Preference Tree Optimization Framework: Enhancing Goal-Oriented Dialogue with Look-Ahead Simulations

Lior Baruch

October 2024

Abstract

Developing dialogue systems capable of engaging in multi-turn, goal-oriented conversations remains a significant challenge, especially in specialized domains with limited data. This research proposes a novel framework called *Preference Tree Optimization Framework (PTOF)*, designed to iteratively improve agent models in such dialogue systems, by generating preference data using a method called *Preference Tree with Look-Ahead*. Focusing on Motivational Interviewing (MI)—a counseling technique aimed at facilitating behavioral change—we leverage virtual patients and an oracle evaluator to simulate conversations and generate rich preference datasets. By combining this method with Direct Preference Optimization (DPO) [1], we aim to enhance the agent's decision-making capabilities over iterative training cycles. The proposed framework addresses data scarcity and advances the development of more nuanced and effective dialogue systems in goal-oriented domains.

1 Introduction

Goal-oriented dialogue systems are designed to achieve specific objectives through interactive conversations. Developing such systems in specialized domains is challenging due to the complexity of interactions and the scarcity of domain-specific data. Motivational Interviewing (MI) is such a domain – it is a counseling approach that facilitates behavioral change through collaborative, client-centered dialogue, requiring nuanced understanding and adaptability from the conversational agent.

This research introduces a framework for iteratively improving agent models in goal-oriented dialogue systems, called *Preference Tree Optimization Framework (PTOF)*, by generating preference data using a novel method called *Preference Tree with Look-Ahead*. This method systematically simulates various conversational paths and evaluates them using an oracle to generate preference data. We use this preference data with Direct Preference Optimization (DPO) [1] to iteratively refine the agent model, enhancing its decision-making capabilities.

Our approach leverages existing virtual patients and evaluators from previous research in MI [2], making it an ideal testbed for our framework. By addressing the challenges of data scarcity and the need for nuanced interactions, we aim to contribute to the advancement of dialogue systems capable of effective, goal-oriented conversations.

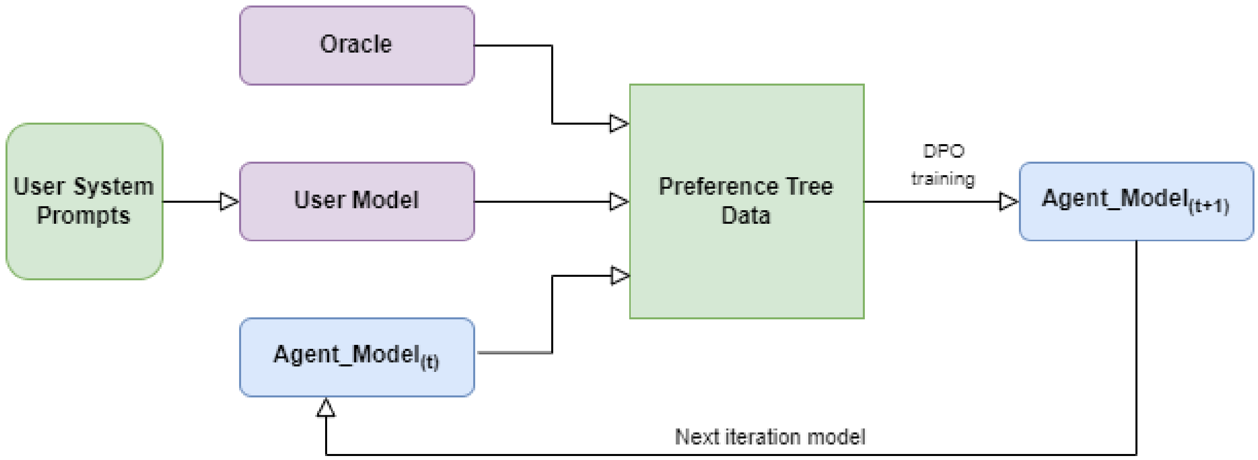

Figure 1: **Preference Tree Optimization Framework (PTOF)**. The framework operates in two iterative steps: (i) **Preference Data Generation**: The User Model is prompted with personalized attributes to simulate diverse user personalities. For each digital user personality, the *Preference Tree with Look-Ahead* 2 method is used in conjunction with the Oracle Evaluator and the current agent model ($Agent\_Model_t$) to generate a preference tree that explores various conversational pathways. These trees are aggregated into a comprehensive preference dataset. (ii) **Model Training**: The current agent model is trained on the newly generated preference dataset using Direct Preference Optimization (DPO), resulting in an improved model ($Agent\_Model_{t+1}$). The updated agent model is then used for the next iteration, repeating the process for continuous improvement.

