

Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models

Xiaolei Wang^{1,4}, Xinyu Tang^{1,4}, Wayne Xin Zhao^{1,4*}, Jingyuan Wang³ and Ji-Rong Wen^{1,2,4}

¹Gaoling School of Artificial Intelligence, Renmin University of China

²School of Information, Renmin University of China

³Beihang University

⁴Beijing Key Laboratory of Big Data Management and Analysis Methods

wx11999@foxmail.com, txy20010310@163.com, batmanfly@gmail.com

Abstract

The recent success of large language models (LLMs) has shown great potential to develop more powerful conversational recommender systems (CRSs), which rely on natural language conversations to satisfy user needs. In this paper, we embark on an investigation into the utilization of ChatGPT for conversational recommendation, revealing the inadequacy of the existing evaluation protocol. It might over-emphasize the matching with the ground-truth items or utterances generated by human annotators, while neglecting the interactive nature of being a capable CRS.

To overcome the limitation, we further propose an interactive **E**valuation approach based on LLMs named **iEvaLM** that harnesses LLM-based user simulators. Our evaluation approach can simulate various interaction scenarios between users and systems. Through the experiments on two publicly available CRS datasets, we demonstrate notable improvements compared to the prevailing evaluation protocol. Furthermore, we emphasize the evaluation of explainability, and ChatGPT showcases persuasive explanation generation for its recommendations. Our study contributes to a deeper comprehension of the untapped potential of LLMs for CRSs and provides a more flexible and easy-to-use evaluation framework for future research endeavors. The codes and data are publicly available at <https://github.com/RUCAIBox/iEvaLM-CRS>.

1 Introduction

Recent advancements in conversational intelligence (Chen et al., 2017; Gao et al., 2018) have paved the way for interactive information seeking through *conversational recommender systems* (CRS) (Jannach et al., 2021; Gao et al., 2021). CRSs aim to provide high-quality recommendation services through natural language conversations

that span multiple rounds. Typically, in CRSs, a *recommender module* suggests suitable items based on user preferences, and a *conversation module* further generates appropriate responses considering the conversation context.

Since CRSs rely on strong abilities of understanding and generating natural language text, capable approaches for CRSs have been built on pre-trained language models in existing literature (Yang et al., 2022; Deng et al., 2023). More recently, large language models (LLMs) (Zhao et al., 2023a), such as ChatGPT, have shown capable of solving various natural language tasks via conversations, which shows great potential to develop more powerful CRSs. Since ChatGPT has acquired a wealth of world knowledge during pre-training and is also specially optimized for conversation, it is expected to be an excellent conversational recommender. While, there still lacks a comprehensive study of how LLMs (e.g., ChatGPT) perform on the task of conversational recommendation.

To investigate the capacity of LLMs in CRSs, this work conducts an empirical study on the performance of ChatGPT on existing benchmark datasets for CRSs. We follow the standard evaluation protocol and compare ChatGPT against state-of-the-art methods for CRSs. Surprisingly, the finding is rather counter-intuitive: ChatGPT shows unsatisfactory performance in this empirical evaluation. To understand why this finding occurs, we manually inspect the failure examples. We find that the root cause lies in that the existing evaluation protocol relies on the matching of manually annotated recommendations and conversations, which might over-emphasize the fitting of the ground-truth items based on conversation contexts. Since most CRS datasets are created in a chit-chat way, we find that these conversations are often vague about the user preference, making it difficult to exactly match the ground-truth items even for human annotation. In addition, the current evaluation protocol is based

*Corresponding author

on fixed conversations, which may not be suited for evaluating the capacities of proactive clarification. Similar findings have also been discussed in existing studies (Bang et al., 2023; Qin et al., 2023): traditional metrics such as BLEU and ROUGE may not reflect the real capacities of LLMs on text generation tasks.

Considering this issue, we aim to improve the evaluation approach, to make it more focused on the interactive capacities of CRSs. Ideally, such an evaluation approach should be conducted by human annotators, since the performance of CRSs would be finally tested by real users in practice. However, conducting user studies is both expensive and time-consuming, making it infeasible to carry out large-scale evaluations. As the surrogate, user simulators can be applied for evaluation. However, existing simulation methods are typically limited to pre-defined conversation flows or template-based utterances (Lei et al., 2020; Zhang and Balog, 2020; Afzali et al., 2023). To address these limitations, a more flexible user simulator that supports free-form interactions in CRSs is actually needed.

To this end, this work further proposes an interactive **E**valuation approach based on LLMs, named **iEvaLM**, in which LLM-based user simulation is conducted to examine the performance of conversational recommendation. Our approach draws inspiration from the remarkable instruction-following capabilities exhibited by LLMs, which have already been leveraged for role-play scenarios (Papazov et al., 2022; Wu et al., 2023). With elaborately designed instructions, LLMs can interact with users in a highly cooperative manner. Thus, we design our user simulators based on LLMs, which can flexibly adapt to different CRSs without further tuning. With our evaluation approach, CRSs are no longer constrained to fit the rigid human-written conversation texts. Instead, they are enabled to interact with users in a more free way: the simulating users can support close-to-real evaluation scenarios for CRSs.

With this new evaluation approach, we observe significant improvements in the performance of ChatGPT, as demonstrated through assessments conducted on two publicly available CRS datasets. Notably, the recall@10 metric has increased from 0.174 to 0.670 on the REDIAL dataset with five-round interaction, even surpassing the recall@50 result achieved by the currently leading CRS baseline. Moreover, in our evaluation approach, we

have taken into consideration the crucial aspect of explainability, wherein ChatGPT exhibits proficiency in providing persuasive explanations for the generated recommendations. In contrast, the existing CRSs perform badly in terms of explainability, which is important for CRSs but overlooked in the traditional evaluation.

We summarize our key contributions as follows:

(1) To the best of our knowledge, it is the first time that the capability of ChatGPT for conversational recommendation has been systematically examined on large-scale datasets.

(2) We provide a detailed analysis of the limitations of ChatGPT under the traditional evaluation protocol, discussing the root cause of why it fails on the existing benchmarks.

(3) We propose a new interactive approach that employs LLM-based user simulators for evaluating CRSs. Through experiments conducted on two public datasets, we demonstrate the effectiveness and reliability of our evaluation approach.

