Stop Playing the Guessing Game! Evaluating Conversational Recommender Systems via Target-free User Simulation

Anonymous ACL submission

Abstract

Recent developments in Conversational Recommender Systems (CRSs) have focused on simulating real-world interactions between users and CRSs to create more realistic evaluation environments. Despite considerable advancements, reliably assessing the capability of CRSs in eliciting user preferences remains a significant challenge. We observe that user-CRS interactions in existing evaluation protocols resemble a ***guessing game***, as they construct *target-biased* simulators pre-encoded with target item knowledge, thereby allowing the CRS to shortcut the elicitation process. Moreover, we reveal that current evaluation metrics, which predominantly emphasize single-turn recall of target items, suffer from ***target ambiguity*** in multi-turn settings and overlook the intermediate process of preference elicitation. To address these issues, we introduce **PEPPER**, a novel CRS evaluation protocol with *target-free* user simulators that enable users to gradually discover their preferences through enriched interactions, along with detailed measures for comprehensively assessing the preference elicitation capabilities of CRSs. Through extensive experiments, we validate PEPPER as a reliable simulation environment and offer a thorough analysis of how effectively current CRSs perform in preference elicitation and recommendation. https://anonymous.4open.science/r/User_Simulator-3906

1 Introduction

Conversational recommender systems (CRSs) have played an increasingly important role in enhancing personalized experiences by providing tailored recommendations through interactive dialogues (Sun and Zhang, 2018; Jannach et al., 2021; Lin et al., 2023a). Throughout the interaction, these systems are expected to perform two key tasks: (1) *preference elicitation* - exploring and uncovering user preferences by encouraging them to express their likes and dislikes, and (2) *recommendation* - retrieving personalized items based on the preferences inferred from the dialogue. In the field of CRSs, automatically evaluating the system's capability has remained challenging (Friedman et al., 2023; Wu et al., 2024; Zhao et al., 2024; Lin et al., 2023b; Zhu et al., 2024). Conventional offline approaches relying on static, pre-collected dialogues from datasets often neglect the system's responsibility to dynamically shape the dialogue itself, whereas evaluating with real user interactions is costly and time-consuming (Zhang and Balog, 2020; Gao et al., 2021; Yoon et al., 2024).

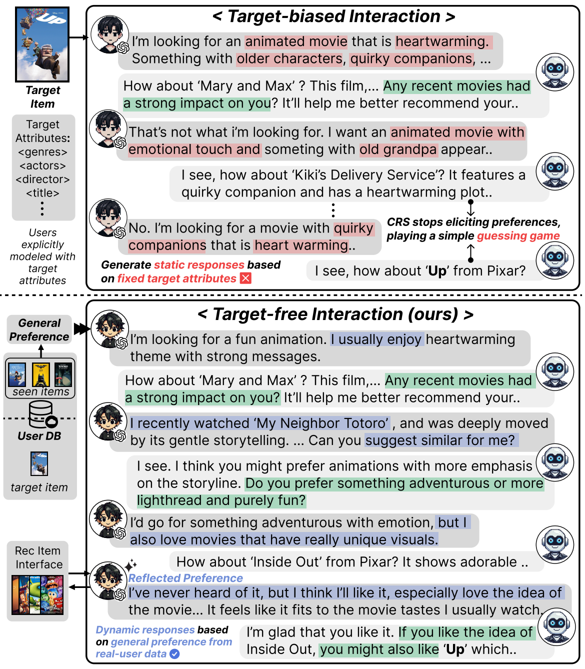

Figure 1: While existing *target-biased* user simulators directly reveal attributes of target items for CRS evaluation (Upper), our *target-free* user simulator engages with more general preference (Lower), making preference elicitation crucial to provide accurate recommendations.

Recently, many studies (Zhang and Balog, 2020; Friedman et al., 2023) have explored leveraging Large Language Models (LLMs) to simulate user conversations with CRSs, creating more realistic evaluation environment that reflect the complexity of human-agent dialogue. However, while effective at assessing recommendation quality, these approaches still face challenges in reliably evaluating the process of preference elicitation. Specifically, we highlight two major limitations in existing user simulation paradigms: **(1) Target-biased user simulation**: Existing methods assume scenarios where users have specific items in mind, thereby constructing user simulators that are explicitly informed by the target item attributes. However, relying on the target items to model the user simulator significantly hinders user-CRS interactions, as it tends to generate static responses that repeatedly expose the same target attributes, causing the CRS to take shortcuts to the target items. **(2) Lack of reliable metrics**: Existing evaluation metrics are typically limited to measuring single-turn recall of target items, without accounting for the intermediate elicitation process. As a result, they fail to fully assess how well the CRS guides the conversation to uncover the user's evolving preferences or how effectively it addresses the user's diverse tastes throughout the interaction.

Motivated by these, this paper begins by investigating two key research questions: **(1) How does reliance on target items affect the quality of user-CRS interactions?** We reveal that target-biased user simulators reduce interactions to a simplistic ***guessing game*** (Yoon et al., 2024), where the CRS succeeds by repeatedly guessing the target items rather than meaningfully eliciting user preferences. This oversimplified interaction inflates CRS performance and leads to substantial performance disparities across target items, ultimately distorting evaluation results. (Figure 1 Upper). **(2) How reliable is Recall@K as a metric for evaluating CRS in multi-turn dialogues?** We observe that Recall@K suffers from ***target ambiguity*** in multi-turn settings, where the system may hit different target items at each turn yet receive the same score—failing to capture meaningful differences in recommendation behavior. This limitation makes it difficult to distinguish whether the CRS is genuinely guiding the conversation to uncover new target items or merely reiterating previous recommendations.

To tackle these challenges, we propose a novel **P**rotocol for **E**valuating **P**ersonal **P**reference **E**licitation and **R**ecommendation of CRS, named **PEPPER**. To address the target-biased interactions of user simulators, PEPPER adopts *target-free* user simulators, modeled on diverse preferences drawn from real user interaction histories and reviews. Built upon real user data, our simulators personalize their initial behavior based on the review-driven user profiles, instead of relying on fixed target item attributes. In particular, we encourage users to actively participate in conversations with the CRS, enabling them to gradually discover their own preferences through interaction (Figure 1 Lower). To achieve this, we simulate users to continuously enrich the responses by incorporating implicit preferences derived from reflecting their general preferences on items emerging within the interaction.

Moreover, we introduce both quantitative and qualitative measures to comprehensively evaluate preference elicitation capabilities of CRSs. For quantitative measure, we propose a new metric, PREFERENCE COVERAGE, to assess how effectively the CRS elicits each user's diverse preferences with high coverage evolving throughout the conversation. For qualitative measure, we propose fine-grained scoring rubrics to evaluate three different aspects of preference elicitation: *proactiveness*, *coherence* and *personalization*.

