Broadening the View: Demonstration-augmented Prompt Learning for Conversational Recommendation

Huy Dao
Singapore Management University
Singapore
qh.dao.2023@phdcs.smu.edu.sg

Yang Deng
National University of Singapore
Singapore
ydeng@nus.edu.sg

Dung D. Le
VinUniversity, Hanoi
Vietnam
dung.ld@vinuni.edu.vn

Lizi Liao
Singapore Management University
Singapore
lzliao@smu.edu.sg

ABSTRACT

Conversational Recommender Systems (CRSs) leverage natural language dialogues to provide tailored recommendations. Traditional methods in this field primarily focus on extracting user preferences from isolated dialogues. It often yields responses with a limited perspective, confined to the scope of individual conversations. Recognizing the potential in collective dialogue examples, our research proposes an expanded approach for CRS models, utilizing selective analogues from dialogue histories and responses to enrich both generation and recommendation processes. This introduces significant research challenges, including: (1) How to secure high-quality collections of recommendation dialogue exemplars? (2) How to effectively leverage these exemplars to enhance CRS models?

To tackle these challenges, we introduce a novel Demonstration-enhanced Conversational Recommender System (DCRS), which aims to strengthen its understanding on the given dialogue contexts by retrieving and learning from demonstrations. In particular, we first propose a knowledge-aware contrastive learning method that adeptly taps into the mentioned entities and the dialogue's contextual essence for pretraining the demonstration retriever. Subsequently, we further develop two adaptive demonstration-augmented prompt learning approaches, involving contextualized prompt learning and knowledge-enriched prompt learning, to bridge the gap between the retrieved demonstrations and the two end tasks of CRS, *i.e.*, response generation and item recommendation, respectively. Rigorous evaluations on two established benchmark datasets underscore DCRS's superior performance over existing CRS methods in both item recommendation and response generation [1].

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Discourse, dialogue and pragmatics**; Intelligent agents.

KEYWORDS

conversational recommendation, demonstration-based learning

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657755

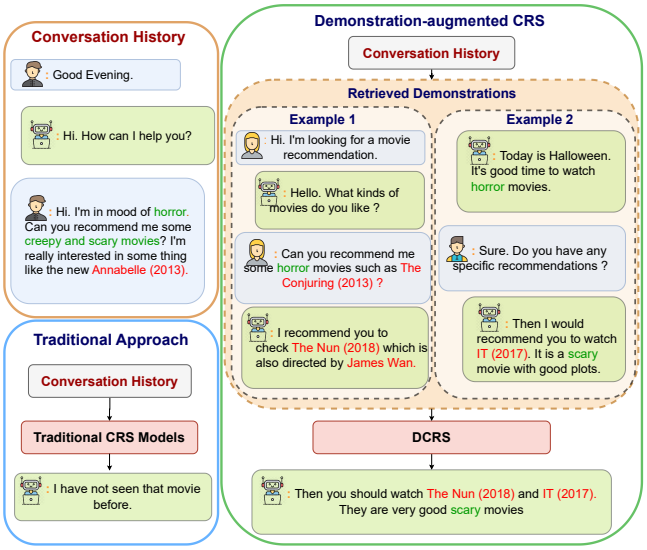


Figure 1: Traditional CRS models function over individual dialogue sessions. On the other hand, our proposed demonstration-augmented approach expands the perspective of the model via a set of collective exemplars.

ACM Reference Format:
Huy Dao, Yang Deng, Dung D. Le, and Lizi Liao. 2024. Broadening the View: Demonstration-augmented Prompt Learning for Conversational Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3626772.3657755

1 INTRODUCTION

The domain of Conversational Recommender Systems (CRSs) has seen a notable evolution in recent years, driven by the integration of natural language processing and user preference analysis [6, 14, 23, 29]. Existing CRS models [13, 25, 34, 36, 50, 61] primarily focus on analyzing user preferences within isolated dialogue contexts. This approach, while effective in certain scenarios, often yields responses that are confined to the limited perspective of individual

conversations, thereby missing the richness of collective dialogue experiences, as illustrated in the lower left-hand part of Figure 1.

To overcome these limitations, some CRS models incorporate external knowledge sources, such as knowledge graphs [1, 43], user reviews [33, 60] or item meta information [52]. These efforts demonstrate a broadening of the CRS scope. However, the dependence on external data presents its own set of challenges, particularly in domains where such data is scarce or costly to acquire. Moreover, a crucial issue persists in the current methods: the struggle to effectively utilize internal training instances to enhance model performance. An initial attempt to tackle this, as seen in [30], focused on retrieving related entities from similar dialogues, but this approach primarily centered on explicit entity mentions, thus neglecting the wider dialogue context and its potential value.

In parallel, the blooming field of demonstration-based and in-context learning offers promising avenues for enhancing language model performance across various tasks, as shown in studies by [3, 4, 18, 27]. This approach involves presenting pre-trained language models with a small number of informative examples, leading to notable improvements in tasks like named entity recognition [18] and text classification [3] etc. Nonetheless, directly applying these approaches to the nuanced and multifaceted sphere of conversational recommendation has been challenging. Two primary hurdles emerge: (1) The acquisition of high-quality recommendation dialogue exemplars, which requires careful curation and validation to ensure their relevance and effectiveness. (2) There's the challenge of how these exemplars can be best employed to improve the performance of CRS models. It not only involves selecting the right exemplars but also determining the optimal way to leverage these exemplars for maximum impact.

To address these challenges, we propose a novel **D**emonstration-enhanced **C**onversational **R**ecommender **S**ystem (**DCRS**), which facilitates an enriched understanding of dialogue contexts by retrieving and learning from demonstrations. DCRS first employs a knowledge-aware contrastive learning method, which adeptly taps into the mentioned entities and the contextual essence of dialogues. This method serves as a foundation for pretraining the demonstration retriever, aligning with our first challenge of securing high-quality dialogue exemplars. Further addressing the second challenge, DCRS implements two innovative demonstration-augmented prompt learning approaches: contextualized prompt learning and knowledge-enriched prompt learning. The former tailors the prompts to the specific nuances of each dialogue for more relevant and engaging response generation, while the latter enriches these prompts with entity-specific knowledge to improve recommendation relevance. These approaches bridge the gap between the retrieved demonstrations and the dual end-goals of CRS, namely response generation and item recommendation, enhancing both recommendation relevance and response quality. Our rigorous evaluations on two established benchmark datasets – ReDial [23] and INSPIRED [11] – highlight the effectiveness of the proposed DCRS. The results demonstrate that our system outperforms existing CRS methods in crucial aspects like recommendation accuracy and response quality.

[1]Code and data are available here: https://github.com/huyquangdao/DCRS.

To the best of our knowledge, this is the initial application of a demonstration-based learning framework within the context of CRS. Our contributions are three-folds:

- We introduce DCRS, a novel demonstration-based method for conversational recommendation. It is armed with a knowledge-aware contrastive retriever that integrates knowledge entities with the dialogue's intrinsic context for demonstration retrieval.
- We introduce two schemes for prompt learning enhanced by demonstrations: contextualized prompt learning and knowledge-enriched prompt learning. They dynamically employ gathered demonstrations to bolster both item recommendation and response generation.
- We empirically validate DCRS's superiority on two benchmark datasets over existing CRS frameworks, emphasizing both recommendation precision and linguistic nuance.

2 RELATED WORK

Conversational Recommender Systems. Recent advancements in CRS can be categorized into two distinct classes: recommendation-centric approaches [5, 19, 20, 36] and dialogue-driven CRS techniques [2, 26, 48, 60]. The former paradigm inherently gravitated towards seeking clarifications on item attributes, progressively refining the candidate item set. In contrast, the latter paradigm placed emphasis not only on the precision of recommendations but also on the caliber of generated natural language expressions.

