Evaluating Zero-Shot Large Language Models Recommenders on Popularity Bias and Unfairness: A Comparative Approach to Traditional Algorithms

Gustavo Mendonça Ortega, Rodrigo Ferrari de Souza, Marcelo Garcia Manzato
gustavo_ortega@usp.br,rodrigofsouza@usp.br,mmanzato@icmc.usp.br
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
São Carlos, SP

ABSTRACT

Large Language Models (LLMs), such as ChatGPT, have transcended technological boundaries and are now widely used across various domains to enhance productivity. This widespread application highlights their versatility, with a notable presence as recommender systems. Existing literature already showcases their capabilities in this area. In this paper, we present a detailed empirical evaluation of the effectiveness of Zero-Shot LLMs, specifically ChatGPT 3.5 Turbo, without special settings, in calibrating popularity bias and ensuring fairness in movie and TV show recommendations when prompted. We particularly focus on how these models adapt their output, comparing them to traditional post-processing algorithms. Our findings indicate that LLMs, evaluated through metrics such as Mean Average Precision (MAP) and Mean Rank Miscalibration (MRMC), not only perform well but also have the potential to surpass conventional recommender systems models like Singular Value Decomposition (SVD) when paired with calibration methods. The results underscore the advantages of using LLMs in more advanced scenarios due to their ease of implementation and performance.

KEYWORDS

Recommender Systems, LLM, Zero-Shot, Popularity Bias, Fairness.

1 INTRODUCTION

Traditional recommendation algorithms often struggle with popularity bias and unfairness, impacting user satisfaction and trust. Popularity bias, where popular items are disproportionately recommended, limits the exposure of less popular but potentially relevant items, reducing the diversity of recommendations [9]. Unfairness arises when certain groups of users consistently receive lower-quality recommendations, leading to unequal access to information and opportunities. Addressing these issues is critical for creating equitable and effective recommender systems that serve all users fairly [14].

Despite the availability of well-established recommender algorithms and bias/unfairness mitigation [15, 17, 18], the emergence and rapid popularization of Large Language Models (LLMs) have introduced novel approaches for a wide range of tasks. These models are being rigorously tested in various domains, including classification problems [7], anomaly detection [2], mathematical problem solving [19], and recommender systems [21].

Unlike traditional algorithms that require extensive training and large collections of interactions, LLMs primarily use their extensive content and contextual understanding to generate recommendations [8]. However, given their growing prominence, it is crucial to study LLMs in the context of bias and unfairness to ensure they provide equitable and diverse recommendations without perpetuating existing biases. Several recent works examine these aspects [5, 6, 10, 12, 16, 20], but there is a lack of studies comparing LLMs with traditional approaches to bias and unfairness mitigation. This comparison is essential to understand the effectiveness of LLMs in this task and to position these models relative to traditional methods developed so far.

Given the current landscape, this paper presents a detailed empirical evaluation of Zero-Shot LLMs, specifically ChatGPT 3.5 Turbo[1], without any special settings, in calibrating popularity bias and ensuring fairness in movie and TV show recommendations. We focus on how these models adapt their outputs compared to traditional post-processing algorithms. By evaluating metrics such as Mean Average Precision (MAP) and Mean Rank Miscalibration (MRMC), we demonstrate that LLMs not only perform well but also have the potential to surpass conventional recommender system models like Singular Value Decomposition (SVD) when paired with calibration methods.

This paper is organized as follows. Section 2 reviews the related work and background information necessary for understanding the context of this study. Section 3 discusses the traditional approaches used in the field and highlights their strengths and limitations. Section 4 describes the experimental setup, including the data, methods, and tools used to conduct the research. Section 5 presents the results of the experiments and provides an analysis of the findings. Section 6 summarizes the study's main contributions and outlines potential directions for future research.

