Fairness Matters: A look at LLM-generated group recommendations

Antonela Tommasel
ISISTAN, CONICET-UNCPBA
Tandil, Argentina
antonela.tommasel@isistan.unicen.edu.ar

ABSTRACT

Recommender systems play a crucial role in how users consume information, with group recommendation receiving considerable attention. Ensuring fairness in group recommender systems entails providing recommendations that are useful and relevant to all group members rather than solely reflecting the majority's preferences, while also addressing fairness concerns related to sensitive attributes (e.g., gender). Recently, the advancements on Large Language Models (LLMs) have enabled the development of new kinds of recommender systems. However, LLMs can perpetuate social biases present in training data, posing risks of unfair outcomes and harmful impacts. We investigated LLMs impact on group recommendation fairness, establishing and instantiating a framework that encompasses group definition, sensitive attribute combinations, and evaluation methodology. Our findings revealed the interaction patterns between sensitive attributes and LLMs and how they affected recommendation. This study advances the understanding of fairness considerations in group recommendation systems, laying the groundwork for future research.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Social and professional topics** → **Sustainability**.

KEYWORDS

fairness, recommender systems, group recommendation, Large Language Models, evaluation

ACM Reference Format:
Antonela Tommasel. 2024. Fairness Matters: A look at LLM-generated group recommendations. In *18th ACM Conference on Recommender Systems (RecSys '24), October 14–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3640457.3688182

1 INTRODUCTION

Group recommenders aim to provide recommendations to groups of users, acknowledging their diverse preferences [27]. Ensuring that recommendation outcomes align with the needs of everyone in the group is crucial for fairness, especially when users have distinct sensitive attributes (e.g., demographic backgrounds). These attributes add complexity to the task, as factors like age, culture, gender, language, and socioeconomic status can significantly influence preferences and the acceptance of recommendations. While utility group fairness has received considerable attention [27], fairness concerning sensitive attributes is often overlooked, even when sensitive attributes have been shown to significantly influence standard user-item recommendations [8].

LLMs represent a paradigm shift towards leveraging advanced NLP techniques to improve recommendations [12], offering the promise of increased personalization, intuitiveness, and interpretability. Despite their benefits, concerns have been raised about their inherent biases [4, 5, 15], potentially leading to unfair treatment of certain user groups. Although users might opt not to disclose their sensitive attributes due to privacy concerns when interacting in different group settings [18] and with LLMs [29], LLMs could inadvertently create biases due to their preferences for specific sensitive attributes based on their training data [21, 29]. