Enhancing Conversational Recommender Systems with Tree-Structured Knowledge and Pretrained Language Models

Yongwen Ren[1,2], Chao Wang[3*], Peng Du[4], Chuan Qin[5,6], Dazhong Shen[7], Hui Xiong[8,9*]

[1]School of Computer Science and Technology, University of Science and Technology of China
[2]iFLYTEK AI Research, iFLYTEK Co.,Ltd
[3]School of Artificial Intelligence and Data Science, University of Science and Technology of China
[4]School of Software and Microelectronics, Peking University
[5]Computer Network Information Center, Chinese Academy of Sciences
[6]University of Chinese Academy of Sciences
[7]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
[8]Thrust of Artificial Intelligence, The Hong Kong University of Science and Technology (Guangzhou)
[9]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology
yovren@mail.ustc.edu.cn, wangchaoai@ustc.edu.cn, pdu@pku.edu.cn,
chuanqin0426@gmail.com, shendazhong@nuaa.edu.cn, xionghui@ust.hk

Abstract

Recent advances in pretrained language models (PLMs) have significantly improved conversational recommender systems (CRS), enabling more fluent and context-aware interactions. To further enhance accuracy and mitigate hallucination, many methods integrate PLMs with knowledge graphs (KGs), but face key challenges: failing to fully exploit PLM reasoning over graph relationships, indiscriminately incorporating retrieved knowledge without context filtering, and neglecting collaborative preferences in multi-turn dialogues. To this end, we propose PCRS-TKA, a prompt-based framework employing retrieval-augmented generation to integrate PLMs with KGs. PCRS-TKA constructs dialogue-specific knowledge trees from KGs and serializes them into texts, enabling structure-aware reasoning while capturing rich entity semantics. Our approach selectively filters context-relevant knowledge and explicitly models collaborative preferences using specialized supervision signals. A semantic alignment module harmonizes heterogeneous inputs, reducing noise and enhancing accuracy. Extensive experiments demonstrate that PCRS-TKA consistently outperforms all baselines in both recommendation and conversational quality.

Code — https://github.com/YovRen/PCRS-TKA

1 Introduction

Recommendation systems play a crucial role in intelligent assistants by helping users efficiently discover relevant items. However, traditional systems lack interactive dialogue abilities, limiting flexibility and explainability (Chen et al. 2017; Ji et al. 2025). Conversational recommender systems (CRS) address this issue by supporting personalized and natural interactions. With recent advances in pretrained language models (PLMs), CRS have achieved notable improvements in conversational fluency and context understanding (Wu et al. 2021; Qin et al. 2025).

However, PLMs still suffer from hallucinations, generating inaccurate or irrelevant information, which damages recommendation reliability (Chen et al. 2023; Zhang et al. 2023). To mitigate this issue, knowledge graphs (KGs) can be integrated with PLMs to provide factual external knowledge, thereby improving both accuracy and robustness of conversational recommendations (Wang et al. 2022, 2018; Tong, Li, and Liu 2024).

In the literature, early CRS primarily relied on structured conversations centered around item attributes such as genre or price (Gao et al. 2021). Later, KG-based CRS (Wang et al. 2019; Petroni et al. 2019; Bouraoui, Camacho-Collados, and Schockaert 2020) integrated external knowledge resources and developed alignment techniques to ensure semantic consistency, e.g., KBRD (Chen et al. 2019) leveraging relational graph convolutional networks (RGCNs) (Schlichtkrull et al. 2018), KGSF (Zhou et al. 2020a) applying mutual information maximization for word-entity alignment. More recent PLM-based approaches, including BARCOR (Wang, Su, and Chen 2022), UniCRS (Wang et al. 2022), KERL (Qiu et al. 2025), and DCRS (Dao et al. 2024), combine prompt learning with KG integration to mitigate hallucination and boost domain knowledge. However, these existing methods fail to fully exploit the rich semantic information embedded in KGs, limiting their ability to capture complex relational patterns and contextual dependencies.

Despite progress, several challenges persist in integrating PLMs and KGs for conversational recommendation. First, many existing approaches depend on GCN-based encoders (e.g., RGCN) (Zhu et al. 2024), which struggle to leverage PLMs' reasoning capabilities over complex graph structures and cannot fully capture rich KG semantics (Shen et al. 2021). Second, current approaches typically integrate all retrieved KG information without filtering out information irrelevant to the current dialogue context. This inevitably in-

*Corresponding author.

troduces noise and compromises the model's performance. Third, dialogue text is often treated merely as textual input data, neglecting valuable latent user collaborative preference information embedded in conversations. This oversight can result in suboptimal personalized recommendations that fail to align with users' true preferences (Wang et al. 2021b).

To address these issues, we propose PCRS-TKA, a prompt-based framework that integrates PLMs with KGs via retrieval-augmented generation (RAG) style strategy. First, to augment the static, global embeddings from RGCN and fully exploit PLM reasoning capabilities, PCRS-TKA constructs dynamic, context-aware knowledge trees in a RAG-style. These trees are selectively built from the KG, serialized into text, and fed into the PLM to enable direct, fine-grained reasoning on dialogue-specific structured information. Second, our approach filters context-relevant semantics through this dialogue-specific knowledge tree construction, ensuring only pertinent information is utilized. Third, we explicitly model collaborative preferences from multi-turn dialogues using specialized supervision signals, capturing latent user preference patterns that are often overlooked in existing methods. Additionally, we employ an RGCN to encode the global semantics of the entire KG, and these multiple sources of knowledge are jointly embedded into prompt representations, enabling the PLM to perform end-to-end graph reasoning in a unified manner. Complemented by a semantic alignment module that harmonizes heterogeneous inputs from dialogues and KGs, PCRS-TKA effectively reduces noise and enhances recommendation accuracy. Extensive experiments demonstrate its superiority in both recommendation and conversational quality.

