PEARL: A Review-Driven Persona-Knowledge Grounded Conversational Recommendation Dataset

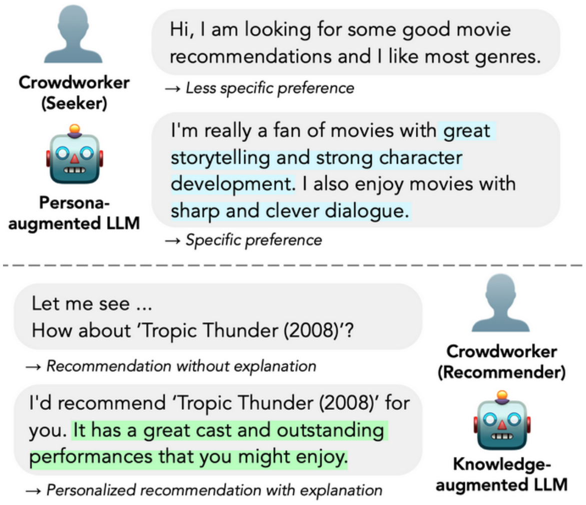
Minjin Kim[1*] **Minju Kim**[1*] **Hana Kim**[1] **Beong-woo Kwak**[1]
SeongKu Kang[3] **Youngjae Yu**[1] **Jinyoung Yeo**[1] **Dongha Lee**[1†]
Yonsei University, Korea[1]
University of Illinois at Urbana-Champaign, USA[3]
`{minjin.kim,minnju,yjy,jinyeo,donalee}@yonsei.ac.kr` `seongku@illinois.edu`

Abstract

Conversational recommender systems are an emerging area that has garnered increasing interest in the community, especially with the advancements in large language models (LLMs) that enable sophisticated handling of conversational input. Despite the progress, the field still has many aspects left to explore. The currently available public datasets for conversational recommendation lack specific user preferences and explanations for recommendations, hindering high-quality recommendations. To address such challenges, we present a novel conversational recommendation dataset named PEARL, synthesized with persona- and knowledge-augmented LLM simulators. We obtain detailed persona and knowledge from real-world reviews and construct a large-scale dataset with over 57k dialogues. Our experimental results demonstrate that PEARL contains more specific user preferences, show expertise in the target domain, and provides recommendations more relevant to the dialogue context than those in prior datasets. Furthermore, we demonstrate the utility of PEARL by showing that our downstream models outperform baselines in both human and automatic evaluations. We release our dataset[1] and code.[2]

1 Introduction

Recently, conversational recommender system (CRS) has become an emerging research topic, which aims to elicit user preferences and offer personalized recommendations by engaging in interactive conversations. Towards this goal, an increasing emphasis has been placed on constructing high-quality dataset (Li et al., 2018; Liu et al., 2020; Zhou et al., 2020b; Hayati et al., 2020). Existing conversational recommendation datasets are mainly collected via crowdsourcing, which is to gather interactions between two crowdworkers where one plays a role as a user (*i.e.*, recommendation seeker) and the other pretends as a recommender.



Hi, I am looking for some good movie recommendations and I like most genres.
→ *Less specific preference*

I'm really a fan of movies with great storytelling and strong character development. I also enjoy movies with sharp and clever dialogue.
→ *Specific preference*

Let me see ...
How about 'Tropic Thunder (2008)'?
→ *Recommendation without explanation*

I'd recommend 'Tropic Thunder (2008)' for you. It has a great cast and outstanding performances that you might enjoy.
→ *Personalized recommendation with explanation*

Figure 1: An example comparing utterances of crowdworkers and our persona-knowledge augmented LLM simulators.

However, existing dialogues have several limitations that impede downstream CRS models from delivering satisfactory user experiences. First, user preferences expressed in the existing CRS datasets are often less specific. An example includes dialogues with statements like, *"I like most genres"* as shown in Figure 1. Such dialogues lead downstream models to offer recommendations that are generic and less personalized (He et al., 2023; Zhou et al., 2020a). This happens as crowdworkers, playing the role of users, often lack specific preferences during their tasks, unlike most real-world users who seek personalized recommendations. Second, conversational recommendation dialogues often provide suboptimal recommendations and insufficient explanations alongside recommendations, due to the limited knowledge of crowdworkers (Guo

---

[*] Equal contribution

[†] Corresponding author

[1] https://huggingface.co/datasets/DLI-Lab/pearl

[2] https://github.com/kkmjkim/PEARL

Findings of the Association for Computational Linguistics ACL 2024, pages 1105–1120
August 11-16, 2024 ©2024 Association for Computational Linguistics

et al., 2023). A specific example is dialogues containing utterances like *"Let me see... How about Tropic Thunder?"* as shown in Figure 1. Such responses can lead to CRS models generating less accurate and relevant suggestions. Additionally, the absence of explanations alongside recommendations can be a crucial hurdle, preventing users from grasping the items and the rationale behind the recommendations.

In this paper, we present a conversational recommendation dataset named PEARL (Persona and knowledgE Augmented Recommendation diaLogues) that addresses the limitations of existing datasets. We transform reviews into persona and item knowledge and incorporate large language model (LLM)-based simulators augmented with persona and knowledge. Each simulator is designed to enhance the preference specificity and informativeness of the collected data, respectively. The *user simulator* plays the role of a recommendation seeker and is equipped with a persona, which is a set of sentences describing features that the user likes and dislikes. Each persona can help the simulator express a distinct user preference of a single real-world user as it is constructed based on the item reviews written by the same user. By simulating users with distinct preferences and specific needs, we can generate dialogues with consistent and clear preferences. The *recommender simulator* is designed to emulate a recommender with domain knowledge. This addresses the lack of proper recommendations and explanations in existing datasets. Specifically, we incorporate item reviews which can not only provide basic information about items but also reveal soft attributes of items that can be only described through experience (*e.g.*, *"feel-good movies"* or *"upbeat music"*) which could be crucial information in the users' decision-making process.

Our collected dataset includes over 57.2k dialogues simulating over 4k users and covering more than 9k items. Our human evaluation results show that our synthesized dataset is preferred by human raters compared to other crowdsourced dialogues. Moreover, our dataset covers a broader spectrum of user needs as we utilize a large set of authentic reviews written by diverse users.

We also conduct extensive experiments assessing the utility of PEARL through human evaluation and automatic evaluation. The results demonstrate that CRS models trained on PEARL show competitive or better performances in recommendation and response generation tasks compared to models trained on human-annotated datasets. Additionally, human judges consistently favor responses from models trained on PEARL over those from models trained on crowdsourced datasets, across all metrics. Furthermore, our experimental results empirically validate that CRS models trained on PEARL generalize better to unseen dialogues than the models trained on existing datasets.

2 Related Work

2.1 Conversational Recommendation

Conversational recommendation is an emerging task where the main goal is to provide high-quality recommendations to users through natural language conversations. Compared to traditional recommendation tasks, conversational recommendation is a more challenging task as it requires the model to not only recommend appropriate items (*i.e.*, recommendation) but also generate engaging and helpful responses (*i.e.*, response generation).

To facilitate the study of conversational recommendation, several datasets have been proposed by previous work. Li et al. (2018) and Hayati et al. (2020) combine the elements of social chitchat and recommendation dialogues. Zhou et al. (2020b) collect human-machine conversation data guided by pre-defined topics. While most existing work has collected conversational recommendation data through crowdsourcing, those datasets are often not scalable and can suffer from less diverse user preferences and uninformative recommendations.

