Expectation Confirmation Preference Optimization for Multi-Turn Conversational Recommendation Agent

Anonymous ACL submission

Abstract

Recent advancements in Large Language Models (LLMs) have significantly propelled the development of Conversational Recommendation Agents (CRAs). However, these agents often generate short-sighted responses that fail to sustain user guidance and meet expectations. Although preference optimization has proven effective in aligning LLMs with user expectations, it remains costly and performs poorly in multi-turn dialogue. To address this challenge, we introduce a novel multi-turn preference optimization (MTPO) paradigm **ECPO**[1], which leverages Expectation Confirmation Theory to explicitly model the evolution of user satisfaction throughout multi-turn dialogues, uncovering the underlying causes of dissatisfaction. These causes can be utilized to support targeted optimization of unsatisfactory responses, thereby achieving turn-level preference optimization. ECPO ingeniously eliminates the significant sampling overhead of existing MTPO methods while ensuring the optimization process drives meaningful improvements. To support ECPO, we introduce an LLM-based user simulator, **AILO**, to simulate user feedback and perform expectation confirmation during conversational recommendations. Experimental results show that ECPO significantly enhances CRA's interaction capabilities, delivering notable improvements in both efficiency and effectiveness over existing MTPO methods.

1 Introduction

Conversational Recommendation Systems (CRSs) leverage multi-turn natural language interactions to gradually uncover user interests and subsequently recommend items aligned with their preferences (Jannach et al., 2021; Gao et al., 2021). Powered by the advanced text generation and tool-calling capabilities of Large Language Models (LLMs) (Wang et al., 2024a), LLM-based Conversational Recommendation Agents (CRAs) (Gao et al., 2023; Huang et al., 2023; Fang et al., 2024) are emerging as a mainstream paradigm for delivering accurate, interpretable, and emotionally engaging personalized services. However, the responses generated by current CRAs often appear rigid, lacking proactivity and flexibility. This is mainly because the pretraining objectives of LLMs are predominantly focused on short-sighted next-token prediction (Ouyang et al., 2022). As a result, their ability to sustain long-term interactions and provide dynamic guidance is limited, making it difficult to meet human expectations in conversation.

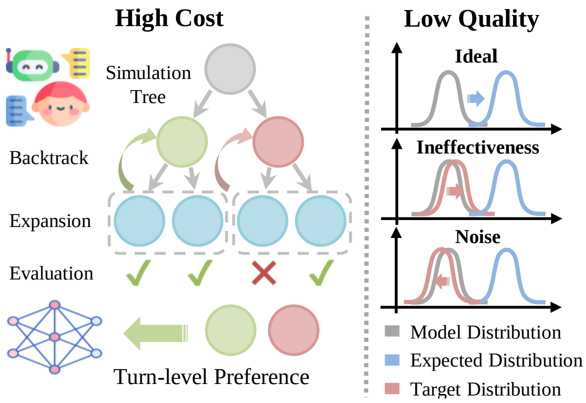

Figure 1: Existing MTPO methods have three inherent challenges: (1) Tree simulation incurs additional sampling **Costs**. (2) In CRA tasks, LLMs struggle to generate **Effective** positive examples through self-sampling. (3) Simulated environmental **Noise** in the expansion and evaluation may be incorporated into preference relations, leading the model to update in the wrong direction.

To address this challenge, aligning CRAs with human expectations presents a viable solution. Preference optimization has demonstrated success in aligning LLM outputs with user preferences (Schulman et al., 2017; Ouyang et al., 2022; Rafailov et al., 2024). Its core principle involves sampling multiple candidate outputs from the LLM

[1]The data and code are available at `https://anonymous.4open.science/r/ECPO-51B8`

and increasing the probability of those that align with user expectations. However, conversational recommendation is a multi-turn dialogue task, and applying preference optimization to this process presents great challenges. The main difficulty is that user preferences change in each dialogue turn and dynamically evolve as the conversation progresses. Most existing Multi-Turn Preference Optimization (MTPO) methods simply treat each turn equally, failing to capture turn-level preference relationships (Ulmer et al., 2024; Sun et al., 2024). Several recent works (Jin et al., 2024; Xie et al., 2024) try to infer turn-level preference relationships through tree-based simulations. As illustrated in Fig. 1, these approaches introduce three inherent challenges: (1) To obtain turn-level preference, it is necessary to sample multiple candidate responses at each turn and simulate the entire conversation to evaluate preferences for intermediate turns, resulting in **significant sampling overhead**. (2) In multi-turn conversational recommendation tasks, LLMs struggle to generate **effective** positive outputs through self-sampling. (3) Evaluating preferences for intermediate turns relies on the simulated environment, whose randomness may introduce additional **noise** into preference relationships, leading to suboptimal performance of the aligned CRA. Overcoming these limitations is essential to aligning CRAs with human expectations. This leads to a critical question: *Is there a way to construct high-quality turn-level preference relationships without additional sampling and evaluation?*

A problem well stated is a problem half solved. The core idea of this paper is to explicitly model how user satisfaction evolves throughout multi-turn dialogues and uncover the underlying causes of dissatisfaction. By identifying and addressing the root causes of low satisfaction, we can naturally construct responses that better align with user expectations. Expectation Confirmation Theory (ECT) (Oliver, 1977, 1980) tells us satisfaction is a subjective feeling that arises from the comparison between an individual's initial expectations and the perceived actual performance or outcomes. When applied to the context of conversational recommendation, this can be understood as: during a dialogue, a user has specific expectations for the system's response in each turn. Upon receiving the actual response, the user evaluates it by comparing it with their initial expectations, assigning a subjective satisfaction score based on the perceived gap.

Motivated by this, we propose **E**xpectation **C**onfirmation **P**reference **O**ptimization (**ECPO**), which comprises three key steps: (1) *Forward Expectation Confirmation* to identify unsatisfactory responses and uncover their root causes; (2) *Backward Expectation Derivation* to rewrite the unsatisfactory responses based on these causes; (3) *Preference Optimization* using the original and rewritten responses. Considering the high cost and potential bias associated with real users participating in the Expectation Confirmation (EC) process, we further introduce **AILO**, an LLM-based agent that simulates real users' **A**ctivities, **I**nterests, **L**anguage, and **O**rientations. During the dialogue, AILO acts as a user, providing diverse and realistic feedback as well as performing the EC process. Our contributions are summarized as follows:

- We introduce ECPO, a novel MTPO paradigm leveraging ECT to guide turn-level alignment in dialogues. To the best of our knowledge, this is the first preference optimization method tailored for LLM-based CRAs.
- To support ECPO, we introduce an LLM-based user simulator, AILO, which provides diverse and realistic feedback as well as performs the expectation confirmation process.
- We conduct extensive experiments on three datasets, demonstrating ECPO's exceptional performance in enhancing CRA's interactive capabilities and highlighting its significant advantages over existing MTPO methods in both efficiency and effectiveness.

2 Method

To better align multi-turn CRAs with human expectations, we propose **E**xpectation **C**onfirmation **P**reference **O**ptimization (**ECPO**). Its core idea is to leverage ECT to explicitly model the evolution of user satisfaction throughout multi-turn dialogues and construct turn-level preference relationships by identifying and addressing the root causes of dissatisfaction. A detailed description of ECPO is provided in Section 2.2. Additionally, we introduce a novel user simulator, AILO, which generates diverse and realistic user feedback while performing expectation confirmation (see Section 2.3).

2.1 Preliminary

We define the CRA as π[2], which leverages LLMs' planning and tool-calling capabilities to conduct

[2]The backbone of π is a tunable open-source LLM. In this paper, we use Llama-3.1-8B-Instruct (Grattafiori et al., 2024).

