LLM-REDIAL: A Large-Scale Dataset for Conversational Recommender Systems Created from User Behaviors with LLMs

Anonymous ACL submission

Abstract

The large-scale conversational recommendation dataset is pivotal for the development of conversational recommender systems (CRS). Most existing CRS datasets suffers from the problems of data inextensibility and semantic inconsistency. To tackle these limitations and establish a benchmark in the conversational recommendation scenario, in this paper, we introduce the LLM-REDIAL dataset to facilitate the research in CRS. LLM-REDIAL is constructed by leveraging large language models (LLMs) to generate the high-quality dialogues. To provide the LLMs with detailed guidance, we integrate historical user behavior data with dialogue templates that are carefully designed through the combination of multiple pre-defined goals. LLM-REDIAL has two main advantages. First, it is the largest multi-domain CRS dataset which consists of 46.9k multi-turn dialogues with 465.9k utterances across 4 domains. Second, dialogue semantics and the users' historical interaction information is highly consistent. Human evaluation are conducted to verify the quality of LLM-REDIAL. In addition, we evaluate the usability of advanced LLM-based models on LLM-REDIAL. Our dataset will be released on github recently.

1 Introduction

In recent years, conversational recommender systems (CRS) have been widely explored in both academia and industry (Zhou et al., 2020a; He et al., 2023; Zhou et al., 2021), which leverage natural language conversations to provide users with personalized and context-aware recommendations. Unlike the conventional recommender systems that rely solely on user-item interactions, CRS incorporates the conversational aspect, allowing users to interact with the system through natural language.

The existing CRS methods are primarily data-driven, requiring large-scale conversational datasets for model training. In this connection, an increasing emphasis has been placed on dataset construction in the field of CRS. There are a few efforts to build datasets for conversational recommendation (Li et al., 2018; Zhou et al., 2020b; Liu et al., 2020; Manzoor and Jannach, 2022). Table 1 lists some commonly known CRS datasets. The REDIAL dataset (Li et al., 2018) consisting of over 10,000 dialogues was realised to the community for conversational movie recommendation. REDIAL was collected by pairing up Amazon Mechanical Turk (AMT) workers and guiding them to engage in a dialogue with the purpose of recommending movies. A topic-guided CRS dataset named TG-ReDial (Zhou et al., 2020b) was constructed with the topic threads-based utterance retrieval and human annotation. DuRecDial (Liu et al., 2020) is a human-to-human recommendation oriented multi-type dialog dataset which was created by manual annotation with pre-defined goals.

While these existing datasets have propelled the development of conversational recommendation to some extent, there are still the following limitations of two aspects: (1) **Data inextensibility.** Most of previous dataset construction require a lot of human annotations significantly limiting the dataset scalability. Additionally, the quality of dialogue texts obtained through sentence retrieval or crowd-workers can not be guaranteed. Insufficient quantity and quality of dialogues would hinder the training of CRS models. Even with the emergence of Large Language Models (LLMs), this situation persists. While current LLMs demonstrate superior capabilities in text generation, they exhibit less promising performance in conversational recommendation. Consequently, large-scale conversational recommendation datasets remain a bottleneck in the development of CRS. (2) **Semantic Inconsistency.** The surge of LLMs making the response generation in CRS less challenging, and the research focus is gradually shifting towards the recommendation aspect. The consistency between dialogues

Table 1: Comparison of LLM-REDIAL with other datasets for conversational recommendation.

Datasets	#Dialogues	#Utterances	#Tokens	#4-Grams	Domains	User Interaction	User-Centric
REDIAL	10k	182k	4.5k	58k	Movie	No	No
TG-REDIAL	10k	129k	50k	7.5k	Movie	No	No
DuRecDial	10.2k	156k	17.6k	461k	Movie, music, food, etc	No	No
INSPIRED	1k	35k	11k	182k	Movie	No	No
OpenDialKG	15k	91k	22k	547k	Movie, book	No	No
LLM-REDIAL	46.9k	465.9k	116k	4.4M	Movie, book, sport, etc	Yes	Yes

and users' actual behaviors is a choke point for the assessment of recommendation. Neither the simulated dialogue generated by crowd-workers nor the user profile-based semi-automatic dialogue generation can maintain semantic consistency between the conversation content and users' historical behaviors. Because these generation methods typically only specify the start of dialogues and the final goal or topic of recommendation, they fail to fully leverage the users' truely historical behaviors to present the recommendation process. Consequently, a dataset that aligns the semantics in dialogue texts with users' behavior is indispensable for the thorough evaluation of conversational recommendation.