To summarize, our contributions are as follows:

- We provide detailed analysis of two key limitations in existing CRS evaluation protocols: (1) target-biased user simulation and (2) lack of reliable metrics.
- We propose PEPPER, a novel CRS evaluation protocol with *target-free* user simulators, enabling realistic user-CRS dialogues without falling into simplistic *guessing games*.
- We present detailed measures for comprehensively evaluating the preference elicitation capabilities of CRSs, encompassing both quantitative and qualitative approaches.
- Through extensive experiments, we demonstrate the validity of PEPPER as a simulation environment and conduct a thorough analysis of how effectively existing CRSs perform in preference elicitation and recommendation.

2 Related Work

2.1 Conversational Recommender Systems

Conversational Recommender System (CRS) aims to elicit user preferences and provide personalized recommendations through conversations. In the field of CRSs, one line of research (Wang et al., 2022a,b) has focused on refining architectural designs to improve recommendation accuracy, while another (Kostric et al., 2021; Ziegfeld et al., 2025)

Method	Dataset (Movie Domain)	User Simulation				CRS Evaluation	
		User Profile Input	Target-free	Free-form	Interaction Strategy	Pref. Elicit.	Recommend.
iEvaLM (Wang et al., 2023)	Redial, OpenDialKG	Target Item Title	✗	✗	✗	✗	✓
SimpleUserSim (Zhu et al., 2024)	Redial, OpenDialKG	Target Item Attr.	✗	✗	✗	✗	✓
CSHI (Zhu et al., 2025)	MovieLens	Target Item Attr., Long-term Pref.	✗	✓	Intent Understanding	✗	✓
CONCEPT (Huang et al., 2024)	LLM-Generated	Target Item Attr., Personality	✗	✓	Feeling Generation	✗	✓
PEPPER (Ours)	IMDB	General Preference	✓	✓	Preference Reflection	✓	✓

Table 1: Comparison of existing CRS evaluation protocols with LLM-based user simulators.

has emphasized enhancing the preference elicitation process to support more personalized interactions. Despite significant advancements, previous evaluation protocols have predominantly focused on measuring final recommendation accuracy using pre-collected dialogue datasets (Chen et al., 2019; Wang et al., 2022b,a), often overlooking the interactive process of preference elicitation. Consequently, automatic evaluation of CRSs has emerged as a key challenge in CRS, as it requires to create more realistic testing environments that reflect the complexity of human-agent dialogue.

2.2 CRS Evaluation with User Simulator

Recently, researchers have focused on developing user simulators for evaluating the performance of CRSs (Zhang and Balog, 2020; Yoon et al., 2024). iEvaLM (Wang et al., 2023) addresses the limitations of traditional offline evaluation methods by dynamically extending pre-collected dialogues through free-form interactions. While effective, concerns have been raised about data leakage, where target item titles are disclosed in existing dialogue histories or user prompt, leading to inflated evaluation results. To mitigate this, (Zhu et al., 2024; Huang et al., 2024; Zhu et al., 2025) have tried to model user preferences using only target item attributes (e.g., genres). However, this simplification still falls short of fully addressing the core issue, as providing target attributes can still shortcut the recommendation process by implicitly narrowing the candidate space. A summary of the existing simulation methods is shown in Table 1.

3 Preliminary Analysis

3.1 Focus and Task

Focus: We focus on unveiling the impact of target-biased user simulation and the limitations of current evaluation metrics in assessing CRS performance. Specifically, we analyze how (1) reliance on predefined target items and (2) the use of Recall as an evaluation metric distort the evaluation process.

Task: CRSs aim to identify a user's target items through multi-turn, preference-eliciting dialogues. Formally, given a user-item dataset, $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items, respectively. For each user $u \in \mathcal{U}$, the preference is modeled with a set of target items $i_u \subset \mathcal{I}$. During interaction, the user provides utterances u_t at each turn, either stating their preferences or giving feedback on prior recommendations. The CRS then generates a response r_t along with a predicted item list $P_t \subset \mathcal{I}$. The ultimate goal of the CRS is to recommend items contained in the user's target set i_u.

3.2 Evaluation Setup

Dataset: We use IMDB[1] movie dataset to initialize user simulators and conduct our experiments on CRSs trained with Redial (Li et al., 2018) and OpenDialKG (Moon et al., 2019) datasets. To ensure a reliable evaluation, we have aligned movie entities in IMDB with each CRS dataset by retaining only the items shared between them. Further details on the dataset is described in A.1.

Metric. To reflect how the CRS performs throughout the interaction, we use Recall@(t, K), which measures the proportion of target items successfully retrieved at the t-th turn.

CRS Baselines. We evaluate four representative CRSs, including three supervised models—KBRD, BARCOR, and UniCRS —and one LLM-based method, ChatGPT. The implementation details of these models are provided in Appendix A.2

Target-biased User Simulation. Following (Zhu et al., 2024), we initialize the preferences of the target-biased user simulators by excluding movie titles and relying solely on item attribute information (i.e., genres, directors, stars, and plot summaries). To explore how target-item reliance impacts user-CRS interaction, we further divide the target item set into two parts: a randomly sampled subset, denoted as the *selected* set, and the remaining subset, denoted as the *residual* set. We then implement target-biased user simulators using only the attributes from the *selected* set. We hypothesize that user preferences modeled solely from

[1]https://www.imdb.com/

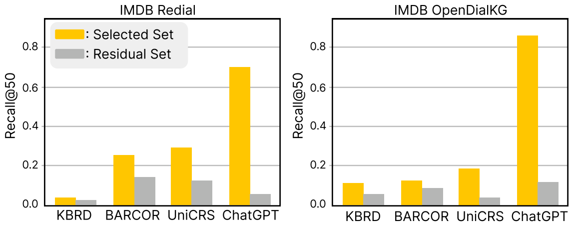

Figure 2: Comparison between selected and residual recall for revealing target-item reliance in user simulators.

the selected target attributes fail to fully capture the diversity of human interests. Otherwise, such attribute-based representations would be sufficient to generalize and allow the CRS to discover the full range of target items, including the *residual* set. To examine this, we compare CRS performance on the *selected* and *residual* sets. Further implementation details are provided in Appendix A.3.

3.3 Results and Analysis

Target-biased user simulation results in a guessing game. As shown in Figure 2, the results reveal a significant performance disparity for target-biased user simulation. For example, on the $\text{IMDB}_{\text{OpenDialKG}}$ dataset, ChatGPT achieves an average score of 0.86 for the *selected* set but only 0.12 for the *residual* set. Similar trends are observed in other CRS models and in the results from the $\text{IMDB}_{\text{ReDial}}$ dataset, further confirming the presence of significant bias. We interpret this bias as a consequence of target disclosure, where target-biased user simulators tend to prioritize certain target items based on their known attributes, resulting in static and narrowly focused preferences that fail to generalize to the *residual* set. Moreover, target-biased simulators tend to provide shortcuts for CRSs by explicitly revealing the target item attributes, reducing the need for meaningful preference elicitation and substantially inflating evaluation results. This calls into question the reliability of existing evaluation protocols and highlights the need for a more realistic user simulation approach.