2 Research Objectives

The primary objectives of this research are:

1. **Propose a Novel Preference Data Generation Method**: Implement the *Preference Tree with Look-Ahead* to simulate and evaluate potential conversational paths in goal-oriented dialogues.

2. **Iteratively Enhance Agent Models**: Employ the Preference Tree Optimization Framework (PTOF) to integrate the generated preference data with Direct Preference Optimization (DPO), iteratively improving the agent's decision-making capabilities.

3. **Apply to Motivational Interviewing (MI)**: Evaluate the effectiveness of the proposed framework in the MI domain, leveraging existing virtual patients and evaluators.

4. **Benchmark Against Existing Models**: Compare the performance of our agent with current state-of-the-art dialogue systems in MI contexts to validate the efficacy of our approach.

5. **Contribute to Goal-Oriented Dialogue Systems**: Provide insights and methodologies that can be generalized to other specialized dialogue domains.

3 Background and Related Work

Advancements in Natural Language Processing (NLP) and the emergence of Large Language Models (LLMs) have significantly transformed dialogue systems. Despite these developments, creating effective goal-oriented dialogue systems in specialized domains—such as Motivational Interviewing (MI)—remains a complex challenge due to the scarcity of domain-specific data and the necessity for nuanced understanding in multi-turn interactions.

3.1 Preference Optimization in Language Models

One of the key approaches to improving language models involves aligning them with human preferences. This alignment helps models generate responses that are not only coherent but also contextually appropriate and tailored to specific conversational objectives. Traditional approaches, such as Reinforcement Learning from Human Feedback (RLHF) [3], involve training a separate reward model based on human evaluations of model outputs. This reward model then guides the language model through reinforcement learning to produce preferred responses. Although effective, RLHF can be complex and resource-intensive due to the necessity of maintaining a distinct reward model and implementing reinforcement learning algorithms [4].

Direct Preference Optimization (DPO) [1] offers a more streamlined alternative by directly optimizing the language model using preference data, eliminating the need for a separate reward model and the complexities of reinforcement learning. DPO establishes a direct mapping between LLM policies and reward functions, enabling the training of an LLM to satisfy preference data through a straightforward cross-entropy loss.

3.2 Synthetic Data Generation and Iterative Self-Improvement

The challenge of data scarcity in specialized dialogue domains has led researchers to explore synthetic data generation and iterative self-improvement techniques to enhance language models. Several notable approaches have been proposed in this context.

Pace et al. [5] introduced *West of N*, a method for synthetic preference data generation. They generate multiple responses from an LLM and use a reward model to score these responses, selecting the best and worst ones to form preference pairs. These synthetic preference pairs are then used to further train the same reward model, improving its ability to align with human preferences. Unlike our framework, which directly trains the agent model using Direct Preference Optimization (DPO) with preference data generated through look-ahead simulations involving a user model, *West of N* focuses on refining the reward model itself.

Madaan et al. [6] proposed *Self-Refine*, a technique where during inference, the model generates an initial output and then provides feedback on its own response to iteratively refine it. This process enhances the output quality without additional training or data generation.

Yuan et al. [7] presented *Self-Rewarding Language Models*, where the model generates multiple responses and uses an LLM-as-a-Judge mechanism to evaluate them, creating its own preference dataset. This preference data is then used to iteratively train the model using DPO, enhancing both response generation and internal reward modeling. While similar in utilizing DPO, our framework differs by relying on an external oracle evaluator for preference data and simulating conversation paths using a user model, rather than self-assessment.

Xie et al. [8] explored integrating Monte Carlo Tree Search (MCTS) with iterative preference learning to improve reasoning capabilities in LLMs. Their method uses MCTS to generate and evaluate step-level reasoning paths, collecting preference data at each step to refine the model using DPO. The focus is on reasoning paths, whereas our framework simulates full conversation trajectories with a *Preference Tree with Look-Ahead* and a user model, emphasizing goal-oriented dialogue.

Liang et al. [9] introduced *I-SHEEP* (Iterative Self-EnHancEmEnt Paradigm), where the model generates synthetic data, self-assesses it, and filters out low-quality responses. The filtered data is then used for Supervised Fine-Tuning (SFT), allowing the model to improve iteratively without human supervision. Unlike our approach, which generates preference data and applies DPO for iterative improvement with look-ahead conversation simulations, *I-SHEEP* focuses on generating general synthetic data for SFT rather than preference data.