To handle this limitation, He et al. (2023) scrape single-turn recommendation dialogues from Reddit. However, their practical usability has been limited due to the low quality of scraped dialogues and lack of active interactions. Similar to our work, Lu et al. (2023) generate recommendation dialogues by converting user-item interactions into dialogues with a data-to-text generation model. Nonetheless, as the data-to-text model is trained on existing recommendation dialogues, the generated dialogues still inherit the previous limitations of the crowdsourced dataset. In contrast, we tackle the limitation by fully utilizing LLMs for collecting recommendation dialogues and leveraging user reviews to enhance the diversity and informativeness of collected dialogues.

	PEARL (this work)	ReDial (Li et al., 2018)	INSPIRED (Hayati et al., 2020)	TG-ReDial (Zhou et al., 2020b)	DuRecDial 2.0 (Liu et al., 2021)
Collection method	Synthesized	Crowdsourced	Crowdsourced	Human-Machine	Crowdsourced
Real-world persona	✓	✗	✗	✓	✗
Explained recommendation	✓	✗	✗	✗	✗
Number of dialogues	**57,277**	10,006	1,001	10,000	16,482
Number of users	**4,680**	956	1,594	1,482	2,714
Number of utterances	**548,061**	182,150	35,811	129,392	255,346

Table 1: A comparison of our synthesized dataset to notable conversational recommendation datasets.

2.2 Synthetic Data Generation

LLMs have been increasingly used to synthesize dialogue datasets (Kim et al., 2022; Lu et al., 2023; Chen et al., 2023; Kim et al., 2023; Chae et al., 2023). Kim et al. (2022) build Blended Skill BotsTalk using multiple agents grounded in target skills. Chen et al. (2023) generate dyadic and multiparty conversations with topic words and show they have comparable quality to human-authored conversations. There are recent approaches that utilize external sources for generating high-quality dialogues with LLMs (Li et al., 2022; Zhou et al., 2022; Kulhánek et al., 2021). Kim et al. (2023) sought to distill conversations from InstructGPT 175B using a commonsense knowledge graph. Compared to existing works, we are the first to utilize dynamic input sources for generating informative responses on a significantly large-scale, which allows us to encompass an exceptionally broad information of user preferences and item information.

3 PEARL Construction

We construct PEARL, a large-scale recommendation dataset covering diverse preferences and detailed item explanations through five steps: (1) grouping real-world reviews with two axes, which are user and item, (§3.1), (2) equipping a user simulator with preferences of a real-world user using reviews (§3.2), (3) infusing item knowledge extracted from reviews to a recommender simulator (§3.3), (4) inferring the simulators to derive a dialogue (§3.4), and (5) filtering dialogues (§3.5). While our dataset construction process is domain-independent, we validate it in the movie domain due to its extensive prior attention (Li et al., 2018; Hayati et al., 2020; Liu et al., 2020; Zhou et al., 2020b), making it easier to evaluate. The overview of PEARL construction process is shown in Figure 2 and an example of PEARL is in Table 2.

3.1 Constructing User-Review and Item-Review Databases

To grant realistic preferences to the user simulator and item knowledge to the recommender simulator, we collect authentic reviews by scraping movie and review data from IMDB website.[3]

For the user simulator, we construct a user-review database where a set of movie title, review text, and rating is grouped by the user who wrote the review. To clarify the preference and remove any noise from the raw review text, we transform the raw text into a high-level review text that focuses on the features the user likes and dislikes by using an LLM. By utilizing a high-level review text instead of the raw text, we can help our user simulator ground on refined text without noise and also filter out personal information in the raw reviews. The prompt is shown in Table 12.

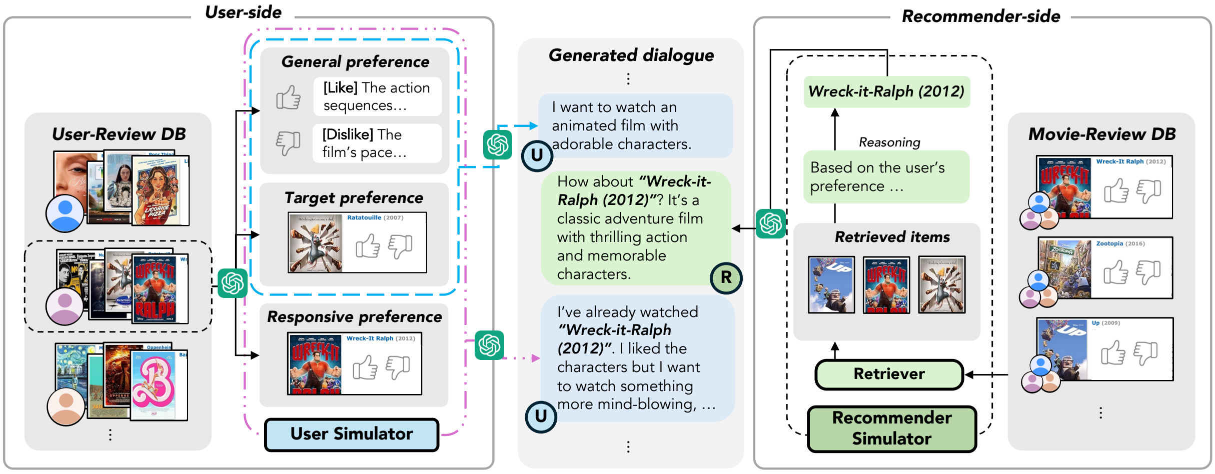
For the recommender simulator which should provide proper recommendations based on rich knowledge about the items, we construct an item-review database where a set of genre, director, cast, and review text is grouped per movie title. By utilizing reviews, we can obtain information about items that cannot be gathered by only utilizing metadata of items. Here, we select up to three most voted reviews on IMDB for efficiency, instead of using all collected reviews of the item. Then, we transform them into a high-level review text, similar to the user-review database construction process. The prompt is shown in Table 13. As a result, the two databases contain 11,839 movies, 68,709 users, and 221,242 reviews in total.

3.2 Persona-augmented User Simulator

Our user simulator uses GPT-3.5 (*i.e.*, `GPT-3.5-turbo-1106`; Ouyang et al. (2022)), though in practice, a different model could be used. We explain how we compose persona and how the user simulator generates an utterance.

---

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



I want to watch an animated film with adorable characters.

How about ***"Wreck-it-Ralph (2012)"***? It's a classic adventure film with thrilling action and memorable characters.

I've already watched ***"Wreck-it-Ralph (2012)"***. I liked the characters but I want to watch something more mind-blowing, …

Figure 2: The overview of PEARL construction method. We synthesize recommendation dialogues with review-driven persona-knowledge grounded simulators. Specifically, our user simulator is equipped with persona and our recommender simulator is augmented by knowledge derived from reviews.

Persona components. We provide the user simulator with persona which includes three types of preferences: general, target, and responsive preferences. For the general preference, we combine three randomly selected review texts of a particular user from the user-review database. This combination represents the user's overall movie taste throughout the dialogue. For the target preference, we select a movie title and its corresponding review text that the user has rated highly (at least 8 out of 10). The user simulator's role is to consistently express the specific attributes of the movie, so that the recommender simulator can eventually suggest the target movie to the user simulator while having a conversation. Lastly, to facilitate the user simulator to provide realistic feedback when a movie is suggested, we obtain the user's review text of the movie from the user-review database as reference.