2 Background and Experimental Setup

In this section, we describe the task definition and experimental setup used in this work.

2.1 Task Definition

Conversational Recommender Systems (CRSs) are designed to provide item recommendations through multi-turn interaction. The interaction can be divided into two main categories: question answering based on templates (Lei et al., 2020; Tu et al., 2022) and chit-chat based on natural language (Li et al., 2018; Zhang et al., 2023b). In this work, we consider the second category. At each turn, the system either presents a recommendation or initiates a new round of conversation. This process continues until the user either accepts the recommended items or terminates the conversation. In general, CRSs consist of two major tasks: recommendation and conversation. Given its demonstrated prowess in conversation tasks (Pan et al., 2023; Zhao et al., 2023b), we focus our evaluation of ChatGPT on its performance in the recommendation task.

2.2 Experimental Setup

Datasets. We conduct experiments on the REDIAL (Li et al., 2018) and OPENDIALKG (Moon et al., 2019) datasets. REDIAL is the most commonly used dataset in conversational recommendation, which is about movie recommendations.

Dataset	#Dialogues	#Utterances	Domains
ReDial	10,006	182,150	Movie
OpenDialKG	13,802	91,209	Movie, Book, Sports, Music

Table 1: Statistics of the datasets.

OPENDIALKG is a multi-domain conversational recommendation dataset covering not only movies, but also books, sports, and music. Both datasets are widely used for CRS evaluation. The statistics of them are summarized in Table 1.

Baselines. We present a comparative analysis of ChatGPT with a selection of representative supervised and unsupervised methods.

- **KBRD** (Chen et al., 2019): It introduces DBpedia to enrich the semantic understanding of entities mentioned in dialogues.
- **KGSF** (Zhou et al., 2020): It leverages two knowledge graphs (KGs) to enhance the semantic representations of words and entities, and utilizes the Mutual Information Maximization method to align the semantic spaces of these KGs.
- **CRFR** (Zhou et al., 2021a): It performs flexible fragments reasoning on knowledge graphs to address their inherent incompleteness. The information from fragments is fused to improve both recommendation and conversation.
- **BARCOR** (Wang et al., 2022b): It proposes a unified CRS framework based on BART, which tackles two tasks using a single model.
- **MESE** (Yang et al., 2022): It formulates the recommendation task as a two-stage item retrieval process, *i.e.*, candidate selection and ranking, and introduces meta-information when encoding items.
- **UniCRS** (Wang et al., 2022c): It designs knowledge-enhanced prompts based on DialoGPT, effectively tackling two tasks within a unified approach.
- **text-embedding-ada-002** (Neelakantan et al., 2022): It is a powerful model provided in the OpenAI API to transform each input into embeddings, which can be used for recommendation.

Among the above baselines, text-embedding-ada-002 is an unsupervised method, while others are supervised and trained on the CRS datasets.

Evaluation Metrics. Following existing studies (Zhang et al., 2023b; Zhou et al., 2022), we adopt Recall@ k to evaluate the recommendation task. Specifically, for the REDIAL dataset, we

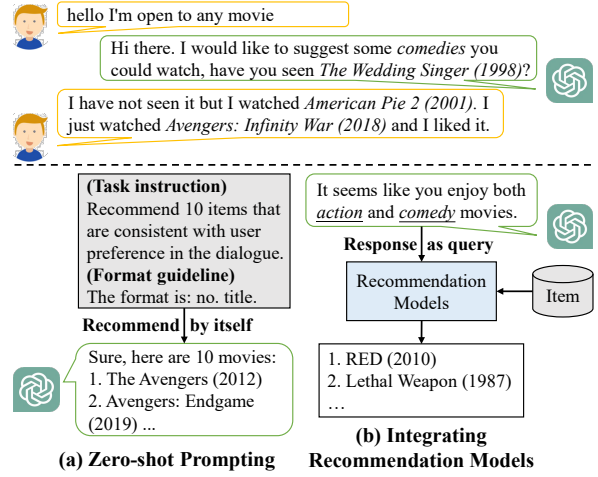


Figure 1: The method of adapting ChatGPT for conversational recommendation.

set $k = 1, 10, 50$ following (Zhang et al., 2023b), while for the OPENDIALKG dataset, we set $k = 1, 10, 25$ following (Zhou et al., 2022). Since requiring too many items at once can sometimes be refused by ChatGPT, we only assess Recall@1 and Recall@10 for it.

Model details. We employ the publicly available model gpt-3.5-turbo provided in the OpenAI API¹, which is the underlying model of ChatGPT. To make the output as deterministic as possible, we set temperature=0 when calling the API. All the prompts we used are detailed in Appendix A.

3 ChatGPT for Conversational Recommendation

In this section, we first discuss how to adapt ChatGPT as a CRS, and then analyze its performance.

3.1 Methodology

Since ChatGPT is specially optimized for dialogue, it possesses significant potential for conversational recommendation. Here we propose two approaches to stimulate this ability as illustrated in Figure 1.

Zero-shot Prompting. We first investigate the ability of ChatGPT through manually designed prompts without any demonstration. The prompt consists of two parts: task instruction and format guideline. The task instruction aims to describe the task of conversational recommendation in natural language, helping ChatGPT understand how to solve it, and the format guideline is designed to

¹<https://platform.openai.com/docs/models>

Datasets	ReDial			OpenDialKG		
Models	Recall@1	Recall@10	Recall@50	Recall@1	Recall@10	Recall@25
KBRD	0.027	0.168	0.369	0.227	0.415	0.540
KGSF	0.039	0.183	0.378	0.119	0.436	0.523
CRFR	0.040	0.202	0.399	0.130	0.458	0.543
BARCOR	0.033	0.173	0.376	0.315	0.459	0.557
UniCRS	0.052	0.218	0.419	0.309	0.509	0.608
MESE	0.056*	0.256*	0.455*	0.279	0.592	0.666*
text-embedding-ada-002	0.025	0.140	0.250	0.279	0.519	0.571
ChatGPT	0.034	0.172	-	0.105	0.264	-
+ text-embedding-ada-002	0.033	0.174	-	0.338*	0.614*	-
+ MESE	0.036	0.195	-	0.240	0.508	-

Table 2: Overall performance of existing CRSs and ChatGPT. Since requiring too any items at once can sometimes be refused by ChatGPT, we only assess Recall@1 and 10 for it. Numbers marked with * indicate that the improvement is statistically significant compared with the best baseline (t-test with p-value < 0.05).

specify the output format, making it easy to understand and assess the recommendation results.

