Evaluating ChatGPT as a Recommendation System: A Comprehensive Study

First Author
peyman.75.naseri@gmail.com

1 Problem statement

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 What you proposed vs. what you accomplished

This section outlines the project's initial goals against the backdrop of actual achievements and the inevitable adjustments made during the course of the study. Our endeavor embarked on a structured timeline, aiming to replicate and extend the findings of "Is ChatGPT a Good Recommender? A Preliminary Study" by navigating through four main phases: reproducing the original study, experimenting under alternate conditions, conducting an in-depth analysis and enhancement, and compiling a comprehensive final report.

- ~~Reproducing the Original Study~~: Successfully replicated experiments and methodologies for three out of the five tasks due to budget constraints limiting API usage, thus focusing on a subset of the main dataset.

- ~~Experimentation Under Alternate Conditions~~: Adjusted the project scope to incorporate the LLAMA model instead of GPT-4, considering the prohibitive costs associated with the latter. This modification was made to assess adaptability across different language models while managing resource limitations.

- *In-depth Analysis and Enhancement*: Partially completed. Detailed analysis was conducted for the tasks we were able to reproduce. However, the strategies intended for model improvement, specifically refining models with the Lora method, were not implemented. The high cost of utilizing the GPT API and the hardware limitations encountered with Google Colab's pro account for LLAMA model fine-tuning significantly hindered these efforts.

- ~~Final Report and Synthesis~~: Compiled findings and analyses into a draft report, albeit with adjustments to the planned content due to the aforementioned constraints. Future research directions were outlined, taking into account the project's scope adjustments and the insights gained from the tasks completed.

- *Neglected Tasks - Explanation Generation and Review Summarization*: Due to the original article's emphasis on the necessity of human evaluation for these tasks and budgetary limitations, these aspects were deprioritized in favor of focusing on reproducible quantitative analysis.

The project encountered unforeseen challenges, particularly with budgetary and hardware constraints, which led to significant adaptations in our approach. Despite these hurdles, the study provides valuable insights into the potential of language models in recommendation systems, setting a foundation for future research under more favorable conditions.

3 Related work

The intersection of language models (LMs) and recommendation systems represents a burgeoning field of research, propelled by significant advancements in NLP. Pioneering models like BERT (Devlin et al., 2019) and GPT (OpenAI, 2023) have laid the groundwork for transformative approaches in understanding and generating human language, catalyzing the exploration of their application within recommendation systems. This surge of interest is epitomized by the P5 model (Geng et al., 2023), which harmonizes various recommendation tasks under a unified language generation framework. P5's ingenuity in leveraging LMs to navigate the complexities of recommendation scenarios has significantly influenced our exploration of ChatGPT's capabilities as a standalone recommendation system.

Our project diverges from traditional methods by positioning ChatGPT as a self-sufficient entity capable of handling diverse recommendation tasks without reliance on external systems. This approach seeks to capitalize on ChatGPT's comprehensive knowledge base and linguistic proficiency, aiming for a more integrated and fluid recommendation experience. In contrast, LMRecSys (Zhang et al., 2021) and M6-Rec (Cui et al., 2022) explore the utility of LMs in recommendation systems through prompt engineering and foundational model development, respectively. While these contributions have underscored the potential of LMs in enhancing recommendation systems, our work endeavors to push the boundaries further by examining ChatGPT's autonomous operation within this domain.

Moreover, the Chat-REC model (Gao et al., 2023) represents a significant stride towards embedding ChatGPT in conversational recommendation systems, emphasizing interactive and explainable recommendations. While Chat-REC focuses on the conversational aspect, our study expands the exploration to include a broader range of recommendation tasks, including those not inherently conversational in nature.

In synthesizing these developments, our study not only seeks to build upon the foundational insights provided by the likes of P5, LMRecSys, and M6-Rec but also to carve a distinct niche by exploring the untapped potential of ChatGPT in autonomously managing a spectrum of recommendation tasks. Through this, we aspire to contribute to the evolving narrative of AI-driven, personalized user experiences within the realm of recommendation systems.