Utterance generation. The goal of the user simulator is to express its detailed preferences consistent with the persona and eventually get the target item as the final recommendation. Given the task description D_u, persona P, and the dialogue context $C_u^t = (u_1, r_1, u_2, ..., r_t)$ which consists of utterances up to t-th turn where u_i and r_i are utterances from user and recommender respectively, GPT-3.5 generates the next response u_t by following the task description under zero-shot setting. Note that the responsive preference in the persona is dynamically obtained from the user-review database when an item is suggested in the middle of the dialogue.

3.3 Knowledge-augmented Recommender Simulator

To generate utterances of recommender with proper recommendations and explanations, we leverage an LLM and a retriever. The retriever first searches top-k context-relevant items, and the LLM performs a step-by-step reasoning to suggest an item the user would like the most based on the expressed preferences and item knowledge. We use GPT-3.5 as the LLM.

Item retrieval for recommendation. Given the dialogue context $C_r^t = (u_1, r_1, u_2, ..., u_t)$, the retriever in the recommender simulator searches context-relevant items among candidate items. Here, we limit the candidate pool to the items for which the user has written a review, as the user simulator of the next turn should be able to provide detailed feedback. In detail, with `text-embedding-ada-002`, we compute the similarity between the text embedding of the dialogue context and the text embedding of each item knowledge in the candidate pool, which includes the metadata (*e.g.*, title, genre, director, and cast) and the review text of the item. Finally, the top-k texts of the retrieved item knowledge are given to the LLM as input. In order to prevent the dialogue from diverging, we gradually decrease k and forcefully include the target item knowledge in the retrieved pool from a specific turn we set.

Item recommendation with reasoning. For a corpus with proper recommendations and sufficient explanations, we utilize an LLM's reasoning ability to recommend the optimal item based on the dialogue context. Given the task description D_u, the retrieved item knowledge I, and the dialogue context C_r^t, an LLM generates the next recommender response r_t. It first infers the user's preference based on the dialogue context. Then, given the retrieved top-k texts of item knowledge, the LLM selects the most preferable item by the user. Lastly, the LLM generates an utterance that recommends the selected item with a plausible explanation. The prompt is shown in Table 15.

3.4 Dialogue Generation

Figure 2 shows the architecture for the interaction between the user and recommender simulators. We first initialize the dialogue context with an utterance pair (*i.e.*, two dialogue turns), and then each simulator starts interacting based on the given utterances. To simulate the real-world recommendation conversation, we provide different inputs in each turn. On the user simulator side, we provide a persona to generate utterances with specific preferences. On the recommender simulator side, we use the retriever to find suitable items for the dialogue context. Then the LLM reasons about the user's preference and chooses the most proper item for the user and recommends the item along with a detailed explanation based on the item knowledge.

3.5 Dialogue Filtering

Basic filtering. Starting with an initial set of dialogues sampled from GPT-3.5, we remove dialogues (1) where utterances are generated repeatedly, (2) where the user simulator explicitly mentions the target item, or (3) where the user accepts other items instead of the target item. Through these data filtering procedures, the gathered dialogues could be better utilized to investigate the conversational recommendation methods.

Preference-based filtering. We also filter out dialogues based on user preference to maintain dialogue consistency and enhance preference-based recommendations. Specifically, we (1) compute the natural language inference (NLI) score between the persona and user utterances, filtering out the dialogues that include any utterance that contradicts the persona, and (2) extract the recommender simulator's guess of user preference to check for contradictions with user utterances. If there is any contradicted utterance, we consider that the recommender simulator fails to model the user's preference and discard such dialogues.

Seeker: Hi there! I'm in the mood to watch a movie. Can you recommend something?

Recommender: Absolutely! What kind of movie are you in the mood for? Any specific genre or theme?

Seeker: I'm generally a fan of movies with strong performances by the cast and enthralling cinematic accomplishments. I also appreciate well-written plots and screenplays.

Recommender: I think you would really enjoy 'Darkest Hour (2017)'. It has a strong performance by Gary Oldman as Winston Churchill and the attention to detail in bringing the period to life is enthralling. Plus, the focus on Churchill's point of view and how he came to power makes for a compelling and well-written plot.

Seeker: I can see how 'Darkest Hour' fits some of my preferences, but I'm really in the mood for something with more action and combat. I also enjoy movies with historical accuracy and realistic depictions of challenges faced by characters in combat situations. Do you have anything that fits those criteria?

Recommender: How about watching 'Godzilla (1998)'? It has great action sequences involving Godzilla, and the movie stays faithful to the original 1950's movie concept. The special effects and pacing are also very well done, and the film balances action, sci-fi, and thriller elements. Plus, it's based on historical events and provides a realistic depiction of characters facing challenges in combat situations.

Seeker: I appreciate the recommendation, but 'Godzilla (1998)' isn't quite what I'm in the mood for right now. I'm really looking for something with a focus on the life of an Army crew and the challenges they face in combat. Do you have anything else in mind that fits that description?

Recommender: I think 'Fury (2014)' would be a perfect fit for what you're looking for. It provides a realistic depiction of the challenges faced by an Army Sherman tank crew in combat situations. The performances by the cast are strong, and the screenplay is well-written, delivering a compelling and enthralling cinematic experience.

Seeker: That sounds like exactly what I'm looking for! I'll definitely give 'Fury (2014)' a watch. Thanks for the recommendation!

Table 2: Turn-by-turn dialogue sample from PEARL.

Final dataset. By applying a series of dialogue filtering, we obtain PEARL with 57.2K dialogues with more than 4k users and 9k items, where 22.5% of the initial dialogues are removed.

4 Experiments

4.1 Evaluation on Dataset Quality

Datasets. We conduct experiments on ReDial (Li et al., 2018) and INSPIRED (Hayati et al., 2020). ReDial is an English CRS dataset about movie recommendations, and is constructed through crowdsourcing workers on Amazon Mechanical Turk (AMT). Similar to ReDial, INSPIRED is also an English CRS dataset about movie recommendations, but with a smaller size. These two datasets

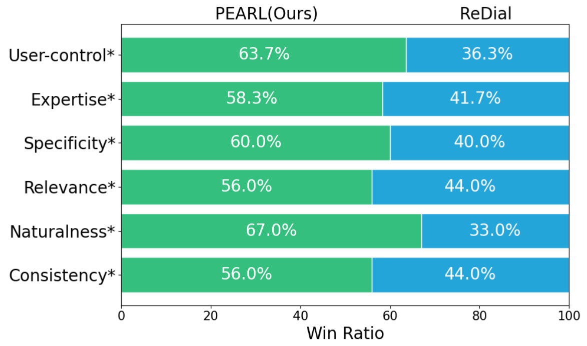


Figure 3: Results of human evaluation on head-to-head comparison between conversations sampled from PEARL and those from ReDial. (*: p-value < 0.05)

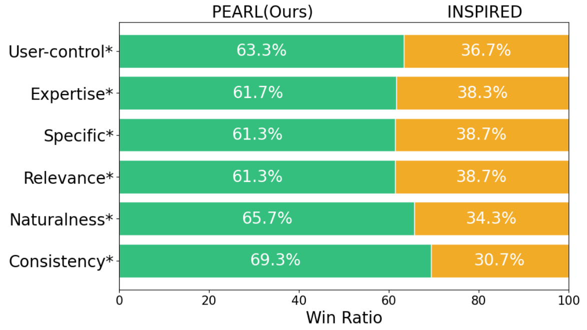


Figure 4: Results of human evaluation on head-to-head comparison between conversations sampled from PEARL and those from INSPIRED. (*: p-value < 0.05)

are widely used for evaluating CRS models.