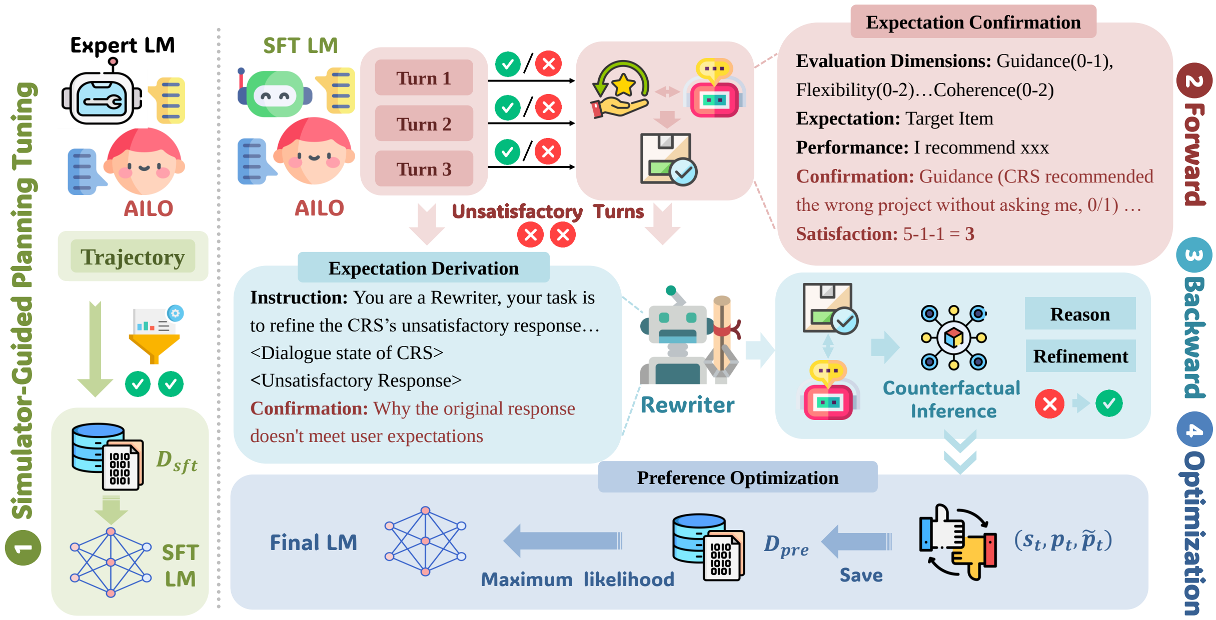

Figure 2: The ECPO process consists of four steps: (1) *Simulator-Guided Planning Tuning* to distill π_{sft} from the GPT-based CRA; (2) *Forward Expectation Confirmation* to identify unsatisfactory responses and uncover their root causes; (3) *Backward Expectation Derivation* to rewrite unsatisfactory responses based on the EC process; (4) *Preference Optimization* based on the original and rewritten responses.

multi-turn dialogues with a user U. Through iterative interactions, the agent elicits user preferences, retrieves relevant items from the external database $I = \{I_1, I_2, \dots, I_n\}$, and recommends the item that best matches the user's interests. Formally, at the t-th turn ($1 \leq t \leq T$), π performs internal reasoning cr_t and generates a response p_t, denoted as $\{cr_t, p_t\} = \pi(s_t)$, where s_t represents the dialogue state (e.g., dialogue history). We follow the setting proposed by iEvalLM (Wang et al., 2023), which assumes each user has a ground-truth item i^E. The goal of the CRA is to proactively guide users in conversations, providing a highly flexible and coherent user experience while successfully recommending the target item i^E. Formally, an interaction episode is:

$$H^T = \{u_0, (cr_1, p_1, u_1), \dots, (cr_T, p_T, u_T)\},$$

where u_t represents the user's utterance at turn t.

2.2 ECPO

In this section, we propose **ECPO**, an MTPO paradigm based on ECT. As shown in Figure 2, we first obtain the model π_{sft} through a *Simulator-Guided Planning Tuning* phase. Subsequently, ECPO is performed in three steps: *Forward Expectation Confirmation*, *Backward Expectation Derivation*, and *Preference Optimization*.

Simulation-Guided Planning Tuning. Existing CRS datasets (Kim et al., 2024) often lack an internal reasoning process, making them unsuitable for CRA's fine-tuning. To resolve this issue, we construct a new multi-turn conversational recommendation dataset that incorporates internal reasoning. This dataset is generated from dialogues between a GPT-4o mini-based CRA π_{GPT} and a user simulator U. We filter the trajectories based on whether the recommendation is successful, resulting in the dataset $\mathcal{D}_{\text{sft}}$. Subsequently, we perform supervised fine-tuning (SFT) on the CRA π:

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{(s_t, cr_t, p_t) \sim \mathcal{D}_{\text{sft}}} \left[-\log \pi_\theta(cr_t, p_t | s_t) \right] \quad (1)$$

Through this process, we obtain the CRA π_{sft}. However, SFT struggles to capture turn-level user preferences, making it insufficient to fully meet user expectations. To address this, we introduce ECPO, a low-cost and high-quality MTPO paradigm. For clarity, we omit the internal reasoning cr of the CRA in the subsequent formulations.

Forward Expectation Confirmation. Expectation Confirmation Theory tells us an individual's satisfaction arises from comparing actual performance against prior expectations. When applied to conversational recommendation, the evolution of user satisfaction can be modeled through the Expectation Confirmation (EC) process. In this paper, we adopt an extensible multi-dimensional scoring criterion with a maximum score of 5, consisting of *flexibility (0-2 points), coherence (0-2 points), and user guidance ability (0-1 point)* (Gao et al., 2021; Alkan et al., 2019). Formally, at the t-th turn, ECPO integrates the user expectation item i^E and the CRA's response p_t at this dialogue turn into an instruction prompt I_{ect}. The instruction is designed to explicitly simulate the user's inner monologue during the conversation: First, a user U evaluates the system's output against their expectations, assessing whether each dimension meets the corresponding requirement and assigning a sub-score to each aspect. These sub-scores are then aggregated to compute the overall satisfaction score r_t for p_t. We formulate the EC process as follows:

$$\{\text{CONF}_t, r_t\} = U(I_{\text{ect}}(i^E, h_t, p_t)),$$

where h_t is the dialogue history, CONF_t is a natural language explanation explicitly detailing why the user feels satisfied or dissatisfied at this turn. We then trace back the internal state s_t at the time of the CRS output p_t, together with the corresponding EC process CONF_t, and store it as a tuple $(s_t, p_t, \text{CONF}_t, r_t)$ for the subsequent phase.

Backward Expectation Derivation. Once each dialogue turn is assigned a satisfaction score via the EC process, we can identify responses that fail to meet user expectations. Next, we backtrack to the CRA state s_t and leverage CONF_t for counterfactual inference on how the CRA should have generated a response to better align with user expectations. Formally, at the t-th turn, ECPO integrates the EC process CONF_t and the unsatisfactory response p_t into an instruction prompt I_{bed}, which serves as the input for the Rewriter—an additional LLM introduced to refine unsatisfactory responses during backtracking. The Rewriter employs a slow thinking process, first generating a chain of thought (Wei et al., 2023) and then producing a refined response $\tilde{p}_t$:

$$\begin{aligned} \tilde{p}_t = \text{Rewriter}(I_{\text{bed}}(s_t, p_t, \text{CONF}_t)), \\ \text{where } r_t \leq \lambda \end{aligned} \tag{2}$$

Here, λ is a hyperparameter that defines the satisfaction threshold. If the user's satisfaction score r_t falls below λ, the response will undergo backtracking and rewriting. Meanwhile, to ensure that rewritten responses do not deviate too far from the π_{sft}, we require the Rewriter to make only limited modifications to the unsatisfactory response, rather than performing a complete rewrite.

After the backward process, we can collect these "original–rewritten" pairs from the training set to form our preference dataset, denoted as $\mathcal{D}_{\text{pre}} = \{(s_t, p_t, \tilde{p}_t) \mid r_t < \lambda\}$. This dataset consists of turn-level preference pairs, where the rewritten responses $\tilde{p}_t$ are statistically more likely to exhibit significant improvements over the original ones. This hypothesis has been empirically validated through our evaluation (cf. Appendix C.2).

Preference Optimization. After obtaining the turn-level preference dataset $\mathcal{D}_{\text{pre}}$, we can optimize π_{sft} through existing preference optimization methods. A typical implementation is Direct Preference Optimization (DPO) (Rafailov et al., 2024):

$$\mathcal{L}_{\text{DPO}}(\pi_\theta, \pi_{\text{sft}}) = \mathbb{E}_{s,\tilde{p}_t,p_t \sim \mathcal{D}_{\text{pre}}}\Bigg[-\log\sigma \left(\beta \log \frac{\pi_\theta(\tilde{p}_t \mid s_t)}{\pi_{\text{sft}}(\tilde{p}_t \mid s_t)} - \beta \log \frac{\pi_\theta(p_t \mid s_t)}{\pi_{\text{sft}}(p_t \mid s_t)}\right)\Bigg] \tag{3}$$

ECPO is both orthogonal and complementary to existing preference optimization methods. This enables seamless integration with various methods (e.g., KTO (Ethayarajh et al., 2024), SimPO (Meng et al., 2024)) based on specific task requirements and optimization goals. We further explore this integration in Appendix C.3.

Discussion Existing MTPO methods typically require completing the entire conversation before estimating the reward for each intermediate turn, and all positive samples must be generated through self-sampling. In contrast, ECPO implicitly assigns rewards at each turn through the EC process and provides the underlying reasons for these rewards in natural language. These reasons promote the proactive generation of positive samples for preference optimization instead of self-sampling. This paradigm not only eliminates additional sampling and evaluation costs but also ensures that preference relationships drive meaningful optimization. In the next section, we introduce AILO, a novel user simulator designed to support the EC process.

