A COMPARATIVE STUDY OF FAIRNESS IN AI-ENABLED AND LLM-BASED RECOMMENDATION SYSTEMS

Jiaqian Zhang
School of Management,
Hefei University of Technology
Hefei, Anhui 230009, China
jqzhang@mail.hfut.edu.cn

Jianshan Sun
School of Management
Hefei University of Technology
Hefei, Anhui 230009, China
sunjs9413@hfut.edu.cn

Ying Xue[1]
School of Management
Hefei University of Technology
Hefei, Anhui 230009, China
xueying@mail.hfut.edu.cn

Yezheng Liu
School of Management
Hefei University of Technology
Hefei, Anhui 230009, China
liuyezheng@hfut.edu.cn

ABSTRACT

The advent of Artificial General Intelligence (AGI) has heralded a new era in e-commerce, empowering recommendation systems with its advanced capabilities, yet concerns about fairness in these systems have emerged. This paper presents a comprehensive study examining user gender fairness in various recommendation algorithms and domains, with a particular focus on AI-enabled and LLM-based recommendation systems. Concretely, we conduct experiments on four datasets from distinct domains to evaluate and compare the gender fairness of eleven recommendation models from six families under several fairness metrics, such as Absolute Difference, Item Coverage, and Gini coefficient. Our findings reveal significant disparities in recommendation accuracy and diversity between male and female users, highlighting the need for fair and unbiased recommendation services in e-commerce. Notably, the latest LLM-based recommendation model demonstrates promising fairness in terms of Item Coverage and Gini coefficient between male and female users, suggesting its potential in mitigating gender bias in recommendations. This study contributes to the understanding of gender fairness in different families of recommendation systems and provides insights for recommendation system design in e-commence platforms.

Keywords: Recommendation systems; Gender fairness; Fairness evaluation; LLM-based recommendation

1. Introduction

Artificial intelligence (AI) has become a cornerstone in e-commerce, particularly through recommendation systems (RecSys) that have transformed people's lives by providing personalized product services across various platforms. With the emergence of artificial general intelligence (AGI), the integration of large language models (LLMs) into recommendation systems has further enhanced their capabilities (L. Wu et al., 2024). However, some social issues

Cite: Zhang, J., Sun, J., Xue, Y., & Liu, Y. (2025, July). A Comparative Study of Fairness in AI-Enabled and LLM-Based Recommendation Systems. *Journal of Electronic Commerce Research*, *26*(3).
[1] Correspongding author

such as fairness have emerged. Over the past two decades of rapid development in recommendation technology, researchers have been dedicated to designing more sophisticated models to provide more accurate and personalized services to meet user needs. As the reliance on AI-enabled recommendation systems grows, so does the scrutiny on their fairness, particularly in how they treat different demographic groups.

Researchers have highlighted potential biases that can lead to unfair treatment of certain groups. Among these biases, gender-based discrimination stands out as a critical area of concern, which is the primary focus of this study. Several studies (M. Ekstrand et al., 2018; M. D. Ekstrand & Kluver, 2021; Melchiorre et al., 2021; Wang & Chen, 2021) have explored gender bias in recommendation systems, finding significant disparity in recommendation effectiveness for male and female users. This gender bias can limit the consumption choices and experiences of disadvantaged groups, undermine consumer trust, and raise ethical and legal concerns (Ren et al., 2024). However, there are still relatively few studies specifically on gender fairness in recommendation systems, and most of the existing researches focus on specific recommendation models and scenarios, such as traditional collaborative filtering models (M. Ekstrand et al., 2018) and music recommendation (Melchiorre et al., 2021), lacking systematization and comprehensiveness. In this paper, we study user gender fairness of recommendation models from various families comparatively in multiple data scenarios, examining the extent to which these models may inadvertently favor one gender over another in their recommendations. Besides, we evaluate fairness in a wider dimension during the comparative study. In other words, we evaluate user gender bias in accuracy measurements and beyond-accuracy measurements, such as diversity.

We begin by reviewing the evolution of recommendation systems and research status of fairness in recommendation systems, our review reveals a growing body of work aimed at enhancing fairness and maintaining the overall performance of recommendation systems, while ignoring several critical questions about whether many of the current recommendation algorithms are unfair inherently, and if so, how and why. To explore and clarify the above questions, we conduct a comprehensive experimental analysis of four traditional recommendation models, six AI-enabled recommendation models, and one state-of-the-art LLM-based recommendation model. We evaluate these models across four diverse datasets, evaluating their accuracy, diversity and fairness of different gender user groups using established metrics such as Absolute Difference, Item Coverage, and Gini coefficient. Our findings provide insights into the gender fairness of recommendation systems, revealing significant differences in recommendation accuracy and diversity between male and female users. We also explore the potential of large language models to offer more equitable recommendations, suggesting that these models may hold the key to addressing gender bias in recommendation systems.

This paper aims to contribute to the understanding of gender fairness in recommendation systems and to inform future research and development efforts. By shedding light on the current state of gender fairness in recommendation systems, we hope to inspire further investigation into the factors that influence fairness and the development of fair models that are more issue-solving oriented.

2. Related Works

To provide a foundation for understanding current recommendation systems and the challenges they face in achieving fairness, we will review the development trajectory of recommendation systems from traditional collaborative filtering to the latest LLM-baesd recommendation systems, as well as the research status of fairness in recommendation systems, including diverse fairness concepts and optimization strategies that have been proposed to address these issues.

2.1. Recommendation Systems

2.1.1 Traditional Recommendation Systems

Traditional recommendation systems, represented by heuristic collaborative filtering (CF) and matrix factorization (MF) methods, mainly rely on user-item historical interaction data to generate recommendations. Heuristic methods, dating back to an email screening system (Goldberg et al., 1992) and popularized by Amazon's collaborative filtering (Linden et al., 2003), calculate similarity between users or items based on their interaction history. CF is efficient and interpretable, yet it may face challenges such as the head effect, poor generalization in sparse data, and scalability (Kim & Ahn, 2011). MF methods, which gained prominence in the Netflix Prize Challenge in 2006 (Bell et al., 2007), decompose the interaction matrix into low-rank matrices to capture the latent user and item features for predicting user preferences. They offer enhanced performance on sparse matrices and better scalability. However, they typically assume linear data patterns and may struggle with generalization for unseen feature combinations, especially in cold start scenarios.

2.1.2 AI-Enabled Recommendation Systems

Recommenders with deep neural networks integrated, which we call AI-enabled recommendation systems, have flourished in the last decade. They excel in processing diverse data types and offering more relevant recommendations

through strong feature learning ability. We mainly review recommendation systems based on neural collaborative filtering, autoencoder, and graph neural network.

Neural collaborative filtering-based recommendation systems, popular in the 2010s, use multi-layer neural networks to automatically learn implicit user and item features, capturing complex non-linear relationships in data. For example, ConvNCF employs CNNs to model higher-order user-item correlations (He et al., 2016), and NeuMF combines matrix factorization with neural networks to capture both linear and non-linear correlations (He et al., 2017). However, these models may lack interpretability as black-box models and are prone to overfitting with sparse data (Wu et al., 2023).

Autoencoder-based recommendation systems use autoencoders to learn low-dimensional representations of users and items for recommendations. They excel in processing sparse data and learning effective representations. For example, CDAE introduces noise to the input layer and employs denoising to improve robustness and generalization capabilities (Wu et al., 2016), while Mult-VAE boosts recommendation by integrating multi-source information to capture multi-dimensional latent features (Aguila et al., 2023).

Graph neural networks-based recommendation systems leverage the graph structure of user-item interactions to capture complex collaborative signals by propagating information across the graphs. For instance, NGCF and LightGCN learn node embeddings and model high-order connectivity through message passing and feature aggregation on the graph. (Wang et al., 2019; He et al., 2020). These models improve recommendation accuracy and address challenges like data sparsity and cold-start problems.

2.1.3 LLM-based Recommendation Systems

At the moment, recommendation systems based on large language models are a prominent emerging direction in the field of recommendation system (Wu et al., 2024). These models leverage pre-trained large language models (LLMs) to capture user needs and preferences through their powerful semantic understanding and generation capabilities, thereby enhancing the performance of recommendation systems. They are adept at processing and understanding a vast amount of textual data and generating richer, more accurate, and more personalized recommendations.

To sum up, although the recommendation technology is updated and iterated rapidly, there is no absolute winner among the old and the new recommendation models, and the early classic model can still be widely used in business combined with the newer technology (Moon et al., 2019; Lin, 2024). Sometimes these different families of recommendation algorithms can be merged into a hybrid model recommendation system according to the needs of business scenarios (Ye et al., 2019). Whether it is a single model or a hybrid model, most of them aim to improve the accuracy of recommendations to enhance the user experience and maximize the benefits of multi-stakeholders.

2.2. Fairness in Recommendation Systems

AI algorithms' bias in sensitive areas such as STEM career advertising (Lambrecht & Tucker, 2019), algorithmic recruiting (Ochmann et al., 2024), credit financing (Fu et al., 2021), emotional AI applications (Rhue, 2024) has sparked widespread fairness concerns. These concerns naturally extend to e-commerce and personalized product consumption (Ren et al., 2024; Weith & Matt, 2023), where fairness is essential not only for ethical considerations but also for enhancing user trust and business sustainability. Recommendation systems, as a prominent application of AI in e-commerce, are no exception and require thorough examination of their fairness implications. Relevant research is burgeoning, covering a wide range of topics. We surveyed existing studies from two aspects: the concepts and evaluations of fairness in recommendation systems (summarized in Table 1) and the improvements and optimizations of fairness in recommendation systems (summarized in Table 2). Based on this, we have identified that there is still room for progress in some critical aspects of recommendation system fairness research, which has motivated us to conduct this work.

2.2.1. The Concepts and Evaluations of Fairness in RecSys

For the fairness evaluation of recommendation systems, researchers have proposed many different fairness concepts, which can be categorized from various perspectives. Generally, from the perspective of demographic unit in recommendation systems, there are Group Fairness (Yao & Huang, 2017) and Individual Fairness (Biega et al., 2018). From the perspective of stakeholders in RSs, fairness is categorized into user-side (consumer) fairness, item-side (provider) fairness and two-side (CP or multi-sided) fairness. User-side fairness typically denotes no significant performance disparities among user groups divided by sensitive attributes. These attributes commonly refer to demographic characteristics (e.g., age, gender), yet some studies explore non-demographic properties, such as user activity (Zhang et al., 2024; Han et al., 2024; Wang et al., 2025), user interest diversity (Y. Zhao, Xu, et al., 2024), the rarity of users' disease (Z. Zhao et al., 2024), user sexual orientation (Y. Zhao, Wang, et al., 2024), etc. Most research focuses on recommendation performance disparities in accuracy (relevance or utility) measurements, with several proposing user bias in beyond-accuracy measurements (Wang & Chen, 2021; Melchiorre et al., 2021). Item-side fairness generally implies equal exposure opportunity among different products (Qi et al., 2022; Shang et al., 2024;

Zhu et al., 2020). Two-sided fairness means that a recommendation system is supposed to treat all stakeholders fairly (Naghiaei et al., 2022; Patro et al., 2020), usually with tradeoff objective between user fairness and item fairness (Greenwood et al., 2024). Some new concepts of fairness have emerged recently, such as long-term fairness (Ge et al., 2021), personalized fairness (Li et al., 2021), selective fairness (Wu et al., 2022), and explainable fairness (Ge et al., 2022). On the basis of various fairness concepts, researchers design metrics respectively to quantify them (Y. Wu et al., 2024), such as Absolute Difference (Fu et al., 2020; Li et al., 2021; Zhu et al., 2018), KS statistic (Kamishima et al., 2018; Zhu et al., 2018), Pairwise Fairness (Beutel et al., 2019), KL-divergence (Steck, 2018; Wan et al., 2020), Entropy (Patro et al., 2020), Gini coefficient (Fu et al., 2020; Ge et al., 2021; Leonhardt et al., 2018), and so on. Several reviews on the fairness of recommendation systems (Wang et al., 2023; Wu et al., 2024; Deldjoo et al., 2024) have classified and summarized these numerous concepts and metrics of fairness in various forms. For example, Wang et al. (2023) summarize the existing fairness metrics as belonging to two fairness concepts Process Fairness and Outcome Fairness. In this paper we focus on the user-oriented group fairness in recommendation systems, with particular attention to gender attribute. We choose common metrics, Absolute Difference, Gini coefficient, Variance, to measure the gender bias of user groups from accuracy and diversity dimension.

Table 1: the Concepts of Fairness in RecSys

Views	Category	References
Demographic Unit	Group Fairness	Burke et al. (2017); Kamishima & Akaho (2017); Yao & Huang (2017); M. Ekstrand et al. (2018); Farnadi et al. (2018); Kamishima et al. (2018); M. D. Ekstrand et al. (2018); Zhu et al. (2018); Geyik et al. (2019); M. D. Ekstrand & Kluver (2021); R. Z. Li et al. (2021); Ge et al. (2021); Islam et al. (2021); Naghiaei et al. (2022); Rahmani et al. (2022); Boratto et al. (2022, 2023); Chen et al. (2023);
	Individual Fairness	Biega et al. (2018); Rastegarpanah et al. (2019); Patro et al. (2020); Mansoury et al. (2020, 2022); Y. Li et al. (2021); J. Li et al. (2022);
Stakeholders	User/Consumer Fairness	Burke et al. (2017); Kamishima & Akaho (2017); Yao & Huang (2017); M. Ekstrand et al. (2018); M. D. Ekstrand et al. (2018); Farnadi et al. (2018); Kamishima et al. (2018); Leonhardt et al. (2018); Steck (2018); Zhu et al. (2018); M. D. Ekstrand & Kluver (2021); R. Z. Li et al. (2021); Y. Li et al. (2021); Sonboli et al. (2021); Boratto et al. (2022, 2023); Rahmani et al. (2022); Zhang et al. (2024); Han et al. (2024); Y. Zhao, Xu, et al. (2024); Z. Zhao et al. (2024); Y. Zhao, Wang, et al. (2024); Wang et al. (2025)
	Item/Provider Fairness	Abdollahpouri et al. (2017, 2019, 2020); Biega et al. (2018); Singh & Joachims (2018); Ferraro (2019); Geyik et al. (2019); Morik et al. (2020); Li et al. (2022); Zhu et al. (2020); Naghiaei et al. (2022); Qi et al. (2022); Chen et al. (2023); Do & Usunier (2023); Jiang et al. (2024); Shang et al. (2024)
	Multi-sided/CP Fairness	Naghiaei et al. (2022); Patro et al. (2020); Wang, et al. (2024); (Greenwood et al., 2024)

2.2.2. The Improvements and Optimizations of Fairness in RecSys

Relatively, a greater number of papers focus on how to enhance fairness while optimizing the overall performance of the model, so that accuracy is not significantly compromised. Towards various recommendation domains and basic recommendation models, researchers have designed a multitude of sophisticated fairness-aware recommendation models employing a variety of classical or advanced optimization techniques at different stages of the recommendation system pipeline.

