

# Towards Explainable Conversational Recommender Systems

Shuyu Guo  
Shandong University  
Qingdao, China  
guoshuyu225@gmail.com

Shuo Zhang  
Bloomberg  
London, United Kingdom  
szhang611@bloomberg.net

Weiwei Sun  
Shandong University  
Qingdao, China  
sunweiwei@gmail.com

Pengjie Ren  
Shandong University  
Qingdao, China  
renpengjie@sdu.edu.cn

Zhumin Chen  
Shandong University  
Qingdao, China  
chenzhumin@sdu.edu.cn

Zhaochun Ren\*  
Shandong University  
Qingdao, China  
zhaochun.ren@sdu.edu.cn

## ABSTRACT

Explanations in conventional recommender systems have demonstrated benefits in helping the user understand the rationality of the recommendations and improving the system's efficiency, transparency, and trustworthiness. In the conversational environment, multiple contextualized explanations need to be generated, which poses further challenges for explanations. To better measure explainability in conversational recommender systems (CRS), we propose ten evaluation perspectives based on concepts from conventional recommender systems together with the characteristics of CRS. We assess five existing CRS benchmark datasets using these metrics and observe the necessity of improving the explanation quality of CRS. To achieve this, we conduct manual and automatic approaches to extend these dialogues and construct a new CRS dataset, namely Explainable Recommendation Dialogues (E-ReDial). It includes 756 dialogues with over 2,000 high-quality rewritten explanations. We compare two baseline approaches to perform explanation generation based on E-ReDial. Experimental results suggest that models trained on E-ReDial can significantly improve explainability while introducing knowledge into the models can further improve the performance. GPT-3 in the in-context learning setting can generate more realistic and diverse movie descriptions. In contrast, T5 training on E-ReDial can better generate clear reasons for recommendations based on user preferences. E-ReDial is available at <https://github.com/Superbooming/E-ReDial>.

## CCS CONCEPTS

• **Information systems** → Users and interactive retrieval; Recommender systems; • **Human-centered computing** → Human computer interaction (HCI).

\*Corresponding Author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9408-6/23/07...\$15.00

<https://doi.org/10.1145/3539618.3591884>

## KEYWORDS

Explainable recommendation, conversational recommendation, conversational information access

### ACM Reference Format:

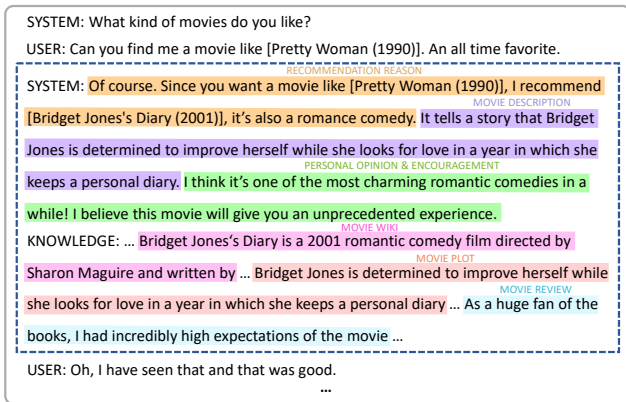
Shuyu Guo, Shuo Zhang, Weiwei Sun, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2023. Towards Explainable Conversational Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*, July 23–27, 2023, Taipei, Taiwan. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3539618.3591884>

## 1 INTRODUCTION

Recommender systems provide personalized suggestions to help users find items based on their preferences and have been widely used in various online applications. Explanations for recommender systems are expected to clarify why such items are recommended. Researchers have pointed out that appropriate explanations can help improve the recommender systems' transparency, persuasiveness, effectiveness, trustworthiness, and user satisfaction [42]. Thus, a growing body of work has been devoted to improving the recommendation explainability through various methods, e.g., incorporating user reviews [6, 22], counterfactual reasoning [37, 40].

Conventional recommender systems primarily predict a user's preference over an item by analyzing their past behaviours, which can neither clarify what a user likes nor explain why a user likes an item. In contrast, conversational recommender systems (CRS) elicit user preferences dynamically and respond to users' needs through real-time multi-turn interactions [15]. Compared to single-turn explanation generation in conventional recommender systems, CRS needs to generate multiple contextualized explanations, which poses further challenges. As of now, there is limited progress for explainable CRS [19]. Wen et al. [41] firstly try to generate explanations on real-world CRS datasets. However, due to the lack of evaluation methods for CRS explanations and datasets with high-quality explanations, the performance of the generated explanations is limited and not effectively evaluated. In this study, we focus on explainable conversation recommender systems and conduct further investigations.

We elucidate the meaning of the explanation for CRS, combing explainable recommender systems [39] and the characteristics of CRS. Formally, it refers to a response from CRS containing relevant details about the recommended items, e.g., the recommendation reasons, the items' descriptions and personal opinions when recommending items. To evaluate the explanation for CRS, we



**Figure 1: An example snippet of dialog in E-ReDial. Above each system’s response with explanation and knowledge is different components. Best seen in colors.**

propose ten evaluation perspectives. The effectiveness, efficiency, persuasiveness, satisfaction, scrutability, transparency, trust, and representativeness are inherited from explainable recommender systems [1, 8]. They are used to evaluate properties common to all recommendation systems, e.g., system transparency, recommendation accuracy, and user satisfaction. In addition, the reasonability and coherence are derived from the characteristics of the CRS. They measure the consistency between explanations and conversations regarding logic and language, respectively.

We measure five widely used CRS datasets using a human annotation task to verify the necessity of explainability in CRS. We design a questionnaire containing 13 questions covering 10 evaluation perspectives we proposed. With 50 dialogues sampled from each dataset, over 20 participants are invited to answer these questions at the dialogue exchange level. According to the questionnaire results, the existing datasets have relatively low-quality explanations. Specifically, lack of explanation, ambiguous recommended reason, unrepresentative item description, low effectiveness, low efficiency, low user satisfaction, trust or willingness to accept the recommendation in most dialogue turns. Meanwhile, we label all sampled dialogues at the dialogue level and compute the correlation between dialogues and explanations. This task covers effectiveness, efficiency, dialogue quality, and satisfaction [19]. The results indicate that the explanation quality is highly correlated with the overall performance of the CRS. Specifically, improving any perspectives of explanation is beneficial to the effectiveness and efficiency of CRS. Dialogues with more diverse movie descriptions will be considered to have higher dialogue quality, while more trustworthy explanations lead to higher satisfaction.

