A Normative Framework for Benchmarking Consumer Fairness in Large Language Model Recommender Systems

Yashar Deldjoo, Fatemeh Nazary

Polytechnic University of Bari, Italy

Abstract

The rapid adoption of large language models (LLMs) in recommender systems (RS) presents new challenges in understanding and evaluating their biases, which can result in unfairness or the amplification of stereotypes. Traditional fairness evaluations in RS primarily focus on collaborative filtering (CF) settings, which may not fully capture the complexities of LLMs, as these models often inherit biases from large, unregulated data. This paper proposes a normative framework to benchmark consumer fairness in LLM-powered recommender systems (RecLLMs). We critically examine how fairness norms in classical RS fall short in addressing the challenges posed by LLMs. We argue that this gap can lead to arbitrary conclusions about fairness, and we propose a more structured, formal approach to evaluate fairness in such systems. Our experiments on the MovieLens dataset on *consumer fairness,* using in-context learning (zero-shot vs. few-shot) reveal fairness deviations in age-based recommendations, particularly when additional contextual examples are introduced (ICL-2). Statistical significance tests confirm that these deviations are not random, highlighting the need for robust evaluation methods. While this work offers a preliminary discussion on a proposed normative framework, our hope is that it could provide a formal, principled approach for auditing and mitigating bias in RecLLMs. The code and dataset used for this work will be shared at **gihub-anonymized**.

Keywords

Large Language Models, Recommender Systems, Fairness Evaluation, Benchmarking, Framework, In-Context Learning, Normative

1. Introduction

Context. Fairness in recommender systems (RS) has garnered significant attention in recent years, driven by the need to mitigate biases that can negatively impact both consumers and providers. Most existing research in the RS field, however, has focused on fairness in classical collaborative filtering (CF) setting, which primarily relies on *in-domain* user-item interaction data to compute recommendations [1, 2]. Since the introduction of ChatGPT in 2023, the RS community has seen unprecedented interest in integrating generative models—particularly those powered by pre-trained large language models (LLMs)–for personalization [3, 4], (see in particular Deldjoo et al. [5] for a frame of reference). These models, which capture vast amounts of knowledge during pre-training on semi-supervised tasks, can be quickly adapted to a variety of contexts within RS, offering notable advantages for personalized recommendations. These benefits include *efficiency* (rapid deployment and adaptability), *precision and context-awareness* (enhanced personalization across diverse tasks), and *robustness in data scarce scenarios* (the ability to perform well under sparse data conditions).

However, integrating LLMs into recommender systems introduces new risks, particularly biases embedded in the training data, which can lead to unfairness or the amplification of stereotypes affecting sensitive or protected groups. Given the unregulated nature of online data, it becomes a critical concern to *(i)* understand, *(ii)* evaluate, and *(iii)* mitigate biases and unfairness of RecLLMs. RecLLMs differ significantly from traditional CF systems in several key areas: the *input space* (where simple star ratings are replaced by more complex inputs like natural language user profiles), *model types* (pre-trained on vast datasets rather than being directly trained using in-domain data), and *output spaces* (offering more structured outputs like complementary items or detailed explanations instead of just item IDs). This work advocates for developing RecLLM fairness evaluation frameworks that account for these differences, as they can influence our understanding of what is fair and unfair. The central question explored in this work revolves around the following:

*How can we audit the fairness of **RecLLMs** (recommender systems powered by large language models), and how does fairness evaluation for RecLLMs differ from traditional CF methods?*

Research Problem. In RecLLMs, there is the ability to incorporate sensitive demographic information, such as gender, directly from natural language (NL) user profiles—something that mainstream collaborative filtering (CF) models typically do not utilize. Figure 1 illustrates this with an example. Building on this, Zhang et al. [6] propose a fairness evaluation framework for RecLLMs that examines how demographic attributes, such as gender, can impact recommender outcomes. Their approach defines unfairness as the difference in recommendations between a sensitive ranker (which considers demographic factors) and a neutral ranker (which does not), based solely on differences in item IDs or their ranking in the list. This approach equates differences across user groups with unfairness, oversimplifying the issue by failing to account for situations where such differences might represent valid personalization.

