Fairness in Collaborative Filtering Recommender Systems

A Comparative Analysis of Trade-offs Across Model Architectures

Author: Jeeyoon Kang[1]
Supervisor: Dr. Masoud Mansoury[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A thesis submitted to the EEMCS Faculty of Delft University of Technology
in partial fulfilment of the requirements for the degree of

Bachelor of Science in Computer Science and Engineering

June 22, 2025

Jeeyoon Kang
Final project course: CSE3000 Research Project
Thesis committee: Dr. Masoud Mansoury, Dr. Nergis Tömen

Abstract

Recommender systems personalize content by predicting user preferences, but this often results in unequal treatment of users and items—for example, some users may receive lower-quality recommendations, while niche items remain underexposed. Although fairness-enhancing interventions exist, they can obscure the extent to which disparities stem from model architecture alone. This study investigates how collaborative filtering architectures affect both accuracy and fairness. We evaluate six models, including two non-personalized baselines, across two public datasets using a unified pipeline without fairness-specific interventions. Our results reveal a general trade-off: models with higher accuracy often exhibit greater fairness disparities, particularly on the user side. For example, `LightGCN` combines strong accuracy with relatively high item-side fairness, while `SLIMElastic` ranks high in accuracy but worsens unfairness. However, this trade-off is not uniform across datasets; `NeuMF` degrades notably on sparser data. These findings demonstrate that model architecture alone can shape fairness–accuracy trade-offs, highlighting the importance of considering dataset characteristics and model design when selecting or developing recommender systems.

1 Introduction

Recommender systems shape user experiences on platforms like *Netflix, Spotify, and Amazon* by suggesting what to watch, listen to, or buy next. Among various approaches, collaborative filtering (CF) is widely used, leveraging historical user–item interactions to generate personalized recommendations. CF models—such as matrix factorization, neural networks, and graph-based architectures—achieve strong performance on ranking metrics like `NDCG` and `Recall` [1].

While CF models excel at optimizing accuracy, they have also raised growing concerns around fairness. Unfairness can arise from imbalanced data, biased model assumptions, or an overemphasis on accuracy during training. These issues affect both users and items, as real-world systems often reproduce or amplify societal biases [2, 3, 4]. This can lead to lower-quality recommendations for certain users and reduced visibility for less popular items.

Empirical studies have shown that recommendation accuracy often varies across both user demographics (e.g., age, gender, location) and behavioral characteristics (e.g., activity level or preference for niche content), with users from majority groups typically receiving more relevant suggestions than those from minority or less represented groups [5]. On the item side, less popular or niche items frequently suffer from reduced exposure and lower ranking quality—a phenomenon known as *popularity bias*. These disparities have been observed across both traditional collaborative filtering methods [5] and modern deep learning-based recommendation models [6].

To mitigate these issues, various fairness-enhancing interventions have been proposed, including re-ranking, regularization, and adversarial training [7, 8, 9, 10, 11]. While often effective, such methods increase model complexity and make it difficult to assess the fairness properties intrinsic to the model architecture. In addition, many existing studies focus on isolated models or lack controlled experimental settings, limiting the generalizability of their findings.

Recent surveys emphasize the need for standardized benchmarks that assess fairness without modifying model architectures [12]. Without such baselines, it is difficult to isolate model-specific effects from variations in datasets, preprocessing, or evaluation—hindering reproducibility and fair comparison across studies.

To address this, we conduct a unified evaluation of six collaborative filtering models using a consistent and controlled experimental framework to assess both accuracy and group fairness across two widely used benchmark datasets: MovieLens 1M and Book-Crossing. By deliberately excluding fairness interventions and debiasing techniques, we aim to isolate the effects of model architecture on fairness outcomes and explore their trade-offs with accuracy.

We investigate the following research questions:

- **RQ1:** How do different collaborative filtering models perform in terms of accuracy, user fairness, and item fairness?
- **RQ2:** What trade-offs arise between accuracy and fairness across different models?
- **RQ3:** How do dataset characteristics—such as interaction sparsity, item popularity bias, and user activity imbalance—impact fairness–accuracy trade-offs in collaborative filtering models?

The remainder of this paper is organized as follows. Section 2 introduces the fairness concepts relevant to this study. Section 3 details the methodology, including datasets, models, and the evaluation setup. Section 4 presents the empirical results, followed by a discussion of broader implications and limitations in Section 5. Section 6 addresses ethical considerations and reproducibility. Finally, Section 7 concludes the paper and outlines directions for future research.

2 Background

Fairness in recommender systems involves both theoretical principles and practical challenges. We adopt the framework of Wang et al. [12], which distinguishes between formal fairness definitions and orthogonal classification dimensions. For comprehensive coverage, we refer readers to their survey and summarize only the aspects relevant to our analysis.

2.1 Fairness Definitions

Fairness is commonly categorized into two types: *process fairness*, which addresses the integrity of data collection and model training, and *outcome fairness*, which concerns the fairness of recommendation outputs and their impact on users and items.

According to Wang et al. [12], outcome fairness can be further analyzed along two dimensions: **target** and **concept**.

Target Defines the entity to which fairness is applied. *Individual fairness* ensures that similar users or items receive similar recommendations [13], while *group fairness* promotes equitable outcomes across groups defined by sensitive attributes such as gender, age, or item popularity.

Concept Captures the underlying fairness objective. *Consistent fairness* requires similar inputs to yield similar outputs. *Calibrated fairness* ensures alignment between recommendations and true relevance. *Counterfactual fairness* assesses stability under hypothetical changes to sensitive attributes. Other notions, including *envy-freeness*, *Rawlsian maximin*, and *maximin-share fairness*, are derived from fair division and welfare economics [12].

2.2 Classification Dimensions

In addition to formal fairness definitions, Wang et al. [12] propose three orthogonal dimensions to systematically classify fairness considerations in recommender systems: **subject**, **granularity**, and **optimization objective**.

Subject The *subject* dimension identifies the entity to which fairness is applied—users, items, or both. *User fairness* assesses whether individuals or user groups receive comparable recommendation quality, using metrics such as accuracy, diversity, or explainability [5, 14]. *Item fairness* evaluates the equitable treatment of items, focusing on aspects like exposure and prediction error [13]. *Joint fairness* considers fairness from both perspectives simultaneously, ensuring that neither users nor items are disproportionately disadvantaged.

