Do Mentioned Items Truly Matter? Enhancing Conversational Recommender Systems with Causal Intervention and Large Language Models

Lingzhi Wang[1], **Xingshan Zeng**[2*] and **Kam-Fai Wong**[3]
[1]Harbin Institute of Technology (Shenzhen), Shenzhen, China
[2]Huawei Noah's Ark Lab, China
[3]The Chinese University of Hong Kong, China
wanglingzhi@hit.edu.cn, zeng.xingshan@huawei.com, kfwong@se.cuhk.edu.hk

Abstract

Conversational Recommender Systems (CRS) have become increasingly important due to their ability to recommend items through interactive dialogue, adapting to user preferences in real time. Traditional CRS approaches face challenges in generating high-quality, diverse responses due to the limited availability of training data and the inherited biases from domain-specific fine-tuning. Furthermore, existing systems often overlook the impact of confounding variables during user interactions, leading to suboptimal recommendations. In this work, we propose a novel hybrid framework that integrates large language models (LLMs) with traditional recommendation techniques to address these limitations. Our approach leverages the strengths of LLMs in generating fluent, contextually appropriate responses while employing a traditional recommendation module to capture complex interaction structures. To ensure unbiased recommendations, we introduce causal interventions that disentangle confounding variables, improving recommendation accuracy. We evaluate our framework on established CRS datasets, demonstrating significant improvements in recommendation quality and response generation. Our results highlight the effectiveness of the causal intervention mechanism in producing more reliable and personalized recommendations, while the LLM-based response generation offers scalability across multiple domains.

1 Introduction

Conversational Recommender Systems (CRS) [Li *et al.*, 2018; Wang *et al.*, 2024] have gained substantial attention for their capacity to recommend items through interactive conversations, capturing users' interests and intents dynamically. A standard CRS [Chen *et al.*, 2019; Zhou *et al.*, 2020] generally consists of two primary components: a recommendation module responsible for recommending items, and a generation module that generates the final user-visible responses containing the recommended items. However, the development of effective generation modules is often hampered by the scarcity of training data [Wang *et al.*, 2022] in the conversational recommendation format. To address this, prior research [Wang *et al.*, 2022; Wang *et al.*, 2024] has leveraged pretrained language models (PLMs), such as DialoGPT [Zhang *et al.*, 2020] and BART [Lewis, 2019], for the generation module. These PLMs, trained on large-scale unsupervised data, are expected to exhibit strong generative capabilities. Nonetheless, during fine-tuning on domain-specific data, these models often inherit limitations from the training corpus, producing responses that adhere to repetitive patterns and lack diversity, reducing the overall quality of the generated responses.

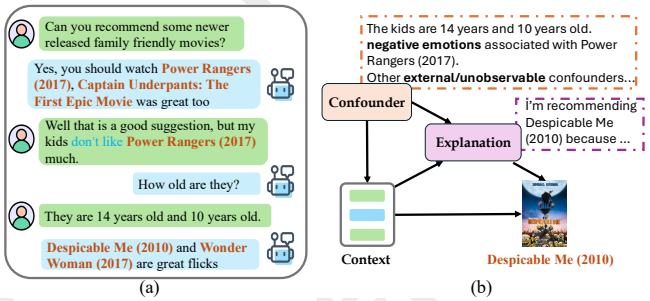


Figure 1: (a) An example of conversational recommendation from the ReDial dataset. (b) The corresponding structural causal model (SCM) for the conversation in (a).

With the emergence of large language models (LLMs) [Ouyang *et al.*, 2022; Touvron *et al.*, 2023], there is a growing trend toward utilizing these models to serve dual roles in both recommendation and response generation. By incorporating relevant external knowledge, LLMs are capable of generating responses that seamlessly integrate recommended items [Di Palma, 2023; Dai *et al.*, 2023; Wang *et al.*, 2023a; Li *et al.*, 2024]. However, external knowledge is often stored in extensive knowledge graphs, and incorporating such data into LLMs poses significant challenges. Directly providing the entire graph as input can overwhelm the model, leading to prohibitively long inputs that impair its performance.

Another challenge prevalent in previous CRS implementations is the ignorance of confounding variables when modeling user interactions. Many CRS systems have relied on extracting mentioned items as key indicators of user preferences, often neglecting the broader context – such as whether the items

* Xingshan Zeng is the corresponding author.

were mentioned positively or negatively, and other text-based expressions of user interest. For instance, in Figure 1(a), the mention of "Power Rangers (2017)" is clearly negative, yet traditional systems would likely treat it as a positive indicative signal for next item recommendation. While the advent of LLMs addresses some of these issues by improving contextual understanding, they still struggle with effectively organizing vast amounts of information. Additionally, real-world user preferences are influenced not only by conversational interactions but also by external, unobserved factors. Existing systems often overlook these confounders, resulting in recommendations that are overly reliant on superficial correlations, ultimately compromising the accuracy of recommendations.

To address these challenges, we propose a hybrid framework that synergizes the strengths of LLMs with traditional recommendation techniques, capitalizing on LLMs' ability to generate fluent, contextually consistent responses, while leveraging traditional recommendation module to model structured knowledge more effectively. In our framework, the recommendation module captures the structural dependencies within interaction data and external knowledge, generating a fixed size of candidate items. The LLM subsequently refines these candidates, producing a re-ranked list and generating a contextual response that highlights the top item along with a justification (i.e., explanation shown in Figure 1(b)) for the recommendation. Crucially, we introduce causal interventions to isolate and mitigate the impact of confounding variables, ensuring that recommendations more accurately reflect user preferences rather than contextually induced biases.