To address the above limitations, in this paper, we construct a new large-scale dataset for CRS created from user behaviors through LLMs (LLM-REDIAL). For the first limitation, we introduce the LLMs to generate a large quantity of high-quality dialogue sentences under the guidance of pre-defined dialogue templates. To the best of our knowledge, this is the largest conversational recommendation dataset with multiple domains. Table 1 shows that our LLM-REDIAL contains 46.9k multi-turn dialogues with 465.9k utterances across 4 domains. For the second limitation, we create a collection of templates by assigning each turn a goal in the dialogues. By filling these dialogue templates with the users' behaviors including both positive and negative feedbacks along with review information, the prompts are derived for the LLMs to generate the complete multi-turn dialogues covering the recommendation process. In this manner, the consistency between the dialogue semantics and the users' actual interactions can be effectively guaranteed. Furthermore, LLM-REDIAL is user-centric, which means the user of each dialogue can be identified and all the dialogues and historical interactions associated with one specific user can be located in our dataset.

2 Dataset Construction

In this section, we present the process of dataset construction. We first introduce the data source for dialogue generation. Then the overview of data construction is described. After that, the details of each step are introduced sequentially.

2.1 Data Source

To approach the realistic conversational recommendation scenario as closely as possible, we construct the dataset based on authentic user historical behaviours. In addition, we aim at naturally incorporating relevant item details, making the dialogues appear more reasonable and real. Therefore, we select the product reviews from Amazon[1] as the data base. The review data contains user reviews along with rating information from Amazon platform. Specifically, the ratings of each user are used to identify the preference which would be combined with the corresponding review texts to generate the dialogues. In this manner, each dialogue is associated with one user's historical interactions. The combination of these elements forms a complete data for conversational recommendation.

2.2 Overview of Dataset Construction

As Figure 1 shows, the overall process of the dataset construction sequentially consists of data preprocessing, template construction, and dialogue generation. First, the raw data of Amazon reviews are processed through the operations of data filtering, grouping, and splitting to obtain the historical interactions and the item list to be predicted for each user. The following template construction module designs the multiple goals for utterances and formulates templates for multi-turn dialogues by combining these goals. In the dialogue generation phase, the publicly available LLM is invoked to generate the dialogues implying the recommendation process based on pre-designed prompts which are derived by filling the dialogue templates with users' behaviors and reviews.

2.3 Data Preprocessing

In order to smoothly utilize the raw review data to generate dialogues that centered around the func-

[1] http://jmcauley.ucsd.edu/data/amazon

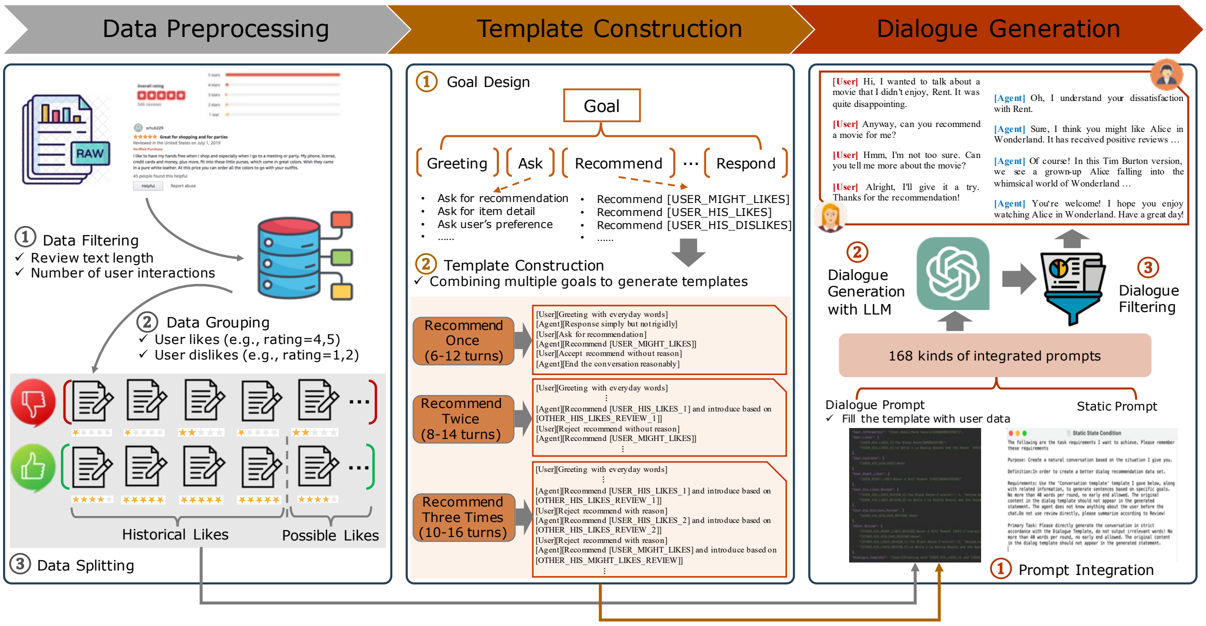

Figure 1: Overview of the LLM-REDIAL dataset construction framework consisting of data preprocessing, template generation, and dialogue generation.