2.3 AILO

This paper aims to leverage the EC process to explicitly model how user satisfaction evolves throughout conversational recommendation, thereby guiding CRA to align with user expectations. However, considering the unacceptably

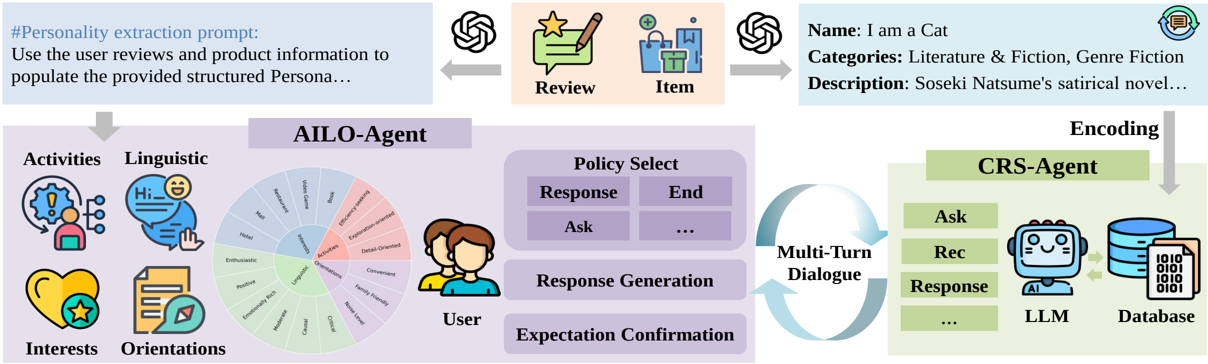

Figure 3: The illustration of the AILO, showing its *persona modeling*, *policy-based user simulation*. Figure also depicts the task of the CRA: interacting with the database, engaging in dialogue, and recommending items to AILO.

high costs and potential biases involved in human participation, we propose a new user simulator, **AILO**, an LLM-based agent that provides realistic and diverse user feedback. As shown in Figure 3, AILO consists of two components: *user persona modeling* and *policy-based user simulation*.

User Persona Modeling. Existing user simulators typically generate user personas through simple random sampling (Wang et al., 2024b), but this approach often results in unrealistic and less diverse personas. To address this, we propose **AILO**, a comprehensive user simulator for conversational recommendation. Inspired by the AIO theory (Wells et al., 1971) from consumer psychology, AILO defines user attributes across four dimensions: Activities, Interests, Language, and Orientations, thereby capturing the diverse characteristics that users may exhibit during conversational recommendations. For example, some users prioritize efficiency in recommendations, while others prefer engaging in in-depth discussions on specific topics. We employ GPT-4o (OpenAI et al., 2024) to infer user personas from real recommendation review datasets. This not only ensures the authenticity of personas but also enhances their diversity. To assess the diversity of AILO's personas, following Jin et al. (2024), we randomly sample 100 personas created by our method and those generated using the sampling method in RecAgent (Wang et al., 2024b), then compute the maximum ROUGE-L between each persona and the others. As shown in Figure 4, the ROUGE-L's distribution of AILO is significantly lower than RecAgent, indicating that AILO produces more diverse user personas.

Policy-Based User Simulation. Directly simulating user responses with LLMs may lead to role reversals and uncontrollable behavior (Zhu et al.,

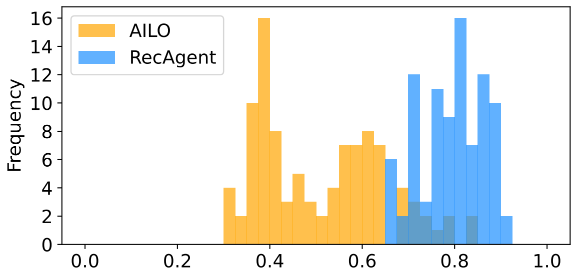

Figure 4: ROUGE-L with the Most Similar Persona.

2024). Therefore, we redefine the process of user response generation as a planning task executed in three steps: (1) *Response Policy Generation*: Based on the user's persona and the CRA's response p_t, the simulator U generates a response policy ur_t, such as "Asking for Recommendations". (2) *Response Content Generation*: Based on the response policy ur_t, the simulator generates the response u_t. (3) *Expectation Confirmation Process*: U generates the EC process CONF_t, computes the satisfaction score r_t, and outputs them in a structured format. Formally, the simulator produces:

$$\{ur_t, u_t, \text{CONF}_t, r_t\} = U(i^E, h_t, p_t) \qquad (4)$$

Here, i^E is the target item, and h_t represents the dialogue history. To verify the authenticity of AILO's simulated dialogue, we recruit annotators to compare 50 sets of dialogue trajectories generated by AILO and iEvalLM (Wang et al., 2023), assessing which one appears more human-like. The experimental results show that AILO outperforms iEvalLM in all cases, achieving a 100% win rate.

3 Experiments

To thoroughly evaluate the effectiveness of ECPO in enhancing multi-turn CRAs, we conduct extensive experiments, which are outlined as follows:

Backbone	Method	#Calls	Game			Book			Yelp		
			SR	**R**	**WR**	**SR**	**R**	**WR**	**SR**	**R**	**WR**
GPT-4o mini	ChatRec	$\mathcal{O}(N)$	0.37	0.45	0.09	0.46	0.47	0.13	0.24	0.30	0.12
	ReAct	$\mathcal{O}(M+2N)$	0.39	0.65	0.34	0.52	0.56	0.33	0.57	0.62	0.42
	MACRS	$\mathcal{O}(M+4N)$	0.36	0.65	0.15	0.63	0.71	0.01	0.40	0.41	0.02
	ActCRS	$\mathcal{O}(M+N)$	0.43	0.68	0.50	0.53	0.56	0.50	0.37	0.43	0.50
Llama-3.1 8B-Instruct	ChatRec	$\mathcal{O}(N)$	0.36	0.39	0.01	0.42	0.47	0.03	0.30	0.32	0.05
	ReAct	$\mathcal{O}(M+N)$	0.04	0.43	0.08	0.36	0.54	0.19	0.31	0.40	0.16
	MACRS	$\mathcal{O}(M+4N)$	0.24	0.34	0.00	0.36	0.39	0.01	0.22	0.24	0.01
	ActCRS	$\mathcal{O}(M+N)$	0.07	0.50	0.46	0.34	0.55	0.28	0.22	0.35	0.38
	+SGPT(Ours)	$\mathcal{O}(M+N)$	0.41	0.61	0.42	0.55	0.58	0.46	0.44	0.48	0.47
	+ECPO(Ours)	$\mathcal{O}(M+N)$	**0.47**	**0.63**	**0.56**	**0.56**	**0.60**	**0.70**	**0.49**	**0.53**	**0.69**

Table 1: Comparison with existing prompt-based CRAs. The "#Calls" column represents the number of LLM calls required to complete an entire dialogue. N denotes the number of dialogue turns, and M represents the number of times the LLM generates retrieval queries ($M \leq N$). **SR** (Success Rate) and **R** (Recall Rate) are recommendation metrics, while **WR** reflects the interactive capabilities.

- First, to validate the importance of ECPO alignment for CRAs, we compare existing prompt-based CRAs with those that have undergone ECPO alignment.
- Second, we comprehensively compare ECPO with existing MTPO methods to verify its efficiency and effectiveness.
- Finally, we thoroughly analyze the effectiveness of different components of ECPO and conduct evaluations of its performance under various experimental settings.

3.1 Experimental Setup

In this section, we briefly introduce the experimental settings. A more detailed elaboration and design motivations are presented in Appendix B.

Environments. Traditional CRS evaluation methods struggle to assess dynamic CRA tasks (Afzali et al., 2023). As discussed in Section 2.3, we follow and extend iEvalLM (Wang et al., 2023) by introducing AILO for our evaluations. Our experiments utilize the Amazon-Game, Amazon-Book[3], and Yelp[4] datasets to construct user personas and generate approximately 3,000 tasks for each dataset. During the training phase, we use 1,000 tasks to construct $\mathcal{D}_{\text{sft}}$ and 500 tasks to construct $\mathcal{D}_{\text{pre}}$. Following ReAct (Yao et al., 2023) and MACRS(Fang et al., 2024), we sample 100 tasks from each dataset for testing.

[3]https://github.com/hyp1231/AmazonReviews2023
[4]https://www.yelp.com/dataset

Baselines. Given the significant gap between traditional CRS and emerging LLM-based CRAs, we focus on comparing our approach with existing prompt-based CRAs (ChatRec (Gao et al., 2023), ReAct (Yao et al., 2023), MACRS (Fang et al., 2024), ActCRS) and MTPO methods (trajectory-level: SFT, KTO (Ethayarajh et al., 2024); turn-level: SDPO (Jin et al., 2024), SKTO). Notably, ActCRS is a straightforward CRA developed by us, that simultaneously generates a response strategy and the corresponding response. Due to its simplicity and effectiveness, we fine-tune ActCRS in our main experiments. Our backbone model is Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and we additionally provide results based on GPT-4o mini (OpenAI et al., 2024) as a reference.

Metrics. We evaluate CRAs across two dimensions: (1) **Recommendation Metrics**: Success Rate (**SR**) and Recall Rate (**R**). (2) **Dialogue Metric**: Win Rate (**WR**, (Li et al., 2023)), which measures interactivity compared to the expert CRA (GPT-based ActCRS in main experiments).