Furthermore, we validate our approach through extensive experiments on established CRS datasets, including ReDial [Li *et al.*, 2018] and OpenDialKG [Moon *et al.*, 2019]. Our results demonstrate overall improvements in recommendation quality compared to baseline models. Additional analyses reveal the effectiveness of our causal intervention through visualizations of the causal effects. In summary, our contributions are:

- We propose a novel hybrid framework that combines the strengths of LLMs and traditional recommendation techniques for conversational recommender systems.
- We introduce a deconfounded recommendation module that effectively mitigates the influence of confounding variables using causal interventions.
- We demonstrate the effectiveness of our approach through extensive experiments, achieving improved recommendation quality and response generation across multiple domains.

2 Related Work

2.1 Conversational Recommender Systems

Conversational recommender systems (CRSs) aim to overcome the limitations of traditional recommendation systems, which rely on sparse implicit feedback. Conventional CRS are divided into two main approaches: question-driven systems, which refine user preferences through clarifying questions [Christakopoulou *et al.*, 2018; Zhang *et al.*, 2018; Aliannejadi *et al.*, 2019; Xu *et al.*, 2021], and multi-turn strategy exploration, which balance exploration and exploitation to optimize recommendations, especially in cold-start scenarios [Li *et al.*, 2010; Christakopoulou *et al.*, 2016; Li *et al.*, 2020]. Recent advances also explore the integration of large language models (LLMs) into CRS, where LLMs can act as direct recommenders [Geng *et al.*, 2022; Hua *et al.*, 2023], assist traditional systems [Xi *et al.*, 2023], or manage recommendation pipelines [Wang *et al.*, 2023b]. These LLM-based CRS enhance user interaction, refine recommendation processes, and incorporate external knowledge sources [Liu *et al.*, 2023; Li *et al.*, 2024; Spurlock *et al.*, 2024], driving further improvements in system performance and user satisfaction.

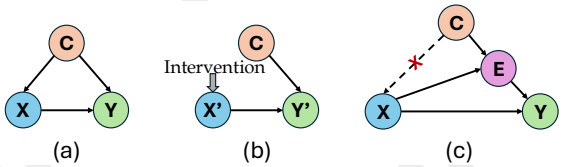


Figure 2: (a) Structural Causal Model (SCM) example, (b) intervention illustration, (c) SCM of Our Work.

2.2 Causality and Interventions

Causal inference has become an increasingly relevant tool in machine learning, providing a framework to identify and control confounding variables that may introduce bias into model predictions [Pearl and others, 2009; Yao *et al.*, 2020]. Unlike associative reasoning, which models conditional probabilities based on observed data, causal inference enables the simulation of interventions to estimate the causal effects of actions, a critical advantage in recommender systems where spurious correlations can lead to suboptimal recommendations. Using tools like the do-operator [Pearl, 2009], causal inference helps evaluate how changes in user preferences affect recommendations while controlling for confounding influences. This approach has been successfully applied to mitigate bias [Bareinboim and Pearl, 2012], enhance generalization [Parascandolo *et al.*, 2018], and modularize reusable features in learning systems [Besserve *et al.*, 2018]. Recent research continues to integrate causal methods into machine learning pipelines, enabling more accurate differentiation between true causal relationships and mere associations [Pearl *et al.*, 2016].

3 Methodology

3.1 Causal Analysis in CRS

We model the causal relationships among history items X, context confounder C, explanation E, and recommended items Y using a structural causal model (SCM, a mathematical framework used to describe causal relationships between variables) [Pearl *et al.*, 2016], shown in Figure 2(c). The direct links represent causal effects.

$\boldsymbol{C \rightarrow X}$: The user's contextual information C (e.g., temporal, social, environmental factors) affects the historical items X they interact with. This suggests that past preferences are shaped by context, with C acting as a confounder for X.

$\boldsymbol{C \rightarrow E \leftarrow X}$: Both C and X influence the generation of explanations E for recommending items. These explanations rationalize why certain items are suggested, drawing on the

user's history and context. For example, if a user liked certain movies (X) during a specific season (C), the explanation for recommending similar movies (E) reflects both factors.

$\boldsymbol{X \rightarrow Y \leftarrow E}$: Historical items X influence the recommendation of new items Y, with E mediating the effect. Explanations help justify the recommendations and align user expectations. For instance, if X represents a preference for action movies and E explains why a particular action movie is recommended, the recommended Y will match this rationale.

In this causal structure, C acts as a confounder, affecting both historical preferences X and recommendation explanations E. The recommendation process, recommending items Y, can be biased by the confounding effect of C. To correct for this, we apply *intervention* on X using the backdoor adjustment technique to estimate $P(Y \mid do(X))$. In observational settings, the causal effect of X on Y is confounded by C, which influences both X and Y, potentially skewing results. To isolate the true effect of X on Y, we intervene with *do-calculus* on X, setting X to specific values (e.g., the last interaction), which breaks the causal link between C and X.

The backdoor adjustment formula [Pearl *et al.*, 2016] for estimating $P(Y \mid do(X))$ is:

$$P(Y \mid do(X)) = \sum_{C} P(Y \mid X, C)P(C) \tag{1}$$

It removes C's influence on X by averaging over C, allowing for an accurate estimate of X's causal effect on Y. By controlling for C, the system can provide recommendations that better reflect the user's true preferences, unaffected by external factors. Implementation details are provided in Section 3.3.

3.2 RGCN-Based Recommendation

A typical CRS consists of two main modules: the *generation* module and the *recommendation* module. The system processes the conversation history $S = (t_1, t_2, \ldots, t_m)$, where each t_i is an utterance from the user or the CRS itself, and m is the total number of utterances. Each utterance is made up of word tokens, $t_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,n_i})$, where n_i is the number of tokens in t_i. The recommendation module uses the conversation context S and external knowledge sources (e.g., knowledge graphs) to suggest items $\mathcal{I}_i$ from a candidate set $\mathcal{I}$. The generation module then creates a natural language response $R = (y_1, y_2, \ldots, y_n)$, where n is the number of tokens in the response, based on the recommended items and the conversation history.

