GenRec: Large Language Model for Generative Recommendation

Jianchao Ji, Zelong Li, Shuyuan Xu, Wenyue Hua, Yingqiang Ge, Juntao Tan, and Yongfeng Zhang(✉)

Department of Computer Science, Rutgers University, New Brunswick, NJ 08854, Canada
{jianchao.ji,zelong.li,shuyuan.xu,wenyue.hua,
yingqiang.ge,juntao.tan,yongfeng.zhang}@rutgers.edu

Abstract. In recent years, Large Language Models (LLMs) have emerged as powerful tools for diverse natural language processing tasks. However, their potential for recommender systems under the generative recommendation paradigm remains relatively unexplored. This paper presents an innovative approach to recommender systems using Large Language Models (LLMs) purely based on raw text data, i.e., using item name or title as item IDs rather than creating meticulously designed user or item IDs. More specifically, we present a novel LLM for Generative Recommendation (GenRec) method that utilizes the expressive power of LLM to directly generate the target item to recommend, rather than calculating the ranking score for each candidate item one by one as in traditional discriminative recommendation. GenRec uses LLM's understanding ability to interpret context, learn user preferences, and generate relevant recommendations. Our proposed approach leverages the vast knowledge encoded in Large Language Models to accomplish recommendation tasks. We formulate specialized prompts to enhance the ability of LLM to comprehend recommendation tasks. Subsequently, we use these prompts to LoRA-fine-tune the LLaMA backbone LLM on the user-item interaction data represented by raw text (using raw item name or title as the item's ID) to capture user preferences and item characteristics. Our research underscores the potential of LLM-based generative recommendation in revolutionizing the domain of recommendation systems and offers a foundational framework for future explorations in this field. We conduct extensive experiments on benchmark datasets, and the experiments show that our GenRec method achieves better results on large datasets. Code and data are are open-source at GitHub (https://github.com/rutgerswiselab/GenRec).

Keywords: Large Language Model · Recommender Systems · Natural Language Processing · Generative Recommendation

1 Introduction

Large Language Models (LLMs) have made a particularly significant milestone in this technological evolution. These LLMs, designed to understand and gen-

N. Goharian et al. (Eds.): ECIR 2024, LNCS 14610, pp. 494–502, 2024.
https://doi.org/10.1007/978-3-031-56063-7_42

erate human-like text, have revolutionized numerous applications, from search engines to chatbots, and have facilitated more natural and intuitive interactions between humans and machines. This paper seeks to explore a relatively new and promising application of these models in recommender systems. Recommender systems have become an integral part of our digital experience. They are the unseen force guiding us through the vast amounts of data, suggesting relevant products on e-commerce websites, recommending movies on streaming platforms, or proposing what news to read or videos to watch. The primary aim of these systems is to predict the individual user preferences and enhance user experience and engagement.

Traditionally, recommender systems have been built around methods such as collaborative filtering [11,19,21], content-based filtering [22,24], and hybrid approaches [1,2,18]. Collaborative filtering leverages user-item interactions, making suggestions based on patterns found in the behavior of similar users or items. On the other hand, content-based filtering uses item features to recommend similar items to those a user has previously interacted with. Hybrid methods attempt to combine the strengths of these two approaches to overcome their limitations.

Despite the progress made with these traditional techniques, there still have some significant challenges. For instance, both collaborative and content-based filtering can hardly handle the issue of data sparsity, given that most users interact with only a small fraction of the total items available. Additionally, because of the computational complexity of processing large interaction matrices, these models often struggle to scale effectively with the growth of users and items.

The integration of language-based LLMs into recommender systems presents an exciting opportunity to address these challenges [3,26]. These models can learn and understand complex patterns in human language, which allows for a more nuanced interpretation of user preferences and a more sophisticated generation of recommendations. A significant number of the prevailing LLM-based recommendation models are trained using meticulously designed user and item IDs [9,15,16,26,27]. These approaches demonstrate an important direction towards LLM-RecSys alignment, since they have the advantage of encoding the crucial collaborative information into the user or item IDs, which helps LLMs to align the content and collaborative information and better learn the relationship between users and items or between items and items, thus enhancing the LLM-based recommendation performance. However, creating effective user or item IDs is not a trivial task, which requires meticulously designed techniques such as sequential indexing, collaborative indexing, content-based indexing, or hybrid indexing [9].