tion of providing recommendations, we design a series of data preprocesses to filter out the interactions that meets the requirements. Due to the presence of non-word tokens in the review texts, we firstly tokenize the texts and remove those irregular tokens. After that, to guarantee the usability of the review content while avoiding excessively long text that may not provide accurate semantic information, we filter the review texts and retain records with a word count between 20 and 400. Besides, to ensure that the interaction quantity for each user is sufficient to support the generation of dialogues representing the recommendation process, we impose restrictions on the number of interactions. Specifically, we remove the users and items with less than 10 interactions. To make the dialogue content more diverse, it is expected to not only reflect situations where users accept recommendations but also those where users reject recommendations. Therefore, we intend to incorporate interactions of both user likes and dislikes into the dialogue. Ratings equal to or higher than 4 are picked out as positive feedbacks, while those equal to or lower than 2 are used as negative ones. Finally, the positive and negative interactions are sorted chronologically to form two collections (LIKES and DISLIKES) that prepare for generating prompts in the subsequent dialogue generation step. It should be noted that the last 10% of the positive interactions of each user are moved to a new collection (MIGHT_LIKES) from which the items are selected as the final golden recommendation in the dialogues.

2.4 Template Construction

2.4.1 Goal Design

To make the dialogue proceed along the lines of recommendations, we carefully design multiple kinds of primary goals for the utterances referring to the communicative functions from the international standard ISO 244617-2. Table 2 shows that there are total 8 primary goals based on which 30 detailed sub-goals are provided. The primary goals are used to decide the function of each utterance. Under each primary goal, there are several sub-goals of two types. One is the fixed instruction that indicates the more specific aspect, such as "Ask for recommendation" and "Ask for item detail". The other type is the flexible instruction, mainly consisting of the fixed instruction and a slot to be filled, such as "Recommend `[USER_HIS_LIKES]`", where `[USER_HIS_LIKES]` would be filled with an item randomly sampled from the LIKES set collected from historical items with positive feedbacks. Table 2 lists the details of all the pre-defined goals.

2.4.2 Template Construction

To offer the LLMs the more instructive inputs for the generation of fluent and natural conversations, we construct various dialogue templates each of which is composed of multiple sub-goals. Specifically, to enhance the diversity of dialogues, we set different templates based on the frequency of recommendations with the count restricted to 1-3

Table 2: The primary goals and sub-goals for the utterances

Primary Goal	Sub-Goal	Description
Greating	Greeting with [USER_HIS_DISLIKES] and [USER_HIS_DISLIKES_REVIEW]	The user starts the conversation with the user's likes item
	Greeting with [USER_HIS_LIKES] and [USER_HIS_LIKES_REVIEW]	The user starts the conversation with the user's dislikes
Ask	Ask for recommendation	The user seeks for recommendations
	Ask for item detail	The user asks for specific information about the item
	Ask for user's preferenc	The system asks for user preferences
	Ask if need more recommend	The user is asked if they want more recommendations
Respond	Responds with [OTHER_REVIEW]	The system uses other people's reviews to reply
	Response simply but not rigidly	The system replies simply and politely
	Responds in detail	The system replies in detail
	Responds according to the user's mood	The system replies according to the user's mood
Recommend	Recommend [USER_MIGHT_LIKES]	The system recommends items that will be accepted
	Recommend [USER_HIS_LIKES]	The system recommends items that will not be accepted but the user likes
	Recommend [USER_HIS_DISLIKES]	The system recommends items that the user dislikes
Feedback	Accept recommendation without reason	The user accepts recommendation without reason
	Accept recommendation with reason	The user accepts recommendation for some reason
	Express interest	The user expresses interest in the item
	Reject recommendation without reason	The user rejects recommendation without reason
	Reject recommendation with reason	The user rejects recommendation for some reason
Chit-Chat	Chit-Chat	Make a transition between the beginning and the end of a conversation
Talk	Lead the converstaion to recommend	The system directs the conversation to the recommended task
	Lead the conversation to end	The user leads the conversation to the end
	Introduce based on [OTHER_DISLIKES_REVIEW]	The system uses other people's comments to introduce items the user dislikes
	Introduce based on [OTHER_HISLIKES_REVIEW]	The system uses other people's comments to introduce items the user likes
	Introduce based on [OTHER_MIGHTLIKES_REVIEW]	The system uses other people's comments to introduce items will be accepted
	Accept and express gratitude	The user accepts recommendation and express gratitude
	Reject and express gratitude	The user rejects recommendation and express gratitude
	Express gratitude	The user expresses gratitude
	End the converstaion reasonably	The system reasonably ends the session
Reason	Have seen the movie before	One of the reasons users reject recommendations
	[USER_HIS_DISLIKES_REVIEW]	One of the reasons users reject recommendations