4 Your dataset

For this project, we utilized the Amazon Beauty dataset due to its extensive coverage and relevance in testing recommendation tasks. The dataset comprises product reviews and ratings in the beauty sector, making it an ideal candidate for exploring the capabilities of language models like ChatGPT and LLAMA in recommendation systems.

4.1 Basic Statistics

- **Total Prompts Created:** Approximately 700,000
- **Prompts Used for Evaluation:**
  - **ChatGPT:** 3,000 (500 per task for both zero-shot and few-shot modes)
  - **LLAMA:** 60 (due to lower performance, indicating less utility in further evaluations)

This dataset's rich textual content and user-generated reviews pose unique challenges, including the variability in user sentiment, diverse linguistic expressions, and the necessity to discern nuanced preferences and product features from free-text reviews.

4.2 Data Preprocessing

The preprocessing stage involved several steps to tailor the dataset for our evaluation tasks:

- **Prompt Generation:** Developed prompts from the dataset to interact with the language models. The prompts were designed to simulate real-world recommendation queries, requiring models to interpret user reviews and ratings.

- **Subsampling:** From the 700,000 generated prompts, we strategically selected samples to ensure a manageable yet representative subset for evaluation, focusing on diversity in user ratings, review lengths, and product types.

- **Normalization:** Standardized review texts to mitigate discrepancies in formatting and language use, enhancing the models' interpretability of the inputs.

4.3 Data Annotation

Given the nature of our tasks, specifically for those requiring qualitative assessments like explanation generation and review summarization, a preliminary annotation process was necessary. However, due to constraints, extensive annotation was not feasible.

The limited use of the LLAMA model for evaluation purposes reflects its preliminary performance, which did not justify extensive dataset application, contrasting with ChatGPT's broader evaluation scope.

5 Baselines

In this project, our baseline models were drawn from the traditional recommendation systems mentioned in the "Is ChatGPT a Good Recommender? A Preliminary Study" article. These models serve as a comparative foundation against which we evaluated the performance of ChatGPT across various recommendation tasks. Our decision to use these specific baselines was driven by their established benchmarks in the literature, providing a clear point of comparison for assessing the innovative approaches employed by ChatGPT.

5.1 Baseline Models and Their Working Mechanisms

The baseline models include collaborative filtering, content-based filtering, and hybrid approaches, each with distinct methodologies for generating recommendations:

- **Collaborative Filtering:** Utilizes user-item interaction data to predict user preferences based on similar users or items.

- **Content-Based Filtering:** Recommends items by analyzing item features and matching them to user profiles.

It's important to note that our work diverges significantly from these traditional models. We introduce a novel approach that can be described as "knowledge-based," leveraging the extensive knowledge and understanding capabilities of ChatGPT to provide recommendations. This method stands apart from Collaborative and Content-Based Filtering by directly tapping into the rich information encoded within the model, offering a unique perspective on recommendation systems.

5.2 Evaluation Metrics

Our evaluation framework comprises specific metrics tailored to each task, ensuring an accurate assessment of performance:

- **Rating Prediction:** Employed Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to gauge prediction accuracy.

- **Sequential and Direct Recommendation:** Utilized Hit Rate (HR@k) and Normalized Discounted Cumulative Gain (NDCG@k) to measure recommendation relevance and diversity.

These metrics were chosen for their ability to capture the efficacy of ChatGPT in navigating the complexities of recommendation scenarios, providing insights into its strengths and limitations.

5.3 Data Split

The dataset was segmented into training, validation, and testing sets with the following distribution:

- **Training Data:** 500,000 samples

- **Validation Data:** 100,000 samples

- **Testing Data:** 100,000 samples

This split was designed to ensure a robust evaluation process, allowing for the fine-tuning of model parameters on the validation set while preserving the integrity of the test set for unbiased performance assessment.

