Leveraging Knowledge Graphs and Large Language Models for Conversational Recommender Systems

A Zero-Shot Approach with Relational Graph-Augmented Reasoning for Movie Recommendation

Derek Yao
University of California, Berkeley
derekyao672@berkeley.edu

Pranav Viswanathan
University of California, Berkeley
viswanathan.pranav@berkeley.edu

Abstract

In this paper, we propose a novel methodology that integrates Knowledge Graphs (KGs) with Large Language Models (LLMs) to enhance contextual reasoning in conversational recommender systems (CRSs). By incorporating both user preferences from conversational histories and external metadata about movies, our approach enables LLMs to perform knowledge-aware reasoning on "hidden relationships" between nodes in the graph, resulting in more relevant recommendations and richer explanations. Our empirical evaluation on a CRS benchmark demonstrates the effectiveness of our method, highlighting its potential to improve zero-shot conversational recommendations and advance personalized, real-time recommendation systems.

1 Introduction

Conversational Recommender Systems (CRSs) represent a critical advancement in the evolution of personalized recommendation technologies. Unlike traditional recommender systems, which primarily rely on static user interaction signals such as clicks, ratings, or purchase history, CRSs engage users in dynamic, multi-turn dialogues. These systems aim to actively elicit user preferences through natural language interactions and provide recommendations that adapt to users' evolving needs and contextual inputs. By incorporating the ability to interpret and respond to nuanced, multi-turn natural language inputs, CRSs introduce the potential for highly personalized and interactive recommendation experiences.

The conversational nature of CRSs imposes unique challenges that differ from those faced by traditional recommended systems. A CRS must not only maintain an understanding of historical user behaviors but also dynamically extract and infer user preferences from ongoing conversations. This requires continuous preference modeling that reflects the evolving nature of user intent across multiple dialogue turns. Additionally, CRSs must generate human-like responses that align with the conversational context, often in zero-shot scenarios where prior training on similar dialogues is unavailable.

Large Language Models (LLMs), with their extensive pre-training on diverse textual corpora, have emerged as powerful tools for addressing these challenges. LLMs excel in generating fluent, context-aware responses and can leverage their expansive knowledge to support zero-shot recommendation tasks. Recent work has provided empirical evidence that LLMs, even in a zero-shot capacity, can outperform even fine-tuned traditional CRS models [3]. Although LLMs have shown promise in CRS tasks such as evaluation [5], and task planning [1], the role of incorporating explainable user preferences into evolving user information and preferences still remains a field largely in its infancy. LLMs still often struggle to incorporate domain-specific knowledge or maintain real-time awareness of users' shifting preferences, which are essential for accurate preference modeling and recommendations.

To overcome these limitations, the integration of Knowledge Graphs (KGs) has shown significant promise in addressing these limitations. Knowledge graphs have proven effective in providing a structured, interconnected representation of domain-specific entities and their relationships, offering rich contextual grounding that enhances both the accuracy and explainability of recommendations [2]. By leveraging KGs, CRSs can supplement the reasoning capabilities of LLMs with domain-specific insights, enabling more precise and interpretable recommendations. However, achieving seamless integration between the structured knowledge of a KG and the generative reasoning capabilities of LLMs remains a major research challenge mainly due to the existence of a

significant modality gap between KGs and LLMs, which hinders a LLM's ability to understand and interpret KG information. While LLMs process sequences of tokens that represent natural language, KGs represent information in a graph-based format. This difference makes it difficult for LLMs to directly interpret the entities and relationships encoded in KGs, limiting their ability to perform this cross-modal reasoning of KG information of user preferences and extraneous data.

2 Overview

Our work focuses on tackling part of this issue by focusing our efforts on providing structured data representation that dynamically captures for contextual reasoning. Our methods range across structured representation, graph-augmented reasoning, and relational data encoding. Recent work has shown that natural language captioning techniques can fail to retain the full complexity of the KG's structure or require transformations that reduce the interpretability and reasoning capacity of the approach needed for real-time adaption and contextual reasoning [4]. The importance of contextual reasoning in conversational systems was emphasized, suggesting that directly encoding relational context—rather than embedding entities into vectors—better enables the model to respond to real-time, dynamically evolving user preferences. By providing relational, structured information directly to the LLM, we allow for richer, more context-aware reasoning at each conversational turn, which is especially valuable in zero-shot settings, where prior knowledge of the specific task is minimal or absent.

In this work, we use the ReDIAL dataset [1] as our benchmark for evaluating conversational recommendation systems (CRS). ReDIAL is a conversational dataset designed specifically for training and testing movie recommendation systems that engage in multi-turn dialogues. The dataset consists of over 10,000 movie recommendations, accompanied by conversational history that simulates a user's preferences and interactions with a CRS over time. Each conversation in ReDIAL contains multiple turns, where users discuss their preferences for movies, and the system suggests relevant films based on the ongoing dialogue. We also use a Movie Metadata dataset from Kaggle [2] which provides structured metadata about movies (such as genres, directors, and actors) and also includes descriptions of movies.