Granularity The *granularity* dimension concerns the level at which fairness is assessed. *Single-instance fairness* evaluates fairness at the level of individual recommendation decisions. In contrast, *amortized fairness* examines fairness over aggregated distributions—across users, items, or interactions—capturing persistent or systemic disparities.

Optimization Objective The *optimization objective* dimension reflects the way fairness is incorporated into system behavior. *Treatment fairness* emphasizes the equitable allocation of exposure, ranking opportunities, or recommendation slots across groups. *Impact fairness*, on the other hand, focuses on long-term or downstream outcomes, such as engagement levels, satisfaction, or economic benefit.

2.3 Scope of This Work

This study is situated within the space of **group-based outcome fairness**, with a specific emphasis on **consistent fairness** evaluated at the **amortized** level. Our analysis targets **treatment fairness**, without applying any fairness-enhancing interventions such as re-ranking, regularization, or adversarial training.

3 Methodology

This section presents the methodology used in this study, covering the selected datasets, the set of collaborative filtering models evaluated, and the experimental setup.

Table 1: Statistics of preprocessed datasets. ML-1M requires no filtering. "Inter." = Interactions. Interaction stats (last two rows) are shown as Average / Median / Standard Deviation of interactions per user and item.

Dataset	**#Users**	**#Items**	**#Inter.**	**Sparsity**
ML-1M	6,040	3,706	1,000,209	95.52%
BX	6,851	9,085	115,219	99.81%

Inter. Stats	**Per User**	**Per Item**
ML-1M	166 / 96 / 193	270 / 124 / 384
BX	17 / 9 / 38	13 / 8 / 16

3.1 Datasets

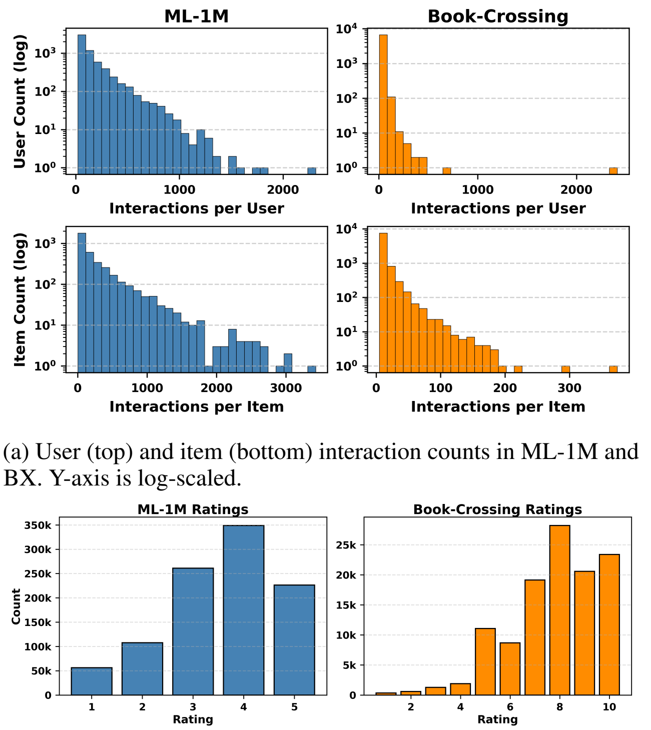
We conduct our experiments on two widely used benchmark datasets: **MovieLens 1M** (ML-1M) and **Book-Crossing** (BX). As shown in Table 1 and Figure 1, the datasets differ along several key dimensions relevant to our analysis, including metadata availability, interaction sparsity, distributional characteristics, and rating behavior.

Metadata availability informs our group definitions, as discussed in Section 3.2. ML-1M provides rich user-side metadata, including age, gender, occupation, and ZIP code, as well as item-level genre annotations. In contrast, BX offers limited and occasionally noisy user information (age and location), but includes more item metadata, such as author and publisher.

The datasets also differ in their interaction distribution characteristics. ML-1M is relatively dense: users engage with an average of 166 items, and items receive around 270 interactions. While it supports a wide range of user and item engagement levels, it exhibits a pronounced *item popularity bias*, where a small fraction of items accounts for the majority of interactions. In contrast, BX is considerably sparser, with users interacting with only 17 items on average and items receiving just 13. It displays a stronger *user activity imbalance*, characterized by a long-tailed distribution in which most users contribute very few interactions, and a small number of highly active users dominate the activity volume.

The rating distributions further underscore these differences. ML-1M uses a 1–5 scale that is relatively balanced, with a mode of 4, a mean of 3.58, and a standard deviation of 1.12—indicating a mild skew toward higher ratings. BX, by contrast, uses a 1–10 scale but shows a notable skew toward high ratings, with a mode of 8, a mean of 7.81, and a standard deviation of 1.78.

Prior to model training, we apply a consistent set of preprocessing steps to both datasets. First, we convert them into RecBole's atomic format, which tokenizes user, item, and interaction IDs. For BX, we remove interactions with a rating of 0, as these represent implicit feedback; no such filtering is needed for ML-1M. We then apply a 5-core filter to BX, retaining only users and items with at least five interactions—ML-1M already meets this criterion. Finally, to enable top-N recommendation, we binarize the ratings: interactions above a threshold are treated as positive, using a threshold of 3 for ML-1M and 7 for BX..



(a) User (top) and item (bottom) interaction counts in ML-1M and BX. Y-axis is log-scaled.

(b) Rating distributions in ML-1M and BX.

Figure 1: Interaction and rating statistics for ML-1M and Book-Crossing. (a) shows the long-tail distributions in user activity and item popularity; (b) illustrates rating skew and scale differences across datasets.

Table 2: Group sizes for users and items in ML-1M and Book-Crossing.

Group Dimension	Category	ML-1M	BX
Age	0–24	1325	–
	25–34	2096	–
	35–44	1193	–
	45+	1426	–
Gender	Female	1709	–
	Male	4331	–
Popularity Preference	Bestseller	–	1370
	Diverse	–	4111
	Niche	–	1370
Activity	Cold	1522	2097
	Moderate	3007	3022
	Active	1511	1732
(Item) Popularity	Head	112	291
	Mid-tail	1068	4611
	Long-tail	2503	4181

3.2 Group Assignment

To evaluate group-level fairness, we assign users and items to predefined groups based solely on the training set to prevent label leakage. Table 2 summarizes group definitions and distributions across datasets.