2 RELATED WORK

The field of recommender systems has been significantly influenced by large language models (LLMs) like GPT (Generative Pre-trained Transformer) [3]. The literature reports works analyzing LLMs in a general context [13], and in a specific context of bias [5, 10, 16] and unfairness [6, 12, 20] in recommender systems. These studies highlight LLMs' potential and limitations, showing their performance in

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo

In: IV Concurso de Trabalhos de Iniciação Científica (CTIC 2024). Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web (CTIC'2024). Juiz de Fora/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

ISSN 2596-1683

rating prediction, explanation generation, and personalized recommendations while identifying significant biases. Various strategies, including tailored prompting and framework designs, are proposed to mitigate these biases and improve fairness. This work compares traditional bias mitigation and fairness approaches with those produced by LLMs, comprehensively evaluating their effectiveness in addressing these critical issues in recommendation systems.

3 TRADITIONAL APPROACHES

There are several traditional approaches for mitigating bias and unfairness in recommendation systems. As noted in the literature [14], works proposing recommendation calibration can apply this strategy at three stages: pre-processing, in-processing, and post-processing. In this paper, we chose the post-processing calibration approach due to its simplicity and independence from the training data.

We selected four calibration approaches according to their relevance and recency:

(1) **CP**: Proposed by [1], this method implements a calibration technique for popularity, similar to our proposed popularity calibration, but using the Jensen-Shannon divergence metric for comparing the profile and recommendation distributions. In our experiments, we followed the authors' method for both datasets to split the popularity into groups and exploited the parameter $\lambda \in [0, 1]$. This method is compared against our proposals using the SVD++ recommender. In this approach, the author divides users into popularity groups and uses a divergence measure to return the best recommendations for each user.
(2) **Steck's Calibration**: proposed by [18], this method works as a post-processing step for genre calibration. This approach aims to provide the best possible recommendation list for the user based on their genre preferences, ensuring it matches their interest proportion.
(3) **Personalized**: proposed by [15], this method implements a switch-based calibration, where some users receive the genre calibration, and others receive the popularity calibration. In our experiments, we followed the authors' methodology and exploited the parameter $\lambda \in [0, 1]$. In this approach, the authors consider that users should receive recommendations calibrated based on popularity if they consume many popular items. This method works similarly to the one implemented by Steck [18]. Still, it uses some weights to balance recommendations between accuracy and calibration and also considers the popularity of items in the calibration process.
(4) **Two Stage**: proposed by [17], this method implements a pipeline of two calibrations based on genres and popularity. In our experiments, we followed the authors' methodology. First, it generates a recommended list from the model, followed by the best possible list that matches the user's interest proportion regarding item popularity. After that, a second calibration is done to produce a new list that meets the user's interests in terms of genre.

As a recommender algorithm, we selected the SVD++ [11] as the model to be combined with these approaches because it effectively incorporates both explicit ratings and implicit feedback, leading to more accurate and personalized recommendations.

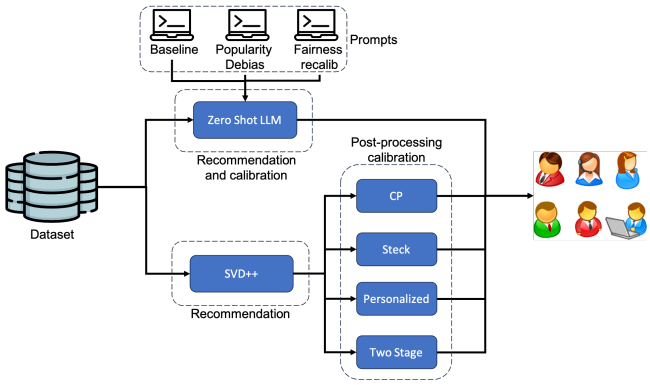

Figure 1: Overview of proposed evaluation scenario.

4 EXPERIMENTAL SETUP

To thoroughly measure the effectiveness of LLMs in this task and compare the performance and the impact of calibration techniques, the LLM model was evaluated as illustrated in Figure 1.

The evaluation process involves two main models: Zero-Shot ChatGPT 3.5 Turbo and SVD++. Data from a dataset is fed into both models to generate recommendations. ChatGPT provides recommendations based on its inherent capabilities without specific pre-training, while SVD++ uses four post-processing calibration techniques (CP, Steck's, Personalized, and Two Stage) to improve fairness and mitigate bias. The performance of both models is then compared, focusing on personalized and fair recommendations for different user profiles, to assess the effectiveness of LLMs versus traditional calibrated recommendation algorithms.