2 Related Work

2.1 Conversational Recommendation Systems

Early CRS methods (Chen et al. 2019, 2017) focused on structured conversations that gathered user preferences through item attributes like genre or price, using predefined templates and algorithms such as multi-armed bandits or reinforcement learning (Wang et al. 2021a). However, these methods lacked flexibility and natural language generation capabilities. KG-based CRS methods were later developed to improve this. KBRD (Chen et al. 2019) introduced KGs and RGCNs to model relationships between items and users. KGSF (Zhou et al. 2020a) enhanced this by incorporating word-level KGs and MIM to align word and entity representations, resulting in more coherent responses. RevCore (Lu et al. 2021) enriched dialogue generation with unstructured review data, while C2CRS (Zhou et al. 2022) used multi-granularity contrastive learning for multimodal data alignment (Zong et al. 2024). Despite these advancements, KG-based methods often treated recommendation and dialogue modules separately, limiting the full use of dialogue content. With the rise of PLMs, prompt learning was introduced to improve conversational capabilities. For example, BARCOR (Wang, Su, and Chen 2022) uses BART (Lewis et al. 2020) for better response generation, UniCRS (Wang et al. 2022) integrates recommendation and dialogue generation via prompt learning, and DCRS (Dao et al. 2024) employs a knowledge-aware retriever for improved capabilities.

2.2 Unifying PLMs and KGs in Recommendations

Recommendation systems often require domain-specific knowledge that PLMs alone cannot provide. KGs supply structured knowledge to fill this gap. Many studies focus on aligning the semantic spaces of PLMs and KGs by modifying Transformer architectures, often incorporating cross-attention mechanisms to jointly model dialogue text and KG information. For instance, KGSF (Zhou et al. 2020a) uses mutual information maximization for embedding alignment, while C2CRS (Zhou et al. 2022) applies contrastive learning at both sentence and word levels to improve semantic consistency. In prompt learning frameworks, the PLM architecture generally remains fixed, and KG information is integrated by concatenating implicit graph embeddings derived from graph neural networks (GNNs) into input prompts, as seen in UniCRS and DCRS. Although this enables KG data incorporation, it does not fully exploit the reasoning abilities of PLMs over KG relations, nor does it provide dynamic knowledge filtering tailored to the dialogue context.

3 Methodology

3.1 Task Formulation

The goal of CRS is to recommend relevant items while maintaining a natural, interactive dialogue with the user. At each turn, the system analyzes the dialogue history, infers the user's preferences, and generates a response that includes recommended items within the natural language utterance. Formally, let u represent a user in the user set $\mathcal{U}$, i an item in the item set I, and w a word in the vocabulary $\mathcal{V}$. A multi-turn conversation C is defined as a sequence of utterances: $C = \{s_t\}_{t=1}^{n}$, where s_t denotes the t-th sentence. Additionally, an external KG G is represented as a set of triples: $G = \{(e_1, r, e_2)\}$, where $e_1, e_2 \in \mathcal{E}$ and $r \in \mathcal{R}$. Here, we assume $I \subset \mathcal{E}$, indicating that all candidate items are included in $\mathcal{E}$. Given an n-turn conversation C, the task of CRS is to generate a response sentence s_{n+1} and select a set of recommended items I_{n+1} from the item set I.

3.2 Overview of the Approach

As illustrated in Figure 1, we propose **PCRS-TKA**, a conversational recommendation framework that unifies PLMs and KGs via knowledge-enhanced prompt learning. The key idea is to augment the PLM with two distinct forms of knowledge: a static, global understanding of the entire KG, and a dynamic, dialogue-specific subgraph for fine-grained reasoning. We also introduce a collaborative signal to guide the PLM's optimization. This is realized through four core components: 1) **User Preference Extraction Module.** We first employ RoBERTa (Liu et al. 2021) , a bidirectional PLM, and RGCN (Schlichtkrull et al. 2018) to encode dialogue text C and mentioned KG entities E into embeddings $\mathbf{C}$ and $\mathbf{E}$, user preference features $\mathbf{U}$ are then extracted explicitly and supervised by an auxiliary recommendation task to ensure that P_{user} effectively learns and incorporates this information. 2) **Knowledge Tree Enhanced Module.** For

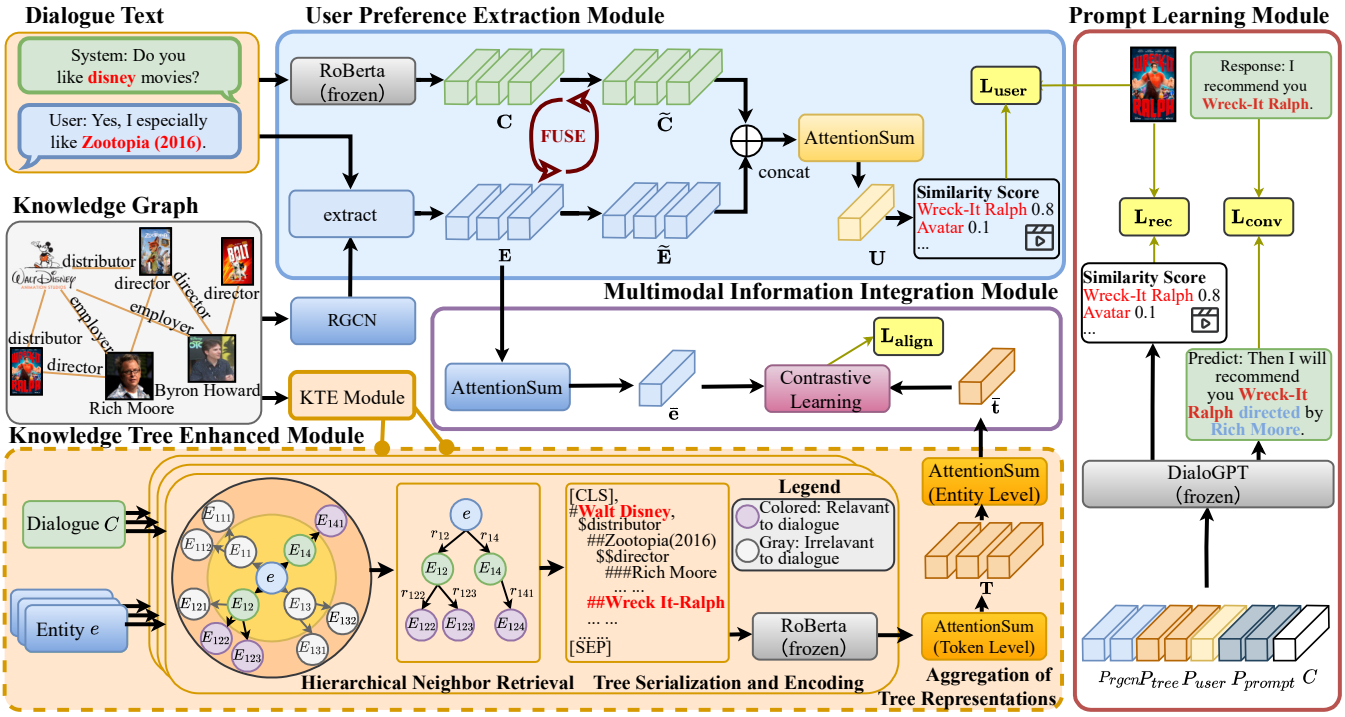

Figure 1: The network architecture of the PCRS-TKA framework.