To enhance entity modeling, we incorporate external knowledge graphs, such as DBpedia, which provide additional context about the items (e.g., actors for movie recommendations). Knowledge graphs help refine the representation of entities to alleviate data sparsity. A knowledge graph triplet is defined as $\langle e_1, r, e_2 \rangle$, where e_1 and e_2 are entities, and r is a relation between them.

We employ a Relational Graph Convolutional Network (R-GCN) [Schlichtkrull *et al.*, 2018] to generate relation-aware entity representations. It aggregates features from neighboring entities in the knowledge graph to learn a better representation. The update rule for an entity e at layer $(l+1)$ is:

$$\boldsymbol{h}_e^{(l+1)} = \mathrm{ReLU}\left(\sum_{r \in \mathcal{R}'} \sum_{e' \in \mathcal{E}_e^r} \frac{1}{Z_{e,r}} \boldsymbol{W}_r^{(l)} \boldsymbol{h}_{e'}^{(l)}\right) \tag{2}$$

where $\boldsymbol{h}_e^{(l)}$ is the representation of entity e at layer l, $\mathcal{E}_e^r$ is the set of neighboring entities under relation r, and $\mathcal{R}'$ includes all relations plus a self-loop. The learned relation-specific transformation matrix $\boldsymbol{W}_r^{(l)}$ and normalization factor $Z_{e,r}$ help adjust the aggregation process.

Through multiple layers of aggregation, the R-GCN incorporates structural and relational information, producing a final entity representation $\boldsymbol{h}_e$ that is used for further recommendation and generation tasks.

With the modeled entity representations, we can summarize the user's preferences from the entities mentioned in context S. These entities, denoted as $\mathcal{T}_u = \{e_1, e_2, \ldots, e_{|\mathcal{T}_u|}\}$, are extracted from two sources: item entities and contextual entities (such as actors in a movie, which may not be part of the item set). Entities not found in the entity set $\mathcal{E}$ are ignored, following prior work [Chen *et al.*, 2019; Zhou *et al.*, 2020; Lu *et al.*, 2021]. Each entity $e_i \in \mathcal{T}_u$ is ordered based on its appearance in the conversation. We then map these entities to their corresponding representations, $\boldsymbol{H} = \{\boldsymbol{h}_e\}_{e=1}^{|\mathcal{E}|}$, resulting in the sequence $(\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_{|\mathcal{T}_u|})$.

To build a user preference vector that captures the evolving nature of the conversation, we use a time-aware attention mechanism [Wang *et al.*, 2024]. This mechanism gives more weight to recent entities:

$$\boldsymbol{h}^E = \sum_{i=1}^{|\mathcal{T}_u|} \frac{\lambda^{i-1}}{\sum_{i=1}^{|\mathcal{T}_u|} \lambda^{i-1}} \boldsymbol{h}_i \tag{3}$$

Here, λ is a hyperparameter controlling the influence of recency, typically set to a value greater than 1. This ensures that more recent entities have a greater impact on the recommendation. Finally, the initial recommendation probability is calculated as:

$$\boldsymbol{p}_e = \mathrm{softmax}(\mathrm{mask}(\boldsymbol{h}^{\boldsymbol{E}} \boldsymbol{H}^\top)) \tag{4}$$

The mask operation filters out non-item entities by setting their values to $-\infty$, ensuring that the recommendation focuses only on the candidate items in $\mathcal{I}$. Thus, $\boldsymbol{p}_e \in \mathbb{R}^{|\mathcal{E}|}$ represents the probability distribution over the item set.

3.3 Deconfounded Recommendation with Causal Intervention

To address potential confounding effects between a user's historical interactions and context C, we apply causal intervention. The goal is to isolate the true causal effect of the user's preferences (captured by X, i.e., historical entities) on recommended items (Y) without interference from confounding factors. We achieve this by introducing an intervention on X using the backdoor adjustment technique ($P(Y \mid do(X))$). In practice, this intervention is implemented through a masking mechanism applied to the representation of the last entity in the user's history, $\mathcal{T}_u$. By selectively masking parts of the historical sequence, particularly the final entity, we simulate how

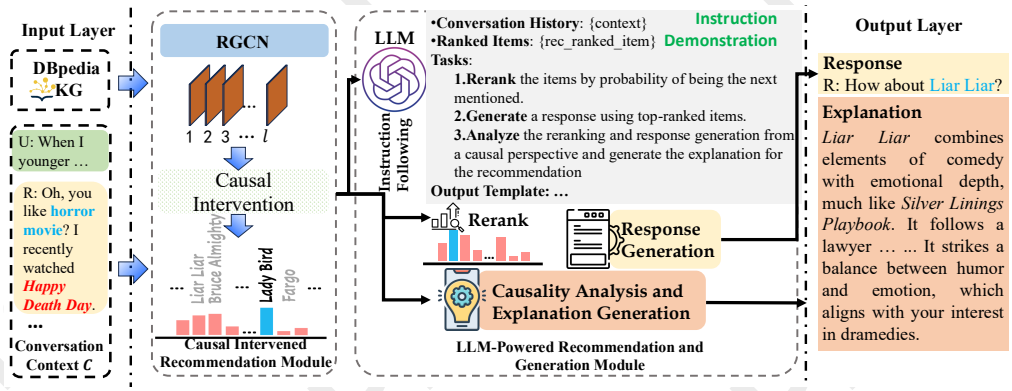


Figure 3: Workflow of our LLM-powered conversational recommender system.

changes in X influence recommendations (Y) and disentangle the effect of confounder C.

