Model Meets Knowledge: Analyzing Knowledge Types for Conversational Recommender Systems

Jujia Zhao
Leiden University
Leiden, Netherlands
zhao.jujia.0913@gmail.com

Yumeng Wang
Leiden University
Leiden, Netherlands
y.wang@liacs.leidenuniv.nl

Zhaochun Ren
Leiden University
Leiden, Netherlands
z.ren@liacs.leidenuniv.nl

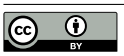
Suzan Verberne
Leiden University
Leiden, Netherlands
s.verberne@liacs.leidenuniv.nl

Abstract

Conversational Recommender Systems (CRSs) often integrate external knowledge to enhance user preference modeling and item representation learning, addressing the challenge of sparse conversational contexts. Traditional methods primarily utilize structured knowledge graphs (KGs) to model entity relationships and capture deep, multi-hop relationships among items. More recent studies employing pre-trained language models (PLMs), however, leverage unstructured text (e.g., customer reviews) to enrich contextual understanding of users and items. Despite reported performance gains from both knowledge types, a question remains: What is the compatibility between specific CRS model architectures and types of external knowledge, and how do different knowledge sources complement each other? We present a reproducibility study evaluating 9 state-of-the-art CRSs, including KG-based and PLM-based paradigms, to systematically investigate model—knowledge compatibility and complementarity. Through a comprehensive evaluation on three datasets, we uncover three key findings: (1) Different model architectures have different compatibility with knowledge types: decoder-only models excel with structured knowledge, whereas encoder-decoder models better utilize unstructured knowledge. (2) Combining multiple knowledge sources isn't always superior to using a single type, but merging similar knowledge types is generally more effective than mixing different ones. (3) Unstructured knowledge broadly benefits all scenario-specific conversations, particularly in genre-specific and descriptive scenarios, whereas structured knowledge demonstrates superior performance in comparative recommendation scenarios. Our study serves as an inspiration for future research on maximizing the benefits of external knowledge across different models in CRSs.

CCS Concepts

• **Information systems** → **Recommender systems**.



RecSys '25, Prague, Czech Republic

ACM ISBN 979-8-4007-1364-4/25/09
https://doi.org/10.1145/3705328.3748152

Keywords

Conversational Recommender System, Knowledge Graphs, Pre-trained Language Models

ACM Reference Format:
Jujia Zhao, Yumeng Wang, Zhaochun Ren, and Suzan Verberne. 2025. Model Meets Knowledge: Analyzing Knowledge Types for Conversational Recommender Systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3705328.3748152

1 Introduction

Conversational Recommender Systems (CRSs) have shown great promise by enabling interactive and personalized recommendations [22, 28]. CRS methods often incorporate external knowledge to enrich user preference modeling and item representation learning given the inherent sparsity of conversational contexts [12]. Early systems primarily integrated structured knowledge — such as knowledge graphs (KGs) — to enrich sparse conversation contexts and better understand complex dialogues [1]. With the advances in pre-trained language models (PLMs), CRS methods have increasingly incorporated PLMs. This has enabled the use of unstructured external knowledge, like user reviews and item descriptions [5, 6, 25]. Figure 1 illustrates the format of different types of external knowledge. This trend raises a question: What is the compatibility between specific CRS model architectures and different types of external knowledge?

The question stems from an observed divergence in knowledge utilization in CRSs: Traditional CRS methods, constrained by weaker textual understanding, relied on the explicit, structured relationships in KGs to infer user preferences (e.g., linking "Christopher Nolan" to "Inception" via director-entity edges) [38]. In contrast, modern PLM-based CRSs increasingly leverage unstructured knowledge (e.g., movie descriptions or user reviews) through natural language prompts, capitalizing on their pre-trained ability to contextualize free-text information [3]. The use of unstructured knowledge relies on the natural language understanding abilities of PLMs to derive semantic connections [27]. While both approaches claim performance gains, their comparative strengths remain unclear: Does PLMs' parametric knowledge reduce the need for structured KGs, or does each knowledge type uniquely address

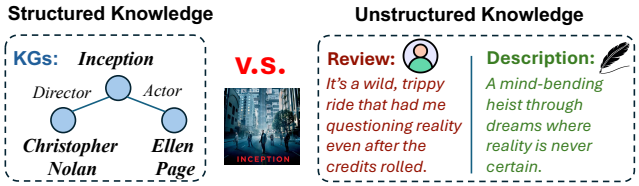


Figure 1: Format of different types of external knowledge.

specific model limitations? Moreover, how can different forms of knowledge effectively complement each other?

To systematically investigate this, we conduct a reproducibility study addressing 4 research questions: (i) **RQ1**: How do state-of-the-art (SOTA) CRS methods compare in recommendation accuracy and response quality? (ii) **RQ2**: Is there inherent compatibility between CRS models and specific types of external knowledge? If so, which knowledge types are most effective for which models? (iii) **RQ3**: How well can different knowledge types complement each other to improve CRS performance? (iv) **RQ4**: Under what conditions (e.g., genre-specific mentions, descriptive preferences) does each knowledge type provide maximal performance gains?

Our experiments provide the following key findings in response to the proposed research questions: (1) The BART-based PLM methods generally exhibit strong performance in both recommendation and response generation tasks, demonstrating robust generalization across all evaluated datasets. (2) Decoder-only models perform better with structured knowledge but struggle with unstructured knowledge, whereas encoder-decoder models show the opposite trend. (3) Combining different types of knowledge rarely outperforms using a single type, whereas integrating multiple sources of the same type tends to be more effective. (4) Descriptions generally benefit all scenario-specific conversations, reviews are particularly effective for genre- and description-based scenarios in certain models, and KGs are most useful in comparative scenarios. Beyond answering these research questions, we also uncovered additional interesting insights. For example, structured knowledge enhances response quality, while unstructured knowledge improves recommendation precision and response diversity. Additionally, we identified a phenomenon of "knowledge dominance" – some knowledge types strongly overshadow others in shaping performance scores.

Our contributions are as follows: (1) We reproduce 9 state-of-the-art methods, providing the first cross-paradigm comparison. (2) We systematically evaluate CRS compatibility with diverse external knowledge types and provide foundational insights into knowledge complementarity. (3) We analyze the role of each knowledge type in different conversation scenarios, which provides practical guidelines for scenario-specific knowledge integration[1].