times. For the settings where recommendations are made 2 or 3 times, except for the final recommendation, all preceding recommendations are assumed to be rejected. Correspondingly, based on the three setting types, the ranges of dialogue lengths are also restricted differently. Referring to the dialogue lengths of most existing CRS datasets that range around 6-16, such as the datasets listed in Table 1, we constrained the dialogue lengths of all the settings within the same range. In setting with a higher number of recommendations, the dialogue length is extended accordingly. The combinations of goals are manually and carefully designed and finally 168 dialuge templates are obtained. Figure 2 (a) displays an example of the template that makes recommendation once with 8 utterances.

2.5 Dialogue Generation

2.5.1 Generation with LLMs

The prompt that fed into the LLMs is formed by integrating a pre-defined static prompt and a concretized template. The static prompt provides the task description and requirements with simple plain language statements as shown in Figure 2 (b). It is worth noting that, to establish a strong connection between dialogue content and item information, we introduce the real users' reviews of the historical interactions to enrich the dialogue, while avoiding verbatim replication of review content. To prevent the dialogue from becoming overly verbose and ensure the quality of sentence generation, we limit the length of each sentence to 60 words.

The concretized template is achieved by filling in user information into the slots of the dialogue template. Specifically, for the generation of each dialogue, user information is obtained by sampling interactions and review texts from the historical behavior of one specified user. Figure 2 (c) shows an example of user information which is structured in a JSON file. By concatenating the static prompt and the concretized template, the complete prompt to be fed into the LLMs are constructed.

To facilitate reproducibility, we adopt the static version of ChatGPT[2], *i.e.,* GPT-3.5-turbo, to generate the dialogues for conversational recommendation. Based on the integrated prompt shown in Figure 2 (b) and (c), Figure 2 (d) presents the complete dialogue output by GPT-3.5-turbo. It can be observed that the dialogue flow smoothly follows the designed dialogue template, and the key steps such as requesting recommendations, providing recommendations, and accepting recommendations are well reflected in the dialogue (the underlined words). Benefitting from the powerful generation capabilities of LLMs, the generated sentences seamlessly incorporate the item information

[2]https://openai.com/blog/chatgpt

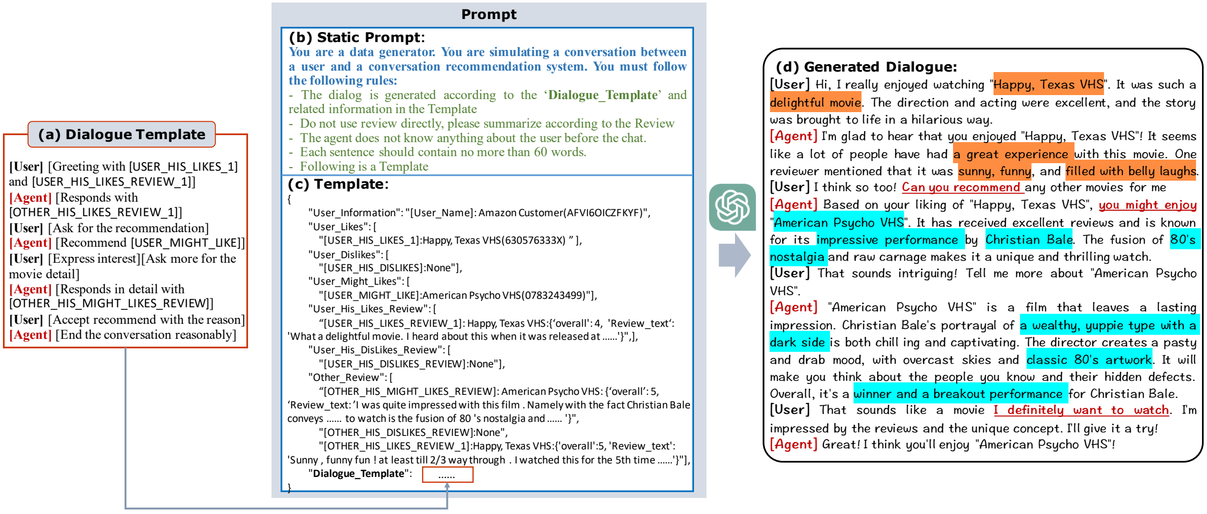

Figure 2: Overview of the LLM-REDIAL dataset construction framework consisting of data preprocessing, template generation, and dialogue generation.

from the relevant review texts and express in a natural and coherent manner, enhancing the diversity and authenticity of the dialogue. More importantly, the incorporation of items and related review information effectively strengthens the consistency between dialogue content and users' historical behaviors, which is more aligned with the scenario of conversational recommender systems.