Formally, the intervention on the last entity's representation $\boldsymbol{h}_{|\mathcal{T}_u|}$ is performed with a masking operation $\mathcal{M}$, resulting in the updated user representation as follows:

$$\bar{\boldsymbol{h}}^E = \sum_{i=1}^{|\mathcal{T}_u|-1} \frac{\lambda^{i-1}}{\sum_{i=1}^{|\mathcal{T}_u|}\lambda^{i-1}}\boldsymbol{h}_i + \frac{\lambda^{|\mathcal{T}_u|-1}}{\sum_{i=1}^{|\mathcal{T}_u|}\lambda^{i-1}}\mathcal{M}\boldsymbol{h}_{|\mathcal{T}_u|} \quad (5)$$

The corresponding intervention-adjusted probability is then:

$$\bar{\boldsymbol{p}}_e = \text{softmax}(\text{mask}(\bar{\boldsymbol{h}}^{\boldsymbol{E}}\boldsymbol{H}^\top)) \quad (6)$$

This approach prevents the model from relying on spurious correlations between context and past interactions (C and X), enhancing the robustness and accuracy of recommendations.

To formalize the causal intervention, we define a causal loss function that penalizes the similarity between intervention-adjusted recommendations and ground truth while maintaining the similarity between non-intervened recommendations and ground truth. The causal loss is defined as:

$$\mathcal{L}^c = \text{Sim}(\bar{\boldsymbol{p}}_e, \hat{\boldsymbol{p}}) - \text{Sim}(\boldsymbol{p}_e, \hat{\boldsymbol{p}}) \quad (7)$$

where $\hat{\boldsymbol{p}}$ represents the true recommendation labels (as a one-hot vector), and Sim measures the similarity between two distributions (using cosine similarity in our experiments).

The overall training objective combines the causal intervention loss and the regular recommendation loss:

$$\mathcal{L}_{rec} = \text{CrossEn}(\boldsymbol{p}_e, \hat{\boldsymbol{p}}) + \alpha \cdot \mathcal{L}^c + \beta \cdot \text{KL}(\bar{\boldsymbol{p}}_e || \boldsymbol{p}_e) \quad (8)$$

Here, CrossEn represents the cross-entropy loss for the regular recommendation task, and the KL-divergence term between the original and deconfounded recommendations limits the extent of modification introduced by the intervention mask. α and β are hyperparameters controlling the impact of the causal loss and KL-divergence terms. This combined objective encourages the model to make accurate recommendations while ensuring robustness against confounding effects.

3.4 LLM-Enhanced Causal Fused Recommendation and Response Generation

To produce response with recommendation, our framework utilizes the capabilities of large language models (LLMs) without the need for extensive tuning on specialized datasets, contrast to traditional CRS. While domain-specific fine-tuning can be effective, it often struggles with poor generalizability and depends on scarce, high-quality data. On the other hand, relying solely on LLMs, though fluent, can lead to irrelevant or impractical recommendations (e.g., suggesting items not in the candidate list, as shown in Table 2). Our approach overcomes these limitations by combining the generalization power of LLMs with a dedicated recommendation module, enabling scalable, effective response generation.

In our approach, the conversational history S and the ranked recommendation list I^{rec} – obtained from our previous recommendation module and ranked based on the predicted probability – are provided as input to the LLM. The LLM is asked to generate the final response based on the information, while simultaneously producing contextualized explanations E and refining the item list from I^{rec} to I^{llm} for enhanced personalization. This process is formalized as follows:

$$E, I^{llm}, R \leftarrow \mathcal{G}(S, I^{rec}), \quad (9)$$

where $\mathcal{G}$ represents the LLM. The output includes E, the generated explanation for the recommended items, I^{llm}, the refined item list, and R, the natural language response incorporating the recommendations and explanations.

We highlight the necessity of generating additional output in the form of contextualized explanations E. This draws inspiration from chain-of-thought (CoT) prompting [Wei *et al.*, 2022], but extends it further. Our system makes the LLM to take specific reasoning procedure informed by causal reasoning, rather than merely prompting for more reasoning steps. Specifically, we require the LLM to analyze the causal relationships between the given candidate items (i.e., I^{rec}) and the context (i.e., S), producing causal-level explanations E prior to generating the final re-ranked recommendation list.

4 Experimental Setup

Datasets. We conduct experiments on two widely-used CRS datasets: ReDial [Li *et al.*, 2018] and OpenDialKG [Moon *et al.*, 2019]. Table 1 provides a summary of the key statistics for both datasets. We follow standard dataset splits: 80%:10%:10% for training, validation, and test sets in ReDial [Li *et al.*, 2018], and 75%:15%:15% in OpenDialKG

	Conv #	Utter #	Avg Utter #	Domain
ReDial	10,006	182,150	18.2	Movie
OpenDialKG	13,802	91,209	6.6	Movie, Book, Sports, Music

Table 1: Statistics of ReDial and OpenDialKG datasets.

[Moon *et al.*, 2019]. For ReDial, we use a subset of DBpedia as the linked knowledge graph provided by [Chen *et al.*, 2019]. For OpenDialKG, we extract a DBpedia subset by mapping items to corresponding DBpedia entities and linking them to dialogue content following [Daiber *et al.*, 2013].

Evaluation Metrics. We evaluate the performance of the recommendation and generation separately. For the recommendation task, we use Recall@K (with K = 1, 10, 50 for ReDial following [Chen *et al.*, 2019], and K = 1, 3, 5, 10, 25 for OpenDialKG following [Moon *et al.*, 2019]). Recall@K checks if ground truth items are in the top-K predictions. For generation, we report Dist-n (n=2, 3, 4), along with case-insensitive BLEU-n (n=2, 4) scores. BLEU scores are calculated using the NLTK package[1]. For LLM-generated recommendations, we employ fuzzy matching, considering 90% token overlap as a match with the ground truth. LLM-based methods report Recall@K for $K \leq 10$ to avoid noise from longer recommendation lists.