In this paper, we propose a novel pure-text-based large language model for generative recommendation (GenRec). GenRec directly uses the textual item name or title as the ID for the item, eliminating the need to create specifically designed IDs for each item. One of the primary benefits of the GenRec model is that it capitalizes on the rich, descriptive information inherently contained within the item names, which often contain features that can be semantically

analyzed, enabling a better understanding of the item's potential relevance to the user. This could potentially provide more accurate and personalized recommendations, thereby enhancing the overall user experience. The key contributions of this paper can be summarized as follows:

- We highlight the promising paradigm of LLM-based generative recommendation, which directly generates the name of the target item to recommend, rather than traditional discriminative recommendation, which has to calculate a ranking score for each and every candidate item one by one and then sorts them for deciding which to recommend.
- We introduce a novel approach, GenRec, to enhance the generative recommendation performance by properly incorporating the textual information into the generative recommendation model.
- We also illustrate the efficacy of GenRec on practical recommendation tasks, underscoring its prospective abilities for a wider scope of applications.

2 Related Work

The use of Large Language Models (LLMs) for recommender systems has gained significant attention recently [8,14]. These models exhibit great potential in the understanding and modeling of user-item interactions, exploiting rich semantics and long-range dependencies present in user activity data.

The pioneering work of P5 [3,26] illustrates the feasibility of using large language models for generative recommendation. It fine-tunes a large language model backbone to create a unified system capable of handling various recommendation tasks, such as sequential recommendation, direct recommendaitaon, rating prediction, explanation generation, etc. This innovative approach highlighted the effectiveness of LLMs in handling multi-task learning in the recommendation context. To enhance the Large Language Model's comprehension of a user's behavior history, some researchers have attempted to incorporate collaborative information into the training process [9,15,16]. For example, collaborative indexing creates item IDs from the user-item or item-item collaborative information based on spectral collaborative learning [9] or graph collaborative learning [16]. Each item is assigned an ID, ensuring that items of that share similar user behavior also share similar IDs. These IDs are then utilized to represent the items during the training process and to generate recommendations.

These recent studies by P5 and item ID creation methods [3,9,13,15,16,27] provide compelling evidence that, with suitable fine-tuning, LLMs can exhibit remarkable performance in recommendation tasks. This advancement underscores the adaptability and potential of LLMs in recommendation domains. Beyond this, some researchers extend LLMs to address cold start or zero-shot problems [6,20,25], where traditional models often falter due to the lack of historical data or prior information. By leveraging the inherent knowledge encoded within their extensive training data, LLMs can generate meaningful insights and predictions even in these challenging situations. This capability is significantly enhanced with carefully crafted prompts, which guide the models to focus on

relevant aspects of the problem at hand, thereby improving the quality and relevance of the output. Such findings highlight the impact of LLMs in areas where data scarcity or the need for immediate insights presents substantial obstacles.

Besides employing Large Language Models (LLMs) as recommender systems, several researchers also explored the use of LLM as auxiliary tools within a traditional recommender system. This approach involves utilizing LLMs as feature encoders and scoring functions [12]. By doing so, LLMs can process and interpret user data, extracting nuanced features that traditional methods might overlook. These features are then integrated into recommendation algorithms, enhancing its accuracy and relevance. Additionally, LLMs can be used to score and rank recommendations, leveraging their advanced natural language processing capabilities to better align suggestions with user preferences and behaviors. This application of LLMs represents a significant shift from their conventional use, opening new avenues for more personalized and efficient recommender systems.

However, the potential of LLMs to understand and generate purely text-described item IDs as recommendations has not been fully explored. In this paper, we propose a novel approach to purely text-ID-based generative recommendation, leveraging the latest advances in LLMs. We aim to address some of the limitations of previous works and push the boundaries of what is possible in the realm of LLM-based recommender systems.