Recommendation-centric CRS methods [7, 44, 55] mainly focused on directly improving the performance of item recommendation, where they aimed to ask clarifying questions about the item attributes and gradually find an optimal candidate set according to the user's preference. To reduce the difficulty of understanding natural language utterances, they leveraged predefined templates to interact with the users. Such systems attempted to learn recommendation strategies and fulfill users' demands while avoiding lengthy conversations since such long-lasting dialogues could hurt the user's experience. Therefore, these works usually utilized reinforcement learning (RL) [7, 19, 20, 40] or bandit-based solutions [24], which could effectively maximize long-term advantages.

In recent years, dialogue-driven CRS models [13, 26] have been investigated more extensively. Despite their promising performance, published dialog-based CRS datasets such as [11, 23, 59], confronted an inherent challenge of data scarcity stemming from the high costs of human annotations. Therefore, an increasing number of efforts [28, 33, 46, 52, 54, 58] enhanced these datasets by using external knowledge resources. For instance, Zhou et al. [58] sought to harness commonsense correlations between entities and terms through sub-graphs derived from DBpedia [1] and ConceptNet [43]. Diverging in approach, Lu et al. [33] incorporated user reviews to enrich the content garnered from user dialogues. Similarly, Yang et al. [52] leveraged rich meta information to better represent items in the database. However, a discernible limitation within these methodologies is their reliance solely on the immediate dialogue context containing limited clues on the user preferences. Moreover, over-emphasizing external resources might restrict the generalization abilities of such approaches to domains where external knowledge is either rare or costly to obtain.

To address such a limitation, we aim to extract informative exemplars from training data and leverage these retrieved exemplars to improve both response quality and recommendation accuracy.

Retrieval-augmented Generation and Recommendation. Recently, the retrieval-augmented generation paradigm (RAG) [22] has attracted increasing attention. At its core, a generative model will integrate data retrieval into the generation process, enhancing its ability to provide accurate and relevant responses. While some pre-trained encoders [8, 15, 39] have been commonly utilized for dense retrieval, existing RAG methods [38, 47, 51] often emphasized more on developing effective strategies for leveraging retrieved candidates. Among these methods, there are two common approaches, namely (1) leveraging a non-parametric integration (e.g., K nearest neighbors) [16] and (2) directly augmenting the current input with retrieved candidates [10, 47, 53]. For example, [16] introduced KNN-LM, which combined two probabilities for predicting the next token, one computed via a Transformer Decoder and the other established via the retrieval results with the KNN algorithm. Similarly, Wu et al. [51] incorporated external key and value vectors into intermediate hidden representations via a KNN-augmented attention layer. On a different angle, Izacard and Grave [12] proposed a RAG model based on encoder-decoder architecture for open-domain question-answering. Specifically, given a set of retrieved passages, they utilized the encoder to produce their intermediate representations. These sequences of hidden vectors are then fused at the decoder part to generate the answer. Wang et al. [47] and Ram et al. [38] prepended retrieved segments of texts to the current input of an LM model. However, concatenating retrieved candidates with the current input prompt might significantly increase the whole sequence's length, raising considerable computational costs. At the same time, it is challenging to fully leverage the retrieved set due to the maximal number of tokens accepted by existing pre-trained models [8, 56]. Last but not least, due to the difference in modality, how to use retrieved documents to improve ranking problems such as item recommendation is still unexplored.

To this end, in this work, we aim to introduce an effective technique for leveraging retrieved candidates. In particular, our proposed method incorporates retrieved demonstrations into both response generation and item recommendation processes via a contextualized prompt learning and a knowledge-enriched prompt-learning, respectively.

Contrastive Learning for Document Representations. Recently, contrastive self-supervised learning [45] has been commonly utilized to produce sentence representations. The intuition is to pull semantically similar samples close and keep dissimilar samples apart. Existing efforts [31, 39, 49] combined general-purpose pretrained language models such as BERT [8] with self-supervised contrastive learning objectives [35**?**] to produce high-quality document embeddings. **?**] introduced a lightweight data augmentation method that utilizes a dropout mask to produce positive examples for training an encoder model. Liu et al. [31] proposed to capture fine-grained relevance between query and target sentences with a ranking system. Moreover, they introduced a method that incorporates ranking consistency and ranking distillation with contrastive learning into a unified framework. Despite their effectiveness, these approaches above did not explicitly consider knowledgeable information, such as entities within the sentences. Nishikawa et al. [35] proposed an entity-ware contrastive learning framework for sentence embeddings. However, their method only considered one entity for each sentence. In this work, we propose a knowledge-aware contrastive learning method for pretraining a dialogue retriever, which can distinguish multiple entity instances in a dialogue session and then integrate the underlying user preference information into dialogue representations for conversational recommendation.

3 PLELIMINARIES

Notations. We denote by $\mathcal{I}, \mathcal{V}$ the set of all items and the generation vocabulary, respectively. Besides, we use d_c, d_e to define the dimensions of token and entity representations in our framework. We denote by $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$ the set of all training examples where N is the total number of instances. Each tuple (X_i, Y_i) consists of a historical context X_i and its corresponding groundtruth response Y_i. Broadly speaking, the goal of CRS models is to recommend appropriate items to the users via natural language dialogues. In particular, given a specific dialogue context X, CRS methods aim to produce a generated response $\hat{Y}$ to manage the conversation with the users. If the recommendation action is triggered, the models additionally recommend a set of candidate items $\mathcal{I}_c \in \mathcal{I}$ ($\mathcal{I}_c \neq \emptyset$) based on user's preferences extracted from the dialogue X. In contrast to existing works that perform these two aforementioned processes solely based on the given context X, we additionally offer the model a collection of informative demonstrations $\mathcal{R} = \{(\bar{X}_j, \bar{Y}_j)\}_{j=1}^{K}$ (K is the size of the collection) to enhance its ability to understand the tasks at hand. The demonstration retrieval step is conducted by a neural text retriever, which is pre-trained with a novel knowledge-aware contrastive learning method. Finally, following modern CRS approaches, we utilize an item-oriented knowledge graph $\mathcal{G} = \{(u, r, v)\}$ (where u, r, v are the head entity, relation, and tail entity respectively) to capture commonsense relationships between items and their associated entities. Formally, we decompose the CRS task into three sub-tasks:
Demonstration Retrieval: For each historical dialogue X_i, we retrieve a collection of instructive demonstrations $\mathcal{R}_i = \{(\bar{X}_j, \bar{Y}_j)\}_{j=1}^{K}$ (K is the size of the collection) from the training corpus by using a neural text retriever. It is worth noticing that we do not consider retrieved candidates that belong to the same conversation with the given context to avoid the data leakage problem.
Response Generation: Given the historical context X_i and the corresponding set of retrieved demonstrations $\mathcal{R}_i$, we aim to produce a natural language response $\hat{Y}_i$ where specific items in the response will be replaced by a special token [MASK].
Item Recommendation: Given the historical context X_i, the corresponding set of retrieved demonstrations $\mathcal{R}_i$ and the masked response template $\hat{Y}_i$, we attempt to predict a set of candidate items $\mathcal{I}_c \in \mathcal{I}$ to recommend to the user. In following sections, we ignore conversation indexes of the training examples for simplicity.