Integrating Recommendation Models. Although ChatGPT can directly generate the items (*i.e.*, the item title or description) in natural language, it may not be well adapted to the downstream recommendation tasks, since it is not specially trained with domain-specific corpus (Zhang et al., 2023a; Hou et al., 2023). In addition, it tends to generate items that are outside the evaluation datasets, which makes it difficult to directly assess the recommendations. To bridge this gap, we incorporate external recommendation models to constrain the output space, and use the conversation history and generated response as inputs. The recommendation models use them to directly predict target items or calculate the similarity with item candidates for matching. We select the CRS model MESE (Yang et al., 2022) as the supervised method and the text-embedding-ada-002 (Nee-lakantan et al., 2022) model provided in the OpenAI API as the unsupervised method. We denote them as **ChatGPT+text-embedding-ada-002** and **ChatGPT+MESE** respectively.

3.2 Evaluation Results

We first compare the accuracy of ChatGPT with representative CRS baselines following existing work (Chen et al., 2019; Zhang et al., 2023b). Then, to examine the inner working principles of ChatGPT, we showcase the explanations generated by ChatGPT for recommendations to assess its explainability as suggested by Jannach et al. (2021).

Dataset	Irrelevant	Partially relevant	Highly relevant
ReDial	8%	20%	72%
OpenDialKG	20%	16%	64%

Table 3: The relevance degree of the explanations generated by ChatGPT to the conversation context.

3.2.1 Accuracy

The performance comparison of different methods for conversational recommendation is shown in Table 2. Surprisingly, ChatGPT does not perform very well as we expect. When using zero-shot prompting, ChatGPT only achieves an average performance among these baselines, and far behind the top-performing methods. When incorporating external recommendation models, the performance can be effectively improved. In particular, on the OPENDIALKG dataset, the performance gap is significantly reduced. It indicates that the responses generated by ChatGPT can help external models understand the user preference. However, there is still a noticeable performance gap on the REDIAL dataset.

3.2.2 Explainability

To better understand how ChatGPT conducts the recommendation, we require it to generate the explanation for its recommendation to examine the inner working principles. Then, we employ two annotators to judge the relevance degree (irrelevant, partially relevant, or highly relevant) of the explanation to the conversation context on randomly sampled 100 failure examples. The Cohen’s Kappa between annotators is 0.77, indicating a good agree-

[Conversation History]
 User: Hello, I am looking for movies for a night with friends that I have coming up. Any suggestions?
[Label] Black Panther (2018)
[Prediction of ChatGPT] The Hangover (2009), Bridesmaids (2011), Superbad (2007), Mean Girls (2004), The Princess Bride (1987), Clueless (1995), The Breakfast Club (1985), Ferris Bueller's Day Off (1986), Pitch Perfect (2012), 21 Jump Street (2012)

(a) Lack of explicit user preference

[Conversation History]
 Bot: What genre of movie do you like?
 User: I like *comedies*
[Response in the Dataset] Have you seen *Talladega Nights: The Ballad of Ricky Bobby* (2006) or *The Hangover* (2009)?
[Response by ChatGPT] That's great! Comedies are a fun and lighthearted genre that can make you laugh and feel good. Do you have a favorite comedy movie or actor?

(b) Lack of proactive clarification

Figure 2: Two failure examples of ChatGPT for conversation recommendation.

ment. The results in Table 3 indicate that ChatGPT can give highly relevant explanations in most of the cases. Here is one example:

[Conversation History]
 User: Hi I want a movie like *Super Troopers* (2001)
[Label] Police Academy (1984)
[Prediction of ChatGPT] Beerfest (2006), The Other Guys (2010), Hot Fuzz (2007), ...
[Explanation of ChatGPT] ... I have recommended movies that share *similar themes* of comedy, law enforcement, and absurdity. ... Some of the movies on the list are also from the *same creators* or feature some of the *same actors* as Super Troopers. ...

We can see that ChatGPT understands the user preference and the recommendations are reasonable, suggesting that it has the ability for conversational recommendation. However, this contradicts its poor performance in terms of accuracy. It motivates us to investigate the reasons for failure.

3.2.3 Why does ChatGPT Fail?

In this part, we analyze *why* does ChatGPT fail in terms of accuracy. Specifically, we manually inspect the incorrect (evaluated according to the annotated labels) recommendations made by ChatGPT. In this way, we identify two main causes and detail them as follows.

Lack of Explicit User Preference. The examples in this class typically have very short conversation turns, in which CRSs may be unable to collect sufficient evidence to accurately infer the user intention. Furthermore, the conversations are mainly collected in a chit-chat form, making it vague to reflect the real user preference. To see this, we present an example in Figure 2(a). As we can

see, the user does not provide any explicit information about the expected recommendations, which is a common phenomenon that occurs at the beginning of a conversation as observed by existing work (Wang et al., 2022a). To verify this, we randomly sample 100 failure examples with less than three turns and invite two annotators to determine whether the user preference is ambiguous. Among them, 51% examples are annotated as ambiguous and the rest 49% are considered clear, which confirms our speculation. The Cohen's Kappa between annotators is 0.75. In this class of examples, ChatGPT cannot produce the exact recommendations as ground-truth items, due to the lack of explicit user preference.