To improve the explainability of CRS, we construct a new dataset named **Explainable Recommendation Dialogues (E-ReDial)**. We conduct a user study to investigate high-quality explanations for CRS and extract four main characteristics, i.e., clear recommendation reason, representative item description, encouragement or personal opinion, and reasonable and contextual. Based on these, we conduct manual and automatic approaches to rewrite the system responses with low-quality explanations sampled from a commonly

used real-world CRS dataset. In the manual method, over 30 participants are involved. We ask them to search for relevant information about movies mentioned in responses (or knowledge for short), e.g., the plots, the trailers, the wikis and the reviews and extend responses to meet all requirements. In the automatic method, we use GPT-3 for rewriting. We input the context with a prompt covering all requirements and collect the generated responses. We observe that the automatic method often fails to meet the rewriting requirements by evaluating the explanations obtained by both methods. Thus E-ReDial is constructed entirely by manual method. It consists of 756 dialogues in the movie domain with 2,058 high-quality rewriting explanations, where each explanation is additionally annotated for corresponding knowledge and different components. Figure 1 presents an illustrative example for our E-ReDial dataset.

We compare training-based and prompt-based approaches to perform explanation generation, i.e., given the context and the ground-truth recommended movies to generate corresponding explanations. Experimental results suggest that models trained on E-ReDial can significantly improve explainability, while introducing knowledge into the models can further improve the performance. In addition, both training-based and prompt-based methods have their own advantages. While GPT-3 can generate more realistic and diverse movie descriptions in an in-context learning setting. T5 training on E-ReDial can better generate clear reasons for recommendations based on user preferences.

In summary, this paper makes the following contributions: (1) We propose the definition and evaluation perspectives of the explanation for CRS and verify its necessity. (2) We collect and share a CRS dataset, E-ReDial, which includes over 2,000 high-quality explanations. (3) We introduce both training-based and prompt-based baseline methods for explanation generation using E-ReDial. The data and code are available at <https://github.com/Superbooming/E-ReDial>.

## 2 RELATED WORK

**Conversational recommendation.** Unlike conventional recommendations, conversational recommenders interact with users through real-time, multi-turn conversations [15]. Two main types of conversational recommender systems (CRS) are being studied: attribute-aware and topic-guided [19]. The former focus on the recommendation strategy [26, 47], while the latter interacts with users through free-form natural language [4, 47]. We focus on topic-guided CRS in this paper. We analyze them from the view of explainability and enhance their corresponding abilities through manual and automatic methods.

**Explainable recommendation.** The explainable recommendation provides the items and explanations for why such items are recommended. Providing explanations for the recommendations can help improve the recommender systems’ transparency, persuasiveness, effectiveness, trustworthiness, and user satisfaction [42]. There are various forms of explanations, such as item features [7, 43], natural language [17, 21], images [5], and prototypes [28]. Since CRS interacts with users through multi-turn conversations using natural language, we focus on the natural language explanations in this paper. To evaluate the explainability of recommender system,

Tintarev and Masthoff [39] have proposed 7 evaluation goals: *transparency*, *scrutability*, *trust*, *effectiveness*, *persuasiveness*, *efficiency* and *satisfaction*. Chen et al. [8] extend on these to propose a new goal, *scrutability*, which has the same name mentioned above but has a different meaning. For the sake of distinction, we rename it as *representativeness* in the remainder of this paper. We integrate these goals into the CRS in Section 3.2.

**Explainable conversational recommendation.** There are only a few papers so far study explainable conversational recommendations. The existing works in CRS aim to make the generated responses more natural and fluent instead of more explanatory [25, 33, 45]. Compared to single-turn explanation generation in conventional recommender systems, CRS generates multiple contextualized explanations, which poses further challenges for explanations. Chen et al. [9] is the first work on this topic. Still, it constructs dialogues using template-based user feedback to improve the performance of single-turn explanation generation rather than in a CRS scenario. Wen et al. [41] firstly propose a framework to generate explanations on a real-world CRS dataset. However, the performance of the generated explanations is limited and not effectively evaluated due to the lack of evaluation methods for CRS explanations and datasets with high-quality explanations. To alleviate these issues, we propose 10 evaluation perspectives and collect a new CRS dataset with high-quality explanations in this work.

### 3 THE EXPLANATION FOR CRS

In this section, we elucidate the meaning of the explanation for CRS, evaluation measurements and detail how to validate the necessity of explainability for CRS.

#### 3.1 Definition

The definition of explanation in conventional recommender systems has been widely accepted as a justification for why items have been recommended [42], or an item description helping the user better understand the qualities of the item at a broader level [39]. CRS differs from conventional recommendation systems in many aspects. The major difference is that CRS dynamically interacts with users through real-time multi-turn conversations to elicit their preferences [15]; thus, the scope of queries is enlarged. Combining these factors, we deem the explanation for CRS as:

**Explanation for CRS** *A response from CRS that contains relevant details about the recommended items in the recommended round.*

We expect the explanations only appear when CRSs are recommending items, i.e., responses in subsequent conversation rounds with details about the item can only be considered as an addition to its information and not as an explanation for the recommendation. In addition, an explanation must contain relevant details of recommended items, e.g., the recommendation reasons, the item descriptions, and personal opinions.

#### 3.2 Evaluation perspectives

Explanations for recommendation can vary in purposes, e.g., providing users with recommendations and reasons to increase their trust in the system or providing detailed item descriptions to help

users make a more informed decision. Thus, evaluating the explanations are expected to be comprehensive, i.e., covering different perspectives. We propose the following evaluation perspectives to better evaluate the explanation quality for CRS and list their definitions and evaluation methods in Table 1.

Among these perspectives, the *effectiveness*, *efficiency*, *persuasiveness*, *satisfaction*, *scrutability*, *transparency*, *trust*, and *representativeness* are inherited from explainable recommender systems. They are used to evaluate properties common to all recommendation systems, e.g., system transparency, recommendation accuracy, and user satisfaction. In addition, the *reasonability* and *coherence* is derived from the characteristics of the CRS. They measure the consistency between explanations and conversations regarding logic and language, respectively.

#### 3.3 Necessity

To validate the necessity of explainability for CRS, we look into five widely used CRS datasets as representatives and check their explanation quality at exchange and dialogue levels, respectively.

**Datasets.** The five CRS datasets are as follows. (1) *ReDial* [23] is an English conversational movie recommendation dataset in realistic dialogue scenarios that contains 10,006 conversations. (2) *TG-ReDial* [46] is a Chinese conversational movie recommendation dataset incorporating topic threads annotated by semi-automatic approaches. It contains 10,000 conversations. (3) *DuRecDial* [24] is a Chinese conversational recommendation dataset with multi-type dialogues which contains 10,190 conversations. (4) *INSPIRED* [18] is an English conversational movie recommendation dataset with good social strategies which contains 1,001 conversations. (5) *OpenDialKG* [29] is an English conversational recommendation dataset pairing each dialogue with corresponding knowledge graph paths, which contains 15,673 conversations. We sample 50 dialogues randomly from each dataset for validation.