4 METHODOLOGY

Figure 2 illustrates the proposed DCRS framework. Overall, our DCRS method consists of three components, namely a text retriever module, which is pre-trained by a knowledge-aware contrastive

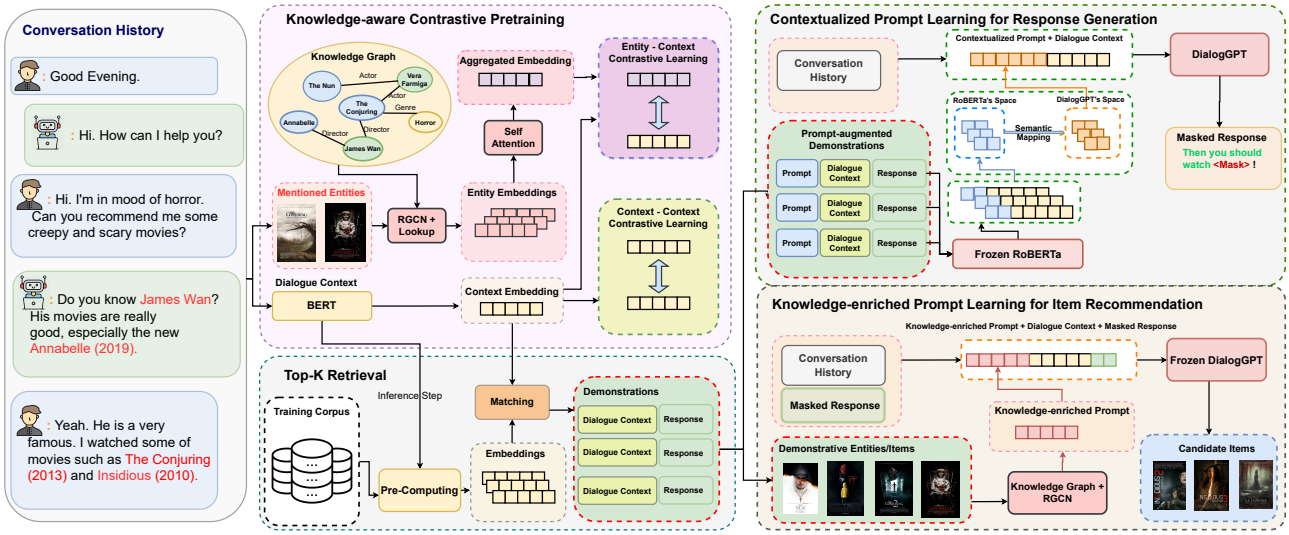


Figure 2: The architecture of the proposed DCRS model, which consists of three main modules: a knowledge-aware contrastive pretraining enhanced text retriever module for soliciting demonstrations, a response generation module augmented by contextualized prompt learning, and an item recommendation module augmented by knowledge-enriched prompt learning.

pretraining method for soliciting demonstrations, a response generation module, and an item recommendation module. In what follows, we describe each of the mentioned components in detail.

4.1 Knowledge-aware Contrastive Pretraining for Demonstration Retrieval

In conversational recommendation scenarios, users often articulate their preferences through specific entities or descriptive terms. As highlighted in Figure 2, the user specifies a liking for *scary and creepy* movies and mentions past experiences with films like *The Conjuring (2013)* and *Insidious (2010)*. Recognizing the importance of such entity-based evidence, it's crucial to seamlessly integrate them into the retrieval process. Additionally, the broader dialogue context can provide even richer insights into user inclinations. To tap into this depth, we utilize a knowledge-aware contrastive learning scheme. With these insights in mind, we introduce our learning method for retrieval from the following aspects.

Entity Modeling. To model representations of knowledge entities, similar to existing works [58, 60], we adopt RGCN model [42] and the item-oriented knowledge graph $\mathcal{G}$. Formally, at the l-th layer, we compute entity embeddings by the following:

$$\mathbf{e}_u^{(l)} = \text{RELU}\left(\sum_{r\in\mathcal{R}}\sum_{v\in\mathcal{N}_r(u)} \mathbf{W}_r^{(l)}\mathbf{e}_v^{(l-1)} + \mathbf{b}_r^{(l)}\right), \quad (1)$$

where $\mathbf{e}_u^{(l)} \in \mathbb{R}^{d_e}$ is the embedding vector of an entity u, $\mathbf{W}_r^{(l)} \in \mathbb{R}^{d_e\times d_e}$, $\mathbf{b}_r^{(l)} \in \mathbb{R}^{d_e}$ are model parameters corresponding to a specfic relation r at the l-th layer. Since a dialogue context X might contain multiple mentioned entities, we aim to obtain a single knowledge-aware entity context representation $\mathbf{e}_c \in \mathbb{R}^{d_e}$ by combining the latent vectors of all mentioned entities in the dialogue with a self-attention layer defined as follows:

$$\alpha = \text{Softmax}\left(\mathbf{b}_\alpha^{\mathrm{T}}\tanh\left(\mathbf{W}_\alpha\mathbf{E}\right)\right), \quad \mathbf{e}_c = \mathbf{E}\alpha^{\mathrm{T}}, \quad (2)$$

where $\mathbf{W}_\alpha \in \mathbb{R}^{d_e\times d_e}$, $\mathbf{b}_\alpha \in \mathbb{R}^{d_e}$ are model parameters, n_e is the number of entities mentioned in the context X, and $\mathbf{E} \in \mathbb{R}^{d_e\times n_e}$ is the entity embedding matrix.

Context Encoding. In addition to knowledge entities, the conversation context X itself contains rich contextual features, which is also crucial for the retrieval process. To leverage the semantic meaning of the current dialogue session, we produce a dense representation by utilizing a contextual encoder. In this work, we adopt the bidirectional Transformer architecture BERT [8] as our encoding model. In particular, to obtain the dialogue representation $\mathbf{h}_c \in \mathbb{R}^{d_c}$, we feed the whole conversational history X through the encoder and utilize the output embedding of the [CLS] token.

Knowledge-aware Contrastive Pretraining. To train our retrieval component, we introduce a novel knowledge-aware contrastive learning paradigm. The key idea is to learn meaningful, dense dialogue representations that also emphasize information from the mentioned knowledge entities. Hence, we propose to optimize two objective functions, namely entity-context and context-context contrastive losses. With the former, we aim to maximize the agreement between the mentioned knowledge entities and the corresponding dialogue context. Specifically, we attempt to maximize the alignment score between the entity context representation $\mathbf{e}_c$ and dialogue context representation $\mathbf{h}_c$ by optimizing the following:

$$L_{c,e} = -\log\frac{\exp(sim(\mathbf{h}_c, \mathbf{W}_c\mathbf{e}_c)/\rho)}{\sum_{(\mathbf{h}_c,\mathbf{e}_c,\hat{\mathbf{e}}_c)}\exp(sim(\mathbf{h}_c, \mathbf{W}_c\hat{\mathbf{e}}_c)/\rho)}, \quad (3)$$

where $sim()$ refers to cosine similarity, $\mathbf{W}_c \in \mathbb{R}^{d_c\times d_e}$ is a linear transformation that aligns two representation spaces and ρ is the temperature parameter. In this work, we obtain negative examples $\hat{\mathbf{e}}_c$ via within-batch negative sampling. For context-context contrastive loss, we aim to learn meaningful contextual representations via self-supervised signals with data augmentation. Specifically, we minimize a loss function to pull the dialogue representations and their augmented views together in their corresponding latent space.

In particular, the context-context objective function is defined as:

$$L_{c,c} = -\log \frac{\exp(sim(\mathbf{h}_c, \mathbf{h}_c^+)/\rho)}{\sum_{(\mathbf{h}_c, \mathbf{h}_c^+, \mathbf{h}_c^-)} \exp(sim(\mathbf{h}_c, \mathbf{h}_c^-)/\rho)} , \tag{4}$$

where $\mathbf{h}_c^-$ is the negative instance. Following [**?**], we adopt different drop-out masks to produce augmented views $\mathbf{h}_c^+$ of the dialogue representation. The final objective function is a weighted combination of the two aforementioned ones and is defined as follows:

$$L_{rev} = \gamma L_{c,e} + \delta L_{c,c} , \tag{5}$$

where γ, δ are predefined hyper-parameters and chosen via cross-validation. We optimize all parameters with this final objective.