Additionally, prior work on preference trees for reasoning tasks [10] involves constructing preference trees for tasks such as planning, multi-turn interaction, and problem-solving in reasoning, coding, and math. While they utilize preference trees for reasoning, our framework employs a *Preference Tree with Look-Ahead* to simulate conversational pathways with a user model, specifically tailored for goal-oriented dialogue systems.

These methodologies highlight the effectiveness of combining synthetic data generation with iterative self-improvement to address data scarcity and complex interaction dynamics. Our proposed *Preference Tree with Look-Ahead* builds upon these approaches by incorporating strategic exploration of future conversational paths with a user model and generating nuanced preference data evaluated by an oracle. This allows us to directly train the agent model using DPO, enhancing its decision-making in specialized goal-oriented dialogue domains such as Motivational Interviewing.

3.3 Motivational Interviewing and AI Dialogue Systems

Motivational Interviewing (MI) is a client-centered counseling approach aimed at eliciting behavioral change by helping clients explore and resolve ambivalence [11]. Implementing MI in AI dialogue systems presents unique challenges due to the need for empathy, adaptability, and the ability to interpret subtle conversational cues.

Previous research has explored the potential of LLMs in simulating MI sessions. Yosef et al. [2] utilized AI-generated patient simulations to assess MI sessions, highlighting the feasibility of virtual patients in training and evaluating therapeutic dialogues. Their work demonstrated that AI agents could engage in MI conversations to a certain extent but also underscored the limitations in capturing the full depth of human therapist-patient interactions.

In addition, Yosef fine-tuned therapist models using existing datasets specific to MI, demonstrating that such fine-tuning can improve the models performance in therapeutic settings [2]. We will compare the performance of these fine-tuned models with our approach, which employs Direct Preference Optimization (DPO) combined with the *Preference Tree with Look-Ahead* method for iterative improvement.

4 Methodology

Our methodology involves two main components: the *Preference Tree with Look-Ahead* method for preference data generation and an iterative training process to refine the agent model using DPO.

4.1 Preference Tree with Look-Ahead

The *Preference Tree with Look-Ahead* method systematically explores potential conversational paths by simulating multiple agent responses and their subsequent dialogue trajectories. This is intended to allow the agent to anticipate the long-term impact of its responses, as shown in Appendix A 2. The process is as follows:

1. **Agent Decision Point**: At each turn, the agent model generates N possible responses.
2. **Branch Initialization**: For each response, a new branch is created, and the response is appended to the conversation history.
3. **Look-Ahead Simulation**: Each branch simulates K future steps, alternating between the agent and the virtual patient, to anticipate the long-term implications of the agent's response.
4. **Oracle Evaluation**: An oracle evaluator assesses each branch based on predefined criteria (e.g., adherence to MI principles, empathy, goal progression) and assigns scores.
5. **Preference Recording**: The response with the highest score is considered the preferred response, and the one with the lowest score is the least preferred. The preference tuple is recorded in the dataset.
6. **Conversation Update**: The conversation continues with the preferred response, and the process repeats until a termination condition is met (e.g., reaching maximum conversation length or achieving the goal).

By considering future conversation trajectories, the agent is expected to learn to make decisions that are not only immediately appropriate but also beneficial in the long term. (Figure 2)