**Explanation quality at exchange level.** We design a questionnaire<sup>1</sup> containing 13 questions covering all evaluation perspectives in Table 1. Over 20 participants are invited to assess system responses from the sampled dialogues resorting to these evaluation metrics at the dialogue exchange level. For more details on annotation, see our repository [https://github.com/Superbooming/E-Redial/blob/main/Exchange\\_level.md](https://github.com/Superbooming/E-Redial/blob/main/Exchange_level.md). The annotation results are shown in Table 2. The metrics are obtained by averaging the corresponding perspectives' scores across all exchange turns. We also compute Cohen's Kappa to measure the annotation disagreement. Since there are few explanations in *OpenDialKG*, its annotation results are highly agreed upon, so we omit Cohen's Kappa. Our results indicate that the existing datasets have relatively low-quality explanations for our proposed metrics. Specifically, lack of explanation, ambiguous recommended reason, unrepresentative item description, low effectiveness, low efficiency, low user satisfaction, trust, or willingness to accept the recommendation in most dialogue turns. The results differ across datasets, with the two Chinese datasets performing the best, *INSPIRED* and *OpenDialKG* the next best, and *ReDial* the worst. The results also differ across metrics, *reasonability* and *coherence* are relatively high, while others are pretty low.

<sup>1</sup>See [https://github.com/Superbooming/E-Redial/blob/main/Exchange\\_level.md](https://github.com/Superbooming/E-Redial/blob/main/Exchange_level.md)

**Table 1: Evaluation perspectives used in CRS. D denotes definition and M is for Measure method.**

Metric	Definition and Measure Method	Reference
<i>Effectiveness</i>	D: whether the explanation can help users to make good decisions [39] M: difference between the ratings of the item before and after consuming	[2, 3, 34, 38, 39]
<i>Efficiency</i>	D: whether the explanation can help users to make decisions faster [39] M: time spent making decisions [16], reading an explanation [3], the number of dialogue turns	[3, 16, 39]
<i>Persuasiveness</i>	D: whether the explanation can convince users to accept the recommendation [39] M: accept rate of recommendation in conversations	[11, 39]
<i>Transparency</i>	D: whether the explanation explains how the system works [39] M: user’s perceived understanding of how the system works	[12, 39]
<i>Scrutability</i>	D: whether the explanation allows users to know their preference for the item [39] M: user’s perceived understanding of how their preference is used to make recommendations	[13, 39]
<i>Trust</i>	D: whether the explanation increases users’ confidence in the system [39] M: related questions on transparency and scrutability	[30, 39]
<i>Satisfaction</i>	D: whether the explanation can make the use of the system enjoyable [39] M: ask users whether they prefer a system with explanations [36] or whether they will choose the system again for the next	[14, 27, 35, 36, 39]
<i>Representativeness</i>	D: whether the explanation can exactly correspond to the items [8] M: question-based	[8]
<i>Reasonability</i>	D: whether the explanation is logically correct M: question-based	-
<i>Coherence</i>	D: whether the explanation is contextual M: question-based	-

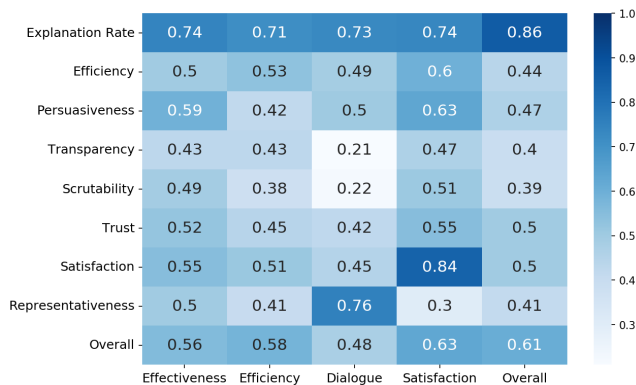
**Table 2: Annotation results resorting to evaluation perspectives we proposed in different CRS datasets. The Explainable Rate ranges from 0 to 1, while the Overall metric ranges from 1 to 5. The other perspective metrics exhibit values between 1 to 3. Higher scores indicate better performance.**

Datasets	ReDial		TG-ReDial		DuRecDial		INSPIRED		OpenDialKG		E-ReDial (Ours)	
	Avg	Kappa	Avg	Kappa	Avg	Kappa	Avg	Kappa	Avg	Kappa	Avg	Kappa
Explainable Rate	0.32	0.97	0.79	0.98	0.64	0.97	0.24	0.96	0.22	-	<b>1.00</b>	1.00
Effectiveness	1.77	0.65	1.91	0.62	2.19	0.95	2.19	0.65	1.92	-	<b>2.96</b>	0.92
Efficiency	1.75	0.79	2.12	0.80	2.37	0.83	2.00	0.57	2.00	-	<b>2.95</b>	0.92
Persuasiveness	1.83	0.79	1.99	0.79	2.21	0.88	1.93	0.65	2.00	-	<b>2.92</b>	0.94
Transparency	1.79	0.78	1.97	0.70	2.21	1.00	1.89	0.54	1.92	-	<b>2.97</b>	0.97
Scrutability	1.69	0.79	1.94	0.65	2.07	0.91	1.78	0.60	2.00	-	<b>2.95</b>	0.93
Trust	1.90	0.71	2.08	0.82	2.26	0.85	2.07	0.64	2.00	-	<b>2.89</b>	0.93
Satisfaction	1.73	0.79	2.02	0.79	2.26	0.83	2.11	0.73	2.08	-	<b>2.92</b>	0.91
Representativeness	1.98	0.63	2.11	0.69	2.19	0.86	2.04	0.89	2.08	-	<b>2.99</b>	0.95
Reasonability	2.79	0.57	2.64	0.74	2.86	1.00	2.52	0.57	2.92	-	<b>2.99</b>	0.99
Coherence	2.54	0.68	2.55	0.83	2.84	0.92	2.48	0.65	2.75	-	<b>2.99</b>	0.99
Overall	2.06	0.74	2.68	0.74	2.98	0.74	2.56	0.48	2.33	-	<b>4.04</b>	0.94

**Explanation quality at dialogue level.** We average the values on all explanation metrics in each dialogue as their overall ratings. To figure out the relationship between explanation quality and dialogue performance, we ask annotators to label all sampled dialogues at the dialogue level. We design another questionnaire<sup>2</sup> containing 5 questions covering the main overall goals of CRS, i.e., the *effectiveness*, *efficiency*, *dialogue quality* and *satisfaction* [19]. These overall

goals evaluate different aspects of the CRS, i.e., users’ acceptance rate of recommendation, time cost to make decisions, the quality of conversation, and the system’s success rate in meeting user needs. For more details on annotation, see our repository [https://github.com/Superbooming/E-ReDial/blob/main/Dialogue\\_level.md](https://github.com/Superbooming/E-ReDial/blob/main/Dialogue_level.md). Based on the annotation result, we compute the Spearman Correlation between the explanation and CRS metrics. The correlation result is

<sup>2</sup>See [https://github.com/Superbooming/E-ReDial/blob/main/Dialogue\\_level.md](https://github.com/Superbooming/E-ReDial/blob/main/Dialogue_level.md)



**Figure 2: Spearman Correlation between the overall goals of CRS and different explanation perspectives. Each column corresponds to one CRS metric, and each row corresponds to one explanation metric.**

shown in Figure 2. We omit the *effectiveness*, *reasonability* and *coherence* of explanations since they are weakly correlated with CRS metrics. Our results indicate that the explanation quality is highly correlated with the overall performance of the CRS. Specifically, the existence (*explanation rate*) and quality of explanations (*overall*) can significantly influence the performance of CRS. The *effectiveness* and *efficiency* of CRS are strongly correlated with almost all explanation metrics, which means the improvement of any perspectives of explanation is beneficial to them. The *dialogue quality* of CRS is mainly related to *representativeness* of explanation, which suggests that dialogues with more diverse movie descriptions will be considered to have higher dialogue quality. The *satisfaction* of CRS has a higher correlation with trust-related metrics of explanation, e.g., the *transparent* and *scrutability*.