Lack of Proactive Clarification. When evaluating the performance of CRSs on a human-annotated dataset, a major limitation is that the tested method has to strictly follow the existing conversation flows. However, in real-world scenarios, a CRS would proactively ask clarification questions when needed, which is not supported by existing evaluation protocols. In Figure 2(b), the response in the dataset directly gives recommendations, while ChatGPT asks for detailed user preference. As we can see, since many items fit the current requirement, it is reasonable to seek clarification before recommendation. While, such cases cannot be well handled in the existing evaluation protocol, since no user responses are available again in this process. To verify this, we randomly sample 100 failure examples for two annotators to classify the type of responses generated by ChatGPT (clarification, recommendation, or chit-chat). We find that 36% of them are clarification, 11% are chit-chat, and

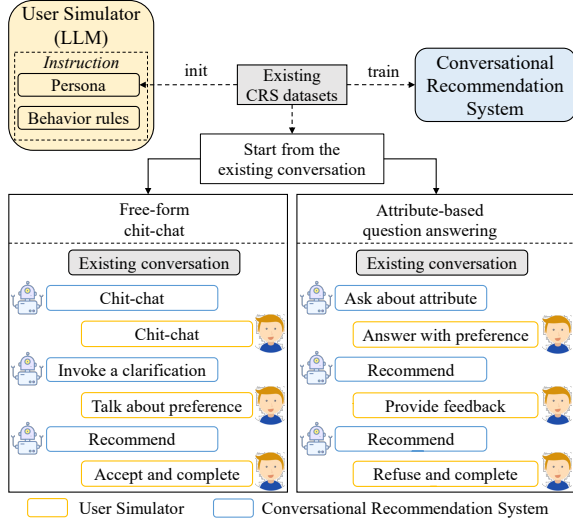


Figure 3: Our evaluation approach **iEvaLM** is based on the existing dataset. It has two settings: free-form chit-chat (left) and attribute-based question answering (right).

only 53% are recommendation, suggesting the importance of allowing clarification when evaluating. The Cohen’s Kappa between annotators is 0.81.

To summarize, there are two potential issues with the existing evaluation protocol in human-annotated CRS datasets (Li et al., 2018; Chen et al., 2019; Zhou et al., 2021b): the conversation contexts tend to be vague about user preferences, and it cannot effectively support proactive clarification as in real-world scenarios. Although conversation-level evaluation (Lei et al., 2020; Zhang et al., 2018) allows the interaction between systems and users, it is limited to pre-defined conversation flows or template-based utterances, which fails to capture the intricacies and nuances of real-world conversations.

4 A New Evaluation Approach for CRSs

Considering the issues with the existing evaluation protocol, in this section, we propose an alternative evaluation approach **iEvaLM** and demonstrate its effectiveness through experiments.

4.1 Overview

Following Jannach et al. (2021), a well-designed CRS evaluation protocol should satisfy the following criteria: (1) *free-form interaction*. The conversations between users and systems should be natural without being restricted to pre-defined conversation flows or utterance templates. This allows users to have more flexibility, enabling the evalua-

tion in line with real scenarios; (2) *mixed-initiative*. Both the user and system can lead the conversation at any point. The system can not only respond to the user utterances but also request the user preference. This allows for a more dynamic and flexible evaluation; (3) *both subjective and objective metrics*. In addition to the accuracy of predicted items, subjective metrics like the persuasiveness of generated explanations should also be considered. This allows for the evaluation of user experience, which is also an important perspective for CRSs.

However, the current evaluation protocol restricts the interaction form and only considers objective metrics like recall, which cannot fully reflect the system capabilities and user experience.

To overcome these limitations, we propose an **interactive Evaluation** approach based on LLMs named **iEvaLM** that harnesses LLM-based user simulators, as illustrated in Figure 3. Our approach is seamlessly integrated with the existing CRS datasets. Each interaction extends over the observed human-annotated conversations in the example. The key idea of our approach is to conduct close-to-real user simulation based on the excellent *role-play* capacities of LLMs (Papazov et al., 2022; Wu et al., 2023). We take the ground-truth items from the example as the user preference through the interaction, and use them to set up the persona of the simulated users by LLMs through instructions. After the interaction, we assess not only the accuracy by comparing with the ground-truth items, but also the explainability by querying an LLM-based scorer.

4.2 Interaction Forms

To make a comprehensive evaluation, we consider two types of interaction: *attribute-based question answering* and *free-form chit-chat*.

In the first type, the action of the system is restricted to choosing one of the k pre-defined attributes to ask the user or making recommendations. At each round, we first let the system decide on these $k + 1$ options, and then the user gives the corresponding template-based response: answering with the attributes of the target item to questions or giving feedback to recommendations. An example interaction round would be like: “System: Which genre do you like? User: Sci-fi and action.”

In contrast, the second type does not impose any restrictions on the interactions of either the system or the user, they are free to take the initiative dur-

ing the interaction. An example interaction round would be like: “System: Do you have any specific genre in mind? User: I’m looking for something action-packed with a lot of special effects.”

These two forms enable a holistic evaluation of the system in both structured and unstructured conversational scenarios, providing valuable insights into its strengths and weaknesses across different types of interactions.

4.3 User Simulation

In our setting, to support the interaction with the system, we employ LLMs to simulate the users. The simulated user can take one of the following three behaviors:

- *Talking about preference.* When the system makes clarification or elicitation about user preference, the simulated user would respond with information about the target item.
- *Providing feedback for the recommendation.* When the system recommends an item list, the simulated user would check each item, and provide positive feedback if finding the target and negative feedback if not.
- *Completing the conversation.* If one certain target item has been recommended by the system or the interaction has reached a certain number of rounds, the conversation would be finished by the simulated user.

Specifically, we use the ground-truth items from existing datasets to construct realistic personas for simulated users. This is achieved by leveraging the text-davinci-003 (Ouyang et al., 2022) model provided in the OpenAI API along with manual instructions (see Appendix A). In these instructions, we first fill the ground-truth items into the persona template and then define their behaviors using a set of manually crafted rules. An example instruction would be like:

You are a seeker chatting with a recommender for movie recommendations. Your target item: “The Godfather (1972)”.

If the recommender recommends the target item, you should accept it.

If the recommender recommends other items, you should refuse and provide the information about the target item.

If the recommender asks for your preference, you should provide the information about the target item.

At each turn, we append the conversation to the instruction as the input. When calling the API,

Setting	Single-turn		Multi-turn	
	Naturalness	Usefulness	Naturalness	Usefulness
DialoGPT	13%	23%	11%	31%
iEvaLM	36%	43%	55%	38%
Tie	51%	34%	34%	31%
Human	10%	34%	17%	28%
iEvaLM	39%	33%	35%	40%
Tie	51%	33%	48%	32%

Table 4: Performance comparison in terms of naturalness and usefulness in the single-turn and multi-turn settings. Each value represents the percentage of pairwise comparisons won by the specific model or ties.

we set max_tokens to 128, temperature to 0, and leave other parameters at their default values. The maximum number of interaction rounds is set to 5.