2.5.2 Dialogue Filtering

Due to the randomness of LLMs and the long review texts that may confuse the model, the conversations directly generated by ChatGPT contains some invalid, noisy cases, which might has detrimental impact on the research using this dataset for conversational recommendation. To tackle this issue, we design the following automatic data filtering processes to filter out the high-quality multi-turn dialogues step by step: (1) We remove the dialogues that are not completely generated; (2) We ignore and discard the dialogues containing garbled or unreadable characters. (3) We remove the dialogues that contains the template information, *i.e.,* the slots in the templates are not successfully filled with the user information. (4) We discard the dialogues that are inconsistent in length with the related dialogue templates. Through the above data filtering procedure, the final large-scale CRS dialogues could be better utilized to investigate the conversational recommendation methods.

3 Dataset Statistics

Our LLM-REDIAL is constructed based on the Amazon review dataset. There are 24 different domains and this work selects 4 of them to be the data sources. More domains will be used to generate more conversations in our future work. The LLM-REDIAL consists of 46,964 dialogues with 465,896 utterances across 4 domains. The statistics of our LLM-REDIAL are shown in Table 3. On average, each dialogue session in 4 domains has 9∼10 utterances since we design three kinds of dialogue template with fixed ranges of dialogue length. One distinctive character of our dataset is its user-centric focus, each user has two corresponding dialogue sessions on average. Compared to Sports and Electronics categories, users in the Books and Movies categories have the higher average numbers of dialogues, possibly due to longer historical interaction sequences for book and movie purchases. Compared to the existing available conversational recommendation datasets, LLM-REDIAL has a significantly larger scale of dialogues. The abundance of unique tokens and 4-grams indicates that the dialogues generated based on LLMs and users' interaction information with reviews convey the richer and more diverse semantic information which is conducive to recommendation.

4 Evaluation

4.1 Human Evaluation on Dataset Quality

To perform a thorough and direct assessment of the quality of our curated dataset, we chose a representative CRS dataset (*i.e.,* REDIAL) for comparative analysis. Subsequently, we conduct a human evaluation to measure the effectiveness and reliability of our constructed dataset, incorporating assessments at both the utterance and conversational levels.

Table 3: Data statistics of our LLM-REDIAL dataset.

	Books	Movies	Sports	Electronics	Total
#Dialogues	25,080	10,093	6,218	6,260	47,651
#Utterances	259,850	106,151	58,289	58,394	482,684
#Tokens	79,540	40,285	35,137	31,331	124,269
#4-Grams	2,385,204	1,100,472	757,201	679,257	4,679,146
# Users	9,893	3,133	5,128	4,469	22,151
# Items	112,913	11,589	34,733	18,034	177,269
Avg. #Dialogues per User	2.54	3.22	1.21	1.40	2.15
Avg. #Utterances per Dialogue	10.36	10.52	9.37	9.33	10.13

Table 4: Human evaluation results on the LLM-REDIAL dataset.

	Fluency(0-2)	**Informative(0-2)**	**Logicality(0-2)**	**Coherence(0-2)**
LLM-REDIAL	1.98	1.28	1.90	1.88
REDIAL	1.83	1.18	1.76	1.77

4.1.1 Utterance-Level Evaluation

In the utterance-level evaluation, we randomly sampled 10 dialogues from each of the two datasets, REDIAL (Li et al., 2018) and our LLM-REDIAL. The order of these 20 dialogues is randomly shuffled. The selected dialogues of LLM-REDIAL totally contain 112 utterances with given contexts while these of REDIAL have 103 utterances.

We then evaluated the quality of utterances based on four aspects: (1) Fluency: Assessing whether a response is organized in regular English grammar and is easy to understand. (2) Informativeness: Determining whether a response is meaningful and not a "safe response", with repetitive responses considered uninformative. (3) Logicality: Evaluating the logical consistency of a response by assessing whether it aligns with common sense reasoning and follows a logical flow. (4) Coherence: Ensuring that a response is coherent with the previous context. We enlisted seven annotators (students) to evaluate responses on these four aspects, using a scale of 0, 1, 2 (a more detailed rating scheme can be found in Appendix A).

The results of the human evaluation on two datasets are presented in Table 4. The utterances of our LLM-REDIAL dataset achieve the higher scores than those in REDIAL in terms of all the four metrics. The utterances in our dataset exhibit extremely high fluency, logicality, and coherence, which benefits from the strong generation capability of LLMs. Compared to REDIAL, the superiority in information expression of utterances in LLM-REDIAL is significant. It is mainly because we incorporate the users' historical interactions with review information in the dialogue templates for LLMs-based generation, while REDIAL relies on the temporarily paired two crowd-workers to generate dialogues, making it challenging to delve into detailed and in-depth topics.

