A SURVEY ON LARGE LANGUAGE MODELS FOR RECOMMENDATION

Madison E. Chester | Dafni Tziakouri

Recommenders Presentation 2024

Content

A Survey on Large Language Models for Recommendation

Likang Wu[1,2*], Zhi Zheng[1,2*], Zhaopeng Qiu[2*], Hao Wang[1†], Hongchao Gu[1], Tingjia Shen[1], Chuan Qin[2], Chen Zhu[2], Hengshu Zhu[2†], Qi Liu[1], Hui Xiong[3†], Enhong Chen[1†]

[1]University of Science and Technology of China, [2]Career Science Lab, BOSS Zhipin, [3]Hong Kong University of Science and Technology (Guangzhou)

{wulk,zhengzhi97,hcgu,jts_stj}@mail.ustc.edu.cn,
{zhpengqiu,chuanqin0426,zc3930155,zhuhengshu}@gmail.com,
{wanghao3,qiliuql,cheneh}@ustc.edu.cn, xionghui@ust.hk

Abstract

Large Language Models (LLMs) have emerged as powerful tools in the field of Natural Language Processing (NLP) and have recently gained significant attention in the domain of Recommendation Systems (RS). These models, trained on massive amounts of data using self-supervised learning, have demonstrated remarkable success in learning universal representations and have the potential to enhance various aspects of recommendation systems by some effective transfer techniques such as fine-tuning and prompt tuning, and so on. The crucial aspect of harnessing the power of language models in enhancing recommendation quality is the utilization of their high-quality representations of textual features and their extensive coverage of external knowledge to establish correlations between items and users. To provide a comprehensive understanding of the existing LLM-based recommendation systems, this survey presents a taxonomy that categorizes these models into two major paradigms, respectively Discriminative LLM for Recommendation (DLLM4Rec) and Generative LLM for Recommendation (GLLM4Rec), with the latter being systematically sorted out for the first time. Furthermore, we systematically review and analyze existing LLM-based recommendation systems within each paradigm, providing insights into their methodologies, techniques, and performance. Additionally, we identify key challenges and several valuable findings to provide researchers and practitioners with inspiration.

1 Introduction

Recommendation systems play a critical role in assisting users in finding relevant and personalized items or content. With the emergence of Large Language Models (LLMs) in Natural Language Processing (NLP), there has been a growing interest in harnessing the power of these models to enhance recommendation systems.

The key advantage of incorporating LLMs into recommendation systems lies in their ability to extract high-quality representations of textual features and leverage the extensive external knowledge encoded within them [Liu *et al.*, 2023b]. And this survey views LLM as the Transformer-based model with a large number of parameters, trained on massive datasets using self/semi-supervised learning techniques, e.g., BERT, GPT series, PaLM series, etc[‡]. Different from traditional recommendation systems, the LLM-based models excel in capturing contextual information, comprehending user queries, item descriptions, and other textual data more effectively [Geng *et al.*, 2022]. By understanding the context, LLM-based RS can improve the accuracy and relevance of recommendations, leading to enhanced user satisfaction. Meanwhile, facing the common data sparsity issue of limited historical interactions [Da'u and Salim, 2020], LLMs also bring new possibilities to recommendation systems through zero/few-shot recommendation capabilities [Sileo *et al.*, 2022]. These models can generalize to unseen candidates due to the extensive pre-training with factual information, domain expertise, and common-sense reasoning, enabling them to provide reasonable recommendations even without prior exposure to specific items or users.

The aforementioned strategies are already well-applied in discriminative models. However, with the evolution of AI learning paradigms, generative language models have started to gain prominence [Zhao *et al.*, 2023]. A prime example of this is the emergence of ChatGPT and other comparable models, which have significantly disrupted human life and work patterns. Furthermore, the fusion of generative models with recommendation systems offers the potential for even more innovative and practical applications. For instance, the interpretability of recommendations can be improved, as LLM-based systems are able to provide explanations based on their language generation capabilities [Gao *et al.*, 2023], helping users understand the factors influencing the recommendations. Moreover, generative language models enable more personalized and context-aware recommendations, such as users' customizable prompts [Li *et al.*, 2023] in the chat-based recommendation system, enhancing user engagement and satisfaction with the diversity of results.