Baselines and Variants. For the ReDial dataset, we compare our model against seven competitive baselines: ReDial [Li *et al.*, 2018], KBRD [Chen *et al.*, 2019], CRWalker [Ma *et al.*, 2020], KGSF [Zhou *et al.*, 2020], RevCore [Lu *et al.*, 2021], C^2-CRS [Zhou *et al.*, 2022], and CTA-CRS [Wang *et al.*, 2024]. While for the OpenDialKG dataset, we compare the following models based on prior works [Moon *et al.*, 2019; Wang *et al.*, 2024]: seq2seq [Sutskever *et al.*, 2014], Tri-LSTM [Young *et al.*,], Ext-ED [Parthasarathi and Pineau, 2018], DialKG Walker [Moon *et al.*, 2019], and CTA-CRS [Wang *et al.*, 2024].

In addition, we establish several LLM-based baselines: (1) ChatGPT Vanilla: A baseline using standard ChatGPT. (2) ChatGPT Few-shot: A few-shot learning variant using ChatGPT. (3) ChatGPT CoT: Chain-of-thought prompting applied to ChatGPT. Finally, we evaluate the performance of several variants of our model to examine the impact of the proposed modules. The FULL MODEL refers to the complete model as described in Section 3. The W/O LLM MODULE variant removes the LLM-based reranking module, while the W/O CAUSAL variant omits the causal interventions and the KL loss during the training of the recommendation module.

Implementation Details and Parameter Settings. For the RGCN, we set both the entity embedding size and the hidden dimension to 128, with the layer number as 1 and the normalization factor $Z_{e,r}$ fixed at 1, consistent with prior work [Chen *et al.*, 2019; Zhou *et al.*, 2020; Wang *et al.*, 2024]. The recency coefficient λ in Eq. 5 is set to 1.5. Additionally, the balance factors in Eq. 8 are set as $\alpha = 2.0$ and $\beta = 0.5$, respectively. For the response generation module, we employ the `gpt-4o-mini` model. During generation, we set the temperature to 1.0. All hyperparameters are tuned based on the model's performance on the validation set.

The recommendation module is trained on an NVIDIA 3090 GPU. We set the batch size to 32, with an update frequency of 4. We employ the Adam optimizer with an initial learning rate of 5e-3, and training is conducted with 1000 warm-up steps followed by a polynomial decay learning rate scheduler. Early stopping is applied based on the validation performance.

[1]We use NLTK (https://www.nltk.org) for BLEU computation.

Models	Input			Rec@1	Rec@10	Rec@50
	Context	KG	Rev			
Baselines						
REDIAL	✓			2.4	14.0	32.0
KBRD	✓	✓		3.1	15.0	33.6
CRWALKER	✓	✓		3.1	15.5	36.5
KGSF	✓	✓		3.9	18.3	37.8
REVCORE	✓	✓	✓	4.6	22.0	39.6
C^2−CRS	✓	✓	✓	5.3	23.3	40.7
CTA-CRS	✓	✓		5.9	24.0	41.3
LLM-Based						
CHATGPT VANILLA	✓			2.6	15.7	-
CHATGPT FEW-SHOT	✓			2.4	11.0	-
CHATGPT COT	✓			2.7	17.3	-
Our Models						
OUR FULL MODEL	✓	✓		**6.1**	**28.0**	-
W/O LLM MODULE		✓		5.9	23.0	**41.8**
W/O CAUSAL		✓		5.2	18.2	34.6

Table 2: **Recommendation** results (in %) on ReDial dataset. Our full model achieves the best performance with less external domain-specific knowledge as input, compared to the previous work. "KG" refers to knowledge graph and "Rev" represents review information.

5 Experimental Results

5.1 Recommendation Result Comparison

We present the recommendation results on the ReDial dataset in Table 2. Several key observations can be drawn:

• *Our full model surpasses the baselines without requiring additional external knowledge.* Our method outperforms previous SOTAs in terms of both Rec@1 and Rec@10, demonstrating that it can produce more precise recommendations without extra domain-specific knowledge, such as reviews.

• *The causal-enhanced recommendation module plays a key role.* Our model variant "W/O LLM MODULE" exhibits competitive performance, especially in terms of Rec@50. It surpasses baselines such as REVCORE and C^2−CRS, both of which leverage additional external knowledge, while further removing causal relative modules ("W/O CAUSAL") results in a larger performance drop.

• *The LLM reranking enhances the overall performance of our full model.* A comparison between the full model and its variant without the LLM module ("W/O LLM MODULE") highlights the critical role of the LLM in refining and reranking the recommendations. The improvement in Rec@10 from 23.0% to 28.0% suggests that the LLM's reranking process helps better surface relevant items in the top recommendations.

Models	Rec@1	Rec@3	Rec@5	Rec@10	Rec@25
Baselines					
seq2seq	3.1	18.3	29.7	44.1	60.2
Tri-LSTM	3.2	14.2	22.6	36.3	56.2
Ext-ED	1.9	5.8	9.0	13.3	19.0
DialKG Walker	13.2	26.1	35.3	47.9	62.2
CTA-CRS	18.0	33.5	41.5	50.0	64.8
Our Models					
OUR FULL MODEL	**20.9**	**40.1**	**53.1**	**60.7**	-
W/O LLM MODULE	19.8	35.6	40.9	52.4	**65.8**
W/O CAUSAL	16.0	28.9	34.3	45.1	57.9

Table 3: **Recommendation** results (in %) on OpenDialKG.