2 Related Work

Conversational Recommender Systems (CRSs) aim to optimize recommendation accuracy and generate coherent, human-like responses [17, 29, 40]. They typically combine a recommendation module with a generation-based dialogue component [2, 18, 20]. Due to limited conversational context, user preference modeling remains challenging, motivating the use of external knowledge as auxiliary signals [6, 19, 30]. From the perspective of external knowledge types, CRSs are broadly categorized into those using structured knowledge (*i.e.,* KGs) and those leveraging unstructured knowledge (*e.g.,* reviews or item descriptions) [32, 33]. Here, external knowledge refers to supplementary inputs to the model beyond the historical conversational context. This section outlines key research directions leveraging these knowledge sources, providing the background for this reproducibility study.

Structured external knowledge. Early CRS methods leverage structured KGs to incorporate semantic information, [24, 36] but face a semantic gap between word-level dialogue context and entity-level KG representations. Effective CRSs require seamless integration of dialogue and recommendation modules with diverse knowledge sources. KBRD [1] facilitates inter-module communication by linking contextual entities to KGs and propagating vocabulary bias to the dialogue system, enhancing contextual understanding and generation. To bridge the semantic gap between natural language expressions and item-level user preferences, KGSF [38] integrates word-oriented and entity-oriented KGs, enriching data representation in CRSs. C2-CRS [40] proposes a coarse-to-fine contrastive learning framework to improve multi-type data semantic fusion. UniCRS [23] unifies recommendations and dialogue generation tasks into the prompt learning paradigm based on a fixed PLM. VRICR [34] addresses challenges of incompleteness, sparsity, and noise in KGs for CRSs. This study focuses on KGs as the sole form of structured external knowledge, reflecting their dominant use in existing research.

Unstructured external knowledge. Unstructured data, unlike structured data which follows a predefined format, is typically expressed in natural language format. Given the natural language understanding capabilities of PLMs, unstructured knowledge naturally suits PLM-based approaches [4, 21]. The unstructured knowledge used in CRS approaches can be categorized into item-specific and user-specific knowledge.

Research on *item-specific knowledge* often incorporates user reviews from conversational history. For instance, RevCore [13] enhances CRSs by extracting sentiment-consistent reviews, enabling review-augmented item recommendation and review-attentive response generation. Similarly, C2-CRS [40] employs contrastive learning to align representations from conversations, KGs, and reviews. Another line of research incorporates item descriptions, such as textual descriptions (e.g., movie titles, actors, directors, genres). MESE [27] introduces an item description encoder to learn semantically aligned item representations with dialogue context. In contrast, PECRS [16] first encodes item embeddings using a PLM and concatenates them with dialogue context for reranking. In this work, we consider both reviews and item descriptions as unstructured external knowledge, as they provide complementary item information for experimentation. In some applications, *user-specific knowledge*, such as historical item interactions or personal knowledge, are also available. UniMIND [3] prepends user profiles to the dialogue context for predicting the next conversation topic and recommending items. However, since not all datasets contain such user-specific information, we exclude the use of user-specific knowledge in our experiments and only include item-specific knowledge as unstructured external knowledge.

[1]We release our code at https://github.com/Polaris-JZ/CRS.

Several previous studies have explored reinforcement learning [31] and provided theoretical surveys of the advances in CRS [7, 14]. In contrast, our study analyzes the role of different types of knowledge and the compatibility between models and knowledge types, an aspect that has not been previously investigated in conversational recommender systems.

3 Problem Formulation

Typically, CRSs consist of two core tasks: a recommendation task to recommend items and a conversation task to formalize natural language response. These tasks are jointly optimized to fulfill user needs during interactions. We formalize the components of CRSs as follows:

Notations for CRSs. Let $\mathcal{U}$ and $\mathcal{I}$ denote user and item sets. A conversation up to turn t is a sequence $C_t = \{s_1, ..., s_t\}$ where each utterance $s_i \in \mathcal{S}$ is the natural language from the user or recommender system, and $\mathcal{S}$ denotes the universe of all possible utterances. The system maintains history $\mathcal{H}_t = (u, C_t, \mathcal{I}_t)$, where $u \in \mathcal{U}$ is the specific user, C_t is the dialogue history, and $\mathcal{I}_t \subseteq \mathcal{I}$ is the subset of candidate items dynamically filtered based on $\mathcal{H}_t$.

External Knowledge. The system may use two kinds of knowledge sources: (1) *Structured Knowledge* is represented as a knowledge graph $G = (N, R, E)$, where N denotes entities (items $\mathcal{I}$ and related concepts like directors/genres), R specifies relations between entities (*e.g., directedBy*, *hasGenre*), and $E \subseteq N \times R \times N$ contains factual triples. For example, the triple (*Inception*, *directedBy*, *Christopher Nolan*) encodes cinematic authorship. A user-specific subgraph $G_u \in G$ is dynamically constructed by extracting entities mentioned in the conversation history $\mathcal{H}_t$ and their connected triples. (2) *Unstructured Knowledge* typically comprises three textual sources: 1) *User Reviews*: Free-text feedback expressing subjective preferences (*e.g.,* "The dream sequences visually redefine sci-fi cinema"), and 2) *User Profiles*: Personal metadata indicating user preferences and demographics (*e.g.,* "User prefers sci-fi thrillers with intricate plots"), and 3) *Item Descriptions*: factual detailing attributes (*e.g.,* "A thriller exploring dream infiltration and subconscious security"). For each item $i \in \mathcal{I}$, let d_i denote its description and $\mathcal{R}_i = r_1, \ldots, r_m$ denote its set of review sentences. For each user $u \in \mathcal{U}$, denote the profile information as P_u.