Recall@K fails to reflect meaningful preference elicitation. Preference elicitation in conversational recommendation involves progressively uncovering users' diverse preferences through interactive dialogue. However, relying solely on Recall exhibits structural limitations that prevent it from properly reflecting this elicitation process. Specifically, Recall@K (1) permits redundancy by allowing repeated counting of identical items across turns (refer to as *target ambiguity*) and (2) measures performance independently at each turn, ignoring previously discovered or missed preferences. For example, as shown in Figure 3, ChatGPT consistently explores new items at each turn, indicated by its high Jaccard distance, whereas KBRD rarely updates its recommendations (low Jaccard distance). Although ChatGPT actively explores new preferences, Recall@K captures only the low hit rate per turn, failing to acknowledge its consistent efforts and treating both models similarly, despite substantial differences in their preference exploration behaviors. Therefore, Recall@K alone fails to capture the process of preference elicitation and points to the need for a metric that reflects diverse preference discovery throughout the dialogue.

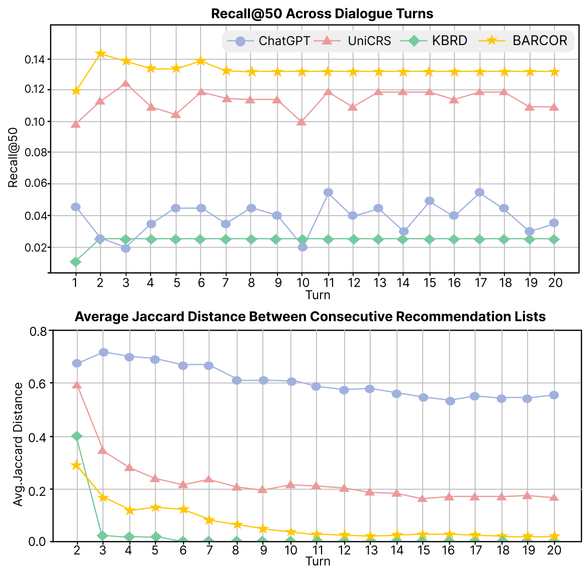

Figure 3: (Upper) Recall@50 of the different CRSs across 20 dialogue turns on the $\text{IMDB}_{\text{ReDial}}$ dataset. (Lower) Average Jaccard distance between consecutive recommendation lists of CRS at each turn.

4 PEPPER: Target-free CRS Evaluation

Guided by the limitations of existing evaluation protocols, we introduce PEPPER, a novel evaluation protocol designed to comprehensively assess both preference elicitation and recommendation abilities of CRSs, addressing key shortcomings of prior approaches. Specifically, it incorporates two key components: (1) ***target-free user simulators*** with richly expressed preferences derived from real user interaction histories and reviews, and (2) ***preference elicitation metrics*** that thoroughly measure a CRS's ability to uncover diverse user preferences and deliver accurate recommendations.

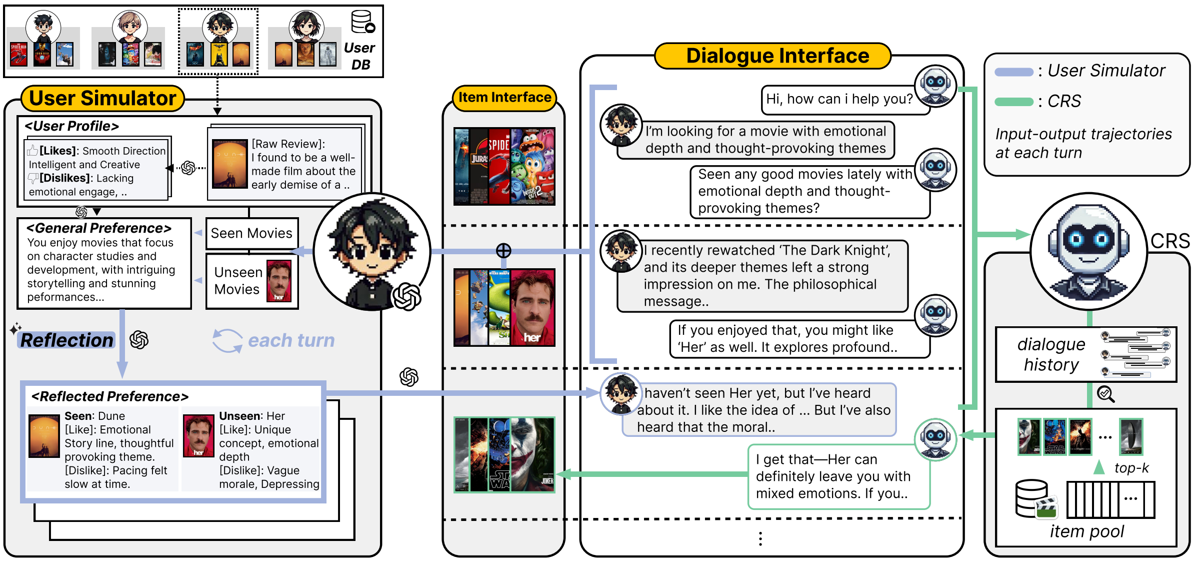

Figure 4: Overview of PEPPER. Within our protocol, a user simulator and a CRS interact via (1) item interface and (2) dialogue interface. The user simulator is initialized with general preferences derived from real-world datasets (*i.e.*, IMDB). [*Blue line*] At each interaction, the user simulator first inspects top-k recommendations in the item interface, classifying the items into seen and unseen sets. It then uses these classifications and the general preferences to generate reflected preferences. Finally, it provides a tailored response enriched with detailed personal preferences. [*Green line*] In response, the CRS generates an utterance and presents new item recommendations.

4.1 Target-free User Simulator

Unlike prior approaches (Wang et al., 2023; Zhu et al., 2024, 2025; Huang et al., 2024), which assume scenarios where users have predefined target items in mind, we design our user simulators with diverse preferences derived from actual user experiences. We aim to construct target-free simulators, instructing them to seek target items without any predefined target information. Instead, these user simulators gradually elaborate on their preferences through ongoing conversations, mirroring how real users naturally articulate and discover their interests. To achieve this, we introduce two core components: ***General Preferences*** and ***Reflected Preferences***. Specifically, general preferences are established as a foundational profile for the user simulator, providing a broad base of interests and inclinations. Reflected preferences, on the other hand, enrich the conversation context by allowing the user simulator to dynamically adapt to the interaction, accordingly refine its preferences, and thoughtfully respond to the CRS. Figure 4 illustrates the overall interaction flow of our framework.