Models	Dist-2	Dist-3	Dist-4	BLEU-2	BLEU-4
TRANSFORMER	14.8	15.1	13.7	-	-
REDIAL	22.5	23.6	22.8	17.8	7.4
KBRD	26.3	36.8	42.3	18.5	7.4
KGSF	28.9	43.4	51.9	16.4	7.4
REVCORE	42.4	55.8	61.2	-	-
CTA-CRS	45.7	65.3	76.1	**19.1**	**8.9**
OURS	**46.2**	**67.2**	**77.8**	8.5	2.3

Table 4: **Generation** results (in %) on the ReDial dataset.

We also present the recommendation results on the OpenDialKG dataset in Table 3, where similar observations can be drawn. The experimental results on both ReDial and OpenDialKG demonstrate the versatility and effectiveness of our framework, consistently outperforming the state-of-the-art baselines without relying on domain-specific knowledge.

5.2 Generation Result Comparison

We evaluate the generation performance of our models and baselines on the ReDial dataset using both automatic evaluation (Table 4) and human evaluation (Table 5). These evaluations provide insights into both the diversity and quality of generated responses, offering a comprehensive assessment.

Automatic Evaluation

Table 4 presents the automatic evaluation results. We can find:

• *Our method outperforms the baselines in diversity.* Our model achieves the highest scores in all Dist-n metrics, indicating that our model generates more varied and less repetitive responses compared to baselines.

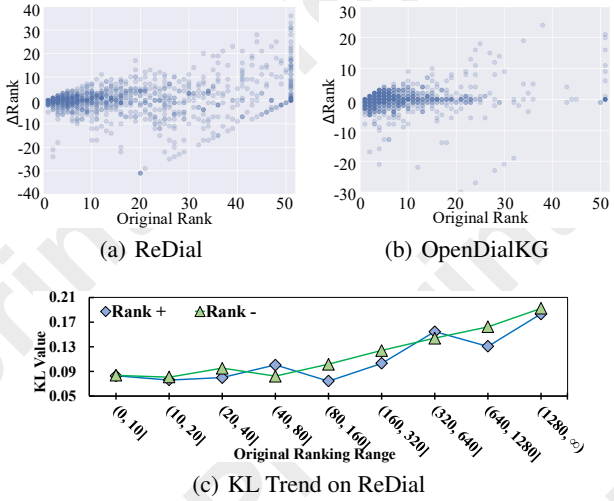
• *BLEU scores of our model are relatively low due to the nature of LLM-based generation.* Since the LLM in our system is not specifically fine-tuned on the ReDial dataset, the reference-based BLEU scores are lower than the baselines. However, we argue that BLEU, being a reference-dependent metric, is not well-suited to capture the true quality of LLM-generated responses, as LLMs prioritize fluency, coherence, and diversity over matching specific reference sentences. We have conducted a human evaluation to further validate it.

Human Evaluation

To evaluate response quality, we conducted a human evaluation using 100 randomly sampled context-response pairs from the test set, comparing our model's outputs with baselines. Two crowd-workers independently rated the responses on three aspects (following [Bao *et al.*, 2020]) using a [0, 1, 2] scale, where higher scores indicate better quality. Table 5 presents the results. The Cohen's kappa coefficient for inter-rater reliability exceeds 0.65, demonstrating substantial agreement. Key observations include:

Models	Fluency	Informativeness	Coherence
Ground Truth	1.95	1.71	1.71
REDIAL	1.92	1.32	1.23
KBRD	1.95	1.39	1.31
KGSF	1.91	1.02	0.95
CTA-CRS	1.95	1.54	1.66
OURS	**1.98**	**1.96**	1.80
CHATGPT VANILLA	**1.98**	1.76	**1.95**

Table 5: **Human evaluation** of the **generation** results on ReDial.



(a) ReDial (b) OpenDialKG

(c) KL Trend on ReDial

Figure 4: Figures 4(a) and 4(b) visualize ranking changes (Δ Rank) relative to the original rank (position of the ground truth item in the recommendation list) after applying causal intervention. Deeper colors indicate higher frequencies, with darker regions showing greater instance concentrations. Figure 4(c) shows KL divergence trends for two instance categories: "Rank +" (higher ground-truth ranking post-intervention) and "Rank -" (lower ground-truth ranking post-intervention), based on the original ranking.

• *LLM-based methods consistently outperform others in overall quality.* Our model scores highest in fluency and informativeness, producing grammatically superior and richer responses than traditional models. Notably, LLM-based methods sometimes surpass ground truth, indicating they can generate responses more polished than human annotations.

• *Our model is more informative than* CHATGPT VANILLA*, with a slight trade-off in coherence.* It achieves a higher informativeness score (1.96 vs. 1.76), likely due to the recommendation module's item list. However, this emphasis on recommendations occasionally reduces coherence, as the model prioritizes item suggestions.

5.3 Effectiveness of Proposed Mechanisms

Analysis of Causal Intervention Via Ranking Changes

Causal intervention helps the model address confounding factors, improving recommendations. To evaluate its case-level impact, we analyze Δrank—the change in the ground-truth

Models	Rec@1	Rec@3	Rec@5	Rec@10
CHATGPT VANILLA	2.6	3.5	8.9	15.7
CHATGPT WITH ORACLE	11.1	18.3	26.6	53.0
CHATGPT WITH ORACLE†	28.7	64.7	85.3	94.7
OUR MODEL	6.1	10.7	22.2	28.0

Table 6: Results (%) on ReDial dataset. CHATGPT WITH ORACLE ensures ground truth items are in the candidate list, while CHATGPT WITH ORACLE† includes all items mentioned in the conversation.

item's ranking in the final recommendation list when causal intervention is applied versus when it is not. Using the recommendation module, we recorded the top-50 recommended items for each sample, both with and without causal intervention, and calculated Δrank for the ground-truth item.