Top-K Retrieval of Demonstrations. We utilize the learned contextual encoder to pre-compute dialogue representations for all training dialogue contexts. Specifically, we utilize the training examples $\mathcal{D} = \{D_i\}_{i=1}^N$ to produce a set of context representations $\mathcal{H} = \{\mathbf{h}_i\}_{i=1}^N$. Given a dialogue session X, we obtain its dialogue embedding $\mathbf{h}_c$ and compute top-K retrieval scores as follows:

$$\mathcal{S}_K = \text{Top-K}\left(\{\mathbf{h}_c^{\mathrm{T}}\mathbf{h}_i\}_{i=1}^N\right). \tag{6}$$

With the top-K computed scores $\mathcal{S}_K$, we then obtain corresponding contexts and responses $\mathcal{R} = \{(\bar{X}_j, \bar{Y}_j)\}_{j=1}^K$ from the training corpus.

4.2 Contextualized Prompt Learning with Demonstrations for Generation

For response generation, simply concatenating demonstrations with the current dialogue history likely results in a lengthy input. Due to limited context length, such an approach hinders the capabilities of the model in fully exploiting retrieved exemplars. Hence, we propose to extract semantic information from demonstrations with a set of prompt tokens via a pre-trained encoder. The prompt vectors are then mapped to the input space of a pre-trained decoder to enrich the current dialogue context for generating responses.

Prompting Contextual Information from Demonstrations. We aim to elicit useful information, such as commonsense knowledge or task-specific instructions from the retrieved demonstrations with a set of contextualized prompts. Specifically, for the j-th demonstration $(\bar{X}_j, \bar{Y}_j)$ in the retrieved set $\mathcal{R}$, we first prepend a sequence of prompt tokens $P_j = [p_{1,j}, p_{2,j}, ..., p_{T,j}]$ (T is the number of prompt tokens) to each retrieved demonstration as follows:

$$I_j = [P_j, \bar{X}_j, \bar{Y}_j] .$$

Afterward, to obtain demonstration-enhanced continuous prompts, we feed the constructed sequence through a prompt generator f_{prompt} based on a pre-trained bidirectional Transformer architecture, which is formulated as follows:

$$\mathbf{P}_j^c = [\mathbf{p}_{1,j}^c, \mathbf{p}_{2,j}^c, ..., \mathbf{p}_{T,j}^c] = f_{\text{prompt}}(I_j) ,$$

where $\mathbf{P}_j^c \in \mathbb{R}^{T \times d_c}$ is our contextualized prompt embeddings. We instance f_{prompt} with a RoBERTa encoder [32], and collect the output vectors from the positions corresponding prompt tokens P_j. For efficiency, we freeze the parameters of the prompt generator during training. We independently employ the same process for K different retrieved demonstrations $\{(\bar{X}_j, \bar{Y}_j)\}_{j=1}^K$ to obtain corresponding contextualized prompts $\mathbf{P}_1^c, \mathbf{P}_2^c, ..., \mathbf{P}_K^c$. We then concatenate these continuous tokens to produce a single prompt sequence $\mathbf{P}^c \in \mathbb{R}^{K*T \times d_c}$. Notably, the length of the final prompt sequence $\mathbf{P}^c$ is $K * T$ (in practice, we can choose the value of T so that $K * T \ll K * L_d$ where L_d is the length of each demonstration).

Semantic Space Mapping for Contextualized Prompts. When generating responses, suppose $\mathbf{X}$ is the matrix of embedding vectors after forwarding the dialogue context X through the input layer of the generation model f_{gen}, one can notice that there is a semantic gap between retrieval-augmented prompt vectors $\mathbf{P}^c$ and input embeddings $\mathbf{X}$ since $\mathbf{P}^c$ are embedding vectors in the prompt generator's output space, while $\mathbf{X}$ belongs to text generator's input vector space $\mathbf{E}_{gen}$. To handle such a semantic gap, we propose to map the contextualized prompts into the input space of the generative model as follows:

$$\mathbf{S} = \mathbf{P}^c \mathbf{W}_{align} \mathbf{E}_{gen}^{\mathrm{T}} ,$$
$$\mathbf{P}_{gen}^c = \text{Softmax}(\mathbf{S}) \cdot \mathbf{E}_{gen} ,$$

where $\mathbf{P}_{gen}^c \in \mathbb{R}^{K*T \times d_c}$ are corresponding prompt embeddings in the target space, $\mathbf{E}_{gen} \in \mathbb{R}^{|\mathcal{V}| \times d_c}$ are the embedding matrix of the input layer of the generation model f_{gen} and $\mathbf{W}_{align} \in \mathbb{R}^{d_c \times d_c}$ is a linear transformation that aligns these two representation spaces. Then we prepend the prompt $\mathbf{P}_{gen}^c$ to the input embeddings $\mathbf{X}$ to obtain augmented input sequence $\mathbf{I}_{gen}$ (i.e $\mathbf{I}_{gen} = [\mathbf{P}_{gen}^c, \mathbf{X}]$) which is subsequently used to generate the desired response $\hat{Y}$.

Parameters Learning. Given the current dialogue embeddings $\mathbf{X}$ and the retrieval-augmented prompts $\mathbf{P}_{gen}^c$, we train the generation model f_{gen} by optimizing the following objective function:

$$L_{gen}(\Theta_{gen}) = -\sum_{t=1}^{N} \sum_{j=1}^{L_o} \log \mathbf{Pr}_{\text{gen}}\left(y_{t,j}^* | y_{t,<j}, \mathbf{P}_{gen}^c, \mathbf{X}_i\right) ,$$

where N is the number of training examples, L_o is the length of the output sequence, Θ_{gen} are parameters of the generation module.

4.3 Knowledge-enriched Prompt Learning with Demonstrations for Recommendation

In this subsection, we describe how to leverage the retrieved demonstrations to enhance the item recommendation task.

Demonstrations for Item Recommendation. To solve the recommendation task, existing works [48, 54, 58] often leveraged mentioned entities/items in the current input, which we denote by $\mathcal{E}_{men} = \{e_i\}_{i=1}^{|\mathcal{E}_{men}|}$, to capture user preferences. However, due to the limited information in the current context, $\mathcal{E}_{men}$ might not be sufficient to properly model the user interests. While retrieved dialogue history-response pairs can be regarded as instructive exemplars for response generation, these pieces of texts might contain a collection of pertinent entities/items that can be viewed as informative clues to hint at possible candidates for item recommendation, which we refer to as item demonstrations. Hence, we propose to extract knowledge entities and items from these retrieved demonstrations, which we denote by $\mathcal{E}_{dem} = \left\{e_j\right\}_{j=1}^{|\mathcal{E}_{dem}|}$, to enrich the mentioned set $\mathcal{E}_{men}$. In particular, the final set of entities/items $\mathcal{E}_{rec}$ is an union of the mentioned set $\mathcal{E}_{men}$ and demonstration set $\mathcal{E}_{dem}$ (i.e $\mathcal{E}_{rec} = \mathcal{E}_{men} \bigcup \mathcal{E}_{dem}$).

Knowledge-enriched Prompts for Item Recommendation. Similar to response generation, we produce a set of prompt tokens to enrich the input for the recommendation task. Specifically, we first look up the latent vectors of entities in $\mathcal{E}_{rec}$ using RGCN

and the item-oriented knowledge graph $\mathcal{G}$ to obtain an embedding matrix $\mathbf{P}^k_{rec} \in \mathbb{R}^{|\mathcal{E}_{rec}| \times d_c}$. It is worth noticing that we use two different sets of entity embeddings for retrieval and recommendation, respectively. Afterward, we construct the input $\mathbf{I}_{rec}$ for the recommendation task as follows:

$$\mathbf{I}_{rec} = [\mathbf{P}^k_{rec}, \mathbf{X}, \hat{\mathbf{Y}}] \,,$$

where $\mathbf{X}, \hat{\mathbf{Y}}$ are corresponding embeddings of the current dialogue context X and the masked response $\hat{Y}$ generated from the response generation module. We aim to feed the constructed input through a frozen DialogGPT [56] and apply a pooling layer on the outputs to obtain a demonstration-enhanced preference vector $\mathbf{u}_{rec} \in \mathbb{R}^{d_c}$.