Human evaluation. To assess the relative quality of PEARL compared to previous datasets, we conduct human evaluation through head-to-head comparison on Amazon Mechanical Turk, comparing PEARL with two widely used open-domain dialogue datasets: ReDial (Li et al., 2018) and INSPIRED (Hayati et al., 2020). We randomly sample 100 dialogues from each dataset and evaluate them according to six criteria: (1) user-control, (2) expertise, (3) specificity of user preference, (4) relevance, (5) flow naturalness, and (6) consistency. Judges are asked to select a better one between two given dialogues, regarding each criterion. Further details are in Appendix B.

Figure 3 and 4 summarizes the head-to-head comparison of PEARL and human-annotated datasets. Despite being fully machine-generated, human raters judge PEARL as better in quality compared to both ReDial and INSPIRED.

User preference analysis. We compare the specificity of user preferences across diverse CRS datasets in Table 3 as the capability to understand specific user preferences is crucial in suggesting personalized and satisfactory recommendations. We measure it by concatenating all user utterances (*i.e.*, preference) of a dialogue and obtaining the number of unique n-grams within it. According to Table 3, PEARL contains more unique expressions than ReDial (Li et al., 2018) or INSPIRED (Hayati et al., 2020), indicating PEARL has less generic and more specific user preferences.

	ReDial	INSPIRED	PEARL
# of dialogues	10,006	1,001	57,277
# of utterances	182,150	35,811	548,061
2-gram specificity	65.44	119.56	141.79
3-gram specificity	65.97	123.01	149.75
4-gram specificity	65.37	122.81	153.00

Table 3: Statistics of PEARL compared to ReDial and INSPIRED. The table shows the number of dialogues, utterances, and n-gram specificities for each dataset.

Inter-dialogue similarity of user utterances	
Ours	0.1900
w/o Persona	0.1962

Table 4: Inter-dialogue similarity of PEARL and ablated dialogues generated without persona.

	ReDial	INSPIRED	PEARL
# of words	11.01	14.62	38.81

Table 5: Average number of words per recommender utterance in ReDial, INSPIRED, and PEARL.

To investigate the effect of utilizing persona in the data synthesis process, we additionally conduct an ablation study on how persona-augmentation yields more diverse and distinct preferences explicitly through user utterances. By comparing the inter-dialogue similarity (*i.e.*, semantic similarity between the concatenated user utterances from two arbitrary dialogues) in Table 4, we observe that synthesizing a dialogue with persona input yields utterances with more distinct preferences that are less similar to each other.

Knowledge-augmented recommendation analysis. We compare the degree of richness in explanations of recommender utterances of ReDial, INSPIRED, and PEARL in Table 5 as it is an important factor for knowledgeable and explainable conversational recommender systems. To quantitatively measure the richness, we calculate number

ReDial: You might like "The Boss Baby (2017)" that was a good movie.
INSPIRED: Have you seen the movie Hustlers yet? It is a little bit of a mix of comedy, drama and thriller.
PEARL: How about "The Addams Family (1991)"? It's a dark comedy with supernatural elements and a great cast, including standout performances from Raul Julia and Christopher Lloyd. Plus, it has a macabre and humorous vibe that I think you'll enjoy.

Table 6: Sample utterances from ReDial, INSPIRED, and PEARL.

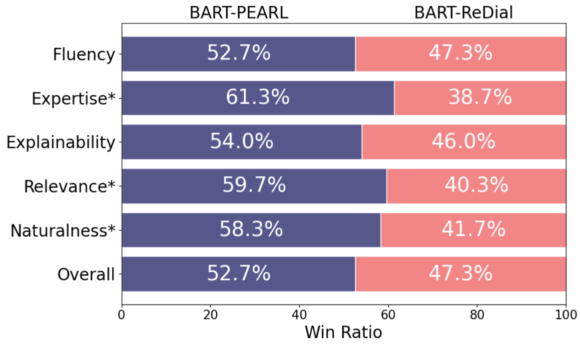


Figure 5: Results of head-to-head comparison human evaluation between responses generated from BART trained on PEARL and on ReDial. (*: p-value < 0.05)

of words used in a single recommender utterance in average. For ReDial which contains several consecutive recommender utterances, we combine them into one utterance as we calculate. Also, we show qualitative examples of each dataset in Table 6. While the utterances of Redial and INSPIRED are relatively shallow and brief, the utterance of PEARL explains about the item in a great detail which may enhance user satisfaction.

Data scale comparison. With 57,277 dialogues, PEARL is the largest in scale compared to existing crowdsourced conversational recommendation datasets (Table 3). It contains more than 500k utterances, each reflecting the preferences of real-world users, thereby providing a rich resource for training conversational recommender systems.

Cost & time-efficiency comparison. Synthesizing PEARL by utilizing the simulators grounded on persona and knowledge is significantly more efficient than traditional dialogue crowdsourcing datasets in both cost and time. For instance, INSPIRED dataset took four months to crowdsource a total of 1,001 dialogues. In contrast, our data generation process with GPT-3.5-turbo-1106 generates more than 57k dialogue datasets in just one week. Furthermore, in terms of cost, the INSPIRED dataset costs an average of $5 per dialogue, whereas the data synthesis process incurs a total cost of about $0.02 per dialogue.

4.2 Evaluation on Dataset Utility

Human evaluation. To qualitatively assess the utility of PEARL, we perform human evaluation that compares the responses of BART (Lewis et al., 2020) trained on ReDial (BART-ReDial) and on PEARL (BART-PEARL) given the same dialogue context from INSPIRED, which is an unseen dataset for the models. We sample 100 dialogue contexts randomly from INSPIRED test set and ask three human judges per each dialogue context to select a better response between the two in terms of six distinct criteria: (1) fluency, (2) expertise, (3) explainability, (4) relevance, (5) naturalness, and (6) overall. Further details are in Appendix B.

Although PEARL is the only machine-generated dataset, Figure 5 shows that BART-PEARL consistently outperforms BART-ReDial across all evaluation criteria. Specifically, BART-PEARL shows the largest gap in terms of *expertise*, highlighting the utility of our approach in enriching recommender responses with a deeper understanding and insight into the movie domain.

Automatic evaluation. We conduct experiments on response generation and recommendation tasks to assess the utility of PEARL, following previous CRS works (Zhou et al., 2020a; Wang et al., 2022). For the response generation task, we employ diverse language models such as BART (Lewis et al., 2020), UniCRS (Wang et al., 2022), and PECRS (Ravaut et al., 2024) to evaluate how effectively PEARL enhances the diversity of the outputs. We evaluate the models using context from an unseen dataset, INSPIRED, and adopt Distinct-n (n=3,4) as the metric. For the recommendation task, BERT (Devlin et al., 2019), UniCRS, and PECRS are employed. We use Recall@k (k=1, 10, 50) as the metric, which indicates the percentage of target items correctly identified within the top-k recommendations. In addition to the downstream models, we also provide performances for zero-shot GPT-3.5 (`GPT-3.5-turbo-1106`) on both tasks to assess the capabilities of large language models in handling

Model	Dist-3	Dist-4
BART-ReDial	0.6220	0.5057
BART-PEARL	**0.9241**	**0.8861**
UniCRS-ReDial	0.5413	0.3667
UniCRS-PEARL	**0.9338**	**0.9007**
PECRS-ReDial	0.6798	0.5906
PECRS-PEARL	**0.9132**	**0.8947**
GPT-3.5	0.9256	0.8910

Table 7: Response generation performances on INSPIRED.