Task Definition. Let Θ denote the model parameters. CRSs jointly optimize two tasks: (1)*Recommendation Task*: Given user u, the recommendation task leverages both the conversation history and external knowledge sources to predict a top-k set of items $\hat{\mathcal{I}}_{t+1} \subseteq \mathcal{I}$ for the next turn. The recommendation score f_{rec} for candidate item $i \in \mathcal{I}$ is computed through the following equation:

$$f_{\text{rec}}(i|\mathcal{H}_t, \mathcal{K}_{u,i}; \Theta) = \phi_{\text{conv}}(\mathcal{H}_t; \Theta) \oplus \phi_{\text{ext}}(\mathcal{K}_{u,i}; \Theta), \quad (1)$$

where $\phi_{\text{conv}}(\cdot)$ encodes the conversational context, $\phi_{\text{ext}}(\cdot)$ encodes the external knowledge $\mathcal{K}_{u,i}$ selected for user u and item i, $\mathcal{K}_{u,i} \subseteq \{G_i, \mathcal{T}_i, P_u\}$ denotes optional knowledge, $\oplus$ denotes method-specific fusion (e.g., concatenation, attention). The recommendation loss is then defined as:

$$\mathcal{L}_{\text{rec}} = -\sum_{i \in \mathcal{I}} [\, y_i \log \sigma(f_{\text{rec}}(i)) + (1 - y_i) \log(1 - \sigma(f_{\text{rec}}(i)))\,], \quad (2)$$

where $y_i \in \{0, 1\}$ indicates ground-truth relevance. (2) *Response Generation*: Given user u, this task generates the next utterance s_{t+1} by conditioning the conversation history, the current user u, the selected item i_{t+1}, and its associated external knowledge. Specifically, the probability of generating s_{t+1} is modeled as:

$$P(s_{t+1}|\mathcal{H}_t, i_{t+1}, \mathcal{K}_{u,i_{t+1}}; \Theta) = \prod_{j=1}^{|s_{t+1}|} P(w_j | w_{<j}, \mathcal{H}_t, \mathcal{K}_{u,i_{t+1}}; \Theta), \quad (3)$$

with the negative log-likelihood generation loss:

$$\mathcal{L}_{\text{gen}} = -\sum_{j=1}^{|s_{t+1}|} \log P(w_j | w_{<j}, \mathcal{H}_t, \mathcal{K}_{u,i_{t+1}}; \Theta). \quad (4)$$

The joint training loss function $\mathcal{L}_{\text{total}}$ is then defined as $\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{rec}} + (1 - \alpha)\mathcal{L}_{\text{gen}}$, where $\alpha \in [0, 1]$.

4 Experimental Setup

4.1 Data

We use three public datasets to evaluate the impact of each knowledge type.

Datasets. (1) **ReDial** [11] is an English-language collection of conversational recommendation sessions. The conversations are manually curated by crowdworkers on the Amazon Mechanical Turk platform, where users engage in recommending movies to one another. The dataset comprises 10,006 sessions, containing a total of 182,150 utterances and covering 51,699 unique movies. (2) **TG-ReDial** [39] is a Chinese conversational recommendation dataset. It features topic-guided conversations for natural transitions into recommendations, semi-automatic annotations for improved data quality, and detailed user profiles with interaction history to enhance personalization modeling. It consists of 10,000 sessions with 129,392 utterances covering 33,834 movies. (3) **INSPIRED** [8] is an English conversational recommendation dataset featuring two-paired crowdworkers in a natural setting, with annotations for sociable recommendation strategies, where each utterance is manually labeled based on social science theory. It contains 1,001 dialogues with 35,811 utterances related to 1,783 movies. Each dataset is randomly split into training, validation, and test datasets in a ratio of 8:1:1.

External Knowledge. We incorporate both structured and unstructured external knowledge for evaluation. For structured knowledge, following [37], we integrate KGs from DBpedia [9] and CN-DBpedia [26]. DBpedia is a multilingual knowledge graph constructed from Wikipedia articles, containing 4,828,418 entities and 18,746,176 relations. CN-DBpedia includes three popular Chinese knowledge sources—Chinese Wikipedia, Hudong Baike, and Baidu Baike—comprising 10,341,196 entities and 88,454,264 relations. We link these external knowledge sources to entities appearing in the dataset conversations. For unstructured knowledge, following [40], we extract item reviews from IMDB[2] and Douban[3], which provide reviews for all items appearing in the dataset. Additionally, we obtain item descriptions from the tmdbv3api library[4].

4.2 Methods reproduced

Traditional CRSs primarily rely on knowledge graphs to enrich semantic information beyond the dialogue context. Recent approaches have shifted the focus toward unifying recommendation

[2]https://www.dbpedia.org/.
[3]https://movie.douban.com/.
[4]https://github.com/AnthonyBloomer/tmdbv3api.

and response generation modules within a single PLM framework. In this work, we select representative baselines from each category by considering their citation impact and widespread adoption in recent literature, and provide detailed descriptions below.

KG-based methods. (1) **KBRD** [1] uses DBpedia to enhance contextual item representations, integrating KG as a vocabulary bias to enrich responses. (2) **KGSF** [38] integrates DBpedia and ConceptNet to bridge the semantic gap between word-level and item-level user preferences. (3) **C2-CRS** [40] epmploys coarse–to-fine contrastive learning to align multi-type semantic information in CRSs. (4) **VRICR** [34] introduces a variational Bayesian method to reconstruct missing entity relations and dynamically select knowledge based on dialogue context.

PLM-based methods. (1) **UniCRS** [23] unifies recommendation and dialogue generation tasks via knowledge-enhanced prompt learning on a fixed PLM. (2) **MESE** [27] employs an item metadata encoder to learn semantic-aligned item embeddings with dialogue context, allowing the PLM to generate review-aware recommendations and responses. (3) **UniMIND** [3] is a unified multi-goal conversational recommender system that utilizes prompt-based learning to integrate Goal Planning, topic prediction, item recommendation, and response generation within a single sequence-to-sequence framework. (4) **PECRS** [16] introduces a unified CRS designed to jointly handle recommendation and conversation tasks through parameter-efficient fine-tuning techniques.

Table 1 shows a comparison of all the CRS methods employed in our study, highlighting their use of external knowledge and the integration of a PLM backbone.

4.3 Evaluation and implementation details

Following previous CRS works [3, 34], we adopt different metrics to evaluate the recommendation and dialogue generation tasks separately. For recommendation, we use Recall@K (R@K) and NDCG@K (N@K) with $K \in 10, 50$ to measure top-K recommendation performance. For dialogue generation, we adopt ROUGE-n with $n \in \{1, 2, l\}$ and BLEU-t with $t \in \{1, 2\}$ to evaluate the word-level recall and precision of generated n-grams against reference responses. We also use Distinct-m with $m \in \{1, 2\}$ to measure diversity. Higher scores on all metrics indicate better performance.