5.4 Hyperparameters and Tuning

Given the project's focus on evaluating ChatGPT without fine-tuning specific to the recommendation tasks, the primary hyperparameters under consideration were related to the prompt design and selection criteria for the few-shot learning scenarios. The tuning process involved iterative adjustments to the prompts based on validation set performance, with an emphasis on optimizing clarity and relevance to the task at hand. It's crucial to note that no hyperparameters were tuned based on the test set, adhering to best practices in model evaluation.

6 Leveraging Large Language Models for Recommendation Tasks: A Prompt-Based Approach

Our project embarked on an exploratory journey to assess the feasibility of using large language models (LLMs) for recommendation tasks through a novel prompt-based approach. The crux of our methodology was to formulate prompts that effectively translate recommendation queries into a format understandable by LLMs, thereby leveraging the vast knowledge embedded within these models to generate recommendations.

6.1 Methodology and Implementation

The primary challenge was crafting prompts that could elicit meaningful responses from the LLMs, serving as recommendations. This process involved iterative refinement to identify prompt structures that maximized the relevance and accuracy of the model's outputs. Unfortunately, the ambitious scope of our project encountered significant hurdles, primarily due to the high computational costs and hardware resource limitations associated with using state-of-the-art LLMs like ChatGPT and LLAMA.

6.2 Tools and Libraries

To facilitate our experiments, we utilized the following libraries and frameworks:

- **Pandas:** For data manipulation and analysis, crucial for processing the outputs of the LLMs and organizing our datasets.
- **PyTorch and Hugging Face:** These libraries were instrumental in loading and interacting with the LLAMA model, providing a flexible environment for NLP tasks.
- **OpenAI's API:** Enabled access to ChatGPT, allowing us to submit prompts and retrieve the model's recommendations.

The experiments were conducted on Google Colab Pro, which provided the necessary GPU and RAM resources. However, even with this enhanced computational capacity, we faced limitations that restricted our ability to fully explore the potential of LLMs for our intended recommendation tasks.

6.3 Challenges

Several challenges impeded our project's progress:

- **Computational Costs:** The financial and computational expenses of utilizing LLMs at scale were prohibitive, curtailing the breadth of our experiments.
- **Hardware Limitations:** Despite leveraging Google Colab Pro, we encountered barriers in processing power and memory that hindered our ability to employ the models as extensively as planned.

7 Results and Comparisons

7.1 Rating Prediction

The following table compares the performance in rating prediction, highlighting our approach against traditional and ChatGPT methods from the article:

Table 1: Performance comparison in rating prediction.

Method	RMSE	MAE
MF	1.1973	0.9461
MLP	1.3078	0.9597
GPT (zero)	1.4059	1.1861
GPT (few)	1.0751	0.6977
Our (zero)	1.6332	1.3665
Our (few)	1.7314	1.1670

The article presented a traditional approach using methods like MF (Matrix Factorization) and MLP (Multi-Layer Perceptron), with ChatGPT variants showing diverse outcomes. Our approach, when employing zero-shot and few-shot techniques, demonstrated higher RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) values compared to those reported for ChatGPT

in the article. This indicates a lower accuracy in prediction, with zero-shot and few-shot learning not effectively bridging the gap towards the more traditional methods or the ChatGPT multi-shot results mentioned.

Reasoning: The higher error rates in our approach likely stem from the complexities of accurately predicting ratings solely based on prompts without extensive fine-tuning or adaptation to the dataset's specific nuances. The gap in performance could be attributed to the limitations of prompt engineering or the inherent challenge of extracting precise numerical predictions from LLMs without tailored training.

7.2 Sequential Recommendations

This table delineates the stark contrast in performance for sequential recommendations, with our approach failing to yield positive results compared to the baseline and ChatGPT methods:

For sequential recommendations, both our zero-shot and few-shot approaches resulted in zero performance across all metrics, contrasting starkly with the varied performance of traditional and ChatGPT methods reported in the article.