For example, if a RecLLM recommends songs like "Hey Young Girl" by Lloyd based on prior interactions in a neutral scenario, but then suggests a song by Jamiroquai when gender is factored in, the system might flag this as unfair. This assumes the recommendation change is driven by gender stereotypes, without considering that such recommendations may align with the user's actual preferences.

To address this limitation, Deldjoo and Di Noia [7] propose the CFairLLM framework, which improves on FaiRLLM

ROEGEN@RECSYS'24. The 1st Workshop on Risks, Opportunities, and Evaluation of Generative Models in Recommender Systems, Colocated with ACM Conference on Recommender Systems (RecSys) in Bari, Italy, October 2024

deldjooy@acm.org (Y. Deldjoo); fatemeh.nazary@poliba.it (F. Nazary)

https://yasdel.github.io/ (Y. Deldjoo); https://sisinflab.poliba.it/people/fatemeh-nazary/ (F. Nazary)

0000-0002-6767-358X (Y. Deldjoo); 0000-0002-6683-9453 (F. Nazary)

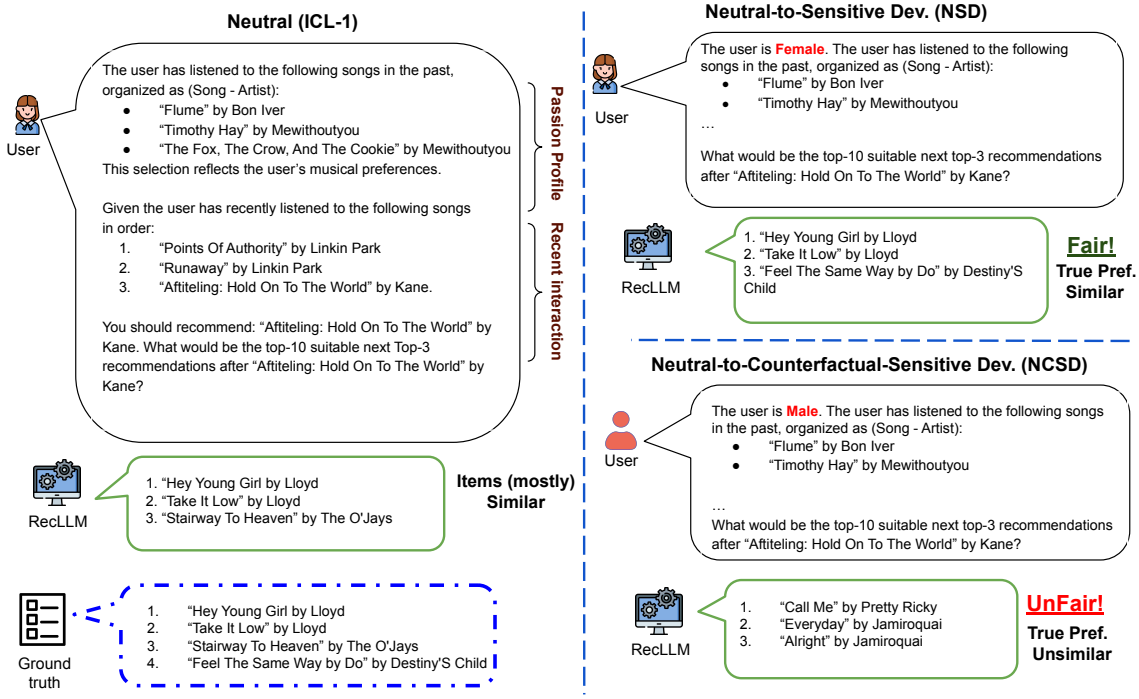

Figure 1: This figure illustrates the direct use of Large Language Models (LLMs) in generating personalized recommendations. It compares outputs under neutral conditions with those generated under scenarios that consider sensitive attributes.

by assessing fairness based on whether the benefits a sensitive ranker provides differ (or more precisely worse) from those of a reference ranker, to flag it unfair. Thus, CFairLLM evaluates recommendation variations by comparing them to the user's true preferences. For instance, if the sensitive ranker suggests songs that better match the user's preferences, even if they differ from the neutral ranker's list, these variations may indicate proper personalization rather than unfairness. This highlights the importance of defining fairness norms clearly; otherwise, research results might be contradictory.