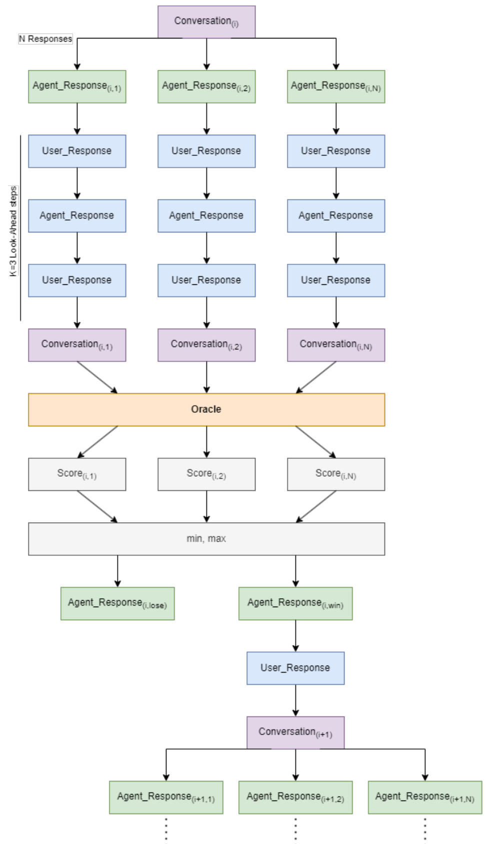

Figure 2: **Preference Tree Generation Process.** The figure shows how a preference tree is used to generate preference data. At each conversation step i, the agent generates N possible responses, and each branch simulates the conversation through several look-ahead steps. These branches represent possible future dialogue paths. An oracle evaluates each path, assigning scores to determine the best ($response_{i,win}$) and worst ($response_{i,lose}$) outcomes. After selecting the winning response, the user model replies, advancing to the next conversation step $conversation_{i+1}$, and the process repeats. This way, each preference tree produces multiple preference samples, with each sample consisting of a tuple $(conversation_i, response_{i,lose}, response_{i,win})$.

4.2 Preference Tree Optimization Framework (PTOF)

This process forms the Preference Tree Optimization Framework (PTOF). The agent model is iteratively improved through cycles of preference data generation and training using DPO.

1. **Initial Training**: The agent model is initially trained on available data or pre-trained weights.
2. **Preference Data Generation**: Using the current agent model, the *Preference Tree with Look-Ahead* method generates new preference data, capturing the agent's strengths and weaknesses.
3. **Model Update**: The agent model is fine-tuned using DPO on the newly generated preference data, optimizing it directly based on preferences without the need for a reward model.
4. **Evaluation**: The updated model is evaluated using predefined metrics to assess improvements.
5. **Iteration**: Steps 2-4 are repeated, allowing the agent to improve over time through continuous learning.

This process balances exploration (generating new conversational paths) and exploitation (refining the agent's responses), leading to incremental enhancements in performance.

Algorithm 1 Preference Tree Optimization Framework (PTOF)

Require: Initial agent model $A^{(0)}$, User model U, Oracle evaluator O, Maximum conversation length L, Look-ahead steps K, Branching factor N, Trees per iteration T, Total iterations I
Ensure: Sequence of optimized agent models $\{A^{(1)}, A^{(2)}, \dots, A^{(I)}\}$
for $i = 1$ to I **do**
    Initialize preference dataset: $D^{(i)} \leftarrow \{\}$
    for $t = 1$ to T **do**
        Assign user role: $U_t \leftarrow U$
        Generate preference tree: $P^{(t)} \leftarrow \text{GeneratePreferenceTree}(A^{(i-1)}, U_t, O, L, K, N)2$
        Aggregate preferences: $D^{(i)} \leftarrow D^{(i)} \cup P^{(t)}$
    end for
    Optimize Agent Model with DPO: $A^{(i)} \leftarrow \text{DPO}(A^{(i-1)}, D^{(i)})$
end for
Return: Optimized agent models $\{A^{(1)}, A^{(2)}, \dots, A^{(I)}\}$

5 Experimental Setup

To evaluate our proposed framework, we conducted a series of initial experiments in the Motivational Interviewing (MI) domain. The experimental setup is detailed as follows:

5.1 Models and Tools

- **Agent Model**: We utilized *Llama-2-7B* as the base model for the therapist agent.
- **User Model**: Virtual patients were simulated using *GPT-3.5*, based on guidelines from previous MI research [2]. Each patient is defined by parameters such as gender, age, problem (smoking/obesity), duration, prior attempts to resolve the issue, and cooperation level, creating 96 unique profiles to capture diverse challenges and attitudes toward counseling.
- **Oracle Evaluator**: GPT-3.5 model is used as the oracle evaluator, using specific questionnaires designed to assess MI adherence and conversational quality based on the guidelines from previous research [2] and detailed in Table 2.

5.2 Experimental Variables

- **Look-Ahead Depths**: We tested two different look-ahead depths: 0 (no look-ahead) and 3. This variable assesses the impact of anticipating future conversational turns on the agent's performance.
- **Iterations per Look-Ahead**: For each look-ahead depth, we conducted 3 iterative training cycles. Each iteration involved:
    1. **Preference Data Generation**: Utilizing the *Preference Tree with Look-Ahead* method to generate preference tuples based on simulated conversational paths.
    2. **Model Fine-Tuning**: Applying Direct Preference Optimization (DPO) to fine-tune the agent model using the newly generated preference data.

5.3 Data Collection

After each iteration, we generated a set of conversations to evaluate the agent's performance:

- **Number of Conversations**: For each trained model, we conducted 96 separate conversations with virtual patients to ensure a comprehensive assessment.
- **Evaluation Metrics**: Each conversation was scored by the oracle evaluator based on two distinct questionnaires designed to measure MI adherence and overall conversational quality, detailed in Table 2.