Figure 4 illustrates the results for both datasets, with the x-axis showing the original rank (without intervention) and the y-axis showing Δrank. A positive Δrank indicates improved ranking due to causal intervention. The trends in both datasets are similar: the area above the x-axis (positive Δrank) is larger and darker than below it, demonstrating that causal intervention generally enhances item rankings.

Analysis of Causal Intervention Via KL Divergence

As discussed in Section 3, the causal loss and KL divergence loss (Equation 8) are in a trade-off relationship. The causal loss pushes the representation after causal intervention to deviate from the original, while the KL divergence loss seeks to maintain similarity between them. A higher KL divergence value indicates a weaker trade-off, suggesting that confounding factors significantly impact the target item. To analyze KL divergence trends across instances with varying difficulty (using the ground-truth item's original rank as a proxy), we plot KL divergence values over different ranking ranges on ReDial in Figure 4(c). Figure 4(c) reveals the following: 1) KL divergence increases for more challenging instances (lower original rankings), confirming its correlation with poorer performance. 2) Instances with performance declines after causal intervention ("Rank -" category) exhibit higher KL divergence than those with improvements ("Rank +"), indicating that KL divergence effectively reflects recommendation performance.

Effectiveness of LLM Re-Ranking

We further investigate the effectiveness of LLMs in re-ranking recommendation lists with varying quality of candidate items. The results presented in Table 6 illustrate the performance of different models in terms of recommendation accuracy at various cutoffs (Rec@1, Rec@3, Rec@5, and Rec@10).

Notably, the CHATGPT WITH ORACLE setting significantly enhances performance across all metrics compared to CHATGPT VANILLA. This improvement indicates that the quality of the candidate list plays a critical role in re-ranking. Our model demonstrates a moderate performance improvement, restricted by the candidate list produced by the recommendation module. This suggests that while our model is capable of generating relevant recommendations, there is still room for enhancement, especially when compared to the CHATGPT WITH ORACLE† setting, which yields the highest performance. The results underscore the importance of candidate quality in maximizing the potential of LLMs for effective recommendation systems.

Conversation History	
Seeker:	I have seen The Conjuring (2013) I very much enjoyed this movie. When I was younger I liked movies like Friday the 13th (1980).
Rec:	I liked that one too!
Seeker:	I also liked to watch the A Nightmare on Elm Street (1984) Series.
Rec:	I loved that as well mainly because Johnny Depp is in it
Seeker:	I love Johnny Depp! So handsome, such a great actor. I love johnny Depp in Alice in Wonderland (2010).
Model Output	
Ground:	Me too one of my favorites from him is The Astronaut's Wife (1999)
ReDial:	i have n't seen that one yet . i 've seen It (2017) .
KBRD:	i have n't seen that one . i will check it out .
KGSF:	I hope you enjoy them !
CTA-CRS:	Yes, I love Johnny Depp too ! Have you seen the new It (2017) movie?
Ours:	Since you love Johnny Depp, I highly recommend you check out Sweeney Todd: The Demon Barber of Fleet Street (2007). In this film, Depp delivers a captivating performance in a gothic, musical thriller, which I believe aligns well with your taste for his unique characters and dramatic storytelling.
	Candidate list from recommendation module: "Alice in Wonderland (2010)", "Charlie and the Chocolate Factory (2005)", "Dark Shadows (2012)", "Edward Scissorhands (1990)", "Pirates of the Caribbean ", "Public Enemies (2009)", "Chocolat (2000)", "Sweeney Todd: The Demon Barber of Fleet Street (2007)", "The Grudge (2004)", ...
	Reranked list from LLM: "Sweeney Todd: The Demon Barber of Fleet Street (2007)", "Alice in Wonderland (2010)", "Charlie and the Chocolate Factory (2005)", "Dark Shadows (2012)", ...
	Explanation: I suggested Sweeney Todd: The Demon Barber of Fleet Street (2007) because it stars Johnny Depp, whom you admire, and it combines his acting talent with a dark, thrilling storyline that aligns with your love for horror and suspense films.

Table 7: An example on *movie item* recommendation from ReDial.

5.4 Case Study

We analyze a conversation example from the ReDial dataset (Table 7) to demonstrate the effectiveness of our method. In this interaction, the **Seeker** expresses enjoyment for movies featuring Johnny Depp, highlighting preferences for horror and suspense genres. like *A Nightmare on Elm Street (1984)*. The seeker explicitly states admiration for Johnny Depp, especially in *Alice in Wonderland (2010)*. The model outputs reveal how different recommender systems respond to the conversation.

The **Ground Truth** response suggests *The Astronaut's Wife (1999)*, aligning well with the seeker's stated preference. In contrast, the **ReDial** model misses the connection with Johnny Depp, recommending *It (2017)*, not aligning with the seeker's preferences. **KBRD** and **KGSF** provide generic responses lacking substantial recommendations. Although **CTA-CRS** acknowledges Johnny Depp, it still pivots to recommending *It (2017)*. Our approach yields a more relevant recommendation: *Sweeney Todd: The Demon Barber of Fleet Street (2007)*. This suggestion not only matches the seeker's admiration for Johnny Depp but also aligns with their interest in darker, suspenseful films. The response is also more informative, providing the rationale behind the recommendation.

6 Conclusion

We propose a hybrid framework for Conversational Recommender Systems that combines large language models with causal intervention. Our approach addresses domain-specific tuning limits and confounding factors in user interactions. A deconfounded recommendation module enhances personalization through causal methods, while an LLM-based response generator ensures fluent, scalable dialogue without retraining.

Acknowledgments

The research described in this paper is partially supported by National Key Research and Development Program of China under Grant 2023YFB3107000 and Shenzhen Science and Technology Program (Grant No.ZDSYS20210623091809029).

References

[Aliannejadi *et al.*, 2019] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference*, pages 475–484, 2019.