General Preferences. To establish general preferences, we leverage a real-world user database with extensive interaction histories and informative reviews. These reviews provide insights into personal preferences that extend beyond simple item attributes, capturing nuanced opinions on aspects such as storyline, pacing, and emotions. However, given that user-generated reviews often contain noise and ambiguous expressions, following (Kim et al., 2024), we employ ChatGPT to extract and transform each collected reviews into clear, structured binary preferences categorized into *Likes* and *Dislikes*. We then partition each user's interaction history into two distinct subsets: *seen* items and *target* items. The *seen* items refer to those the user has previously interacted with. In contrast, the *target* set, reserved for CRS evaluation, consists exclusively of highly rated items, ensuring a reasonable basis for their use as the evaluation set. When generating general preference, we provide ChatGPT with metadata and corresponding binary preferences derived solely from the *seen* items. The model is then instructed to generate descriptive narratives highlighting the most representative features. These narratives are subsequently used to initialize our simulators, each tailored to mimic a distinct instance from the user database. Through this approach, we ensure that user simulators remain uninformed of target items while being robustly grounded in detailed general preferences. This grounding allows their preferences to be sufficiently generalizable to discover target items, thereby closely emulating real users.

Reflected Preferences. Beyond simply articulating general preferences, real users evaluate items through the lens of their past interactions. They tend to uncover their implicit preferences while interacting with recommendation systems, showing a dynamic and adaptable nature. Reflected preference functions to capture this nuanced user behavior, enabling user simulators to reflect their preferences with regard to current recommendations responsively. To achieve this, we categorize the items recommended by the CRS at each turn into two sets: a *seen* set and an *unseen* set. For *seen* items, we allow the user simulators to revisit their corresponding reviews and recalling what they liked or disliked. For *unseen* items, we prompt the user simulators to shape opinions based on their general preferences, identifying what they are expected to like or dislike. These reflected preferences are then provided as additional input for the user's subsequent response. This approach enables user simulators to proactively provide feedback on both previously interacted items and newly encountered ones, consequently enriching the dialogue.

4.2 Evaluation on Preference Elicitation

Since the preference elicitation ability can be defined as *"how proactively a CRS leads the conversation in a natural and engaging manner, guiding the user through discovering a diverse range of preferences to achieve a satisfactory experience"*, we consider the following key aspects:

(1) ***Preference Coverage***: evaluates how effectively CRS discover the diverse preferences of users through the dialogue. (2) ***Proactiveness*** (Deng et al., 2024): characterizes a CRS that actively guides the conversation by making suggestions or asking relevant questions to actively uncover and clarify the user's preferences. (3) ***Coherence*** Dziri et al. (2019): reflects the CRS's proficiency in maintaining fluid and natural interactions, providing contextually appropriate responses. (4) ***Personalization*** (Lin et al., 2023a): refers to how well the system provides recommendations and information that align with the user's preferences, ensuring a satisfying interaction experience.

Based on these key aspects, we analyze CRSs both quantitatively and qualitatively. For quantitative analysis, we measure PREFERENCE COVERAGE to assess how the CRS identifies each user's target items with high coverage throughout the conversation. For qualitative analysis, we evaluate Proactiveness, Coherence, and Personalization to assess how effectively the CRS integrates the preference elicitation process into the conversation.

Quantitative Metric. To quantitatively measure how well the system understands user's evolving preferences and makes accurate recommendations as the conversation progresses, we propose novel metrics, PREFERENCE COVERAGE (PC) and PREFERENCE COVERAGE INCREASE RATE (PCIR). Specifically, PC is defined as follows:

$$\mathrm{PC}_t = \frac{1}{|U|} \sum_{u \in U} \frac{|(\bigcup_{x=1}^{t} P_x^u) \cap Y(u)|}{|Y(u)|} \tag{1}$$

Here, U denotes the set of users, $Y(u)$ is the set of target items for user $u \in \mathcal{U}$, and P_x^u represents the list of items recommended to user u at turn x. This metric cumulatively measures the capability of a CRS to address diverse user preferences and provide accurate recommendations. Building on this concept, we additionally define PREFERENCE COVERAGE INCREASE RATE at round t as follows:

$$\mathrm{PCIR}_t = \mathrm{PC}_t - \mathrm{PC}_{t-1} \tag{2}$$

PCIR_t indicates the change of PREFERENCE COVERAGE between round $t-1$ and t. The incremental rate of PC reflects how effectively the system discovers new preferences and delivers corresponding recommendations at each turn.

Qualitative Metric. To qualitatively analyze the preference elicitation ability of CRSs, following (Liu et al., 2023), we adopt an automated approach, employing an LLM (i.e., GPT-4o) as the evaluator. Specifically, we task the LLM with fine-grained 1-to-5 scoring rubrics with specified criteria for each rating to evaluate *Proactiveness*, *Coherence*, and *Personalization* based on generated dialogues and each simulator's general preferences.

5 Experiments

We conduct comprehensive experiments to demonstrate the reliability of PEPPER. Detailed settings for user simulation are provided in Appendix A.4, and the implementation details for qualitative evaluation are presented in Appendix A.5.1.

5.1 Reliability of PEPPER

Target-free user simulator of PEPPER closely reflects human preferences. We investigate the extent to which our target-free user simulator can truly represent human preferences. To achieve this, we structure our experiments using rating information, as it provides a clear and quantifiable indication of user preferences for items. For comparison, we provide baseline user simulators initialized with raw reviews and binary preferences (e.g., Likes and Dislikes) to study the effectiveness of *general preference* described in Section 4.1. As shown in Table 2, we observe that our simulator impressively identifies high-rated items that align with its actual user ratings, achieving an accuracy of 69.5%. In contrast, our findings reveal that raw reviews and binary preferences are less effective at representing real user preferences. This highlights the importance of reducing noise and ambiguity in raw reviews and modeling user preference with detailed narratives rather than simplistic binary expressions.

User Preference Representation Types	Accuracy (%)
Raw review	50.6
Binary preference	60.8
PEPPER (General Preference)	69.5

Table 2: Evaluation results of target-free user simulator's capability to reflect human preference.

Target-free user simulator of PEPPER closely emulates human behavior. To further demonstrate the efficacy of our approach, we conduct a human evaluation via Amazon Mechanical Turk (AMT). Specifically, we compare the quality of generated dialogues from target-biased and target-free user simulations, focusing on how effectively the user simulators provide meaningful feedback and how naturally the dialogue flows without resembling a *guessing game*. We compare 100 randomly sampled dialogues from both user simulations. The results, shown in Figure 5, demonstrate that our approach achieves superior performance in capturing diverse user behaviors and maintaining a fluid dialogue progression, highlighting its effectiveness in producing realistic interactions.

