ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202305.00121

Predicting User Book Rating Preferences Using ChatGPT: A Study

Authors: Chen Yanfang, Li Zhiyu, Li Zhiyu

Date: 2023-05-12T00:00:00+00:00

Abstract

Purpose/Significance: With the continuous development and transformation of large language model technology represented by ChatGPT, many classic scenarios across various fields are being revitalized with new opportunities. Meanwhile, an increasing number of scholars are focusing on how to apply the intelligent capabilities and technologies of large language models to existing scenarios, and analyzing the challenges and opportunities brought by these technologies. Method/Process: This paper takes ChatGPT as the modeling object, and for the first time introduces large language model technology into the typical application scenario of user book rating preference prediction in the library and information science field, and puts it into practice. By constructing a ChatGPT-based user book rating prediction model (CUBR, ChatGPT-based model for User Book Rating Prediction), this study explores the feasibility of implementing and deploying large language model technology in the book recommendation domain. Simultaneously, based on different evaluation schemes for book rating tasks, this paper compares CUBR with existing classical recommendation models, discusses and presents its advantages and disadvantages in the user book rating prediction scenario, and analyzes potential research opportunities for large language models in other book recommendation scenarios. Results/Conclusion: Experimental studies in this paper demonstrate that (1) the CUBR model can achieve commendable recommendation performance on existing user book rating preference prediction tasks, particularly in few-shot scenarios like One-shot where target information for recommendation is limited, its performance approaches or exceeds current classical recommendation algorithms, with strong generalization capabilities, making it more suitable for cold-start recommendation scenarios. (2) As the content of individual user prompt samples increases (e.g., from One-shot to Ten-shot), the prediction performance of CUBR improves significantly, indicating that CUBR possesses considerable in-context learning capabilities. Limitations: The research scenario of this paper is limited to user book rating preference understanding and recommendation; future work will

attempt to apply and adapt existing large language model technologies to more library and information science scenarios to achieve better practical results.

Full Text

Preamble

Journal: Information Science: Theory & Application
Title: A ChatGPT-based Model for User Book Rating Preference Prediction
Authors: Chen Yanfang[1], Li Zhiyu[2]
Affiliations:
1. Renmin University of China Library, Beijing, 100872, China
2. AI for Science Institute, Beijing, 100084, China

Abstract

[**Purpose/Significance**] The continuous development and transformation of Large Language Models (LLMs), exemplified by ChatGPT, have revitalized classical scenarios across numerous fields with new opportunities. Concurrently, an increasing number of scholars are focusing on how to apply the intelligent capabilities and technologies of LLMs to existing contexts while analyzing the challenges and opportunities these technologies present. [**Method/Process**] This study, for the first time, introduces LLM technology into user book rating preference prediction—a typical application scenario in library and information science—and implements it in practice. By constructing a ChatGPT-based model for user book rating prediction (CUBR), we explore the feasibility of applying LLM technology in book recommendation. Additionally, this paper compares different evaluation schemes for book rating tasks against existing classical recommendation models, discusses the advantages and disadvantages of CUBR in user book rating prediction scenarios, and analyzes potential research opportunities for future LLM applications in other book recommendation contexts. [**Results/Conclusions**] Our experimental research demonstrates that: (1) The CUBR model can achieve commendable recommendation performance on existing user book rating preference prediction tasks, particularly in few-shot scenarios like one-shot where target information is limited. Its performance approaches or exceeds current classical recommendation algorithms, demonstrating strong generalization ability and suitability for cold-start recommendation scenarios. (2) As the number of prompt samples for individual users increases (e.g., from one-shot to ten-shot), CUBR's prediction performance improves significantly, indicating robust real-time in-context learning capabilities. [**Limitations**] This study is limited to understanding and recommending user book rating preferences. Future work will attempt to apply and adapt existing LLM technologies to more library and information science scenarios to achieve better practical results.

Keywords: ChatGPT; Large Language Model; Book Rating; Generative Response

This work is supported by the Renmin University of China Scientific Research Fund (Central University Basic Research Operating Expenses Special Fund) Project (No. 23XNQT24).