Overall, these approaches lack a formal discussion of key elements essential for determining fairness in RecLLMs, such as establishing a clear reference point for evaluation and defining precise benefits and metrics for measuring fairness.

Contributions. The current work advocates for the need for a "**normative framework**" to address the above issues–one that ensures fairness is evaluated according to well-established, principled standards, rather than subjective or arbitrary assumptions (e.g., assuming all differences indicate unfairness). Additionally, it highlights the importance of having clear, objective criteria for consistently and meaningfully assessing fairness.

The main contributions of this paper are as follows:

1. **Distinguishing Fairness Frameworks:** We differentiate between two types of fairness evaluation in RecLLMs: *(i)* fairness when *sensitive attributes* are involved in generating recommendations (referred to as *sensitive rankers*), and *(ii)* fairness when comparing recommendations from neutral rankers, which do not use sensitive attributes, to predefined target distributions. This distinction is crucial for understanding the nuances in fairness evaluation between RecLLMs and traditional CF models, which typically rely on (i). Our work considers both approaches.
2. **Introduction of Novel Fairness Metrics:** Building on the previous distinction, we introduce three fairness evaluation metrics, which essentially operate on the same principle: comparing the deviation and difference between the RS ranker and another ranker (either a reference or a target representation). The metrics are as follows: Neutral vs. Sensitive Ranker Deviation (NSD), Neutral vs. Counterfactual Sensitive Deviation (NCSD), and Intrinsic Fairness (IF). While NSD and NCSD are specific to RecLLMs, IF is typically applied in CF scenarios.
3. **Introducing of Reference Rankers:** A key aspect of fairness evaluation in RecLLMs is the use of reference rankers. Each fairness notion relies on a specific reference ranker, such as the neutral ranker (for NSD and NCSD) or a predefined target distribution (for IF). The choice of the reference ranker plays a critical role in measuring deviations, as it influences how fairness or unfairness is perceived. In contrast to previous work that focuses merely on differences between rankers, as suggested by [7], we consider preference alignment with ground truth data.
4. **Quantification of Fairness Deviations:** To quantify ranker deviation, we propose metrics that assess the quality of recommendations through both set-based and rank-based measures. These measures examine the accuracy of the ranking and the overall benefit derived by different demographic groups. In addition to evaluating benefit deviation, we apply statistical significance tests to determine whether observed differences in recommendation benefits across groups are meaningful. This provides a comprehensive mechanism to quantify fairness deviations and ensures that the quality of recommendations is assessed based on rigorous, statistically sound methods.

Overall, this work advocates the need for a *normative framework* for fairness evaluation of RecLLMs, proposing a more formal approach than previous research. In this framework, fairness is evaluated against clear, well-defined standards, aiming to avoid arbitrary assumptions and provide a structured method for assessing how sensitive information affect recommendation outcomes. Our focus in this work is specifically on **consumer fairness**.

2. Evaluation Framework for Consumer Fairness in RecLLMs

We present a multi-faceted framework designed to evaluate fairness in Recommender Systems powered by Large Language Models (RecLLMs). This work builds significantly upon and extends previous research [6, 7], but offers a more formal approach to fairness evaluation in RecLLMs.

2.1. Definitions.

Definition of Groups. In this study, fairness discussions are conducted at the group level. We denote by $\mathcal{A}$ the set of all sensitive attributes, represented as $\mathcal{A} = \{a, b\}$, where each attribute a and b corresponds to specific characteristics. We specifically use

- The attribute a to represent "**gender**," with the values a_1 and a_2, where $a_1 =$ Male and $a_2 =$ Female.
- The attribute b to represent "**age-groups**," with the values b_1 and b_2, where $b_1 =$ Young and $b_2 =$ Old.