each entity in E, we retrieve a multi-hop knowledge tree from G using a RAG-style strategy. These trees are then hierarchically aggregated based on structural and contextual relevance, forming a serialized textual sequence T, which is encoded into P_{tree}. 3) **Multimodal Information Integration Module** This module aligns $\mathbf{T}$ with $\mathbf{E}$ through pairwise contrastive learning, effectively integrating information across modules. 4) **Prompt Learning Module.** The concatenated prompt sequence $(P_{\text{prompt}}, P_{\text{rgcn}}, P_{\text{tree}}, P_{\text{user}})$, including learnable soft prompts P_{prompt}, is prepended to the dialogue context and fed into DialoGPT.

3.3 User Preference Extraction Module

This module learns user preference representations $\mathbf{U}$ from multi-turn dialogue C and injects them into the PLM as an auxiliary supervision signal for prompt learning.

Feature Encoder. For natural language understanding, we use RoBERTa to encode the dialogue text C into embeddings $\mathbf{C}$ as follows:

$$\mathbf{C} = \text{RoBERTa}(C). \tag{1}$$

To provides structured information about the entities mentioned in the conversation, we incorporate a knowledge graph G and adopt RGCN to encode it into embeddings $\mathbf{G}$, the embeddings $\mathbf{E}$ corresponding to E are retrieved from matrix $\mathbf{G}$ as below:

$$\mathbf{G} = \text{RGCN}(G), \tag{2}$$
$$\mathbf{E} = \text{Retrieve}(\mathbf{G}, E). \tag{3}$$

User Collaborative Information Extraction. Since $\mathbf{C}$ and $\mathbf{E}$ originate from different sources, a semantic gap exists between them. To bridge this gap, we employ a cross-interaction mechanism (Wang et al. 2022) to align their semantic spaces via linear transformations and bilinear interaction:

$$\mathbf{C} = \mathbf{C}W_C, \mathbf{E} = \mathbf{E}W_E, \mathbf{A} = \mathbf{C}W\mathbf{E}^\top, \tag{4}$$
$$\widetilde{\mathbf{C}} = \mathbf{C} + \mathbf{E}\mathbf{A}, \widetilde{\mathbf{E}} = \mathbf{E} + \mathbf{C}\mathbf{A}^\top, \tag{5}$$

This interaction fuses contextual semantics from the dialogue and relational knowledge from the KG, generating enriched representations $\widetilde{\mathbf{C}}$ and $\widetilde{\mathbf{E}}$. Furthermore, we concatenate them and apply a self-attention mechanism to capture dependencies across the entire input sequence. We define the resulting attention aggregation as a function $\text{ASum}(\cdot)$, and compute the user preference embedding $\mathbf{U}$ as follows:

$$\text{ASum}(\mathbf{X}) = \sum_{i=1}^{n_X}\sum_{j=1}^{n_X} \text{softmax}\left(\frac{\mathbf{X}_i W_Q \cdot (\mathbf{X}_j W_K)^\top}{\sqrt{d}}\right) \cdot \mathbf{X}_j W_V, \tag{6}$$

$$\mathbf{X} = \text{concat}(\widetilde{\mathbf{C}}, \widetilde{\mathbf{E}}), \quad \mathbf{U} = \text{ASum}(\mathbf{X}) \tag{7}$$

where W_Q, W_K, W_V are learnable parameters.

Supervision Recommendation Task. To introduce collaborative preference supervision, we define an auxiliary task. Instead of implicitly fetching neighbor features (Xie et al. 2024; Liu et al. 2025), this task learns to explicitly encode a collaborative signal by computing relevance scores between the user embedding $\mathbf{U}$ and candidate item embeddings $\mathbf{I}$, which are retrieved from the graph entity embeddings $\mathbf{G}$. We then project $\mathbf{I}$ and compute the scores:

$$\mathbf{I} = \text{Retrieve}(\mathbf{G}, I), \quad \mathbf{I} = \mathbf{I}W_I, \quad R = \text{softmax}(\mathbf{U}\mathbf{I}^\top), \tag{8}$$

where W_I is a learnable projection matrix and R represents the predicted rating scores. Given N conversations and ground-truth preference labels $\mathbf{Y}$, we can compute the predicted rating score R in all conversations, and the cross-entropy loss is defined as:

$$L_{user} = -\sum_{j=1}^{N}\sum_{i=1}^{n_I}\left[Y_j^i \cdot \log R_j^i + (1-Y_j^i)\cdot\log(1-R_j^i)\right]. \tag{9}$$

3.4 Knowledge Tree Enhanced Module

To better utilize the structured semantics in KGs, we dynamically construct a dialogue-specific knowledge tree that captures context-relevant information in a hierarchical form, enabling PLMs to conduct structure-aware reasoning.

Hierarchical Neighbor Retrieval. Given a dialogue context C and the mentioned entities $E = \{e_1, \ldots, e_{n_E}\}$, we construct a personalized knowledge tree G_{tree_e} for each entity $e \in E$. We first encode the dialogue into a fixed-length vector by applying a pooling operation (e.g., mean pooling) over word embeddings:

$$\mathbf{c} = \text{Pooling}(\mathbf{C}). \tag{10}$$

Starting from the root $\mathcal{N}_e^0 = \{e\}$, each layer l expands by retrieving one-hop neighbors $\mathcal{M}(v)$ for every node $v \in \mathcal{N}_e^l$. Their relevance to the dialogue is measured via cosine similarity and we select the top-N most relevant neighbors to form the next layer until reaching depth L:

$$\text{sim}(\mathbf{c}, \mathbf{E}_m) = \frac{\mathbf{c}\cdot\mathbf{E}_m^\top}{\|\mathbf{c}\|\|\mathbf{E}_m\|}, \tag{11}$$

$$\mathcal{N}_v^{l+1} = \text{TopN}\left(\mathcal{M}(v), \text{sim}(\mathbf{c}, \mathbf{E}_m), N\right), \tag{12}$$

$$\mathcal{N}_e^{l+1} = \bigcup_{v\in\mathcal{N}_e^l} \mathcal{N}_v^{l+1}, \tag{13}$$

$$G_{\text{tree}_e} = \left(\bigcup_{l=0}^{L}\mathcal{N}_e^l, \text{corresponding edges}\right). \tag{14}$$