Motivated by the remarkable effectiveness of the afore-

*Equal Contribution.
†Corresponding Author.

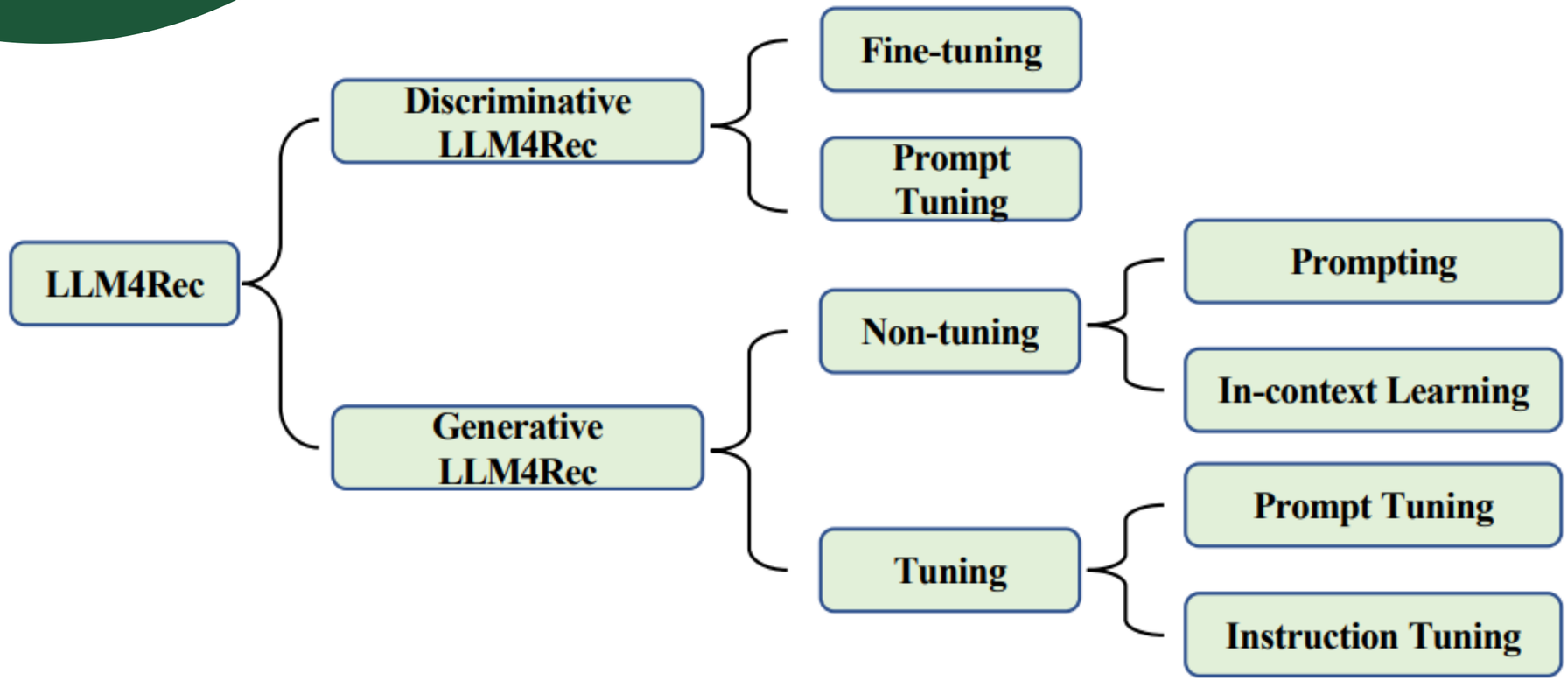
‡https://en.wikipedia.org/wiki/Large_language_model

https://github.com/hongleizhang/RSPapers/blob/master/16LLM_for_RS/%5B23.05%5D%20A%20Survey%20on%20Large%20Language%20Models%20for%20Recommendation.pdf