All experiments are conducted on an NVIDIA A100 GPU with 40GB of memory, providing a consistent and high-performance environment for all methods. For KBRD and KGSF, we use the CRSLab framework for implementation [37], while for the other approaches, we directly adopt the publicly available code released by their respective authors. To ensure a fair comparison, we strictly follow the hyperparameter tuning ranges specified in the original papers and document the tuning details in our GitHub repository. For TG-Redial, we replace PLMs with their Chinese versions for methods that are not specifically designed for Chinese scenarios, as the dataset is in Chinese. Since prior works often use different sub-datasets or adopt varying data-splitting strategies (e.g., some methods lack a validation set), we standardize the evaluation by unifying the dataset and applying a consistent splitting strategy across all methods.

Table 1: Comparison of different approaches based on external knowledge and model usage.

Approach	External Knowledge				PLM
	KG	Rev.	Descr.	Prof.	
KBRD [1]	✓	✗	✗	✗	✗
KGSF [38]	✓	✗	✗	✗	✗
C2-CRS [40]	✓	✓	✗	✗	✗
VRICR [34]	✓	✗	✗	✗	✗
UniCRS [23]	✓	✗	✗	✗	DialoGPT [35]
MESE [27]	✗	✗	✓	✗	GPT-2 [15]
UniMIND-N [3]	✗	✗	✗	✓	BART [10]
UniMIND-S [3]	✗	✗	✗	✓	BART [10]
PECRS [16]	✗	✗	✓	✗	GPT-2 [15]

5 Experiments

5.1 Overall performance (RQ1)

We reproduce all methods on the three public datasets listed in Section 4.1 and report their recommendation and response generation performance separately in Table 2 and Table 3.

Answer to RQ1: The BART-based PLM methods generally exhibit strong performance in both recommendation and response generation tasks, demonstrating robust generalization across all evaluated datasets.

Recommendation. According to the results shown in Table 2, we make several key observations: (1) UniMIND-N and UniMIND-S consistently achieve the best recommendation performance across all datasets, potentially due to the superior capability of BART [10] and multi-task learning, with UniCRS ranking second on the INSPIRED dataset. Additionally, we find that all methods achieve their highest performance on the ReDial dataset and their lowest on the TG-ReDial dataset.

(2) KG-based methods exhibit greater robustness across diverse datasets, whereas PLM-based methods show higher variability in performance. Among KG-based methods, VRICR consistently outperforms others, likely due to its ability to dynamically refine KGs. Despite lacking rich semantic information, KG-based methods leverage structured representations that mitigate semantic differences across languages. This structural consistency enhances their stability across datasets, making them less sensitive to dataset-specific characteristics.

(3) In contrast, PLM-based methods exhibit significant performance variation. For instance, UniCRS performs the worst on the ReDial dataset but achieves competitive results on the INSPIRED dataset, even approaching the best-performing models. This inconsistency may arise from their heavy reliance on external knowledge and the natural language understanding capabilities of PLMs. Since PLMs encode pre-trained knowledge differently depending on the dataset –particularly across languages – their performance can fluctuate significantly. These observations highlight the importance of model–knowledge compatibility in PLM-based approaches, warranting further analysis in Section 5.2.

Response generation. As shown in Table 3, UniMIND methods consistently achieve high performance across all datasets in response generation in terms of ROUGE and Distinct scores. However, the KG-based method VRICR achieves the highest BLEU@1 and

Table 2: Recommendation performance on three datasets. The best results are highlighted in bold, while the second-best results are underlined. * denotes statistically significant improvements (t-test, $p < 0.01$) over the second best result.

Models	Datasets	ReDial				TG-ReDial				INSPIRED			
	Metrics	R@10	R@50	N@10	N@50	R@10	R@50	N@10	N@50	R@10	R@50	N@10	N@50
KG-based	KBRD	0.1806	0.3357	0.0966	0.1310	0.0245	0.0561	0.0127	0.0195	0.0420	0.0970	0.0197	0.0315
	KGSF	0.1794	0.3637	0.0955	0.1364	0.0200	0.0721	0.0091	0.0204	0.0258	0.0970	0.0150	0.0296
	C2-CRS	0.2219	0.4047	0.1243	0.1648	0.0325	0.0770	0.0161	0.0258	-	-	-	-
	VRICR	0.2520	0.4080	0.1420	0.1770	0.0320	0.0810	0.0160	0.0260	0.0540	0.0880	0.0350	0.0420
PLM-based	UniCRS	0.1493	0.3409	0.0772	0.1192	0.0298	0.0657	0.0135	0.0219	0.1445	**0.3242***	0.0863	0.1259
	MESE	0.2118	0.4020	0.1172	0.1591	0.0152	0.0402	0.0085	0.0168	0.0981	0.2585	0.0692	0.0917
	UniMIND-N	**0.6600***	**0.7064***	**0.6045***	**0.6148***	0.0454	0.1247	0.0250	0.0416	0.1225	0.2806	0.0776	0.1131
	UniMIND-S	0.4225	0.4994	0.3602	0.3772	**0.0552***	**0.1377***	**0.0305***	**0.0479***	**0.1548***	0.3064*	**0.1230***	**0.1553***
	PECRS	0.1844	0.3583	0.1053	0.1435	0.0173	0.0439	0.0096	0.0172	0.0903	0.1710	0.0602	0.0781

Table 3: Response generation performance on three datasets. We exclude C2-CRS from the INSPIRED analysis due to the absence of review knowledge. The best results are highlighted in bold, while the second-best results are underlined. * denotes statistically significant improvements (t-test, $p < 0.01$) over the second best result.