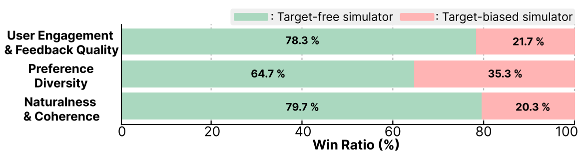

Figure 5: Human evaluation on the quality of generated dialogues from Target-free vs Target-biased simulator.

Target-free user simulator of PEPPER mitigates bias. We provide a comparative analysis to further reveal the extent of bias introduced by target-biased user simulations. Our findings in Section 3.3 shows that target-biased simulations result in significant performance disparities; this limitation becomes even more evident when measured with PC. As shown in Table 3, biased simulators significantly inflate the performance for the *selected* set. In contrast, target-free simulators demonstrate consistent PC, indicating balanced exploration across all target items. This suggests that target-free approach ensures unbiased simulation, providing a reliable framework for evaluating preference elicitation.

Dataset	CRS	Target-biased			Target-free		
		PC_{sel}	PC_{res}	Δ	PC_{sel}	PC_{res}	Δ
IMDB Redial	KBRD	0.040	0.030	-0.010	0.060	0.050	-0.010
	BARCOR	0.285	0.165	-0.120	0.135	0.160	+0.025
	UniCRS	0.410	0.160	-0.250	0.130	0.110	-0.020
	ChatGPT	0.850	0.090	-0.760	0.120	0.125	+0.005
IMDB OpenDialKG	KBRD	0.125	0.095	-0.030	0.125	0.090	-0.035
	BARCOR	0.155	0.115	-0.040	0.155	0.155	+0.000
	UniCRS	0.305	0.105	-0.200	0.125	0.120	-0.005
	ChatGPT	0.950	0.245	-0.705	0.200	0.200	+0.000

Table 3: Recommendation Accuracy of CRSs under target-biased and target-free user simulations. We report $PC_{selected}$@50, $PC_{residual}$@50, and their difference (Δ) from 100 randomly sampled user instances.

Qualitative measure of PEPPER aligns with human judgement. To further validate the reliability of the qualitative metric in PEPPER, we conduct a meta-evaluation to verify its alignment with human judgments. Specifically, we collect human ratings for a total of 100 samples. Each response is evaluated by three human annotators based on the same rubric for *Proactiveness*, *Coherence*, and *Personalization*. We then compute the percentage of agreement and Randolph's Kappa between the human ratings and the automatic scores produced by PEPPER. From the results in Table 4, the agreement rates between PEPPER and human annotators reach 88% for Proactiveness, 92% for Coherence, and 96% for Personalization, with corresponding Cohen's Kappa of 0.81, 0.87, and 0.93, respectively, indicating a strong alignment between the model's judgments and human assessments.

Evaluation Criteria	Agreement	Cohen's Kappa (95%CI)
Proactiveness	88.00	0.81
Coherence	92.00	0.87
Personalization	96.00	0.93

Table 4: Both human evaluators and PEPPER rate the samples on a 1–5 Likert scale. We report the agreement rate and Cohen's Kappa between PEPPER and human.

5.2 CRS Evaluation with PEPPER

Leveraging the PEPPER, we evaluate and analyze the performance of existing CRS baselines with both quantitative and qualitative measures.

Dataset	CRS	Evaluation Metric							
		$PC_{20}@5$	$PC_{20}@10$	$PC_{20}@20$	$PC_{20}@50$	Recall@5	Recall@10	Recall@20	Recall@50
$IMDB_{Redial}$	KBRD	0.0081	0.0127	0.0194	0.0477	0.0066	0.0120	0.0178	0.0353
	BARCOR	0.0155	0.0307	0.0472	0.0911	**0.0072**	**0.0128**	**0.0225**	**0.0525**
	UniCRS	0.0097	0.0186	0.0447	0.0905	0.0035	0.0052	0.0177	0.0375
	ChatGPT	**0.0334**	**0.0495**	**0.0671**	**0.1041**	0.0011	0.0035	0.0053	0.0135
$IMDB_{OpenDialKG}$	KBRD	0.0128	0.0197	0.0422	0.0926	0.0081	0.0096	0.0126	0.0281
	BARCOR	0.0128	0.0300	0.0567	0.1220	0.0040	**0.0133**	**0.0261**	**0.0651**
	UniCRS	0.0163	0.0275	0.0534	0.1174	0.0060	0.0067	0.0106	0.0200
	ChatGPT	**0.0573**	**0.0826**	**0.1236**	**0.2082**	**0.0083**	0.0083	0.0168	0.0480

Table 5: Evaluation of CRSs under our evaluation protocol. We report PREFERENCE COVERAGE and Avg.Recall across 20 conversation turns to evaluate both the preference elicitation and recommendation accuracy of CRSs.

	CRS	Evaluation Metric			
		$PCIR_{avg}$	*Proactiveness*	*Coherence*	*Personalization*
IMDB ReDial	KBRD	0.0019	1.88	2.54	2.04
	BARCOR	0.0019	2.03	2.03	1.46
	UniCRS	0.0030	2.14	2.74	2.25
	ChatGPT	**0.0043**	**4.59**	**4.83**	**4.89**
IMDB OpenDialKG	KBRD	0.0016	1.70	1.93	1.76
	BARCOR	0.0030	2.31	2.82	2.11
	UniCRS	0.0050	2.15	2.62	2.20
	ChatGPT	**0.0081**	**4.51**	**4.95**	**4.35**

Table 6: Comparison on preference elicitation performances of the CRSs. The $PCIR_{avg}$ denotes the average PCIR value per turn across the entire conversation.

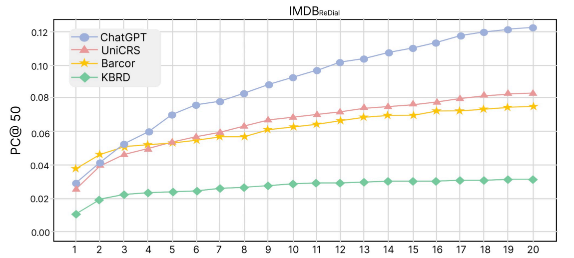

Figure 6: PC values of the CRSs for every turn t in the $IMDB_{ReDial}$ dataset.

Quantitative Evaluation. From Table 5, ChatGPT outperforms other models in terms of PC, benefiting from its advanced conversational abilities that support effective preference elicitation. BARCOR and UniCRS show moderate PC improvements over turns, attributed to the basic conversational understanding of their underlying PLMs (BART and DialoGPT). In terms of the difference between PC and Recall, KBRD and ChatGPT show similar Recall values but present a clear gap in PC. This further strengthen the findings in Section 3.3, indicating that while Recall is effective for measuring per-turn target item accuracy, it fails to assess preference elicitation at the dialogue level.