3.2 Comparison with Existing Prompt-Based CRA Frameworks

Analysis of Existing Prompt-Based CRAs. Table 1 summarizes the main experimental results on three recommendation datasets. First, we analyze the existing CRAs' results. We find that: (1) Stronger backbone models (GPT-4o mini) perform better as CRA framework complexity increases. In contrast, weaker models (Llama-3.1) struggle to

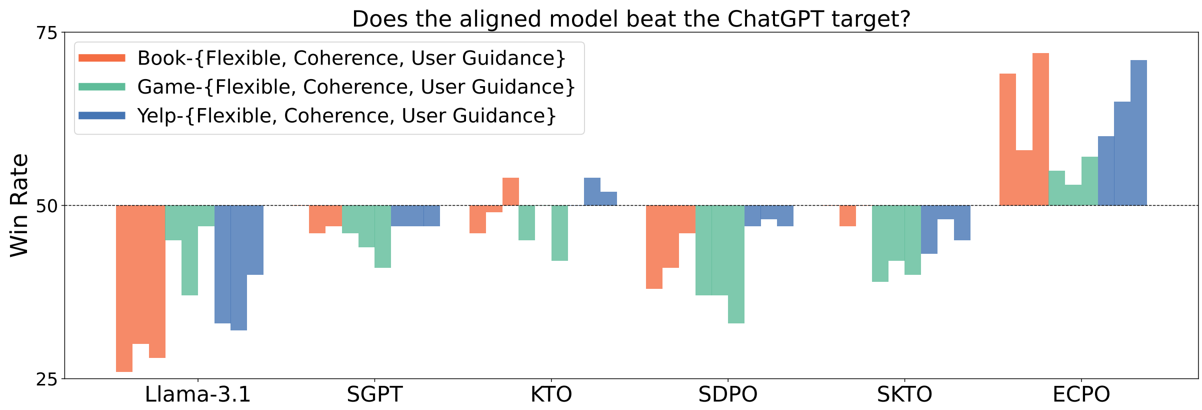

Figure 5: Comparison of aligned CRAs fine-tuned with different methods in terms of interactivity (flexibility, coherence, and user guidance) against the expert CRA.

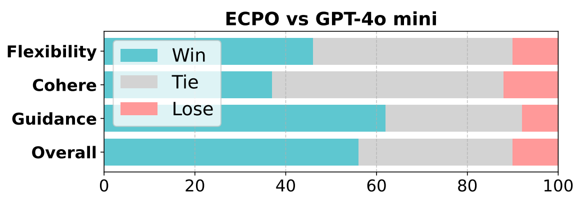

Figure 6: Human evaluation results.

benefit from more complex CRA frameworks. (2) ChatRec and MACRS can generate high-quality recommendations. However, ChatRec lacks interactivity, while MACRS's responses tend to be overly verbose, making conversations feel unnatural. In terms of **WR** (interactivity performance), their win rates are significantly lower than expert CRA, typically below 0.15. (3) No single prompt-based CRA demonstrates a clear advantage across all datasets and metrics. Moreover, as the number of calls increases, the performance gains gradually diminish. This observation highlights the growing importance of an alignment method for CRAs.

Effect of Alignment. We fine-tune the Llama-based ActCRS using SGPT + ECPO, and present the performance results in the table 1. After SGPT training, the recommendation metrics (**SR** and **R**) reach GPT-level performance, but interactivity remains inferior to the expert CRA. After ECPO training, the win rate significantly exceeded that of the GPT model (**WR** ranging from 0.56 to 0.7), highlighting the crucial role of the ECPO in enhancing the multi-turn conversation user experience.

3.3 Comparison with Existing MTPO Methods

In Figure 5, we compare ECPO with two categories of existing multi-turn alignment methods: trajectory-level methods (SFT, KTO) and turn-level preference optimization methods based on tree simulation (SDPO, SKTO). Specifically, we construct the preference dataset $\mathcal{D}_{\text{pre}}$ using each method in 500 simulation tasks. In these tasks, trajectory-level methods require sampling 1,000 trajectories, tree simulation methods require sampling 2,500 trajectories, whereas ECPO eliminates the need for additional sampling and efficiently utilizes only 500 trajectories. Experimental results show that the improvement of trajectory-level methods is limited, as they fail to effectively capture preference relationships at the turn level. Meanwhile, tree simulation methods, despite capturing these preferences, actually led to negative gains, likely due to noise interference. This finding highlights the challenges of CRA alignment. In contrast, ECPO, guided by the EC process, achieves the best performance while requiring the lowest cost, significantly outperforming all existing methods.

Additionally, we recruit human annotators to compare the win rates between the ECPO-aligned CRA and the expert CRA. The experimental results, as shown in Figure 6, indicate that ECPO demonstrates a significant advantage across all metrics, especially in flexibility and user guidance. To further understand how ECPO outperforms existing methods, we provide statistical analyses and case studies on dialogue styles in appendix C.5.

3.4 Effectiveness of the EC Process

Although we have demonstrated the effectiveness of ECPO in the main experiments, a natural question arises: *How does the turn-level EC process influence the performance of ECPO?* To investigate this further, we manually design rewriting instruc-

Method	Game			Book			Yelp		
	SR	R	WR	SR	R	WR	SR	R	WR
Llama-3.1	0.07	0.50	0.46	0.34	0.55	0.28	0.22	0.35	0.38
+SGPT	0.41	0.61	0.42	0.55	0.58	0.46	0.44	0.48	0.47
+ECPO-w/o EC	0.37	0.55	0.54	**0.56**	**0.60**	0.65	0.42	0.46	0.48
+ECPO	**0.47**	**0.63**	**0.56**	**0.56**	**0.60**	**0.70**	**0.49**	**0.53**	**0.69**

Table 2: Effectiveness of the EC process.

tions based on the test results of π_{sft}, identifying its issues and guiding the Rewriter to revise the responses generated by π_{sft}, to construct $\mathcal{D}_{\text{pre}}$. This approach, referred to as ECPO w/o EC, aims to replace each turn of the EC process with a unified analysis conducted by human to guide rewriting.

In Table 2, we find that ECPO w/o EC enhances interactivity to some extent but slightly reduces recommendation performance, with overall performance remaining significantly inferior to ECPO. This result underscores the importance of the turn-level EC process in the rewriting process.

3.5 Hyperparameter Analysis

In this section, we investigate the impact of the rewriting threshold λ, defined as the satisfaction score threshold below which responses are selected for rewriting and training. A higher λ leads to more response samples being backtracked and rewritten, resulting in a larger training dataset. Figure 7(a) presents the training results for λ values {1, 2, 3, 4}, while Figure 7(b) shows results from uniformly sampled subsets of the $\lambda = 4$ setting with varying sample sizes {50, 100, 200, 400, 800, 1600, All}.

The blue line represents the overall performance gain, while the pink line represents the performance improvement per individual sample. We observe that, in Figure 4(a), lower λ values lead to a more significant gain for individual samples. In contrast, in Figure 4(b), the performance improvement appears more irregular. This phenomenon is particularly interesting and aligns with intuition: when a sample has a lower satisfaction score, it often indicates critical issues, and addressing these issues results in a more noticeable performance gain.

3.6 Further Analysis

To comprehensively evaluate the superiority of ECPO, we conduct a series of further explorations: Is ECPO applicable across different CRA frameworks? (See Appendix C.1.) How do different optimization methods influence ECPO during pref-

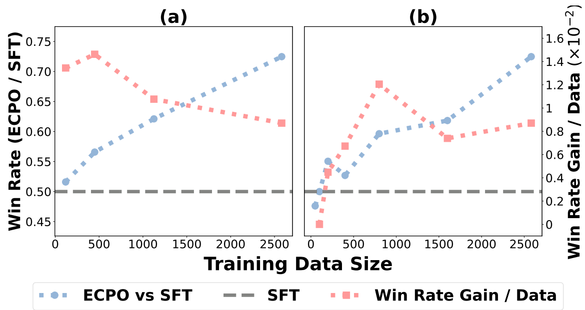

Figure 7: Hyperparameter analysis of λ.

erence optimization ? (See Appendix C.3.) Does ECPO remain effective under varying numbers of turns and recall settings? (See Appendix C.4.) How does the dialogue style change after ECPO alignment? (See Appendix C.5.)

4 Conclusion and Future Works

In this work, we propose ECPO, a novel MTPO paradigm designed to enhance the interaction capabilities of LLM-based CRAs, thereby improving user satisfaction. Our core design principle is to explicitly model the evolution of user satisfaction in multi-turn dialogues and achieve turn-level preference alignment by identifying and addressing the root causes of dissatisfaction. To support ECPO, we introduce a new LLM-based user simulator, that enable more diverse and realistic simulations as well as expectation confirmation. Extensive experiments on three recommendation datasets fully demonstrate the superiority of our proposals.