## 4 DATASET AUGMENTATION

This section elaborates on our efforts to improve the explainability of CRS. We conduct a user study to investigate the characteristics of good explanations for CRS and conduct both manual and automatic methods to rewrite the low-quality explanations on a commonly used CRS dataset. Finally, we propose a new dataset with high-quality explanations, *Explainable Recommendation Dialogues* (E-ReDial). Below, we detail the creation of the dataset.

### 4.1 User study

To explore the approaches to improving the explanation quality of CRS, we conduct a user study to investigate the characteristics of good explanations. Specifically, we invite over 20 participants and let each participant annotate 30 system responses containing explanations of different qualities, which we have evaluated in Section 3.3. We ask them to explain each response why the explanation is good or bad. By analyzing the explanations collected, we arrive at four main characteristics of high-quality explanations for CRS: (1) *Clear recommendation reason* is a rational explanation about why the system recommends these items and what preferences the recommendation is based on. (2) *Representative item description*, which means the detailed information about recommended items, e.g.,

the directors, the actors, the genres, and the plot of recommended movies. The description should be true and have a cinematic characteristic. (3) *Encouragement or personal opinion on recommended items*, which means persuading users to accept recommendations from the view of language or personal experience. (4) *Be reasonable and contextual*, meaning the explanations must be logically and linguistically consistent with the conversation. We aspire to enhance the overall explanation quality of the dataset by generating better explanations, as the existing explanations within the dataset are commonly lacking in meeting the four aforementioned characteristics.

We consider three methods to generate good explanations: re-collecting dialogues and rewriting low-quality explanations manually and automatically. Given the difficulty and cost of re-collecting, we choose the latter two and will explore the approach to re-collecting in future work. This work extends the *ReDial* and rewrites its low-quality explanations. We choose *ReDial* since it is more commonly used and more realistic than a dataset constructed by some specific schemas. Meanwhile, according to annotation results in Section 3.3, it performs the worst explanation quality. We conduct manual and automatic approaches to rewrite its explanations and compare the two methods’ performance.

### 4.2 Manual method

Our manual method involves over 30 participants. They are expected to search for relevant information and rewrite the response to meet all requirements in Section 4.1. The annotation pipeline includes three steps: searching for information, rewriting the response and checking quality. Figure 3 illustrates the steps.

**Searching for information:** To facilitate the needed information for rewriting the responses, such as the movie plot, the trailer, the reviews from *IMDb*, and the wiki knowledge from *Wikipedia*, our annotation pipeline provides the search function to retrieve them. Specifically, the search function leverages *Tf-IDF similarity* between the context and all reviews and returns the three most relevant reviews. For diversity, external information sources are also allowed.

**Rewriting the response:** The annotators are asked to rewrite the less explainable responses without changing the original meaning of the sentence. We further ask the annotators to label the rewritten responses’ different aspects, i.e., the recommendation reasons, movie descriptions, encouragement or personal opinions.

**Checking quality:** Another 10 participants are instructed to check the quality of the rewritten responses. They are asked to answer the questions to determine whether the responses meet the requirements. Responses having no qualified explanations are excluded and resent for annotations.

### 4.3 Automatic method

In addition to the manual approach, we extend the explanations by applying automatic text generation models. We choose *GPT-3*<sup>3</sup> as our benchmark model for an extension since it has achieved outstanding performance on lots of text generation tasks. We test two versions of *GPT-3*, namely *text-davinci-003* and *code-davinci-002*. We input the dialogue context and the system

<sup>3</sup><https://openai.com/api/>

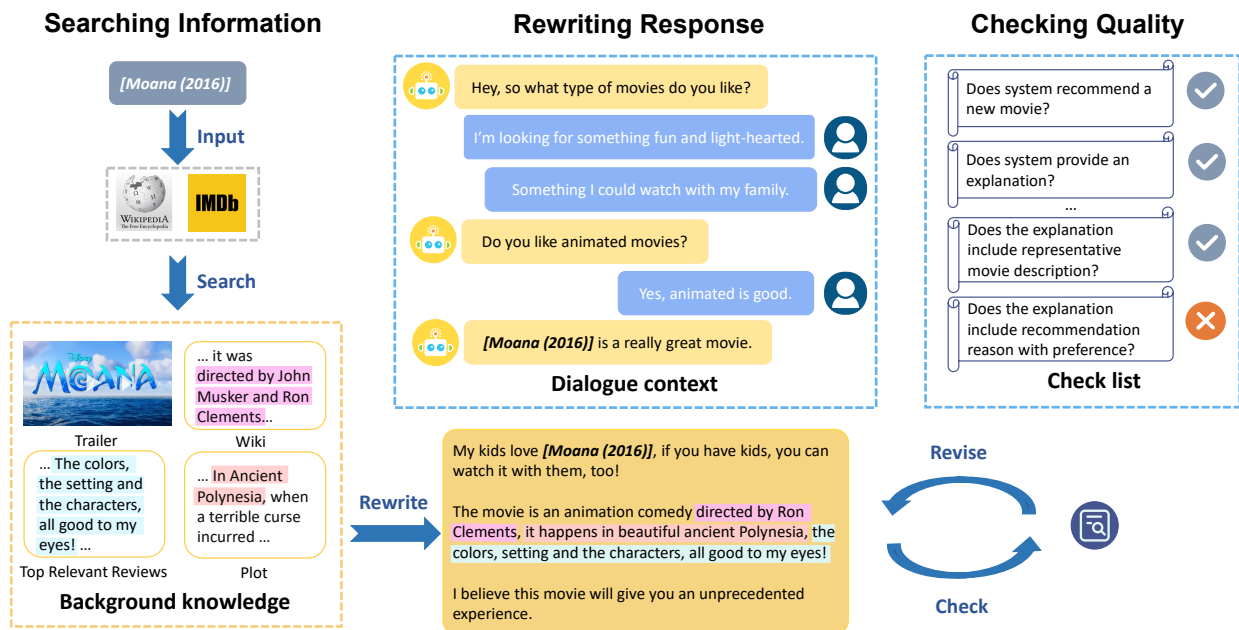


Figure 3: Illustration of the manual annotation pipeline. The response is structured into separate paragraphs for the recommendation reason, movie description, and encouragement, with different coloured highlights indicating varied information sources.