For simplicity, in this work we only consider groups that are **independent** and **binary**. However, overlapping groups can also be considered by defining combinations of attributes, such that the set of possible overlapping groups is given by $\mathcal{G} = \{(a_1, b_1), (a_1, b_2), (a_2, b_1), (a_2, b_2)\}$. Here, each pair represents a distinct demographic group (e.g., (a_1, b_1) for young males, (a_2, b_2) for older females), allowing to analyze fairness across intersectional identities [7].

Ranking Lists. We first introduce primary ranker types for fairness definitions.

Neutral Ranker ($\mathcal{R}_N$): Referred to as the *neutral ranking list*, this term describes a sequence of items $\{i_1, i_2, \ldots, i_k\}$ ranked by a Recommender Language Learning Model (RecLLM), using NL profile (as prompts) that do not incorporate sensitive user attributes. The neutral ranker is designed to reflect scenarios based purely on non-sensitive demographic data. It bases recommendations solely on the historical interaction of the user with the system.

Sensitive Ranker ($\mathcal{R}_S^a$): Short for *sensitive ranking list*, it denotes a sequence of items $\{i_1^a, i_2^a, \ldots, i_k^a\}$ ranked by a RecLLM using prompts that do utilize sensitive attributes such as gender, age, etc. They aim to capture scenarios where the LLM is potentially influenced by sensitive attributes, whether positively (providing more relevant recommendations) or negatively (recommending less relevant items).

Counterfactual Sensitive Ranker ($\mathcal{R}_{CS}^{do(a)}$): This ranker represents a sequence of items ranked by a RecLLM under the counterfactual scenario where the sensitive attribute a is set to a specific hypothetical value through the $do()$ operation. For example, $\mathcal{R}_{CS}^{do(\text{Male})}$ tests the recommendations as if the gender of every user were male, regardless of their actual gender. This method allows us to explore "**what-if**" scenarios, examining how different assumed values of sensitive attributes impact the recommendations, thereby exploring **counterfactual outcomes**. See also Section 2.1, the discussion of NCSD.[1]

Fairness Frameworks. On the consumer side, we consider the following fairness notions, each linked to the corresponding rankers:

Neutral vs. Sensitive Ranker Deviation (NSD): This notion measures disparities between the neutral ranker ($\mathcal{R}_N$) and the sensitive ranker ($\mathcal{R}_S^a$), evaluating how the inclusion of sensitive attributes influences the recommendations. *Thus, the neutral ranker $\mathcal{R}_N$ serves as the 'reference' against which fairness is measured.*

Neutral vs. Counterfactual Sensitive Deviation (NCSD): This concept assesses changes in recommendations when a sensitive attribute is counterfactually altered using the $do()$ operation, setting the attribute to a specific hypothetical value. The comparison is made between the counterfactual sensitive ranker ($\mathcal{R}_{CS}^{do(a)}$) and the neutral ranker ($\mathcal{R}_N$). *Here, we select $\mathcal{R}_N$ as the reference ranker to evaluate how assumptions about changes in a affect the recommendations.*[2]

Intrinsic Fairness (IF): Focusing on qualities intrinsic to recommendations, IF evaluates the fairness of distributions generated by the neutral ranker ($\mathcal{R}_N$), and evaluates the benefits provided by the recommender across sensitive groups (e.g., male vs. female). Since no direct comparisons between sensitive ranker(s) are conducted, this analysis is essentially testing where the prevalence of certain sensitive groups in training data skew LLM outputs. *Thus, a predefined 'target distribution', e.g., uniform, serves as the reference against which fairness is measured.*

It could be noted that both NSD and NCSD evaluate fairness across *two types of rankers*, examining the potential biases introduced by sensitive attributes and their counterfactual adjustments, while IF focuses on a *single ranker*, the Neutral Ranker.

Fairness quantification. To quantify unfairness in RecLLMs, we start by defining the general concept of benefit deviation, which serves as the foundation to *quantify* unfairness in our framework, given by:

$$\Delta\mathcal{B} = \mathcal{B}(\mathcal{R}_X) - \mathcal{B}(\mathcal{R}_{ref}) \tag{1}$$

where $\mathcal{R}_X$ represent ranking generated by the target recommender (e.g., sensitive ranker), $\mathcal{R}_{ref}$ is the reference ranker (e.g., $\mathcal{R}_N$ in NSD), and $\mathcal{B}$ represents the benefit derived from each list. A lower value of $\Delta\mathcal{B}$ in every scenario below indicates a higher amount of unfairness.