3 Methodology

Our work here can be best summarized as being three-fold: consisting of parts focusing on Structured Representation, Graph-Augmented Reasoning, and Relational Data Encoding.

3.1 Knowledge Graph Construction

The Knowledge Graph (KG) is the core of our methodology, providing a structured representation of the movie domain. It encodes detailed relationships between entities such as movies, actors, directors, and genres, enabling rich domain-specific reasoning.

3.1.1 Primary Relationships

- **Actor-Movie**: Captures the participation of an actor in a movie.
- **Director-Movie**: Links directors to the movies they directed.
- **Genre-Movie**: Represents the genre(s) associated with a given movie.

3.2 Schema-Augmented Reasoning

Schema-Augmented Reasoning involves embedding unique domain-specific knowledge directly into our graph nodes. By providing structured contextual knowledge from the KG, our approach here functions as a form of knowledge injection. This explicit method allows models to leverage symbolic relationships encoded within the graph.

3.3 Relational Data Encoding

Relationships between entities in the nodes are explicitly encoded as attributes, making the hidden structure of the knowledge graph interpretable for reasoning tasks. By "hidden structure," we refer to newly defined relationships that emerge after linking nodes based on their attributes. For instance, after establishing connections between directors and movies, and between movies and genres, we infer a new relationship between directors and genres. These hidden relationships, which are not directly present in the raw data but become apparent through the graph structure, are critical for enriching the context provided to the LLM. We assign

[1] https://redialdata.github.io/website/

[2] https://www.kaggle.com/datasets/thedevastator/imdb-movie-ratings-dataset

weights to these relationships based on their frequency, allowing us to prioritize the most significant connections for inclusion in the LLM context.

3.3.1 Hidden Relationships

- **Actor-[movie]-Actor**: Captures common pairs of actors who frequently collaborate across movies.
- **Actor-[genre]-Actor**: Identifies actors who consistently perform within the same genre.
- **Actor-[movie]-Director**: Links actors and directors who have worked together on the same movie.
- **Director-[genre]-Director**: Connects directors who commonly create movies within shared genres.

By structuring these relationships alongside the metadata of the movies in our dataset, we aim to directly inject structural relational data as additional context when prompting our LLM. This approach enables context-aware dynamic reasoning over recommendations based on user preferences within the conversation.

4 Evaluation

We evaluate the the proposed KG-enhanced LLM framework, against a baseline zero-shot LLM model on the ReDIAL dataset to assess the effectiveness of the KG in improving recommendation accuracy and relevance.

4.1 Evaluation Setup

To evaluate recommendation quality, we use a masking strategy applied to the last movie mentioned recommendation of each dialogue turn where the movie suggestion was not explicitly disliked by the user. The rationale behind this approach is to evaluate how well the model can handle unstructured dialogue while relying on explicitly, structured knowledge from the KG. This masking process helps isolate the movie suggestions and assess their relevance based on the information available in the preceding dialogue context, without the influence of user feedback where negative preferences are already clear.

We then evaluate the recommendation performance using Recall at 5 (Rec@5), a metric commonly used in recommendation systems to measure the fraction of relevant items that appear in the top-5 ranked results. Rec@5 is an ideal choice here as it evaluates how effectively the model suggests 5 movie recommendations that align with user preferences and the relational data embedded in the KG.

4.2 Removal of Repeated Items

Similar to Qiu in 2024 [3], we also argue it's beneficial we observe a high pattern of repeated items in the corpus of dialogues in the ReDIAL dataset. Given the nature of recommendation conversations between two users, it is more probable that items repeated during a conversations are intended for discussion rather than serving as recommendations. The presence of repeats in recommendations would not only serve to overly complicate our KG framework but could also risk shortcut learning where results would be inconsistent with a designer's original intent of evaluation. Because of this, we remove repeated items in order to better understand a model's recomemndation ability.

4.3 Evaluation Results

4.3.1 Recall @ 5

To assess the impact of integrating the KG, we compare the KG-augmented LLM model's Rec@5 performance against that of a baseline zero-shot LLM model. The baseline model generates recommendations purely based on the unstructured dialogue context present in the ReDIAL dialogue context without any external structured knowledge. In contrast, the KG-enhanced model integrates relational data about movies, directors, actors, genres, as well as the "hidden" relationships between these nodes. Using OpenAI's Chat GPT 4 model, we find that

Finding 1 - The KG-enhanced model achieves a higher recall at 5 compared to the baseline model, demonstrating the benefit of incorporating structured domain knowledge.

4.3.2 Hallucinations

To evaluate hallucinations in recommendations, we examine the fraction of predicted movie IDs and names that do not match any entries in the dataset. This analysis highlights the model's ability to generate valid recommendations grounded in the provided context.