Table 3: An example of extension methods. The automatic method misses [Meet the Parents (2000)] and fails to generate the description of [Meet the Fockers (2004)], while the manual method meets all requirements.

**System:** Hi.  
**User:** Hello! Could you recommend some comedy's? I've had a tough day.  
**System:** Sure, have you seen [Meet the Parents (2000)]? It is soooo funny.  
**User:** No, I haven't.

---

**ReDial:** You have to see it. There is also [Meet the Fockers (2004)] which is the second part, they both are hilarious.

---

**Automatic:** You should definitely watch [Meet the Fockers (2004)], since it's a hilarious comedy. It will surely make you laugh and bring some joy to your tough day.

---

**Human:** Since you want some comedies, you should definitely watch these movies: [Meet the Parents (2000)], in this movie, male nurse Greg Focker meets his girlfriend's parents before proposing, but her suspicious father is every date's worst nightmare; [Meet the Fockers (2004)], which is the second part, this time, Focker came back again, and all hell breaks loose when the Byrnes family meets the Focker family for the first time! Both are hilarious, so don't miss them! They can surely bring you a happy day!

response to be rewritten alongside a task instruction<sup>4</sup> covering all requirements in Section 4.1. Then, we collect the generated text as the new response to the dialogue.

We invite five annotators to check the quality of generated responses. Analyzing the results shows that more than 30% of the

<sup>4</sup>The complete prompt we used is in <https://github.com/Superbooming/E-ReDial/blob/main/Prompt.md>

generated responses need to meet the rewriting requirement. Specifically, changing the original meaning of the sentence, the lack of recommended movies, user preference, and movie descriptions in most cases. An example of extension results generated by the manual and automatic methods is shown in Table 3.

#### 4.4 The E-ReDial dataset

After collecting all eligible manually rewritten responses from *ReDial*<sup>5</sup>, we put them together with the corresponding context to form new dialogues. This results in a new CRS dataset, Explainable Recommendation Dialogues (E-ReDial). It comprises 756 dialogues with 12,003 utterances, each with 15.9 turns on average. 2,058 high-quality explanations are included, each with 79.2 tokens on average. We annotate it in the same way as in Section 3.3 to measure the explanation quality. The detailed evaluation results of E-ReDial are shown in Table 2. The results show that our dataset is significantly better than others in all metrics.

## 5 EXPERIMENTS

### 5.1 Dataset and metrics

**Training and testing data.** We divide the newly labelled data into training and test sets at an 8:2 ratio. The test set comprises 150 conversations and 1121 system responses. Out of the system responses in the test set, 823 are idle or interrogative, with no movie recommendations, while the remaining 418 responses include movie recommendations and explanations. We have categorized the subset

<sup>5</sup>The homepage of ReDial is <https://redialdata.github.io/website/>. ReDial is published under the CC BY 4.0 License, allowing redistribution.



of tests that contain recommendations and explanations as *Test-Rec*, while the entire set of tests is referred to as *Test-Full*.

**Automatic evaluation metrics.** Following previous studies on text generation, we utilize the following automatic evaluation metrics: *F1*, unigram F1 score that measures the similarity between the generated text and the ground-truth response. *BLEU*, we employ BLEU-2 (B2) and BLEU-4 (B4) implemented in the NLTK toolkit <https://www.nltk.org/>. *METEOR*, or MT, is a widely-used metric demonstrating good consistency in human evaluation. *Distinct* assesses text diversity by measuring the number of distinct n-grams in the generated text. Our experiments employ Distinct-2 (D2) and Distinct-3 (D3).

**Human Evaluation Metrics.** In addition to automatic evaluation, we conduct a human evaluation of various benchmark models. Specifically, we randomly sample 30 dialogues generated by the tested models and engage three well-informed annotators to evaluate the responses based on the metrics introduced in Section 3.2.

## 5.2 Benchmark models

We test various benchmark dialogue generation models in E-ReDial:

- **T5-Base** [32], a pre-trained encoder-decoder Transformer with 220M parameters;
- **BART-Base** [20], a pre-trained encoder-decoder Transformer with 140M parameters;
- **GPT-2** [31], a decoder-only Transformer with 117M parameters;
- **DialoGPT-Small** [44], a decoder-only language model that continues pre-trains GPT-2 on 2 billion dialogue corpus;
- **Flan-T5-XXL** [10], an 11B-parameter language model that is finetuned on various NLP datasets using instruction tuning;
- **Davinci-002**, a large language model developed by OpenAI with about 170B parameters<sup>6</sup> and the model is pre-trained on a mixed corpus including both code and text.

This paper aims to compare the efficacy of different approaches for generating responses with explanations rather than improving the accuracy of movie recommendations. Therefore, we keep the recommendation results as the ground truth for all benchmark models to ensure a fair comparison. The inputs for the models are the dialogue context and the names of the recommended movies, while the output is the response to the dialogue.

To further enhance the models, we propose a **knowledge-grounded (KG)** approach that utilizes the background knowledge of the recommended movies when generating recommendation explanations. For the KG approach, we incorporate each movie’s plot, wiki, and reviews into the model input, as described in Section 4.2. We only select the most relevant reviews based on the context to account for input length limitations.

Finetuning can be extremely costly in benchmark models incorporating large language models such as Flan-T5-XXL and Davinci-002. Therefore, we use a **prompt-based approach** with in-context learning (ICL). Specifically, we select eight (*context, response*) examples from the E-ReDial training set to serve as in-context exemplars,

<sup>6</sup>We use the code-davinci-002 API, and OpenAI has not disclosed the details of the parameters of the model until the time of writing the paper.

**Table 4: Performance for response generation among different training-based models in *Test-Full*. Bold face indicates the best result in terms of the corresponding metric.**

	F1	B2	B4	MT	D2	D3
T5-Base						
+ ReDial	20.35	3.65	1.70	11.34	34.04	39.23
+ ReDial + KG	19.34	3.56	1.44	11.01	32.39	36.62
+ E-ReDial	24.31	6.75	2.38	15.14	26.69	39.18
+ E-ReDial + KG	<b>26.01</b>	<b>7.89</b>	<b>4.10</b>	<b>16.29</b>	<b>47.60</b>	<b>59.44</b>
BART-Base						
+ ReDial	18.63	3.12	1.39	10.38	30.57	32.27
+ ReDial + KG	18.04	2.71	1.37	9.86	31.91	32.97
+ E-ReDial	23.63	6.26	2.52	14.11	26.96	35.60
+ E-ReDial + KG	<b>24.46</b>	<b>7.87</b>	<b>4.61</b>	<b>16.48</b>	<b>47.38</b>	<b>59.11</b>
GPT-2						
+ ReDial	19.82	3.61	1.54	11.80	28.10	34.51
+ ReDial + KG	18.71	3.08	1.54	10.65	34.15	39.62
+ E-ReDial	22.87	6.02	2.53	13.86	25.29	34.44
+ E-ReDial + KG	<b>25.38</b>	<b>8.75</b>	<b>4.55</b>	<b>17.83</b>	<b>40.33</b>	<b>52.63</b>
DialoGPT-Small						
+ ReDial	18.67	3.49	1.39	11.29	31.37	36.72
+ ReDial + KG	18.12	3.11	1.46	10.33	29.34	33.77
+ E-ReDial	22.80	6.06	2.13	14.23	19.02	26.83
+ E-ReDial + KG	<b>25.70</b>	<b>8.12</b>	<b>4.22</b>	<b>16.61</b>	<b>30.87</b>	<b>40.34</b>

which are then inputted into the model alongside a task instruction<sup>7</sup>. Due to the model’s limited context size, we do not input background knowledge to prompt-based models.