6 Preliminary Results

To assess the efficacy of the proposed Preference Tree Optimization Framework (PTOF), we conducted initial experiments concentrating on two distinct look-ahead depths: 0 and 3. Each configuration was subjected to three iterative training cycles, and their performances were benchmarked against the baseline model, *Llama-2-7B*.

6.1 Performance Metrics

The agent's effectiveness was evaluated using two primary metrics derived from the oracle evaluator's questionnaires (Table 2):

- **Session Satisfaction (Q1)**: This metric aggregates scores from Questionnaire 1, assessing overall satisfaction, content relevance, motivation facilitation, learning outcomes, and applicability to everyday life.
- **Working Alliance (Q2)**: This metric aggregates scores from Questionnaire 2, evaluating the therapist's interpersonal skills, empathy, communication effectiveness, and ability to establish a collaborative relationship.
- **Final Score**: Calculated as the average of Session Satisfaction and Working Alliance scores, this provides a comprehensive indicator of overall performance.

6.2 Results Overview

Table 1 presents the average scores and standard deviations for **Session Satisfaction (Q1)**, **Working Alliance (Q2)**, and the **Final Score** across the baseline model and the PTOF-enhanced models with look-ahead depths of 0 and 3 over three iterative training cycles.

Table 1: Average Performance Scores and Standard Deviations Across Iterations

Model	**Session Satisfaction (Q1)**		**Working Alliance (Q2)**		**Final Score**	
	Mean	SD	Mean	SD	Mean	SD
Llama-2-7B (Baseline)	2.952	1.359	3.185	0.944	3.069	1.103
Look-Ahead Depth 0						
M1_L0	3.262	1.396	2.979	1.044	3.121	1.186
M2_L0	3.575	1.177	3.230	0.770	3.402	0.916
M3_L0	3.473	1.228	3.277	0.858	3.375	0.992
Look-Ahead Depth 3						
M1_L3	3.175	1.436	3.194	0.977	3.185	1.160
M2_L3	3.331	1.236	3.388	0.712	3.360	0.917
M3_L3	**3.704**	1.026	**3.581**	0.639	**3.642**	0.788

6.3 Performance Evaluation

Baseline Performance:
The baseline *Llama-2-7B* model showed high variability in session satisfaction scores, indicating inconsistent handling of diverse conversational scenarios.

Look-Ahead Depth 0:
Models with a look-ahead depth of 0 (*M3_L0*) consistently outperformed the baseline across all metrics. While improvements in the Working Alliance were modest, this setting enhanced overall stability and dialogue quality.

Look-Ahead Depth 3:
Models with a look-ahead depth of 3 (*M3_L3*) achieved the highest scores in Session Satisfaction, Working

Alliance, and Final Score. Additionally, reduced standard deviations in the Working Alliance metric indicate greater consistency and reliability. These improvements suggest that deeper look-ahead mechanisms enable the agent to better anticipate and manage future conversational turns, aligning closely with Motivational Interviewing principles.

Significance Testing:

We performed an Analysis of Variance (ANOVA) followed by Tukey's Honest Significant Difference (HSD) tests to evaluate statistical significance among the baseline (*Base*), look-ahead depth 0 (*M3_L0*), and depth 3 (*M3_L3*) models.

- **ANOVA Results**: Significant differences were observed across all metrics:
  - **Final Score**: $F(2, XYZ) = 8.408$, $p = 0.00028$
  - **Session Satisfaction (Q1)**: $F(2, XYZ) = 9.702$, $p = 0.00008$
  - **Working Alliance (Q2)**: $F(2, XYZ) = 6.055$, $p = 0.00266$

  Note: An F-statistic with $p < 0.05$ indicates statistical significance.

- **Tukey's HSD Tests**:
  - **Final Score**: *M3_L3* significantly outperformed *Base* ($\Delta = 0.5737$, $p = 0.0002$). Differences between *Base* and *M3_L0* ($p = 0.0751$) and between *M3_L0* and *M3_L3* ($p = 0.1376$) were not significant.
  - **Session Satisfaction (Q1)**: Both *M3_L0* ($\Delta = 0.5208$, $p = 0.0088$) and *M3_L3* ($\Delta = 0.7521$, $p < 0.0001$) significantly improved over *Base*. The difference between *M3_L0* and *M3_L3* was not significant ($p = 0.384$).
  - **Working Alliance (Q2)**: *M3_L3* significantly outperformed both *Base* ($\Delta = 0.3953$, $p = 0.0029$) and *M3_L0* ($\Delta = 0.3036$, $p = 0.03$).