Firstly, existing researches consider fairness in different recommendation scenarios. For example, Qi et al. (2022) and Wu et al. (2021) aim to improve the fairness in news recommendation for both consumer and provider; Zhao et al. (2024) addressed the unfairness issue in medication recommendation, enabling patients with rare diseases to obtain accurate recommendations. Ferraro (2019), Melchiorre et al. (2021) and Dinnissen (2024) studied fairness in music recommendation. Geyik et al. (2019), Islam et al. (2021) and Sühr et al. (2021) designed fairness-aware models to advance fairness in employment recommendation. Lee et al. (2014), Liu et al. (2019) and Smith et al. (2023) paid attention to fairness issue in lending platform.

Meanwhile, researchers achieve their fairness goals based on diverse basic models from various recommendation system families, as described in the previous section. The optimization approaches (also known as optimization

strategies or techniques) may vary depending on different families of recommendation models according to their respective characteristics. For instance, Burke et al. (2017, 2018) created a balanced neighborhood in which recommendations for all users are generated from neighborhoods that are balanced with respect to the protected and unprotected classes. For matrix factorization models, most studies (Kamishima et al., 2018; Kamishima & Akaho, 2017; Yao & Huang, 2017; Zhu et al., 2018) use regularization constraints to remove sensitive information from models. For deep learning recommendation models, one of the typical approaches is to obtain unbiased user or item representations through adversarial learning (Li et al., 2021; Wu et al., 2021), or especially, Islam et al. (2021) use a pre-training and fine-tuning approach with bias correction techniques. Towards graph-based recommendation models, additional techniques tailored to graph-learning, like re-wiring (Wang et al., 2022) and graph modification (Current et al., 2022), are employed to obtain fair graph embeddings and thus achieve fairness. With regard to fairness in LLM-based recommendation, the latest research introduces fairness evaluation frameworks suitable for LLM recommendations, such as FaiRLLM (Zhang et al., 2023) and CFaiRLLM (Deldjoo & Nazary, 2024), to identify and quantify potential biases, and Deldjoo & di Noia (2024) explores how to reduce stereotypical recommendation biases through various user profiling strategies, concretely, by constructing user profiles that more accurately reflect their preferences.

Generally, the existing optimization methods for fairness in recommendation systems are divided into Pre-processing, In-processing and Post-processing, which is a simple division that is widely accepted. That means to design optimization strategies before model training (usually by re-balancing the input data, such as re-labeling, re-sampling and data modification), on the model itself (usually by adjusting the model structure or training objective), and on the recommendation results generated initially (usually by re-ranking the recommendation lists).

Table 2: the Optimizations of Fairness in RecSys

Dimension	Category	References
Recommendation Family	Memory-based RecSys	Burke et al. (2017, 2018); Ekstrand et al. (2018)
	MF-based RecSys	Yao & Huang (2017); Kamishima & Akaho, (2017); Farnadi et al. (2018); Kamishima et al. (2018); Zhu et al. (2018); Rastegarpanah et al. (2019); Li, Chen, Fu, et al. (2021)
	DL-based RecSys	Islam et al. (2021); R. Z. Li et al. (2021); Y. Li et al. (2021); C. Wu et al. (2021)
	GNN-based RecSys	Current et al. (2022); Fu et al. (2020); L. Wu et al. (2021); Xu et al. (2021); Chen et al. (2023)
	LLM-based RecSys	Deldjoo & di Noia (2024); Deldjoo & Nazary (2024); Zhang et al. (2023)
Recommendation Pipline	Pre-processing	Rastegarpanah et al. (2019);
	In-processing	Yao & Huang (2017); Burke et al. (2018); Morik et al. (2020); Wan et al. (2020); C. Wu et al. (2021); L. Wu et al. (2021); R. Z. Li et al. (2021); Li, Chen, Xu, et al. (2021); Ge et al. (2021); Islam et al. (2021); Ge et al. (2022); Qi et al. (2022); Li et al. (2022)
	Post-processing	Biega et al. (2018); Geyik et al. (2019); Liu et al. (2019); Fu et al. (2020); Patro et al. (2020); Li, Chen, Fu, et al. (2021); ; Naghiaei et al. (2022)
Optimazation Technique	Data modification	Ekstrand et al. (2018); Rastegarpanah et al. (2019)
	Regularization	Burke et al. (2017); Zhu et al. (2018); Fu et al. (2020); Kamishima & Akaho (2017); Li, Chen, Xu, et al. (2021); Yao & Huang (2017); Beutel et al. (2019); Morik et al. (2020); Wan et al. (2020)
	Adversarial learning	Bose & Hamilton (2019); Wu et al. (2021); R. Z. Li et al. (2021); Li, Chen, Xu, et al. (2021); C. Wu et al. (2021); L. Wu et al. (2021)
	…	

Broadly speaking, these studies aimed at achieving a fairer recommendation system may have addressed some potential unfairness issues, and recently, several research efforts (Boratto et al., 2022, 2023; Rahmani et al., 2022) have been made to evaluate the proposed fairness-aware recommendation models. Some research has also suggested that algorithms cannot achieve fairness alone, and it needs human-technology collaboration (Boston College et al., 2021). However, while pursuing to make it fairer, we would like to ask that ***would many current AI-enabled***

recommendation algorithms be considered inherently unfair? If so, which models are more unfair and which are less so? Does unfairness vary across different datasets? What factors might influence fairness? Most relevant research in this field pay attention to how to improve various fairness odjectives excessively, while neglecting to inspect the fairness status of numerous existing recommendation models. Several studies (Mansoury et al., 2019; Deldjoo et al., 2021; Guo et al., 2023; Anelli et al., 2023) carried out comparison and evaluation works on the fairness of some typical recommendation models, but they did not cover the whole recommendation families. For example, Mansoury et al. (2019) and Deldjoo et al. (2021) involved a series of models in the unified fairness evaluation, including non-personalized models, memory-based models and matrix factorization models, whereas more state-of-the-art deep recommenders were not considered. Guo et al. (2023) proposed a framework called FairRec to support fairness testing of recommendation systems, but the framework contains only deep learning recommendation models. Besides, Anelli et al. (2023) audited fairness in graph collaborative filtering, exploring how different graph CF strategies affect various model performance including fairness. They found in the motivation example that there was no distinct winner when it comes to user fairness, with traditional CF models performing better, while some graph CF models did not achieve significant results. Nevertheless, they only compared two traditional CF models BPRMF and RP$^3\beta$, and a more systematic fairness comparison is not the purpose of this study. On the whole, there is still a lack of comprehensive research in a unified way on fairness comparison and evaluation of the existing different families of recommendation models, which this paper aims to complete.

3. Recommendation Models

For a general recommendation task, U denotes the set of users and I denotes the set of items. Let $U \times I \to R$, where R is a totally order set. And $r_{ui} \in R$ represent the score of user u for product i. m and n denote the total number of users and items, respectively.

As technological approaches continue to evolve, personalized recommendation models continue to be iterated and optimized. In this process, we selected two classical algorithms at several key stages in the development of personalized recommendation methods, respectively. Specifically, we have selected ten classical models from heuristics, matrix factorization based, neural collaborative filtering based, autoencoder based, and graph neural network-based recommendation models, as well as one LLM-baesd recommendation model. The next sections describe each of the selected recommendation methods.

3.1. Heuristic Recommendation Models

Heuristic recommendation methods are simple but efficient methods to generate recommendation results from some rules. These methods do not require complex training, and the recommendation results are highly interpretable. We chose two KNN-based recommendation methods, UserKNN and ItemKNN, to generate personalized recommendation lists from similar neighbors by calculating the similarity between users or items. Similarity can be calculated by cosine similarity, Pearson's correlation coefficient, and other approaches. Take the cosine similarity as an example, the formula is as follows:

$$sim(i,j) = \frac{\sum_{u \in U_{ij}} r_{u,i} \cdot r_{u,j}}{\sqrt{\sum_{u \in U_i} r_{u,i}^2} \cdot \sqrt{\sum_{u \in U_j} r_{u,j}^2}} \tag{1}$$

$$\hat{r}_{u,i} = \frac{\sum_{i \in I_u} sim(i,j) \cdot r_{u,i}}{\sum_{i \in I_u} |sim(i,j)|} \tag{2}$$

Where U_i means the set of users who interact with item i, I_u means the set of items which interact with user u.

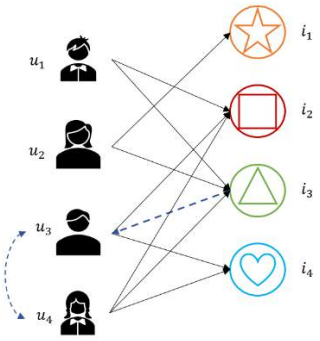

Figure 1: Basic Structure for UserKNN

ItemKNN (Deshpande & Karypis, 2004): Item-based k-nearest neighbor method. This approach first calculates the similarity between items using methods such as cosine similarity. It then leverages this similarity to identify and recommend items that are most similar to the user's previously interacted items.

UserKNN (Breese et al., 1998): User-based k-nearest neighbor method. Similar in principle to the item-based approach, this method calculates the similarity between users. It then identifies users with similar preferences and recommends items that those similar users have liked or interacted with.

3.2. Matrix Factorization-Based Recommendation Models

Matrix factorization is the process of decomposing a complex matrix into the product of two or more simple matrices. Matrix factorization-based methods decompose the user-item matrix into two low-dimensional matrices P and Q, and then predict ratings and recommended items by computing the inner product of the user and item vectors.

$$\hat{R}_{u,i} = P_u \cdot Q_i^T \quad (3)$$

Bayesian Personalized Ranking （*BPR*） (Rendle et al., 2009): BPR is a pairwise ranking method that optimizes personal recommendation by maximizing the posterior probability of observed user-item interactions. The training data for BPR contains positive and negative pairs (missing values).

Factorization Machines (FM) (Rendle, 2010): FM is a popular solution for efficiently utilizing second-order feature interactions. It embeds features into the hidden space and models the interaction between features by the inner product of the embedding vectors.

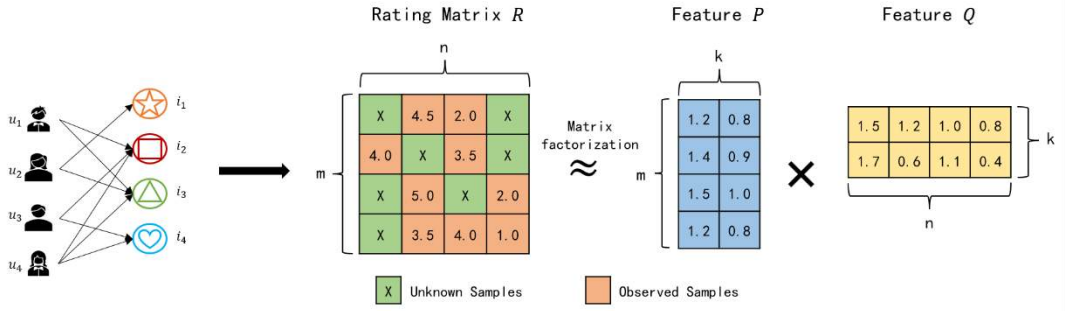

Figure 2: Basic Framework for Matrix Factorization

3.3. Neural Collaborative Filtering-Based Recommendation Models

Neural collaborative filtering is a generalized framework which replaces the inner product with a neural architecture that can learn arbitrary functions from data. Compared to traditional matrix factorization methods, Neural collaborative filtering-based methods are able to capture nonlinear and higher-order interactive signals through multi-layer neural networks.

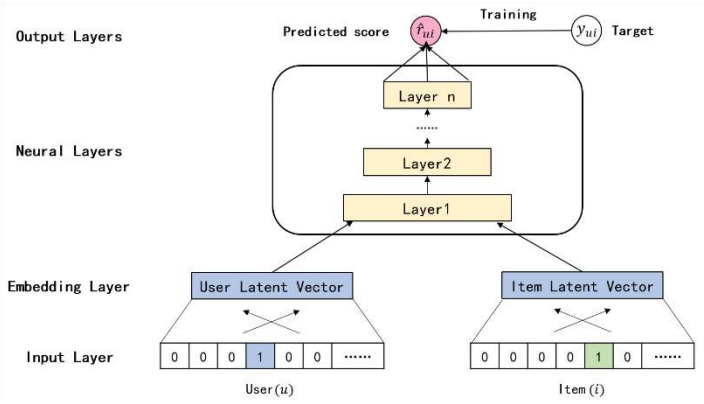

Figure 3: Basic Framework for Neural collaborative filtering-based Recommendation

The predicted rating of product i by user u is calculated as follows:

$$\hat{r}_{u,i} = e_u^T \cdot e_i \quad (4)$$

Where e_u and e_i is the embedding vector of user u and product i obtained by the model.