1 Introduction

In recent years, Natural Language Processing (NLP) technology has advanced rapidly, with daily transformations occurring in both model parameter scales and training data richness. In early December 2022, OpenAI released ChatGPT (Chat Generative Pre-trained Transformer), a conversational chatbot built upon and fine-tuned from the GPT-3.5 series of large language models. This model not only enables efficient and accurate interactive responses for multi-turn conversations but also performs various natural language processing tasks, including code assistance, document summarization, and story continuation. Its release immediately sparked enthusiastic discussion in both industry and academia.

[Figure 1: see original paper] Baidu search index trend for ChatGPT and typical important events

As shown in Figure 1, during the initial release period (November 2022–February 2023), ChatGPT maintained relatively low popularity due to imperfect model performance and user interfaces. Starting in February, as OpenAI iterated on the model and several important events were reported—such as ChatGPT passing Google engineer interviews and widespread participation from major internet companies—the technological foundation of ChatGPT, large language models, attracted significant academic attention.

Current popular LLM versions mainly include the GPT-3/4 series, the LLaMA model, and the GLM-130B model from Tsinghua University. Among these, ChatGPT, built on the GPT-3/3.5/4 series and supported by Microsoft and OpenAI's extensive promotion with API access, has been increasingly adopted by researchers and vendors for various real-world scenarios such as intelligent customer service, interactive translation, and personal assistants.

Particularly in the library and information science field, theoretical research surrounding ChatGPT-like models is growing, covering technical ethics and risk studies, as well as application scenario research. However, these studies primarily focus on theoretical discussions and application scenario analyses without testing ChatGPT in practical implementations. Critical questions remain unexplored: Can we consider building recommendation models based on ChatGPT-like models to solve existing recommendation problems in library and information science? How do ChatGPT-like models compare to traditional recommendation models when applied to corresponding recommendation scenarios? These questions merit thorough investigation.

The innovations of this paper are primarily threefold: (1) We apply large language models (such as ChatGPT) to a typical task in library and information science—user book rating preference prediction—to explore the feasibility of

implementing ChatGPT-like models in this domain. (2) We design prompt engineering paradigms for user book rating preference tasks, providing inspiration for similar implementation research. (3) Through one-shot and few-shot modeling experiments on the GoodBooks dataset, we demonstrate the feasibility of applying ChatGPT-like LLMs to user book rating preference prediction scenarios using various evaluation metrics.

2 Related Research

As library collections continue to expand in volume, variety, and reader interactions, library services are evolving toward greater intelligence, with recommendation models serving as a crucial method to address information overload. Personalized recommendation systems for library resources primarily rely on machine learning recommendation models. Research in this direction includes:

User-Item collaborative filtering recommendation models: Yu Yisheng et al. improved interpretability and accuracy in book recommendation systems by introducing biases into item-based collaborative filtering models, incorporating the meaning of biases and contributions from similar books to predicted ratings. Yang Chen et al. further introduced user social relationship similarity beyond book content similarity, optimizing recommendation effectiveness through heuristic unsupervised methods for fusing similarity measures.

User-Book interaction behavior sequence recommendation models: From an individual user perspective, series of interactions between users and books over time form corresponding book sequences. These models primarily address the recommendation problem of predicting the next (or multiple) interaction objects. For example, Wang Dailin et al. proposed a personalized recommendation model based on a table-of-contents attention mechanism, modeling users' historical browsing interactions through user ratings and attention mechanisms, and incorporating reader interest preferences via BiLSTM to improve recommendation accuracy. However, this model's performance strongly depends on dense reader behavior matrices and is limited in sparse scenarios.

Graph neural network recommendation models based on user-book networks. Benefiting from graph models' feature representation and high-order extraction capabilities, graph neural networks have been widely applied across recommendation systems. For instance, Chen Zhi et al. modeled interaction histories constructed as user-book bipartite graphs based on graph convolutional neural networks, capturing high-order connectivity between nodes to better model readers' domain preference information and improve recommendation effectiveness.

3 ChatGPT-Based User Book Rating Preference Prediction Model

[Figure 2: see original paper] Framework of the user book rating preference

prediction model

3.1 Model Overview

This paper proposes a user book rating preference prediction model based on ChatGPT-like large language models. By integrating existing LLMs with user rating preference prediction tasks, constructing appropriate prompt strategies, and combining data validation, backtracking, and retry methods, we explore the feasibility of applying LLMs in user book rating preference prediction scenarios. The model consists of four modules: (1) Task formalization definition, (2) Task prompt engineering design, (3) Model interaction and response parsing and validation, and (4) Task metric evaluation.