Dataset	Model	ROUGE@1	ROUGE@2	ROUGE@L	BLEU@1	BLEU@2	Distinct@1	Distinct@2
ReDial	KBRD	0.2745	0.0534	0.2691	0.1841	0.0533	0.0079	0.0209
	KGSF	0.2502	0.0459	0.2440	0.1740	0.0491	0.0309	0.1375
	C2-CRS	0.0535	0.0063	0.0528	0.0329	0.0062	0.0299	0.1450
	VRICR	0.1459	0.0279	0.1388	**0.3514***	0.2464	0.0113	0.1115
	UniCRS	0.2937	0.0684	0.2887	0.2005	0.0993	**0.3125***	**0.4769***
	MESE	0.3076	0.0960	0.3040	0.1300	0.0837	0.0156	0.0488
	UniMIND-N	0.3560	0.2462	0.3549	0.3333	0.2659	0.1233	0.2375
	UniMIND-S	**0.3580***	**0.2474***	**0.3570***	0.3386	**0.2709***	0.1206	0.2294
	PECRS	0.0932	0.0221	0.0875	0.0449	0.0245	0.0320	0.1226
TG-ReDial	KBRD	0.3539	0.0505	0.3050	0.2727	0.0805	0.0256	0.0776
	KGSF	0.3457	0.0594	0.2952	0.2671	0.0918	0.0090	0.0366
	C2-CRS	0.1911	0.0202	0.1697	0.1568	0.0295	0.0297	0.1678
	VRICR	0.3325	0.0928	0.2917	**0.4492***	**0.3453***	0.0146	0.1644
	UniCRS	0.2368	0.0391	0.2011	0.2912	0.1007	0.0301	0.1708
	MESE	0.3412	0.1082	0.3046	0.1923	0.0728	0.0121	0.1038
	UniMIND-N	0.3660	0.0941	0.3210	0.3167	0.1440	**0.0425***	**0.1883***
	UniMIND-S	**0.3672***	0.0979	**0.3230***	0.3159	0.1480	0.0408	0.1801
	PECRS	0.3572	**0.1153***	0.3186	0.2055	0.0852	0.0358	0.1277
INSPIRED	KBRD	**0.2707***	0.0397	**0.2573***	0.1534	0.0386	0.0027	0.0125
	KGSF	0.2091	0.0048	0.2037	0.0748	0.0035	0.0069	0.1000
	VRICR	0.1464	0.0194	0.1368	**0.3761***	**0.2602***	0.0269	0.1455
	UniCRS	0.2164	0.0324	0.2106	0.1447	0.0574	**0.4724***	**0.8267***
	MESE	0.1385	0.0359	0.1297	0.1034	0.0540	0.0102	0.0249
	UniMIND-N	0.2270	**0.0777***	0.2154	0.1866	0.0862	0.1491	0.3235
	UniMIND-S	0.1211	0.0251	0.1207	0.1944	0.0932	0.2068	0.4375
	PECRS	0.0704	0.0129	0.0665	0.0548	0.0252	0.1260	0.3534

BLEU@2 scores, indicating its strength in generating n-gram precise responses. UniCRS demonstrates superior capability in producing diverse responses, as reflected in its high Distinct scores. This is largely due to its use of DialoGPT [35] as the PLM backbone, which is optimized for open-domain dialogue and prioritizes varied responses over repetitive or generic phrases.

5.2 Model–knowledge compatibility (RQ2)

We examine the compatibility of CRS models with different types of external knowledge.

Scope of this study. We focus on PLM-based methods in this analysis, as traditional KG-based methods are specifically designed around KGs, and replacing KGs with other external knowledge

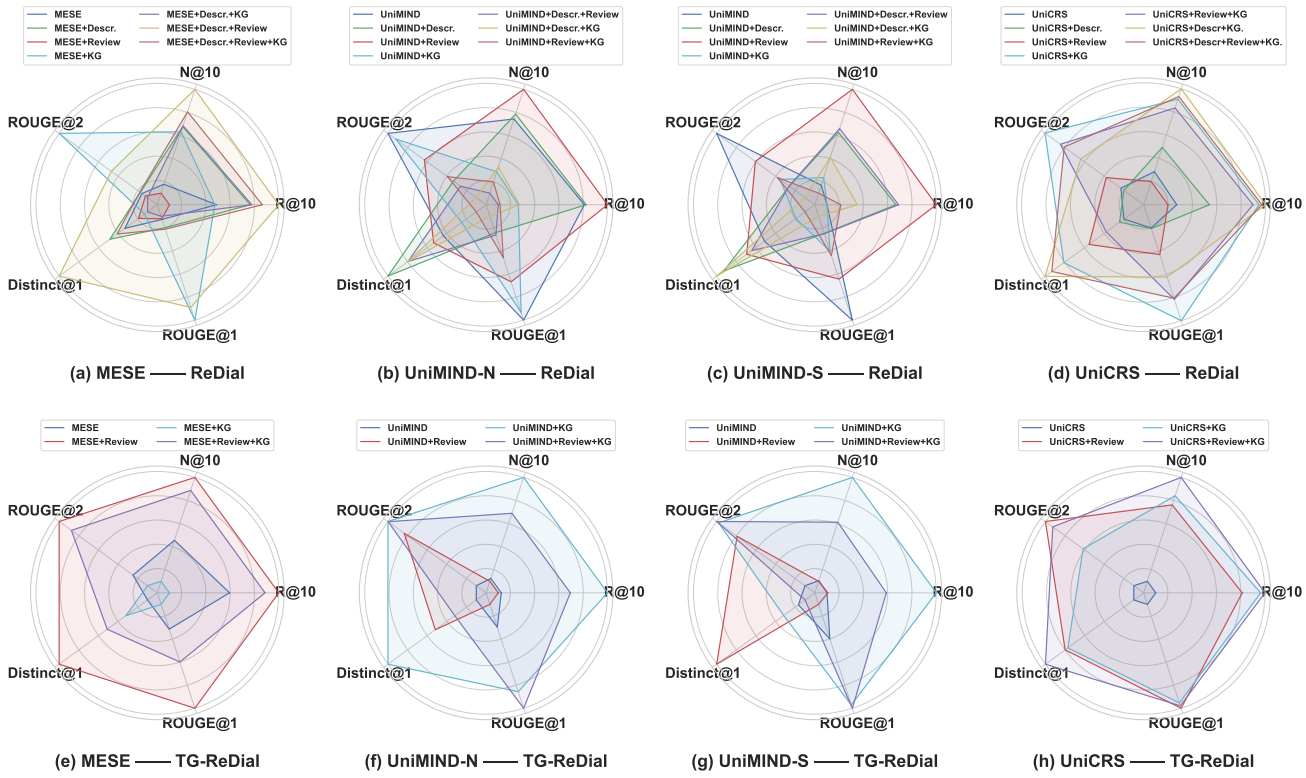


Figure 2: Overall performance of CRS methods while incorporating different kinds of external knowledge. Note that the evaluation metric values in each plot are normalized relative to the highest value.

sources would deviate from their original methodology. We select three distinct PLM backbones from Table 1 for our analysis: DialoGPT (UniCRS), GPT-2 (MESE), and BART (UniMIND). PECRS is excluded due to computational inefficiency and architectural overlap with MESE (both use GPT-2). For UniMIND, we analyze both its -N (natural language prompt) and -S (special token prompt) sub-models, as the two divergent designs offer insights into knowledge compatibility under varying prompt configurations. This setup enables systematic exploration of how knowledge integration efficacy depends on PLM architecture choices. We incorporate datasets in different languages: ReDial as the English dataset and TG-ReDial as the Chinese dataset, excluding INSPIRED due to its smaller size and redundancy as another English dataset. For the fusion method, we examine prompt concatenation, embedding MLP fusion, and cross-attention fusion, reporting the best performance among these strategies. The results are shown in Figure 2.