Tree Serialization and Encoding. To retain structural information, virtual relation nodes are inserted between parent–child pairs. The tree is serialized into a natural language sequence via depth-first traversal, where entities and relations are formatted as "#EntityName" or "$RelationName" to indicate depth. The serialized sequence T_e is encoded using RoBERTa to produce contextualized token embeddings:

$$\mathbf{T}_e = \text{RoBERTa}(T_e). \tag{15}$$

Aggregation of Tree Representations. To obtain a condensed representation for each tree, we apply token-level self-attention followed by weighted summation to get the representation of T_e. The entity-level embeddings for all trees are then concatenated as follows:

$$\mathbf{t}_e = \text{ASum}(\mathbf{T}_e), \tag{16}$$

$$\mathbf{T} = [\mathbf{t}_1; \mathbf{t}_2; \ldots; \mathbf{t}_{n_E}]. \tag{17}$$

A further self-attention layer captures interactions among entities and yields a unified knowledge representation. The final vector $\mathbf{t}_E$ encodes structured, dialogue-specific knowledge for downstream recommendation and generation:

$$\mathbf{t}_E = \text{ASum}(\mathbf{T}). \tag{18}$$

3.5 Multimodal Information Integration Module

Our framework incorporates information from both conversations and KGs. The KG provides two complementary forms of embeddings: entity embeddings $\mathbf{E}$ and knowledge tree embeddings $\mathbf{T}$. Aligning them helps ensure that representations of the same entity remain close in the semantic space. Since the entity embedding matrix $\mathbf{E} = [\mathbf{e}_1; \mathbf{e}_2; \ldots; \mathbf{e}_{n_E}]$ contains all mentioned entities, we obtain a single entity-level representation using the same attention-based aggregation introduced in Section 3.4:

$$\mathbf{e}_E = \text{ASum}(\mathbf{E}). \tag{19}$$

To align $\mathbf{e}_E$ and $\mathbf{t}_E$, we employ contrastive learning(Ma and Collins 2018), encouraging positive pairs to be close and negative pairs to be separated. Specifically, given two entity sequences E_i and E_j from a mini-batch, $\mathbf{e}_{E_i}$ and $\mathbf{t}_{E_j}$ form a positive pair if $E_i = E_j$, and a negative pair otherwise. For a batch of b conversations with sequences $[E_1, \ldots, E_b]$, the contrastive loss is:

$$M_{i,j} = \mathbb{I}[E_i = E_j], \tag{20}$$

$$L_{align} = \sum_{i=1}^{b}\sum_{j=1}^{b} -\log\frac{e^{(\mathbf{e}_{E_i}\cdot\mathbf{t}_{E_j})/\tau}\cdot M_{i,j}}{\sum_{k=1}^{b}(e^{(\mathbf{e}_{E_i}\cdot\mathbf{t}_{E_k})/\tau}\cdot(1-M_{i,k}))}, \tag{21}$$

where τ is a temperature hyperparameter that controls the sharpness of similarity discrimination.

3.6 Prompt Learning Module

The overall model is divided into four parameter groups: the base PLM Θ_{plm}, the user preference extraction module Θ_{user}, the knowledge tree enhanced module Θ_{tree}, and the task-specific soft prompts Θ_{prompt}. We adopt DialoGPT (Zhang et al. 2019), a Transformer-based autoregressive model pretrained on large-scale Reddit dialogues, as the backbone PLM. Its parameters Θ_{plm} remain frozen during training, while the other modules are learnable. To reduce hallucinations in recommendation, we follow UniCRS (Wang et al. 2022) and append the generated conversation response to the dialogue context C when constructing the recommendation input. The final inputs for generation and recommendation tasks are:

$$\widetilde{C}_{\text{gen}} = \text{concat}(P_{\text{rgcn}}, P_{\text{tree}}, P_{\text{user}}, P_{\text{prompt(C)}}, C), \tag{22}$$

$$\widetilde{C}_{\text{rec}} = \text{concat}(P_{\text{rgcn}}, P_{\text{tree}}, P_{\text{user}}, P_{\text{prompt(E)}}, C). \tag{23}$$

The training is conducted in two stages. In Stage 1, Θ_{user} and Θ_{tree} are pretrained via self-supervised response generation. In Stage 2, we initialize Θ_{prompt} randomly and jointly optimize Θ_{user}, Θ_{tree}, and Θ_{prompt}. For the recommendation task, we compute:

$$\mathbf{O} = \text{Pooling}(f(\widetilde{C}_{\text{rec}} \mid \Theta)), \quad \hat{R} = \text{softmax}(\mathbf{O}\mathbf{I}^\top), \tag{24}$$

$$L_{\text{rec}} = -\sum_{j=1}^{N}\sum_{i=1}^{M}[Y_j^i \log\hat{R}_j^i + (1-Y_j^i)\log(1-\hat{R}_j^i)], \tag{25}$$

$$L_{\text{all}} = L_{\text{rec}} + \alpha L_{\text{user}} + \beta L_{\text{align}}, \tag{26}$$

where $f(\cdot \mid \Theta)$ is the output of DialoGPT with input tokens and parameters. Pooling can be mean, max, or [CLS] embedding. α and β are weighting hyperparameters. For the conversation task, we follow a similar two-stage training strategy, replacing L_{rec} with standard generation loss L_{conv}.