Neural Matrix Factorization (*NeuMF)* (He et al., 2017): NeuMF utilizes both MF and neural network MLP to fit the matching score, using the vector inner product to learn the association between user and item, while the MLP

partially captures other higher-order information about the two. The model can be divided into two parts: GMF and MLP.

Convolutional Neural Collaborative Filtering (ConvNCF) (He et al., 2016): ConvNCF captures higher-order correlations between embedded dimensions through outer product operations and use convolutional neural networks (CNNs) to learn these higher-order correlations.

3.4. Autoencoder-Based Recommendation Models

Autoencoders extract useful features by compressing the input data into a low-dimensional space through an encoder and recovering the original data through a decoder. Recommendation systems use this approach to extract user and item characteristics to optimize recommendation performance.

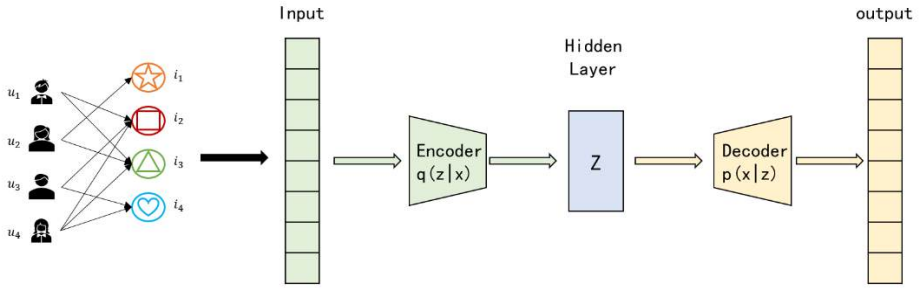

Figure 4: Basic Framework for Autoencoder-based Recommendation

The predicted rating of product i by user u is calculated as follows:

$$\hat{r}_{u,i} = f(W_i'^T z_u + b_i') \quad (5)$$

where W' and b' are the weight matrix and the offset vector for the output layer, respectively, z_u is the latent representation, and $f(\cdot)$ is a mapping function.

Collaborative Denoising Auto-Encoder (CDAE) (Wu et al., 2016): CDAE uses the idea of a noise-reducing self-encoder to construct the Top-n recommendation problem, where random noise is added to the input to improve the robustness of the model.

Variational Autoencoders for Collaborative Filtering (Mult-VAE) (Liang et al., 2018): Mult-VAE is a collaborative filtering model based on implicit feedback and variational autoencoders (VAE) that uses polynomial likelihood variational autoencoders to solve the problem of too many parameters when variational inference is used for recommendations.

3.5. Graph Neural Network-Based Recommendation Models

Graph neural networks (GNN) utilize neural networks to learn graph-structured data. A recommendation system based on graph neural networks constructs user and item information into a graph, learns the knock-in representations of nodes by aggregating neighboring information through the GNN approach, and extracts and mines features and patterns in the graph for interest prediction.

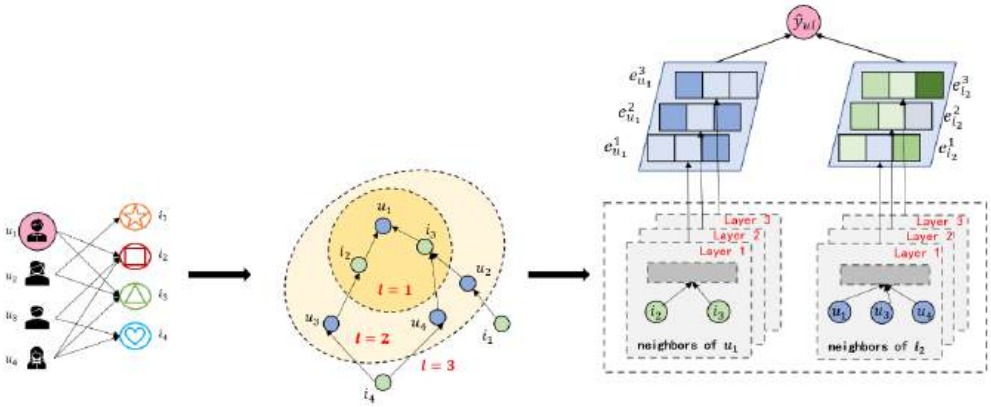

Figure 5: Basic Structure for GNN-based Recommendation

The predicted rating of product i by user u is calculated in the same way as in equation (4).

Neural Graph Collaborative Filtering (NGCF) (X. Wang et al., 2019): NGCF is a graph-based recommendation model that exploits higher-order connectivity information in user-item graphs by propagating embeddings over the

graph. The embeddings of user and item are allowed to interact with each other to obtain collaborative signals.

Light Graph Convolution Network (LightGCN) (He et al., 2020): LightGCN simplifies the GCN model, retaining only the neighborhood aggregation part for collaborative filtering. It employs simple weighting schemes and aggregators, while omitting feature transformations and nonlinear activations.

3.6. Large Language Model-Based Recommendation Models

GLM-4 (GLM et al., 2024): The GLM-4 model is pre-trained on a 10 trillion corpus consisting mainly of Chinese and English, and is further aligned for both Chinese and English usage to better understand the user's intent to effectively accomplish complex tasks.GLM-4 supports 128K contexts, which is well adapted to long context scenarios in our recommendation tasks. At the same time, GLM-4 is close to the state-of-the-art models (GPT-4-Turbo, Gemini 1.5 Pro, and Claude 3 Opus) in terms of standardized benchmarks as well as instruction adherence, long contexts, code problem solving, and agent capabilities.

We designed three different prompt word templates to explore the impact of gender variations on top-k recommendation results. Specifically, prompt word 1 (P1) serves as a baseline without explicitly specifying the user's gender; prompt word 2 (P2) explicitly assumes that the user's gender is male; and prompt word 3 (P3) explicitly assumes that the user's gender is female. In addition, to avoid the effect of duplicate recommendations, we remove items that the user has already interacted with from the list of candidate recommendations for prompt words, as follows in the template.

System: You are a movie recommendation expert. Based on the user's viewing history, please select 10 movies from the following movie list as the recommendation list to recommend to the user. Please note that each movie you recommend must be in the following movie list and do not output unrelated movie names. The movies in the list are sorted in descending order according to the user's preferences, and the recommendation list is output in the format of nested strings in the list, such as ['movie_name_1', 'movie_name_2', ...]

P1: The viewing history of an user is as follows:《Dumbo》,《Star Wars》……

P2: The viewing history of a male user is as follows:《Dumbo》,《Star Wars》……

P3: The viewing history of a female user is as follows:《Dumbo》,《Star Wars》……

Please select 10 movies from the following to recommend to users:

1. Zeus and Roxanne

2. Maverick

……

We called GLM-4-FLASH through the API provided by zhupu AI. To output a specified format that can be subsequently processed and to avoid over-conservative and repetitive recommendation results, we set temperature to 0.95 and top_p to 0.7. On top of that, we turned off the tool calls to prevent the introduction of external knowledge that leads to discrepancies in the recommendations.

Note that in this paper our goal is not to improve the performance of recommendation methods, but to compare the difference in fairness between different recommendation methods. Therefore, we have chosen a generalized recommendation framework and a classical recommendation model.

4. Experiment

In this section, we provide details on the datasets (Section 4.1), evaluation metrics (Section 4.2), and experimental setup (Section 4.3) used to compare the accuracy and fairness of different recommendation models.

4.1. Datasets

The fairness of recommendation systems is also usually related to the recommendation context and dataset properties. Hence, we chose four common recommendation and user behavior scenarios for movies, music, e-commerce, and Q&A, corresponding to four different datasets. Each dataset includes the user's sensitive attribute "gender". However, the proportion of men and women in each dataset is different. Since the size of different datasets varies very much, even by a factor of 100, this not only creates huge computational problems for the model, but also introduces new biases. Therefore, we sampled the data based on timestamps and selected data from different time periods in different datasets. To ensure that the amount of interaction data in these datasets does not vary too much, we took the data volume of the Movielens 100k dataset as a benchmark, and sample about 100,000 interactions data in each dataset.

The basic information about the four datasets is as follows.

Movielens (ML): a widely used dataset in movie recommendation research for developing and evaluating recommendation algorithms. It contains user ratings for movies, information about users and movies. Several versions of Movielens are available. In this study we use the ML-100K dataset, which consists of 100,000 ratings from 943 users on 1,682 movies.

Last.fm: a dataset that provides music recommendations, containing information about the user's favorite music, artists, albums and user profiles. We chose the Last.fm-artists dataset to recommend favorite artists for users, focusing on the period July-December 2006. This subset contains 111394 "favorites-artists" interactions between 1361 users and 51940 artists.

Ali_Display_Ad_Click (AliEC): log data provided by Alibaba for some users clicking on ads over an eight-day period, including advertising information, user information and user behavior logs, where each advertisement represents a specific product. We chose data from the first four days, including 104,295 click logs from 6,548 users on 56,111 ads.

ZhihuRec (Hao et al., 2021): a large-scale text query and recommendation dataset released by the Q&A community platform Zhihu, including users, questions, answers, authors, topics, and user search logs. We chose topics as recommendation targets and filtered the interaction data of users who registered in 2018, including 107,189 interaction records from 3,220 users on 6,797 topics.

Table 3 and Table 4 summarize more detailed characteristics of these datasets.

Table3: Statistics on the Number of Users of the Dataset

Dataset	inter_num	item_num	user_num	inter/user	inter/item	sparsity
Movielens	100000	1682	943	106.05	59.45	93.69%
Last.FM	111394	51940	1361	81.85	2.15	99.84%
AliEC	104295	56111	6548	15.93	1.86	99.97%
ZhihuRec	107189	6796	3220	33.29	15.77	99.51%

inter_num: Number of user interactions with the item
item_num: Number of items
user_num: Number of users

Table 4: Statistics on User Interactions by Gender in the Dataset

Dataset	male_ num	male_i nter	male_ ratio	male_inter/ user	female num	female inter	female _ratio	female_inter /user
Movielens	670	74260	0.71	110.84	273	25740	0.29	94.29
Last.FM	1096	90118	0.81	82.22	265	21276	0.20	80.29
AliEC	1339	21143	0.20	15.80	5209	83152	0.80	15.96
ZhihuRec	3032	101306	0.94	33.41	188	2883	0.06	15.34

male_num: Number of male users in the dataset
male_inter: Number of interactions with item by male users
male_ratio: male_num/user_num
female_num: Number of female users in the dataset
female_inter: Number of interactions with item by female users
female_ratio: female_num/user_num

4.2. Metrics

The evaluation metrics package is divided into two categories: accuracy metrics and fairness metrics.

4.2.1. Accuracy metrics

These metrics measure the quality of personalized recommendations.

Precision: Defined as the ratio of the user's favorite items to all recommended items in the system's recommendation list, i.e., how many of the recommended (predicted) items are actually of interest to the user.

$$Precision = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|} \quad (6)$$

Recall: Defined as the probability that a user's favorite item is recommended, i.e., how many of the items the user likes (clicks on) are recommended. This metric is more responsive to the performance of recommender systems in real-world scenarios.

$$Recall = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |T(u)|} \quad (7)$$

Mean Reciprocal Rank (MRR): Take the inverse of the standard answer's ranking among the results given by the evaluated system as its accuracy, and average it over all the questions. Specifically, for each user query, the system finds the first relevant item in the recommended list, and the reciprocal rank is computed as the inverse of that rank (i.e., $\frac{1}{rank}$). If a relevant item is found at rank 1, the reciprocal rank is 1; at rank 2, it is 0.5, and so on.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (8)$$

Where $|Q|$ is the number of queries, $rank_i$ is the rank position of the first relevant item for the ith query.

Normalized Discounted Cumulative Gain (NDCG): Evaluate the quality of the recommendation list by considering the rank order and relevance score of each item in the list. It gives more weight to higher-ranked relevant items by applying a logarithmic discount, emphasizing that relevant items shown earlier in the list are more valuable to users. The formula for NDCG for a single query or user is:

$$NDCG = \frac{DCG}{IDCG} \quad (9)$$

where DCG (Discounted Cumulative Gain) is calculated as:

$$DCG = \sum_{i=1}^{k}\frac{2^{rel_i}-1}{\log_2(i-1)} \quad (10)$$

where rel_i means the relevance score of the item at position i, k is the length of the recommended list. IDCG (Ideal DCG) is the maximum DCG value in the ideal case.

4.2.2. Coverage metrics

Coverage is a measure of the domain of the recommendation item. Low coverage can limit decision-help capabilities and thus be of less value to users (Herlocker et al., 2004).

Item Coverage (Ge et al., 2010): The coverage of a recommender system is a measure of the domain of items in the system over which the system can form predictions or make recommendations, can be represented as a percentage of the recommended projected scores.

$$Item\ Coverage = \frac{N_{rec-list\_items}}{N_{items}} \quad (11)$$

4.2.3. Fairness metrics

Different fairness metrics serve different scopes of application. For consistent fairness, it is required that all individuals or groups should be treated similarly(Y. Wang et al., 2023). Thus, the associated metrics primarily measure the inconsistency of the utility distribution, such as Absolute Difference, Gini coefficient, and Variance.

Absolute Difference (AD) (Fu et al., 2020; Li et al., 2021; Zhu et al., 2018): is the absolute difference of the utility between the protected group G0 and the unprotected group G1. For user group, the recommendation utility f (G) is often defined as the average predicted rating or the average recommendation performance in the group G (Fu et al., 2020; Li et al., 2021). The lower the value, the fairer the recommendations.

$$AD = |f(G_0 - f(G_1))| \quad (12)$$

In this paper, we chose the NDCG to measure each group's utility.

$$Item\ Coverage = \frac{N_{rec-list\_items}}{N_{items}} \quad (13)$$

Here, $N_{rec-list\_items}$ means the number of products in the recommended list, and N_{items} means the number of all items.