[Bao *et al.*, 2020] Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*, 2020.

[Bareinboim and Pearl, 2012] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108. PMLR, 2012.

[Besserve *et al.*, 2018] Michel Besserve, Arash Mehrjou, Rémy Sun, and Bernhard Schölkopf. Counterfactuals uncover the modular structure of deep generative models. *arXiv preprint arXiv:1812.03253*, 2018.

[Chen *et al.*, 2019] Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. Towards knowledge-based recommender dialog system. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1803–1813, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Christakopoulou *et al.*, 2016] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD*, pages 815–824, 2016.

[Christakopoulou *et al.*, 2018] Konstantina Christakopoulou, Alex Beutel, Rui Li, Sagar Jain, and Ed H Chi. Q&r: A two-stage approach toward interactive recommendation. In *Proceedings of the 24th ACM SIGKDD*, pages 139–148, 2018.

[Dai *et al.*, 2023] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. Uncovering chatgpt's capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1126–1132, 2023.

[Daiber *et al.*, 2013] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th international conference on semantic systems*, pages 121–124, 2013.

[Di Palma, 2023] Dario Di Palma. Retrieval-augmented recommender system: Enhancing recommender systems with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1369–1373, 2023.

[Geng *et al.*, 2022] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 299–315, 2022.

[Hua *et al.*, 2023] Wenyue Hua, Yingqiang Ge, Shuyuan Xu, Jianchao Ji, and Yongfeng Zhang. Up5: Unbiased foundation model for fairness-aware recommendation. *arXiv preprint arXiv:2305.12090*, 2023.

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

[Li *et al.*, 2010] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th WWW*, pages 661–670, 2010.

[Li *et al.*, 2018] Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. Towards deep conversational recommendations. *arXiv preprint arXiv:1812.07617*, 2018.

[Li *et al.*, 2020] Shijun Li, Wenqiang Lei, Qingyun Wu, Xiangnan He, Peng Jiang, and Tat-Seng Chua. Seamlessly unifying attributes and items: Conversational recommendation for cold-start users. *arXiv preprint arXiv:2005.12979*, 2020.

[Li *et al.*, 2024] Chuang Li, Yang Deng, Hengchang Hu, Min-Yen Kan, and Haizhou Li. Incorporating external knowledge and goal guidance for llm-based conversational recommender systems. *arXiv preprint arXiv:2405.01868*, 2024.

[Liu *et al.*, 2023] Yuanxing Liu, Weinan Zhang, Yifan Chen, Yuchi Zhang, Haopeng Bai, Fan Feng, Hengbin Cui, Yongbin Li, and Wanxiang Che. Conversational recommender system and large language model are made for each other in e-commerce pre-sales dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9587–9605, 2023.

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

[Ma *et al.*, 2020] Wenchang Ma, Ryuichi Takanobu, Minghao Tu, and Minlie Huang. Bridging the gap between conversational reasoning and interactive recommendation. *arXiv preprint arXiv:2010.10333*, 2020.

[Moon *et al.*, 2019] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th ACL*, pages 845–854, 2019.

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

[Parascandolo *et al.*, 2018] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pages 4036–4044. PMLR, 2018.

[Parthasarathi and Pineau, 2018] Prasanna Parthasarathi and Joelle Pineau. Extending neural generative conversational model using external knowledge sources. *arXiv preprint arXiv:1809.05524*, 2018.

[Pearl and others, 2009] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[Pearl *et al.*, 2016] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Schlichtkrull *et al.*, 2018] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer, 2018.

[Spurlock *et al.*, 2024] Kyle Dylan Spurlock, Cagla Acun, Esin Saka, and Olfa Nasraoui. Chatgpt for conversational recommendation: Refining recommendations by reprompting with feedback. *arXiv preprint arXiv:2401.03605*, 2024.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Wang *et al.*, 2022] Lingzhi Wang, Huang Hu, Lei Sha, Can Xu, Daxin Jiang, and Kam-Fai Wong. Recindial: A unified framework for conversational recommendation with pretrained language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 489–500, 2022.

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

[Wang *et al.*, 2023b] Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. Recmind: Large language model powered agent for recommendation. *arXiv preprint arXiv:2308.14296*, 2023.

[Wang *et al.*, 2024] Lingzhi Wang, Shafiq Joty, Wei Gao, Xingshan Zeng, and Kam-Fai Wong. Improving conversational recommender system via contextual and time-aware modeling with less domain-specific knowledge. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[Xi *et al.*, 2023] Yunjia Xi, Weiwen Liu, Jianghao Lin, Xiaoling Cai, Hong Zhu, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, Rui Zhang, et al. Towards open-world recommendation with knowledge augmentation from large language models. *arXiv preprint arXiv:2306.10933*, 2023.

[Xu *et al.*, 2021] Kerui Xu, Jingxuan Yang, Jun Xu, Sheng Gao, Jun Guo, and Ji-Rong Wen. Adapting user preference to online feedback in multi-round conversational recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 364–372, 2021.

[Yao *et al.*, 2020] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *arXiv preprint arXiv:2002.02770*, 2020.

[Young *et al.*,] T Young, E Cambria, I Chaturvedi, M Huang, H Zhou, and S Biswas. Augmenting end-to-end dialog systems with commonsense knowledge (2017). *arXiv preprint arXiv:1709.05453*.

[Zhang *et al.*, 2018] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 177–186, 2018.

[Zhang *et al.*, 2020] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, 2020.

[Zhou *et al.*, 2020] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD*, pages 1006–1014, 2020.

[Zhou *et al.*, 2022] Yuanhang Zhou, Kun Zhou, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and He Hu. C2-crs: Coarse-to-fine contrastive learning for conversational recommender system. *arXiv preprint arXiv:2201.02732*, 2022.