Model	INSPIRED						ReDial					
	R@10	R@50	N@10	N@50	M@10	M@50	R@10	R@50	N@10	N@50	M@10	M@50
ReDial	0.106	0.223	0.049	0.075	0.031	0.037	0.050	0.186	0.024	0.053	0.015	0.021
KBRD	0.151	0.278	0.102	0.128	0.086	0.091	0.189	0.372	0.101	0.141	0.074	0.082
KGSF	0.178	0.294	0.109	0.133	0.088	0.093	0.177	0.369	0.094	0.137	0.069	0.078
TG-ReDial	0.173	0.331	0.110	0.144	0.091	0.098	0.179	0.353	0.101	0.140	0.078	0.086
UNICRS	0.262	0.406	0.159	0.193	0.131	0.138	0.213	0.414	0.119	0.163	0.090	0.100
KERL	0.206	0.380	0.128	0.168	0.113	0.117	0.216	0.421	0.122	0.161	0.087	0.097
DCRS	0.267	0.410	0.162	0.196	0.132	0.139	0.217	0.426	0.119	0.166	0.090	0.101
PCRS-TKA	**0.278***	**0.438***	**0.187***	**0.222***	**0.158***	**0.166***	**0.221***	**0.432***	**0.123***	**0.170***	**0.093***	**0.103***
Improvement (%)	*4.12%*	*6.83%*	*15.43%*	*13.27%*	*19.70%*	*19.42%*	*1.84%*	*1.41%*	*3.36%*	*2.41%*	*3.33%*	*1.98%*

Table 1: Evaluation results for recommendation task across two datasets (INSPIRED and ReDial). "R", "N", and "M" refer to "recall", "ndcg", and "mrr", respectively.

4 Experiments

4.1 Experimental Setup

Dataset. We conducted experiments on the ReDial (Li et al. 2018) and INSPIRED (Hayati et al. 2020) datasets. Both datasets were constructed using the Amazon Mechanical Turk (AMT) platform. The ReDial dataset contains 10,006 dialogues, with an average of 18.2 utterances per dialogue and 14.5 words per utterance, covering 6,281 movies. In contrast, the INSPIRED dataset includes 1,001 dialogues, with an average of 35.7 utterances per dialogue and 19 words per utterance, encompassing 1,472 movies. In the experiment, we split each dataset into training, validation, and test sets in an 8:1:1 ratio. For each dialogue, we incrementally added one round of utterances starting from the first round to create new data, thereby expanding the dataset.

Knowledge Graph. DBpedia (Auer et al. 2007) is a large-scale KG extracted from the structured content of Wikipedia. It contains 5,040,986 high-frequency entities with their corresponding 927 relations and 24,267,796 triplets. Since the entire DBpedia graph is too large, we collected all entities in the dataset corpus using TagMe (Ferragina and Scaiella 2010) tool and extracted their one-hop triples, forming the subgraph used as the external KG in our experiment.

Evaluation Metrics. We conducted two types of evaluations: recommendation and conversation. For the recommendation task, we used recall@k (k=10, 50), ndcg@k (k=10, 50), and mrr@k (k=10, 50) as metrics. For the conversation task, we employed both automatic and manual evaluations. The automatic evaluation used word-level distinct-n (n=2, 3, 4) to measure response diversity. Additionally, for manual evaluation, we randomly selected 100 conversations and their model-generated responses and invited ten annotators to score them. The evaluation assessed four aspects: *Fluency*, *Informativeness(Inform.)*, *Consistency*, and *Accuracy*, with scores ranging from 0 to 5. Further details are provided in our public code repository.

Benchmark Models. We compared PCRS-TKA with several state-of-the-art models, including: **ReDial**: Integrates an autoencoder-based recommendation module with an HRED-based conversation module. **KBRD**: Utilizes DBpedia to enhance entity representations, combining a self-attention recommendation module with a Transformer-based conversation module. **KGSF**: Integrates ConceptNet (Speer, Chin, and Havasi 2017) and DBpedia, using mutual information maximization to align their semantic spaces. **TG-ReDial** (Zhou et al. 2020b): Introduces a topic prediction task, using SASRec (Kang and McAuley 2018) for recommendation, BERT (Devlin et al. 2019) for topic prediction, and GPT-2 (Gao, Fisch, and Chen 2021) for response generation. **UniCRS**: Uses prompt learning to guide a PLM for both tasks, integrating DBpedia information into the prompt. **KERL**: Introduces a Wikipedia KG with entity descriptions and a PLM encoder for better entity representations. **DCRS**: Incorporates a knowledge-aware retriever to collect selective analogues from dialogue histories.

Implementation Details. We conducted the experiment on one single NVIDIA-V100 GPU. We used grid search to choose the hyperparameters. After searching, we used AdamW with epsilon set to 0.01, learning rate set to 5e-4 for first-stage pre-training, and 1e-4 for second-stage training for both recommendation and conversation tasks. The batch size was set to 64 for the recommendation task and 8 for the conversation task. The soft prompt token length was set to 10 for the recommendation task and 20 for the conversation task. For all baseline methods, hyperparameters were also tuned using grid search. More details can be found in our public code repository.

4.2 Overall Performance

Recommendation Task. As shown in Table 1, our model consistently outperforms all baseline models on both INSPIRED and ReDial datasets. Specifically, it achieves significant improvements over the best-performing baseline, DCRS, with improvements of up to 19.70% on INSPIRED and 3.33% on ReDial in key metrics like mrr@10. The larger improvements on the INSPIRED dataset are likely due to the greater number of utterances per dialogue, which enables PCRS-TKA to capture more complex user preferences. Those using external KGs, such as KBRD and KGSF, generally outperform the basic ReDial model. And KGSF outperforms KBRD for the integration of ConceptNet. However, these methods show limited improvement compared to UniCRS, which leverages PLMs and prompt learning.

Model	INSPIRED			ReDial		
	Dist-2	Dist-3	Dist-4	Dist-2	Dist-3	Dist-4
ReDial	0.313	1.237	2.562	0.070	0.279	0.643
KBRD	0.567	2.017	3.621	0.094	0.488	1.004
KGSF	0.657	2.822	5.992	0.110	0.656	1.729
TG-ReDial	0.778	2.825	5.511	1.016	1.487	1.642
UniCRS	3.949	6.004	7.082	0.899	1.267	1.390
KERL	3.102	5.194	6.548	0.797	1.511	1.783
DCRS	4.383	6.140	7.312	0.922	1.313	1.422
PCRS-TKA	**6.805***	**9.906***	**10.804***	**1.162***	**1.676***	**1.845***
Improvement (%)	*55.23%*	*61.33%*	*47.74%*	*26.02%*	*27.65%*	*29.74%*

Table 2: Evaluation results for conversation task across two datasets (INSPIRED and ReDial). "Dist" refers to "distinct".