Concerning the LLM, we designed three scenarios:

(1) **Baseline**: no post-processing algorithms were applied. The prompt consists of a user's historical profile, and the LLM is requested for recommendations.
(2) **Popularity Debiasing**: the LLM is informed through the prompt about the items' popularity and their distribution. The LLM is requested to provide recommendations without popularity bias.
(3) **Fairness Recalibration**: a genre distribution of the user's profile is added to the prompt, and the LLM is requested to provide recommendations according to this distribution.

This structured approach facilitates a comprehensive understanding of the calibration effectiveness and enables the empirical data collection demonstrating the LLM's ability to adjust its outputs based on input prompts.

4.1 Dataset

To conduct the tests, a subset of the **MovieLens-20M**[2] dataset was used, containing 2,809,860 interactions. Users with at least 30 interactions were selected, with the last 30 interactions divided into

[2]https://grouplens.org/datasets/movielens/20m

10 for the test set and 20 for the training set. From this, 4,000 users were randomly selected for the test sample.

The training sample contains 2,769,860 interactions, 138,493 distinct users, and 12,366 unique movies/TV shows. The test sample comprises 40,000 interactions with 4,000 unique users and 599 items.

Both training and test sets use additional information related to genres and popularity for calibration techniques. The dataset includes 19 genres (e.g., Action, Drama, Thriller), used to ensure fairness in genres. A genre distribution is computed for each user based on the training data, utilized by post-processing calibration methods for traditional approaches or incorporated into the prompts for the LLM.

For popularity, items are divided into three categories: **Head (H)** for the top 20% of interactions, **Tail (T)** for the bottom 20%, and **Mid (M)** for the rest, based on Pareto's principle.

Users are categorized into three groups based on [9]: **Block-Buster (BB)** for users consuming at least 50% of the most popular items, **Niche (N)** for users consuming at least 50% of the least popular items, and **Diverse (D)** for users with preferences differing from the other two groups.

4.2 ChatGPT 3.5 Turbo

Our experiments aim to emulate a traditional model closely, ensuring the capabilities of the LLM are measurable while minimizing prompt engineering influence. Recommendations are generated through a minimal prompt and the user's past interactions without a conversational context.

For bias evaluation, items are categorized into blockbuster, medium, and niche groups for popularity bias, and similar approaches are applied for gender fairness. Four prompts were designed and tested with 15 random users for each scenario to avoid bias from random prompts.

For the Baseline execution (1), the prompt with the highest Mean Average Precision (MAP) score was chosen. The same process was used for scenarios (2) and (3), with prompts designed to address popularity bias or fairness constraints. The best prompt for each scenario was selected based on empirical evaluation, using the MRMC of popularity for scenario (2) and the MRMC of genre for scenario (3). Figure 2 shows the three resulting prompts used in our experiments[3].

Finally, a mechanism was implemented to ensure that the recommendations returned by the model are contained within the dataset so that the metrics are not influenced. The process starts with a request to the OpenAI API[4]. The response is checked to see if at least 5 titles are in the dataset. If fewer than 5 titles exist, it checks whether fewer than 2 interactions exist for the same user. If this condition is true, it increments the prompt, reports which recommendations were in the dataset, and then repeats the process. If there are at least 5 titles in the dataset or no fewer than 2 interactions for the same user, the recommendations are saved, and the process iterates to the following user.

[3]All tested prompts can be found in the project's GitHub: https://github.com/CuriousGu/llm_zeroshot_calibration.
[4]https://api.openai.com/v1/

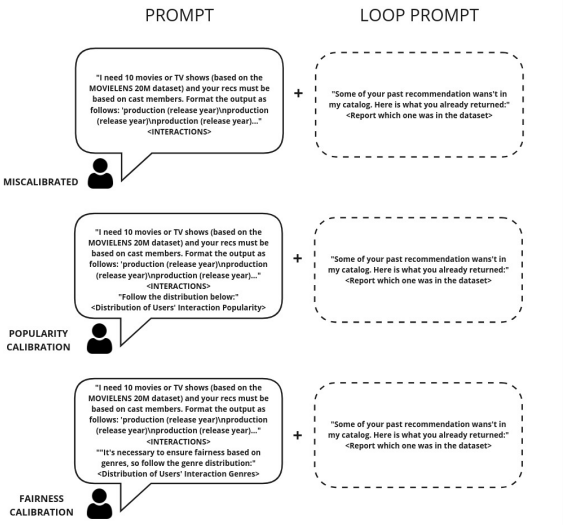

Figure 2: Template of prompts used in three different contexts.

