This instability is likely caused by the user simulator, which can exhibit erratic behavior at times.

Table 1: Experiment Results with Mean Rewards (* = Statistical difference with baseline)

Experiment	Mean Reward (± SD)
Baseline	235± 106
Without MAML	303 ± 93
Without HRL	460 ± 102
Full model	**526 ± 161** *

6.1 Ablation Studies

We conduct two ablation studies to evaluate the impact of each design choice. In the first ablation study, we perform the same training procedure without employing MAML to train the master policy. In the second ablation study, we remove the hierarchical reinforcement learning (HRL) framework and train solely with the SAC algorithm (see Figure 3 and Table 1).

Effect of MAML The inclusion of MAML improves the accumulated reward (see Table 1), suggesting that it enhances the learning of the master policy by explicitly accounting for variations across user types. Standard training can be biased by the sequence in which different user types are encountered, whereas MAML mitigates this by guiding the master policy toward an initialization that enables rapid adaptation to diverse users.

Effect of HRL The effectiveness of HRL is further supported by the experimental results. Training with only the SAC algorithm leads to a lower accumulated reward, likely because the phase-based structure of MI dialogues is more challenging to capture without hierarchical modeling.

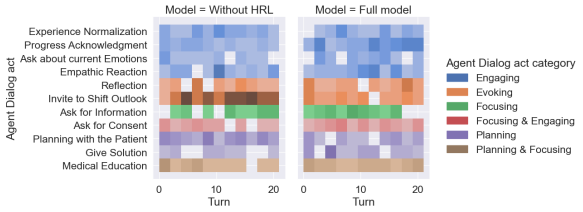

Figure 4: Dialogue act distribution over time, highlighting different dialogue phases for the full model. The intensity of the color is proportional to the use of the corresponding dialog act in the turn.

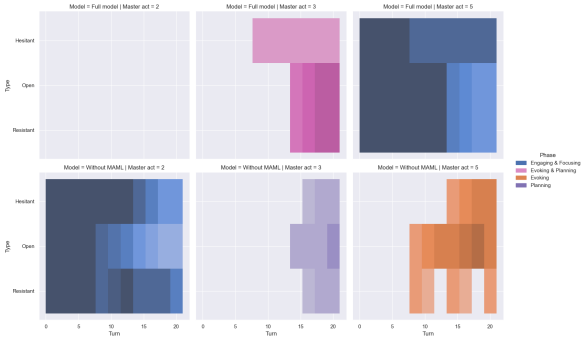

Figure 5: Distribution of phase activation over time across different user types. The intensity of the color is proportional to the use of the corresponding dialog act in the turn.

7 Interpretation

In this section, we analyze the generated dialogues and examine how our design choices influence them. Specifically, we investigate the different MI phases to determine whether they emerge as expected and assess the impact of HRL. Additionally, we explore variations across user types and evaluate the effect of MAML on the generated dialogues.

Differences Between Phases To analyze the MI phases, we examine the distribution of dialogue acts across different dialogue turns. Dialogue acts associated with the Engaging phase, such as asking about emotions or sharing emotions, should be more prevalent at the beginning of the conversation, whereas those related to the Planning phase, such as providing solutions or promoting behavior change, should appear more frequently towards the end (Miller and Rollnick, 2012). While Engaging should occur throughout the entire dialogue, the later stages should be more focused on Planning. The phases are interwoven rather than strictly sequential (see Figure 1). Figure 4 shows that both the full model and the HRL ablation employ engagement-related dialogue acts throughout the conversation. However, the HRL ablation tends to use Evoking and Planning-related acts too early, rather than focusing primarily on Engaging and Focusing at the beginning. This issue is particularly noticeable with the Invite to Shift Outlook dialogue act. Notably, in the full model, Asking for Information and Focusing gradually decrease once sufficient information has been gathered, reflecting a more structured and adaptive dialogue flow.

Differences between user types In this section, we examine the impact of meta-learning on training the master policy. Figure 5 illustrates the distribution of master actions activation over time for different user types. The models effectively differentiates between distinct phases, initially engaging in Engaging/Focusing phases (Master action 5 for the full model and 2 for the ablation) before transitioning into Evoking/Planning phases (Master action 3 for the full model and 3 and 5 for the ablation). For the full model, this transition occurs earlier in the dialogue for Hesitant users, as they require additional motivation, the Evoking phase then tends to be prolonged, aligning with prior findings (anonymous, 2024a). In contrast, in the MAML ablation, the Evoking (Master action 5) and Planning (Master action 3) phases are less stable, with the model frequently oscillating between them, especially for Resistant users. This instability suggests that the ablation struggles to maintain a well-structured dialogue flow, making it more susceptible to variations in user types. This qualitative analysis helps explain the significant difference in reward observed between the full model and the MAML ablation. It highlights the benefits of incorporating meta-learning, as it enhances the model's ability to structure MI phases effectively and provide a more personalized dialogue experience.

8 Conclusion

In this paper, we present a dialogue manager for MI dialogue design, addressing the unique challenges posed by this type of interaction. We leverage HRL to model the structured phases of MI and employ meta-learning to enhance adaptability across diverse user types. Our findings demonstrate that the proposed dialogue manager outperforms an LLM baseline in terms of reward. Additionally, our analysis of the generated conversations provides valuable insights into how HRL and meta-learning contribute to the structured yet adaptive nature of the dialogue.