Models	Fluency	Inform.	Consistency	Accuracy
ReDial	2.57	2.11	1.96	3.04
KGSF	3.08	1.98	1.86	2.76
UniCRS	3.43	3.36	2.98	3.14
KERL	3.47	3.44	3.28	3.23
DCRS	3.93	3.36	3.66	3.07
PCRS-TKA	**4.18***	**3.94***	**3.77***	**3.27***
Improvement (%)	*6.36%*	*14.53%*	*3.01%*	*1.24%*

Table 3: Human evaluation results on INSPIRED dataset for conversation task.

KERL introduces a Wikipedia KG with entity descriptions and DCRS uses a knowledge-aware retriever. Our model, utilizes PLMs to process tree-structured KGs and incorporates user preference extraction from multi-turn dialogues, leading to its superior performance over all baselines.

Conversation Task. Table 2 presents the evaluation results for the conversation task, where PCRS-TKA outperforms all baselines across both datasets in distinct-n metrics. Specifically, it improves up to 55.23% in distinct-2 on the INSPIRED dataset and 29.74% in distinct-4 on the ReDial dataset compared to the best baseline, DCRS. These results demonstrate that PCRS-TKA generates more diverse responses, crucial for engaging and natural conversations. While distinct-n metrics measure lexical diversity, they don't fully capture dialogue quality. Table 3 presents human evaluation results, where PCRS-TKA outperforms all baselines in fluency, informativeness, consistency, and accuracy. Our model achieves improvements of 6.36% in fluency, 14.53% in informativeness, 3.01% in consistency, and 1.24% in accuracy over DCRS. The superior performance is driven by two factors: integrating structured dialogue-specific knowledge from the KG and leveraging user preferences across multi-turn conversations for better alignment. While DCRS excels at retrieving contextualized information, PCRS-TKA captures conversational dynamics more effectively. UniCRS surpasses KGSF and ReDial by leveraging PLMs for text generation. KGSF benefits from external KGs but doesn't fully capture the nuances of dialogue like PCRS-TKA does.

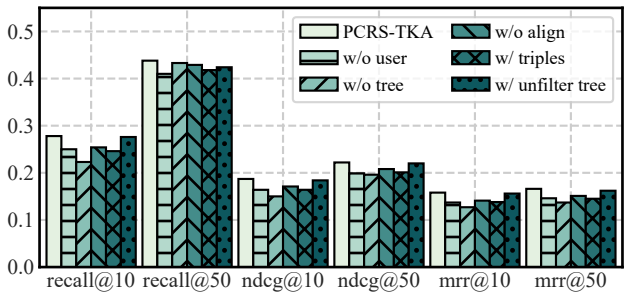

Figure 2: Ablation study on INSPIRED dataset for recommendation task.

Model	R@10	R@50	N@10	N@50	M@10	M@50
PCRS-TKA (BERT)	0.254	0.441	0.174	0.214	0.148	0.156
-all (BERT)	0.250	0.426	0.162	0.201	0.138	0.143
PCRS-TKA (GPT2)	0.258	0.430	0.162	0.204	0.140	0.147
-all (GPT2)	0.230	0.422	0.153	0.196	0.129	0.138
PCRS-TKA (origin)	0.278	0.438	0.187	0.222	0.158	0.166
-all (origin)	0.262	0.406	0.159	0.193	0.131	0.138

Table 4: Generalizability analysis on INSPIRED for recommendation task.

4.3 Ablation Study

We performed ablation experiments to evaluate the contribution of each component in our approach. Specifically, we excluded the knowledge tree enhanced prompts, user preference prompts, and the multimodal information integration module, separately, during both the pre-training and training phases. As shown in Figure 2, removing any component leads to performance degradation, confirming that all components are crucial for improving the recommendation task. To further validate the design of our core knowledge tree module, we conducted deeper ablations. For instance, replacing the hierarchical tree with simple unordered triples caused a significant performance drop, confirming that the tree structure is critical for preserving a clear reasoning path for the PLM, a finding that aligns with prior work like TREA (Li et al. 2023). Similarly, removing our context-aware filtering mechanism also degraded performance. This result, combined with our hyperparameter analysis showing that performance declines as tree degree increases, provides strong evidence that our RAG-style filtering is significant for reducing noise. The ability to isolate such a relevant, filtered subgraph also gives our framework strong potential for robustness against dynamic or noisy KGs.

4.4 Generalizability Analysis

To evaluate the generalizability of PCRS-TKA across various PLMs, we substituted RoBERTa with BERT and DialoGPT with GPT-2 to assess whether the framework could be seamlessly adapted to different model architectures while maintaining its effectiveness. The experimental results, presented in Table 4, demonstrate the consistent performance PCRS-TKA across these alternative PLMs. These findings

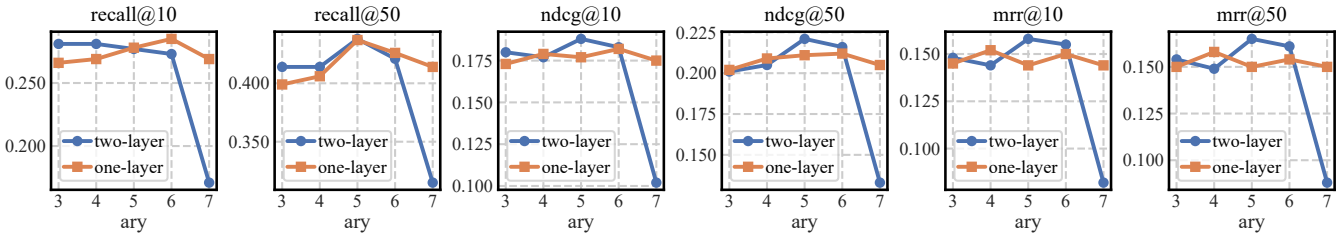

Figure 3: Model performance comparison with varing degree and depth of knowledge tree on INSPIRED dataset.