4.1.2 Conversation-Level Evaluation

For the conversation-level evaluation, we assess quality through direct pair comparisons, asking annotators to determine which of the two provided conversations (note that the order of sourced datasets is randomized) exhibits higher quality. Specifically, we randomly select 50 dialogues from each of the two datasets, forming 50 pairs through random matching. We assign seven annotators to annotate all the 50 pairs of dialogues and select the one with overall higher quality for each pair. Finally, we obtained 350 annotations out of which 88% indicate our LLM-REDIAL has the better quality.

4.2 Evaluation on Conversational Recommendation

We conduct a series of experiments on the dataset of Movie domain to show the applicability of LLM-REDIAL on the task of conversational recommendation and emphasize the importance of user-centric dialogues with interactions. Since generating dialogue texts is not a particularly challenging task for LLMs, our focus is on the recommendation task. We use Recall@K and NDCG@K ($K = 10, 50$) as evaluation metrics.

4.2.1 Baselines

To verify the practicable of the constructed LLM-REDIAL, we consider the following baselines for performance comparison: (1) ChatGPT-based model: We use GPT-3.5-turbo from OPENAI[3] as recommender. (2) Vicuna-based model: Refer

[3]https://openai.com/

Table 5: Performance of the LLM-based models on our LLM-REDIAL and REDIAL.

Methods		REDIAL						LLM-REDIAL					
		R@5	R@10	R@50	N@5	N@10	N@50	R@5	R@10	R@50	N@5	N@10	N@50
ChatGPT-based													
Zero-Shot	Dial. Only	0.0100	0.0100	0.0150	0.0072	0.0071	0.0085	0.0000	0.0000	0.0400	0.0000	0.0000	0.0086
	Dial. + H. I	/						0.0000	0.0050	0.0350	0.0000	0.0015	0.0077
Few-Shot	Dial. Only	0.0100	0.0150	0.0200	0.0100	0.0115	0.0130	0.0000	0.0000	0.0350	0.0000	0.0000	0.0075
	Dial. + H. I	/						0.0000	0.0000	0.0400	0.0000	0.0000	0.0087
Fine-Tuning	Dial. Only	0.2000	0.2600	0.4400	0.1757	0.1953	0.2021	0.2625	0.3150	0.5175	0.1716	0.1768	0.2353
	Dial. + H. I	/						0.4500	0.4600	0.5100	0.4270	0.4295	0.4265
Vicuna-based													
Zero-Shot	Dial. Only	0.0005	0.0007	0.0013	0.0001	0.0003	0.0004	0.0010	0.0013	0.0027	0.0007	0.0006	0.0010
	Dial. + H. I	/						0.0033	0.0080	0.0507	0.0025	0.0034	0.0128
Few-Shot	Dial. Only	0.0004	0.0007	0.0053	0.0005	0.0007	0.0016	0.0000	0.0027	0.0100	0.0000	0.0009	0.0026
	Dial. + H. I	/						0.0080	0.0133	0.0553	0.0073	0.0089	0.0172
Fine-Tuning	Dial. Only	0.1945	0.3018	0.4993	0.1397	0.1642	0.2080	0.2869	0.3325	0.6090	0.2624	0.2684	0.2988
	Dial. + H. I	/						0.3260	0.3980	0.6940	0.2569	0.2655	0.3108

to (He et al., 2023), we consider the representative open-sourced Vicuna (Chiang et al., 2023) to be the recommender.

Both ChatGPT and Vicuna-based models take the preceding context of each dialogue as input to predict the item that will appear in the next response. Specially, we consider three settings which are zero-shot, few-shot, and fine-tuning. For the ChatGPT-based model, we randomly select 200 dialogues for testing. In the few-shot setting, we offer 5 case as examples. In the fine-tuning setting, we use 200 training examples to fine-tune. For the Vicuna-based model, we randomly select 1,500 dialogues for testing. In the few-shot setting, we offer 5 case as examples. In the fine-tuning setting, we use the remaining 8,593 training examples to fine-tune. As the LLM-based models offer the recommendation through the way of generative retrieval, we follow (He et al., 2023) and apply a fuzzy matching to transfer the generated textual recommendation list to a item ranking list.

4.2.2 Results and Analysis

Table 5 reports the performance of different baseline models on recommendation task. 'Dial. Only' indicates that only the dialogue texts are fed into the LLMs to generate the results, and 'Dial. + H. I' represents that both dialogue texts and user's historical interactions are considered to be the inputs. It can be observed that both ChatGPT-based and Vicuna-based models achieve clear improvements from fine-tuning on training data. Compared to the fine-tuning setting of ChatGPT, fine-tuning Vicuna obtains the better performance. It is probably because considering the scale of ChatGPT, the training data we provide is much less than that of Vicuna. The incorporation of users' historical interactions effectively improves the recommendation performance for all the three settings, with the most significant enhancement in the fine-tuning setting. The experimental results demonstrates that the user's historical interaction records are quite crucial in the scenario of CRS. However, most existing CRS datasets predominantly focus on the dialogue text. The conversations in these datasets often can not associated with the specific users, making it impossible to identify the corresponding historical interaction information.