3 The GenRec Method

Our proposed GenRec framework for LLM-based generative recommendation is simple and effective. The architecture of the proposed framework is illustrated in Fig. 1. Given a user's item interaction sequence, the large language model for Generative Recommendation (GenRec) will format the text-based item names or titles with a prompt. This reformatted sequence is subsequently employed to fine-tune a Large Language Model (LLM). The fine-tuned LLM can then predict subsequent items the user is likely to interact with. In our paper, we select the LLaMA [23] language model as the backbone and use Low-Rank Adaptation (LoRA) [7] for fine-tuning. However, our framework retains flexibility, allowing for seamless integration with any other LLM, thus broadening its potential usability and adaptability.

3.1 Sequence Generation

The initial component of GenRec is a generative function, tasked with producing various sequences that encapsulate user interests. To enhance the model's comprehension of the recommendation task, we have devised multiple prompts that facilitate sequence generation. Take Fig. 2 as an example, we use the user's movie watching history as the training data and use this information to format the training sequence. The sequence consists of three parts: instruction, input and output. The instruction element outlines the specific task of movie recommendation, for which we have created several directives to enhance the LLM's

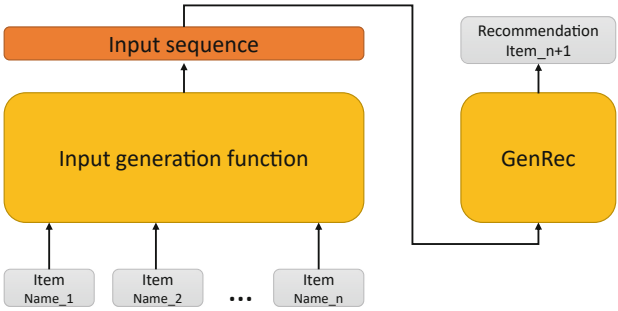

Fig. 1. An illustration of GenRec's simple architecture. Our model generates an input sequence based on the interaction history. Then the model predicts the next item the user may interact with.

comprehension of the ongoing recommendation task. The input represents the history of the user's interactions, excluding the most recent instance. The output is the latest interaction in this record. The primary task for the LLM here is to accurately predict this final interaction.

Interaction history: Pinocchio (1940), Legends of the Fall (1994), Once Were Warriors (1994), In the Name of the Father (1993), Shadowlands (1993), Heavenly Creatures (1994), Quiz Show (1994), In the Line of Fire (1993)
Recommendation Prompt Example:
Instruction: Given the movie viewing habits, what is the most probable movie they will choose to watch next?
input: Pinocchio (1940), Legends of the Fall (1994), Once Were Warriors (1994), In the Name of the Father (1993), Shadowlands (1993), Heavenly Creatures (1994), Quiz Show (1994)
output: In the Line of Fire (1993)

Fig. 2. GenRec prompt and data. GenRec converts the interactive history to a training sequence consisting of instruction, input and output.

Refer to Fig. 2 for an illustration. This figure represents how we utilize a user's history of watched movies as interaction data. Given the prompt "Based on the movie viewing habits, what is the most likely movie they will select to watch next?" and the provided input, we then allow GenRec to generate the subsequent output.

3.2 Training Strategy

In this paper, we use the LLaMA large language model [23] as the backbone for the training of GenRec. The LLaMA model is pre-trained on an expansive

language corpus, offering a valuable resource for our intended purpose of efficiently capturing both user interests and item content information. However, it is important to note that the memory requirements for GPU to fine-tune LLaMA, even the smallest 7-billion parameter version, are pretty substantial.

To circumvent this challenge and conserve GPU memory, we adopt the Low-Rank Adaptation (LoRA) [7] method for fine-tuning and inference tasks over the LLaMA-7b model within the scope of this study. By this measure, we have achieved a significant reduction in the GPU memory requirements. With this optimized approach, we can fine-tune the LLaMA-LoRA model on a single GPU with a memory capacity of 24GB.

4 Experiments

We conduct extensive experiments on two real-world datasets from Amazon [17] and MovieLens [4], respectively, to evaluate the performance of our proposed GenRec approach on recommendation tasks.