Users are grouped by both demographic and behavioral attributes. For ML-1M, demographic groups include gender and age (consolidated from seven to four bins). Behavioral groupings are based on implicit feedback. *Activity* groups—Cold, Moderate, and Active—are determined using the 25th and 75th percentiles of total interactions. *Popularity preference* groups are based on the share of interactions with Head items: Bestseller-oriented (top 20%), Diverse (middle 60%), and Niche-oriented (bottom 20%).

Items are grouped by popularity based on cumulative interaction mass. The top 20% of interactions define the Head, the next 60% the Mid-tail, and the bottom 20% the Long-tail.

Although BX offers more item-side metadata, these attributes primarily reflect provider-related characteristics (e.g., author, publisher) rather than item content. As such, they are more relevant to *provider fairness* than to item-level fairness, and are not used in this analysis.

3.3 Collaborative Filtering Models

Collaborative filtering (CF) methods are commonly classified into *memory-based* and *model-based* approaches. Memory-based methods compute similarities between users or items directly from the interaction matrix. In contrast, model-based methods learn latent user and item representations by optimizing task-specific objectives.

This study focuses exclusively on model-based approaches, which span several model families, including *matrix factorization*, *clustering-based*, *neural*, and *graph-based* models [15].

These models differ not only in how they represent user–item interactions but also in how they are trained. Depending on the formulation, training objectives can be either *pointwise*, where models predict binary labels or explicit scores, or *pairwise*, where the goal is to rank observed items above unobserved ones for each user.

To cover diverse modeling paradigms and training strategies, we evaluate four representative model-based collaborative filtering algorithms alongside two non-personalized baselines:

- **Pop** (*Non-personalized*): Recommends the most frequently interacted-with items in the training set, regardless of individual user preferences.
- **Random** (*Non-personalized*): Recommends items uniformly at random, serving as a trivial baseline with no personalization or learned structure.
- **BPR** (*Matrix Factorization*, *Pairwise*): Learns latent user and item embeddings by optimizing a pairwise ranking loss over implicit feedback [16].
- **LightGCN** (*Graph-based*, *Pairwise*): Propagates user–item embeddings through simplified graph convolutions, removing feature transformations and nonlinear activations to retain pure neighborhood aggregation [17]. .
- **SLIMElastic** (*Linear Models*, *Pointwise*): Learns a sparse item–item similarity matrix W via ℓ_1/ℓ_2-regularized linear regression on binary feedback [18].
- **NeuMF** (*Neural*, *Pointwise*): Combines GMF and MLP to capture linear and nonlinear user–item interactions, optimized with binary cross-entropy [19].

Table 3: Evaluation metrics used in this study. Arrows (↑, ↓) indicate whether higher or lower values are preferred.

Metric	Interpretation
Accuracy	
Recall@K ↑	Share of relevant items retrieved in top-K
Precision@K ↑	Share of top-K items that are relevant
NDCG@K ↑	Position-aware gain for relevant items
MAP@K ↑	Mean of average precision scores
Hit@K ↑	Fraction of users with at least one hit
Item Fairness	
IC@K ↑	Share of unique items recommended
Entropy@K ↓	Inconsistency of item exposure
Gini@K ↓	Inequality in item exposure concentration
AvgPop@K ↓	Average popularity of recommended items
Tail%@K ↑	Share of long-tail items in recommendations
Head%@K ↓	Share of head items in recommendations
User Fairness	
MAD@K ↓	Median deviation of group-level accuracy
STD@K ↓	Standard deviation of group-level accuracy

AvgPop = Average Popularity, IC = Item Coverage, Gini = Gini Index; Tail%@K and Head%@K are based on cumulative interaction mass: bottom 20% (tail), top 20% (head).

3.4 Evaluation Metrics

We assess model performance along three dimensions: *accuracy*, *item fairness*, and *user fairness*. Accuracy metrics evaluate the relevance and ranking quality of top-K recommendations. Item fairness metrics capture the diversity and balance of item exposure. User fairness metrics quantify disparities in recommendation quality across user groups based on demographic or behavioral attributes. Arrows (↑, ↓) indicate whether higher or lower values are preferred. Table 3 summarizes the full set of evaluation metrics used in this study.

3.5 Training and Evaluation

All models are trained and evaluated using the RecBole framework [20]. User–item interactions are split into training, validation, and test sets using an 8:1:1 ratio with grouped by user, ensuring each user appears in all three splits. Training data is shuffled at the start of each epoch. Adam optimizer with a fixed evaluation batch size of 4096 is used for all models.

For models trained with pairwise objectives, we adopt a standard uniform negative sampling strategy, where one negative item is randomly sampled per positive interaction. Although NeuMF is inherently a pointwise binary classification model, we apply the same negative sampling strategy to align with current best practices in implicit feedback settings, where relying on static labels alone is insufficient for effective training.

Model evaluation follows a full-ranking protocol: for each user, the model ranks the ground-truth item against all unobserved items. This protocol is applied consistently to both validation and test sets. Top-K performance is reported at $K{=}10$, with all metrics averaged over users. Evaluation metrics are described in Section 3.4.

3.6 Hyperparameter Settings

All models are tuned using W&B Sweeps with Hyperband-based Bayesian optimization, targeting validation NDCG@10 as the objective. Each trial is early-stopped after 10 consecutive iterations without improvement. The best configuration is selected based on the highest validation NDCG@10.

A shared hyperparameter search space is used for all models except SLIMElastic, including learning rate ({1e-4, 3e-4, 5e-4, 1e-3}), L_2 regularization weight ({1e-5, 5e-5, 1e-4, 3e-4, 1e-3}), embedding dimension ({64, 128, 256}), and batch size ({512, 1024, 2048}).

SLIMElastic is tuned over alpha ({0.1, 0.2, 0.4, 0.6}) and ℓ_1 ratio ({0.001, 0.01, 0.1, 0.2, 0.4}). LightGCN additionally searches over the number of graph convolution layers ({1, 2, 3, 4, 5}). NeuMF is tuned for dropout rate ({0.1, 0.2, 0.3}), MF/MLP embedding sizes ({8, 16, 32}), and MLP hidden layers ({[64, 32], [64, 32, 16]}).