Gini Coefficient: The Gini coefficient is often used to measure inequality in fields such as sociology (Fu et al., 2020; Mansoury et al., 2020). At the same time, there have been a number of studies on the fairness of recommendations that have used the Gini coefficient to measure individual fairness (Mansoury et al., 2020, 2022; Sun et al., 2019). The value of the Gini coefficient ranges from 0 to 1, with smaller values indicating greater fairness. The Gini coefficient can be calculated as (Sun et al., 2019):

$$Gini = \frac{\sum_{i=1}^{n}(2i-n-1)x_i}{n\sum_{i=1}^{n}x_i} \quad (14)$$

where all items are sorted in ascending order by the number of times they have been recommended, i denotes the position of the item after sorting, x_i represents the number of times item i has been recommended, and n represents the total number of items.

Variance: Variance is a commonly used indicator of dispersion (Lin et al., 2017; Rastegarpanah et al., 2019; Wu et al., 2021). The utilities assessed can be scoring prediction error, recommendation outcome performance, and exposure (Wu et al., 2021). The lower the value, the fairer the recommendation.

4.3. Experiment design

In this section, we show the general recommendation model and the experimental design and process of using the Large Language Model for recommendation.

We use the RecBole (Zhao et al., 2022) recommendation framework to implement general recommendation

models and GLM-4 (GLM et al., 2024) to implement LLM-based recommendation. RecBole is developed based on Python and PyTorch for reproducing and developing recommendation algorithms in a unified, comprehensive and efficient framework for research purposes. We directly implemented the 10 personalized recommendation models introduced in Sections 3.1-3.5 using RecBole.

We chose the 'gender' attribute of the user as a sensitive attribute and based on this we grouped the data and calculated the corresponding metrics. In order to be able to calculate the accuracy and fairness metrics for the different subgroups, the specific experimental procedure was as follows:

(1) Data filtering. To enable grouping by gender, we filtered the data of users with gender attributes. In addition, in order to reduce the amount of computation and to prevent the introduction of additional biases, we sampled the data as described in Section 4.1.

(2) Dividing training set, validation set, test set. In order to allow for the evaluation of the metrics, the filtered data was divided into training, validation and test sets in a ratio of 8:1:1.

(3) Model training and testing. The segmented dataset is first trained using RecBole's default parameters. The model parameters were then auto-tuned using the 'hyper_tuning' function provided by RecBole. For LLM-based recommendations, we first input the user's temporal interaction behavior with items (train set) into the model by organizing them into prompt words to facilitate the model's inference of user preferences. On this basis, we also input the titles of all items in the dataset into the model to limit the range of recommended items given by the model. The specific prompt words are described in Section 3.6.

(4) Evaluation: Recbole provides the evaluation of the accuracy metrics we need in Section 4.2.1. To calculate the metrics for the different gender groups, we output the evaluation results of all users on the test set and then calculate the accuracy metrics for the male and female groups separately. Then, by calculating the absolute value of the NDCG values for the different groups, the fairness metric AD can be obtained. Further, we calculated the values of variance, item coverage, Gini coefficient.

5. Results

In this section, we show the performanc and fairness results of ten classical recommendation models and one LLM-based recommendation in all the experiments on four datasets. In order to answer the four research questions posed in this paper, we first observe whether there is any unfairness by comparing the performance of users of different genders in three categories of metrics: accuracy, item coverage, and Gini coefficient. Then compare the differences in unfairness across models and data in terms of the model and data dimensions, respectively. Finally, the results are summarized and the main factors affecting fairness are analyzed in relation to the dataset characteristics.

5.1. Accuracy

5.1.1. Accuracy Values

First, we compared the accuracy of recommendation algorithms' recommendations for male and female users. Table 5 shows the performance of the 11 recommendation algorithms on the Movielens dataset on five accuracy metrics. As seen in Table 5, for each of the recommendtion methods, males outperform females in most metrics. This is consistent with the previous research findings (Ekstrand et al., 2018; Melchiorre et al., 2021), and we found this gender bias in more models and datasets. For the first research question, it is intuitively clear from the results that there is an unfairness in the recommendation results between male and female users.

Table 5: Accuracy Results in Movielens

Model	Gender	Recall@10	MRR@10	NDCG@10	Precision@10
ItemKNN	Male	**0.2385**	**0.4916**	**0.301**	**0.2215**
	FeMale	0.238	0.4553	0.2719	0.1875
UserKNN	Male	0.235	**0.5344**	**0.31**	**0.2201**
	FeMale	**0.2468**	0.4637	0.2753	0.1813
BPR	Male	**0.2331**	**0.5145**	**0.2988**	**0.241**
	FeMale	0.2283	0.4561	0.2686	0.1839
FM	Male	**0.2359**	**0.5201**	**0.3067**	**0.2222**
	FeMale	0.2338	0.4804	0.2775	0.1883
NeuMF	Male	**0.1912**	**0.4167**	**0.2398**	**0.181**
	FeMale	0.1815	0.3605	0.2078	0.1469
ConvNCF	Male	0.16	**0.3852**	**0.2037**	**0.1464**
	FeMale	**0.1616**	0.3252	0.1784	0.1205
CDAE	Male	0.2423	**0.5234**	**0.3079**	**0.3079**
	FeMale	**0.2433**	0.4974	0.2837	0.2837
MultVAE	Male	0.2424	0.5096	**0.3052**	**0.2204**

	FeMale	**0.255**	**0.5193**	0.3028	0.1927
LightGCN	Male	0.2373	**0.5399**	**0.3061**	**0.2151**
	FeMale	**0.2471**	0.4845	0.2907	0.196
NGCF	Male	**0.2121**	**0.477**	**0.2687**	**0.1915**
	FeMale	0.2096	0.4318	0.2453	0.1648

Then, we tabulated the results for the other three datasets, as shown in Figures 6-8.

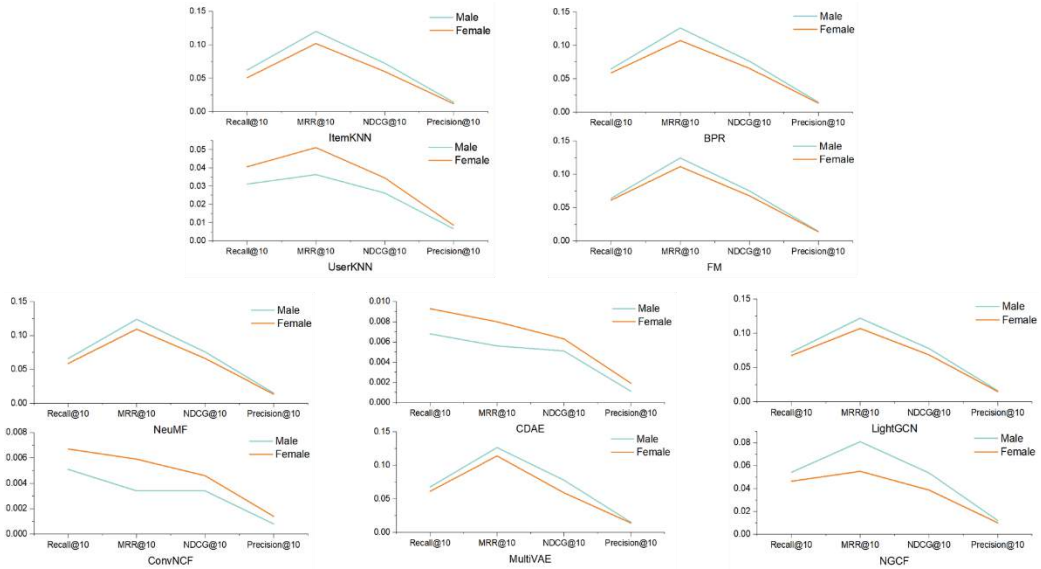

Figure 6: Accuracy Results in AliEC

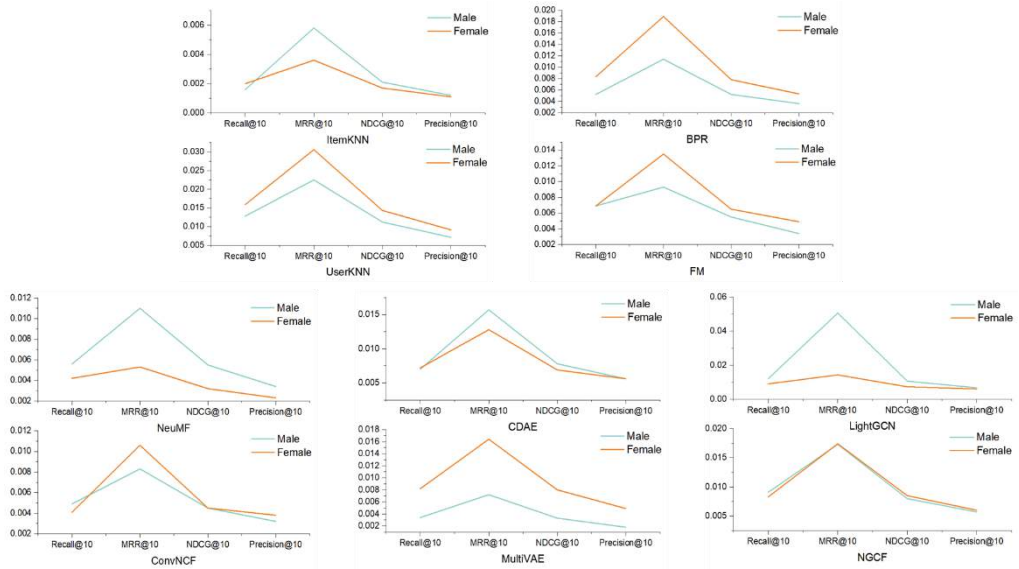

Figure 7: Accuracy Results in Last.fm

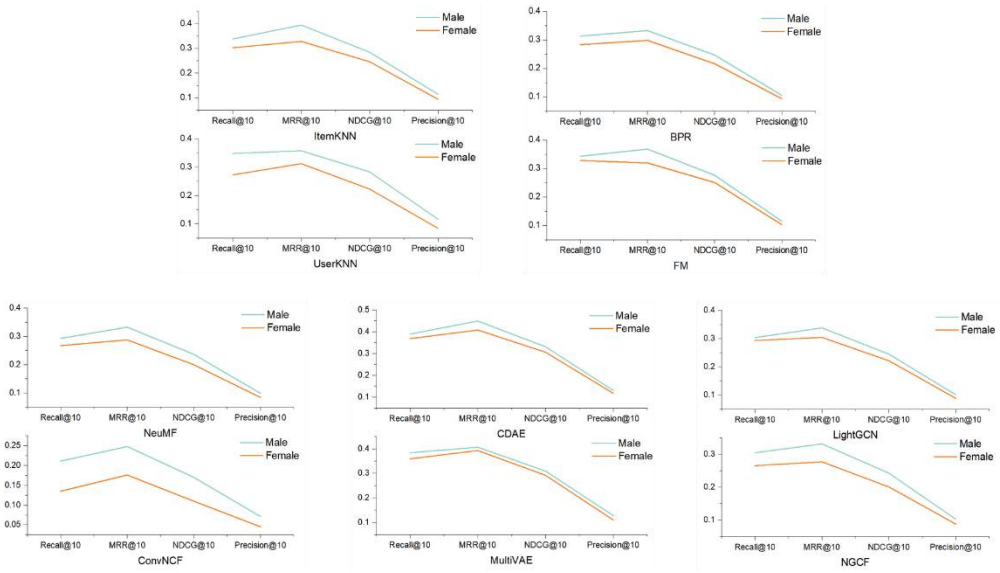

Figure 8: Accuracy Results in ZhihuRec

For the second research question, for different recommendation models, it can be seen that only three recommendation methods, UserKNN, ConvNCF and CDAE, have higher recommendation accuracy for female users than male users in the AliEC dataset. In the Last.fm dataset, more than half of the recommendation methods show better recommendation results for female users. In the ZhihuRec dataset, as with Movielens, male users outperform female users on all recommendation methods.

When analyzed in combination with the attributes of the datasets, the Movielens, Last.fm, and ZhihuRec datasets have significantly more male users and more interaction data than female users. The AliEC dataset has more female than male users. However, in the metric inter/num (i.e., how many interactions are recorded per user on average), the difference between males and females is not significant in the Last.fm and AliEC datasets, while the average number of interactions for male users is also significantly higher than that for female users in the Movielens and ZhihuRec datasets. Based on the above analysis, we can conclude that the merit of recommendation performance for users of different genders is not affected by the number of users but by the average number of user interactions. Resampling so that each group has the same number of interactions may eliminate the difference, which is the same conclusion as Ekstrand et al. (2018) reached.

5.1.2. Absolute Difference in NDCG

In this section, we analyze the absolute difference in NDCG metrics, i.e., we calculate the absolute value of the difference in NDCG metrics for male and female users. The results are shown in Figure 9.

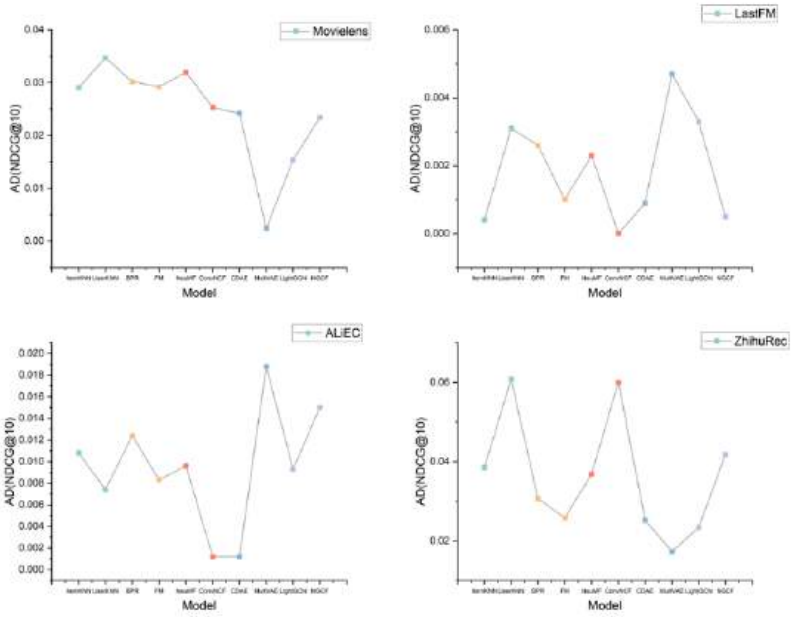

Figure 9: AD Results in NDCG@10

In Figure 9, dots of the same color indicate the same type of recommended methods. However, we do not seem to be able to find a very clear pattern from Figure 9. That is, there is no significant pattern in the absolute difference in recommendation performance between male and female users across different recommendation models.

