Evaluating ChatGPT as a Recommendation System: A Comprehensive Study

Author
peyman.75.naseri@gmail.com

1 Introduction

Recommendation systems have long been integral to enhancing user experiences across various digital platforms, with traditional methods focusing on task-specific approaches. However, these methods often lack the flexibility and generalization capability required for evolving user demands. This project seeks to explore the potential of ChatGPT, an advanced conversational model in Natural Language Processing (NLP), as a versatile tool in the recommendation domain. We aim to investigate how ChatGPT, utilizing its extensive linguistic and world knowledge from large-scale text corpora, can be applied to diverse recommendation scenarios. By evaluating ChatGPT's performance across multiple contexts – including rating prediction, sequential recommendation, direct recommendation, explanation generation, and review summarization – without resorting to model fine-tuning, this study intends to uncover innovative methodologies in recommendation tasks. This approach could revolutionize how recommendation systems understand and address user needs, inspiring further research in leveraging language models to enhance the efficacy of recommendation systems and contribute significantly to the field's advancement.

All related content for the project, including codes, prompts, the reference article, and other utilized articles, are available at this GitHub repository.

2 Related work

Advancements in Language Models (LMs) such as BERT (Devlin et al., 2019) and GPT (OpenAI, 2023) have revolutionized the field of NLP, significantly improving the capability of various tasks. The success of these models has sparked a growing interest in their application within recommendation systems. A key development in this realm is the P5 model (Geng et al., 2023), which has been a major source of inspiration for our work. The P5 model ingeniously integrates different recommendation tasks within a unified language generation framework, exemplifying the adaptability of LMs in complex recommendation scenarios. This innovative approach in P5 has laid the groundwork for our study, where we aim to explore ChatGPT as an autonomous recommendation system. Unlike other models that depend on external systems, our approach positions ChatGPT as a self-contained system, capable of independently handling various recommendation tasks. Other notable contributions include LMRecSys (Zhang et al., 2021), which utilizes prompts for redefining recommendation tasks for efficiency, and M6-Rec (Cui et al., 2022), aiming to establish a foundational model for a range of recommendation tasks. Chat-REC (Gao et al., 2023) has also made significant strides by using ChatGPT in conversational recommendation contexts. Our study, heavily influenced by the P5 model, seeks to further the understanding of ChatGPT's standalone capabilities in the recommendation domain, potentially leading to more personalized and AI-driven user experiences in recommendation systems.

3 approach

We employ prompt engineering and few-shot learning techniques with ChatGPT for experiments in five recommendation tasks. Below are the baseline algorithms and evaluation metrics for each task:

3.1 Tasks

3.1.1 Rating Prediction

Baseline Algorithm: Average rating algorithm.
Evaluation Metric: Mean Absolute Error (MAE) or Root Mean Square Error (RMSE) to measure

the prediction accuracy.

3.1.2 Sequential Recommendation

Baseline Algorithm: Most recently interacted items as recommendations.
Evaluation Metric: Precision and Recall at K (P@K, R@K) to assess relevance.

3.1.3 Direct Recommendation

Baseline Algorithm: Popularity-based recommendation.
Evaluation Metric: Hit Rate and Coverage to evaluate recommendation relevance and diversity.

3.1.4 Explanation Generation

Baseline Algorithm: Template-based method.
Evaluation Metric: BLEU Score or ROUGE for assessing fluency and relevance of explanations.

3.1.5 Review Summarization

Baseline Algorithm: Extraction-based summarization.
Evaluation Metric: ROUGE Score to evaluate the overlap with key points in original reviews.

Each task's evaluation metrics are chosen to best reflect the quality and effectiveness of ChatGPT's performance in that specific area, providing a comprehensive understanding of its capabilities compared to traditional methods. These metrics will enable us to draw informed conclusions about the applicability and advantages of using ChatGPT in diverse recommendation scenarios.

3.2 Schedule

The project is structured into distinct phases, each with specific objectives and timelines, ensuring a comprehensive exploration and timely completion within approximately 1.5 months.

3.2.1 Reproducing the Original Study (2 Weeks)

- Replicating experiments and methodologies from ”Is ChatGPT a Good Recommender? A Preliminary Study.”
- Analyzing results to ensure fidelity to the original study.

3.2.2 Experimentation Under Alternate Conditions (2 Weeks)

- Utilizing various language models like GPT-4 and modifying prompts for testing.
- Experimenting with different datasets to examine adaptability.

3.2.3 In-depth Analysis and Enhancement (2 Weeks)

- Detailed analysis of underperforming tasks.
- Implementing and testing improvement strategies.

3.2.4 Final Report and Synthesis (1 Week)

- Compiling findings, analyses, and conclusions into a comprehensive report.
- Drafting future research directions and project expansions.

4 Data

Utilizing the Amazon Beauty dataset for its relevance in testing recommendation tasks.

5 Tools

- **Primary Tool**: ChatGPT for recommendation tasks.
- **Computational Resources**: Google Colab Pro for enhanced processing capabilities.
- **Supporting Tools**: Data analysis and visualization tools, and LaTeX for report preparation.

References

Cui, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. (2022). M6-rec: Generative pretrained language models are open-ended recommender systems.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Gao, Y., Sheng, T., Xiang, Y., Xiong, Y., Wang, H., and Zhang, J. (2023). Chat-rec: Towards interactive and explainable llms-augmented recommender system.

Geng, S., Liu, S., Fu, Z., Ge, Y., and Zhang, Y. (2023). Recommendation as language processing (rlp): A unified pretrain, personalized prompt predict paradigm (p5).

OpenAI (2023). Gpt-4 technical report.

Zhang, Y., DING, H., Shui, Z., Ma, Y., Zou, J., Deoras, A., and Wang, H. (2021). Language models as recommender systems: Evaluations and limitations. In *I (Still) Can't Believe It's Not Better! NeurIPS 2021 Workshop*.