9 Limitations

The current framework is trained using a single implementation of a simulated user, which limits its generalizability. To fully assess its effectiveness, the model should be tested with human participants or on diverse datasets that capture a broader range of user behaviors and characteristics.

Moreover, the simulator is based on a Mistral LLM, which requires significant processing time to generate user behaviors. This limitation constrains the training capacity, as complex reinforcement learning (RL) problems like this one require extensive trial and error. As a result, certain design choices—such as small batch sizes and a limited number of conversations per epoch—were necessary, contributing to the observed training instability. Addressing this issue in future work could lead to more robust and efficient training.

Additionally, the analysis of dialogue phases currently relies on predefined heuristics, making it inherently subjective. A more rigorous approach would involve annotation and validation by professional MI annotators to ensure alignment with clinical practices, thereby improving the system's reliability.

10 Ethical Implications

This work introduces a dialogue manager designed for Motivational Interviewing interactions. Its objective is not to replace therapists but to provide supplementary support or serve as an introduction to therapy. The focus of this research is exclusively on the dialogue management component, which operates within a constrained set of possible actions. All the LLMs are run locally and no sensitive information is sent to outside services.

Given the sensitive nature of such applications, careful examination and validation of the language model outputs remain essential. It is imperative to emphasize in both the codebase and accompanying documentation that these interactions are not intended to replace professional therapists but rather to complement their efforts in appropriate contexts. Ensuring transparency and adherence to ethical standards is fundamental to responsibly deploying this technology.

11 Acknowledgement

Ai writing assistance has been used purely with the language of the paper.

References

Hasan Abu-Rasheed, Mohamad Hussam Abdulsalam, Christian Weber, and Madjid Fathi. 2024. Supporting student decisions on learning recommendations: An llm-based chatbot with knowledge graph contextualization for conversational explainability and mentoring. *arXiv preprint arXiv:2401.08517*.

anonymous. 2024a. anonymous.

anonymous. 2024b. anonymous. pages 1–4.

anonymous. 2024c. anonymous. pages 192–203.

Jawid Ahmad Baktash and Mursal Dawodi. 2023. Gpt-4: A review on advancements and opportunities in natural language processing. *arXiv preprint arXiv:2305.03195*.

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) 31*, pages 1–7.

Kerstin Denecke, Sayan Vaaheesan, and Aaganya Arulnathan. 2020. A mental health chatbot for regulating emotions (sermo)-concept and usability test. *IEEE Transactions on Emerging Topics in Computing*, 9(3):1170–1182.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Amelia Fiske, Peter Henningsen, and Alena Buyx. 2019. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of medical Internet research*, 21(5):e13216.

Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2024. Simulating patient oral dialogues: A study on naturalness and In *Proceedings of the ACM International Conference on Intelligent Virtual Agents*, pages 1–4.

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 *International conference on machine learning*, pages 1861–1870. PMLR.

Muhammad Usman Hadi, Qasem Al Tashi, Abbas Shah, Rizwan Qureshi, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, et al. 2024. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

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. *arXiv preprint arXiv:2406.05374*.

Toshikazu Kanaoka and Bilge Mutlu. 2015. Designing a motivational agent for behavior change in physical activity. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1445–1450.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Molly Magill, Timothy R Apodaca, Brian Borsari, Jacques Gaume, Ariel Hoadley, Rebecca EF Gordon, J Scott Tonigan, and Theresa Moyers. 2018. A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. *Journal of consulting and clinical psychology*, 86(2):140.

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.

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

Stefan Olafsson, Byron C Wallace, and Timothy W Bickmore. 2020. Towards a computational framework for automating substance use counseling with virtual agents. In *AAMAS*, volume 19, pages 9–13. Auckland.

Florian Pecune and Stacy Marsella. 2020. A framework to co-optimize task and social dialogue policies using reinforcement learning. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8.

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.

Kazuhiro Shidara, Hiroki Tanaka, Hiroyoshi Adachi, Daisuke Kanayama, Yukako Sakagami, Takashi Kudo, and Satoshi Nakamura. 2020. Analysis of mood changes and facial expressions during cognitive behavior therapy through a virtual agent. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 477–481.

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.

Ryuichi Takanobu, Runze Liang, and Minlie Huang. 2020. Multi-agent task-oriented dialog policy learning with role-aware reward decomposition. *arXiv preprint arXiv:2004.03809*.

Marilyn A Walker. 2000. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416.

Klaus Weber, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser, and Elisabeth André. 2018. How to shape the humor of a robot-social behavior adaptation based on reinforcement learning. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 154–162.

Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. 2023. Small models are valuable plug-ins for large language models. *arXiv preprint arXiv:2305.08848*.

Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, et al. 2023. Retroformer: Retrospective large language agents with policy gradient optimization. *arXiv preprint arXiv:2308.02151*.

Xiao Yu, Maximillian Chen, and Zhou Yu. 2023. Prompt-based monte-carlo tree search for goal-oriented dialogue policy planning. *arXiv preprint arXiv:2305.13660*.

Zheng Zhang, Lizi Liao, Xiaoyan Zhu, Tat-Seng Chua, Zitao Liu, Yan Huang, and Minlie Huang. 2020. Learning goal-oriented dialogue policy with opposite agent awareness. *arXiv preprint arXiv:2004.09731*.