4 Empirical Results

This section presents the empirical findings from experiments conducted on the ML-1M and Book-Crossing (BX) datasets, addressing the research questions outlined earlier.

4.1 Accuracy and Fairness Across Models (RQ1)

Accuracy Performance

Model accuracy rankings remain consistent across all evaluation metrics and both datasets, despite an expected decline in absolute scores on the sparser BX dataset (Tables 4 and 5).

The baselines behave as expected: Pop achieves moderate accuracy by exploiting popularity bias, while Random performs the poorest due to its uniform exposure strategy.

Among model-based models, SLIMElastic consistently achieves the highest accuracy across all five metrics and both datasets. LightGCN consistently ranks second in accuracy, followed by BPR, which delivers moderate but stable performance. NeuMF performs weakest among personalized models, especially on BX, where its neural architecture appears less robust under high sparsity.

Notably, the performance gap between models is more pronounced on BX, which is likely due to the amplifying effect of data sparsity on architectural differences.

Item-side Fairness Performance

Item-side fairness rankings vary more across models than accuracy rankings (Tables 4 and 5).

As expected, the baselines show contrasting behavior: Pop heavily concentrates exposure on head items, resulting in low exposure of long-tail items and high inequality of item exposure. Random, by contrast, achieves maximal coverage and the lowest disparity.

Among model-based models, LightGCN demonstrates the strongest item-side fairness on both datasets, outperforming others across all metrics. In contrast, SLIMElastic ranks lowest in item-side fairness on ML-1M. Despite moderate gains in long-tail exposure (Tail%) and average popularity (AvgPop) on BX, the model still suffers from limited item coverage and high exposure inequality.

Table 4: Top-10 ranking accuracy and item-side fairness comparison on the ML-1M dataset. Metrics marked ↑ are better when higher; ↓ when lower. Best and worst (excluding baselines) are in **bold** and underlined, respectively.

Model	Accuracy					Item-side Fairness					
	Recall↑	Precision↑	NDCG↑	Hit↑	MAP↑	IC↑	Tail%↑	Head%↓	AvgPop↓	Gini↓	Entropy↓
Random	0.0030	0.0046	0.0051	0.0449	0.0017	99.97%	64.56%	2.85%	207.52	0.1453	0.0021
BPR	0.0725	0.0595	0.0771	0.4535	0.0318	33.29%	2.43%	63.09%	1380.23	0.9361	**0.0043**
LightGCN	0.0755	0.0608	0.0785	0.4634	0.0322	**34.53%**	**2.44%**	**56.60%**	**1257.36**	**0.9248**	0.0044
SLIMElastic	**0.0922**	**0.0713**	**0.1017**	**0.5233**	**0.0462**	14.21%	0.04%	85.10%	1689.52	0.9762	0.0086
NeuMF	0.0710	0.0580	0.0755	0.4492	0.0311	24.30%	1.01%	66.16%	1426.66	0.9534	0.0057
Pop	0.0463	0.0450	0.0565	0.3545	0.0229	0.26%	0.00%	100.0%	2257.20	0.9974	0.2303

Table 5: Top-10 accuracy and item-side fairness comparison on the BX dataset. Metrics marked ↑ are better when higher; ↓ when lower. Best and worst (excluding baselines) are in **bold** and underlined, respectively.

Model	Accuracy					Item-side Fairness					
	Recall↑	Precision↑	NDCG↑	Hit↑	MAP↑	IC↑	Tail%↑	Head%↓	AvgPop↓	Gini↓	Entropy↓
Random	0.0010	0.0002	0.0005	0.0016	0.0003	99.92%	46.35%	3.21%	10.094	0.2062	0.0010
BPR	0.0427	0.0058	0.0198	0.0559	0.0114	54.74%	4.35%	74.99%	95.481	0.9195	**0.0012**
LightGCN	0.0525	0.0072	0.0253	0.0699	0.0151	**57.45%**	**7.23%**	**61.06%**	**73.204**	**0.8882**	0.0013
SLIMElastic	**0.0589**	**0.0087**	**0.0377**	**0.0840**	**0.0276**	11.95%	4.62%	70.20%	84.459	0.9791	0.0049
NeuMF	0.0344	0.0050	0.0179	0.0486	0.0112	23.02%	1.63%	83.63%	111.49	0.9731	0.0025
Pop	0.0198	0.0031	0.0117	0.0309	0.0081	0.11%	0.00%	100.0%	165.90	0.9990	0.2303

Item fairness of NeuMF declines on BX, where it ranks lowest in Tail% and AvgPop, indicating increased bias under sparse and limited feedback. BPR maintains stable mid-tier fairness across both datasets: it outperforms SLIMElastic on ML-1M in terms of Tail% but falls slightly behind on BX in balancing head and tail exposure.

User-side Fairness Performance

User-side fairness, reflected in group-level accuracy disparities, varies most across models (Figure 2). Disparities are larger for behavioral groups (activity level and popularity preference) than for demographic groups like gender or age.

Baseline behavior aligns with expectations: Random yields minimal group-level disparity, while Pop exhibits substantial disparity, as its reliance on head items systematically favors certain user groups over others.

Among personalized models, SLIMElastic demonstrates the poorest user-side fairness across both datasets, showing the largest disparities across activity groups. NeuMF and BPR display moderate levels of disparity, while LightGCN frequently ranks as the second least fair model. On BX, its fairness gaps are especially pronounced—likely due to the amplification of behavioral biases through neighborhood aggregation under data sparsity.

As shown in Figure 2, fairness rankings vary depending on group type and metric. SLIMElastic shows the highest dispersion in Recall and MAP, whereas Random and Pop exhibit the largest disparities in NDCG. Fairness rankings are generally more consistent on BX than on ML-1M. In particular, disparities across popularity preference groups remain stable, suggesting that item popularity exerts a strong and systematic influence on user-side fairness in BX.