[1] Note that we recognize this might be a naive way of implementing the "*what-if*" scenario, since e.g., with $do(\text{Male})$ and $do(\text{Female})$, only part of the population is hypothetically altered. It nonetheless provides a framework for exploring how altering a single attribute could influence outcomes.

[2] Note that for NCSD, $\mathcal{R}_S^a$ could also be used as the reference ranker.

1. **Quantifying $\Delta\mathcal{B}$ for NSD.**

$$\Delta\mathcal{B} = \mathcal{B}(\mathcal{R}_S^a) - \mathcal{B}(\mathcal{R}_N) \tag{2}$$

This metric compares the benefits derived from comparing a sensitive ranker $\mathcal{R}_S^a$, and a neutral ranker $\mathcal{R}_N$. It evaluates how the inclusion of sensitive attributes impacts the benefits of the recommendation. A positive $\Delta\mathcal{B}$ could indicate enhanced personalization due to the introduction of sensitive attributes, while a negative deviation could suggest unfairness due to stereotypes or biases.

For NSD, we focus on comparing the changes across different groups, specifically:

- For gender. $\Delta\mathcal{B}_{a_1}$ and $\Delta\mathcal{B}_{a_2}$ where a_1 and a_2 correspond to Male and Female, respectively;
- For age categories: $\Delta\mathcal{B}_{b_1}$ and $\Delta\mathcal{B}_{b_2}$ where b_1 and b_2 represent the Young and Adult groups, respectively.

We could utilize a numerical threshold (thr) set at a predefined value, to gauge the magnitude of deviations, providing a quantitative measure of potential unfairness. In our work, beyond this, we adopt a more robust approach by using statistical significance tests to measure whether the means of two distributions—specifically $\Delta\mathcal{B}_{a_1}$ and $\Delta\mathcal{B}_{a_2}$ for gender, and $\Delta\mathcal{B}_{b_1}$ and $\Delta\mathcal{B}_{b_2}$ for age categories—are significantly different. We employ the t-test for independent samples to ascertain differences between these distributions ($p < 0.05$).[3]

Example. Suppose the benefit deviation $\Delta\mathcal{B}_{a_1}$ (Male) is 0.12, and $\Delta\mathcal{B}_{a_2}$ (Female) is -0.15. The positive deviation for males suggests an enhanced personalization effect, while the negative deviation for females, and particularly a large deviation from Male, indicates potential unfairness due to biased or stereotypical recommendations favoring males over females.

To measure NSD, we calculate the disparity in benefit deviations as $\delta_{\text{gender}} = \Delta\mathcal{B}_{a_1} - \Delta\mathcal{B}_{a_2}$ and $\delta_{\text{age}} = \Delta\mathcal{B}_{b_1} - \Delta\mathcal{B}_{b_2}$, using these differences as the main measures of unfairness. We intentionally use the signed version of the metric to discern the direction of unfairness.

Note. The threshold set for differentiating the levels of fairness concerns are inherently subjective and may vary depending on the specific task, system, or analysis objectives. In this work, we chose a threshold value that is reasonably suitable but acknowledge that what makes an "appropriate" value could differ widely based on context. Moreover, we introduce Table 1 to contribute to a more systematic and organized approach to categorize fairness metrics, employing "**color coding**" to visually distinguish between the various levels of concern.

Table 1 essentially aims to present a structured assessment of fairness, organizing different levels of disparity based on $\Delta\mathcal{B}$ and associated p-values into categories ranging from 'Safe' to 'Significant Issue'. This categorization helps stakeholders quickly identify potential biases in the recommendation system and determine the urgency of needed interventions. One might choose to adjust the number of levels or the criteria for each level based on their particular needs, regulations, and the nuances of their data.

[3]Additionally, other statistical significance tests such as the Mann-Whitney U test, a non-parametric test could be used when the data does not meet the assumptions necessary for the t-test.