  Note: Δ denotes the mean difference between models.

These findings demonstrate that the Preference Tree Optimization Framework (PTOF), particularly with a look-ahead depth of 3, significantly enhances dialogue performance compared to the baseline. The superior performance and consistency of *M3_L3* highlight the effectiveness of deeper look-ahead mechanisms in achieving high-quality, goal-oriented conversations.

Graphical Representations

To gain deeper insights into the agent's performance, we present a series of graphical analyses that complement the quantitative results. These visualizations illustrate the distribution, comparison, and efficiency of different models under various configurations.

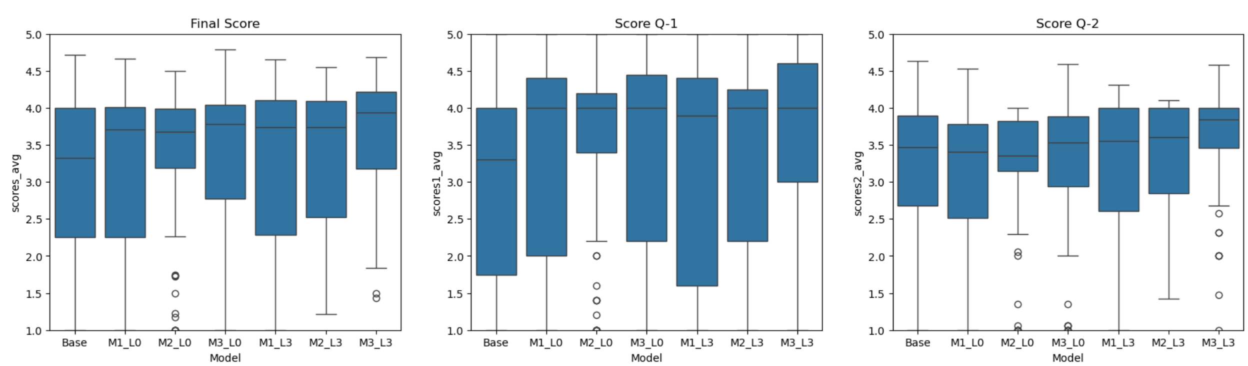

Figure 3: **Distribution of Performance Scores**

Boxplots showing the distribution of scores for **Questionnaire 1 (Session Satisfaction)**, **Questionnaire 2 (Working Alliance)**, and the **Final Score** across all evaluated models. These plots highlight the central tendency and variability of each metric, providing a comprehensive view of model performance.

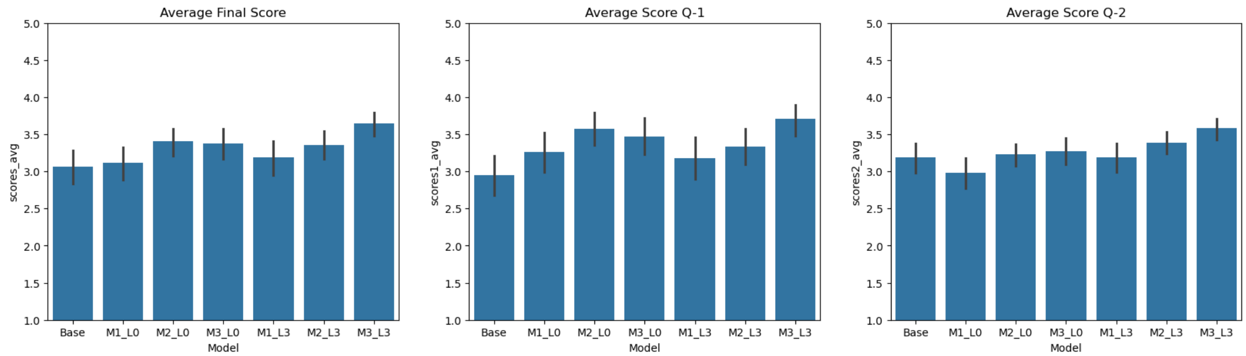

Figure 4: **Comparative Performance Analysis**
Bar charts comparing the average scores of **Questionnaire 1 (Session Satisfaction)**, **Questionnaire 2 (Working Alliance)**, and the **Final Score** between the *Baseline* model (*Llama-2-7B*) and the *PTOF-enhanced* models with different look-ahead depths. This comparison underscores the relative improvements achieved through the Preference Tree Optimization Framework.