Explanation: This outcome suggests a fundamental limitation in leveraging LLMs for sequential recommendation tasks without customized model training or adaptation. The zero performance indicates that the models were unable to grasp the sequential nature of user interactions or the contextual relevance needed to make meaningful recommendations. This could be due to the abstract nature of sequential recommendation tasks, which require understanding complex user behaviors over time—something that may not be directly inferable through general-purpose LLMs without specific tuning.

7.3 Direct Recommendations

The table below showcases our direct recommendation task performance, indicating areas of both challenge and unexpected success against traditional and ChatGPT methods:

[ht]

Method	HR@5	NDCG@5	HR@10	NDCG@10
BPR-MF	0.1426	0.0857	0.2573	0.1224
BPR-MLP	0.1392	0.0848	0.2542	0.1215
SimpleX	0.2247	0.1441	0.3090	0.1711
P5-B	0.1564	0.1096	0.2300	0.1332
GPT (zero)	0.0217	0.0111	0.0652	0.0252
GPT (shot)	0.0349	0.0216	0.0930	0.0398
Our (zero)	0.05	0.0361	0.08	0.0458
Our (few)	0.034	0.0249	0.046	0.0287

In the task of direct recommendation, our results showed some improvement over the zero-shot and few-shot ChatGPT results reported in the article but still lagged behind more specialized methods like SimpleX and P5-B. Interestingly, in some metrics, our approach outperformed the article's ChatGPT implementations, suggesting that there were instances where our model could capture product similarities or user preferences more effectively.

Discussion: The relative success in this area might be due to the direct recommendation task's reliance on understanding immediate user preferences and product features, which can be somewhat inferred through well-constructed prompts. However, the inconsistency and the occasional outperformance could indeed be attributed to the variability inherent in LLM responses, where certain prompts might accidentally align more closely with the underlying patterns in the data.

Overall Comment

The divergent results across tasks underscore the challenges of applying LLMs to recommendation systems without significant customization or fine-tuning. While promising in theory, the practical application reveals critical limitations, especially in tasks requiring nuanced understanding or sequential analysis. The success in direct recommendations, albeit inconsistent, highlights potential pathways for refining LLM-based recommendation approaches, possibly through more sophisticated prompt engineering or hybrid models that combine LLM insights with task-specific data processing.

The performance disparities can largely be attributed to the LLMs' generalist nature, designed to handle a wide range of NLP tasks but not optimized for the specificities of recommendation systems. This underscores the importance of targeted model training and the development of more adaptable LLM frameworks for specialized tasks like recommendations.

8 Conclusion

our exploration into using LLMs for recommendation tasks through a prompt-based approach highlighted both the potential and the pitfalls of this innovative methodology. While we faced significant challenges, the insights gained lay a foundation for future research in this direction, underscoring the need for more accessible and cost-effective solutions for leveraging LLMs in recommendation systems.

Table 2: Comparison of performance in sequential recommendations.

Method	HR@5	NDCG@5	HR@10	NDCG@10
Caser	0.0205	0.0131	0.0347	0.0176
HGN	0.0325	0.0206	0.0512	0.0266
GRU4Rec	0.0164	0.0099	0.0283	0.0137
BERT4Rec	0.0203	0.0124	0.0347	0.0170
FDSA	0.0267	0.0163	0.0407	0.0208
SASRec	0.0387	0.0249	0.0605	0.0318
S3-Rec	0.0387	0.0244	0.0647	0.0327
P5-B	0.0493	0.0367	0.0645	0.0416
GPT (zero)	0.0000	0.0000	0.0000	0.0000
GPT (shot)	0.0135	0.0135	0.0135	0.0135
Our (zero)	0.0	0.0	0.0	0.0
Our (few)	0.0	0.0	0.0	0.0

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*.