Model	R@1	R@10	R@50
BERT-PEARL	0.0018	0.0208	0.0736
UniCRS-PEARL	**0.0310**	**0.0697**	**0.1202**
PECRS-PEARL	0.0151	0.0339	0.0798
GPT-3.5	0.0071	0.0355	0.0709

Table 8: Recommendation performances on PEARL.

these tasks without task-specific fine-tuning.[4]

Table 7 compares the response generation performances of models trained on PEARL and on ReDial, when evaluated on INPSIRED. The results show the effectiveness of PEARL in improving the diversity of responses generated by downstream models. Moreover, GPT-3.5 shows lower performances than UniCRS-PEARL and PECRS-PEARL, despite GPT-3.5's much larger model size, indicating PEARL is challenging and of high-quality.

Table 8 presents the recommendation performances of models trained and evaluated on PEARL. As the recommendation ability of a trained model highly depends on the distribution of the training data, we do not evaluate with an unseen dataset as done in response generation experiments. We find all models exhibit low performances, indicating that PEARL presents a significant challenge. Interestingly, GPT-3.5 shows a poor performance, suggesting that although PEARL is generated by the same model (*i.e.*, GPT-3.5), there is no inherent advantage for the model. This observation underscores the need for future research to focus on developing models specifically tailored to PEARL, which leverages real-world data rather than relying on parametric knowledge.

4.3 Case Study

Table 9 and 10 present cherry-picked and lemon-picked responses from BART-ReDial and BART-PEARL, respectively. BART-PEARL consistently provides explanations that elucidate what the recommended item is and why the recommender suggests the item, while BART-ReDial offers shorter responses without such details. In the lemon-picked example, BART-PEARL seems to forgot the previous mention of *"The Avengers (2012)"*. However, it still manages to explain the recommended item, demonstrating the model's explainability trained on our synthesized dataset.

Dialogue context

...

Seeker: I accept this recommendation and would love to watch it with my family! Thank you for your help.

Recommender: You're very welcome! I hope you enjoy it!

Seeker: If we finish this movie early, do you have children's Christmas movie to recommend? So far our favorite is Elf.

Model responses

GPT-3.5: I would recommend the movie "The Polar Express" as a great children's Christmas movie. It's a heartwarming and magical adventure that is perfect for the holiday season. Enjoy watching it with your family!

BART-ReDial: Yes, I love Elf! It's a great family movie!

BART-PEARL (Ours): How about watching 'The Lion King'? It's a classic animated film with a heartwarming story and memorable characters. It's perfect for family viewing and has a great message about family and friendship.

Ground-truth: I would absolutely recommend Rise of the Guardians it's a kid's movie, but definitely funny so adults can enjoy too.

Table 9: Cherry-picked sample responses from BART-ReDial and BART-PEARL to a context in INSPIRED. The full example is in Table 16.

5 Conclusion

In this work, we introduce PEARL, a large-scale conversational recommendation dataset constructed using LLM simulators augmented with persona or knowledge from real-world reviews. Our comprehensive experiments validate PEARL's superior quality and utility in training models compared to existing datasets.

PEARL paves the way for future research in developing effective conversational recommender systems, particularly those emphasizing explainability, knowledge retrieval, and reasoning abilities. Also, PEARL offers a practical opportunity for developing small and deployable systems capable of handling specific user feedback and providing satisfactory recommendations.

---

[4] GPT-3.5 is evaluated in zero-shot setting on recommendation and response generation tasks following He et al. (2023).

Dialogue context

...

Seeker: I like both, but I'm looking for something super hero related, like Guardians of the Galaxy.

Recommender: Oh! I have a few ideas of what you have in mind. Are comic book characters interesting at all?

Seeker: Yes! I love comic book characters, like Spider man, Ironman and Captain America.

Recommender: Okay. Would you like to watch this trailer for the Avengers?

Seeker: Oh that would be great!

Model responses

GPT-3.5: I would recommend you watch the trailer for "Avengers: Endgame." It features all your favorite comic book characters like Iron Man, Captain America, and Spider-Man, and it's a super hero-related movie just like Guardians of the Galaxy. Enjoy watching the trailer!

BART-ReDial: Okay. I would also recommend 'The Avengers (2012)'

BART-PEARL (Ours): How about 'The Avengers (2012)'? It's an action-packed superhero movie with a great cast and some really interesting characters.

Ground-truth: I haven't actually watched it yet. It is the one with that Thanos character and the Justice league, right?

Table 10: Lemon-picked sample responses from BART-ReDial and BART-PEARL to a context in INSPIRED. The full example is in Table 17.

Limitations

As we generate recommendation dialogues using LLM-based simulators, the choice of a language model (*i.e.*, GPT-3.5 in this work) will impact the quality of dialogue created. One of the possible future directions may include curating recommendation dialogues by using simulators based on different language models and investigating the difference between generated dialogues and utterances.

Ethical Considerations

Our work utilizes a large language model and real-world reviews for recommendation dialogue generation and filters out generated dialogues in terms of user preference. As our filtering mechanism does not address considerations related to dialogue safety, the users who employ our data generation process should be mindful of this limitation and consider the incorporation of additional filtering steps to mitigate potential biases or toxic content. We ensure that workers hired through Amazon Mechanical Turk receive fair compensation. We offer an effective hourly rate exceeding $15, based on the estimated time required to complete the tasks.

Acknowledgements

This work was supported by the IITP grant funded by the Korea government (MSIT) (No. RS-2020-II201361) and the NRF grant funded by the Korea government (MSIT) (No. RS-2023-00244689). We thank Hyunseo Kim and Soyeon Chun for their participation in experiments

References

Hyungjoo Chae, Yongho Song, Kai Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5606–5632.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shuyu Guo, Shuo Zhang, Weiwei Sun, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2023. Towards explainable conversational recommender systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2786–2795.

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

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.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and

Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Minju Kim, Chaehyeong Kim, Yong Ho Song, Seungwon Hwang, and Jinyoung Yeo. 2022. Botstalk: Machine-sourced framework for automatic curation of large-scale multi-skill dialogue datasets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5149–5170.

Jonáš Kulhánek, Vojtěch Hudeček, Tomáš Nekvinda, and Ondřej Dušek. 2021. AuGPT: Auxiliary tasks and data augmentation for end-to-end dialogue with pre-trained language models. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 198–210, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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.

Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. 2022. Controllable dialogue simulation with in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4330–4347, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. Durecdial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4335–4347.

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, Zichen Ma, Xiaoguang Han, Youzheng Wu, Shuguang Cui, and Xiaodong He. 2023. August: an automatic generation understudy for synthesizing conversational recommendation datasets. *arXiv preprint arXiv:2306.09631*.

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.

Mathieu Ravaut, Hao Zhang, Lu Xu, Aixin Sun, and Yong Liu. 2024. Parameter-efficient conversational recommender system as a language processing task. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 152–165, St. Julian's, Malta. Association for Computational Linguistics.

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

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

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.