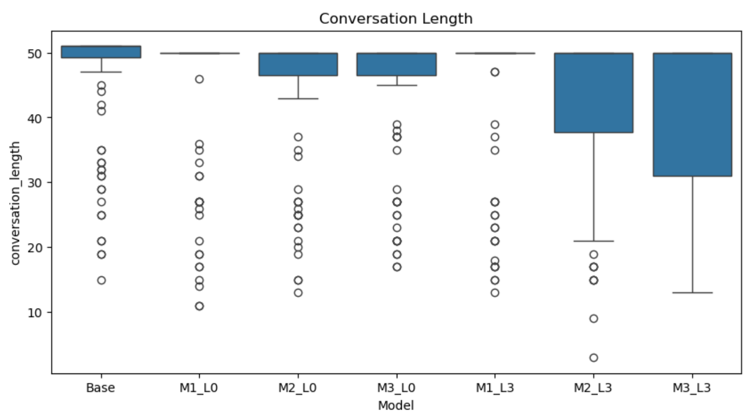

Figure 5: **Conversation Length Distribution**
Boxplot illustrating the distribution of conversation lengths across different models. This figure highlights the agent's ability to efficiently conclude conversations.

Conclusion:
Preliminary results confirm that PTOF effectively improves goal-oriented dialogue systems. Models with a look-ahead depth of 3 not only surpass the baseline but also show enhanced consistency and reliability, which are essential for specialized applications like Motivational Interviewing. These outcomes support incorporating strategic foresight in training dialogue systems to achieve superior conversational quality and user satisfaction.

7 Expected Contributions

The expected contributions are:

1. **A Novel Preference Data Generation Method**: Introduce the Preference Tree with Look-Ahead as an effective way to generate rich preference data for goal-oriented dialogues.

2. **An Iterative Training Framework**: Demonstrate how iterative refinement using DPO can enhance agent models in specialized domains.

3. **Advancements in MI Dialogue Systems**: Improve the effectiveness of AI agents in MI conversations, with potential implications for mental health interventions.

8 Research Road-map

1. **Phase 1 (May - June 2024)**: Develop the *Preference Tree with Look-Ahead* algorithm and validate its ability to generate meaningful preference data.
2. **Phase 2 (July - August 2024)**: Implement the iterative training process using DPO and conduct initial experiments to assess improvements.
3. **Phase 3 (September - November 2024)**: Perform extensive evaluations against baselines using the defined metrics, analyzing the agent's performance in MI dialogues.
4. **Phase 4 (December 2024 - February 2025)**: Refine the models and methods based on findings, explore scalability, and prepare for potential real-world applications.
5. **Phase 5 (March - April 2025)**: Publish results, share methodologies, and consider applications to other domains.

References

[1] Rafael Rafailov et al. "Direct Preference Optimization: Your Language Model is Secretly a Reward Model". In: *arXiv preprint arXiv:2305.18290* (2023). Accepted at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023). URL: `https://arxiv.org/abs/2305.18290`.

[2] Stav Yosef et al. "The Journey Towards an Automatic Mental Health Therapist". In: *Preprint* (2024).

[3] Paul F. Christiano et al. "Deep Reinforcement Learning from Human Preferences". In: *arXiv preprint arXiv:1706.03741* (2017). Presented at the 31st Conference on Neural Information Processing Systems (NeurIPS 2017). URL: `https://arxiv.org/abs/1706.03741`.

[4] Long Ouyang et al. "Training Language Models to Follow Instructions with Human Feedback". In: *arXiv preprint arXiv:2203.02155* (2022). Work by the OpenAI team. URL: `https://arxiv.org/abs/2203.02155`.

[5] Alizée Pace et al. "West-of-N: Synthetic Preference Generation for Improved Reward Modeling". In: *arXiv preprint arXiv:2401.12086* (2024). URL: `https://arxiv.org/abs/2401.12086`.

[6] Aman Madaan et al. "SELF-REFINE: Iterative Refinement with Self-Feedback". In: *arXiv preprint arXiv:2303.17651* (2023). Preprint, under review. URL: `https://selfrefine.info/`.

[7] Weizhe Yuan et al. "Self-Rewarding Language Models". In: *arXiv preprint arXiv:2401.10020* (2024). URL: `https://arxiv.org/abs/2401.10020`.

[8] Yuxi Xie et al. "Monte Carlo Tree Search Boosts Reasoning via Iterative Preference Learning". In: *arXiv preprint arXiv:2405.00451v2* (2024). Preprint, under review. URL: `https://github.com/YuxiXie/MCTS-DPO`.