It is worth mentioning that the structures are very similar to ensure that the effects were caused by the model and not only by the prompt differences.

4.3 Metrics

In our experiments, we evaluated the effects of different calibrations in terms of Mean Reciprocal Rank (MRR) for accuracy, Mean Rank Miscalibration (MRMC) [4] for fairness, and Long Tail Coverage (LTC) [15] for popularity bias. MRR ranges from 0 to 1, where higher values are better and Lower values for LTC mean recommended items are popular. In the case of MRMC, which covers the interval $[0, 1]$ (lower is better), we also use the harmonic mean (or F1 score) between MRMC of genres and popularity, where higher values are better:

$$F1 = 2\frac{(1 - MRMC\ Genre) * (1 - MRMC\ Pop)}{(1 - MRMC\ Genre) + (1 - MRMC\ Pop)} \quad (1)$$

5 RESULTS

Table 5 presents the results of our experiments. Analyzing the **LTC** metric, which indicates long-tail coverage, it is evident that the SVD++ recommender algorithm calibrated with Steck's [18] and Personalized [15] approaches achieved the best values, meaning they returned more diverse recommendations. Regarding the LLM, the Fairness Recalibration LLM obtained better results. In this approach, the prompt asks the model to recalibrate recommendations according to a genre distribution, resulting in a better coverage of the long-tail curve.

Regarding **MRMC**, this metric evaluates how much the recommendations distribution differs from the user's profile in terms of genre distribution (MRMC_g) and popularity (MRMC_p). In the first case, the results show that the two traditional approaches involving genre calibration, SVD++ calibrated with Steck's [18] and SVD++

calibrated with Two Stage [17], achieved the best results. Among the LLM approaches, Popularity Debiasing yielded the most fair values in this aspect. Concerning popularity, the table shows that the LLM effectively provided calibrated recommendations to users according to their profiles. We note a slight improvement in MRMC Pop for the Popularity Debiasing approach compared to the other two LLM-based approaches. Regarding the traditional methods, SVD++ calibrated with CP [1] performed better.

The **F1 Score** metric confirms the good results of the LLM approaches concerning genres and popularity fairness. The Popularity Debiasing approach obtained the best values in this aspect. Among the traditional approaches, the SVD++ calibrated with CP performed better.

Another important aspect to analyze is accuracy, which can be verified by the **MRR** metric. This metric shows that the LLM approaches could recommend more relevant items to the users, with the Baseline and Fairness Recalibration performing almost identically. On the other hand, the traditional approaches had significantly lower values in terms of accuracy, indicating that the LLM approaches managed to balance fairness, diversity, and accuracy in returning recommendations.

Table 1: Comparison of LLM approaches with traditional approaches on the MovieLens 20M dataset.

Algorithm	LTC	MRMC_g	MRMC_p	F1	MRR
Baseline LLM	0.034	0.318	0.022	0.803	0.076
Pop. Debiasing LLM	0.024	0.300	0.020	0.816	0.075
Fairness Recalib. LLM	0.046	0.327	0.027	0.796	0.078
SVD++ + CP	0.032	0.530	0.189	0.595	0.061
SVD++ + Genres	0.050	0.204	0.649	0.484	0.018
SVD++ + Personalized	0.050	0.217	0.647	0.486	0.015
SVD++ + Two Stage	0.049	0.200	0.653	0.484	0.011

6 CONCLUSION

This paper has presented a comprehensive empirical evaluation of Zero-Shot LLMs, particularly focusing on ChatGPT 3.5 Turbo, in addressing popularity bias and ensuring fairness in movie and TV show recommendations. Our study fills a critical gap in the existing literature by directly comparing the performance of LLMs with traditional post-processing algorithms used in recommender systems.