For this result, we argue this outcome is explainable. First, the NDCG metrics used for the calculation are an overall result, and some differences and characteristics among users will be averaged out, reflecting only the overall recommendation performance of users of different genders. Further, directly calculating the difference in NDCG using absolute values only yields an averaged difference. Besides that, the four datasets we used have different attributes, and it can be seen from the results of 5.1.1 that the dataset characteristics have some influence on the results. This is one of the reasons why there is no clear pattern in the AD metrics.

In order to explore the differences in NDCG metrics among users of different genders, we further calculated the variance of NDCG values among male and female users. The variance can indicate the dispersion of user recommendation performance within a group, with lower values indicating greater fairness. The results on the four datasets are shown in Figure 10 and Table 6.

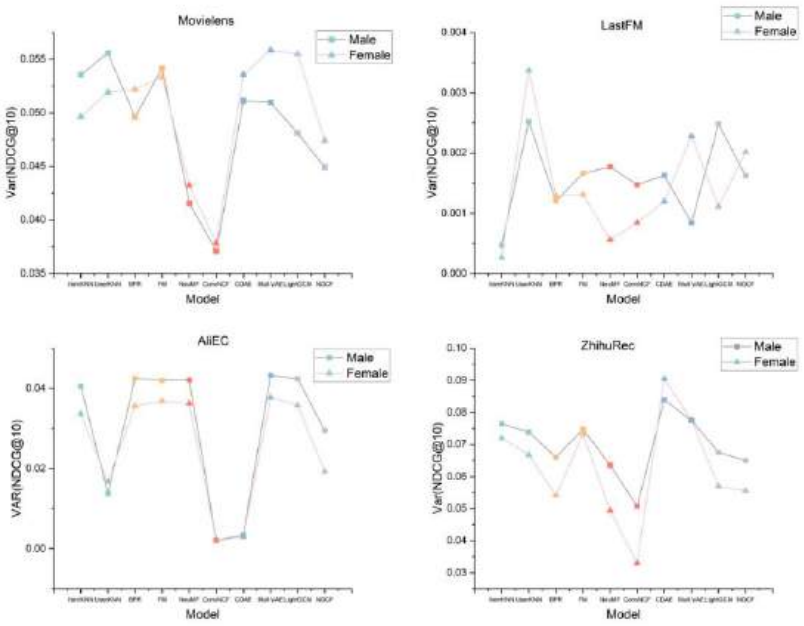

Figure 10: Variance Results by Gender

Table 6: Variance Results by Gender

Model	**Gender**	Variance			
		Movielens	Last.fm	AliEC	ZhihuRec
ItemKNN	Male	0.0526	0.0005	0.0405	0.0764
	FeMale	0.0496	0.0003	0.0336	0.0720
UserKNN	Male	0.0556	0.0025	0.0138	0.0739
	FeMale	0.0519	0.0034	0.0167	0.0668
BPR	Male	0.0496	0.0012	0.0425	0.0660
	FeMale	0.0522	0.0013	0.0356	0.0541
FM	Male	0.0542	0.0017	0.0420	0.0748
	FeMale	0.0533	0.0013	0.0368	0.0732
NeuMF	Male	0.0415	0.0018	0.0421	0.0635
	FeMale	0.0432	0.0006	0.0362	0.0494
ConvNCF	Male	0.0371	0.0015	0.0021	0.0507
	FeMale	0.0378	0.0008	0.0020	0.0330
CDAE	Male	0.0511	0.0016	0.0034	0.0840
	FeMale	0.0536	0.0012	0.0030	0.0904
MultVAE	Male	0.0510	0.0008	0.0432	0.0775
	FeMale	0.0559	0.0023	0.0377	0.0777
LightGCN	Male	0.0480	0.0025	0.0424	0.0676
	FeMale	0.0555	0.0011	0.0358	0.0570
NGCF	Male	0.0449	0.0016	0.0295	0.0650
	FeMale	0.0474	0.0020	0.0193	0.0555

As can be seen from Figure 10, the NDCG variance metrics for male users are not significantly better (lower values) than those for female users when compared to the recommendation performance, and even in more cases female users have smaller variance values. It can be seen that the fairness problem posed by the recommendation model in terms of discretization is smaller than that in terms of recommendation accuracy.

Meanwhile, the variances of the deep learning methods corresponding to the red dots (NeuMF and ConvNCF) are relatively small across the different datasets. That is, the NeuMF and ConvNCF methods are fairer in terms of dispersion. In contrast, the autoencoder methods represented by the blue dots (CDAE and Mult-VAE) and the matrix decomposition methods represented by the orange dots (BPR and FM) have relatively large variance. That is, in terms

of dispersion, these methods create a larger inequity problem. In order to better compare the variance changes between models, we also calculated the average of the variances of each model over the four datasets, as shown in Figure 11. The variances of the models are shown in the following figure. The results are the same as above.

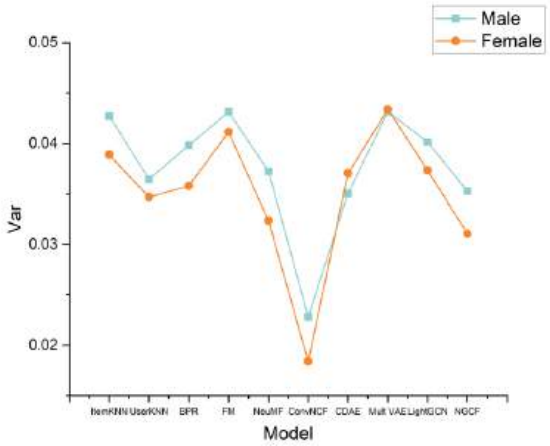

Figure 11: Average Variance Results by Gender

5.1.3. Results for LLM-based recommendations

Since recommendation using LLM requires the input of content information such as the name of the item. Also, considering the limitation of computing power, we choose Movielens data to realize LLM-based recommendation.

Table 7 shows the accuracy of the recommendation results for users of different genders obtained in scenario 1. It can be seen that, consistent with other recommendation methods, female users have lower recommendation performance than male users. It shows that in LLM-based recommendation, there is also the problem of unfairness in the recommendation results for users of different genders.

Table 7: Accuracy Results for Scenario 1 Recommendations Based on Large Models

Model	Gender	Recall@10	MRR@10	NDCG@10	Precision@10
LLM	Male	**0.0082**	**0.0259**	**0.0107**	**0.0078**
	FeMale	0.0059	0.0182	0.0073	0.0051

Then, we analyze the AD values of LLM-based recommendation system, and the relevant results are shown in Table 8.

Table 8: AD(NDCG) Results for Scenario 1 Recommendations Based on Large Models

Model	ItemKNN	User KNN	BPR	FM	NeuMF	Conv NCF	CDAE	Mult VAE	Light GCN	NGCF	LLM
AD	0.0291	0.0347	0.0302	0.0292	0.032	0.0253	0.0242	**0.0024**	0.0154	0.0234	**0.0034**

As can be seen from the table, for the Movielens dataset, LLM has a relatively low AD value. However, since the recommendation performance of LLM is also smaller compared to other methods, it is not straightforward to conclude that the recommendation results of LLM are fairer in terms of the AD of NDCG.

5.2. Item Coverage

5.2.1. Item Coverage Values

In this section, we analyze the fairness of recommendation results for users of different genders from the perspective of recommendation product diversity. Item Coverage calculates the percentage of total products in a user's recommendation list, and can indicate the diversity of recommendations. First, we counted the Item Coverage values obtained by different recommendation methods for different datasets, as shown in Figure 12 and Table 9.

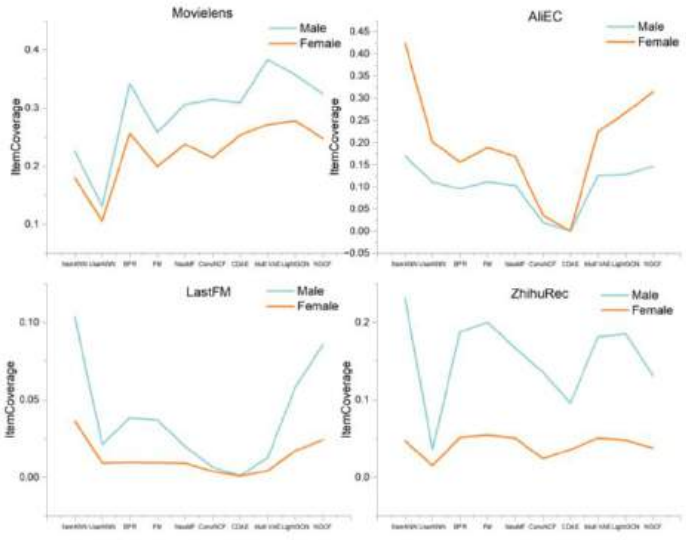

Figure 12: Item Coverage Results by Gender

Table 9: Item Coverage Results by Gender

Model	**Gender**	Item Coverage			
		Movielens	Last.fm	AliEC	ZhihuRec
ItemKNN	Male	0.2260	0.1037	0.1698	0.2310
	FeMale	0.1801	0.0364	0.4247	0.0474
UserKNN	Male	0.1308	0.0213	0.1110	0.0360
	FeMale	0.1052	0.0094	0.2011	0.0152
BPR	Male	0.3419	0.0385	0.0961	0.1879
	FeMale	0.2562	0.0097	0.1560	0.0513
FM	Male	0.2574	0.0371	0.1115	0.1999
	FeMale	0.1998	0.0096	0.1886	0.0547
NeuMF	Male	0.3062	0.0200	0.1032	0.1667
	FeMale	0.2378	0.0092	0.1690	0.0505
ConvNCF	Male	0.3145	0.0066	0.0203	0.1355
	FeMale	0.2146	0.0040	0.0364	0.0243
CDAE	Male	0.3092	0.0012	0.0002	0.0959
	FeMale	0.2533	0.0009	0.0002	0.0358
MultVAE	Male	0.3835	0.0127	0.1258	0.1811
	FeMale	0.2717	0.0044	0.2243	0.0506
LightGCN	Male	0.3567	0.0583	0.1276	0.1849
	FeMale	0.2782	0.0172	0.2677	0.0475
NGCF	Male	0.3240	0.0856	0.1471	0.1309
	FeMale	0.2473	0.0241	0.3142	0.0378

As can be seen from the figure, Item Coverage is higher for male users than for female users in all recommendation models on all datasets except for the AliEC dataset. The first conclusion that can be drawn is that there is also inequity in the recommendation methods in terms of diversity. Further combining the dataset attributes, the AliEC dataset has a surplus of female users over male users. Furthermore, according to the calculation method of Item Coverage, when there are more users, the range of recommended results is generally wider, so we infer that the number of users affects Item Coverage.

5.2.2. Absolute Difference in Item Coverage

We also calculated the absolute difference in Item Coverage for users of different genders. The results are shown in Figure 13.

As can be seen in the figure, the ItemKNN method, as well as the graph neural network methods represented by

the purple dots (Light GCN and NGCF), have higher AD values, i.e., these methods bring more unfairness in terms of Item Coverage. The CDAE and UserKNN methods have smaller AD values in all four datasets, i.e., they are relatively fairer in terms of Item Coverage.

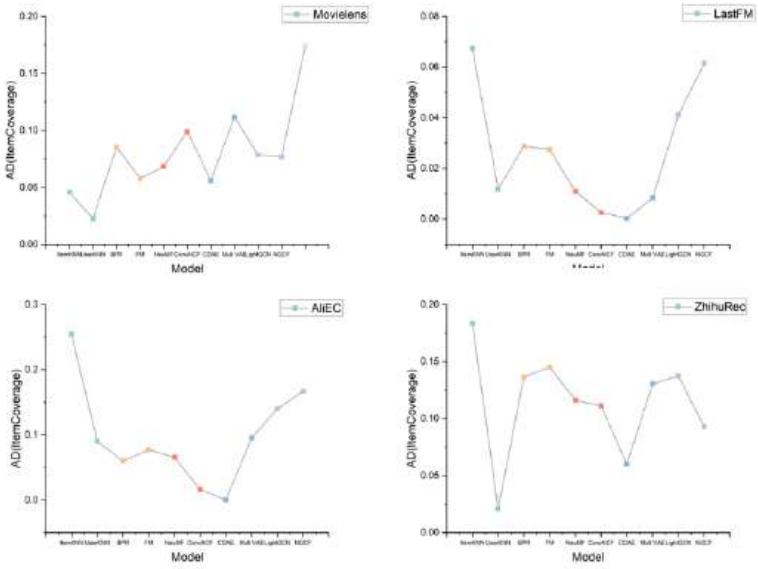

Figure 13: AD Results in Item Coverage

To better compare the fairness in terms of Item Coverage between the different models, we also calculated the average of the AD values for each model on the four datasets, as shown in Figure 14. Results are the same as above.