3.2 Task Formalization Definition

User rating preference prediction estimates users' future preferences for interacting with other books based on historical interactions or rating behaviors. This task has broad applications in book recommendation, such as user book preference prediction for e-commerce sales scenarios and interest preference prediction for library readers' borrowing, clicking, and browsing behaviors. The task typically uses historical reader-book interactions (clicks, views, borrowings, collections, reviews, ratings, etc.) as features and data sources, combined with user attributes and book attributes, to build accurate recommendations using various machine learning models. In this paper, the specific tasks are defined as follows:

One-shot recommendation modeling for users: Given a user u_i's historical book behavior sample sequence (e.g., rating sequence) $H_{u_i} = \{b_1, b_2, ..., b_n\}$, the model is provided with only a single training sample as a prompt or training set and is required to rate the preference for all remaining samples in the behavior sequence. The final evaluation assesses the consistency between the model's ratings and the original sample results.

Few-shot recommendation modeling for users: Given a user u_i's historical book behavior sequence (e.g., rating sequence) $H_{u_i} = \{b_1, b_2, ..., b_n\}$, a certain proportion of data is selected as the training set (this paper selects 10 and 20 prompt examples as the prompt set). The model is required to rate the preferences for the remaining sequence, with final evaluation measuring consistency between predicted and original ratings.

3.3 Task Prompt Engineering Design

Since ChatGPT-like LLMs are typical generative models, the quality of their generated content typically depends on the quality of input prompt content. Therefore, constructing effective prompt engineering for book recommendation tasks is the core discussion of this subsection. As shown in Figure 3, the prompt engineering example for user book rating preference prediction tasks typically includes four core components:

(1) **Identity injection prompt**: This prompt primarily informs the LLM of its current role type, guiding the LLM to produce different behavioral responses according to specific role types. For example, in certain tasks, without identity injection prompts such as "Assume you are an expert in profession XXX," ChatGPT may refuse to respond to direct requests like "Please evaluate the rating preference for XXX." Identity injection helps circumvent such refusals for safety or fairness reasons.

(2) **Task description prompt**: While identity injection prompts set the possible behavior cluster for ChatGPT-like models, task description prompts inform the LLM of the specific task background, framework, and possible task examples (few-shot scenarios). Typically, content prompts based on task examples (few-shot scenarios) provide additional learning samples to the model, enhancing its fitting and understanding of the task to ultimately produce better prediction results.

4 Experimental Setup and Results

4.3 Evaluation Metrics

As described in Section 3.2, user book rating preference recommendation can be treated as either a regression problem or a ranking problem. Therefore, to verify model performance across different test samples, we evaluate model effectiveness using the following metrics for both regression and ranking:

Metric 1: Mean Absolute Error (MAE), which measures the absolute deviation between user rating preference predictions and actual results, focusing on the average of absolute errors between true and predicted values. Calculation: $MAE = |y_i - \tilde{y}_i|$

Metric 2: Mean Absolute Percentage Error (MAPE), which focuses on percentage deviation of prediction errors relative to each sample's true value through dimensional scaling. Calculation: $MAPE = |y_i - \tilde{y}_i|/|y_i|$

Metric 3: Root Mean Square Error (RMSE), which differs from MAE by focusing more on the relative weight of different error magnitudes. Calculation: $RMSE = \sqrt{(y_i - \tilde{y}_i)^2}$

Metric 4: Normalized Discounted Cumulative Gain (NDCG), which primarily observes differences in relative position quality in ranking results. In this paper, we consider NDCG@{5,10,15,20} performance results.

4.4 Results Analysis and Discussion

This subsection analyzes the performance results of the CUBR model and baseline models across different tasks, addressing two core questions: (1) Can the CUBR model achieve effectiveness in user book rating preference recommendation scenarios, and how does its performance compare to other recommendation

models? (2) Can increasing prompt samples improve CUBR's recommendation capability, and are there noticeable changes compared to baseline models?

Comparative performance of user rating preference prediction models

As shown in Table 1, the test results for the CUBR model and baseline models across three sub-tasks (1-few-shot, 10-few-shot, and 20-few-shot) are presented, where gray-background numbers indicate the optimal model for each sub-task and underlined numbers correspond to the second-best model.