Pei Zhou, Hyundong Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2022. Reflect, not reflex: Inference-based common ground improves dialogue response quality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10468, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Implementation Details

A.1 Large language model

In this work, we utilize GPT-3.5 (GPT-3.5-turbo-1106) for our user simulator and recommender simulator. GPT-3.5 is an LLM based on InstructGPT (Ouyang et al., 2022).[5] The prompts used to generate persona, item knowledge, seeker's utterance and recommender's response are in Table 12, Table 13, Table 14, and Table 15, respectively.

A.2 Computational Resources and API Cost

We run the BART and NLI models on eight NVIDIA RTX A5000 GPUs. For ChatGPT API usage, we use about 0.02 dollars per one dialogue generation.

A.3 Natural language inference model

We leverage an external natural language inference (NLI) model in the dialogue filtering process to obtain higher-quality recommendation conversation data. In particular, in preference-based filtering, we filter it out if the NLI model predicts the logical relationship between persona and user's utterance is *contradiction* with $\delta > 0.7$. As for the NLI model, we use RoBERTa (Liu et al., 2019) fine-tuned on the DNLI dataset (Welleck et al., 2019).

A.4 Downstream models

We utilize source codes from `https://github.com/RUCAIBox/UniCRS` for UniCRS and `https://github.com/Ravoxsg/efficient_unified_crs` for PECRS experiments. For BERT, we take a naive method to predict the most plausible movie based on the dialogue context. For BART, we also take a general setting of sequence-to-sequence modeling to generate the next response when given the dialogue context.

B Human Evaluation Metrics

B.1 Dataset Quality

We outsource a human evaluation comparing our PEARL and previous datasets via Amazon Mechanical Turk (AMT). We show the interface for the evaluation in Figure 7. We ask the human judges to compare the dialogues based on the following criteria:

- **User-control**: Which seeker participates more actively and provides feedback to receive recommendations in dialogues?
- **Expertise**: Which recommender seems to be more of an expert in the movie domain?
- **Specificity**: Which seeker is better at expressing preferences that help the recommender suggest a personalized movie?
- **Relevance**: Which dialogue provides more relevant recommendations according to the seeker's preferences?
- **Flow Naturalness**: Which one is more natural in the progression of the dialogue?
- **Consistency**: Which dialogue is more consistent in terms of the seeker's preferences?

During each stage of voting, human judges are given two dialogue candidates and asked to judge which one is of a higher quality based on the above criteria.

B.2 Dataset Utility

We also outsource a human evaluation comparing the responses of the BART model trained on PEARL and ReDial dataset. We show the interface for this evaluation in Figure 8.

We ask the human judges to compare the responses and select the better one based on the following criteria:

- **Fluency**: Which response is more fluent?
- **Expertise**: Which response seems to have more expertise in the movie domain?
- **Explainability**: Which response offers more sufficient explanations with the recommendations?
- **Relevance**: Which response provides more relevant recommendation according to the seeker's preference?
- **Naturalness**: Which response is more natural?
- **Overall**: Which response do you prefer overall?

---

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

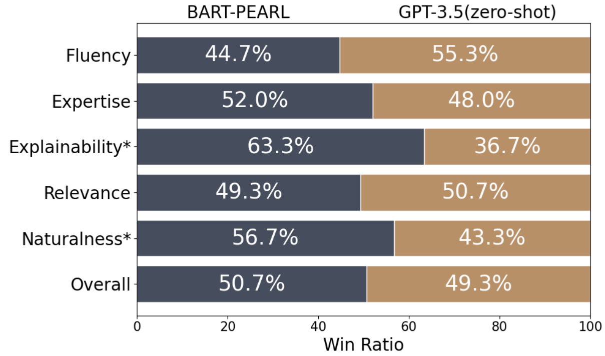


Figure 6: Results of head-to-head comparison human evaluation between responses generated from BART trained on PEARL and from GPT-3.5 zero-shot inference. (*: p-value < 0.05)

Model	ROUGE-1	ROUGE-2	Dist-1	Dist-2	Dist-3	Dist-4
BART-ReDial	0.1370	**0.0264**	**0.9826**	0.7596	0.6208	0.4964
BART-PEARL	**0.1642**	0.0241	0.8803	**0.9594**	**0.9365**	**0.9047**
UniCRS-ReDial	0.0925	0.0097	**0.9545**	0.7916	0.6288	0.4635
UniCRS-PEARL	**0.2153**	**0.0218**	0.7551	**0.8997**	**0.9352**	**0.9027**
PECRS-ReDial	0.1979	**0.0573**	**0.7995**	0.7649	0.6801	0.6039
PECRS-PEARL	**0.2619**	0.0497	0.7637	**0.9058**	**0.9182**	**0.9074**
GPT-3.5	0.2256	0.0330	0.8990	0.9620	0.9354	0.9046

Table 11: Response generation performances on E-ReDial (Guo et al., 2023) *Test-Rec* subset.

C Additional Experimental Results

Dataset utility for practical use Our research aims to tailor a model specific to the conversational recommendation task, by utilizing our proposed dataset. While a large language model (*e.g.*GPT-3.5) may seem plausible as a conversational recommender due to its capability to generate human-like text, PEARL and its downstream models are still necessary. First of all, a large language model (*e.g.*GPT-3.5) is practically inappropriate to be deployed in real-time services due to its latency, cost, and bias issues. Also, while constructing the dialogues in PEARL, we have taken into account the relationship between the actual user history and the the target recommendation. We believe that large language models lack in such collaborative interaction knowledge, as shown in Table 8 where it achieves only 0.0355 in Recall@10 in our test set.

In addition, we ran a human evaluation between responses of GPT-3.5 and BART trained on PEARL(BART-PEARL), given INSPIRED contexts. We randomly sampled 50 responses from each model and used the same criteria as dataset utility human evaluation in the paper. As shown in Figure 6, responses from BART-PEARL are generally preferred to those from GPT-3.5. Although BART shows inferior fluency and decent relevance, it outperforms GPT-3.5 in expertise and explainability, the two most important aspects emphasized throughout our paper. We believe that augmenting responses with domain-specific knowledge in PEARL helped the downstream model enrich explanations beyond the basic information of the movie.

Explainability of downstream models Table 11 shows experimental results for downstream models (BART, UniCRS, and PECRS) evaluated on E-ReDial (Guo et al., 2023). E-ReDial consists of high-quality explanations manually annotated by human workers, based on ReDial dialogues. To asses the models' ability in providing explanations for recommendations, we utilize the *Test-Rec* subset of the test set, which always contains recommendations and explanations as suggested by the authors.

According to Table 11, models trained on PEARL outperforms those trained on ReDial on most metrics. For the reference-based metrics (ROUGE-1, ROUGE-2) where they underperform, we speculate that models trained on ReDial benefit from sharing the recommendation distribution with E-ReDial.

Prompt

Given a review written by you, summarize what you liked and disliked about the movie, under [Like] and [Dislike] respectively. If there is nothing to mention about like/dislike, simply write "None." under the corresponding tag.

Here is the movie review written by you:
{review}

Table 12: The prompt for summarizing a review of a user.

Prompt

Given some popular reviews about {movie title}, describe what people liked and disliked about the movie, under [Like] and [Dislike] respectively. If there is nothing to mention about like/dislike, simply write "None." under the corresponding tag.