Finding 2 - KG-augmented LLM hallucinated more on average for out-of-dataset item titles than the baseline zero-shot LLM model.

Model	Recall@5	Hallucinations in IDs (%)	Hallucinations in Titles (%)
GPT-4o w/ KG Context	8.16%	31.33%	2.86%
GPT-4o	3.06%	16.16%	1.63%

Table 1: Comparison of Rec@5 and hallucination rates for KG-enhanced and baseline models.

This was an unusual finding and deserves more attention in future work to understand why. We theorize that the hidden relationship information injected from our knowledge graph encourages the model to extrapolate beyond the constraints explicitly defined in the model's prompt. By incorporating relational data about actors, directors, genres, and other features, the model may attempt to generate recommendations that align with inferred relationships, even if those items are not present in the dataset.

Discussion

4.4 Temporal Knowledge Graphs

While our results were not as strong as we had hoped for, we still believe the area of KGs and LLMs to be a strong field of research still in its infancy. As reflected in our original project proposal, we had invested a considerable amount of time and work trying to implement Graphiti as a python library solution to leverage the power of Temporal Knowledge Graphs and LLMs for Agentic downstream tasks. While ultimately this solution became infeasible due to cost, in future work we can certainly see a use case for revisiting our work here and adding a temporal aspect to our KGs, a key piece of information that could help our LLMs understand how user's preferences change over time.

4.5 Alternative Models

We also believe that it would have been interesting to see how other LLMs would've performed as alternatives to just OpenAI. BAIZE and Vicuna, which are representative open-sourced LLMs fine-tuned on LLAMA-13B stand out as possible alternatives to OpenAI's zero-shot recommender. Comparing metrics across the board could yield insights how aept each model is for movie recommendation.

4.6 Graph Entity Captioning

Future work can also more heavily emphasize the Graph Entity Captioning aspect of the project by attempting more advanced methods such as fine-tuning a Transformer-based-model such as T5 or BERT to generate natural language captions to pass into the LLM's prompt. At each conversational turn, the KG could be queried to retrieve relevant entities and relationships based on the user's inputs and produce captions to incorporate into more "natural" conversational context.

4.7 Evaluation Metrics

Our primary evaluation, informed by the Netflix research paper, was recall at K, specifically with K = 5. A recommendation is considered successful if at least one of the 5 provided recommendations matches our ground truth value, which is hidden from the LLM. However, this definition of success has inherent limitations when assessing the broader goal of a conversation recommendation system. A recommendation system intends to provide high-quality recommendations that align with the initiator's preferences. While recall at K provides a concrete measure of how well a model identifies a hidden recommendation, it does not account for situations where a recommended movie might be relevant or liked by the user. This is a limitation in our dataset and process, as the dataset lacks the granularity to evaluate the quality of recommendations outside of the defined ground truth value. Ideally, recommendations would be evaluated using explicit feedback from the receiver of the recommendation.

In terms of hallucination for our predictions, a hallucination is defined as the recommendation of a movie that does not fall within our dataset. However, it is important to note that, in this case, hallucinations are consistently movies that exist, and could potentially be high-quality recommendations. It is important to highlight the limitations of measuring hallucination purely based on dataset presence, as it does not consider quality, similar to Recall @ K.

References

[1] Jiabao Fang et al. *A Multi-Agent Conversational Recommender System*. 2024. arXiv: 2402.01135 [cs.IR]. URL: https://arxiv.org/abs/2402.01135.

[2] Zhankui He et al. “Large Language Models as Zero-Shot Conversational Recommenders”. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM ’23. Birmingham, United Kingdom: Association for Computing Machinery, 2023, pp. 720–730. ISBN: 9798400701245. DOI: 10.1145/3583780.3614949. URL: https://doi.org/10.1145/3583780.3614949.

[3] Zhangchi Qiu et al. *Unveiling User Preferences: A Knowledge Graph and LLM-Driven Approach for Conversational Recommendation*. 2024. arXiv: 2411.14459 [cs.CL]. URL: https://arxiv.org/abs/2411.14459.

[4] Muzamil Hussain Syed, Tran Quoc Bao Huy, and Sun-Tae Chung. “Context-Aware Explainable Recommendation Based on Domain Knowledge Graph”. In: *Big Data and Cognitive Computing* 6.1 (2022). ISSN: 2504-2289. DOI: 10.3390/bdcc6010011. URL: https://www.mdpi.com/2504-2289/6/1/11.

[5] Xiaolei Wang et al. “Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*. Ed. by Houda Bouamor, Juan Pino 0001, and Kalika Bali. Association for Computational Linguistics, 2023, pp. 10052–10065. ISBN: 979-8-89176-060-8. URL: https://aclanthology.org/2023.emnlp-main.621.