**Implementation details.** We implement the models using PyTorch and Huggingface Transformers. All the finetuned models are optimized using AdamW optimizer with  $lr=5e-5$ , batch size of 16, and are trained up to 20 epochs. During testing, for finetuned-based models, we employ greedy decoding. For prompt-based models, i.e., Flan-T5-XXL and code-davinci-002, we employ nucleus decoding with  $top-p=0.9$ ,  $temperature=0.8$ .

## 5.3 Automatic evaluation results

**Performance of training-based models.** Table 4 displays the results of various training-based models for response generation. *+ReDial* indicates that the models are trained on the original ReDial datasets, while *+E-ReDial* indicates that they are trained on the newly collected E-ReDial dataset. The addition of *+KG* indicates using a knowledge-grounded approach.

From the results, we have three key findings: Firstly, the models trained on E-ReDial outperform those trained on ReDial across all backbone models and metrics. This suggests that the proposed E-ReDial dataset can enhance the quality of CRS responses. Secondly, incorporating knowledge-grounded methods (*+KG*) significantly<sup>8</sup> improves the performance of models trained on E-ReDial data, but no improvement is witnessed in the models trained on ReDial. This may be because the ReDial responses merely mention the name of movies without providing explanations using background knowledge, unlike E-ReDial. Lastly, the *T5-Base + E-ReDial + KG*

<sup>7</sup>The complete prompt we used is in <https://github.com/Superbooming/E-ReDial/blob/main/Prompt.md>

<sup>8</sup>Improvements are significant at  $p < 0.05$  level using t-test.

**Table 5: Performance for response generation among models training on different data size in *Test-Full*. Bold face indicates the best result in terms of the corresponding metric.**

	F1	B2	B4	MT	D2	D3
<i>From T5-Base</i>						
+ E-ReDial (0%)	9.74	1.70	0.40	6.17	27.80	36.12
+ E-ReDial (5%)	20.46	6.23	3.01	14.11	43.75	57.71
+ E-ReDial (10%)	21.46	6.92	3.69	15.46	41.29	52.52
+ E-ReDial (50%)	24.92	<b>8.45</b>	<b>4.76</b>	<b>16.73</b>	46.53	58.61
+ E-ReDial (100%)	<b>26.01</b>	7.89	4.10	16.29	<b>47.60</b>	<b>59.44</b>
<i>From T5-Base finetuned on full ReDial</i>						
+ E-ReDial (0%)	22.61	4.38	2.16	12.46	32.83	40.65
+ E-ReDial (5%)	25.50	7.92	3.90	17.12	42.92	57.68
+ E-ReDial (10%)	27.04	8.69	4.66	17.37	44.47	58.41
+ E-ReDial (50%)	27.63	9.65	<b>5.00</b>	18.65	<b>47.72</b>	<b>60.28</b>
+ E-ReDial (100%)	<b>27.98</b>	<b>9.71</b>	4.84	<b>18.88</b>	47.07	59.20

**Table 6: Performance for response generation between training-based and prompt-based models in *Test-Rec*. Bold face indicates the best result.**

	F1	B2	B4	MT	D2	D3
T5-Base (+ReDial)	14.76	0.47	0.21	8.37	41.70	48.74
T5-Base (+E-ReDial)	29.53	12.69	5.22	19.83	27.17	40.95
T5-Base (+E-ReDial +KG)	<b>34.62</b>	<b>18.10</b>	<b>11.22</b>	<b>26.65</b>	48.66	62.24
Flan-T5-XXL (8-shot ICL)	24.52	7.53	2.45	15.34	51.08	70.33
Davinci-002 (8-shot ICL)	33.81	14.86	6.07	25.55	<b>53.07</b>	<b>74.60</b>

performs the best overall, and thus, we conduct a detailed analysis of it in our subsequent experiments.

**Impact of data size.** Since the proposed E-ReDial data size is relatively small, we conduct ablation experiments on data size to analyze its impact. Table 5 presents the results of training the *T5-Base+KG* model on different proportions of E-ReDial data, ranging from 0% to 100%. We also compare the models finetuned from the vanilla T5 or the T5 pre-trained on ReDial data, with the former representing the first group in Table 5 and the latter representing the second group. From the results, we see that as the amount of data increases, the models demonstrate improvement on all metrics until 50% of the data is used. Beyond that point, the models show a decrease in some metrics. This finding suggests that the labelled data size of E-ReDial is sufficient to train the knowledge-grounded T5-Base model to generate explanations, and increasing the annotated data may provide limited gains. Moreover, using the ReDial pre-trained model effectively enhances the model’s performance in low-resource settings. For example, using the pre-trained model, it is possible to use only 10% of the E-ReDial data to outperform vanilla T5-Base trained with 100% of the E-ReDial data.

**Performance of prompt-based methods.** Table 6 compares the results of two prompt-based models, Flan-T5-XXL and Davinci-002, with training-based T5 models. We calculate the results only on *Test-Rec* subset instead of the full test set. The results show that prompt-based models outperform training-based models in diversity (measured by D2 and D3). Furthermore, Davinci-002 achieves

**Table 7: Human evaluation results of explanations generated by different models. Bold face indicates the best result in terms of the corresponding metric.**

	T5+ReDial	T5+E-ReDial	T5+E-ReDial+KG	Davinci-002
Explanation rate	0.23	1.00	1.00	1.00
Effectiveness	1.00	1.77	2.33	<b>2.50</b>
Efficiency	1.00	2.10	2.43	<b>2.50</b>
Persuasiveness	1.00	2.00	<b>2.30</b>	<b>2.30</b>
Transparency	1.71	<b>2.97</b>	<b>2.97</b>	2.50
Scrutability	1.71	2.47	<b>2.63</b>	2.17
Trust	1.00	2.03	<b>2.47</b>	2.37
Satisfaction	1.14	2.00	2.47	<b>2.57</b>
Representativeness	1.43	2.10	2.70	<b>2.83</b>
Reasonability	1.71	2.60	2.73	<b>2.77</b>
Coherence	1.57	2.63	2.73	<b>2.77</b>
Overall	1.14	3.10	<b>3.67</b>	3.43

comparable results to the best-performing training model T5-Base (+E-ReDial +KG), despite not explicitly inputting the background knowledge.