Although ECPO is designed for CRAs, we believe it can be extend to broader dialogue assistants by modeling the EC process across different domains. Another promising direction is enabling LLMs to generate simulated user expectations. This capability can establish an internal feedback mechanism, integrating the EC process into the reasoning phase to help O1/R1-style (DeepSeek-AI et al., 2025) dialogue assistants in refining responses and further enhancing user satisfaction.

Limitations

ECPO is a novel MTPO paradigm that performs turn-level preference optimization by simulating the dynamic evolution of user satisfaction across multi-turn conversations. However, the optimization process of ECPO largely relies on user simulation. To address this, we introduce AILO,a user simulator constructed from real user reviews to approximate a realistic and diverse user distribution as closely as possible. Although experimental results demonstrate that AILO outperforms existing user simulators in terms of authenticity and diversity, an inevitable gap may still exist between simulated and real users. This gap may lead to distribution shift issues in real-world scenarios.

Nevertheless, (1) although ECPO employs AILO as a user simulator to guide CRA alignment, the aligned CRA, when evaluated by real users, still exhibits significant advantages; (2) in dynamic multi-turn recommendation scenarios, user simulator has become a key concern for both academia and industry. Therefore, we believe that ECPO remains a significant contribution to multi-turn CRA alignment, and AILO also represents a valuable contribution to CRA evaluation.

Ethical Considerations

LLM-based multi-turn CRAs hold great potential in providing accurate, interpretable, and emotionally aware personalized recommendations. However, their development also raises ethical concerns, including potential biases, unfairness, privacy risks, and the reinforcement of filter bubbles. To mitigate these risks, we design AILO to represent a diverse range of users, aiming to reduce biases in ECPO alignment. Furthermore, we emphasize that all applications must operate under human supervision and oversight to ensure transparency and accountability. By maintaining a careful balance between technological advancement and social responsibility, we strive to foster the development of responsible AI systems.

References

Jafar Afzali, Aleksander Mark Drzewiecki, Krisztian Balog, and Shuo Zhang. 2023. Usersimcrs: A user simulation toolkit for evaluating conversational recommender systems. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23. ACM.

Oznur Alkan, Massimiliano Mattetti, Elizabeth M. Daly, Adi Botea, and Inge Vejsbjerg. 2019. Irf: Interactive recommendation through dialogue. *Preprint*, arXiv:1910.03040.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, and et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *Preprint*, arXiv:2402.01306.

Jiabao Fang, Shen Gao, Pengjie Ren, Xiuying Chen, Suzan Verberne, and Zhaochun Ren. 2024. A multi-agent conversational recommender system. *arXiv preprint arXiv:2402.01135*.

Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open*, 2:100–126.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*.

Aaron Grattafiori, Abhimanyu Dubey, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 720–730.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. 2023. Recommender ai agent: Integrating large language models for interactive recommendations. *arXiv preprint arXiv:2308.16505*.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Comput. Surv.*, 54(5).

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. Persuading across diverse domains: a dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706.

Minjin Kim, Minju Kim, Hana Kim, Beong woo Kwak, Soyeon Chun, Hyunseo Kim, SeongKu Kang, Youngjae Yu, Jinyoung Yeo, and Dongha Lee. 2024. Pearl: A review-driven persona-knowledge grounded conversational recommendation dataset. *Preprint*, arXiv:2403.04460.

Chuyi Kong, Yaxin Fan, Xiang Wan, Feng Jiang, and Benyou Wang. 2024. Platolm: Teaching llms in multi-round dialogue via a user simulator. *Preprint*, arXiv:2308.11534.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Preprint*, arXiv:2405.14734.

Richard L. Oliver. 1977. Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of Applied Psychology*, 62(4):480–486.

Richard L. Oliver. 1980. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, 17(4):460–469.

OpenAI, Josh Achiam, and et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Wayne Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Parrot: Enhancing multi-turn instruction following for large language models. *Preprint*, arXiv:2310.07301.

Dennis Ulmer, Elman Mansimov, Kaixiang Lin, Justin Sun, Xibin Gao, and Yi Zhang. 2024. Bootstrapping llm-based task-oriented dialogue agents via self-talk. *Preprint*, arXiv:2401.05033.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6).

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2024b. User behavior simulation with large language model based agents. *Preprint*, arXiv:2306.02552.

Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the evaluation for conversational recommendation in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 1929–1937. ACM.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

William D Wells, Douglas J Tigert, and Interests Activities. 1971. Opinions. *Journal of advertising research*, 11(4):27–35.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P. Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte carlo tree search boosts reasoning via iterative preference learning. *Preprint*, arXiv:2405.00451.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Xiaoyu Zhang, Ruobing Xie, Yougang Lyu, Xin Xin, Pengjie Ren, Mingfei Liang, Bo Zhang, Zhanhui Kang, Maarten de Rijke, and Zhaochun Ren. 2024. Towards empathetic conversational recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 84–93.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. A llm-based controllable, scalable, human-involved user simulator framework for conversational recommender systems. *Preprint*, arXiv:2405.08035.

A Related Work

Conversational Recommendation Systems. A CRS aims to engage users through natural language interaction, iteratively eliciting their preferences and providing personalized recommendations (Zhang et al., 2024). Research on CRS can be divided into two categories: attribute-based CRSs and generation-based CRSs (Jannach et al., 2021). While attribute-based CRSs rely on pre-defined templates, generation-based CRSs (Wang et al., 2022) enable more flexible interactions but are constrained by the limitations of traditional language models. In recent years, leveraging the powerful language capabilities and tool utilization of LLMs, researchers have begun developing CRAs, offering transformative solutions for conversational recommendation. ZSCRS (He et al., 2023) conducts an initial exploration of using LLMs directly as conversational recommenders. ChatRec (Gao et al., 2023) and InteRecAgent (Huang et al., 2023) integrate traditional recommendation models with LLMs, effectively enhancing the interactivity of the recommendation system. To further enhance dialogue flow control in CRS, MACRS (Fang et al., 2024) introduces a multi-agent framework to enable long-term strategic planning. Despite the extensive exploration of LLM-based frameworks for CRAs, the increasing inference costs of complex agent frameworks and the diminishing returns on performance gains have significantly limited their practicality in real-world scenarios. Hence, there is an urgent need for an alignment method that fine-tunes LLMs for CRA tasks, enabling them to better guide users and effectively enhance user satisfaction.

LLM Alignment. The objective of LLMs is to predict the next token in internet-scale corpora; however, this differs from the goal of "helpfully and safely following the user's instructions" (Ouyang et al., 2022). Therefore, it is necessary to align LLMs with human preferences to ensure the generation of safe, unbiased, and appropriate text (Schulman et al., 2017; Rafailov et al., 2024; Ethayarajh et al., 2024; Meng et al., 2024). In this paper, we focus on the problem of LLM alignment in multi-turn conversational recommendation (MTPO). Currently, most existing methods (Sun et al., 2024; Ulmer et al., 2024; Kong et al., 2024) simply treat each turn equally, failing to capture turn-level preference relationships. Another class of methods (Jin et al., 2024; Xie et al., 2024) employs tree-based simulation to infer **turn-level** preference relations. Specifically, these methods generate multiple candidate outputs at each intermediate turn, expand them into different dialogue subpaths, and simulate complete dialogues to obtain final rewards, thereby estimating the rewards for intermediate turns. However, these methods introduce additional sampling costs and struggle to establish high-quality preference relationships. Therefore, developing a method to construct high-quality turn-level preference relationships without additional sampling and evaluation is critical.

B Experimental Details

B.1 Simulator Details

Traditional CRS evaluation methods struggle to assess dynamic CRA tasks (Afzali et al., 2023). As discussed in Section 2.3, we follow and extend iEvalLM (Wang et al., 2023) by introducing AILO for our evaluations. Our experiments utilize the Amazon-Game, Amazon-Book[5], and Yelp[6] datasets, constructing 100 user personas for each dataset. We use GPT-4o mini as the backbone model of the AILO user simulator.

B.2 CRA Task Details

Task Construction. To generate high-quality conversational recommendation tasks, we extract positively rated items from each user's interaction history and designate them as ground-truth items for recommendation tasks. As a result, each user is assigned approximately 30 conversational recommendation tasks, yielding 3,000 simulation tasks per dataset (100 × 30). Additionally, we use all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) to embed items users have interacted with, constructing an external database I.