To gain deeper insights into how preference elicitation unfolds over time, we analyze PC at each turn of the dialogue. As shown in Figure 6, ChatGPT maintains a consistently upward trend in PC over turns, suggesting a sustained effort to explore user preferences incrementally rather than relying solely on previously revealed information. In contrast, BARCOR and UniCRS exhibit slower PC growth, reflecting more reactive conversational strategies. These trends are further supported by the PCIR scores in Table 6, where ChatGPT achieves the highest score, highlighting its proactive exploration of evolving user preferences and its ability to adapt recommendations throughout the dialogue.

Qualitative Evaluation. According to Table 6, ChatGPT significantly outperforms other baseline models in terms of *Proactiveness*, *Coherence*, and *Personalization*. Specifically, ChatGPT demonstrates superior *Proactiveness*, attributed to its advanced language understanding capabilities, whereas other models are constrained by their reliance on fixed datasets. In terms of *Coherence*, ChatGPT generates more fluent and natural responses, closely resembling human dialogues. It also excels in *Personalization*, effectively tailoring contexts to individual user preferences. These results, also supported quantitatively by Figure 6, show that ChatGPT rapidly achieves higher PC scores, demonstrating its ability to effectively capture context shifts throughout the dialogue and seamlessly adapt to user feedback.

6 Conclusion

In this work, we proposes PEPPER, a novel evaluation protocol that comprehensively assesses both preference elicitation and recommendation accuracy in CRSs. PEPPER incorporates target-free user simulators, along with both quantitative and qualitative metrics, targeting four distinct aspects of the preference elicitation process. Through extensive experiments, we demonstrate the effectiveness of PEPPER, offering valuable insights into the limitations of existing CRS evaluation protocols.

Limitations

While our study offers valuable insights into evaluating preference elicitation in CRS, it is not without limitations. One limitation is that our experiments are conducted solely in the movie domain, where user preferences are well-articulated through reviews. This setting allows us to simulate nuanced behaviors in a controlled environment, but generalizing to other domains remains an open challenge. We believe the design of our simulator is domain-agnostic and can be adapted to new settings, though further validation is required.

Another limitation lies in our reliance on proprietary LLMs such as GPT-4o-mini for both simulation and evaluation, which may introduce generation patterns not fully representative of other models. To reduce this concern, we provide additional results using LLaMA-3.1-8B-Instruct, confirming the robustness of our framework across different architectures. Nonetheless, broader comparisons with diverse model backbones are encouraged to further establish generalizability.

A further limitation is that while PEPPER presents new evaluation metrics and perspectives for understanding CRS behaviors, it does not explore methods for improving CRS models themselves. The focus of this work is to analyze how existing systems perform in eliciting user preferences through dialogue. Future work could build on these insights to develop CRS architectures that better support preference elicitation and adapt more effectively to evolving user needs.

Ethical Consideration

Text generated by LLMs may contain content that is harmful, biased, or offensive. However, in our research, we take several steps to minimize these risks. The source dataset, IMDb Movies, is publicly available under the CC0 Public Domain license and includes human-annotated data. Additionally, we manually inspect and filter the dialogues generated through user–CRS interactions to eliminate toxic, offensive, or biased language. For human evaluation, we recruit three independent annotators per unit task via Amazon Mechanical Turk (AMT), ensuring fair compensation. Each annotator is paid $0.15 per task. The textual content presented in this paper contains no personally identifiable information and poses no risk of re-identifying individuals or groups.

References

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. *arXiv preprint arXiv:1908.05391*.

Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 807–818.

Nouha Dziri, Ehsan Kamalloo, Kory W Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. *arXiv preprint arXiv:1904.03371*.

Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, and 1 others. 2023. Leveraging large language models in conversational recommender systems. *arXiv preprint arXiv:2305.07961*.

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.

Chen Huang, Peixin Qin, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Tat-Seng Chua. 2024. Concept–an evaluation protocol on conversation recommender systems with system-and user-centric factors. *arXiv preprint arXiv:2404.03304*.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, 54(5):1–36.

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. *arXiv preprint arXiv:2403.04460*.

Ivica Kostric, Krisztian Balog, and Filip Radlinski. 2021. Soliciting user preferences in conversational recommender systems via usage-related questions. In *Proceedings of the 15th acm conference on recommender systems*, pages 724–729.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. *Advances in neural information processing systems*, 31.

Allen Lin, Ziwei Zhu, Jianling Wang, and James Caverlee. 2023a. Enhancing user personalization in conversational recommenders. In *Proceedings of the ACM Web Conference 2023*, pages 770–778.

Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, and 1 others. 2023b. How can recommender systems benefit from large language models: A survey. *arXiv preprint arXiv:2306.05817*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 845–854.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, and 1 others. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 235–244.

Ting-Chun Wang, Shang-Yu Su, and Yun-Nung Chen. 2022a. Barcor: Towards a unified framework for conversational recommendation systems. *arXiv preprint arXiv:2203.14257*.

Xiaolei Wang, Xinyu Tang, Wayne Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the evaluation for conversational recommendation in the era of large language models. *arXiv preprint arXiv:2305.13112*.

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

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, and 1 others. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5):60.

Se-eun Yoon, Zhankui He, Jessica Maria Echterhoff, and Julian McAuley. 2024. Evaluating large language models as generative user simulators for conversational recommendation. *arXiv preprint arXiv:2403.09738*.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024. On generative agents in recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval*, pages 1807–1817.

Shuo Zhang and Krisztian Balog. 2020. Evaluating conversational recommender systems via user simulation. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining*, pages 1512–1520.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and 1 others. 2024. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*.

Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. How reliable is your simulator? analysis on the limitations of current llm-based user simulators for conversational recommendation. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1726–1732.

Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2025. A llm-based controllable, scalable, human-involved user simulator framework for conversational recommender systems. In *Proceedings of the ACM on Web Conference 2025*, pages 4653–4661.

Liv Ziegfeld, Daan Di Scala, and Anita HM Cremers. 2025. The effect of preference elicitation methods on the user experience in conversational recommender systems. *Computer Speech & Language*, 89:101696.

A Appendix

A.1 Dataset

IMDB is a comprehensive movie database that features extensive user profiles with rich interaction histories and detailed reviews. Redial is a CRS dataset focused on movie recommendations, created using crowd-sourced dialogues through Amazon Mechanical Turk (AMT). OpenDialKG is also a CRS dataset with a broader range of domains, including movies, sports, books and music. However, in this study, we focus on the movie domain due to its accessibility and prominence in CRS research (Jannach et al., 2021). We have manually enriched the OpenDialKG dataset by collecting movie plots from the IMDB website, as it does not provide movie plots in its metadata. To ensure reliable preference modeling, we also excluded users with fewer than 10 interactions. The statistics of the processed IMDB user dataset are summarized in Table 7.