Answer to RQ2: Decoder-only models perform better with structured knowledge but struggle with unstructured knowledge, while encoder-decoder models show the opposite trend.

From the model perspective, PLMs with different backbones exhibit varying effectiveness in leveraging structured and unstructured knowledge. Decoder-only models (e.g., GPT-2, DialoGPT) handle structured KGs better but struggle with unstructured sources like reviews and descriptions (see red-highlighted areas in the first row of Figure 2). In contrast, encoder-decoder models (such as BART) demonstrate superior understanding and utilization of unstructured knowledge due to their ability to capture richer semantic representations, whereas incorporating structured KGs information tends to degrade their performance.

UniCRS differs from other PLM-based methods by using DialoGPT solely for inference without fine-tuning, which limits the benefits of external knowledge. In contrast, other models show varying performance gains depending on the type of external knowledge used. This suggests that fine-tuning the PLM is essential for maximizing performance improvements.

For the ReDial dataset, we obtain additional insights. (1) In general, structured knowledge influences ROUGE scores, while unstructured knowledge primarily impacts NDCG, Recall, and Distinct scores. Each type influences recommendation and dialogue differently: KGs boost ROUGE, whereas reviews and descriptions improve NDCG, Recall, and Distinct, with UniCRS being an exception. For UniCRS, adding KGs helps retain ROUGE, while descriptions significantly increase Distinct but, like reviews, reduce NDCG and Recall.

(2) BART-based methods generally prefer natural language prompts over special token prompts when utilizing a single external knowledge type. From the perspective of the model input, we observe a preference difference in how the models respond to different prompting methods. UniMIND-N employs a natural language prompt with a guidance sentence for each task, such as "Recommend an item:", while UniMIND-S uses task-specific prompts with special tokens, like `[item]` for the item recommendation task. When using a single external knowledge type, UniMIND-N outperforms UniMIND-S on most knowledge types across all

Table 4: Statistics across different categories in ReDial.

Category	Count	Proportion	Example
Genre-Specific	2555	10.67%	"I'm looking for a good horror movie."
Descriptive	7424	31.00%	"I found it very boring."
Comparative	3326	13.89%	"I want something like The Matrix."
Other	13033	54.41%	"I haven't seen it yet."

metrics, except for review knowledge, where both models show similar performance. This suggests that BART tends to favor natural language prompts over special token prompts when working with a single external knowledge source.

(3) The impact of knowledge varies by language and method. From the language perspective, we observe that the influence of knowledge sources like KGs and reviews differs across languages and methods, without consistently contributing significantly more to any particular metric. For instance, as shown in Figure 2 (e-h), incorporating reviews into MESE leads to better performance across multiple metrics compared to KGs-only. Interestingly, as opposed to the English dataset, adding KG information to UniMIND in the Chinese context results in significant improvements across most metrics, particularly in the recommendation task, while adding reviews primarily enhances performance in the dialogue task.

Additionally, UniMIND in Chinese can incorporate KG information better than other types of knowledge. However, in recommendation tasks, it still lags behind the English UniMIND (with UniMIND-N showing 0.1715 vs. 0.4787 on Recall@10). This discrepancy could be due to language differences or the use of different training strategies for the same model across languages. Therefore, when designing CRS methods, it is crucial to consider the language factor in both model and knowledge selection and a high-quality PLM is essential for effective recommendations.

5.3 Knowledge complementarity (RQ3)

To address RQ3, we examine the impact of combining different knowledge sources on CRS and uncovering additional insights.

Answer to RQ3: Combining different types of knowledge does not always lead to better performance than using a single type, but combining the same type of knowledge tends to be more effective than combining different types.

For MESE, combining descriptions and reviews yields the best results, while pairing descriptions with KGs reduces performance relative to single-source input. In UniMIND-N and -S, most two-knowledge combinations degrade performance, with structured-unstructured pairs performing worse than unstructured-only ones. This may stem from structured knowledge like KGs requiring specialized encoding, which can clash with unstructured data, introducing noise or misalignment. For UniCRS, KGs perform best alone, though combining them with descriptions improves some metrics at the cost of ROUGE scores.

We also find two additional insights: (1) Description knowledge dominates other knowledge types across all PLM-based methods. Through further analysis, we identified a phenomenon called "knowledge dominance", where certain knowledge types exert a stronger influence on metric scores, overshadowing the contributions of others.

In UniMIND, adding descriptions maintains high performance on Distinct, NDCG, and Recall, regardless of additional sources. This dominance follows the order: *description* > *review* > *KG*, suggesting that additional sources only help if they complement rather than conflict with the dominant one. (2) BART-based methods generally prefer special token prompts over natual language prompts when utilizing multiple external knowledge types. In contrast to single knowledge type case, we find that UniMIND-S performs better than UniMIND-N on mixed knowledge occasions for recommendation task. Especially for the mixed of review and description information, UniMIND-S can achieve 0.6266 at Recall@10, compared to 0.4234 for UniMIND-N, and 0.5673 on NDCG@10 compared to 0.3552. However, in this case, different prompt designs have little impact on the response generation task, highlighting the need for careful prompt design when integrating diverse knowledge sources.

5.4 Scenario-specific analysis (RQ4)

We evaluate CRS performance across dialogue scenarios to analyze the contextual impact of structured and unstructured knowledge. Conversations are grouped into three categories based on keywords: (1) *Genre-Specific Mentions*—dialogues referencing explicit genres (e.g., "action," "comedy"); (2) *Descriptive Preference*—segments with evaluative language (e.g., "boring," "exciting"); (3) *Comparative Judgments*—utterances with comparative phrases (e.g., "like," "better than"). Category statistics are in Table 4, and results of the scenario comparisons in terms of recommendation metrics appear in Figure 3. Without external knowledge, genre-specific conversations achieve the highest Recall and NDCG across most methods, except for UniCRS, which performs best in descriptive scenarios. When external knowledge is incorporated, we observe diverse effects in different scenarios:

Answer to RQ4: Description knowledge demonstrates broad positive effects across all scenarios, review knowledge significantly benefits UniMIND in genre-specific and descriptive scenarios, and KGs primarily improve comparative scenarios.