Semantic Alignment for Knowledge-enriched Prompts. Similar to response generation, there is a natural semantic gap between the knowledge-enriched prompt $\mathbf{P}^k_{rec}$ and context, response embeddings $\mathbf{X}, \hat{\mathbf{Y}}$ in the item recommendation task. To address such a problem, we propose to associate the knowledge-enriched prompt $\mathbf{P}^k_{rec}$ with contextual information in $\mathbf{X}, \hat{\mathbf{Y}}$ via another pretraining step. Specifically, with the demonstration enhanced user preference representation $\mathbf{u}_{rec}$, we pre-train the item recommendation module to predict the entities contained in the current input. Formally, we compute the probability of an entity e as follows:

$$\mathbf{Pr}_{entity}(e) = \text{Softmax}\left(\mathbf{u}^{\mathrm{T}}_{rec}\mathbf{E}_{entity}\right)_e \,, \tag{7}$$

where $\mathbf{E}_{entity} \in \mathbb{R}^{d_c \times N_{entity}}$ is the embedding matrix of all entities. Finally, we optimize model parameters to minimize a standard cross-entropy loss of ground-truth entities defined as follows:

$$L_{pre}(\Theta_{rec}) = -\sum_{j=i}^{N} \sum_{e \in \mathcal{N}(X_j)} \log \mathbf{Pr}^j_{entity}\left(e|\mathbf{P}^k_{rec,j}, \mathbf{X}_j, \hat{\mathbf{Y}}_j\right), \tag{8}$$

where N is the total number of training instances and $\mathcal{N}(X_j)$ is the set of mentioned entities in the input context X_j. $\Theta(rec)$ are parameters of the recommendation engine.

Parameters Learning for Item Recommendation. After the aforementioned semantic mapping step, given the demonstration enhanced user preference vector $\mathbf{u}_{rec}$, we compute the probability of recommending an item $\mathbf{Pr}_{item}(i)$ similar to E.q 7. Then, we train the recommendation engine by optimizing the cross-entropy loss of ground-truth items defined below.

$$L_{rec}(\Theta_{rec}) = -\sum_{j=1}^{N} \sum_{i \in \mathcal{V}} y^j_{rec,i} \log \mathbf{Pr}^j_{item}\left(i|\mathbf{P}^k_{rec,j}, \mathbf{X}_j, \hat{\mathbf{Y}}_j\right), \tag{9}$$

where $y^j_{rec,i}$ is the corresponding label of the item i-th in the item set $\mathcal{V}$ at the j-th training instance. During recommendation training, we freeze the parameters of the DialogGPT model.

5 EXPERIMENTS

5.1 Experimental Setup

5.1.1 Datasets. We conduct extensive experiments on two prominent datasets: **ReDial** [23] and **INSPIRED** [11]. These datasets are centered around movie recommendations, with ReDial being curated through Amazon Mechanical Turk, while INSPIRED focuses on sociable recommendation strategies for persuasive outcomes. Table 2 presents detailed statistics for these datasets.

5.1.2 Baseline methods. We compare with four groups of baselines: (1) Traditional recommendation approaches:

- **Item Populalarity**: A simple baseline that ranks the items according to historical recommendation frequencies in the corpus.
- **TextCNN** [17]: A basic convolutional neural network for text classification.

(2) General pre-trained language models:

- **BERT** [8]: A widely used pre-trained model for text classification tasks. We train BERT and use it to predict a set of candidate items.
- **GPT2** [37]: is a basic but strong text generation baseline that gains from large pre-trained language modeling.
- **DialogGPT** [56]: is a dialogue generative pre-trained GPT model trained on large-scale conversation-like exchanges from Reddit.
- **BART** [21]: is a more recent denoising autoencoder pretrained model for language generation.

(3) State-of-the-art CRS methods:

- **ReDial** [23]: An early CRS model adopts an auto-encoder for recommendation and hierarchical RNN for generation.
- **KBRD** [2]: An knowledge-enhanced CRS method that adopts a sub-graph from DBpedia to enrich extracted features.
- **KGSF** [58]: A CRS model that leverages an item-oriented and a word-oriented knowledge graph to improve its capabilities.
- **UNICRS** [48]: A CRS model unifies recommendation and generation tasks with a unified prompt tuning method.
- **TREA** [26]: A CRS model that adopts tree-structure reasoning to solve item recommendation and response generation tasks.
- **COLA** [30]: A CRS model that adopts BM25 and mentioned entities to retrieve collaborative entities.
- **VRICR** [54]: A CRS whose recommendation module is pretrained with a variational Bayesian method.

(4) Different variants of our model, namely: **DCRS Concat** - a simple version using the concatenation of retrieved demonstrations and current inputs. **DCRS w/ BM25, w/ Rand** which are variants equipped BM25 [41] and random selection as the retrieval modules respectively. For ablation study, we also report the results of different variants including: **DCRS w/o ID** - our DCRS without item/entities extracted from retrieved demonstrations (i.e w/o $\mathcal{E}_{dem}$). **DCRS w/o SMR** - our model without the semantic mapping step for response generation. **DCRS w/o SAI** - our variant without the semantic alignment step for item recommendation. **DCRS w/o CP** - our model without contextualized prompts for generation (i.e. w/o $\mathbf{P}^c_{gen}$). **DCRS w/o KP** - our model without knowledge-enriched prompts for recommendation (i.e. w/o $\mathbf{P}^k_{rec}$).

5.1.3 Evaluation Metrics: In this work, we report the performance of both item recommendation and response generation sub-tasks. For *recommendation task*, we report several metrics including **Recall@k** (k=1,10,50), **NDCG@k** (k=10,50) and **MRR@k** (k=10,50). For *response generation*, we conduct both automatic and human evaluation. For automatic assessment, we report **BLEU-N** (N=2,3), **ROUGE-N** (N=2,L) and **Distinct-N** (N=2,3,4). For human study, we randomly sample twenty generated conversations from each model. We then invite two annotators and ask them to score the generated responses. We report the results on three aspects including **Informativeness** and **Fluency**, whose range is from 1 to 3.

Table 1: Automatic evaluation results on the item recommendation task (t-test with p-value < 0.05).

Model	ReDial							INSPIRED						
	Recall			NDCG		MRR		Recall			NDCG		MRR	
	@1	@10	@50	@10	@50	@10	@50	@1	@10	@50	@10	@50	@10	@50
Popularity	0.011	0.053	0.183	0.029	0.057	0.021	0.027	0.031	0.155	0.322	0.085	0.122	0.064	0.071
TextCNN [17]	0.010	0.066	0.187	0.033	0.059	0.023	0.028	0.025	0.119	0.245	0.066	0.094	0.050	0.056
BERT [8]	0.027	0.142	0.307	0.075	0.112	0.055	0.063	0.049	0.189	0.322	0.112	0.141	0.088	0.095
Redial [23]	0.010	0.065	0.182	0.034	0.059	0.024	0.029	0.009	0.048	0.213	0.023	0.059	0.015	0.023
KBRD [2]	0.033	0.150	0.311	0.083	0.118	0.062	0.070	0.042	0.135	0.236	0.088	0.109	0.073	0.077
KGSF [58]	0.035	0.175	0.367	0.094	0.137	0.070	0.079	0.051	0.132	0.239	0.092	0.114	0.079	0.083
TREA [26]	0.045	0.204	0.403	0.114	0.158	0.087	0.096	0.047	0.146	0.347	0.095	0.132	0.076	0.087
COLA [30]	0.048	0.221	0.426	-	-	0.086	0.096	-	-	-	-	-	-	-
VRICR [54]	0.054	0.244	0.406	0.138	0.174	0.106	0.114	0.043	0.141	0.336	0.091	0.134	0.075	0.085
UNICRS [48]	0.065	0.241	0.423	0.143	0.183	0.113	0.121	0.085	**0.230**	0.398	0.149	0.187	0.125	0.133
DCRS	**0.076**	**0.253**	**0.439**	**0.154**	**0.195**	**0.123**	**0.132**	**0.093**	0.226	**0.414**	**0.153**	**0.192**	**0.130**	**0.137**

Table 2: Statistics of the ReDial and INSPIRED datasets.