Dataset	#Users	#Interaction Histories
$\text{IMDB}_{\text{ReDial}}$	3,306	66,075
$\text{IMDB}_{\text{OpenDialKG}}$	2,666	47,337

Table 7: Statistics of processed datasets.

A.2 CRS Baselines

We follow (Wang et al., 2023) for the implementation of CRS models, including KBRD (Chen et al., 2019), BARCOR (Wang et al., 2022a), UniCRS (Wang et al., 2022b) and ChatGPT. We integrate a recommender module using the `text-embedding-ada-002` model (Neelakantan et al., 2022) for ChatGPT to constrain the output space of LLM based methods, as they tend to generate items that are beyond the scope of evaluation datasets. Inspired by (Friedman et al., 2023; Zhang et al., 2024), we introduce an item interface, enabling users to view and interact with the current recommendations. This approach more closely mirrors real-world scenarios, where users actively engage with recommendations and provide implicit feedback, facilitating the dynamic refinement of their preferences. Recommendations are retrieved using each CRS's specific retrieval model. Once retrieved, the items are manually augmented with corresponding plots and incorporated into the reflection generation prompts of our user simulators.

- **KBRD** (Chen et al., 2019): enhances the semantic understanding of entities mentioned in conversation history by bridging the recommendation module and transformer-based conversation module through knowledge propagation.
- **BARCOR** (Wang et al., 2022a): presents a unified framework based on BART (Lewis, 2019) that integrates both recommendation and response generation tasks into a single model.
- **UniCRS** (Wang et al., 2022b): proposes a unified framework based on DialoGPT (Zhang et al., 2019) that incorporates a semantic fusion module and knowledge-enhanced prompt learning to improve the association between dialogue history and knowledge graphs.
- **ChatGPT**: is an LLM that demonstrates remarkable text understanding and generation abilities. In this study, we employ gpt-4o-mini (Ouyang et al., 2022) as the conversation module and text-embedding-ada-002 (Neelakantan et al., 2022) as the retrieval module.

A.3 Target-biased User Simulation

we use `gpt-4o-mini` as the backbone language model to simulate the target-biased user simulator. Following prior work (Zhu et al., 2024; Huang et al., 2024; Zhu et al., 2025; Wang et al., 2023), the user simulator is modeled with target item attributes, including genres, directors, stars, and plot summaries, with the item title intentionally excluded. Each dialogue is simulated for up to 20 turns, allowing sufficient interaction for preference elicitation. We evaluate the performance using 100 sampled user instances from each dataset.

A.4 Target-free User Simulation

A.4.1 Interaction Environment

Our interaction environment comprises two generative agents: a target-free user simulator and a CRS. These agents engage through a dialogue interface and an item interface. The dialogue interface bridges communication between the user and the CRS, while the item interface presents top-K recommendations predicted by the CRS at each turn, along with their metadata (*i.e.,* movie plots). By incorporating the item interface, we closely emulate real-world scenarios where users can access detailed information about the recommended items.

For user simulation, we start by extracting the most representative preferences from a user's raw reviews and categorize them into *Likes* and *Dislikes*. These preferences are then transformed into descriptive narratives, depicting the general preferences of the user simulator. Next, the user simulator initiates a entirely new conversation by requesting recommendations that align with its general preferences. In response, the CRS generates an utterance and presents the top-K item suggestions through the item interface. As interactions continue, the user simulator not only communicates with the recommender but also engages with the item interface by carefully examining each suggested item. For previously interacted items (i.e., seen), it retrieves past reviews, while for newly encountered items (i.e., unseen), it shapes opinions based on its general preferences. This dual engagement allows the simulator to elicit its own preferences and provide detailed feedback during subsequent interactions, thereby enriching the dialogue to better align with the user's interests and facilitating the discovery of relevant items.

A.4.2 Implementation Detail

We conduct experiments using 500 user simulators for each dataset. We adopt `gpt-4o-mini` for our target-free user simulations, comprising (1) preference extraction, (2) general preference generation, (3) reflected preference generation, and (4) response generation. We leverage zero-shot prompting to guide the model through each process. To maintain consistent and deterministic outputs, we fix the temperature parameter for user simulation at 0. The number of items presented in the item interface is set to 4, and each simulated dialogue continues up to 20 turns.

A.4.3 Evaluating Simulator's Capability to Represent Human Preference

In Section 5.1, we provide a study to evaluate how closely the proposed target-free user simulator reflects real human preferences. The experiment is conducted as follows: first, a user simulator takes a pair of target items rated by its corresponding user. Then, we instruct the simulator to select the item that aligns more closely with its general preferences. Afterward, we assess the simulator's ability to correctly identify the item with the higher rating based on the actual user scores.

A.4.4 Target-free user simulation with an open-source LLM

We verify the reproducibility of PEPPER through experiments using Llama-3.1-8B-Instruct as the base model for our target-free user simulators. The experiments involve 100 user samples, with each conversation simulated for up to 15 turns. The results, presented in Table 8, reveal that `LLaMA-3.1-8B-Instruct` shows consistent evaluation performance across different CRSs as ChatGPT. These findings validate not only the reproducibility of our framework with open-source models but also its effectiveness for CRS evaluation.

	CRS	PEPPER$^{\text{Llama}}$			
		PC$_{15}$@5	**PC$_{15}$@10**	**PC$_{15}$@20**	**PC$_{15}$@50**
IMDB ReDial	KBRD	0.0020	0.0020	0.0020	0.0091
	BARCOR	0.0165	0.0215	0.0365	0.0737
	UniCRS	0.0115	0.0185	0.0400	0.0835
	ChatGPT	**0.0287**	**0.0318**	**0.0523**	**0.0926**
IMDB OpenDialKG	KBRD	0.0050	0.0091	0.0320	0.0670
	BARCOR	0.0167	0.0207	0.0498	0.0993
	UniCRS	0.0233	0.0350	0.0617	0.1022
	ChatGPT	**0.0287**	**0.0545**	**0.0877**	**0.1829**

Table 8: Recommendation Accuracy of CRS Models with Target-Free User Simulations using LLaMA-3.1-8B-Instruct.

A.5 Qualitative Evaluation

A.5.1 Implementation Details

Following (Liu et al., 2023), We employing an LLM (i.e., GPT-4o) as the evaluator. We task the LLM with fine-grained scoring rubrics on a 1-to-5 scale, with clear criteria for each rating. The inputs to our qualitative evaluation process comprise generated dialogues and the general preferences unique to each user simulator. In assessing *Proactiveness* and *Coherence*, the LLM is instructed to carefully analyze the full dialogue history, examining how proactively the system discovers user needs while maintaining a fluent conversational tone. For *Personalization*, we leverage the LLM to evaluate whether the recommender's responses, including recommendations and explanations, are consistent with the simulator's general preferences.