4.4 Performance Measurement

We consider both subjective and objective metrics to measure the recommendation performance as well as user experience. For objective metrics, we follow the existing practice to use *recall* as stated in Section 2.2, to evaluate every recommendation action in the interaction process. For subjective metrics, following Chen et al. (2022), we use *persuasiveness* to assess the quality of explanations, which aims to evaluate whether the user can be persuaded to accept recommendations. The value range of this metric is 0 to 2. To reduce the need for humans, we propose an LLM-based scorer, which can automatically give the score according to the given prompts.

Specifically, text-davinci-003 (Ouyang et al., 2022) provided in the OpenAI API is utilized as the scorer with the conversation, explanation, and scoring rules concatenated as prompts (see Appendix A). Other parameters remain the same as the simulated user.

5 Evaluation Results

In this section, we assess the quality of the user simulator and the performance of CRSs using our proposed evaluation approach.

5.1 The Quality of User Simulator

To evaluate the performance of CRSs in an interactive setting, we construct user simulators based on the ground-truth items from the existing datasets. The simulated users should cooperate with the system to find the target item, e.g., answer clarification questions and provide proper feedback to the recommendations. However, it is not easy to directly

Model		KBRD			BARCOR			UniCRS			ChatGPT		
Evaluation Approach		Original	iEvaLM (attr)	iEvaLM (free)	Original	iEvaLM (attr)	iEvaLM (free)	Original	iEvaLM (attr)	iEvaLM (free)	Original	iEvaLM (attr)	iEvaLM (free)
ReDial	R@1	0.028	0.039 (+39.3%)	0.035 (+25.0%)	0.031	0.034 (+9.7%)	0.034 (+9.7%)	0.050	0.050 (+0.0%)	0.107 (+114.0%)	0.033	0.408* (+1136.3%)	0.137 (+315.2%)
	R@10	0.169	0.196 (+16.0%)	0.187 (+10.1%)	0.170	0.196 (+15.3%)	0.198 (+17.1%)	0.215	0.231 (+7.4%)	0.316 (+47.0%)	0.174	0.670* (+285.1%)	0.488 (+180.5%)
	R@50	0.366	0.435 (+18.9%)	0.453 (+23.8%)	0.372	0.416 (+11.8%)	0.463 (+24.5%)	0.413	0.500 (+21.1%)	0.596* (+44.3%)	-	-	-
OpenDialKG	R@1	0.231	0.124 (-46.3%)	0.234 (+1.3%)	0.312	0.238 (-23.7%)	0.314 (+0.6%)	0.308	0.172 (-44.2%)	0.315 (+2.2%)	0.338	0.384 (+13.6%)	0.448* (+32.5%)
	R@10	0.423	0.277 (-34.5%)	0.432 (+2.1%)	0.453	0.377 (-16.8%)	0.459 (+1.3%)	0.513	0.366 (-28.7%)	0.538 (+4.9%)	0.614	0.705* (+14.8%)	0.681 (+10.9%)
	R@25	0.492	0.356 (-27.6%)	0.510 (+3.7%)	0.510	0.437 (-14.3%)	0.526 (+3.1%)	0.574	0.435 (-24.2%)	0.609 (+6.1%)	-	-	-

Table 5: Performance of CRSs and ChatGPT using different evaluation approaches, where “attr” denotes attribute-based question answering and “free” denotes free-form chit-chat. “R@ k ” refers to recall@ k . Since requiring too any items at once can sometimes be refused by ChatGPT, we only assess recall@1 and 10 for it. Numbers marked with * indicate that the improvement is statistically significant compared with the rest methods (t-test with p-value < 0.05).

Model	Evaluation Approach	ReDial	OpenDialKG
KBRD	Original	0.920	1.019
	iEvaLM	0.829 (-9.9%)	0.896 (-12.1%)
BARCOR	Original	0.953	1.260
	iEvaLM	0.836 (-12.3%)	1.221 (-3.1%)
UniCRS	Original	0.991	1.318
	iEvaLM	1.036 (+4.5%)	1.311 (-0.5%)
ChatGPT	Original	1.097	1.191
	iEvaLM	1.316* (+20.0%)	1.407* (+18.1%)

Table 6: Persuasiveness of explanations generated by CRSs and ChatGPT. We only consider the setting of free-form chit-chat in our proposed evaluation approach (New). Numbers marked with * indicate that the improvement is statistically significant compared with the rest methods (t-test with p-value < 0.05).

evaluate the quality of user simulators.

Our solution is to make use of the manually annotated conversations in the datasets. We first use the ground-truth items to set up the persona of user simulators, and let them interact with recommenders played by humans to complete the conversations based on the starting round. Then, we can compare the completed conversations with the original ones of the datasets for evaluation. Following Sekulić et al. (2022), we assess the quality based on the *naturalness* and *usefulness* of the generated utterances in the settings of single-turn and multi-turn free-form chit-chat. Naturalness refers to that the utterances are fluent and likely to be generated by humans, and usefulness means that the utterances

are consistent with the user preference. We compare our user simulator with a fine-tuned version of DialoGPT and the original conversations on the REDIAL dataset.

Specifically, we first invite five annotators to play the role of the recommender and engage in interactions with each user simulator. The interactions are based on the first round of conversations from randomly sampled 100 examples, and the persona of each user is set up with the ground-truth items. Then, we employ another two annotators to make pairwise evaluations, where one is generated by our simulator and the other comes from DialoGPT or the original dataset. We count a win for a method when both annotators agree that its answer is more natural (or useful); otherwise, we count a tie. The Cohen’s Kappa between annotators is 0.73. Table 4 demonstrates the results. We can see that our simulator significantly outperforms DialoGPT, especially in terms of naturalness in the multi-turn setting, which demonstrates the strong language generation capability of LLMs. Furthermore, the usefulness of our simulator is better than others, which indicates that it can provide helpful information according to the preference and cooperate with the system for finding target items.