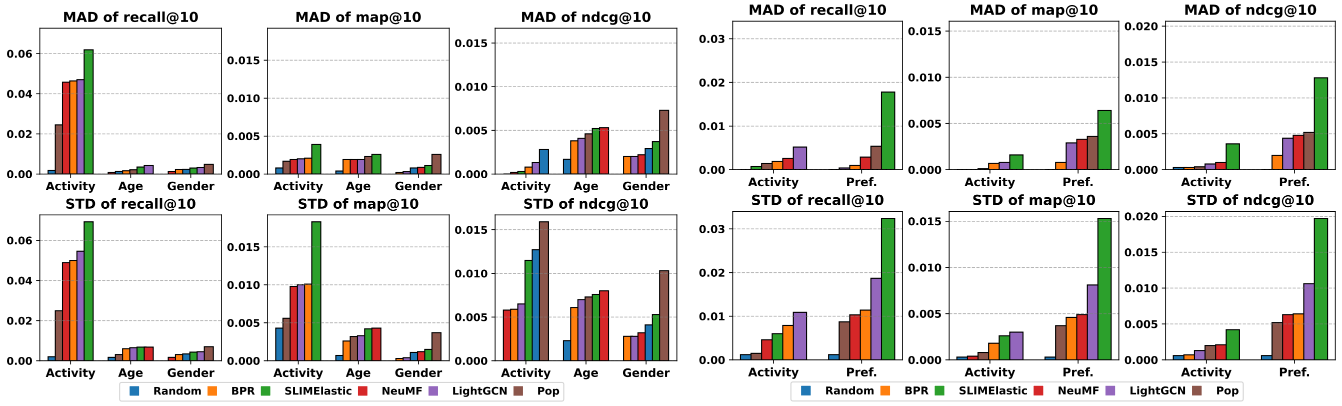
Table 6 reveals how models differ in treating user groups across datasets. On ML-1M, male users consistently achieve higher NDCG than female users, with variation by model. Similarly, on BX, users preferring bestsellers perform best across all models except Random, highlighting a popularity-driven accuracy bias. In terms of user activity, active users generally receive better recommendation accuracy on both datasets. However, LightGCN exhibits a slight preference for cold users on ML-1M. SLIMElastic favors cold users on ML-1M but active users on BX, whereas BPR shows the opposite pattern. All other models show consistently higher performance for active users across both datasets. Moderate users experience the lowest performance on ML-1M, unlike on BX.

4.2 Trade-offs Between Accuracy and Fairness (RQ2)

To address RQ2, we examine trade-offs between accuracy and fairness across models. Figure 3 shows the relationship between NDCG and item-side fairness metrics, while Figure 4 presents the corresponding user-side trade-offs.

A trade-off is evident across all model types. The baselines represent two extremes: Pop achieves moderate accuracy but low fairness due to its popularity-driven design, whereas Random ensures high fairness through uniform exposure at the expense of accuracy.

Among personalized models, trade-offs are still present, but some models offer a more favorable balance. SLIMElastic achieves the highest accuracy but ranks lowest on both item- and user-side fairness. In contrast, LightGCN strikes a more favorable balance—ranking second in accuracy, leading in item-side fairness, and performing moderately in user-side fairness. This suggests that its simplified GCN architecture



(a) ML-1M

(b) BX

Figure 2: Dispersion of top-10 accuracy metrics across user groups for ML-1M and BX. `MAD` (top row) and `STD` (bottom row) by group: Activity, Age, Gender (ML-1M); Activity, Preference (BX). Higher values indicate greater disparity.

Table 6: Group-wise NDCG@10 scores on ML-1M and BX, reported by Activity and Gender (ML-1M) and Activity and Popularity Preference (BX). Best and worst group scores per model (where applicable) are in **bold** and underlined, respectively.

Model	ML-1M					BX					
	Cold	Moderate	Active	Female	Male	Cold	Moderate	Active	Niche	Bestseller	Diverse
Random	0.0016	0.0022	**0.0075**	0.0051	**0.0053**	0.0000	0.0009	**0.0012**	0.0000	0.0000	**0.0011**
BPR	0.0817	0.0719	**0.0825**	0.0742	**0.0782**	**0.0206**	0.0196	0.0193	0.0162	**0.0282**	0.0182
LightGCN	**0.0848**	0.0729	0.0834	0.0756	**0.0797**	**0.0267**	0.0249	0.0241	0.0186	**0.0388**	0.0230
SLIMElastic	**0.1166**	0.0965	0.0968	0.0963	**0.1038**	0.0340	0.0376	**0.0424**	0.0223	**0.0610**	0.0351
NeuMF	0.0805	0.0704	**0.0807**	0.0723	**0.0768**	0.0164	0.0174	**0.0204**	0.0125	**0.0250**	0.0173
Pop	0.0495	0.0496	**0.0772**	0.0460	**0.0606**	0.0106	0.0110	**0.0143**	0.0065	**0.0169**	0.0117

helps distribute exposure more equitably while maintaining strong accuracy. `BPR` shows stable but unspecialized performance across all metrics. Despite its expressive architecture, `NeuMF` underperforms in both accuracy and item fairness, particularly on the sparse BX dataset.

Trade-offs between accuracy and fairness are observed on both the item and user sides; however, the trade-off is less pronounced on the item side than on the user side (Figures 3 and 4). `LightGCN` clearly achieves the best balance between item-side accuracy and fairness across both datasets, whereas no model demonstrates a similarly strong balance on the user side—particularly in terms of `Recall`-based dispersion.

Notably, model complexity does not guarantee better accuracy or fairness. For example, the more complex `NeuMF` performs worse than `SLIMElastic` in accuracy, while `LightGCN` outperforms the simpler `BPR`. Moreover, item-side fairness does not imply user-side fairness: `LightGCN` ranks highest in item fairness but lower in user fairness, whereas `NeuMF` shows the reverse pattern.

In summary, the accuracy–fairness trade-off persists. `SLIMElastic` achieves the highest accuracy but performs worst in fairness. In contrast, models like `LightGCN` strike a more effective balance. Importantly, item- and user-side fairness do not always align, and higher model complexity does not guarantee fairer or more balanced outcomes.