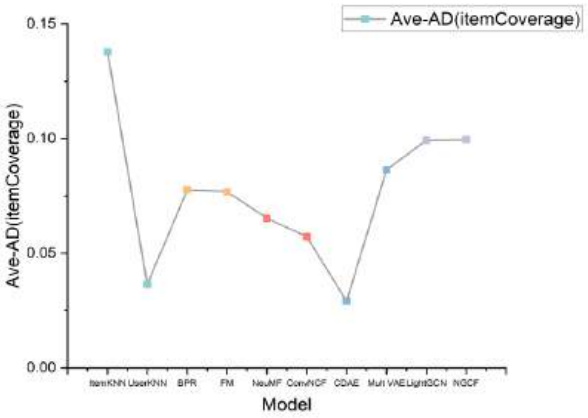

Figure 14: Average AD Results in Item Coverage

5.2.3. Item Coverage Values for LLM-based recommendations

The results of the three LLM-based recommendations for Item Coverage are shown in Table 10.

Table 10: Item Coverage Results for LLM Scenarios

Scenario	Item Coverage
S1-Male	0.5143
S1-Female	0.3401
S2	0.4727
S3	0.4893

In Scenario 1, even though neither the prompt nor the input data included any information related to the user's

gender, the Item Coverage of the recommended lists still showed a significant disparity between genders (0.5143 for males and 0.3401 for females). In contrast, the difference between the results of Scenarios 2 and 3 is much smaller. This suggests that explicitly including the user's gender in the input prompt has minimal impact on the Item Coverage of the recommendation results.

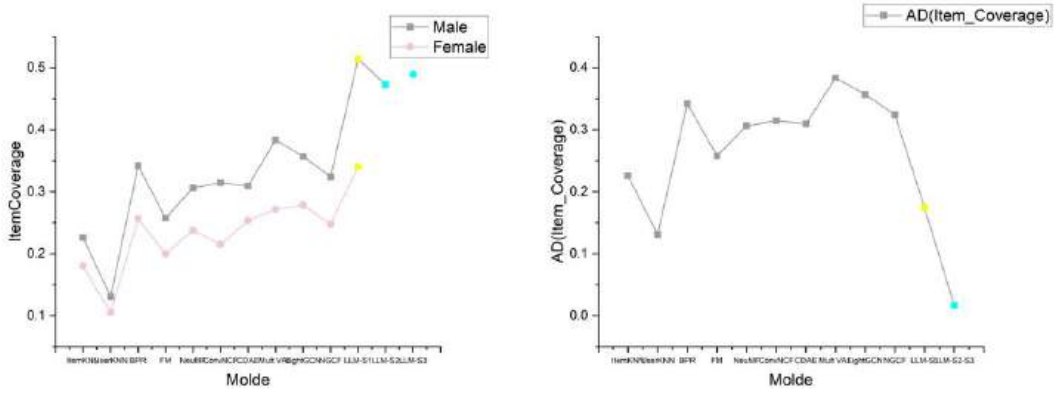

Figure 15: Item Coverage and AD Results in Item Coverage for LLM Scenarios

The values of Item Coverage and the absolute difference values of Item Coverage are shown in Figure 15. In general, the LLM-based recommendation model acheived a somewhat higher Item Coverage. At the same time, the absolute difference in Item Coverage was smaller for LLM-based recommendation results than for other models. Therefore, overall, LLM-based recommendation methods are relatively fairer in terms of Item Coverage.

5.3. Gini Coefficient

Gini coefficient measures the unevenness of distribution and is calculated based on the ranking of items in the recommendation list.

5.3.1. Overall Gini Coefficient

Firstly, we compare the difference of the overall Gini coefficient of different recommendation methods and calculate the average Gini coefficient value of different recommendation methods on the four datasets, and the statistical results are shown in Figure 16.

As can be seen from the Figure 16, UserKNN, ConvNCF and CDAE methods have the highest Gini coefficients, ItemKNN methods have the lowest Gini coefficients, and GNN-based methods (Light GCN and NGCF) are relatively low. That is, the unfairness problem of the overall recommendation order of products caused by UserKNN, ConvNCF and CDAE methods is greater, and the unfairness caused by ItemKNN method and GNN-based methods is relatively small.

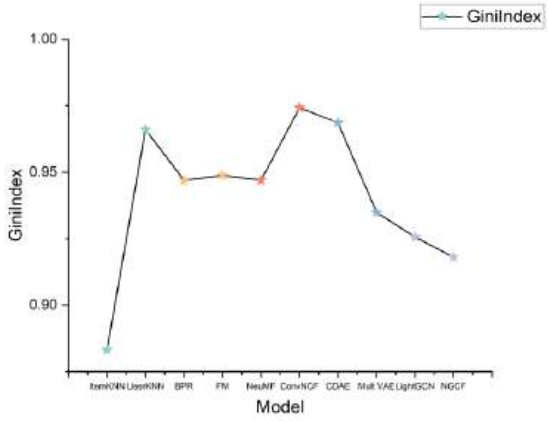

Figure 16: Average Gini Coefficient

5.3.2. Gini coefficient in gender

We calculated the average of the Gini coefficient for users of different genders across the four datasets. As can be seen from Figure 17, in general, female users have a higher Gini coefficient, indicating greater inequality.

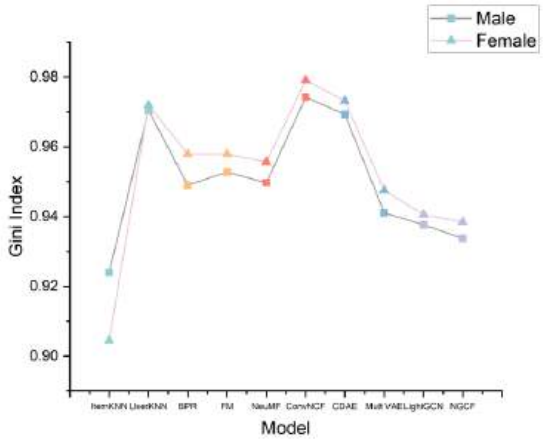

Figure 17: Average Gini Coefficient by Gender

We also calculated the Gini coefficients for different genders for all results, and organized the results as shown in Figure 18 and Table 11. This was used to compare the differences in fairness in the order of recommended products among users of different genders.

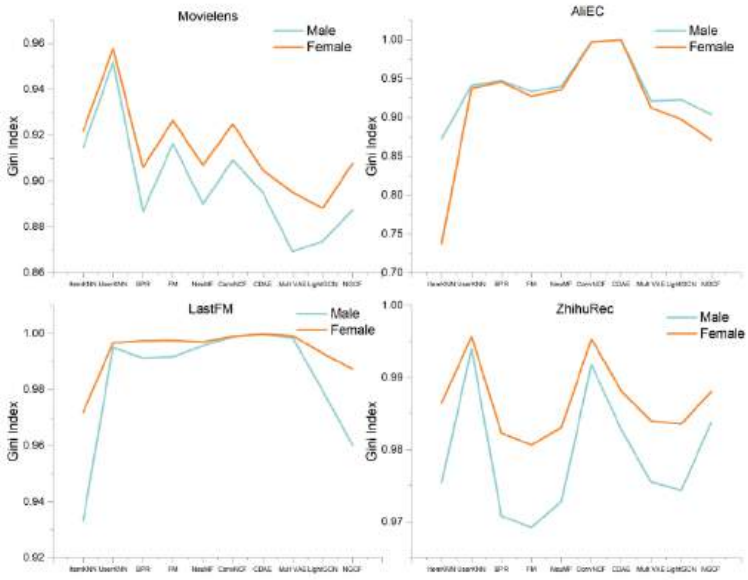

Figure 18: Gini Coefficient by Gender

Table 11: Gini Coefficient Results by Gender

Model	Gender	Gini Coefficient			
		Movielens	Last.fm	AliEC	ZhihuRec
ItemKNN	Male	0.9146	0.9333	0.8726	0.9754
	FeMale	0.9218	0.9720	0.7375	0.9865
UserKNN	Male	0.9516	0.9950	0.9413	0.9940
	FeMale	0.9576	0.9966	0.9377	0.9957
BPR	Male	0.8866	0.9912	0.9474	0.9708
	FeMale	0.9059	0.9974	0.9461	0.8923
FM	Male	0.9163	0.9917	0.9336	0.9692
	FeMale	0.9265	0.9975	0.9272	0.9806
NeuMF	Male	0.8899	0.9956	0.9405	0.9728
	FeMale	0.9069	0.9969	0.9357	0.9831
ConvNCF	Male	0.9092	0.9989	0.9972	0.9918
	FeMale	0.9248	0.9990	0.9973	0.9953
CDAE	Male	0.8949	0.9998	0.9998	0.9828
	FeMale	0.9048	0.9998	0.9998	0.9882
MultVAE	Male	0.8693	0.9984	0.9210	0.9755
	FeMale	0.8951	0.9991	0.9120	0.9840
LightGCN	Male	0.8737	0.9796	0.9229	0.9744
	FeMale	0.8881	0.9929	0.8978	0.9836
NGCF	Male	0.8874	0.9601	0.9038	0.9838
	FeMale	0.9077	0.9872	0.8710	0.9880

The figure shows that the Gini coefficient is slightly higher for female users on nearly all datasets except for AliEC, reflecting greater unfairness. In AliEC, although female users greatly outnumber male users, the gender bias in the Gini coefficient does not differ much except for ItemKNN, Light GCN, and NGCF. It can be concluded that inequity persists in the Gini coefficient aspect for users of different genders.

5.3.3. Gini coefficient for LLM-based recommendations

The results of the three LLM-based recommendations for Gini coefficient are shown in Table 12.

Table 12: Gini Coefficient Results for LLM Scenarios

Scenario	Gini coefficient
S1-Male	0.8362
S1-Female	0.8827
S2	0.8438
S3	0.8052

From the results, it can be seen that in Scenario 1, the Gini coefficient for female users is higher than that of male users, indicating a more unfair recommendation outcome. Whereas in scenarios 2 and 3, the Gini coefficient of female users is less than that of male users.

When compared to other recommendation methods (as shown in Figure 19), LLM-based recommendations consistently achieve a lower Gini coefficient, indicating a fairer distribution of recommendations overall.

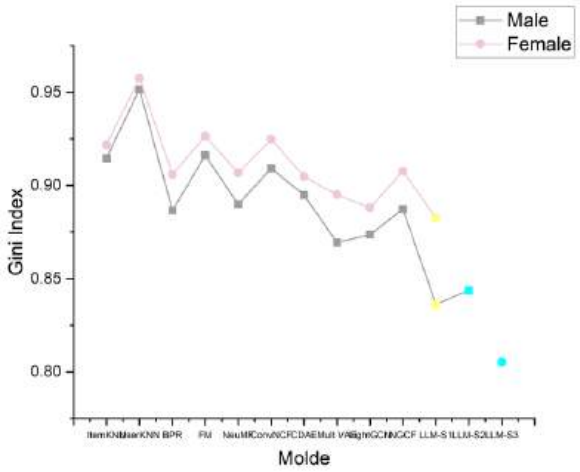

Figure 19: Gini coefficient Results for LLM Scenarios

6. Conclusions and Future Work

6.1. Conclusions

In this work, we investigated user gender fairness of recommendation methods. We analyzed the fairness of recommendation results in terms of accuracy and diversity by comparing the results of 10 general and one LLM-based recommendation method on four datasets. Additionally, we set up three recommendation scenarios based on a large language model to analyze the fairness issue in different contexts. The key findings are as follows:

First, regarding recommendation accuracy, our experiment revealed significant gender disparities (including LLM-based methods). Across most datasets, male users outperform female users in terms of recommendation results, although there are more female users than male users in the AliEC dataset. An exception is observed in the Last.fm dataset, where nearly half of the models were more accurate for female users than for male users. By analyzing the statistical properties of the dataset, we conclude that this gender bias is not related to the number of users, but may be related to the average number of user interactions. This showed that the recommendation method can introduce unfairness in terms of accuracy, a finding consistent with perior work.

We also analyzed the absolute difference (AD) in recommendation accuracy for male and female users, using NDCG as the metric. The results showed that AD failed to reveal clear patterns. We attribute this to the fact that AD is an aggregate measure of overall recommendation performance, in which inter-individual differences are ignored in the calculation process. Therefore, we further analyzed the NDCG variance across genders and found that the neural collaborative filtering-based methods (NeuMF and ConvNCF) exhibited the smallest variance, while the MF-based and autoencoder-based methods exhibited larger variance.

Then, we evaluated diversity fairness using Item Coverage. The experiments revealed significant disparities: in the dataset with more male users, males enjoyed higher Item Coverage, indicating that unfairness in recommendation diversity may correlate with user group size. We also calculated the absolute difference in Item Coverage. The results showed that ItemKNN and GNN-based methods (LightGCN and NGCF) produced the largest gaps, whereas UserKNN and CDAE yielded the smallest.

For Gini coefficient, in general, ItemKNN and GNN-based recommendation methods yielded lower values, indicating greater fairness, whereas UserKNN, ConvNCF and CDAE methods produced higher values. This result is similar to the overall result of Item Coverage. Since both metrics are calculated from the product distribution in users' recommendation lists, the results are expected. Additionally, we also observed pronounced gender bias in this respect: the Gini coefficient for female users was usually higher than that of male users, indicating greater inequality. In the AliEC dataset, despite the large imbalanc in user group size (5,209 female vs. 1,339 male users), the difference in Gini coefficient between the two groups was small, confirming that gender unfairness persists in terms of the Gini coefficient of the recommendation lists.

Finally, we summarized the findings from the three LLM-based recommendation scenarios. In Scenario 1, we calculated the recommendation accuracy metrics for users and found that male users had significantly higher accuracy than female users, highlighting the inequity problem. For fairness regarding diversity, the Item Coverage difference between male and female users was larger in Scenario 1 and much smaller in Scenarios 2 and 3. This suggests that using LLM for recommendation introduces some unfairness in terms of diversity, but setting the user's gender directly when using LLM for recommendation does not have much effect on diversity. It can be inferred that LLM itself does

not make unfair recommendations for male and female users, but the differences in the data between users of different genders can still lead to unfairness issues. In terms of the Gini coefficient of the recommendation list, the difference between the Gini coefficient of users of different genders in Scenario 1, as well as between Scenario 2 and Scenario 3 users, is more obvious. This is inconsistent with the results of the Item Coverage results. In addition, the overall Item Coverage and Gini coefficient in the LLM-based recommendation results were better than other recommendation models. Moreover, the overall Item Coverage and Gini coefficient achieved by LLM-based recommendations outperformed those of other recommendation models, demonstrating their superior ability to balance diversity and fairness at the aggregate level.