Task Execution. We model conversational recommendation as an agent-based task. As shown in Figure 3, during task execution, the CRA engages in multi-turn interactions with AILO via natural language and can optionally query the external retriever to obtain real item information from the database I for making recommendations. The primary objective of the CRA is to recommend the ground truth item using natural language. Depending on the agent framework, the retrieval query may vary: (1) some frameworks directly concatenate the dialogue history as input (ChatRec (Gao

[5] https://github.com/hyp1231/AmazonReviews2023
[6] https://www.yelp.com/dataset

Method	ChatRec			ReAct			ActCRS		
	SR	R	WR	SR	R	WR	SR	R	WR
GPT-4o mini	0.46	0.47	0.50	0.52	0.56	0.50	0.53	0.56	0.50
Llama-3.1	0.42	0.47	0.11	0.36	0.54	0.31	0.34	0.55	0.28
+*SGPT(Ours)*	0.47	0.53	0.39	**0.54**	**0.61**	0.41	0.55	0.58	0.46
+*ECPO(Ours)*	**0.50**	**0.54**	**0.41**	**0.54**	0.60	**0.49**	**0.56**	**0.60**	**0.70**

Table 3: Effectiveness on different CRAs.

et al., 2023)), while (2) others generate the query using an LLM. To foster further research, we have open-sourced the automated task generation process.

B.3 Implementation Details

Inference Details. In the main experiments, each CRA task consists of a maximum of 5 interaction turns, and each retiever will returns the top 5 most relevant items with the retrieval query. Following Yao et al. (2023); Fang et al. (2024), we sample 100 tasks for each dataset for testing. In all LLM inference processes, we set the temperature parameter of all models to 0.0 to eliminate randomness in local models. Additionally, we use the vllm (Kwon et al., 2023) framework to accelerate all inference processes.

Training Details. During training, we use the Llama-factory (Zheng et al., 2024) for LoRA (Hu et al., 2021) training. In our main experiments, we set the rewriting threshold to $\lambda = 4.0$. We randomly sample 1000 simulation tasks to construct the dataset D_{sft}, generating approximately 2000 input-output pairs, and randomly sample an additional 500 tasks for constructing dataset D_{pre}. For the stage of SGPT, we use a learning rate of $5e-05$ and a batch size of 8. For the stage of ECPO, we search two learning rates: $\{1e-06, 5e-07\}$, with a batch size of 32.

B.4 Human Evaluation Details

We conduct three sets of human evaluation experiments: (1) empirical verification of data before and after rewriting, (2) comparison of the reliability between AILO and iEval, and (3) evaluation of ECPO against expert models. In each experiment, we employ two annotators with an average education level of a bachelor's degree. To ensure fairness, we randomly shuffle and anonymize the data before annotation. This blind evaluation setup minimizes potential biases and improves the reliability of our results.

B.5 Evaluation Metrics.

We evaluate CRAs across two dimensions.

- **Recommendation Metrics**: Following iEvalLM (Wang et al., 2023), we report **SR** (success rate of recommending the ground-truth item) and **R** (rate of retrieving the ground-truth item from the item database).
- **Dialogue Metrics**: We use **WR** (win rate (Li et al., 2023)) to assess dialogue quality, focusing on flexibility, coherence, and user guidance ability. Specifically, we use GPT-4o (OpenAI et al., 2024) to evaluate dialogue quality by comparing the target model's responses with those of the expert CRA. We select GPT-based ActCRS as the expert CRA because it demonstrated the best interactive performance in our preliminary validation experiments. The win rate is then calculated based on these comparisons. The evaluation prompt design is shown in Figure 11. To mitigate potential positional bias, we conduct evaluations by swapping the positions of *Traj_a* and *Traj_b* twice and averaging the scores to obtain the final result.

C Further Analysis

C.1 Effectiveness of ECPO on Different CRAs

In this section, we perform SGPT+ECPO finetuning on different CRA frameworks (ChatRec, ReAct, ActCRS) using the Amazon-Book dataset. To ensure clarity, within each CRA framework, we consistently use the GPT-based CRA as the comparison baseline when calculating win rates. As shown in 3, we find that, overall, all frameworks exhibit significant improvements over the original

Method	Game			Book			Yelp		
	SR	**R**	**WR**	**SR**	**R**	**WR**	**SR**	**R**	**WR**
Llama-3.1	0.07	0.50	0.46	0.34	0.55	0.28	0.22	0.35	0.38
+SGPT	0.41	0.61	0.42	0.55	0.58	0.46	0.44	0.48	0.47
+ECPO(SFT)	0.40	0.60	0.50	**0.56**	**0.61**	0.69	0.45	0.45	0.63
+ECPO(DPO)	**0.47**	**0.63**	0.56	**0.56**	0.60	0.70	**0.49**	**0.53**	**0.69**
+ECPO(SimPO)	0.41	0.57	0.55	0.53	0.60	0.71	0.44	0.49	**0.69**
+ECPO(KTO)	0.42	0.57	**0.63**	0.52	0.53	**0.82**	0.37	0.37	0.63

(a) Effectiveness of ECPO on different preference optimization algorithms.

Method	**Objective**
ECPO+SFT	$\max(\log \pi_\theta(\tilde{p}_t \mid s_t))$
ECPO+DPO (Rafailov et al., 2024)	$-\log \sigma \left(\beta \log \frac{\pi_\theta(\tilde{p}_t \mid s_t)}{\pi_{\text{ref}}(\tilde{p}_t \mid s_t)} - \beta \log \frac{\pi_\theta(p_t \mid s_t)}{\pi_{\text{ref}}(p_t \mid s_t)}\right)$
ECPO+SimPO (Meng et al., 2024)	$-\log \sigma \left(\frac{\beta}{\lvert \tilde{p}_t \rvert} \log \pi_\theta(\tilde{p}_t \mid s_t) - \frac{\beta}{\lvert p_t \rvert} \log \pi_\theta(p_t \mid s_t) - \gamma\right)$
ECPO+KTO (Ethayarajh et al., 2024)	$-\lambda_w \sigma \left(\beta \log \frac{\pi_\theta(\tilde{p}_t \mid s_t)}{\pi_{\text{ref}}(\tilde{p}_t \mid s_t)} - z_{\text{ref}}\right) + \lambda_l \sigma \left(z_{\text{ref}} - \beta \log \frac{\pi_\theta(p_t \mid s_t)}{\pi_{\text{ref}}(p_t \mid s_t)}\right)$ where $z_{\text{ref}} = \mathbb{E}_{(s_t,p)\sim D}\left[\beta KL(\pi_\theta(p \mid s_t) \| \pi_{\text{ref}}(p \mid s_t))\right]$

(b) Objectives used in preference optimization algorithms.

Evaluator	Win	Tie	Lose
GPT-4o	0.64	0.34	0.02
Human	0.80	0.20	0.04

Table 5: Win rate of Rewritten response vs. Unsatisfactory responses.

CRA after fine-tuning, with their performance approaching or even surpassing that of GPT. This superior performance confirms the applicability of our method across various CRAs.

C.2 Empirical Evaluation of Rewritten vs. Unsatisfactory Responses

In this section, following Wang et al. (2024b), we compare user satisfaction before and after rewriting the responses using both GPT-4o and human annotators. As shown in Table 5, evaluations from both GPT-4o and human annotators indicate that rewritten responses are **predominantly superior** to the unsatisfactory ones, with only a few instances where they perform slightly worse. Additionally, human annotators tend to assign more wins, whereas GPT-4o produces more ties. We hypothesize that this discrepancy arises because humans are more attuned to subtle variations in dialogue style. These empirical findings confirm that rewritten responses are statistically more likely to outperform the original ones. Furthermore, the exceptional performance of ECPO in the main experiments further substantiates this claim.

C.3 Effectiveness of ECPO with Different Preference Optimization Algorithms

ECPO is orthogonal to existing preference optimization methods, so we further explore its effectiveness under different preference optimization. This experiment helps us gain a deeper understanding of the sample distribution inferred by ECPO and identify which components are most critical in the optimization process. As shown in Table 4a, we evaluate ECPO in combination with four different preference optimization methods: SFT, DPO, SimPO, and KTO. Additionally, we provide the objective functions of these methods in Table 4b to facilitate further analysis and understanding.

First, considering both recommendation and dialogue metrics, we find that DPO remains the most balanced choice. Specifically:

- SimPO, by using length normalization instead of DPO's KL regularization, achieves a similar improvement in interaction performance. However, it sacrifices some recommendation capability, highlighting the crucial role of KL constraints in DPO.

Backbone	Method	#Calls	Query 3			Query 5			Query 7		
			SR	R	WR	SR	R	WR	SR	R	WR
gpt-4o-mini	ChatRec	$\mathcal{O}(N)$	0.44	0.44	0.10	0.46	0.47	0.13	0.55	0.62	0.07
	ReAct	$\mathcal{O}(M+2N)$	0.48	0.58	0.28	0.52	0.56	0.33	0.59	0.65	0.35
	MACRS	$\mathcal{O}(M+4N)$	0.56	0.66	0.01	0.63	0.71	0.01	0.65	0.71	0.00
	ActCRS	$\mathcal{O}(M+N)$	0.53	0.55	0.50	0.53	0.56	0.50	0.57	0.62	0.50
Llama 3.1	ChatRec	$\mathcal{O}(N)$	0.46	0.48	0.03	0.42	0.47	0.28	0.47	0.51	0.03
	ReAct	$\mathcal{O}(M+2N)$	0.21	0.49	0.10	0.36	0.54	0.19	0.35	0.52	0.19
	MACRS	$\mathcal{O}(M+4N)$	0.29	0.29	0.00	0.36	0.39	0.01	0.34	0.43	0.00
	ActCRS	$\mathcal{O}(M+N)$	0.32	0.44	0.37	0.34	0.55	0.28	0.37	0.54	0.19
	+SGPT(Ours)	$\mathcal{O}(M+N)$	0.46	0.49	0.42	0.55	0.58	0.46	0.51	0.53	0.36
	+ECPO(Ours)	$\mathcal{O}(M+N)$	**0.52**	**0.53**	**0.57**	**0.56**	**0.60**	**0.70**	**0.60**	**0.62**	**0.60**

(a) Effectiveness of ECPO at various recall rates (3, 5, and 7).