Specifically, review knowledge drives the most substantial improvements for UniMIND-S across all scenario types, particularly in genre-specific and descriptive scenarios. However, it degrades performance for other models, aligning with insights from Subsection 5.2. Description knowledge consistently enhances both Recall and NDCG metrics across all scenarios for the three evaluated models. KGs have diverse effects: it generally improves comparative scenario conversations for all models and significantly enhances other scenario types for UniCRS. This observation suggests that the effectiveness of external knowledge depends both on the model architecture and scenario requirements, highlighting the importance of aligning knowledge type with conversational context to optimize recommendation performance.

5.5 Discussion on LLM-based CRSs

The emergence of large language models (LLMs) has revolutionized CRSs by enabling unprecedented open-domain dialogue capabilities and contextual understanding. As the field transitions from (smaller) PLM-based to LLM-based paradigms, our experimental insights

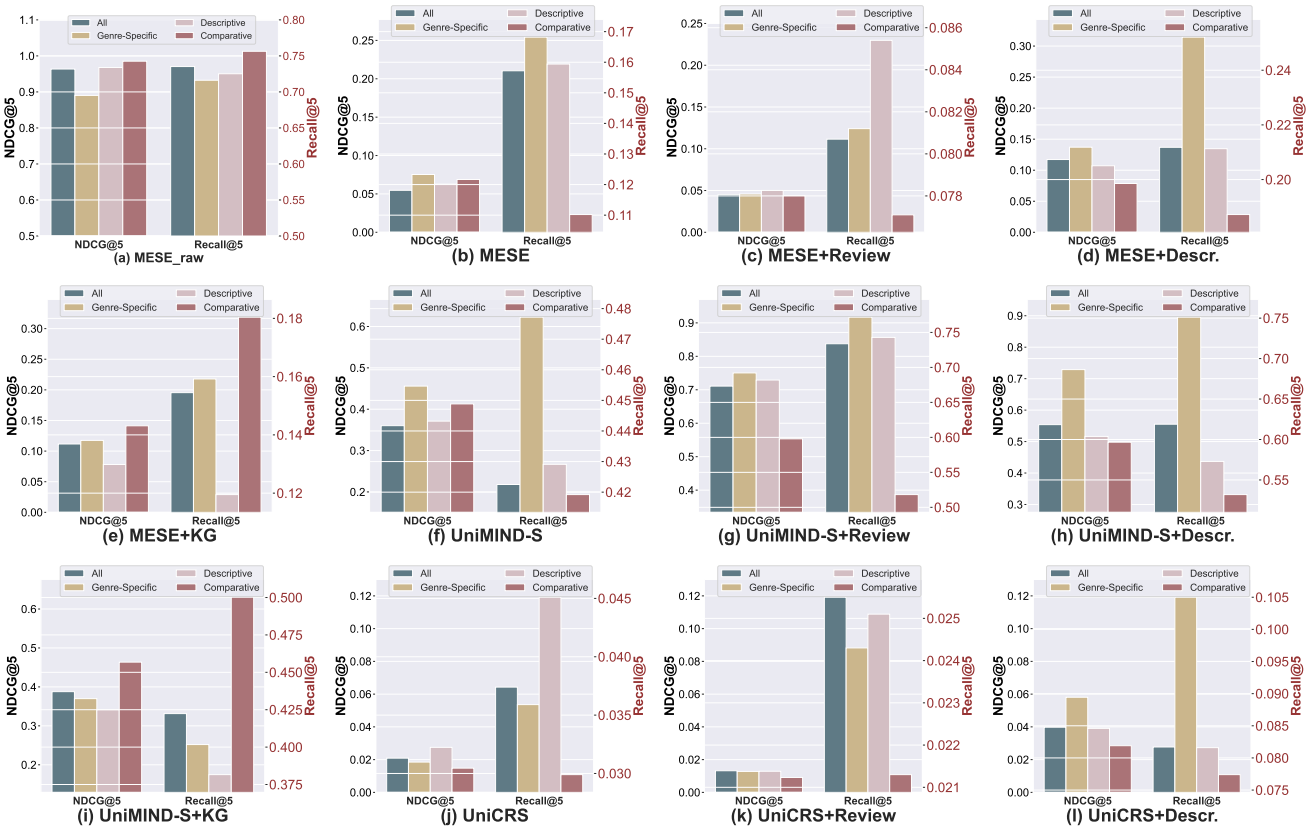


Figure 3: Recommendation performance of different CRS methods under various dialogue scenarios in ReDial.

provide critical design inspiration for next-generation LLM-based CRSs. Here, we discuss some potential key directions tailored to LLMs' unique strengths and challenges:

Designing scenario-aware knowledge routing to improve robustness. LLMs' large parameter space and in-context learning make them well-suited for dynamic knowledge selection. Our RQ4 findings show performance gaps across conversation scenarios. Future CRSs could leverage LLMs' advanced contextual understanding to identify specific conversation scenarios and activate the most appropriate knowledge sources accordingly. For example, it is possible to deploy knowledge routers that parse user utterances to trigger relevant knowledge sources (*e.g.,* KGs for "Find movies like *Inception*" vs. reviews for "Is this film emotionally impactful?").

Leveraging Retrieval-Augmented Generation (RAG) for knowledge integration. Another compelling direction is the integration of RAG architectures to dynamically incorporate external knowledge within LLM-based CRS. Unlike static knowledge fusion techniques that pre-encode knowledge into the model or input embeddings, RAG enables the system to retrieve contextually relevant information – such as item reviews, knowledge graph entities – at inference time based on the ongoing dialogue context. This capability is particularly valuable for domains with frequently changing content or long-tail items with sparse data, as it allows the CRS to ground its responses in up-to-date, personalized evidence. Moreover, RAG facilitates explainable recommendations by explicitly showcasing the sources of retrieved knowledge, thus enhancing transparency and user trust.