5.2 The Performance of CRS

In this part, we evaluate and compare the performance of existing CRSs and ChatGPT using different approaches. For ChatGPT, we use **ChatGPT+text-embedding-ada-002** due to its superior performance in the traditional evaluation.

5.2.1 Main Results

The evaluation results are presented in Table 5 and Table 6. Overall, most models demonstrate improved accuracy compared to the traditional approach, indicating their ability to interact with users for better recommendations, which is an important aspect overlooked in the traditional evaluation. Among the existing CRSs, the order of performance improvement is *UniCRS* > *BARCOR* > *KBRD*. Both UniCRS and BARCOR utilize pre-trained models to enhance conversation abilities. Additionally, UniCRS incorporates external KGs into prompts to enrich entity semantics and better understand user preferences. For ChatGPT, there is a significant performance improvement, with the Recall@10 value even surpassing the Recall@50 value of some CRSs. This indicates that ChatGPT exhibits superior interaction abilities compared to existing CRSs and can provide high-quality recommendations with sufficient information. The results demonstrate the effectiveness of our approach in evaluating the accuracy of recommendations for CRSs developed with LLMs.

Comparing the two interaction settings, the performance changes are related to specific models and datasets. For existing CRSs, they perform better in the setting of free-form chit-chat than attribute-based question answering, and even worse than the traditional setting on the OPENDIALKG dataset. One possible reason is that they are trained on the datasets with natural language conversations, which is consistent with the free-form chit-chat setting. In contrast, ChatGPT performs better in the setting of attribute-based question answering on the REDIAL dataset, while better in the other setting on the OPENDIALKG dataset. The results suggest that the optimal interaction strategies are not the same across different scenarios, and emphasize the importance of adaptively adopting suitable strategies based on the conversation progress.

Regarding the quality of explanations, we evaluate this aspect using the traditional approach and the chit-chat setting of our new approach since question answering does not involve this action. As shown in Table 6, existing CRSs generally perform worse in our interactive setting compared to the traditional setting, indicating their limited ability to generate persuasive responses during interactions. Since user experience in interaction is an important perspective in CRSs, this demonstrates the limitations of traditional evaluation. In contrast, although

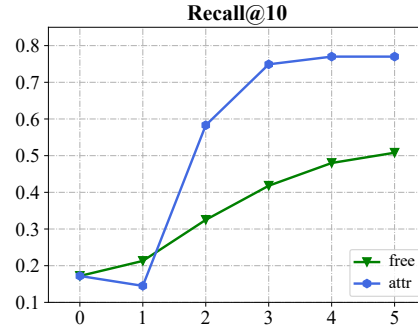


Figure 4: The performance of ChatGPT with different interaction rounds under the setting of attribute-based question answering (attr) and free-form chit-chat (free) on the REDIAL dataset.

ChatGPT is not outstanding in the traditional setting, it performs significantly better than baselines in our interactive setting. This improvement suggests that ChatGPT can effectively persuade users to accept its recommendations, which is a highly desirable trait for CRSs. The results demonstrate the effectiveness of our approach to assessing the explainability of CRSs developed with LLMs.

5.2.2 Influence of Interaction Rounds

Interacting with the user for multiple rounds typically leads to more information and improved recommendation accuracy. However, users have limited patience and may leave the interaction when they become exhausted. It is necessary to investigate the relationship between the number of interaction rounds and performance. Following the setting in our approach, the interaction between ChatGPT and users is start from the observed human-annotated conversation in each dataset example, and we set the maximum interaction rounds to values from 1 to 5, in order to evaluate the changes of recommendation accuracy.

Figure 4 shows the results of Recall@10 on the REDIAL dataset. In attribute-based question answering, the performance keeps increasing except round 1 and reaches saturation at round 3. This observation aligns with our conversation setting, as the REDIAL dataset only has three attributes to inquire about. In free-form chit-chat, the performance curve is steep between rounds 1 and 3, while it is relatively flat between rounds 3 and 5. This pattern may be attributed to insufficient information in the initial round and marginal information in the last rounds. Since the user will gradually get exhausted with

the progress of interaction, how to optimize the conversation strategy remains further research.

5.2.3 The Reliability of Evaluation

Recall that LLMs are utilized in the user simulation and performance measurement part of our evaluation approach as alternatives for humans in Section 4. Considering that the generation of LLMs can be unstable, in this part, we conduct experiments to assess the reliability of the evaluation results compared with using human annotators.

First, recall that we introduce the subjective metric *persuasiveness* for evaluating explanations in Section 4.4. This metric requires human evaluation and we propose an LLM-based scorer as the alternative. Here we evaluate the reliability of the score generated by our LLM-based scorer by comparing it with human annotators. We randomly sample 100 examples with the explanations generated by ChatGPT and ask our scorer and two annotators to rate them separately. The Cohen’s Kappa between annotators is 0.83. It can be observed from Table 7 that the two score distributions are similar, indicating the reliability of our LLM-based scorer as a substitute for human evaluators.

Then, since we propose an LLM-based user simulator as a replacement for humans to interact with CRSs, we examine the correlation between the values of metrics when using real vs. simulated users. Following Section 4.3, both real users and our user simulator receive the same instruction to establish their personas based on ground-truth items. Each user can interact with different CRSs for at most 5 rounds to evaluate their performance in terms of recall and persuasiveness. We randomly select 100 instances and employ five annotators and our user simulator to engage in free-form chit-chat with different CRSs. The results are shown in Table 8. We can see that the ranking of models is consistent with that obtained from real users, and the absolute scores are also comparable. It suggests that our LLM-based user simulator is capable of providing convincing evaluation results and serves as a reliable alternative to human evaluators.

6 Conclusion

In this paper, we systematically examine the capability of ChatGPT for conversational recommendation on existing benchmark datasets and propose a new interactive evaluation approach **iEvaLM**. First, we show that the performance of the vanilla ChatGPT was unsatisfactory, which is contrary to our

Method	Unpersuasive	Partially persuasive	Highly persuasive
iEvaLM	1%	5%	94%
Human	4%	7%	89%

Table 7: The score distribution of persuasiveness (“unpersuasive” for 0, “partially persuasive” for 1, and “highly persuasive” for 2) using our LLM-based scorer and humans on a random selection of 100 examples from the REDIAL dataset.