ReDial		INSPIRED	
# of convs	10,006	# of convs	1,001
# of utterances	182,150	# of utterances	35,811
# of words/utterance	14.5	# of words/utterance	19.0
# of entities/items	64,364/6924	# of items	17321/1,123
# of users	956	# of users	1,482

5.1.4 Implementation Details: We train the framework on 1 GPU NVIDIA A100 40G card. For each model, we run experiments three times with different random seeds and compute the averaged results. In this work, we use the DialogGPT-small (114M) and RoBERTa-base (114M) as our generation model and prompt generator, respectively. We empirically set the number of demonstrations and prompt length to 3 and 50, respectively. For the RGCN model, we set the number of GNN layers to 1. For demonstration retrieval, we set the dimensions of entity and context representations to 128 and 768, respectively. We empirically set the values of γ, δ to 1.0, 0.1 respectively. We use a learning rate of 1e-5 to train the retrieval module. We train the generation model with a learning rate of 1e-4 with 10 epochs. For each model, we set the maximum number of tokens to 400. We utilize the same data split for every model. For COLA [30], since its source code is not published and is not directly adaptable to the INSPIRED dataset, we hence report the official performance in the paper. For other baselines, we leverage the CRS-Lab toolkit [57] [2] to reproduce their results.

5.2 Main Results

5.2.1 Automatic Evaluation on Item Recommendation. We show the results of the item recommendation task in Table 1. First, our proposed DCRS model exhibits outstanding performance across all metrics for both ReDial (achieving the best result in all 7 metrics) and INSPIRED (achieving the best result in 6 out of 7 metrics). This remarkable improvement over the strong baseline UNICRS can be observed in terms of Recall@1 (+16.9% for ReDial, +14.4% for INSPIRED), Recall@10 (+4.97 % for ReDial), and Recall@50 (+3.78% for ReDial, +4.20% for INSPIRED). The superiority of DCRS can be attributed to 2 reasons: (1) Its ability to provide reliable evidence to the recommendation module, specifically through the inclusion of demonstrative entities and items extracted from highly relevant dialogues. (2) The proposed retrieval-augmented prompt learning method can effectively incorporate informative demonstrations to improve the item recommendation performance. Moreover, compared to COLA [30], which is another retrieval-enhanced CRS model, our DCRS consistently achieves better results on all metrics. A potential explanation could be that COLA's retrieval component solely relies on the BM25 [41] and mentioned entities while neglecting rich contextual information of the given dialogue, which limits its ability to fully exploit the provided information for the retrieval process. At the same time, knowledge-enhanced CRS models such as KBRD, KGSF, VRICR, and UNICRS achieve better performance than Redial. This is expected since these methods leverage external knowledge graphs to enhance their recommendation module. However, these CRS models only rely on the current context to make predictions, which might not be sufficient to capture users' interests properly. Therefore, their performance is inferior compared to our DCRS model, which can effectively leverage additional clues/evidence from collections of informative exemplars.

5.2.2 Automatic Evaluation on Response Generation. We show the results of the response generation task in Table 3. Noticeably, our DCRS model consistently outperforms all baseline methods across two published benchmarks. Specifically, compared to the most competitive method, UNICRS, our DCRS shows considerable improvements on DIST-2 (+79.9% for Redial, +47.1 % for INSPIRED), DIST-3 (+56.8% for ReDial, +31.9 % for INSPIRED) and Dist-4 (+38.1% for ReDial, + 12.9 % for INSPIRED). Such substantial improvement could be attributed to the proposed contextualized prompt learning that captures contextual information of both input-output correlations and task-specific instructions from a collection of retrieved demonstrations. At the same time, we noticed that general language models, including GPT2, DialogGPT, and BART, show commendable performance in this task. This observation is expected given the extensive pre-training of these generative models on vast volumes

[2] https://github.com/RUCAIBox/CRSLab

Table 3: Automatic evaluation results on the response generation task (t-test with p-value < 0.05).

Model	ReDial							INSPIRED						
	BLEU		ROUGE		DIST			BLEU		ROUGE		DIST		
	-2	-3	-2	-L	-2	-3	-4	-2	-3	-2	-L	-2	-3	-4
DialogGPT [56]	0.041	0.021	0.054	0.258	0.436	0.632	0.771	0.031	0.014	0.041	0.207	1.954	2.750	3.235
GPT2 [37]	0.031	0.013	0.041	0.244	0.405	0.603	0.757	0.026	0.011	0.034	0.212	2.119	3.084	3.643
BART [21]	0.024	0.011	0.031	0.229	0.432	0.615	0.705	0.018	0.008	0.025	0.208	1.920	2.501	2.670
Redial [23]	0.004	0.001	0.021	0.187	0.058	0.204	0.442	0.001	0.000	0.004	0.168	0.359	1.043	1.760
KBRD [2]	0.038	0.018	0.047	0.237	0.070	0.288	0.488	0.021	0.007	0.029	0.218	0.416	1.375	2.320
KGSF [58]	0.030	0.012	0.039	0.244	0.061	0.278	0.515	0.023	0.007	0.031	0.228	0.418	1.496	2.790
COLA [30]	0.026	0.012	-	-	0.387	0.528	0.625	-	-	-	-	-	-	-
VRICR [54]	0.021	0.008	0.027	0.137	0.107	0.286	0.471	0.011	0.001	0.025	0.187	0.853	1.801	2.827
TREA [26]	0.022	0.008	0.039	0.175	0.242	0.615	1.176	0.013	0.002	0.027	0.195	0.958	2.565	3.411
UNICRS [48]	0.045	0.021	0.058	0.285	0.433	0.748	1.003	0.022	0.009	0.029	0.212	2.686	4.343	5.520
DCRS	**0.048**	**0.024**	**0.063**	**0.285**	**0.779**	**1.173**	**1.386**	**0.033**	**0.014**	**0.045**	**0.229**	**3.950**	**5.729**	**6.233**

of unstructured text data. UNICRS surpasses other CRS methods due to its utilization of a generation module established with DialogGPT, which is effectively learned with a knowledge-enhanced prompt learning technique.

5.2.3 Human Evaluation. We conducted a thorough human evaluation with two annotators on 20 randomly sampled dialogues to gain deeper insights into the performance of all the models. The results are presented in Table 4 with kappa scores over 0.7, which indicates a very high level of agreement among annotators. First, UNICRS, which unified both general pre-trained language models and knowledge graphs via a prefix tuning method, shows the most competitive performance compared to other CRS models, such as VRICR, KGSF, and KBRD. Notably, on both metrics *Fluency* and *Inform*, our DCRS outperforms all baseline methods, demonstrating that our model produces more coherent and meaningful responses. We hypothesize that solely generating responses with dialogue histories might result in safe and less diverse responses due to limited contextual information. Our DCRS alleviates this problem by leveraging demonstrations to enhance current dialogue contexts.