6.2. Future Work

In this paper, we conducted LLM-based recommendation experiments using a single dataset due to dataset limitations. In future work, additional datasets can be utilized to enhance the robustness of the results. While this paper examined unfairness with respect to user gender, other attributes such as age and geography can be further studied for fairness evaluation. Building on insights in this study, we can further investigate the potential factors such as data characteristics or algorithm design that may influence recommendation fairness, so as to better mitigate unfairness in recommendation systems.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (72271084, 72342011, 72271083, 72101076, and 72101072) and the Fundamental Research Funds for the Central Universities of China (JZ2023YQTD0075).

REFERENCES

Abdollahpouri, Himan & Burke, Robin. (2019). Multi-stakeholder Recommendation and its Connection to Multi-sided Fairness [Preprint]. 10.48550/arXiv.1907.13158.

Aguila et al. (2023). Multi-view-AE: A Python package for multi-view autoencoder models. *Journal of Open Source Software*, 8(85), 5093. https://doi.org/10.21105/joss.05093

Amigó, E., Deldjoo, Y., Mizzaro, S., & Bellogín, A. (2023). A unifying and general account of fairness measurement in recommender systems. I*nformation Processing & Management*, 60(1), 103115.

Anelli, V. W., Deldjoo, Y., Di Noia, T., Malitesta, D., Paparella, V., & Pomo, C. (2023). Auditing consumer- and producer-fairness in graph collaborative filtering. In J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, & A. Caputo (Eds.), ECIR (Vol. 13980, pp. 33 – 48). *Springer Nature Switzerland*. https://doi.org/10.1007/978-3-031-28244-7_3

Bell, R. M., Koren, Y., & Volinsky, C. (2007). The BellKor solution to the Netflix Prize. *Technical report, AT&T Labs Research*. http://www.netflixprize.com/assets/

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., & Goodrow, C. (2019). Fairness in Recommendation Ranking through Pairwise Comparisons. *In proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2212 – 2220. https://doi.org/10.1145/3292500.3330745

Biega, A. J., Gummadi, K. P., & Weikum, G. (2018). Equity of Attention: Amortizing Individual Fairness in Rankings. *In Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 405–414). ACM. https://doi.org/10.1145/3209978.3210063*

Boratto, L., Fenu, G., Marras, M., & Medda, G. (2022). Consumer Fairness in Recommender Systems: Contextualizing Definitions and Mitigations. *In Lecture Notes in Computer Science (Vol. 13185, pp. 552–566). Springer. https://doi.org/10.1007/978-3-030-99736-6_37*

Boratto, L., Fenu, G., Marras, M., & Medda, G. (2023). Practical perspectives of consumer fairness in recommendation. *IPM*, 60(2), 103208. https://doi.org/10.1016/j.ipm.2022.103208

Teodorescu, M., Morse, L., Awwad, Y., Kane, G., & et al. (2021). Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation. *MIS Quarterly*, *45*(3), 1483–1500. https://doi.org/10.25300/MISQ/2021/16535

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering. 43-52. https://arxiv.org/abs/1301.7363

Burke, R., Sonboli, N., Mansoury, M., & Ordonez-Gauger, A. (2017). Balanced neighborhoods for fairness-aware collaborative recommendation. *In Proceedings of the FATREC Workshop*. https://doi.org/10.18122/B2GQ53

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2020). Bias and Debias in Recommender System: A Survey and Future Directions. *IEEE Transactions on Knowledge and Data Engineering*, 20.

Current, S., He, Y., Gurukar, S., & Parthasarathy, S. (2022). FairMod: Fair Link Prediction and Recommendation via Graph Modification. Equity and Access in Algorithms, Mechanisms, and Optimization, (pp. 1–14). https://doi.org/10.1145/3551624.3555287

Deldjoo, Y., & di Noia, T. (2024). CFaiRLLM: Consumer Fairness Evaluation in Large-Language Model Recommender System (No. arXiv:2403.05668). arXiv. http://arxiv.org/abs/2403.05668

Deldjoo, Y., & Nazary, F. (2024). FairEvalLLM : A Normative Framework for Benchmarking Consumer Fairness in Large Language Model Recommender System [Preprint]. arXiv. http://arxiv.org/abs/2405.02219

Deldjoo, Y., Anelli, V. W., Zamani, H., Bellogin, A., & Noia, T. D. (2019). Recommender Systems Fairness Evaluation via Generalized Cross Entropy (No. arXiv:1908.06708). arXiv. http://arxiv.org/abs/1908.06708

Deldjoo, Y., Anelli, V. W., Zamani, H., Bellogín, A., & Di Noia, T. (2021). A flexible framework for evaluating user and item fairness in recommender systems. *User Modeling and User-Adapted Interaction,* 31(3), 457–511. https://doi.org/10.1007/s11257-020-09285-1

Deldjoo, Y., Bellogin, A., & Di Noia, T. (2021). Explaining recommender systems fairness and accuracy through the lens of data characteristics. *Information Processing & Management*, 58(5), 102662. https://doi.org/10.1016/j.ipm.2021.102662

Deldjoo, Y., Jannach, D., Bellogin, A., Difonzo, A., & Zanzonelli, D. (2024). Fairness in recommender systems: Research landscape and future directions. *User Modeling and User-Adapted Interaction,* 34(1), 59–108. https://doi.org/10.1007/s11257-023-09364-z

Deshpande, M., & Karypis, G. (2004). Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems*, 22(1), 143–177. https://doi.org/10.1145/963770.963776

Dinnissen, K. (2024). Fairness and transparency in music recommender systems: Improvements for artists. *In Proceedings of 18th ACM Conference on Recommender Systems*, 1368–1375. https://doi.org/10.1145/3640457.3688024

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214-226). ACM.

Ekstrand, M. D., & Kluver, D. (2021). Exploring author gender in book rating and recommendation, *User Modeling and User-Adapted Interaction,*31(3), 377–420. https://doi.org/10.1007/s11257-020-09284-2

Ekstrand, M. D., Tian, M., Kazi, M. R. I., Mehrpouyan, H., & Kluver, D. (2018). Exploring author gender in book rating and recommendation. *RecSys*, 242–250. https://doi.org/10.1145/3240323.3240373

Fan, W., Zhao, X., Chen, X., Su, J., Gao, J., Wang, L., Liu, Q., Wang, Y., Xu, H., Chen, L., & Li, Q. (2022). A Comprehensive Survey on Trustworthy Recommender Systems. arXiv. http://arxiv.org/abs/2209.10117

Farnadi, G., Kouki, P., Thompson, S. K., Srinivasan, S., & Getoor, L. (2018). A Fairness-aware Hybrid Recommender System. *FATRec*. http://arxiv.org/abs/1809.09030

Ferraro, A. (2019). Music cold-start and long-tail recommendation: Bias in deep representations. *In Proceedings of the 13th ACM Conference on Recommender Systems*, 586–590. https://doi.org/10.1145/3298689.3347052

Fu, R., Aseri, M., Singh, P. V., & Srinivasan, K. (2022). "Un" Fair Machine Learning Algorithms. *Management Science*, 68(6), 4173–4195. https://doi.org/10.1287/mnsc.2021.4065

Fu, R., Huang, Y., & Singh, P. V. (2021). Crowds, Lending, Machine, and Bias. *Information Systems Research*, *32*(1), 72–92. https://doi.org/10.1287/isre.2020.0990

Fu, Z., Xian, Y., Gao, R., Zhao, J., Huang, Q., Ge, Y., Xu, S., Geng, S., Shah, C., Zhang, Y., & de Melo, G. (2020). Fairness-Aware Explainable Recommendation over Knowledge Graphs. *SIGIR*. http://arxiv.org/abs/2006.02046

Ge, M., Delgado-Battenfeld, C., & Jannach, D. (2010). Beyond accuracy: Evaluating recommender systems by coverage and serendipity. *RecSys ' 10 -In proceedings of the 4th ACM Conference on Recommender Systems*, 257–260. https://doi.org/10.1145/1864708.1864761

Ge, Y., Liu, S., Gao, R., Xian, Y., Li, Y., Zhao, X., Pei, C., Sun, F., Ge, J., Ou, W., & Zhang, Y. (2021). Towards Long-term Fairness in Recommendation. *In poceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM)*, 445–453. https://doi.org/10.1145/3437963.3441824

Ge, Y., Tan, J., Zhu, Y., Xia, Y., Luo, J., Liu, S., Fu, Z., Geng, S., Li, Z., & Zhang, Y. (2022). Explainable Fairness in Recommendation. *In proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 681–691. https://doi.org/10.1145/3477495.3531973

Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. *In proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2221–2231. https://doi.org/10.1145/3292500.3330691

GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas, D., Feng, G., Zhao, H., Lai, H., Yu, H., Wang, H., Sun, J., Zhang, J., Cheng, J., Gui, J., Tang, J., … Wang, Z. (2024). ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. 1 – 19. https://doi.org/https://doi.org/10.48550/arXiv.2406.12793

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12), 61-70. https://dl.acm.org/doi/pdf/10.1145/138859.138867

Greenwood, S., Chiniah, S., & Garg, N. (2024). User-item fairness tradeoffs in recommendations. *NIPS*.

Guo, H., Li, J., Wang, J., Liu, X., Wang, D., Hu, Z., Zhang, R., & Xue, H. (2023). FairRec: Fairness Testing for Deep Recommender Systems. *ISSTA*, 310 – 321. https://doi.org/10.1145/3597926.3598058

Guo, P., & Xiao, K. (2024). From efficiency to equity: A multi-user paradigm in mobile route optimization. *Electronic Commerce Research and Applications*, *68*, 101459. https://doi.org/10.1016/j.elerap.2024.101459

Hao, B., Zhang, M., Ma, W., Shi, S., Yu, X., Shan, H., Liu, Y., & Ma, S. (2021). A Large-Scale Rich Context Query and Recommendation Dataset in Online Knowledge-Sharing. 3. https://doi.org/https://doi.org/10.48550/arXiv.2106.06467

Han, Z., Chen, C., Zheng, X., Li, M., Liu, W., Yao, B., Li, Y., & Yin, J. (2024). Intra- and Inter-group Optimal Transport for User-Oriented Fairness in Recommender Systems. *AAAI*, *38*, 8463–8471. https://doi.org/10.1609/aaai.v38i8.28689

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & Wang, M. (2020, July 7). LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. *In proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, http://arxiv.org/abs/2002.02126

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T.-S. (2017). Neural collaborative filtering. *In proceedings of the 26th International Conference on World Wide Web*, 173 – 182. https://doi.org/http://dx.doi.org/10.1145/3038912.3052569

He, X., N, Du, X., Wang, X., Tian, F., Tang, J., & Chua, T. (2016). Outer Product-based Neural Collaborative Filtering. *In proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2227 – 2233. https://doi.org/https://doi.org/10.48550/arXiv.1808.03912

Islam, R., Keya, K. N., Zeng, Z., Pan, S., & Foulds, J. (2021). Debiasing Career Recommendations with Neural Fair Collaborative Filtering. *WWW*, 3779 – 3790. https://doi.org/10.1145/3442381.3449904

Jiang, M., Bao, K., Zhang, J., Wang, W., Yang, Z., Feng, F., & He, X. (2024, May). Item-side fairness of large language model-based recommendation system. *In Proceedings of the ACM Web Conference 2024* (pp. 4717-4726).

Jin, D., Wang, L., Zhang, H., Zheng, Y., Ding, W., Xia, F., & Pan, S. (2023). A survey on fairness-aware recommender systems. *Information Fusion*, 100, 101906. https://doi.org/10.1016/j.inffus.2023.101906

Kamishima, T., & Akaho, S. (2017). Considerations on Recommendation Independence for a Find-Good-Items Task. *RecSys*. https://doi.org/10.18122/B2871W

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2018). Recommendation Independence. FAT.

Kim, K., & Ahn, H. (2011). Collaborative Filtering with a User-Item Matrix Reduction Technique. *International Journal of Electronic Commerce*, *16*(1), 107–128. https://doi.org/10.2753/JEC1086-4415160104

Lambrecht, A., & Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science*, 65(7), 2966 – 2981. https://doi.org/10.1287/mnsc.2018.3093

Leal, F., Veloso, B., Malheiro, B., González-Vélez, H., & Carlos Burguillo, J. (2020). A 2020 perspective on "Scalable modelling and recommendation using wiki-based crowdsourced repositories:" Fairness, scalability, and real-time recommendation. *Electronic Commerce Research and Applications*, *40*, 100951. https://doi.org/10.1016/j.elerap.2020.100951

Lee, E. L., Lou, J. K., Chen, W. M., Chen, Y. C., Lin, S. D., Chiang, Y. S., & Chen, K. T. (2014, August). Fairness-aware loan recommendation for microfinance services. *In proceedings of the 2014 international conference on social computing* (pp. 1-4).