Our experiments show that LLMs demonstrated robust performance in calibrating recommendations, with improvements in MRR, MAP, and MRMC metrics. On the contrary, the long-tail coverage was improved by traditional methods, indicating better diversity on recommendations.

In future work, we plan to explore the application of Zero-Shot LLMs across a wider range of domains and content types to validate and extend our findings. We also plan to conduct user-centric studies to assess the real-world impact of LLM-generated recommendations on user satisfaction and trust.

ACKNOWLEDGMENT

The authors would like to thank the financial support from FAPESP and CNPq.

REFERENCES

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. 2021. User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM conference on user modeling, adaptation and personalization*. 119–129.

[2] Sarah Alnegheimish, Linh Nguyen, Laure Berti-Equille, and Kalyan Veeramachaneni. 2024. Large language models can be zero-shot anomaly detectors for time series? *arXiv preprint arXiv:2405.14755* (2024).

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] Diego Corrêa da Silva, Marcelo Garcia Manzato, and Frederico Araújo Durão. 2021. Exploiting personalized calibration and metrics for fairness recommendation. *Expert Systems with Applications* 181 (2021), 115112.

[5] Dario Di Palma, Giovanni Maria Biancofiore, Vito Walter Anelli, Fedelucio Narducci, Tommaso Di Noia, and Eugenio Di Sciascio. 2023. Evaluating chatgpt as a recommender system: A rigorous approach. *arXiv preprint arXiv:2309.03613* (2023).

[6] Mateo Gutierrez Granada, Dina Zilbershtein, Daan Odijk, and Francesco Barile. 2023. VideolandGPT: A User Study on a Conversational Recommender System. *arXiv preprint arXiv:2309.03645* (2023).

[7] Joshua Harris, Timothy Laurence, Leo Loman, Fan Grayson, Toby Nonnenmacher, Harry Long, Loes WalsGriffith, Amy Douglas, Holly Fountain, Stelios Georgiou, et al. 2024. Evaluating Large Language Models for Public Health Classification and Extraction Tasks. *arXiv preprint arXiv:2405.14766* (2024).

[8] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM international conference on information and knowledge management*. 720–730.

[9] Abdollahpouri Himan, Mansoury Masoud, Burke Robin, and Mobasher Bamshad. 2019. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286* (2019).

[10] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*. Springer, 364–381.

[11] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 426–434.

[12] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023. A preliminary study of chatgpt on news recommendation: Personalization, provider fairness, fake news. *arXiv preprint arXiv:2306.10702* (2023).

[13] Junling Liu, Chao Liu, Peilin Zhou, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is chatgpt a good recommender? a preliminary study. *arXiv preprint arXiv:2304.10149* (2023).

[14] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2022. Fairness in rankings and recommendations: an overview. *The VLDB Journal* (2022), 1–28.

[15] Andre Sacilotti, Rodrigo Ferrari de Souza, and Marcelo Garcia Manzato. 2023. Counteracting popularity-bias and improving diversity through calibrated recommendations. In *Proceedings*.

[16] Yubo Shu, Hansu Gu, Peng Zhang, Haonan Zhang, Tun Lu, Dongsheng Li, and Ning Gu. 2023. RAH! RecSys-Assistant-Human: A Human-Central Recommendation Framework with Large Language Models. *arXiv preprint arXiv:2308.09904* (2023).

[17] Rodrigo Souza and Marcelo Manzato. 2024. A Two-Stage Calibration Approach for Mitigating Bias and Fairness in Recommender Systems. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. 1659–1661.

[18] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*. 154–162.

[19] Xin Xu, Tong Xiao, Zitong Chao, Zhenya Huang, Can Yang, and Yang Wang. 2024. Can LLMs Solve longer Math Word Problems Better? *arXiv preprint arXiv:2405.14804* (2024).

[20] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 993–999.

[21] Lemei Zhang, Peng Liu, Yashar Deldjoo, Yong Zheng, and Jon Atle Gulla. 2024. Understanding Language Modeling Paradigm Adaptations in Recommender Systems: Lessons Learned and Open Challenges. *arXiv preprint arXiv:2404.03788* (2024).