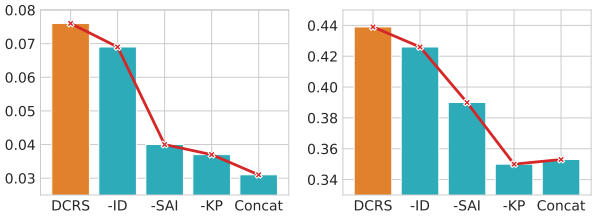
5.3 In-depth Analyses and Discussions

5.3.1 Ablation Study on Item Recommendation. In Figure 3 (a) and (b), we show the performance of our DCRS model without item demonstrations (- ID), semantic alignment step for demonstrations (- SAI), and knowledge-enriched prompts (- KP), respectively. First, removing either retrieved demonstrative entities/items (i.e., DCRS w/o ID) or the semantic mapping step (i.e., DCRS w/o SMI) results in significant drops in performance, which indicates the effectiveness of our designs. This can be attributed to two factors: (1) Indicative demonstrations serve as explicit cues that enable the model to generate more precise recommendations, and (2) with the semantic alignment step, the model manages to alleviate the natural semantic gap by associating representations of entities/items with contextual information produced by the DialogGPT model, which improves the quality of learned embeddings. Second, we also compare DCRS with its variant using a simple concatenation of demonstrations and the input context (DCRS w/ Concat). It indicates that our DCRS performance is significantly better, which demonstrates our proposed designs are indeed more effective than simply concatenating retrieved exemplars with the current dialogue context.

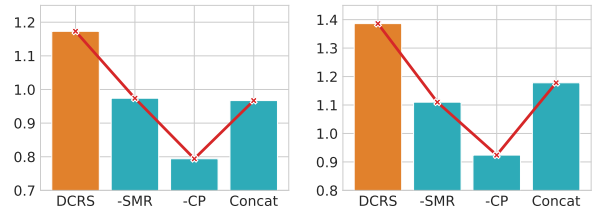
Table 4: Human evaluation results about the conversation task on the ReDial dataset. κ denotes Fleiss' Kappa [9], indicating substantial agreement ($0.61 < \kappa < 0.8$).

Model	Fluency	Inform	κ
KBRD	2.32	1.97	0.70
KGSF	2.46	2.05	0.73
VRICR	2.37	2.17	0.76
UNICRS	2.71	2.53	0.75
DCRS	**2.80**	**2.65**	0.78

5.3.2 Ablation Study on Response Generation. In Figure 3 (c) and (d), we show the performance of our DCRS model without semantic mapping for response demonstration (-SMR) and contextualized prompts (-CP), respectively. Overall, when we remove the semantic mapping (i.e., DCRS w/o SMR) and contextualized prompts (i.e., DCRS w/o CP), the performance of DCRS decreases significantly. This can be explained by the following: (1) Semantic mapping serves to mitigate the disparity between the semantic spaces of the prompt generator and the text generation model. Consequently, this facilitates quicker and more effective convergence of the training process. (2) The contextualized semantic prompts offer valuable information, such as input-output correlations or contextual semantics of retrieved exemplars, which enriches dialogue context representations. Moreover, our DCRS performance is still better than the simple concatenation version, which demonstrates the effectiveness of our proposed designs. The reason is simply combining retrieved demonstrations, and the current context could easily exceed the maximal number of tokens (e.g., 512 in the case of DialogGPT), which limits the capability of DCRS to fully exploit retrieved exemplars. Nevertheless, the concatenation version still manages to achieve surprisingly good results, which indicates that incorporating demonstrations could improve the generation task.

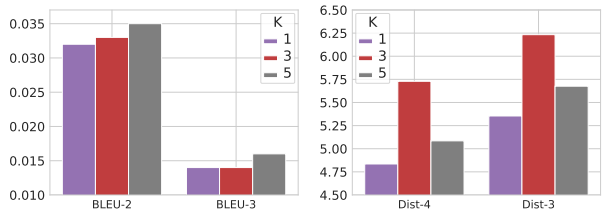


(a) Recall@1 (b) Recall@50

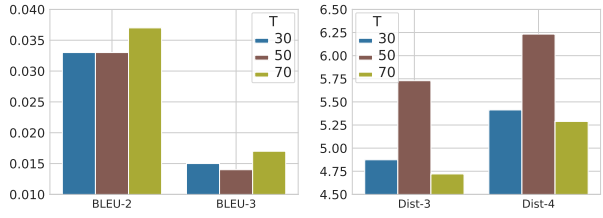


(c) Dist-3 (d) Dist-4

Figure 3: Ablation study on item recommendation and response generation. The results are reported on the Redial dataset (t-test with p-value < 0.05).



(a) BLEU-2,3 (T = 50) (b) DIST-3,4 (T = 50)



(c) BLEU-2,3 (K = 3) (d) Dist-3,4 (K = 3)

Figure 4: Analysis on the prompt length (T) and number of demonstrations (K) respectively (t-test with p-value < 0.05).

Table 5: Performance comparison of DCRS with different retrieval methods (t-test with p-value < 0.05).

	Model	Recall@1	Recall@50	DIST@3	DIST@4
ReDial	DCRS	**0.076**	**0.439**	**1.173**	**1.386**
	- *w/ BM25*	0.071	0.428	0.981	1.118
	- *w/ Rand*	0.069	0.426	0.975	1.115
INSPIRED	DCRS	**0.093**	**0.414**	**5.729**	**6.233**
	- *w/ BM25*	0.078	0.390	4.901	5.460
	- *w/ Rand*	0.074	0.382	4.785	5.339

5.3.3 Performance with Different Retrieval Methods. To demonstrate the effectiveness of our retrieval algorithm, we carry out an experiment in which we replace the DCRS's retrieval part with either the BM25 algorithm or a random retrieval module. As can be seen in Table 5, the results significantly drop when we substitute DCRS's retrieval part with either BM25 or a random retriever. This can be attributed to the inability of these methods to retrieve suitable demonstrations, thereby leading to a notable deterioration in recommendation performance. By integrating knowledge entities and the contextual meaning of the dialogue context into the retrieval process, our proposed method has the potential to generate demonstrative entities/items that are more pertinent compared to the BM25 algorithm and random methods.

5.3.4 Analyses on prompt length T and number of demonstrations K. To probe the influence of the prompt length (T) and the number of demonstrations (K), we also conducted comprehensive experiments on DCRS for the response generation task on the INSPIRED dataset. Figure 4 shows the results of DCRS across varying values of $T \in \{30, 50, 70\}$ and $K \in \{1, 3, 5\}$. In general, the results show that a higher number of demonstrations and longer prompt length yield higher BLEU and DIST scores (as the performance improved when we increased the number of demonstrations and prompt length). This seems reasonable since the larger the number of demonstrations, the richer contextual features the model can utilize. Longer sequences of prompts might offer better capability for compressing semantic information in the demonstrations. Despite improving the performance, it also incurs a significant increase in computational cost as the total length of contextualized prompts scales linearly with both the numbers of demonstrations K and each individual length T (i.e., $K * T$ in general). However, a too-long retrieval-augmented prompt could dominate the information coming from the current dialogue context and make the generated responses less diverse (as Dist metrics decrease when either T or K increases too much) since different instances might share the same set of retrieved demonstrations.

6 CONCLUSION

In this work, we proposed a novel demonstration-based conversational recommendation framework, namely DCRS, which employed a knowledge-aware contrastive retriever to collect selective analogues from dialogue histories to enrich both response generation and recommendation processes. Then we introduced two adaptive demonstration-augmented prompt learning methods for bridging the gap between the retrieved examplar and the ongoing conversational recommendation tasks. Experimental results on two benchmark datasets demonstrated the superiority of the proposed DCRS framework over existing CRS methods, exemplifying advances in both recommendation precision and linguistic coherence.

ACKNOWLEDGMENTS

This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (Proposal ID: 23-SIS-SMU-010).

REFERENCES

[1] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia-a crystallization point for the web of data. *Journal of web semantics* (2009).

[2] Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards Knowledge-Based Recommender Dialog System. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 1803–1813.

[3] Xiang Chen, Lei Li, Ningyu Zhang, Xiaozhuan Liang, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Decoupling Knowledge from Memorization: Retrieval-augmented Prompt Learning. In *Advances in Neural Information Processing Systems*.