Here are some basic information about the movie and reviews about it:
Movie Title: {movie title}
Genre: {genre}
Director: {director}
Cast: {cast}
Reviews:
{reviews}

Table 13: The prompt for summarizing some popular reviews of a movie.

Prompt

You are a Seeker who interacts with a Recommender to get a movie recommendation that suits your preferences well. You will be given a dialogue context, and you must follow the instructions below to interact with the Recommender:

- The recommender may ask for your preference or recommend a movie to you.
- In the beginning, express your general preference on movies based on your past reviews about movies you have watched before.
- If you are recommended a movie which is not {gt movie title}, you should reject it with a reason based on your thought about the recommended movie. Also, express some common features of the movies you have watched before and you should be recommended (DO NOT explicitly mention the movie title!).
- If you are recommended {gt movie title}, you should accept it as if you haven't watched it before, and end the conversation by generating [EOD] token.
- Continue the conversation for at least six turns.

Here are your reviews about movies you have watched before:
{user persona}

Some features of the movie you should be recommended:
{gt abstract}

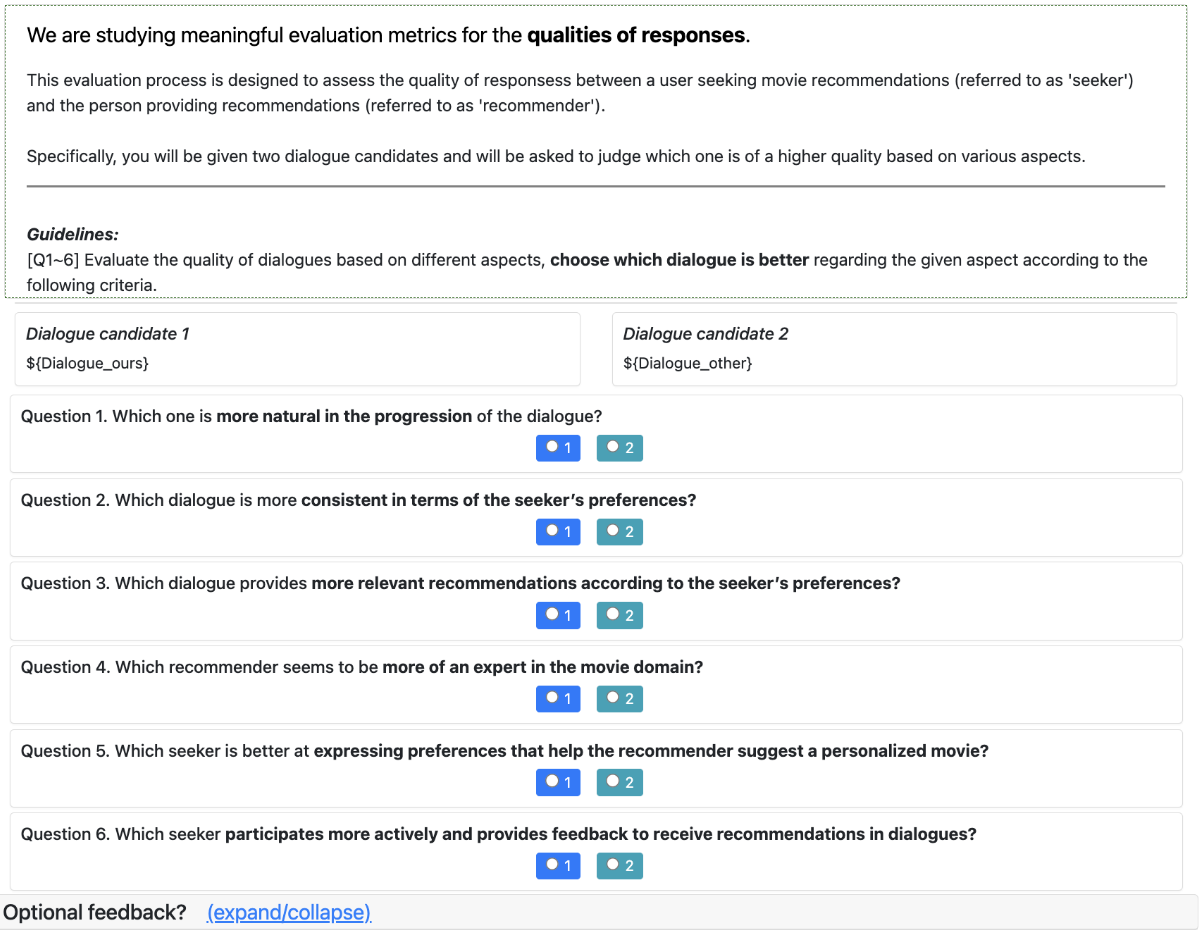
{rec movie abstract}

Here is the dialogue context:
{dialogue context}

Table 14: The prompt for Seeker simulator.

Prompt
You are a Recommender who recommends a Seeker a movie that he/she will enjoy, among the three candidates and based on the dialogue context given. Follow the instructions below to complete the task: - In the beginning of the conversation, engage with the Seeker to discover his/her movie preferences (regardless of the given three candidates). Follow this format: Think: (think about which question to ask the seeker) Recommender: (utterance that asks for the seeker's movie preference) - After some interactions, choose and suggest a movie from the three candidates and make the recommendation in the specified format: Think: (think about the seeker's movie preference based on the context) Movie: (Movie title (Year)) Recommender: (response to the seeker's utterance) - Do not recommend a movie that has been mentioned before in the dialogue context. Here are the three movie candidates: {k movies info} Here is the dialogue context: {dialogue context} Think:

Table 15: The prompt for Recommender simulator.



We are studying meaningful evaluation metrics for the **qualities of responses**.

This evaluation process is designed to assess the quality of responsess between a user seeking movie recommendations (referred to as 'seeker') and the person providing recommendations (referred to as 'recommender').

Specifically, you will be given two dialogue candidates and will be asked to judge which one is of a higher quality based on various aspects.

Guidelines:
[Q1~6] Evaluate the quality of dialogues based on different aspects, **choose which dialogue is better** regarding the given aspect according to the following criteria.

Dialogue candidate 1
${Dialogue_ours}

Dialogue candidate 2
${Dialogue_other}

Question 1. Which one is **more natural in the progression** of the dialogue?
○ 1 ○ 2

Question 2. Which dialogue is more **consistent in terms of the seeker's preferences?**
○ 1 ○ 2

Question 3. Which dialogue provides **more relevant recommendations according to the seeker's preferences?**
○ 1 ○ 2