LLMs & Advantages

Textual Information	*Performance*	*User Engagement*
Comprehending user queries	*Improved accuracy*	*Satisfaction of results*
Comprehending item descriptions	*Improved relevance*	*Diversity of results*

Overview


LLM4Rec
Discriminative LLM4Rec
Fine-tuning
Prompt Tuning
Generative LLM4Rec
Non-tuning
Prompting
In-context Learning
Tuning
Prompt Tuning
Instruction Tuning

DISCRIMINATIVE LLMS

Fine-Tuning

- Model: BERT series initialized with learned parameters
- Fine-tuning done with domain-specific dataset (user-item interactions, user profiles, textual descriptions of items)
- Parameters are then updated
- Particularly important for sequential or session-based recommendation system

Prompt-Tuning

- Align the tuning object with pre-trained loss through hard/soft prompts and label word verbalizer
- Model: BERT Masked Language Modelling (MLM) and Next Sentence Prediction (NSP)
- Experiments revealed that BERT can prioritize relevant items in the ranking process without fine-tuning
- Performance notably enhanced with the utilization of multi-prompt ensembling (better than single-prompt results on discrete and continuous templates

GENERATIVE LLMS

Non-Tuning : Prompting

- Aimed at designing more suitable prompts
- Role instructions enhance the domain adaption ability
- LLMs can function as the controller of the entire recommendation system
- Can control the database and retrieve relevant content to address the cold start problem

5 Most Common Tasks:
rating prediction, sequential recommendation, direct recommendation, explanation generation, review summarization

Prompt Construction Framework:
1) task description
2) behavior injector
3) format indicator

Use Cases:

GENRE - refine news titles according to abstract, extract profile keywords from user history, generate synthetic news

NIR - generate user preference keywords & extract representative movies from user interaction history

Non-Tuning : In-Context Learning

Improves the recommendation ability of LLMs on most tasks

Demonstration input label-pairs are used to predict the label for an unseen output without additional parameter updates

Remaining Questions: Selection and Number Demonstration Examples

Tuning

parameters of LLMs are trained on a specific task, and the LLMs are fine tuned for multiple tasks

Prompt Tuning

- Input: user/item information and Output: User preference or ratings of items
- Rating predictions require multi-class classification and regression
- LLMs can learn how to recommend but language biases still exist for bottom recommendations

Instruction Tuning

- LLMs can align with human intent
- Models can achieve zero-shot generalization to unseen personalized prompts and new items
- Main tasks include respectively scoring, generation, and retrieval

FINDINGS & CHALLENGES

Model Bias

Position

- The infromation is stored in textual sequential descriptions
- LLMs prioritize the items in the top order
- Random sampling-based bootstrapping

Popularity

- The ranking results of LLMs are influenced by the popularity levels of the candidates
- Popular items tend to rank higher

Fairness

- Fairness issues related to sensitive attributes
- Influenced by the training data and the demographics of the individuals
- Assume users belong to a specific group (race or gender)

Recommendation Prompt Designing

User/Item Representation

- LLMs are not using sufficient representation of data
- Name to represent items. List of item names to represent users
- Challenging to use ID-like features

Limited Context Length

- Constrain the length of users' behavioral sequences
- Number of candidate items
- Selecting representative items from user behavior sequences

Promising Ability

Zero/Few-shot Recommendation

- Few-shot learning does not change the parameters of LLMs
- Reduce the cold-start problem with limited data
- Guidance in selecting representative demonstration examples

Explainable

- LLMs have a remarkable ability for natural language generation
- ChatGPT performs better than supervised traditional methods
- Promising performance in explainable Rec from LLMs

Evaluation Issues

Generation Controlling

- LLMs may not follow the desired output format
- Answers in different formats or refuses to answer
- Crucial to find a solution in controlling the output of LLMs

Evaluation criteria

- LLMs have strong generative capabilities
- Generate items that have never appeared in the historical data
- Still not any possible solution

Datasets

- MovieLens, Amazon Books, etc.
- Relatively small in scale, may not reflect the Rec capability of LLMs
- Related information in the pre-training data
- Introduce bias

THANK YOU FOR YOUR ATTENTION!