First, from an overall perspective: The MF (FunkSVD) model achieves good performance across different sub-tasks, particularly optimal in the 1-shot scenario. The core reason is that FunkSVD's recommendation strategy models the user-book rating interaction matrix through matrix factorization, with an optimization objective of minimizing residuals between actual ratings and matrix product ratings. Therefore, even with limited reference rating information from target users (e.g., 1-few-shot), FunkSVD achieves good results on metrics like RMSE. However, as the number of effective prompt examples per user increases, clustering-based models like KNN (means) begin to demonstrate advantages. During prediction, KNN (means) relies on modeling the target user's historical rating habits to generate final predictions, making its prediction accuracy gradually increase with more prompt examples. Notably, the CUBR model achieves second-best recommendation results based on NDCG metrics in the one-shot scenario, indicating good generalization capability in few-shot recommendation scenarios. However, CUBR also shows a performance gap compared to classical recommendation models in personalized understanding and modeling at the user dimension when directly applying general LLMs.

Second, examining different prompt-level sub-tasks: FunkSVD's NDCG metrics are not sensitive to the number of prompt samples, showing consistent performance across 1-few-shot, 10-few-shot, and 20-few-shot sub-tasks. In contrast, other baseline models like KNN (means), SlopeOne, and CUBR show significant metric changes as the number of prompt samples increases. The core reason is that these models reference the target user's historical ratings during prediction. For example, the CUBR model references provided historical ratings for different books to model background knowledge about user interest preferences and applies this knowledge comprehensively in new prediction scenarios. As shown in Figure 5 [Figure 5: see original paper], this illustrates CUBR's explicit rating decision process. Through appropriate example prompts, LLMs can typically learn corresponding contextual knowledge and apply it in prediction scenarios—a capability known as in-context learning, which is one of the important foundational abilities of large language models. After increasing prompt sample references, CUBR's MAE, MAPE, and RMSE metrics on the 20-few-shot sub-task improved significantly, with errors reduced by 21.67%, 16.53%, and 16.84%, respectively.

Finally, examining different metric types: With fewer reference samples (e.g., 1-few-shot), the CUBR model shows good performance in ranking capa-

bility (NDCG) but no advantage over baseline models in score prediction error metrics (MAE/MAPE/RMSE). Additionally, excessive increases in individual user prompt samples (e.g., from 10-few-shot to 20-few-shot) do not produce similar performance gains in NDCG metrics as observed from 1-few-shot to 10-few-shot. However, improvements in error metrics remain substantial. Therefore, if the application scenario prioritizes absolute score preferences for each recommended user, effectiveness can be improved by increasing prompt samples. If the scenario only focuses on relative ranking capability, modeling with few samples suffices, allowing further savings in inference resources.

5 Conclusion and Future Work

This paper proposes a ChatGPT-based large language model technology for user book rating preference prediction (CUBR), which for the first time introduces LLM technology into a classical task in library and information science and implements it in practice. In user book rating preference prediction tasks, we tested three sub-tasks with different sample prompt levels: 1-few-shot, 10-few-shot, and 20-few-shot. Experimental results demonstrate that CUBR achieves good recommendation performance with few prompt samples and without any fine-tuning, with significant improvements in prediction results after increasing prompt sample quantities. Future research will focus on the following aspects:

Research Direction 1: Prompt construction integrating multi-source data. In current explorations, we only used user rating interaction data to construct prompt sets. Future research should investigate how to integrate multi-source user attributes and features, even cross-modal data, and express them within LLMs to build unified recommendation systems. The fundamental challenge is how to consistently represent heterogeneous feature data from multiple sources as natural language encodings that LLMs can understand, thereby fully utilizing various contextual information to further improve model prediction performance.

Research Direction 2: Modeling research for task-specific instruction fine-tuning. The current CUBR modeling approach directly applies pretrained LLMs for recommendation and application, which typically tests LLMs' generalization capabilities. However, existing research shows that constructing targeted instruction training sets can further enhance LLM performance on specific tasks. Therefore, future research should explore how to construct efficient fine-tuning instruction sets based on proprietary tasks and special business scenarios in library and information science, and integrate prediction processes with training processes to ultimately improve LLM performance in recommendation systems.

References

[1] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[A]. 2023: 1-58.