Table 1
Fairness Evaluation Based on the threshold δ, $\Delta\mathcal{B} < \delta$ and p-value

Metric	(δ, p-value)	Status
Level 1	Small - (p>0.05)	Safe
Level 2	Fairly large - (p>0.05)	Attention Needed
Level 3	Large - (p>0.05)	Likely Issue
Level 4	Large/Small - (p<0.05)	Significant Issue

(2) **Neutral vs. Counterfactual Sensitive Deviation (NCSD).**

$$\Delta\mathcal{B} = \mathcal{B}(\mathcal{R}_{CS}^{do(a)}) - \mathcal{B}(\mathcal{R}_N) \tag{3}$$

NCSD essentially measures the difference in recommendation performance in a Hypothetical Scenario, asking how recommendations would perform if everyone were considered to be of the same gender (e.g., male or female). As stated before, although we could use the correct gender of the user (the sensitive ranker), we chose to use the neutral recommender as the reference.

To explore the impact of each attribute value in a controlled, hypothetical scenario, we symbolically use the causal $do()$ operator:

- $do(\textbf{Gender} = \textbf{Male})$ — Simulating the scenario where every individual, regardless of their original gender, is considered as male.
- $do(\textbf{Gender} = \textbf{Female})$ — Simulating the scenario in which every individual is considered as female.

This method allows us to assess the outcomes if the gender of every individual was hypothetically set to Male and then to Female, (and similar for age-categories), exploring the robustness and fairness of the system under these gender-altered conditions.

Finally, Intrinsic Fairness (IF) examines the fairness of recommendation distributions by a neutral ranker, $\mathcal{R}_N$, *across sensitive groups* such as male versus female. While the previous approaches may be more specific to RecLLMs due to the integration of "demographic information," IF represents a more general approach that can also be and has been widely applied to traditional recommendation models, such as collaborative filtering models [8, 9, 10]. Essentially, IF evaluates whether the outcomes provided by $\mathcal{R}_N$ are fair by comparing the actual distribution of recommendations to a target (uniform) distribution across different demographic groups.

2.1.1. Benefit types.

To provide a nuanced assessment of the benefits derived from recommendations, we implement two specific measures:

Hit ($\mathcal{B}_{hit}$). Measures whether the items in a recommendation list are relevant to the user. Specifically, the hit rate evaluates if any of the top k items recommended by the system appear in the ground truth list of user preferred items.

Ranking Quality ($\mathcal{B}_{rank}$). Assesses the alignment between the order of items in the recommendation list and their actual relevance to the user, as determined

by their position in the ground truth list. This metric indicates how effectively the recommendation system orders items in a way that corresponds to the user preferences.

These metrics serve as specific instances of $\mathcal{B}$ in our framework, allowing us to measure the practical benefits of the recommendations provided by different RecLLM scenarios.

3. Experiment

3.1. Setup

We conducted a series of experiments to evaluate the fairness and effectiveness of recommendations generated by our proposed RecLLM fairness evaluation framework. The experiments focused on two key aspects: (i) understanding the behavior of the model when no prior examples are available (0-shot learning) versus (ii) the effect of providing one (ICL-1) and two examples (ICL-2) in context for generating recommendations.

Table 2
Summary statistics of the LastFM and MovieLens datasets after filtering. $|R|$ represents the number of interactions, while $|R|/|U|$ denotes the average number of interactions per user $|U|$. The training and testing data statistics are shown for each dataset.

Dataset / Statistic	Training Data	Testing Data
MovieLens	$\|R\| = 16,757$ |R|/|U| = 209.46	$\|R\| = 4,230$ |R|/|U| = 52.88

The experiments utilize two datasets, LastFM and MovieLens, which offer a combination of music and movie recommendation tasks. However, due to space constraints, we report only the results for the MovieLens dataset. We follow the procedure outlined in [11, 7] to build a sequential recommender system focused on next-item prediction.

We evaluated the model performance based on fairness metrics related to gender and age group, using both Neutral vs. Sensitive Ranker Deviation (NSD) and Neutral vs. Counterfactual Sensitive Deviation (NCSD). These metrics help quantify how incorporating sensitive attributes, such as gender and age, affects the fairness and relevance of the recommendation system.