A.6 Impact of Item Quantity in Item Interface

We explore whether changing the number of items in the item interface influences the quality of user-CRS interactions, as having more items allows the user simulator to better generate its reflected preferences. We conduct experiments using 100 user simulators, with the number of items set to 0, 4, 7, and 10, where 0 is the setting in which preference reflection is excluded. Each dialogue is simulated for 15 turns, and the results are shown in Table 9.

We observe a significant performance gap when the preference reflection process is excluded from the interaction, indicating its critical role in enhancing the quality of user-CRS interactions. However, when preference reflection is included, we observe that increasing the item count has no measurable impact on the interactions. We attribute this to the behavior of our user simulators, which tend to prioritize reflecting preferences for the most relevant recommendations rather than engaging with all available options. In fact, some CRSs, such as UniCRS, exhibit a slight decrease in performance as the item count increases. This indicates that simply adding more items may instead introduce noise into the interaction process.

	CRS	# of items			
		0	**4**	**7**	**10**
$\text{IMDB}_{\text{ReDial}}$	KBRD	0.0199	0.0121	0.0138	0.0129
	BARCOR	0.0715	0.0825	0.0873	0.0842
	UniCRS	0.0860	0.0938	0.0936	0.0772
	ChatGPT	**0.1038**	**0.1130**	**0.1039**	**0.1187**
$\text{IMDB}_{\text{OpenDialKG}}$	KBRD	0.1060	0.0845	0.0737	0.0662
	BARCOR	0.0817	0.0968	0.1043	0.1005
	UniCRS	0.1275	0.1485	0.1410	0.1278
	ChatGPT	**0.1865**	**0.2262**	**0.2243**	**0.2033**

Table 9: CRS performance in user simulations with different numbers of items presented in the item interface. We assess $PC@_{50}$ for evaluation.

B Case Study

Table 10 presents a dialogues generated from $\text{IMDB}_{\text{OpenDialKG}}$, demonstrating a clear alignment between our user simulator's responses and the corresponding real user preferences. For instance, regarding the movie "Ocean's Eleven," the user mentions in their raw review an appreciation for the film's photography, describing it as *"simply wonderful: tinsel colours, beguiling bright lights, tight framing,"* and praises that *"the heist itself is pleasingly clever, it's a charming film."* Correspondingly, the user simulator generates reflected preferences stating: *"[like] Vibrant animation style, entertaining heist plot"*, which aligns with the user's original sentiments. Notably, the user simulator achieves this without any target item information, validating the effectiveness of our approach in representing diverse human preference and evaluating CRSs.

Ground Truth (Target Items): "Ocean's Eleven", "Armageddon", ...

Dialogue Context

...

Recommender: I would recommend the movie "Ocean's Eleven." This film features a group of charismatic and intelligent characters who plan and execute a heist in a stylish and entertaining manner.

User: Ocean's Eleven seems like a promising recommendation with its cool characters and entertaining heist plot. I appreciate the potential for a stylish and engaging storyline. I'm a bit concerned about the possibility of lackluster acting and unsatisfying resolutions, but I'm intrigued to see how it all unfolds.

...

Item Interface

1. [*Ocean's Eleven*] Danny Ocean gathers a group of his World War II compatriots to pull off the ultimate Las Vegas heist ...
2. [*Inside Out*] An agoraphobic must give up his sheltered life and venture outside after a series of personal and financial problems. ...
3. ...

User's General Preferences

You enjoy vibrant animation styles, entertaining heist plots, and cool, intelligent characters with clever dialogue. On the other hand, you tend to dislike movies with contrived endings...

Raw Review

[Ocean's Eleven] : ... but the photography in "Ocean's Eleven" is, at heart, simply wonderful: tinsel colours, beguiling bright lights, tight framing ... Danny and Tess shouldn't have got back together - reaching an amicable understanding would have been better for both them and the audience ... And the heist itself is pleasingly clever. It's a charming film ...

Reflected Preferences

Item: "*Ocean's Eleven*":

Like: Vibrant animation style, entertaining heist plot.

Dislike: Possibility of lackluster acting, unsatisfying

...

Table 10: An example of interactions between our user simulator and CRS (ChatGPT).

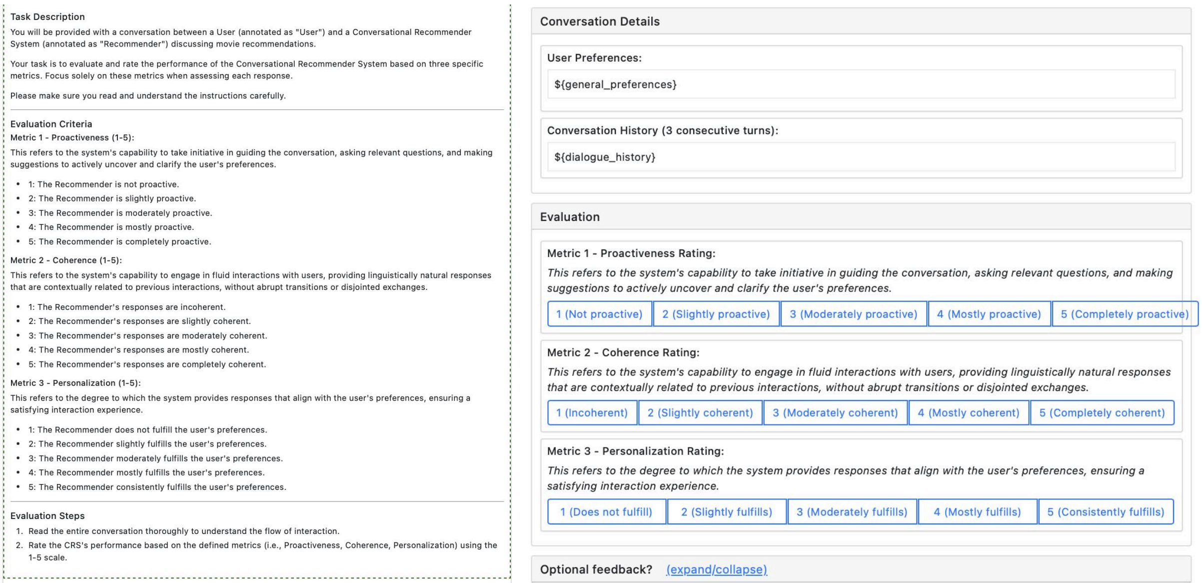

Figure 7: Human evaluation interface.