[4] Jishnu Chowdhury, Yong Zhuang, and Shuyi Wang. 2022. Novelty Controlled Paraphrase Generation with Retrieval Augmented Conditional Prompt Tuning. *Proceedings of the AAAI Conference on Artificial Intelligence* 36 (06 2022), 10535–10544.

[5] Zhendong Chu, Hongning Wang, Yun Xiao, Bo Long, and Lingfei Wu. 2023. Meta Policy Learning for Cold-Start Conversational Recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 222–230.

[6] Huy Dao, Lizi Liao, Dung Le, and Yuxiang Nie. 2023. Reinforced Target-driven Conversational Promotion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 12583–12596.

[7] Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. Unified Conversational Recommendation Policy Learning via Graph-based Reinforcement Learning. In *ACM SIGIR Conference on Research and Development in Information Retrieval*. 1431–1441.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.

[9] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76 (1971), 378–382.

[10] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 3816–3830.

[11] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward Sociable Recommendation Dialog Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 8142–8152.

[12] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. arXiv:2007.01282 [cs.CL]

[13] Yeongseo Jung, Eunseo Jung, and Lei Chen. 2023. Towards a Unified Conversational Recommendation System: Multi-task Learning via Contextualized Knowledge Distillation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 13625–13637.

[14] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul Crook, Y-Lan Boureau, and Jason Weston. 2019. Recommendation as a Communication Game: Self-Supervised Bot-Play for Goal-oriented Dialogue. 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)*. 1951–1961.

[15] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 6769–6781.

[16] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations (ICLR)*.

[17] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1746–1751.

[18] Dong-Ho Lee, Akshen Kadakia, Kangmin Tan, Mahak Agarwal, Xinyu Feng, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, and Xiang Ren. 2022. Good Examples Make A Faster Learner: Simple Demonstration-based Learning for Low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2687–2700.

[19] Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. Estimation-Action-Reflection: Towards Deep Interaction Between Conversational and Recommender Systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 304–312.

[20] Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020. Interactive Path Reasoning on Graph for Conversational Recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2073–2083.

[21] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.

[22] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A Survey on Retrieval-Augmented Text Generation. arXiv:2202.01110 [cs.CL]

[23] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Advances in Neural Information Processing Systems*.

[24] Shijun Li, Wenqiang Lei, Qingyun Wu, Xiangnan He, Peng Jiang, and Tat-Seng Chua. 2021. Seamlessly Unifying Attributes and Items: Conversational Recommendation for Cold-Start Users. *ACM Transactions on Information Systems (TOIS)* (2021).

[25] Shuokai Li, Ruobing Xie, Yongchun Zhu, Xiang Ao, Fuzhen Zhuang, and Qing He. 2022. User-Centric Conversational Recommendation with Multi-Aspect User Modeling. In *Proceedings of the 45nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[26] Wendi Li, Wei Wei, Xiaoye Qu, Xian-Ling Mao, Ye Yuan, Wenfeng Xie, and Dangyang Chen. 2023. TREA: Tree-Structure Reasoning Schema for Conversational Recommendation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2970–2982.

[27] Jinggui Liang, Lizi Liao, Hao Fei, Bobo Li, and Jing Jiang. 2024. Actively Learn from LLMs with Uncertainty Propagation for Generalized Category Discovery. In *NAACL-HLT*.

[28] Zujie Liang, Huang Hu, Can Xu, Jian Miao, Yingying He, Yining Chen, Xiubo Geng, Fan Liang, and Daxin Jiang. 2021. Learning Neural Templates for Recommender Dialogue System. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 7821–7833.

[29] Lizi Liao, Ryuichi Takanobu, Yunshan Ma, Xun Yang, Minlie Huang, and Tat-Seng Chua. 2020. Topic-guided conversational recommender in multiple domains. *IEEE Transactions on Knowledge and Data Engineering* 34, 5 (2020), 2485–2496.

[30] Dongding Lin, Jian Wang, and Wenjie Li. 2023. Cola: Improving conversational recommender systems by collaborative augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 4462–4470.

[31] Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. 2023. RankCSE: Unsupervised Sentence Representations Learning via Learning to Rank. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 13785–13802.

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

[33] Yu Lu, Junwei Bao, Yan Song, Zichen Ma, Shuguang Cui, Youzheng Wu, and Xiaodong He. 2021. RevCore: Review-Augmented Conversational Recommendation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 1161–1173.

[34] Wenchang Ma, Ryuichi Takanobu, and Minlie Huang. 2021. CR-Walker: Tree-Structured Graph Reasoning and Dialog Acts for Conversational Recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 1839–1851.

[35] Sosuke Nishikawa, Ryokan Ri, Ikuya Yamada, Yoshimasa Tsuruoka, and Isao Echizen. 2022. EASE: Entity-Aware Contrastive Learning of Sentence Embedding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3870–3885.

[36] Mingjie Qian, Yongsen Zheng, Jinghui Qin, and Liang Lin. 2023. HutCRS: Hierarchical User-Interest Tracking for Conversational Recommender System. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 10281–10290.

[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[38] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics* 11 (2023), 1316–1331.

[39] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*. 3982–3992.

[40] Xuhui Ren, Hongzhi Yin, Tong Chen, Hao Wang, Zi Huang, and Kai Zheng. 2021. Learning to ask appropriate questions in conversational recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 808–817.

[41] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. 3, 4 (2009), 333–389.

[42] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2017. Modeling Relational Data with Graph Convolutional Networks.

[43] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *ASSOCIATION FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE*.

[44] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*. 235–244.

[45] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748 [cs.LG]

[46] Lingzhi Wang, Huang Hu, Lei Sha, Can Xu, Daxin Jiang, and Kam-Fai Wong. 2022. 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)*. 489–500.

[47] Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Training Data is More Valuable than You Think: A Simple and Effective Method by Retrieving from Training Data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 3170–3179.

[48] Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Towards Unified Conversational Recommender Systems via Knowledge-Enhanced Prompt Learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1929–1937.

[49] Xing Wu, Chaochen Gao, Zijia Lin, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. InfoCSE: Information-aggregated Contrastive Learning of Sentence Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 3060–3070.

[50] Yuxia Wu, Lizi Liao, Gangyi Zhang, Wenqiang Lei, Guoshuai Zhao, Xueming Qian, and Tat-Seng Chua. 2022. State graph reasoning for multimodal conversational recommendation. *IEEE Transactions on Multimedia* (2022).

[51] Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. 2022. Memorizing Transformers. In *International Conference on Learning Representations*.

[52] Bowen Yang, Cong Han, Yu Li, Lei Zuo, and Zhou Yu. 2022. Improving Conversational Recommendation Systems' Quality with Context-Aware Item Meta-Information. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 38–48.

[53] Chenchen Ye, Lizi Liao, Suyu Liu, and Tat-Seng Chua. 2022. Reflecting on experiences for response generation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 5265–5273.

[54] Xiaoyu Zhang, Xin Xin, Dongdong Li, Wenxuan Liu, Pengjie Ren, Zhumin Chen, Jun Ma, and Zhaochun Ren. 2023. Variational Reasoning over Incomplete Knowledge Graphs for Conversational Recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 231–239.

[55] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 177–186.

[56] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 270–278.

[57] Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2021. CRSLab: An Open-Source Toolkit for Building Conversational Recommender System. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. 185–193.

[58] 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 *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1006–1014.

[59] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards Topic-Guided Conversational Recommender System. In *Proceedings of the 28th International Conference on Computational Linguistics*. 4128–4139.

[60] Yuanhang Zhou, Kun Zhou, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and He Hu. 2022. C^2-CRS: Coarse-to-Fine Contrastive Learning for Conversational Recommender System. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1488–1496.

[61] Jie Zou, Evangelos Kanoulas, Pengjie Ren, Zhaochun Ren, Aixin Sun, and Cheng Long. 2022. Improving Conversational Recommender Systems via Transformer-Based Sequential Modelling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. 2319–2324.