Our approach involves a sequential recommendation task where we employ timestamps to ensure the data is split temporally. Initially, we randomly select a subset of 80 users who exhibit a moderate level of interaction within the datasets. This allows us to handle the data efficiently while ensuring that the users selected have enough interactions to inform the training process but are not so many as to skew the representativeness of typical user behavior. The data for these users is then divided into training and test sets by sorting their interactions over time and splitting them such that 80% of a user's interactions are used for training, with the remaining 20% held out for testing. This method respects the chronological order of interactions, thereby simulating a realistic scenario where a model can only learn from past data to make predictions about future user behavior.

We assessed model performance across various conditions:

- **0-shot learning**: No examples of past recommendations are provided to guide the system.
- **ICL-1**: One example of past user interaction is provided to improve contextual understanding.
- **ICL-2**: Two examples are provided to further enhance recommendation relevance and fairness.

The main goal is to test the extent to which including sensitive attributes (gender and age) and providing in-context examples influences both the fairness and relevance of the recommendations generated by the model.

The recommendation generation was tested under different strategies for profile sampling stragies:

- **rand**: : Uniform selection of tracks or movies to provide a stochastic view of user preferences.
- **freq**: Prioritization of tracks/movies based on their frequency of playback, and rating provided emphasizing the main preferences of the user
- **rec-freq**: This hybrid approach combines the recency and frequency of track interactions using a weighted score formula.

To save space, we present a snapshot of the results as initial support for our framework using the MovieLens dataset. A more detailed extension will be provided later. Note that the recommendation scenario here is for *sequential item recommendation* task.

3.2. Results and Discussion

Table 3 shows the fairness evaluation results across various conditions. Key findings include:

Table 3
Key Fairness Results for Gender and Age Groups with NSD and NCSD, Including ΔB_1 and ΔB_2

NSD/Gender			
Condition	ΔB_1	ΔB_2	$\delta_{\text{gender}}(p\text{-value})$
0-shot/rand	-0.0074	-0.019	0.0116 (p=0.730)
ICL-1/rand	-0.0222	-0.0476	0.0254 (p=0.464)
ICL-2/rand	-0.037	-0.019	-0.018 (p=0.386)
0-shot/freq	0.0	-0.019	0.019 (p=0.609)
ICL-1/freq	0.0148	-0.0095	0.0243 (p=0.368)
ICL-2/freq	0.0148	-0.019	0.0339 (p=0.425)
NCSD/Age-Group			
Condition	ΔB_1	ΔB_2	$\delta_{\text{age-gr}}(p\text{-value})$
0-shot/rand	0.0	-0.0046	0.0046 (p=0.954)
ICL-1/rand	0.0	-0.0365	0.0365 (p=0.812)
ICL-2/rand	0.0952	-0.0228	**0.1181 (p=0.108)**
0-shot/freq	0.0476	-0.0091	0.0568 (p=0.307)
ICL-1/freq	0.0476	-0.0091	0.0568 (p=0.312)
ICL-2/freq	0.0	-0.0548	**0.0548 (p=0.022)**

- **Gender-Based Fairness (NSD).** Gender fairness was mostly stable across conditions, with minor deviations observed. The most noticeable case was under the **ICL-2/freq** condition, where the system slightly favored one gender group (i.e., males) ($\delta = 0.0339$, p = 0.425). While not statistically significant, this result suggests the model may introduce slight gender biases when more contextual examples are provided.
- **Age-Based Fairness (NCSD).** Age fairness showed more pronounced issues. Under the **ICL-2/freq** condition, the deviation was statistically significant

($\delta = 0.0548$, p = 0.022), indicating the system significantly favored one age group (Old) when two contextual examples were used. Similarly, **ICL-2/rand** displayed a notable deviation ($\delta = 0.1181$, p = 0.108), though it was not statistically significant.

- Impact of Contextual Information. As more contextual examples were introduced (moving from 0-shot to ICL-1 and ICL-2), deviations became more pronounced, particularly for age groups. This indicates that while context improves recommendation relevance, it can also exacerbate biases.