4.2.3 Case Study

To more intuitively explore the effect of response generation for recommendation based on the LLMs under different settings, we provide an example of generating response with recommendation by ChatGPT-based model in Figure 3. As demonstrated, in both the zero-shot and few-shot settings, the generated responses are coherent and natural while the recommendation performance is relatively poor. In other words, the introduction of LLMs makes the task of response generation in traditional CRS more straightforward, while there is still significant room for improvement for the recommendation task. After fine-tuning, it is more likely to make recommendations meeting users' requirements in the generated responses.

5 Related Work

5.1 Conversational Recommender Systems

Dialogue recommendation systems can be classified based on the number of dialogue turns, distinguishing between single-turn dialogue recommendation and multi-turn dialogue recommendation.This paper is focus on multi-turn dialogue recommendation systems (Srivastava et al., 2023) (Balaraman and Magnini, 2020) (He et al., 2023). The multi-round dialog recommendation system was studied by (Lei et al., 2020a) which

Input	
Historical Interactions	['Sleuth', '**Game Change**', 'Eat Pray Love', "All the King's Men", 'The Double']
Dialogue	**User:** Hey, I recently watched "Vicky Cristina Barcelona" and I loved it. The movie was a fun trip with … Scarlet and Bardem's chemistry was amazing. **Agent:** Hello, I'm glad to hear that you enjoyed "Vicky Cristina Barcelona." … Speaking of recommendations, how about checking out "**Game Change**"? It's an awesome movie that … **User:** Thanks for the recommendation, but I've already seen "**Game Change.**" Although the film was amusing and scary at times, I found it concerning to witness the lack of qualifications of some political figures.
Output	
Zero-Shot	**Agent:** I understand your concerns about the lack of qualifications in politics. If you\'re interested in exploring more thought-provoking films on this topic, I would recommend "**The Ides of March**." It delves into the dark side of political campaigns and the ethical dilemmas faced by those involved.
Few-Shot	**Agent:** I completely understand your perspective on "Game Change." If you're interested in exploring more political themes, you might enjoy "**The Ides of March**." It's a gripping political drama that delves into the complexities of campaigns and power struggles. Ryan Gosling and George Clooney deliver stellar performances.
Fine-Tuning	**Agent:** I understand your perspective. In that case, how about considering "**Ghost Dog - The Way of the Samurai**"? It\'s a unique gangster film with a touch of absurdity. The cinematography is exceptional, and it offers a fresh take on the genre.
Ground Truth	**Agent:** I understand your concerns. In that case, may I suggest another film? "**Ghost Dog - The Way of the Samurai**" is a unique gangster movie with a parallel social commentary. Forest Whitaker's portrayal of the hitman is exceptional.

Figure 3: Case study of response generation for recommendation based on LLMs under different settings.

allowing the Conversational Recommendation System (CRS) to pose multiple questions or recommend items across turns until the user accepts or exits the recommendation. To address challenges associated with multi-turn CRS, (Lei et al., 2020b) leveraged knowledge graphs to select more relevant attributes for cross-turn inquiries. (Xu et al., 2021) dynamically adjusted user embeddings based on user feedback on attributes and items, extending the work of (Lei et al., 2020a). (Deng et al., 2021) (Chu et al., 2023) unified the problem selection module and recommendation module in a reinforcement learning-based CRS solution. However, all the aforementioned works rely on carefully designed heuristic reward functions, which may lead to strategies deviating from the optimal solution.

5.2 Datasets for Conversational Recommendation

In order to enhance the performance of conversational recommendation systems (CRS) and facilitate dialogue recommendation, many researchers have curated dialogue datasets based on specific rules (Li et al., 2018) (Chen et al., 2019) (Jannach et al., 2021) (Lu et al., 2021). (Hayati et al., 2020) manually annotated each utterance using social strategies to validate the effectiveness of social recommendation strategies in CRS. (Moon et al., 2019) provided a parallel dialogue Knowledge Graph (KG) corpus, where each mention of an entity is manually linked to its corresponding KG path. (Liu et al., 2020) created a multi-type dialogue dataset, aiming for bots to naturally guide conversations from non-recommendation types to recommendation types. Similarly, (Zhou et al., 2020b)introduced a topic-guided dialogue recommendation dataset to facilitate the transition of dialogue topics. However, Some studies (Liu et al., 2016) (Novikova et al., 2017) (Gao et al., 2021)pointed out that existing datasets lack the qualification to develop CRS that meet industrial application requirements due to the following reasons: 1) these datasets are insufficient in scale to cover real-world entities and concepts; 2) dataset construction is carried out under strict conditions, making it challenging to generalize to complex and dynamic real-world dialogues. Therefore, developing a large-scale, generalizable, and naturally occurring dialogue dataset is a crucial task.