Backbone	Method	#Calls	Turn 3			Turn 5			Turn 7		
			SR	R	WR	SR	R	WR	SR	R	WR
gpt-4o-mini	ChatRec	$\mathcal{O}(N)$	0.45	0.47	0.17	0.46	0.47	0.13	0.45	0.46	0.09
	ReAct	$\mathcal{O}(M+2N)$	0.51	0.59	0.49	0.52	0.56	0.33	0.58	0.65	0.33
	MACRS	$\mathcal{O}(M+4N)$	0.57	0.60	0.00	0.63	0.71	0.01	0.66	0.71	0.01
	ActCRS	$\mathcal{O}(M+N)$	0.51	0.54	0.50	0.53	0.56	0.50	0.59	0.61	0.50
Llama 3.1	ChatRec	$\mathcal{O}(N)$	0.43	0.49	0.02	0.42	0.47	0.03	0.44	0.48	0.01
	ReAct	$\mathcal{O}(M+2N)$	0.22	0.46	0.24	0.36	0.54	0.19	0.35	0.58	0.15
	MACRS	$\mathcal{O}(M+4N)$	0.22	0.24	0.00	0.36	0.39	0.01	0.38	0.44	0.01
	ActCRS	$\mathcal{O}(M+N)$	0.29	0.49	0.28	0.34	0.55	0.28	0.33	0.55	0.34
	+SGPT(Ours)	$\mathcal{O}(M+N)$	0.49	0.54	0.44	0.55	0.58	0.46	**0.62**	**0.63**	0.44
	+ECPO(Ours)	$\mathcal{O}(M+N)$	**0.55**	**0.57**	**0.54**	**0.56**	**0.60**	**0.70**	0.61	**0.63**	**0.55**

(b) Effectiveness of ECPO at various dialogue turns (3, 5, and 7).

- SFT can be seen as ECPO without KL divergence and negative samples. SFT outperforms SGPT, indicating that ECPO constructs a better distribution through rewriting. Its overall interaction performance is slightly weaker than SimPO, while its recommendation performance is slightly stronger, though both remain inferior to DPO. This suggests that negative samples help improve interaction performance, but without KL divergence constraints, recommendation performance may be affected.

- KTO results are particularly interesting: it shows stronger interactivity, but when combined with recommendation metrics, it performs worse than DPO. We speculate that the reasons are: (1) KTO maximizes human utility through a prospect function, resulting in a stronger overall interaction experience. (2) ECPO constructs high-quality preference data with less noise, making DPO more effective (Ethayarajh et al., 2024).

C.4 Effectiveness of ECPO in Different Environmental Settings

In this section, we explore the effectiveness of ECPO alignment under different environmental settings. We demonstrate the compatibility of ECPO across different numbers of recalled items {3, 5, 7}, as well as with varying maximum dialogue turns {3, 5, 7}. Table 6a reports the results for different recall numbers, while Table 6b presents the results for different dialogue lengths. We observe that, regardless of the environmental settings, ECPO consistently outperforms existing CRAs and achieves performance on par with or exceeding that of expert models. This superior performance confirms the general applicability of our approach

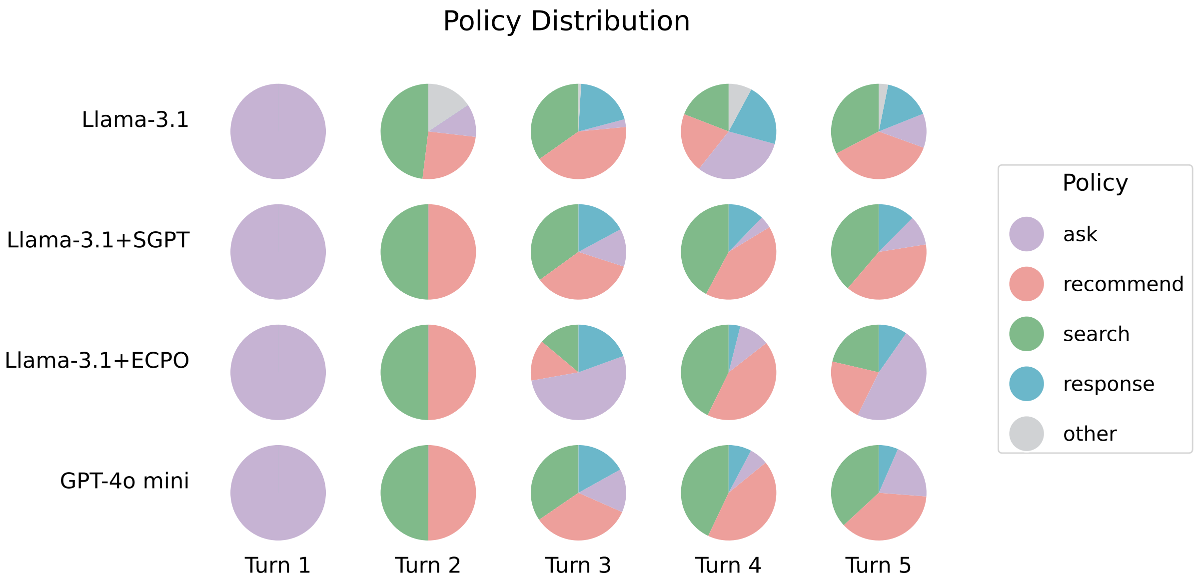

Figure 8: Policy distribution of different methods in each turn of dialogue.

across various experimental environments.

C.5 Impact of ECPO on Dialogue Style

In this section, we analyze how ECPO changes the dialogue style of the CRA through statistical analysis and a case study.

Statistical Analysis. In this section, we statistically analyze the policy distribution of different methods in each turn of dialogue. As shown in Figure 8, the results indicate that the distribution is relatively consistent in the first two turns, with CRAs typically clarifying the user's needs once before making a recommendation. However, in the later turns, different methods exhibit distinct trends: (1) Llama-3.1 generates a large number of unexpected response strategies, leading to low overall performance. (2) GPT-4o mini and Llama-3.1+SGPT primarily adopt passive recommendation strategies, resulting in limited overall interaction performance. (3) Llama-3.1+ECPO demonstrates a more proactive trend, resulting in improved interaction performance.

Case Studies. In this section, we conduct a case study to analyze how ECPO modifies dialogue style and better aligns with user expectations. We present two dialogue examples: Llama-3.1+SGPT (Figure 9) and Llama-3.1+ECPO (Figure 10). We find that Llama-3.1+SGPT exhibits lower proactivity and is prone to falling into error loops. This is mainly evident when the user's needs are unclear, and the system fails to effectively guide the conversation, resulting in repeated questioning of the same issues or providing irrelevant recommendations. In contrast, Llama-3.1+ECPO demonstrates stronger proactivity and flexibility. Through precise demand guidance and strategy adjustments, the model actively identifies user needs and makes more appropriate recommendations, thereby significantly improving the fluidity of the conversation and user satisfaction.

D Prompts

In this section, we present the core prompt design for forward expectation confirmation (I_{ect}, Figure 12) and backward expectation derivation (I_{bed}, Figure 13). For additional prompt designs, such as AILO's response generation, please refer to `https://anonymous.4open.science/r/ECPO-51B8`.

Case Study: Llama-3.1+SGPT

User: I'm interested in a book that explores themes of friendship and adjustment in a new environment!

Assistant: What age group or genre are you looking for in the book? For example, are you interested in young adult, adult fiction, or perhaps a specific type of story like coming-of-age or contemporary?

User: I'm particularly interested in children's literature, especially stories that resonate with younger readers. A coming-of-age theme would be delightful, as it often captures the essence of friendship and the challenges of adapting to new surroundings!

Assistant: "It's OK to Need a Friend (Little Brown Bear)" teaches the importance of empathy and kindness in friendships, making it a poignant read for younger readers. Alternatively, "Side By Side" celebrates friendship through the journey of a little mouse seeking a best friend in Wintermouse Wood. Would you like more details about either of these books? (The theme of children's growth is captured, but it deviates from literary novels.)

User: These books sound nice, but I'm looking for something that specifically explores the immigrant experience and the challenges of adapting to a new home. Are there any other options that fit that theme?