## 5.4 Human evaluation results

Table 7 shows the human evaluation results of explanations generated by four benchmark models. The results suggest that the T5 model trained on ReDial data (T5+ReDial) not only has a low explanation generation rate, accounting for only 23% but produces low-quality responses, with an overall score of only 1.14. Training the model on the proposed E-ReDial significantly improves the explanation generation rate and quality. Furthermore, comparing the results of the explanations generated by T5-Base with and without knowledge-grounded generation indicates that introducing knowledge can substantially enhance almost all metrics. This finding is consistent with the automatic evaluation results shown in Table 4. Additionally, T5-Base+KG and Davinci-002 exhibit their strengths, with T5-Base+KG having a slightly higher overall rating. By analyzing the generated responses, we infer that Davinci-002 generates more human-like and diverse movie descriptions, improving the effectiveness and efficiency of conversation and leading to higher user satisfaction and acceptance. T5-Base+KG generates more precise reasons for recommendations based on user preferences, reasons increasing user trust and persuasiveness.

## 5.5 Case study

Table 8 shows two examples of the model outputs. In the first example, T5+ReDial recommends the movie without explanation. T5-E-ReDial explains, but its description of the movie is vague and unconvincing due to a lack of background knowledge. T5-E-ReDial+KG, quoting classic scenes from the movie, better recommends the movie [Spider-Man (2002)] to users. Davinci-002 also explains the recommendations by acknowledging the user preferences and introducing the movie plot. A similar phenomenon can be seen in the second example. T5+ReDial gives a short explanation that is non-factual and unconvincing. T5+E-ReDial’s response is more explanatory but also includes factual mistakes. T5+E-ReDial+KG and Davinci-002 generate responses that satisfy the users, whereas



**Table 8: Case study.**

Example 1
<p><b>User:</b> Do you recommend any movies? <b>System:</b> Do you like Marvel movies? <b>User:</b> Sure! What do you recommend to me?</p>
<p><b>T5+ReDial:</b> [Spider-Man (2002)] is a good one.</p>
<p><b>T5+E-ReDial:</b> I will recommend [Spider-Man (2002)] since you like Marvel movies. It tells that Spider-Man, a mutant, must step forward to lead his people into a new future and must confront a challenger from his country's past. You will love this movie!</p>
<p><b>T5+E-ReDial+KG:</b> Since you like love movies, then I think these movies will suit your taste: [Spider-Man (2002)]. With great power comes great responsibility. These words, spoken by Peter Parker's uncle Ben (Cliff Robertson) during their last conversation and which have become Spider-Man's motto, also perfectly represent the challenge Sam Raimi was facing when asked to bring one of America's most loved superheroes to the screen. I believe that you will enjoy yourself while watching it.</p>
<p><b>Davinci-002:</b> I will recommend you [Spider-Man (2002)] since you like Marvel movies. After being bitten by a genetically altered spider, nerdy high school student Peter Parker is endowed with amazing powers. I will recommend you this movie and hope you like it.</p>
Example 2
<p><b>User:</b> See any good films lately? Any dramas or comedies you'd like to suggest? <b>System:</b> I recommend [Wonder Woman (2017)] and [The Big Sick (2017)] because they are both drama or comedy films, [...] <b>User:</b> I was disappointed in the last two films I saw [The Shape of Water (2017)] and [Lady Bird (2017)]. haven't seen [Wonder Woman (2017)] . Loved [The Big Sick (2017)]</p>
<p><b>T5+ReDial:</b> I also recommend [Get Out (2017)], which is a drama film about a young girl who falls in love with a man who is a teen.</p>
<p><b>T5+E-ReDial:</b> [Get Out (2017)] is a good one. Since you want a good drama, [Get Out (2017)] is a good one. The story is that a young man who is trying to get out of a house with his girlfriend and her friend is forced to make a decision that will affect their future. I highly recommend it.</p>
<p><b>T5+E-ReDial+KG:</b> Since you like dramas or comedies, then I think these movies well suit your taste [Get Out (2017)] It's nostalgic in its embracing of contemporary and classic horror and also brave in the accomplished handling of the film's combative themes. This movie is worth watching, you will like it.</p>
<p><b>Davinci-002:</b> [Get Out (2017)] is a great horror film. It's about Chris and his girlfriend meeting her parents for the first time. At first, Chris reads the family's overly accommodating behavior as nervous attempts to deal with their daughter's interracial relationship, but as the weekend progresses, a series of increasingly disturbing discoveries lead him to a truth that he could have never imagined. I recommend this movie to you, you won't be disappointed.</p>

T5+E-ReDial+KG gives more apparent reasons based on user preferences, and Davinci-002's response is more realistic and contains diverse movie descriptions.

## 6 CONCLUSION

To elaborate on more explainable CRS, we have elucidated the meaning of the explanation for CRS and proposed ten evaluation perspectives to evaluate them. We further assessed existing benchmarks and verified the necessity of improving the explanation quality for CRS. To improve the explanation quality, both manual and automatic methods are employed. Our manual method has collected and released a new CRS dataset, E-ReDial, which includes 756 dialogues with 2,058 high-quality explanations. In addition, we compared two baselines for explanation generation: training-based and prompt-based methods. Experiments conducted on the newly collected dataset suggest that models trained on E-ReDial can significantly improve explainability while introducing knowledge into the models can further improve the performance. GPT-3 in the in-context learning setting can generate more realistic and diverse movie descriptions. In contrast, T5 training on E-ReDial can better generate clear reasons for recommendations based on user preferences.

**Resource utilization.** The E-ReDial dataset can be used for explanation generation and other conversational information access tasks. As a CRS dataset, it can be used for any CRS-related tasks.

Since each explanation is annotated with the knowledge, it can also be used in knowledge-grounded conversation. The components annotated on explanations, i.e., the recommendation reasons, the items' descriptions, encouragement, or personal opinions, can be used for corresponding generation tasks.

**Limitations.** There are some limitations in this work. Firstly, we improve the explainability of CRS by extending an existing CRS dataset. We rewrite the responses with low-quality explanations via manual and automatic methods. All rewritten responses passing the quality check will be collected, and we have put them with the corresponding context to construct our dataset. Since the dialogue is obtained by extending the existing dialogue rather than from a real-world conversation scene, its explanations may not be realistic enough. Secondly, We only use the commonly used text generation models to generate explanations without designing a dedicated module, lacking further exploration.

**Future work.** This work opens up several directions for future studies in building explainable conversational recommender systems, as follows. Firstly, we will explore more effective approaches to improving the explainability of CRS datasets, e.g., constructing dialogues from scratch. Secondly, we will investigate more complex evaluation methods for explanation quality, e.g., automatic metrics. Finally, we will explore more methods to generate better explanations automatically.