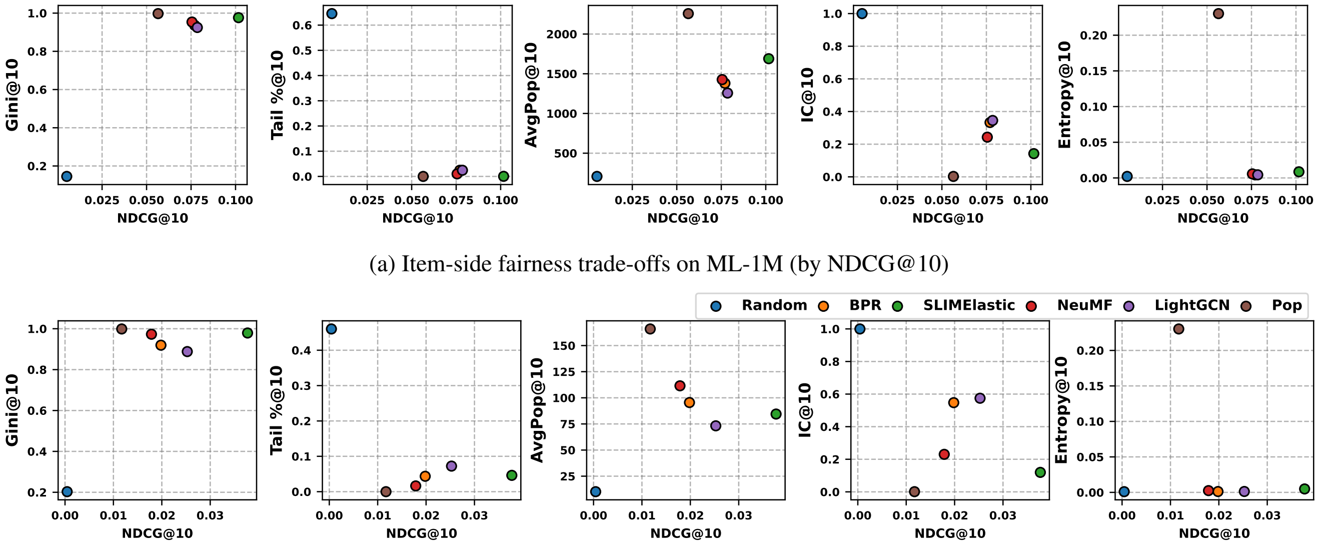
4.3 Trade-offs Across Datasets (RQ3)

To address RQ3, we compare the accuracy, fairness, and trade-off rankings of models across datasets with differing characteristics, as detailed in Section 3.1, examining how variations in data properties influence these relationships.

Accuracy rankings remain consistent across both datasets: `SLIMElastic` outperforms all other models, followed by `LightGCN`, while `NeuMF` yields the lowest accuracy among personalized models. The baselines, `Pop` and `Random`, remain the least effective throughout. Despite these stable rankings, absolute accuracy declines noticeably on BX due to its higher sparsity and fewer interactions per user and item.

Item-side fairness generalizes moderately across datasets: model rankings remain fairly consistent, with some dataset-specific shifts. On ML-1M, rankings are stable across all metrics, while on BX, they remain similar for most metrics but diverge in `Tail%` and `AvgPop`, where `NeuMF` degrades sharply, ranking last. In contrast, `BPR` and `LightGCN` show marked improvements in item coverage (`IC`), increasing from 33–34% on ML-1M to over 54–57% on BX. These shifts suggest that certain models—possibly due to their pairwise learning—are better equipped to sustain exposure diversity under sparsity.

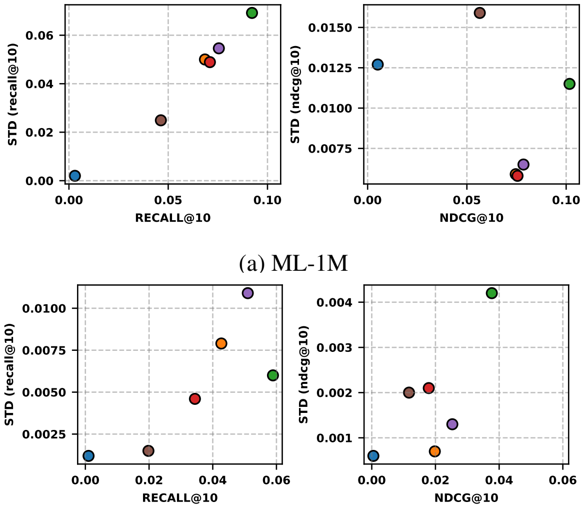
User-side fairness shows greater sensitivity to dataset characteristics than item-side fairness. In ML-1M, disparities are most pronounced across activity groups, whereas in BX, they



(a) Item-side fairness trade-offs on ML-1M (by NDCG@10)

(b) Item-side fairness trade-offs on BX (by NDCG@10)

Figure 3: Accuracy–fairness trade-offs on the item side for (a) ML-1M and (b) BX. Each point represents a model's NDCG@10 (x-axis) and item-side fairness score (y-axis).



(a) ML-1M

(b) BX

Figure 4: Accuracy–fairness trade-offs on the user side by activity group, via STD of group-wise NDCG@10. Lower values indicate more equal recommendation quality.

are largest across popularity preference groups. Models also vary in which user groups they favor, and this preference shifts across datasets—even within the same group type. For example, SLIMElastic favors cold users in ML-1M but active users in BX (Table 6). The magnitude of disparity further reflects these differences. As shown in Figure 2, SLIMElastic and Random yield the widest group gaps in ML-1M, whereas in BX, LightGCN and SLIMElastic exhibit the highest dispersion—most notably for LightGCN.

The accuracy–fairness trade-off persists across datasets, though its magnitude varies. For item-side fairness, models exhibit similar trade-off patterns on ML-1M, clustering closely (Figure 3). In contrast, BX displays greater dispersion, revealing sharper differences in how models balance accuracy and fairness under sparse conditions. The most notable differences appear in item coverage: on BX, pairwise models—BPR and LightGCN—achieve nearly double the coverage of pointwise models. LightGCN consistently achieves the best overall balance across both datasets, with its advantage particularly pronounced on BX.

For user fairness, the trade-off with accuracy is more strongly linear on BX than on ML-1M—particularly with respect to NDCG (Figure 4). This suggests that improving user fairness under sparse and skewed conditions may demand greater compromises in accuracy. Still, the general pattern holds: higher accuracy often comes at the cost of increased fairness disparity.