In conclusion, while the system demonstrates relatively strong performance in terms of gender fairness, concerns remain regarding age-based fairness, particularly in the **ICL-2/freq** condition. The experiments related to IF focus on measuring the fairness of consumers using sensitive attributes; details on these experiments will be provided in future work. Our primary objective here is not to present detailed experiments on all the elements used in this study (such as ICL, profile sampling type, or sensitive attributes), but rather to illustrate the effectiveness of the proposed framework through empirical analysis.

4. Conclusion

In this paper, we introduced a normative framework for benchmarking consumer fairness in large language model (LLM)-based recommender systems (RecLLMs), addressing the limitations of traditional fairness evaluations applied to collaborative filtering models. We provide a more formal and structured approach to auditing fairness by introducing key elements such as Neutral vs. Sensitive Ranker Deviation (NSD), Neutral vs. Counterfactual Sensitive Deviation (NCSD), and Intrinsic Fairness (IF). These metrics offer a principled way to assess fairness by clearly defining the reference point for fairness evaluation, whether it is a neutral ranker or a target distribution, and by quantifying fairness deviations through statistical tests. Additionally, we highlight the importance of specifying the underlying benefit types, such as hit rate and ranking quality, which provide a clear foundation for measuring fairness in relation to user preferences.

Our experiments on the MovieLens dataset demonstrate that while fairness remains stable in gender-based groups, age-based fairness deviations become more pronounced, especially when contextual examples are introduced (ICL-2). This suggests a potential amplification of biases when more contextual information is provided to the model. Future work should focus on refining these formal metrics, expanding the framework to cover more diverse sensitive attributes, and exploring further strategies to mitigate bias

References

[1] Y. Deldjoo, D. Jannach, A. Bellogin, A. Difonzo, D. Zanzonelli, Fairness in recommender systems: research landscape and future directions, User Modeling and User-Adapted Interaction 34 (2024) 59–108.

[2] M. D. Ekstrand, A. Das, R. Burke, F. Diaz, Fairness in information access systems, Found. Trends Inf. Retr. 16 (2022) 1–177. URL: https://doi.org/10.1561/1500000079. doi:10.1561/1500000079.

[3] Y. Deldjoo, Z. He, J. McAuley, A. Korikov, S. Sanner, A. Ramisa, R. Vidal, M. Sathiamoorthy, A. Kasirzadeh, S. Milano, A review of modern recommender systems using generative models (gen-recsys), in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 6448–6458.

[4] Z. He, Z. Xie, R. Jha, H. Steck, D. Liang, Y. Feng, B. P. Majumder, N. Kallus, J. McAuley, Large language models as zero-shot conversational recommenders, in: Proceedings of the 32nd ACM international conference on information and knowledge management, 2023, pp. 720–730.

[5] Y. Deldjoo, Z. He, J. McAuley, A. Korikov, S. Sanner, A. Ramisa, R. Vidal, M. Sathiamoorthy, A. Kasrizadeh, S. Milano, F. Ricci, Recommendation with generative models, arXiv preprint arXiv:2409.15173 (2024).

[6] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, X. He, Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation, in: Proceedings of the 17th ACM Conference on Recommender Systems, 2023, pp. 993–999.

[7] Y. Deldjoo, T. Di Noia, Cfairllm: Consumer fairness evaluation in large-language model recommender system, arXiv preprint arXiv:2403.05668 (2024).

[8] L. Boratto, G. Fenu, M. Marras, G. Medda, Consumer fairness in recommender systems: Contextualizing definitions and mitigations, in: European Conference on Information Retrieval, Springer, 2022, pp. 552–566.

[9] M. D. Ekstrand, A. Das, R. Burke, F. Diaz, Fairness in recommender systems, in: Recommender systems handbook, Springer, 2012, pp. 679–707.

[10] Y. Li, H. Chen, Z. Fu, Y. Ge, Y. Zhang, User-oriented fairness in recommendation, in: Proceedings of the web conference 2021, 2021, pp. 624–632.

[11] Y. Deldjoo, Understanding biases in chatgpt-based recommender systems: Provider fairness, temporal stability, and recency, ACM Transactions on Recommender Systems (2024). doi:/10.1145/3690655.