## REFERENCES

- [1] Krisztian Balog and Filip Radlinski. 2020. Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations. In *SIGIR*. 329–338.
- [2] Mustafa Bilgic and Raymond J. Mooney. 2005. Explaining Recommendations: Satisfaction vs. Promotion. In *IUI*. 153.
- [3] Shuo Chang, F. Maxwell Harper, and Loren G. Terveen. 2016. Crowd-Based Personalized Natural Language Explanations for Recommendations. In *RecSys*. 175–182.
- [4] Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards Knowledge-Based Recommender Dialog System. In *EMNLP*. 1803–1813.
- [5] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *SIGIR*. 765–774.
- [6] Xu Chen, Zheng Qin, Yongfeng Zhang, and Tao Xu. 2016. Learning to Rank Features for Recommendation over Multiple Categories. In *SIGIR*. 305–314.
- [7] Xu Chen, Zheng Qin, Yongfeng Zhang, and Tao Xu. 2016. Learning to rank features for recommendation over multiple categories. In *SIGIR*. 305–314.
- [8] Xu Chen, Yongfeng Zhang, and Jingxuan Wen. 2022. Measuring “Why” in Recommender Systems: a Comprehensive Survey on the Evaluation of Explainable Recommendation. *ArXiv abs/2202.06466* (2022).
- [9] Zhongxia Chen, Xiting Wang, Xing Xie, Mehul Parsana, Akshay Soni, Xiang Ao, and Enhong Chen. 2021. Towards explainable conversational recommendation. In *IJCAI*. 2994–3000.
- [10] Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *ArXiv abs/2210.11416* (2022).
- [11] Dan Cosley, Shyong K. Lam, I Edwin Albert, Joseph A. Konstan, and John Riedl. 2003. Is seeing believing?: how recommender system interfaces affect users’ opinions. In *CHI*. 585–592.
- [12] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob J. Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *UMUAI* 18 (2008), 455–496.
- [13] Marek Czarkowski and Judy Kay. 2002. A Scrutable Adaptive Hypertext. In *AH*. 384–387.
- [14] Alexander Felfernig and Bartosz Gula. 2006. An Empirical Study on Consumer Behavior in the Interaction with Knowledge-based Recommender Applications. In *CEC/EEE*. 37–37.
- [15] Chongming Gao, Wenqiang Lei, Xiangnan He, M. de Rijke, and Tat-Seng Chua. 2021. Advances and Challenges in Conversational Recommender Systems: A Survey. *AI Open* 2 (2021), 100–126.
- [16] Fatih Gedikli, D. Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *Int. J. Hum. Comput. Stud.* 72 (2014), 367–382.
- [17] Shijie Geng, Zuohui Fu, Yingqiang Ge, Lei Li, Gerard de Melo, and Yongfeng Zhang. 2022. Improving Personalized Explanation Generation through Visualization. In *ACL*. 244–255.
- [18] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyang Shi, and Zhou Yu. 2020. INSPIRED: Toward Sociable Recommendation Dialog Systems. In *EMNLP*. 8142–8152.
- [19] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2020. A Survey on Conversational Recommender Systems. *CSUR* 54 (2020), 1–36.
- [20] 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 *ACL*. 7871–7880.
- [21] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Extra: Explanation ranking datasets for explainable recommendation. In *SIGIR*. 2463–2469.
- [22] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural Rating Regression with Abstractive Tips Generation for Recommendation. In *SIGIR*. 345–354.
- [23] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *NeurIPS*. 9748–9758.
- [24] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards Conversational Recommendation over Multi-Type Dialogs. In *ACL*. 1036–1049.
- [25] Yu Lu, Junwei Bao, Yan Song, Zichen Ma, Shuguang Cui, Youzheng Wu, and Xiaodong He. 2021. RevCore: Review-Augmented Conversational Recommendation. In *ACL*. 1161–1173.
- [26] Kai Luo, Hojin Yang, Ga Wu, and Scott Sanner. 2020. Deep Critiquing for VAE-Based Recommender Systems. In *SIGIR*. 1269–1278.
- [27] Sean M. McNee, Shyong K. Lam, Joseph A. Konstan, and John Riedl. 2003. Interfaces for Eliciting New User Preferences in Recommender Systems. In *UM*. 178–187.
- [28] Alessandro B. Melchiorre, Navid Rekasaz, Christian Ganhör, and Markus Schedl. 2022. ProtoMF: Prototype-based Matrix Factorization for Effective and Explainable Recommendations. In *RecSys*. 246–256.
- [29] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialogKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs. In *ACL*. 845–854.
- [30] Roobina Ohanian. 1990. Construction and Validation of a Scale to Measure Celebrity Endorsers’ Perceived Expertise, Trustworthiness, and Attractiveness. *Journal of Advertising* 19 (1990), 39–52.
- [31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1 (2019), 9.
- [32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67.
- [33] Zhaochun Ren, Zhi Tian, Dongdong Li, Pengjie Ren, Liu Yang, Xin Xin, Huasheng Liang, M. de Rijke, and Zhumin Chen. 2022. Variational Reasoning about User Preferences for Conversational Recommendation. In *SIGIR*. 165–175.
- [34] Amit Sharma and Dan Cosley. 2013. Do social explanations work?: studying and modeling the effects of social explanations in recommender systems. In *WWW*. 1133–1144.
- [35] Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating User Satisfaction for the Evaluation of Task-oriented Dialogue Systems. In *SIGIR*. 2499–2506.
- [36] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. 2008. Justified Recommendations based on Content and Rating Data. In *WebKDD*.
- [37] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. 2021. Counterfactual Explainable Recommendation. In *CIKM*. 1784–1793.
- [38] Nava Tintarev and Judith Masthoff. 2008. Over- and underestimation in different product domains. In *ECAL*. 14–19.
- [39] Nava Tintarev and Judith Masthoff. 2015. Explaining Recommendations: Design and Evaluation. In *Recommender Systems Handbook*. 353–382.
- [40] Khanh Tran, Azin Ghazimatin, and Rishiraj Saha Roy. 2021. Counterfactual Explanations for Neural Recommenders. In *SIGIR*. 1627–1631.
- [41] Bingbing Wen, Xiaoning Bu, and Chirag Shah. 2022. EGCR: Explanation Generation for Conversational Recommendation. *ArXiv abs/2208.08035* (2022).
- [42] Yongfeng Zhang and Xu Chen. 2018. Explainable Recommendation: A Survey and New Perspectives. *Found. Trends Inf. Retr.* 14 (2018), 1–101.
- [43] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *SIGIR*. 83–92.
- [44] Yizhe Zhang, Siqu Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B. Dolan. 2020. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *ACL*. 270–278.
- [45] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji rong Wen, and Jingsong Yu. 2020. Improving Conversational Recommender Systems via Knowledge Graph based Semantic Fusion. In *SIGKDD*. 1006–1014.
- [46] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji rong Wen. 2020. Towards Topic-Guided Conversational Recommender System. 4128–4139.
- [47] Lixin Zou, Long Xia, Yulong Gu, Xiangyu Zhao, Weidong Liu, Xiangji Huang, and Dawei Yin. 2020. Neural Interactive Collaborative Filtering. In *SIGIR*. 749–758.