In summary, the fairness–accuracy trade-off is evident in both item and user dimensions, but its severity and clarity vary across datasets. Accuracy and item-side fairness generalize more consistently, whereas user fairness is highly sensitive to group definitions and data sparsity. On the sparser BX dataset, models exhibit greater variation in how they balance accuracy and fairness. LightGCN emerges as the most robust overall, though achieving user-side fairness remains more challenging across conditions.

5 Discussion

In this section, we reflect on the empirical findings, drawing broader implications beyond results and addressing key limitations and challenges encountered in this study.

Model Architecture and Fairness Dynamics The observed differences in accuracy and fairness outcomes can be explained by the inductive biases inherent in each model's architecture. SLIMElastic, with its pointwise regression objective and reliance on sparse co-occurrence modeling, achieves consistently high accuracy by reinforcing frequently co-rated item pairs. However, this emphasis on popular co-occurrences results in skewed exposure patterns and significant disparities across user groups. In contrast, LightGCN employs simplified graph convolutions, a form of neighborhood smoothing, which improves item-side fairness by increasing exposure of less popular items but may amplify behavioral biases, leading to increased user-side disparity. BPR, which optimizes pairwise rankings, avoids explicit modeling of absolute popularity, resulting in balanced, moderate performance in both accuracy and fairness. As noted in the previous section, both BPR and LightGCN show more favorable treatment of inactive users on the sparser BX dataset, which may be attributed to their pairwise training objective. Pairwise models can be fairer to sparse or low-activity users because they learn from relative preferences, making them less dependent on the volume of user interactions. Finally, NeuMF, combining deep neural components with a pointwise objective, struggles with sparse data such as BX, showing accuracy decline and pronounced bias toward head items due to difficulty diversifying recommendations. This likely stems from its reliance on deep parameterized components, which may require more interactions per user and item to learn stable representations and avoid overfitting to dominant patterns in sparse data.

Dataset and Group Imbalance Effects User fairness disparities can also stem from imbalanced group sizes within the training data. Since models are optimized to maximize overall performance, larger or more active user groups can dominate the training signal, inadvertently biasing model behavior or amplifying existing model biases. For example, as noted in Section 4, male users consistently achieve higher accuracy on ML-1M, suggesting that observed unfairness may be driven more by dataset imbalance than by model architecture. This highlights the importance of accounting for dataset composition when interpreting fairness evaluations.

Stability and Trade-Offs in User-Side Fairness User-side fairness is notably less consistent and stable than item-side fairness across datasets, metrics, and user group definitions. Model fairness rankings vary depending on both the metric and the group examined, with activity-based disparities dominating on ML-1M and popularity preference driving larger gaps on BX. Moreover, disparities computed using different accuracy metrics can yield varying conclusions. For instance, MAD may report low dispersion if only a subset of groups receives similar scores, while STD can be inflated by a few outlier groups even when most perform similarly. While item-side fairness can sometimes improve alongside accuracy, user-side fairness more often exhibits a negative correlation—indicating a trade-off. This may reflect genuine unfairness, where accuracy improvements disproportionately benefit dominant groups, but it can also arise from the sensitivity of dispersion metrics to group size and distribution, as well as the global objectives used during model training. These complexities underscore the importance of evaluating fairness through multiple lenses and interpreting outcomes with caution.

Hyperparameter Tuning and Its Impact Accuracy and fairness outcomes are also influenced by hyperparameter tuning. Variations in optimization targets (e.g., maximizing Recall versus NDCG) and tuning effectiveness can cause some models to benefit more than others, contributing to observed performance differences. This underlines the need for consistent, transparent, and fairness-aware tuning protocols in recommender system evaluations.

Limitations and Directions for Future Work This study does not incorporate statistical significance testing or confidence intervals, which are important for robustly assessing subtle fairness disparities that may be dataset-dependent. Future work should apply statistical methods such as bootstrapped confidence intervals or paired tests. Additionally, our evaluation covers representative models and two datasets but does not encompass the full range of fairness metrics, model types, or application domains. Expanding to include more models, diverse user attributes, and additional fairness notions—such as individual or counterfactual fairness—would enhance understanding of fairness in recommendation systems.

6 Responsible Research

Ethical Considerations This study utilizes two publicly available benchmark datasets—MovieLens 1M (ML-1M) and Book-Crossing (BX)—both of which are anonymized and widely adopted in academic research. No personally identifiable information (PII) beyond de-identified demographic and content metadata is processed. Group definitions (e.g., age, gender, item popularity) are derived solely from these available fields and employed strictly for aggregate fairness evaluation, without any intent to reinforce sensitive or socially constructed categories. We acknowledge that these datasets may contain inherent biases; however, this reflects real-world data conditions commonly encountered in practical systems. Consequently, we interpret group-based disparities as indicative of model behavior and the characteristics of the underlying data, rather than as inherent attributes of individual users or groups.

Reproducibility and Integrity All experiments were conducted using RecBole with reproducibility=True and a fixed random seed (42) to ensure deterministic behavior. We extended RecBole to support group-based fairness evaluation by reconstructing the data objects passed to a custom evaluator, containing only the users belonging to each group rather than the entire user set. Additionally, we implemented two custom metrics—Head% and a cumulative-popularity–based Tail%—while all other metrics are part of RecBole's default implementation. Key experimental settings, including ranking mode, filtering thresholds, group definitions, and evaluation metrics, are fixed as detailed in Section 3. Minor changes in preprocessing (e.g., filtering criteria or group assignment rules) or ranking strategy (e.g., full vs. sampled ranking) may shift data distributions and affect fairness outcomes.

7 Conclusions and Future Work

This study examined how different collaborative filtering model architectures behave in terms of accuracy, fairness, and their trade-offs. We evaluated four personalized models and two non-personalized baselines across two benchmark datasets using a unified framework. By excluding fairness interventions, we isolate architectural effects and reveal how models balance accuracy and fairness under varying data conditions.

Our findings indicate a consistent trade-off: higher accuracy often coincides with increased fairness disparities, particularly on the user side. However, some models achieve a more favorable balance. `SLIMElastic`, for example, delivered the highest accuracy but exhibited substantial unfairness, while `LightGCN` combined strong accuracy with better item-side fairness.