6 Conclusion

This paper presents a large-scale multi-turn dialogue dataset for conversational recommendation which is constructed with LLMs based on the users' historical behaviours. We fill the user behaviour data into the well-designed dialogue template to guide the LLMs to generate high-quality dialogues. Benefitting from the powerful generation capability of LLMs, LLM-REDIAL is the largest multi-domain CRS dataset with 46.9k dialogues covering recommendation process. Comprehensive experiments are conducted to verify the quality and usability of our LLM-REDIAL. We believe that LLM-REDIAL can serve as a rich resource for advancing research in CRS, assisting the community in proposing better methods for conversational recommendation within the context of LLMs.

References

Vevake Balaraman and Bernardo Magnini. 2020. Proactive systems and influenceable users: Simulating proactivity in task-oriented dialogues. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue-Full Papers, Virually at Brandeis, Waltham, New Jersey, July. SEMDIAL.*

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.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023).*

Zhendong Chu, Nan Wang, and Hongning Wang. 2023. Multi-objective intrinsic reward learning for conversational recommender systems. *arXiv preprint arXiv:2310.20109*.

Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. Unified conversational recommendation policy learning via graph-based reinforcement learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1431–1441.

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.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. Inspired: Toward sociable recommendation dialog systems. In *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 8142–8152. Association for Computational Linguistics (ACL).

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

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

Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020a. 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.

Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020b. Interactive path reasoning on graph for conversational recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2073–2083.

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.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049.

Yu Lu, Junwei Bao, Yan Song, Zichen Ma, Shuguang Cui, Youzheng Wu, and Xiaodong He. 2021. Revcore: Review-augmented conversational recommendation. *arXiv preprint arXiv:2106.00957*.

Ahtsham Manzoor and Dietmar Jannach. 2022. Inspired2: An improved dataset for sociable conversational recommendation. In *KaRS Workshop at RecSys '22*, Seattle, USA.

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.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Harshvardhan Srivastava, Kanav Pruthi, Soumen Chakrabarti, et al. 2023. Core-cog: Conversational recommendation of entities using constrained generation. *arXiv preprint arXiv:2311.08511*.

Kerui Xu, Jingxuan Yang, Jun Xu, Sheng Gao, Jun Guo, and Ji-Rong Wen. 2021. Adapting user preference to online feedback in multi-round conversational recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 364–372.

Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2021. 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. 2020a. 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.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020b. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4128–4139.

A Example Appendix

In this manual evaluation, four key metrics are employed to assess the quality of a dialogue, namely: 1) Fluency; 2) Informativeness; 3) Logical; and 4) Coherence. Each metric is graded on a scale from 0 to 2, with 0 indicating poor performance, 1 signifying moderate performance, and 2 denoting excellent performance. The specific grading criteria for each metric are delineated below:

Fluency:

0 (poor): The dialogue exhibits severe grammatical errors, spelling mistakes, vocabulary issues, or incoherent expressions, rendering it difficult to comprehend.

1 (normal): The dialogue contains some grammar errors, spelling mistakes, vocabulary problems, or lacks fluency, yet remains generally understandable.

2 (good): The dialogue is fluent, devoid of noticeable grammar errors, spelling mistakes, or vocabulary issues, presenting clear and comprehensible expression.

Informativeness:

0 (poor): A dialogue lacking information, essentially falling into 'safe response' territory or consisting of repeated answers.

1 (normal): The dialogue offers some information but still lacks detailed or in-depth responses, requiring additional information for user understanding.

2 (good): The dialogue furnishes rich, detailed, and in-depth information, answering user queries while providing additional relevant content.

Logical:

0 (poor): The dialogue features severe logical errors, responses or suggestions are unrelated to user queries or context, or there are blatant self-contradictions.

1 (normal): The dialogue has some logical issues, with responses or suggestions insufficiently related or reasonable, though still understandable.

2 (good): The dialogue maintains logical coherence, with responses or suggestions logically related and reasonable in the context of user queries.

Coherence:

0 (poor): The dialogue is highly incoherent, lacking clear contextual connections, with no logical links between various parts of responses or suggestions.

1 (normal): The dialogue is moderately coherent, exhibiting some coherence but with occasional ruptures or insufficient logical connections between contexts.

2 (good): The dialogue is highly coherent, with clear logical connections between responses or suggestions, ensuring smooth transitions between contexts.