Evaluation Approach		KBRD	BARCOR	UniCRS	ChatGPT
iEvaLM	Recall@10	0.180	0.210	0.330	0.480
	Persuasiveness	0.810	0.860	1.050	1.330
Human	Recall@10	0.210	0.250	0.370	0.560
	Persuasiveness	0.870	0.930	1.120	1.370

Table 8: The evaluation results using our simulated and real users on a random selection of 100 examples from the REDIAL dataset.

initial expectations. Through analysis of failure cases, the root cause is the limitations of the existing evaluation protocol, which overly emphasizes the fitting of ground-truth annotations based on the conversation context. To address these issues, we propose an interactive evaluation approach that leverages LLM-based user simulators.

Through experiments using this new approach, we have the following findings: (1) ChatGPT is powerful and much better than the currently leading CRSs in both accuracy and explainability; (2) The existing CRSs get improved in accuracy from the interaction, but perform worse in explainability; and (3) The optimal interaction strategies vary across different scenarios. We also demonstrate the effectiveness and reliability of our evaluation approach.

Overall, our work contributes to the understanding and evaluation of LLMs such as ChatGPT for conversational recommendation, paving the way for further research in this field in the era of LLMs.

Limitations

A major limitation of this work is the design of prompts for ChatGPT and LLM-based user simulators. We manually write several prompt candidates and select the one with the best performance on some representative examples to reduce the cost of calling model APIs. More effective prompting strategies like chain-of-thought can be explored for better performance, and the robustness of the evaluation framework to different prompts remains to

be assessed.

In addition, our evaluation framework primarily focuses on the accuracy and explainability of recommendations, but it may not fully capture potential issues related to fairness, bias, or privacy concerns. Future work should explore ways to incorporate these aspects into the evaluation process to ensure the responsible deployment of CRSs.

References

- Jafar Afzali, Aleksander Mark Drzewiecki, Krisztian Balog, and Shuo Zhang. 2023. Usersimcrs: A user simulation toolkit for evaluating conversational recommender systems. *arXiv preprint arXiv:2301.05544*.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
- Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.
- Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1803–1813.
- Xu Chen, Yongfeng Zhang, and Ji-Rong Wen. 2022. Measuring "why" in recommender systems: a comprehensive survey on the evaluation of explainable recommendation. *arXiv preprint arXiv:2202.06466*.
- Yang Deng, Wenxuan Zhang, Weiwen Xu, Wenqiang Lei, Tat-Seng Chua, and Wai Lam. 2023. A unified multi-task learning framework for multi-goal conversational recommender systems. *ACM Transactions on Information Systems*, 41(3):1–25.
- 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.
- Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1371–1374.
- Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2023. Large language models are zero-shot rankers for recommender systems. *arXiv preprint arXiv:2305.08845*.
- Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, 54(5):1–36.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.
- Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 304–312.
- 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.
- 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, et al. 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, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Wenbo Pan, Qiguang Chen, Xiao Xu, Wanxiang Che, and Libo Qin. 2023. A preliminary evaluation of chatgpt for zero-shot dialogue understanding. *arXiv preprint arXiv:2304.04256*.
- Stefan Papazov, Wesley Gill, Marta Garcia Ferreiro, Andrew Zhu, Lara J Martin, and Chris Callison-Burch. 2022. Using language models to convert between natural language and game commands. In *The Third Wordplay: When Language Meets Games Workshop*.
- Gustavo Penha and Claudia Hauff. 2020. What does bert know about books, movies and music? probing bert for conversational recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 388–397.

- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.
- Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Evaluating mixed-initiative conversational search systems via user simulation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 888–896.
- Quan Tu, Shen Gao, Yanran Li, Jianwei Cui, Bin Wang, and Rui Yan. 2022. Conversational recommendation via hierarchical information modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2201–2205.
- Lingzhi Wang, Huang Hu, Lei Sha, Can Xu, Daxin Jiang, and Kam-Fai Wong. 2022a. Recindial: A unified framework for conversational recommendation with pretrained language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 489–500.
- Ting-Chun Wang, Shang-Yu Su, and Yun-Nung Chen. 2022b. Barcor: Towards a unified framework for conversational recommendation systems. *arXiv preprint arXiv:2203.14257*.
- Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022c. 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.
- Le Wu, Xiangnan He, Xiang Wang, Kun Zhang, and Meng Wang. 2022. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. *IEEE Transactions on Knowledge and Data Engineering*.
- Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. Large language models are diverse role-players for summarization evaluation. *arXiv preprint arXiv:2303.15078*.
- Bowen Yang, Cong Han, Yu Li, Lei Zuo, and Zhou Yu. 2022. Improving conversational recommendation systems’ quality with context-aware item meta-information. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 38–48.
- Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023a. Recommendation as instruction following: A large language model empowered recommendation approach. *arXiv preprint arXiv:2305.07001*.
- 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.
- Xiaoyu Zhang, Xin Xin, Dongdong Li, Wenxuan Liu, Pengjie Ren, Zhumin Chen, Jun Ma, and Zhaochun Ren. 2023b. Variational reasoning over incomplete knowledge graphs for conversational recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 231–239.
- Yizhe Zhang, Siqu Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.
- Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 177–186.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiao-lei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023a. A survey of large language models. *arXiv preprint arXiv:2303.18223*.
- Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023b. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.
- Jinfeng Zhou, Bo Wang, Ruifang He, and Yuexian Hou. 2021a. Crfr: Improving conversational recommender systems via flexible fragments reasoning on knowledge graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4324–4334.
- Jinfeng Zhou, Bo Wang, Minlie Huang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2022. Aligning recommendation and conversation via dual imitation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 549–561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2021b. Crslab: An open-source toolkit for building conversational recommender system. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 185–193.

Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1006–1014.

A Prompts

A.1 Prompts for ChatGPT in the Traditional Evaluation

We use the following prompts for zero-shot prompting in section 3.1.

- ReDial

Recommend 10 items that are consistent with user preference. The recommendation list can contain items that the dialog mentioned before. The format of the recommendation list is: no. title (year). Don't mention anything other than the title of items in your recommendation list.