6 Conclusion

We have conducted a reproducibility study to investigate the interplay between conversational recommender system architectures and different forms of external knowledge. We have revealed 3 fundamental findings for CRS design: (1) Encoder-decoder PLMs (*e.g.,* BART) achieve superior performance with unstructured knowledge (reviews/descriptions), while decoder-only models (*e.g.,* GPT-2) better utilize structured knowledge. (2) Combining multiple knowledge sources does not always outperform using a single type, but merging similar knowledge types is generally more effective than mixing different ones. (3) The effectiveness of external knowledge varies by model and scenario, with description knowledge offering broad benefits, and knowledge graph information improving comparative scenarios. This work provides practical guidance on external knowledge selection and underscores the need to account for different conversational scenarios, offering empirical insights for developing more adaptive and robust CRS.

Our study primarily focuses on text-based knowledge sources (*e.g.,* knowledge graphs, reviews, and descriptions), leaving multimodal knowledge (*e.g.,* images and social networks) unexplored. Future work could incorporate a more diverse range of knowledge types and domains to further validate and extend our findings.

References

[1] 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 1803–1813. doi:10.18653/v1/D19-1189

[2] Zhongxia Chen, Xiting Wang, Xing Xie, Mehul Parsana, Akshay Soni, Xiang Ao, and Enhong Chen. 2021. Towards explainable conversational recommendation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 2994–3000.

[3] Yang Deng, Wenxuan Zhang, Weiwen Xu, Wenqiang Lei, Tat-Seng Chua, and Wai Lam. 2023. A unified multi-task learning framework for multi-goal conversational recommender systems. *ACM Transactions on Information Systems* 41, 3 (2023), 1–25.

[4] Zhikang Dong, Xiulong Liu, Bin Chen, Pawel Polak, and Peng Zhang. 2024. Musechat: A conversational music recommendation system for videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12775–12785.

[5] Jiabao Fang, Shen Gao, Pengjie Ren, Xiuying Chen, Suzan Verberne, and Zhaochun Ren. 2024. A multi-agent conversational recommender system. *arXiv preprint arXiv:2402.01135* (2024).

[6] Yue Feng, Shuchang Liu, Zhenghai Xue, Qingpeng Cai, Lantao Hu, Peng Jiang, Kun Gai, and Fei Sun. 2023. A large language model enhanced conversational recommender system. *arXiv preprint arXiv:2308.06212* (2023).

[7] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI open* 2 (2021), 100–126.

[8] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward Sociable Recommendation Dialog Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 8142–8152. https://www.aclweb.org/anthology/2020.emnlp-main.654

[9] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web* 6, 2 (2015), 167–195.

[10] 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7871–7880. doi:10.18653/v1/2020.acl-main.703

[11] 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 31 (NIPS 2018)*.

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

[13] 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*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1161–1173. doi:10.18653/v1/2021.findings-acl.99

[14] Dhanya Pramod and Prafulla Bafna. 2022. Conversational recommender systems techniques, tools, acceptance, and adoption: a state of the art review. *Expert Systems with Applications* 203 (2022), 117539.

[15] 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.

[16] Mathieu Ravaut, Hao Zhang, Lu Xu, Aixin Sun, and Yong Liu. 2024. Parameter-Efficient Conversational Recommender System as a Language Processing Task. *arXiv preprint arXiv:2401.14194* (2024).

[17] Xuhui Ren, Tong Chen, Quoc Viet Hung Nguyen, Lizhen Cui, Zi Huang, and Hongzhi Yin. 2024. Explicit knowledge graph reasoning for conversational recommendation. *ACM Transactions on Intelligent Systems and Technology* 15, 4 (2024), 1–21.

[18] 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.

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

[20] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st international acm sigir conference on research & development in information retrieval*. 235–244.

[21] Lingzhi Wang, Huang Hu, Lei Sha, Can Xu, Kam-Fai Wong, and Daxin Jiang. 2021. RecInDial: A unified framework for conversational recommendation with pretrained language models. *arXiv preprint arXiv:2110.07477* (2021).

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

[23] 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.

[24] Chi-Man Wong, Fan Feng, Wen Zhang, Chi-Man Vong, Hui Chen, Yichi Zhang, Peng He, Huan Chen, Kun Zhao, and Huajun Chen. 2021. Improving conversational recommender system by pretraining billion-scale knowledge graph. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2607–2612.

[25] Yunjia Xi, Weiwen Liu, Jianghao Lin, Bo Chen, Ruiming Tang, Weinan Zhang, and Yong Yu. 2024. MemoCRS: Memory-enhanced Sequential Conversational Recommender Systems with Large Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2585–2595.

[26] Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. 2017. CN-DBpedia: A never-ending Chinese knowledge extraction system. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 428–438.

[27] 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.

[28] Dayu Yang, Fumian Chen, and Hui Fang. 2024. Behavior alignment: a new perspective of evaluating LLM-based conversational recommendation systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2286–2290.

[29] Li Yang, Anushya Subbiah, Hardik Patel, Judith Yue Li, Yanwei Song, Reza Mirghaderi, and Vikram Aggarwal. 2024. Item-Language Model for Conversational Recommendation. *arXiv preprint arXiv:2406.02844* (2024).

[30] Gangyi Zhang. 2023. User-centric conversational recommendation: Adapting the need of user with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1349–1354.

[31] Lu Zhang, Chen Li, Yu Lei, Zhu Sun, and Guanfeng Liu. 2024. An Empirical Analysis on Multi-turn Conversational Recommender Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 841–851.

[32] Xiaotong Zhang, Xuefang Jia, Han Liu, Xinyue Liu, and Xianchao Zhang. 2024. A Goal Interaction Graph Planning Framework for Conversational Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19578–19587.

[33] Xiaoyu Zhang, Ruobing Xie, Yougang Lyu, Xin Xin, Pengjie Ren, Mingfei Liang, Bo Zhang, Zhanhui Kang, Maarten de Rijke, and Zhaochun Ren. 2024. Towards empathetic conversational recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 84–93.

[34] 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.

[35] Y Zhang. 2019. Dialogpt: Large-Scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536* (2019).

[36] Jinfeng Zhou, Bo Wang, Ruifang He, and Yuexian Hou. 2021. CRFR: Improving conversational recommender systems via flexible fragments reasoning on knowledge graphs. In *Proceedings of the 2021 conference on empirical methods in natural language processing*. 4324–4334.

[37] 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.

[38] 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 *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1006–1014.

[39] 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, COLING 2020, Barcelona, Spain, December 8-11, 2020*.

[40] 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.