[9] Yiming Liang et al. "I-SHEEP: Self-Alignment of LLM from Scratch through an Iterative Self-Enhancement Paradigm". In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2024). Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved. URL: `https://www.arxiv.org/abs/2408.08072`.

[10] Lifan Yuan et al. "Advancing LLM Reasoning Generalists with Preference Trees". In: *arXiv preprint arXiv:2404.02078* (2024). Preprint. URL: `https://github.com/OpenBMB/Eurus`.

[11] William R. Miller and Stephen Rollnick. "[Book Review] Motivational Interviewing, Preparing People to Change Addictive Behavior". In: *Journal of Studies on Alcohol* 54 (1993), pp. 507–507.

9 Appendix

9.1 Appendix A: Preference Tree with Look-Ahead Algorithm

Algorithm 2 Preference Tree with Look-Ahead

Require:

- Agent model A
- User model U
- Oracle evaluator O
- Maximum conversation length L
- Look-ahead depth K
- Number of agent response candidates N

Ensure: Preference dataset D

Initialize preference dataset: $D \leftarrow \emptyset$
Initialize conversation history: $C \leftarrow \emptyset$
Set initial context in C
while length of $C < L$ **do**
  Agent's Decision Point
  Generate N candidate responses: $R \leftarrow \{r_1, r_2, \ldots, r_N\}$
  Initialize list to store branch scores: $S \leftarrow \emptyset$
  for each response r_i in R **do**
    Initialize Branch
    Create a copy of current history: $C_i \leftarrow C$
    Append agent's response: $C_i \leftarrow C_i \cup \{r_i\}$
    Simulate Look-Ahead Steps
    Initialize step counter: $steps \leftarrow 0$
    Set next turn to User: $current\_turn \leftarrow$ User
    while $steps < K$ **and** not termination condition met **do**
      if $current\_turn =$ User **then**
        Generate user response: $u \leftarrow U(C_i)$
        Append user response: $C_i \leftarrow C_i \cup \{u\}$
        Switch turn to Agent: $current\_turn \leftarrow$ Agent
      else
        Generate agent response: $a \leftarrow A(C_i)$
        Append agent response: $C_i \leftarrow C_i \cup \{a\}$
        Switch turn to User: $current\_turn \leftarrow$ User
      end if
      Increment step counter: $steps \leftarrow steps + 1$
    end while
    Evaluate Branch
    Compute branch score: $s_i \leftarrow O(C_i)$
    Store score: $S \leftarrow S \cup \{s_i\}$
  end for
  Determine Preferences
  Identify index of preferred response: $w \leftarrow \arg\max(S)$
  Identify index of least preferred response: $l \leftarrow \arg\min(S)$
  Extract preferred and least preferred responses: $r_w \leftarrow R[w]$, $r_l \leftarrow R[l]$
  Record Preference Tuple
  Add to dataset: $D \leftarrow D \cup \{(C, r_w, r_l)\}$
  Update Conversation History
  Append preferred response: $C \leftarrow C \cup \{r_w\}$
  Generate user reply: $u \leftarrow U(C)$
  Append user reply: $C \leftarrow C \cup \{u\}$
  if termination condition met **then**
    Exit Loop
    break
  end if
end while
Return Preference dataset D

9.2 Appendix B: Evaluation Questionnaires for Therapist Performance

Table 2: The questions posed to the LLM for evaluating the performance of the therapist.

	Questionnaire 1 (session satisfaction)
Q1	Your overall satisfaction with the chat?
Q2	Your overall satisfaction with the content of the chat?
Q3	To what extent do you feel the chat facilitated motivation?
Q4	Did you learn anything?
Q5	To what extent was this learning relevant to your everyday life?
	Questionnaire 2 (working alliance)
Q1	The therapist gave me a sense of who it was.
Q2	The therapist revealed what it was thinking.
Q3	The therapist shared its feelings with me.
Q4	The therapist seemed to know how I was feeling.
Q5	The therapist seemed to understand me.
Q6	The therapist put itself in my shoes.
Q7	The therapist seemed comfortable talking with me.
Q8	The therapist seemed relaxed and secure when talking with me.
Q9	The therapist took charge of the conversation.
Q10	The therapist let me know when it was happy or sad.
Q11	The therapist didn't have difficulty finding words to express itself.
Q12	The therapist was able to express itself verbally.
Q13	I would describe the therapist as a "warm" communication partner.
Q14	The therapist did not judge me.
Q15	The therapist communicated with me as though we were equals.
Q16	The therapist made me feel like it cared about me.
Q17	The therapist made me feel close to it.