Moreover, item-side fairness generalized more consistently across metrics and datasets, whereas user-side fairness was more sensitive to metric choice and dataset characteristics. Notably, fairness in one dimension (e.g., items) does not imply fairness in another (e.g., users), and greater model complexity does not necessarily lead to more equitable outcomes.

Overall, the results reveal an inverse relationship between accuracy and fairness, though its strength varies across models and datasets. Since model architecture alone can influence these trade-offs, understanding dataset characteristics and selecting appropriate models is essential. Moreover, as item-side and user-side fairness do not always align, both should be jointly considered to support fairer recommender systems.

Future work could incorporate statistical significance testing, broaden the architectural scope, refine group definitions, and explore alternative user-side fairness metrics. Advancing toward individual-level fairness and integrating complementary notions—such as calibrated or counterfactual fairness—may offer deeper insights into fairness dynamics in recommendation.

References

[1] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Item-based collaborative filtering recommendation algorithms,” in *Proceedings of the 10th International Conference on World Wide Web*, ser. WWW ’01. New York, NY, USA: Association for Computing Machinery, 2001, pp. 285–295. [Online]. Available: https://doi.org/10.1145/371920.372071

[2] Z. Zhu, Y. He, X. Zhao, Y. Zhang, J. Wang, and J. Caverlee, “Popularity-opportunity bias in collaborative filtering,” in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, ser. WSDM ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 85–93. [Online]. Available: https://doi.org/10.1145/3437963.3441820

[3] Álvaro González, F. Ortega, D. Pérez-López, and S. Alonso, “Bias and unfairness of collaborative filtering based recommender systems in movielens dataset,” *IEEE Access*, vol. 10, pp. 68 429–68 439, 2022.

[4] H. Abdollahpouri, M. Mansoury, R. Burke, and B. Mobasher, “The unfairness of popularity bias in recommendation,” *CoRR*, vol. abs/1907.13286, 2019. [Online]. Available: http://arxiv.org/abs/1907.13286

[5] M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera, “All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness,” in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, Feb 2018, pp. 172–186. [Online]. Available: https://proceedings.mlr.press/v81/ekstrand18b.html

[6] R. Islam, K. N. Keya, Z. Zeng, S. Pan, and J. Foulds, “Debiasing career recommendations with neural fair collaborative filtering,” in *Proceedings of the Web Conference 2021*, ser. WWW ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 3779–3790. [Online]. Available: https://doi.org/10.1145/3442381.3449904

[7] R. Burke, “Multisided fairness for recommendation,” 2017. [Online]. Available: https://arxiv.org/abs/1707.00093

[8] A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi, “Putting fairness principles into practice: Challenges, metrics, and improvements,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 453–459. [Online]. Available: https://doi.org/10.1145/3306618.3314234

[9] S. Yao and B. Huang, “Beyond parity: Fairness objectives for collaborative filtering,” in *Advances in Neural Information Processing Systems (NeurIPS 2017)*. Curran Associates Inc., 2017, pp. 2925–2934. [Online].

Available: https://dl.acm.org/doi/pdf/10.5555/3294996.3295052

[10] A. Singh and T. Joachims, “Fairness of exposure in rankings,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD ’18. New York, NY, USA: Association for Computing Machinery, Jul 2018, pp. 2219–2228. [Online]. Available: http://dx.doi.org/10.1145/3219819.3220088

[11] J. Huang, H. Oosterhuis, M. Mansoury, H. van Hoof, and M. de Rijke, “Going beyond popularity and positivity bias: Correcting for multifactorial bias in recommender systems,” in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR 2024. New York, NY, USA: Association for Computing Machinery, Jul 2024, pp. 416–426. [Online]. Available: http://dx.doi.org/10.1145/3626772.3657749

[12] Y. Wang, W. Ma, M. Zhang, Y. Liu, and S. Ma, “A survey on the fairness of recommender systems,” *ACM Transactions on Information Systems (TOIS)*, vol. 41, no. 3, Feb. 2023. [Online]. Available: https://doi.org/10.1145/3547333

[13] A. J. Biega, K. P. Gummadi, and G. Weikum, “Equity of attention: Amortizing individual fairness in rankings,” in *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’18. ACM, June 2018. [Online]. Available: https://doi.org/10.1145/3209978.3210063

[14] Z. Fu, Y. Xian, R. Gao, J. Zhao, Q. Huang, Y. Ge, S. Xu, S. Geng, C. Shah, Y. Zhang, and G. de Melo, “Fairness-aware explainable recommendation over knowledge graphs,” 2020. [Online]. Available: https://arxiv.org/abs/2006.02046

[15] M. F. Aljunid, D. Manjaiah, M. K. Hooshmand, W. A. Ali, A. M. Shetty, and S. Q. Alzoubah, “A collaborative filtering recommender systems: Survey,” *Neurocomputing*, vol. 617, p. 128718, 2025.

[16] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, “Bpr: Bayesian personalized ranking from implicit feedback,” in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, ser. UAI ’09. Arlington, Virginia, USA: AUAI Press, 2009, pp. 452–461.

[17] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, “Lightgcn: Simplifying and powering graph convolution network for recommendation,” in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 639–648. [Online]. Available: https://doi.org/10.1145/3397271.3401063

[18] X. Ning and G. Karypis, “Slim: Sparse linear methods for top-n recommender systems,” in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ser. ICDM ’11. USA: IEEE Computer Society, 2011, p. 497–506. [Online]. Available: https://doi.org/10.1109/ICDM.2011.134

[19] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, “Neural collaborative filtering,” in *Proceedings of the 26th International Conference on World Wide Web*, ser. WWW ’17. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017, p. 173–182. [Online]. Available: https://doi.org/10.1145/3038912.3052569

[20] W. X. Zhao, S. Mu, Y. Hou, Z. Lin, Y. Chen, X. Pan, K. Li, Y. Lu, H. Wang, C. Tian, Y. Min, Z. Feng, X. Fan, X. Chen, P. Wang, W. Ji, Y. Li, X. Wang, and J.-R. Wen, “Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms,” in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, ser. CIKM ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 4653–4664. [Online]. Available: https://doi.org/10.1145/3459637.3482016