Leonhardt, J., Anand, A., & Khosla, M. (2018). User Fairness in Recommender Systems. WWW, 101 – 102. https://doi.org/10.1145/3184558.3186949

Li, R. Z., Urbano, J., & Hanjalic, A. (2021). Leave No User Behind: Towards Improving the Utility of Recommender Systems for Non-mainstream Users. *WSDM*, 103 – 111. https://doi.org/10.1145/3437963.3441769

Li, Y., Chen, H., Fu, Z., Ge, Y., & Zhang, Y. (2021). User-oriented Fairness in Recommendation. *WWW*, 624 – 632. https://doi.org/10.1145/3442381.3449866

Li, Y., Chen, H., Xu, S., Ge, Y., & Zhang, Y. (2021). Towards Personalized Fairness based on Causal Notion. *In proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*,, 1054 – 1063. https://doi.org/10.1145/3404835.3462966

Li, Y., Chen, H., Xu, S., Ge, Y., Tan, J., Liu, S., & Zhang, Y. (2023). Fairness in Recommendation: Foundations, Methods and Applications. *TIST (Q3)*. http://arxiv.org/abs/2205.13619

Liang, D., Krishnan, R. G., Hoffman, M. D., & Tony. (2018). Variational Autoencoders for Collaborative Filtering. *International World Wide Web Conference Committee*, 689 – 698. https://doi.org/https://doi.org/10.1145/3178876.3186150

Lin, L. F. (2024). Social Referral Mechanism For Context-aware Mobile Advertising. *Journal of Electronic Commerce Research*, 25(2).

Lin, X., Zhang, M., Zhang, Y., Gu, Z., Liu, Y., & Ma, S. (2017). Fairness-aware group recommendation with pareto-efficiency. *In proceedings of the 11th ACM Conference on Recommender Systems*, 107 – 115. https://doi.org/10.1145/3109859.3109887

Linden, G., Smith, B., & York, J. (2003). Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, 7(1), 76-80.

Liu, W., Guo, J., Sonboli, N., Burke, R., & Zhang, S. (2019). Personalized fairness-aware re-ranking for microlending. *In proceedings of the 13th ACM Conference on Recommender Systems*, 467 – 471. https://doi.org/10.1145/3298689.3347016

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2020). FairMatch: A Graph-based Approach for Improving Aggregate Diversity in Recommender Systems. *In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 154 – 162. https://doi.org/10.1145/3340631.3394860

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2022). A Graph-Based Approach for Mitigating Multi-Sided Exposure Bias in Recommender Systems. *ACM Transactions on Information Systems*, 40(2). https://doi.org/10.1145/3470948

Mansoury, M., Mobasher, B., Burke, R., & Pechenizkiy, M. (2019, August 2). Bias Disparity in Collaborative Recommendation: Algorithmic Evaluation and Comparison. *RecSys*. http://arxiv.org/abs/1908.00831

Masrour, F., Wilson, T., Yan, H., Tan, P.-N., & Esfahanian, A. (2020). Bursting the Filter Bubble: Fairness-Aware Network Link Prediction. *AAAI*, 34(01), 841 – 848. https://doi.org/10.1609/aaai.v34i01.5429

Melchiorre, A. B., Rekabsaz, N., Parada-Cabaleiro, E., Brandl, S., Lesota, O., & Schedl, M. (2021). Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management*, 58(5), 102666. https://doi.org/10.1016/j.ipm.2021.102666

Moon, H. S., Ryu, Y. U., & Kim, J. K. (2019). Enhanced collaborative filtering: A product life cycle approach. *Journal of Electronic Commerce Research*, 20(3), 155-168.

Misztal-Radecka, J., & Indurkhya, B. (2021). Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems. *Information Processing & Management*, 58(3), 102519. https://doi.org/10.1016/j.ipm.2021.102519

Naghiaei, M., Rahmani, H. A., & Deldjoo, Y. (2022, April 17). CPFair: Personalized Consumer and Producer Fairness Re-ranking for Recommender Systems. *In proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, http://arxiv.org/abs/2204.08085

Ochmann, J., Michels, L., Tiefenbeck, V., Maier, C., & Laumer, S. (2024). Perceived algorithmic fairness: An empirical study of transparency and anthropomorphism in algorithmic recruiting. *Information Systems Journal*, *34*(2), 384-414.

Patro, G. K., Biswas, A., Ganguly, N., Gummadi, K. P., & Chakraborty, A. (2020). FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. *In proceedings of The Web Conference 2020*, 1194 – 1204. https://doi.org/10.1145/3366423.3380196

Qi, T., Wu, F., Wu, C., Sun, P., Wu, L., Wang, X., Huang, Y., & Xie, X. (2022, April 10). ProFairRec: Provider Fairness-aware News Recommendation. http://arxiv.org/abs/2204.04724

Rahmani, H. A., Naghiaei, M., Dehghan, M., & Aliannejadi, M. (2022). Experiments on Generalizability of User-Oriented Fairness in Recommender Systems. *In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2755 – 2764. https://doi.org/10.1145/3477495.3531718

Rastegarpanah, B., Gummadi, K. P., & Crovella, M. (2019). Fighting Fire with Fire: Using Antidote Data to Improve Polarization and Fairness of Recommender Systems. *WSDM*, 231 – 239. https://doi.org/10.1145/3289600.3291002

Rendle, S. (2010). Factorization machines. *In proceedings of the IEEE International Conference on Data Mining,* ICDM, 995 – 1000. https://doi.org/10.1109/ICDM.2010.127

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). BPR: Bayesian personalized ranking from implicit feedback. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, UAI 2009, 452 – 461. https://doi.org/https://doi.org/10.48550/arXiv.1205.2618

Ren, Q., Jiang, Z., Cao, J., Li, S., Li, C., Liu, Y., Huo, S., He, T., & Chen, Y. (2024). A survey on fairness of large language models in e-commerce: Progress, application, and challenge (No. arXiv:2405.13025). arXiv. http://arxiv.org/abs/2405.13025

Rhue, L. (2024). The Anchoring Effect, Algorithmic Fairness, and the Limits of Information Transparency for Emotion Artificial Intelligence. *Information Systems Research*, *35*(3), 1479–1496. https://doi.org/10.1287/isre.2019.0493

Satinet, C., Fouss, F., Saerens, M., & Leleux, P. (2024). In-processing and post-processing strategies for balancing accuracy and sustainability in product recommendations. *Electronic Commerce Research and Applications*, *67*, 101433. https://doi.org/10.1016/j.elerap.2024.101433

Shang, Y., Gao, C., Chen, J., Jin, D., & Li, Y. (2024). Improving Item-side Fairness of Multimodal Recommendation via Modality Debiasing. *WWW*, 4697–4705. https://doi.org/10.1145/3589334.3648156

Smith, J. J., Buhayh, A., Kathait, A., Ragothaman, P., Mattei, N., Burke, R., & Voida, A. (2023). The Many Faces of Fairness: Exploring the Institutional Logics of Multistakeholder Microlending Recommendation. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1652 – 1663. https://doi.org/10.1145/3593013.3594106

Sonboli, N., Smith, J. J., Cabral Berenfus, F., Burke, R., & Fiesler, C. (2021). Fairness and Transparency in Recommendation: The Users' Perspective. *UMAP*, 274 – 279. https://doi.org/10.1145/3450613.3456835

Steck, H. (2018). Calibrated recommendations. *RecSys*, 154 – 162. https://doi.org/10.1145/3240323.3240372

Sühr, T., Hilgard, S., & Lakkaraju, H. (2021). Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring. *AAAI / AI*, Ethics, and Society, 989 – 999. https://doi.org/10.1145/3461702.3462602

Wang, N., & Chen, L. (2021). User Bias in Beyond-Accuracy Measurement of Recommendation Algorithms. *RecSys,* 133–142. https://doi.org/10.1145/3460231.3474244

Wan, M., Ni, J., Misra, R., & McAuley, J. (2020). Addressing Marketing Bias in Product Recommendations. *In proceedings of the 13th International Conference on Web Search and Data Mining*, 618 – 626. https://doi.org/10.1145/3336191.3371855

Wang, X., He, X., Wang, M., Feng, F., & Chua, T.-S. (2019). Neural Graph Collaborative Filtering. *In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 165 – 174. https://doi.org/10.1145/3331184.3331267

Wang, Y., Ma, W., Zhang, M., Liu, Y., & Ma, S. (2023). A Survey on the Fairness of Recommender Systems. *ACM TOIS*, 41(3), 1 – 43. https://doi.org/10.1145/3547333

Wang, Y., Sun, P., Ma, W., Zhang, M., Zhang, Y., Jiang, P., & Ma, S. (2024). Intersectional Two-sided Fairness in Recommendation. *WWW*, 3609–3620. https://doi.org/10.1145/3589334.3645518

Wang, Y., Zhou, H., Lu, G.-F., Gao, C., & Meng, S. (2025). Improving user-oriented fairness in recommendation via data augmentation: Don't worry about inactive users. *JSS (Journal of Systems and Software)*, *225*, 112387. https://doi.org/10.1016/j.jss.2025.112387

Weith, H., & Matt, C. (2023). Information provision measures for voice agent product recommendations—The effect of process explanations and process visualizations on fairness perceptions. *Electronic Markets*, *33*(1), 57. https://doi.org/10.1007/s12525-023-00668-x

Wu, C., Wu, F., Wang, X., Huang, Y., & Xie, X. (2021, April 15). Fairness-aware News Recommendation with Decomposed Adversarial Learning. *AAAI*. http://arxiv.org/abs/2006.16742

Wu, L., Chen, L., Shao, P., Hong, R., Wang, X., & Wang, M. (2021, April 23). Learning Fair Representations for Recommendation: A Graph-based Perspective. *WWW*. http://arxiv.org/abs/2102.09140

Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q., Xiong, H., & Chen, E. (2024). A survey on large language models for recommendation. *World Wide Web*, 27(5), 60. https://doi.org/10.1007/s11280-024-01291-2

Wu, S., Sun, F., Zhang, W., Xie, X., & Cui, B. (2023). Graph Neural Networks in Recommender Systems: A Survey. *ACM Computing Surveys*, 55(5), 1 – 37. https://doi.org/10.1145/3535101

Wu, Y., Cao, J., & Xu, G. (2024). Fairness in Recommender Systems: Evaluation Approaches and Assurance Strategies. *In proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 18(1), 1 – 37. https://doi.org/10.1145/3604558

Wu, Y., Cao, J., Xu, G., & Tan, Y. (2021). TFROM: A Two-sided Fairness-Aware Recommendation Model for Both Customers and Providers. *In proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1013 – 1022. https://doi.org/10.1145/3404835.3462882

Wu, Y., DuBois, C., Zheng, A. X., & Ester, M. (2016). Collaborative denoising auto-encoders for top-N recommender systems. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (pp. 153-162). *ACM*.

Wu, Y., Xie, R., Zhu, Y., Zhuang, F., Ao, X., Zhang, X., Lin, L., & He, Q. (2022, July 5). Selective Fairness in Recommendation via Prompts. *In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, http://arxiv.org/abs/2205.04682

Xu, B., Cui, Y., Sun, Z., Deng, L., & Zheng, K. (2021). Fair Representation Learning in Knowledge-aware Recommendation. *ICBK*, 385 – 392. https://doi.org/10.1109/ICKG52313.2021.00058

Yao, S., & Huang, B. (2017). Beyond Parity: Fairness Objectives for Collaborative Filtering. *NIPS*.

Ye, B. K., Tu, Y. J. T., & Liang, T. P. (2019). A hybrid system for personalized content recommendation. *Journal of Electronic Commerce Research*, 20(2), 91-104.

Zhang, C., Chen, S., Zhang, X., Dai, S., Yu, W., & Xu, J. (2024). Reinforcing Long-Term Performance in Recommender Systems with User-Oriented Exploration Policy. *In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*,, 1850–1860. https://doi.org/10.1145/3626772.3657714

Zhao, Z., Jing, Y., Feng, F., Wu, J., Gao, C., & He, X. (2024). Leave No Patient Behind: Enhancing Medication Recommendation for Rare Disease Patients. *In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 533–542. https://doi.org/10.1145/3626772.3657785

Zhang, J., Bao, K., Zhang, Y., Wang, W., Feng, F., & He, X. (2023). Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation. *RecSys*, 993 – 999. https://doi.org/10.1145/3604915.3608860

Zhao, W. X., Hou, Y., Pan, X., Yang, C., Zhang, Z., Lin, Z., Zhang, J., Bian, S., Tang, J., Sun, W., Chen, Y., Xu, L., Zhang, G., Tian, Z., Tian, C., Mu, S., Fan, X., Chen, X., & Wen, J. R. (2022). RecBole 2.0: Towards a More Up-to-Date Recommendation Library. *In proceedings of the International Conference on Information and Knowledge Management,* 4722 – 4726. https://doi.org/10.1145/3511808.3557680

Zhao, Y., Wang, Y., Zhang, Y., Wisniewski, P., Aggarwal, C., & Derr, T. (2024). Leveraging Opposite Gender Interaction Ratio as a Path towards Fairness in Online Dating Recommendations Based on User Sexual Orientation. *AAAI*, *38*, 22547–22555. https://doi.org/10.1609/aaai.v38i20.30263

Zhao, Y., Xu, M., Chen, H., Chen, Y., Cai, Y., Islam, R., Wang, Y., & Derr, T. (2024). Can One Embedding Fit All? A Multi-Interest Learning Paradigm Towards Improving User Interest Diversity Fairness. *WWW*, 1237–1248. https://doi.org/10.1145/3589334.3645662

Zhao, Z., Jing, Y., Feng, F., Wu, J., Gao, C., & He, X. (2024). Leave No Patient Behind: Enhancing Medication Recommendation for Rare Disease Patients. *In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 533–542. https://doi.org/10.1145/3626772.3657785

Zhdanov, D., Bhattacharjee, S., & Bragin, M. A. (2022). Incorporating FAT and privacy aware AI modeling approaches into business decision making frameworks. *Decision Support Systems*, *155*, 113715. https://doi.org/10.1016/j.dss.2021.113715

Zhou, M., Zhang, J., & Adomavicius, G. (2023). Longitudinal Impact of Preference Biases on Recommender Systems' Performance. Information Systems Research, isre.2021.0133. https://doi.org/10.1287/isre.2021.0133

Zhu, Z., Hu, X., & Caverlee, J. (2018). Fairness-Aware Tensor-Based Recommendation. *CIKM*, 1153 – 1162. https://doi.org/10.1145/3269206.3271795

Zhu, Z., Wang, J., & Caverlee, J. (2020). Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems. *In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 449 – 458. https://doi.org/10.1145/3397271.3401177