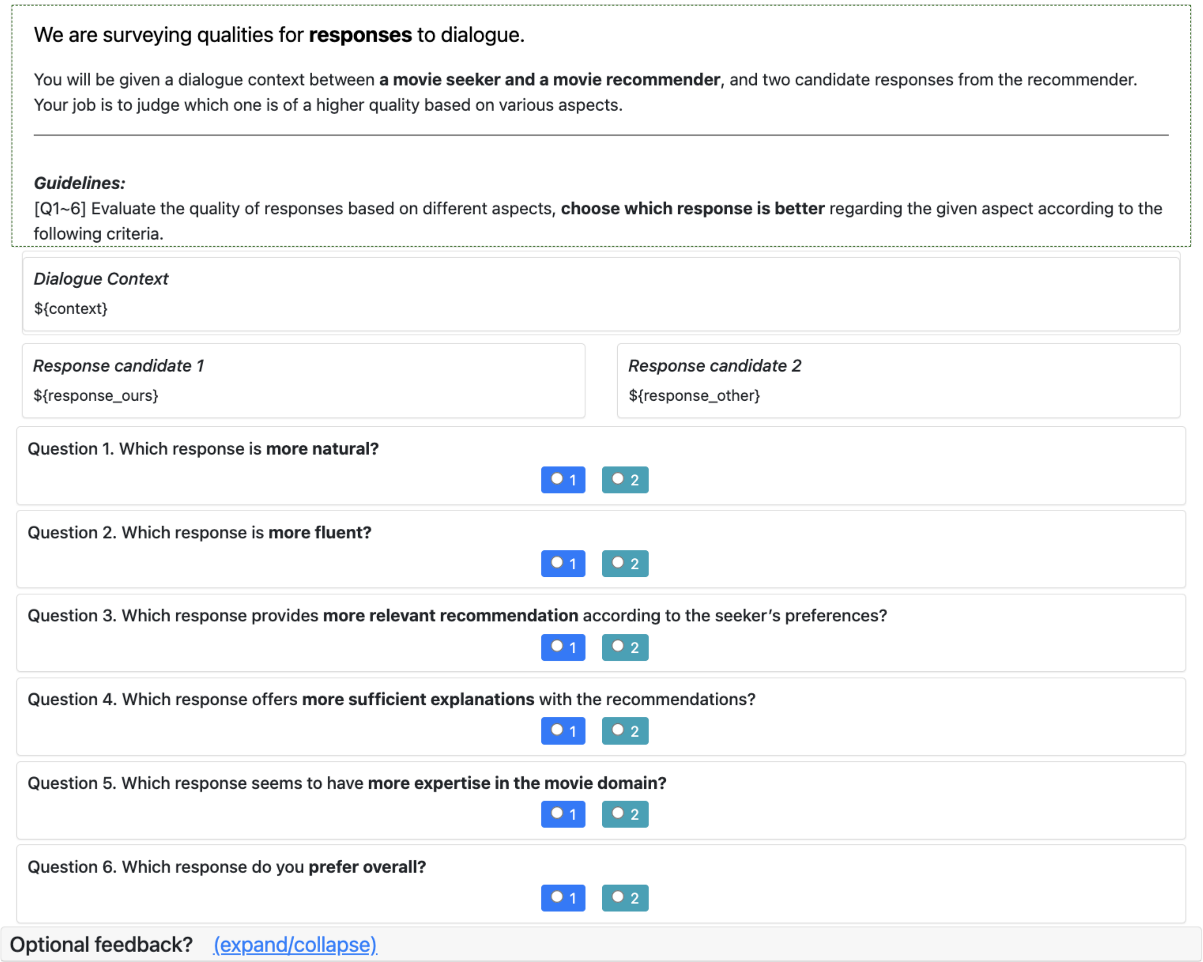
Question 4. Which recommender seems to be **more of an expert in the movie domain?**
○ 1 ○ 2

Question 5. Which seeker is better at **expressing preferences that help the recommender suggest a personalized movie?**
○ 1 ○ 2

Question 6. Which seeker **participates more actively and provides feedback to receive recommendations in dialogues?**
○ 1 ○ 2

Optional feedback? (expand/collapse)

Figure 7: Interface for human evaluation on dataset quality.



We are surveying qualities for **responses** to dialogue.

You will be given a dialogue context between **a movie seeker and a movie recommender**, and two candidate responses from the recommender. Your job is to judge which one is of a higher quality based on various aspects.

Guidelines:
[Q1~6] Evaluate the quality of responses based on different aspects, **choose which response is better** regarding the given aspect according to the following criteria.

Dialogue Context
${context}

Response candidate 1
${response_ours}

Response candidate 2
${response_other}

Question 1. Which response is **more natural?**
○ 1 ○ 2

Question 2. Which response is **more fluent?**
○ 1 ○ 2

Question 3. Which response provides **more relevant recommendation** according to the seeker's preferences?
○ 1 ○ 2

Question 4. Which response offers **more sufficient explanations** with the recommendations?
○ 1 ○ 2

Question 5. Which response seems to have **more expertise in the movie domain?**
○ 1 ○ 2

Question 6. Which response do you **prefer overall?**
○ 1 ○ 2

Optional feedback? (expand/collapse)

Figure 8: Interface for human evaluation on dialogue utility.

Dialogue context

Recommender: Hey there

Seeker: Hi! I'm looking for recommendations on a great holiday movie, do you have any recommendations?

Recommender: For sure! First, do you have certain preferences or aversions? Like are you into musicals?

Seeker: I do not really like musicals.

Recommender: Good to know... I would have recommended White Christmas, it's a classic and a bit of a tradition in our family, but there are so many to choose from! The old Tim Allen Santa Clause movie is super fun and so is Home Alone.

Seeker: Is the Tim Allen Santa Clause movie funny? I would love to watch a funny movie.

Recommender: Yes it is! It's a heart warming family comedy :)

Seeker: I accept this recommendation and would love to watch it with my family! Thank you for your help.

Recommender: You're very welcome! I hope you enjoy it!

Seeker: If we finish this movie early, do you have children's Christmas movie to recommend? So far our favorite is Elf.

Model responses

GPT-3.5: I would recommend the movie "The Polar Express" as a great children's Christmas movie. It's a heartwarming and magical adventure that is perfect for the holiday season. Enjoy watching it with your family!

BART-ReDial: Yes, I love Elf! It's a great family movie!

BART-PEARL (Ours): How about watching 'The Lion King'? It's a classic animated film with a heartwarming story and memorable characters. It's perfect for family viewing and has a great message about family and friendship.

Ground-truth: I would absolutely recommend Rise of the Guardians it's a kid's movie, but definitely funny so adults can enjoy too.

Table 16: A Cherry-picked dialogue example from PEARL dataset. This shows the difference in response generation between BART-ReDial and BART-PEARL. This case illustrates that BART-PEARL offers greater informativeness and explainability for the movie.

Dialogue context

Recommender: Hello

Seeker: hello

Recommender: What can I help you with

Seeker: Can you help recommend a movie trailer to watch?

Recommender: I would recommend horror movies, but I would like to know what movies you like first

Seeker: Oh, I love action movies.

Recommender: The types with very serious actors (Stallone, etc) or, like Guardians of the Galaxy?

Seeker: I like both, but I'm looking for something super hero related, like Guardians of the Galaxy.

Recommender: Oh! I have a few ideas of what you have in mind. Are comic book characters interesting at all?

Seeker: Yes! I love comic book characters, like Spider man, Ironman and Captain America.

Recommender: Okay. Would you like to watch this trailer for the Avengers?

Seeker: Oh that would be great!

Model responses

GPT-3.5: I would recommend you watch the trailer for "Avengers: Endgame." It features all your favorite comic book characters like Iron Man, Captain America, and Spider-Man, and it's a super hero-related movie just like Guardians of the Galaxy. Enjoy watching the trailer!

BART-ReDial: Okay. I would also recommend 'The Avengers (2012)'

BART-PEARL (Ours): How about 'The Avengers (2012)'? It's an action-packed superhero movie with a great cast and some really interesting characters.

Ground-truth: I haven't actually watched it yet. It is the one with that Thanos character and the Justice league, right?

Table 17: Lemon-picked dialogue example from PEARL dataset. This shows that BART-PEARL recommends Avengers again which is recommended already in the previous dialog.