- OpenDialKG

Recommend 10 items that are consistent with user preference. The recommendation list can contain items that the dialog mentioned before. The format of the recommendation list is: no. title. Don't mention anything other than the title of items in your recommendation list.

A.2 Prompts for ChatGPT in Our New Evaluation Approach

We use the following prompts for ChatGPT in our new evaluation approach.

A.2.1 Recommendation

- Free-Form Chit-Chat

- ReDial

You are a recommender chatting with the user to provide recommendation. You must follow the instructions below during chat.
If you do not have enough information about user preference, you should ask the user for his preference.
If you have enough information about user preference, you can give recommendation. The recommendation list must contain 10 items that are consistent with user preference. The recommendation list can contain items that the dialog mentioned before. The format of the recommendation list is: no. title (year). Don't mention anything other than the title of items in your recommendation list.

- OpenDialKG

You are a recommender chatting with the user to provide recommendation. You must follow the instructions below during chat.
If you do not have enough information about user preference, you should ask the user for his preference.
If you have enough information about user preference, you can give recommendation. The recommendation list must contain 10 items that are consistent with user preference. The recommendation list can contain items that the dialog mentioned before. The format of the recommendation list is: no. title. Don't mention anything other than the title of items in your recommendation list.

- Attribute-Based Question Answering: “{ }” refers to the options that have been selected.

- ReDial

To recommend me items that I will accept, you can choose one of the following options.
A: ask my preference for genre
B: ask my preference for actor
C: ask my preference for director
D: I can directly give recommendations
You have selected {}, do not repeat them. Please enter the option character.

– OpenDialKG

To recommend me items that I will accept, you can choose one of the following options.
A: ask my preference for genre
B: ask my preference for actor
C: ask my preference for director
D: ask my preference for writer
E: I can directly give recommendations
You have selected {}, do not repeat them. Please enter the option character.

A.2.2 Explainability

Please explain your last time of recommendation.

A.3 Prompts for the User Simulator in Our Proposed Evaluation Approach

We use the following prompts for text-davinci-003 to play the role of the user during interaction.

- Free-Form Chit-Chat: “{}” refers to the item labels of each example in the datasets.

You are a seeker chatting with a recommender for recommendation. Your target items: {}. You must follow the instructions below during chat.
If the recommender recommends {}, you should accept.
If the recommender recommends other items, you should refuse them and provide the information about {}. You should never directly tell the target item title.
If the recommender asks for your preference, you should provide the information about {}. You should never directly tell the target item title.

- Attribute-Based Question Answering

- When the recommended item list contains at least one of the target items:

That’s perfect, thank you!

- When the recommended item list does not contain any target item:

I don’t like them.

- When the system asks about the preference over pre-defined attributes, we use the attributes of target items as the answer if they exist, otherwise:

Sorry, no information about this.

A.4 Prompts for the LLM-based Scorer in Our Proposed Evaluation Approach

We use the following prompts for text-davinci-003 to score the persuasiveness of the explanations generated by CRSs. “{}” refers to the item labels of each example in the datasets.

Does the explanation make you want to accept the recommendation? Please give your score.
If mention one of [{}], give 2.
Else if you think recommended items are worse than [{}], give 0.
Else if you think recommended items are comparable to [{}] according to the explanation, give 1.
Else if you think recommended items are better than [{}] according to the explanation, give 2.
Only answer the score number.

B Related Work

In this section, we summarize the related work from the following perspectives.

B.1 Conversational Recommender System

The field of conversation intelligence (Chen et al., 2017; Gao et al., 2018) and recommendation systems (Wu et al., 2022) has seen significant progress in recent years. One promising development is the integration of these two fields, leading to the emergence of conversational recommender systems (CRSs) (Jannach et al., 2021; Gao et al., 2021). CRSs provide recommendations to users through conversational interactions, which has the potential to significantly improve the user experience.

One popular approach (Lei et al., 2020; Tu et al., 2022) assumes that interactions with users primarily take the form of question answering, where users are asked about their preferences for items and their attributes. The goal is to learn an optimal interaction strategy that captures user preferences and provides accurate recommendations in as few turns as possible. However, this approach often relies on hand-crafted templates and does not explicitly model the language aspect of CRSs. Another approach (Zhang et al., 2023b; Zhou et al., 2022) focuses on engaging users in more free-form natural language conversations, such as chit-chat. The aim is to capture user preferences from the conversation context and generate recommendations using persuasive responses.

In this work, we systematically evaluate the performance of ChatGPT for conversational recommendation on large-scale datasets.

B.2 Language Models for Conversational Recommendation

There have been recent studies on how to integrate language models (LMs) into CRSs. One notable investigation by Penha and Hauff (2020) evaluates the performance of the pre-trained language model (PLM) BERT (Kenton and Toutanova, 2019) in conversational recommendation. Other studies (Wang et al., 2022c; Yang et al., 2022; Deng et al., 2023) primarily utilize PLMs as the foundation to build unified CRSs, capable of performing various tasks using a single model, instead of multiple components. However, the current approaches are confined to small-size LMs like BERT (Kenton and Toutanova, 2019) and DialoGPT (Zhang et al., 2020).

In this paper, we focus on the evaluation of CRSs in the era of LLMs and propose a new evaluation approach iEvaLM.

B.3 Evaluation and User Simulation

The evaluation of CRSs remains an area that has not been thoroughly explored in existing literature. Previous studies have primarily focused on turn-level evaluation (Li et al., 2018; Chen et al., 2019), where the system output of a single turn is compared against ground-truth labels for two major tasks: conversation and recommendation. Some researchers have also adopted conversation-level evaluation to assess conversation strategies (Lei et al., 2020; Zhang et al., 2018). In such cases, user simulation is often employed as a substitute for human evaluation. These approaches typically involve collecting real user interaction history (Lei et al., 2020) or reviews (Zhang et al., 2018) to represent the preferences of simulated users. However, due to the intricate and interactive nature of conversational recommendation, the evaluation is often constrained by pre-defined conversation flows or template-based utterances. Consequently, this limitation hinders the comprehensive assessment of the practical utility of CRSs.

In our work, we propose an interactive evaluation approach iEvaLM with LLM-based user simulators, which has a strong instruction-following ability and can flexibly adapt to different CRSs based on the instruction without further tuning.