[2] OPENAI. Introducing chatgpt[R/OL]. https://openai.com/blog/chatgpt.
[3] DREIBELBIS E. Chatgpt passes google coding interview for level 3 engineer with $183k salary[EB/OL]. http://985.so/mny2k.
[4] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
[5] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[A]. 2023.
[6] ZENG A, LIU X, DU Z, et al. GLM-130b: An open bilingual pre-trained model[C/OL]//The Eleventh International Conference on Learning Representations (ICLR). 2023. https://openreview.net/forum?id=-Aw0rrrPUF.
[7] GEORGE A S, GEORGE A H. A review of chatgpt ai's impact on several business sectors[J]. Partners Universal International Innovation Journal, 2023, 1(1): 9-23.
[8] LU Q, QIU B, DING L, et al. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt[A]. 2023.
[9] SHAFEEG A, SHAZHAEV I, MIHAYLOV D, et al. Voice assistant integrated with chat gpt[J]. Indonesian Journal of Computer Science, 2023, 12(1).
[10] You J. Application risks and control measures of ChatGPT-like generative AI in research scenarios[J]. Information Theory and Practice, 2023: 01-11.
[11] Cai S, Yang L. Research on risks and collaborative governance of ChatGPT intelligent robot applications[J]. Information Theory and Practice, 2023: 01-11.
[12] Zhang H, Tong T, Ye Y. GPT technology-driven innovation for smart libraries in the AI 2.0 era[J]. Library Journal, 2023: 01-07.
[13] Guo Y, Guo Y, Li S, Feng S. ChatGPT empowering smart library services: Characteristics, scenarios, and paths[J]. Library Construction, 2023.
[14] Zhou W. Application and significance of ChatGPT in the archives field[J]. China Archives, 2023, 593(03): 62-63.
[15] Wang B, Niu C. From ChatGPT to GovGPT: Building a government service ecosystem driven by generative AI[J]. E-Government, 2023.
[16] Yu Y, Wei R, Liu X. Research on interpretable real-time book information recommendation models[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(2): 209-216.
[17] Yang C, Liu T, Liu L, Niu B, Sun J. Research on electronic document resource recommendation method integrating semantic and social features[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(6): 632-640.
[18] WANG S, HU L, WANG Y, et al. Sequential recommender systems: Challenges, progress and prospects[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. International Joint Conferences on Artificial Intelligence Organization, 2019: 6332-6338.
[19] Wang D, Liu L, Liu M, Liu Y. Research on reader preference analysis and recommendation model based on table-of-contents attention mechanism[J]. Data Analysis and Knowledge Discovery, 2022, 6(9): 138-152.

[20] Chen Z, Zhang W, Liu T. Research on book recommendation method based on graph convolutional neural networks[J]. Information Exploration, 2022, 300(10): 1-5.
[21] SARAVIA E. Prompt Engineering Guide[J]. https://www.promptingguide.ai, 2022.
[22] ZAJAC Z. Goodbooks-10k: a new dataset for book recommendations[J/OL]. FastML, 2017. http://fastml.com/goodbooks-10k.
[23] MNIH A, SALAKHUTDINOV R R. Probabilistic matrix factorization[J]. Advances in neural information processing systems (NPIS), 2007, 20.
[24] KOREN Y. Factor in the neighbors: Scalable and accurate collaborative filtering[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2010, 4(1): 1-24.
[25] LEMIRE D, MACLACHLAN A. Slope one predictors for online rating-based collaborative filtering[C]//Proceedings of the 2005 SIAM International Conference on Data Mining. SIAM, 2005: 471-475.
[26] WANG Y, WANG L, LI Y, et al. A theoretical analysis of ndcg type ranking measures[C]//Conference on learning theory. PMLR, 2013: 25-54.

Author Biographies:

Chen Yanfang, female, born 1992, Librarian, Ph.D. Research interests: user behavior, health information, data mining.

Li Zhiyu (Corresponding author: zhiyulee@icloud.com), male, born 1991, Algorithm Expert, Ph.D. Research interests: machine learning, network representation, natural language processing.

Author Contribution Statement: Chen Yanfang is responsible for paper drafting, research framework formulation, and main content writing. Li Zhiyu is responsible for experimental data collection, model implementation, and experimental discussion and analysis.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv — Machine translation. Verify with original.