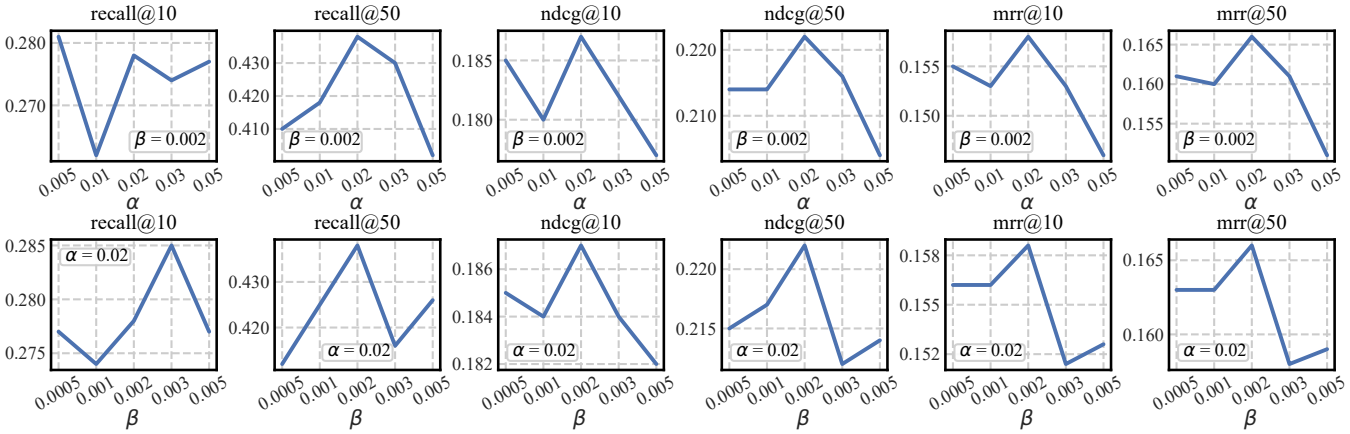

Figure 4: Model performance comparison with varing α and β on INSPIRED dataset.

underscore the flexibility and adaptability of our approach, confirming its applicability to diverse model architectures.

Notably, in principle, we could use larger LLMs to replace RoBERTa and DialoGPT. However, LLMs require substantial computational resources, and prior studies (Schick and Schütze 2021; Subramanian, Elango, and Gungor 2025) suggest in tasks with modest semantic complexity and generalization demands, such as movie recommendations, LLMs may not surpass PLMs. In contrast, PLMs like RoBERTa and DialoGPT strike a balance between performance and efficiency, making them more practical in resource-constrained environments. In this paper, we focus on leveraging PLMs to achieve high-quality recommendations while maintaining efficiency.

4.5 Hyperparameter Sensitivity Analysis

Analysis on the degree and depth of the knowledge tree. We conduct parameter-tuning experiments to study the impact of the knowledge tree's depth and degree on recommendation performance. As shown in Figure 3, one-layer and two-layer trees yield similar optimal results, suggesting that a two-hop path captures sufficient information—e.g., (movie, starring, actor). For one-layer trees, performance is stable across degrees. However, for two-layer trees, performance declines sharply as the degree increases. We attribute this to deeper and wider trees introducing contextually irrelevant entities, diluting critical first-hop signals with noise. This highlights the importance of our context-aware filtering and suggests that the tree's depth and degree should be carefully adapted to the KG structure and dialogue context to minimize noise.

Analysis on loss balancing. The recommendation task loss function includes two hyperparameters, α and β, to balance its components. The parameter α controls collaborative signal weight: as α increases, user preference influence grows, enhancing performance until excessive emphasis over shadows dialogue context, leading to a decline. Similarly, β tunes alignment: as β rises, alignment improves performance up to a point, beyond which can disrupt information learned by entity embeddings via RGCN, while a smaller proportion facilitates better integration of entity and tree embeddings. As shown in Figure 4, optimal performance occurs with $\alpha \approx 0.02$ and $\beta \approx 0.002$. These values remained stable across datasets, indicating consistent performance trends and an efficient tuning process.

5 Conclusion

In this paper, we presented PCRS-TKA, a novel framework for conversational recommendation featuring a synergistic architecture that integrates pretrained language models with knowledge graphs. Our approach uniquely augments static, global KG embeddings with dynamic, context-aware knowledge trees constructed in a RAG-style, enabling direct and fine-grained reasoning by the PLM. Furthermore, it explicitly models collaborative preferences by learning to encode this signal into the prompts via an auxiliary task. Extensive experiments demonstrate that this design allows PCRS-TKA to consistently outperform strong baselines in both recommendation quality and dialogue generation.

Acknowledgments

This work was supported in part by the Natural Science Foundation of Anhui Province (Grant No. 2508085QF211), the National Key R&D Program of China (Grant No. 2023YFF0725001), the National Natural Science Foundation of China (Grant Nos. 62506348, 92370204, 62406141), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023B1515120057), the Key-Area Special Project of Guangdong Provincial Ordinary Universities (Grant No. 2024ZDZX1007), the CCF-1688 Yuanbao Cooperation Fund (Grant No. CCF-Alibaba2025005), the China Postdoctoral Science Foundation (Grant No. GZC20252740), and the Education Bureau of Guangzhou.

References

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. DBpedia: a nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, 722–735. Berlin, Heidelberg: Springer-Verlag. ISBN 3540762973.

Bouraoui, Z.; Camacho-Collados, J.; and Schockaert, S. 2020. Inducing Relational Knowledge from BERT. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 7456–7463.

Chen, H.; Liu, X.; Yin, D.; and Tang, J. 2017. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2): 25–35.

Chen, Q.; Lin, J.; Zhang, Y.; Ding, M.; Cen, Y.; Yang, H.; and Tang, J. 2019. Towards Knowledge-Based Recommender Dialog System. In *Conference on Empirical Methods in Natural Language Processing*.

Chen, Y.; Fu, Q.; Yuan, Y.; Wen, Z.; Fan, G.; Liu, D.; Zhang, D.; Li, Z.; and Xiao, Y. 2023. Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, 245–255. ISBN 9798400701245.

Dao, H.; Deng, Y.; Le, D. D.; and Liao, L. 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*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

Ferragina, P.; and Scaiella, U. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). *Proceedings of the 19th ACM international conference on Information and knowledge management*.

Gao, C.; Lei, W.; He, X.; De Rijke, M.; and Chua, T.-S. 2021. Advances and Challenges in Conversational Recommender Systems: A Survey. *AI Open*, 2: 100–126.

Gao, T.; Fisch, A.; and Chen, D. 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.

Hayati, S. A.; Kang, D.; Zhu, Q.; Shi, W.; and Yu, Z. 2020. INSPIRED: Toward Sociable Recommendation Dialog Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8142–8152.