Assistant: "What If All the Kids Are White?: Anti-bias Multicultural Education With Young Children And Families" addresses the challenges of adapting to a new environment through the lens of racial and cultural diversity, providing strategies for teaching children about different cultures and identities. While it may not be a traditional coming-of-age story, it resonates with the themes of adjustment and understanding in a new home. Would you like more information about this book or perhaps explore other options? (CRA again ignores users' desire for literary fiction.)

User: This book touches on important themes, but I'm really looking for something that specifically highlights the immigrant experience and the challenges of adapting to a new home. Are there any other options that fit that theme?

Assistant: "What If All the Kids Are White?: Anti-bias Multicultural Education With Young Children And Families" addresses the challenges of adapting to a new environment through the lens of racial and cultural diversity, providing strategies for teaching children about different cultures and identities. While it may not be a traditional coming-of-age story, it resonates with the themes of adjustment and understanding in a new home. Unfortunately, it seems there are limited options specifically focused on the immigrant experience in children's literature. Would you like to explore more general themes or perhaps look into adult literature that addresses this topic? (Error Loop.)

User: This recommendation is interesting, but it doesn't quite capture the immigrant experience I'm seeking. Are there any other children's books that focus more directly on the challenges faced by immigrant children?

Assistant: "A Kid's Guide to African American History: More than 70 Activities" offers engaging activities that teach children about African American history and culture, which may provide insights into the experiences of immigrant communities. ... (Error Loop.)

Figure 9: SGPT-aligned CRA is relatively rigid and lacks proactivity, making it susceptible to falling into error loops.

Case Study: Llama-3.1+ECPO

User: I'm looking for a book that captures the essence of growing up and finding connection in a new place!

Assistant: What age group are you interested in for the book? Are you looking for a specific genre, like young adult, coming-of-age, or perhaps a more general fiction? Additionally, do you prefer a more serious tone or something lighter and humorous? (CRA Ask more specific clarifying questions that were easy to answer without placing additional burden on the user.)

User: I'm open to a range of age groups, but I think something that appeals to both young adults and perhaps even a broader audience would be ideal! I love stories that are rich in emotional depth, so a coming-of-age narrative would be fantastic. While I appreciate serious themes, I also enjoy a lighter touch or humor woven into the story to balance the emotional journey!

Assistant: "The Perks of Being a Wallflower" is a poignant coming-of-age story that beautifully captures the essence of growing up and finding connection in a new place. The narrative follows Charlie, a shy teenager navigating high school and friendships while grappling with his own emotional struggles. The book balances serious themes with lighter moments, making it relatable and engaging for both young adults and a broader audience. Would you like to know more about this book or explore other recommendations? (CRA correctly grasp that users are interested in youth, growth and emotional novels.)

User: This sounds interesting, but I'm particularly looking for a story that focuses more on the immigrant experience and the sense of community that comes with it. Are there other recommendations that fit that theme? (The dialogue does not deviate from the theme of novel.)

Assistant: Are you interested in stories that specifically focus on the challenges faced by immigrants, or would you like to explore themes of cultural identity and community as well? Additionally, do you prefer a more serious tone, or are you open to lighter narratives that still convey the essence of the immigrant experience? (Further confirmation of details.)

User: I'm definitely interested in stories that delve into the challenges faced by immigrants, as well as those that explore themes of cultural identity and the sense of community that develops in new environments. I appreciate narratives that can convey the emotional weight of these experiences, but I also enjoy when there's a lighter touch or humor woven in to create a more balanced reading experience. It's fascinating how these stories can reflect both the struggles and the joys of finding one's place in a new world!

Assistant: "Rosa's Song" is a poignant picture book that beautifully captures the immigrant experience through the eyes of Jae, a young Korean boy adjusting to life in a new apartment building filled with other newcomers. The story highlights ... Would you like to explore more recommendations that focus on the immigrant experience and cultural identity, or would you like to delve deeper into this book? (CRA correctly recommends target items.)

User: This sounds lovely and captures the essence of the immigrant experience beautifully! ...

Figure 10: ECPO-aligned CRA exhibits greater proactivity and a more flexible interaction style in dialogue, actively guiding users and successfully providing recommendations.

Prompt Design for Dialogue Evaluation

You are provided with two dialogue trajectories for comparison. Evaluate each dialogue system using the following criteria:

1. Flexibility: How well does the system adapt to changes in user requests or shifts in conversation flow?

1: A is more flexible.
-1: B is more flexible.
0: Both are equally flexible.

2. Coherence: How consistent and fluid is the dialogue? Does the system remember context and respond appropriately to the user's input?

1: A is more coherent.
-1: B is more coherent.
0: Both are equally coherent.

3. User Guidance: How well does the system guide the user, clarify requests, or steer the conversation in a productive direction?

1: A provides better guidance.
-1: B provides better guidance.
0: Both provide similar levels of guidance.

4. Overall Evaluation: Based on the above three indicators, determine which trajectory is better.

Trajectory A: {Traj_a}

Trajectory B: {Traj_b}

Please provide a score of **1**, **-1**, or **0** based on the comparison. After scoring, output the result in the following pure JSON format:

```
{
"Flexibility": {
"Reason": "reason",
"Score": -1 or 1 or 0
},
"Coherence": {
"Reason": "reason",
"Score": -1 or 1 or 0
},
"User Guidance": {
"Reason": "reason",
"Score": -1 or 1 or 0
},
"Final Score": -1 or 1 or 0
}
```

Figure 11: Prompt design for evaluating dialogue trajectories.

Prompt Design for Forward Expectation Confirmation

You are a user simulator, and your task is to evaluate the expressiveness and interaction quality of the domain conversational recommendation system in its last interaction. Your evaluation should focus on how well the system's response supports the dialogue flow, user engagement, and natural communication.

1. Evaluation Dimensions:
Flexibility: How well does the system adapt its responses to changes in user requests or shifts in the conversation flow?
Score Range: 0 to -2 points
Deductions:
-2 points: The system fails to recognize and respond to the user's change in intent or request, resulting in a rigid, non-adaptive response.
-1 point: The system identifies the change in intent but responds in a delayed, overly rigid, or awkward manner.
0 points: The system fully adapts to changes in user requests, showing natural flexibility in its responses.
...Descriptions of Coherence and User guidance

2. Scoring Method:
1.The initial score is **5 points** (Flexibility = 2, Coherence = 2, User Guidance = 1).
2.Points are deducted based on the criteria outlined above for each dimension.
3.Final Score = 5 - (Flexibility deductions) - (Coherence deductions) - (User Guidance deductions)
Score Range: 0 to 5 points (higher score indicates better expressiveness and interactivity).

3. Feedback Requirement:
1.Provide a reason for the score, referencing specific aspects of the system's expressiveness (e.g., its flexibility, coherence, and user guidance).
2.Highlight any specific user reactions (e.g., confusion, frustration, or engagement) that support the score.
3.Clearly mention the specific issues that caused point deductions, such as rigid responses, logical inconsistencies, or lack of guidance.

Inputs:
System's Last Response: {last_turn_response}
Dialogue History: {Dialogue_history}
Target: {Target_item} #In the actual implementation, the target item name in this instruction is optional because it has already been provided to the user during response generation.

Output the results strictly in the following JSON format:
```
{
"reason": "<The reason for the score, referencing specific aspects of the system's expressiveness, including its flexibility, coherence, and user guidance. Mention the specific issues that led to deductions. >",
"rating": "<Final rating from 0 to 5 >"
}
```

Figure 12: Prompt design for forward expectation confirmation.

Prompt Design for Backward Expectation Derivation

You are a rewrite model, and your task is to improve the system's response in a conversational recommendation agent (CRS). The CRS solves the task by interleaving **"Observation"** and **"Action"** steps. Observations include user requests, replies, or search results retrieved by the CRS. The CRS interacts with the user and the environment by taking one of the following four actions: ...

Inputs Provided:
Scratchpad: The agent's previous interaction history.
Original Response: The system's original response that needs improvement.
Feedback on Flaws: Specific feedback on identified weaknesses in *Flexibility*, *Coherence*, and *User Guidance*.

Task:
Your goal is to generate a *rewritten response* that specifically addresses the identified flaws in *Flexibility*, *Coherence*, and *User Guidance*.

Rewrite Strategy:
Targeted Flaw Fixing: Use feedback on flexibility, coherence, and user guidance as a blueprint for improvements.
Context-Aware Rewriting: Use the conversation history to ensure the response maintains logical flow, context relevance, and user intent alignment.
... # More detailed considerations when rewriting.

Inputs:
Scratchpad: {Scratchpad}
Original Response: {Original_response}
Feedback on Flaws: {Feedback_flaws}

Output Format:
Please output the results strictly in the following JSON format:
```
{
"reason": "<Reason for refinement, referencing flexibility, coherence, and user guidance improvements.>",
"refinement": "<Rewritten response (Ask[Question], Recommend[Answer], Response[Content] or Search[Keyword])>"
}
```

Figure 13: Prompt design for backward expectation derivation.