4.1 Baseline Methods

In the following, we introduce the baseline models used in this research. These baselines serve as foundational benchmarks against which the performance and efficacy of our proposed methods are evaluated. By introducing these established methods, we aim to provide a clear context and point of reference for understanding the relative strengths and advancements of our contributions.

GRU4Rec [5]: The GRU4Rec model utilizes a session-based recommendation strategy, harnessing the Gated Recurrent Unit (GRU) to discern user preferences. Within a session, this model draws on past preferences as context to predict the subsequent item that a user may interact with.

SASRec [10]: The Self-attentive Sequential Recommendation (SASRec) method incorporates the self-attention technique into its sequential recommendation framework, appropriately applying the usage of both Markov Chains and RNN-driven methods.

P5 [3,26]: The Pre-train, Personalized Prompt, and Predict Paradigm (P5) for LLM-based generative recommendation, which incorporates an array of templates for input and target sequences throughout the training process. This approach dissolves the boundaries between different tasks, promoting a more fluid and integrated training procedure. It has showcased noteworthy performance in the domain of sequential recommendation tasks, underlining its effectiveness and applicability. In this work, we compare with the sequential recommendation performance of P5.

4.2 Performance Comparison

As we can see in Table 1, P5 gains better performance on Amazon Toys dataset, while our GenRec approach has significantly better performance on MovieLens-25M dataset. The possible reasons behind the different performances could be attributed to the distinct nature of the datasets: the MovieLens-25M dataset, unlike Amazon Toys dataset, contains a much richer amount of interaction information due to its larger scale, which provides a more robust understanding of the user's preferences and behavior, thus likely leading to more accurate recommendations even without the need to create meticulously designed item IDs. This observation implies that when the training data is abundant enough, then LLM may be able to directly learn the user-item collaborative information into the purely name- or title-based item IDs by re-optimizing the word embeddings. However, when the amount of training data is not large enough, it would be necessary to create meticulously designed item IDs such as sequential IDs or collaborative IDs [9] so as to "pre-encode" the collaborative information into the IDs for better facilitating LLMs to learn from the training data.

Table 1. Experimental results on Normalize Discounted Cumulative Gain (NDCG@k) and Hit Ratio (HR@k). Bold numbers represent best performance. We use * to indicate that the performance is significantly better than baselines. The significance is at 0.05 level on paired t-test.

Methods	**MovieLens 25M**				**Amazon Toys**			
	HR@5	NDCG@5	HR@10	NDCG@10	HR@5	NDCG@5	HR@10	NDCG@10
GRU4Rec	0.0439	0.0241	0.0753	0.0312	0.0125	0.0076	0.0171	0.0105
SASRec	0.0517	0.0344	0.0878	0.0408	0.0184	0.0124	0.0218	0.0142
P5	0.0688	0.0464	0.1040	0.0577	**0.0239***	**0.0145***	**0.0411***	**0.0201***
GenRec	**0.1034***	**0.0716***	**0.1311***	**0.0837***	0.0190	0.0136	0.0251	0.0157

5 Conclusion

This paper proposes GenRec, a Large Language Model approach for Generative Recommendation based on textual item name or title as IDs. By focusing on the semantic richness of item names as input, GenRec promises personalized and contextually relevant recommendations. Our practical demonstrations highlight GenRec's efficacy and point towards its adaptability across different recommendation scnarios. Furthermore, the flexibility of the GenRec framework facilitates integration with any Large Language Model, hence widening its sphere of potential utility. In terms of future work, there are several directions to explore. We intend to refine the generation of sequences by developing more sophisticated prompts, which could further enhance the model's understanding of recommendation tasks. Additionally, we plan to extend our research to incorporate more complex user interaction data, such as ratings or reviews, which could provide deeper insights into user behavior and preferences. Another direction is to see

how GenRec works with different Large Language Models, since we are curious about the benefits and downsides of using different models.