Ji, Y.; Sun, Y.; Zhang, Y.; Wang, Z.; Zhuang, Y.; Gong, Z.; Shen, D.; Qin, C.; Zhu, H.; and Xiong, H. 2025. A comprehensive survey on self-interpretable neural networks. *arXiv preprint arXiv:2501.15638*.

Kang, W.-C.; and McAuley, J. 2018. Self-Attentive Sequential Recommendation . In *2018 IEEE International Conference on Data Mining (ICDM)*, 197–206.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 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.

Li, R.; Kahou, S.; Schulz, H.; Michalski, V.; Charlin, L.; and Pal, C. 2018. Towards deep conversational recommendations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 9748–9758.

Li, W.; Wei, W.; Qu, X.; Mao, X.-L.; Yuan, Y.; Xie, W.; and Chen, D. 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. Toronto, Canada: Association for Computational Linguistics.

Liu, Y.; Pei, J.; Zhang, W.-N.; Li, M.; Che, W.; and de Rijke, M. 2025. Augmentation with Neighboring Information for Conversational Recommendation. *ACM Trans. Inf. Syst.*, 43(3).

Liu, Z.; Lin, W.; Shi, Y.; and Zhao, J. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Chinese Computational Linguistics: 20th China National Conference, CCL 2021, Hohhot, China, August 13–15, 2021, Proceedings*, 471–484. ISBN 978-3-030-84185-0.

Lu, Y.; Bao, J.; Song, Y.; Ma, Z.; Cui, S.; Wu, Y.; and He, X. 2021. RevCore: Review-Augmented Conversational Recommendation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1161–1173.

Ma, Z.; and Collins, M. 2018. Noise Contrastive Estimation and Negative Sampling for Conditional Models: Consistency and Statistical Efficiency. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3698–3707.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? 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), 2463–2473.

Qin, C.; Chen, X.; Wang, C.; Wu, P.; Chen, X.; Cheng, Y.; Zhao, J.; Xiao, M.; Dong, X.; Long, Q.; et al. 2025. Scihorizon: Benchmarking ai-for-science readiness from scientific data to large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 5754–5765.

Qiu, Z.; Tao, Y.; Pan, S.; and Liew, A. W. 2025. Knowledge Graphs and Pretrained Language Models Enhanced Representation Learning for Conversational Recommender Systems. *IEEE Trans. Neural Networks Learn. Syst.*, 36(4): 6107–6121.

Schick, T.; and Schütze, H. 2021. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2339–2352.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; van den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling Relational Data with Graph Convolutional Networks. In Gangemi, A.; Navigli, R.; Vidal, M.-E.; Hitzler, P.; Troncy, R.; Hollink, L.; Tordai, A.; and Alam, M., eds., *The Semantic Web*, 593–607. Cham: Springer International Publishing. ISBN 978-3-319-93417-4.

Shen, D.; Qin, C.; Wang, C.; Dong, Z.; Zhu, H.; and Xiong, H. 2021. Topic modeling revisited: A document graph-based neural network perspective. *Advances in neural information processing systems*, 34: 14681–14693.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, 4444–4451.

Subramanian, S.; Elango, V.; and Gungor, M. 2025. Small Language Models (SLMs) Can Still Pack a Punch: A survey. *CoRR*, abs/2501.05465.

Tong, G.; Li, D.; and Liu, X. 2024. An Improved Model Combining Knowledge Graph and GCN for PLM Knowledge Recommendation. *Soft Computing*, 28(6): 5557–5575.

Wang, C.; Liu, Q.; Wu, R.; Chen, E.; Liu, C.; Huang, X.; and Huang, Z. 2018. Confidence-aware matrix factorization for recommender systems. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 32.

Wang, C.; Zhu, H.; Hao, Q.; Xiao, K.; and Xiong, H. 2021a. Variable interval time sequence modeling for career trajectory prediction: Deep collaborative perspective. In *Proceedings of the Web Conference 2021*, 612–623.

Wang, C.; Zhu, H.; Wang, P.; Zhu, C.; Zhang, X.; Chen, E.; and Xiong, H. 2021b. Personalized and explainable employee training course recommendations: A bayesian variational approach. *ACM Transactions on Information Systems (TOIS)*, 40(4): 1–32.

Wang, T.-C.; Su, S.-Y.; and Chen, Y.-N. 2022. BARCOR: Towards A Unified Framework for Conversational Recommendation Systems.

Wang, X.; He, X.; Cao, Y.; Liu, M.; and Chua, T.-S. 2019. KGAT: Knowledge Graph Attention Network for Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 950–958.

Wang, X.; Zhou, K.; Wen, J.-R.; and Zhao, W. X. 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.

Wu, C.; Wu, F.; Qi, T.; and Huang, Y. 2021. Empowering News Recommendation with Pre-trained Language Models. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, 1652–1656.

Xie, Z.; Wu, J.; Jeon, H.; He, Z.; Steck, H.; Jha, R.; Liang, D.; Kallus, N.; and Mcauley, J. 2024. Neighborhood-Based Collaborative Filtering for Conversational Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, 1045–1050. ISBN 9798400705052.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; Wang, L.; Luu, A. T.; Bi, W.; Shi, F.; and Shi, S. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *CoRR*, abs/2309.01219.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, W. B. 2019. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Annual Meeting of the Association for Computational Linguistics*.

Zhou, K.; Zhao, W. X.; Bian, S.; Zhou, Y.; rong Wen, J.; and Yu, J. 2020a. Improving Conversational Recommender Systems via Knowledge Graph based Semantic Fusion. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Zhou, K.; Zhou, Y.; Zhao, W. X.; Wang, X.; and Wen, J.-R. 2020b. Towards Topic-Guided Conversational Recommender System. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4128–4139.

Zhou, Y.; Zhou, K.; Zhao, W. X.; Wang, C.; Jiang, P.; and Hu, H. 2022. C²-CRS: Coarse-to-Fine Contrastive Learning for Conversational Recommender System. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, 1488–1496. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391320.

Zhu, Y.; Wang, C.; Zhang, Q.; and Xiong, H. 2024. Graph signal diffusion model for collaborative filtering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1380–1390.

Zong, Z.; Ma, B.; Shen, D.; Song, G.; Shao, H.; Jiang, D.; Li, H.; and Liu, Y. 2024. Mova: Adapting mixture of vision experts to multimodal context. *Advances in Neural Information Processing Systems*, 37: 103305–103333.