Acknowledgement. The work was supported in part by NSF IIS-2046457 and IIS-2007907. Any opinions, findings, conclusions or recommendations in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

1. Basilico, J., Hofmann, T.: Unifying collaborative and content-based filtering. In: Proceedings of the Twenty-First International Conference on Machine Learning, p. 9 (2004)
2. Burke, R.: Hybrid recommender systems: survey and experiments. User Model. User-Adap. Inter. **12**, 331–370 (2002)
3. Geng, S., Liu, S., Fu, Z., Ge, Y., Zhang, Y.: Recommendation as language processing (RLP): a unified pretrain, personalized prompt & predict paradigm (p5). In: Proceedings of the 16th ACM Conference on Recommender Systems, pp. 299–315 (2022)
4. Harper, F.M., Konstan, J.A.: The MovieLens Datasets: history and context. ACM Trans. Interact. Intell. Syst. (TIIS) **5**(4), 1–19 (2015)
5. Hidasi, B., Quadrana, M., Karatzoglou, A., Tikk, D.: Parallel recurrent neural network architectures for feature-rich session-based recommendations. In: Proceedings of the 10th ACM Conference on Recommender Systems, pp. 241–248 (2016)
6. Hou, Y., et al.: Large language models are zero-shot rankers for recommender systems. arXiv preprint arXiv:2305.08845 (2023)
7. Hu, E.J., et al.: LoRA: low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
8. Hua, W., Li, L., Xu, S., Chen, L., Zhang, Y.: Tutorial on large language models for recommendation. In: Proceedings of the 17th ACM Conference on Recommender Systems, pp. 1281–1283 (2023)
9. Hua, W., Xu, S., Ge, Y., Zhang, Y.: How to index item IDs for recommendation foundation models. SIGIR-AP (2023)
10. Kang, W.C., McAuley, J.: Self-attentive sequential recommendation. In: 2018 IEEE International Conference on Data Mining (ICDM), pp. 197–206. IEEE (2018)
11. Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., Riedl, J.: GroupLens: applying collaborative filtering to Usenet news. Commun. ACM **40**(3), 77–87 (1997)
12. Li, J., Zhang, W., Wang, T., Xiong, G., Lu, A., Medioni, G.: GPT4Rec: a generative framework for personalized recommendation and user interests interpretation. arXiv preprint arXiv:2304.03879 (2023)
13. Li, L., Zhang, Y., Chen, L.: Prompt distillation for efficient LLM-based recommendation. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, pp. 1348–1357 (2023)
14. Li, L., Zhang, Y., Liu, D., Chen, L.: Large language models for generative recommendation: a survey and visionary discussions. arXiv:2309.01157 (2023)
15. Lin, X., Wang, W., Li, Y., Feng, F., Ng, S.K., Chua, T.S.: A multi-facet paradigm to bridge large language model and recommendation. arXiv:2310.06491 (2023)
16. Mei, K., Zhang, Y.: LightLM: a lightweight deep and narrow language model for generative recommendation. arXiv:2310.17488 (2023)

17. Ni, J., Li, J., McAuley, J.: Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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), pp. 188–197 (2019)
18. Pazzani, M.J.: A framework for collaborative, content-based and demographic filtering. Artif. Intell. Rev. **13**, 393–408 (1999)
19. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: GroupLens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, pp. 175–186 (1994)
20. Sanner, S., Balog, K., Radlinski, F., Wedin, B., Dixon, L.: Large language models are competitive near cold-start recommenders for language-and item-based preferences. In: Proceedings of the 17th ACM Conference on Recommender Systems, pp. 890–896 (2023)
21. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web, pp. 285–295 (2001)
22. Son, J., Kim, S.B.: Content-based filtering for recommendation systems using multiattribute networks. Expert Syst. Appl. **89**, 404–412 (2017)
23. Touvron, H., et al.: LlaMA: open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
24. Van Meteren, R., Van Someren, M.: Using content-based filtering for recommendation. In: Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 workshop, vol. 30, pp. 47–56. Barcelona (2000)
25. Wang, L., Lim, E.P.: Zero-shot next-item recommendation using large pretrained language models. arXiv preprint arXiv:2304.03153 (2023)
26. Xu, S., Hua, W., Zhang, Y.: OpenP5: benchmarking foundation models for recommendation. arXiv:2306.11134 (2023)
27. Zheng, B., Hou, Y., Lu, H., Chen, Y., Zhao, W.X., Wen, J.R.: Adapting large language models by integrating collaborative semantics for recommendation. arXiv:2311.09049 (2023)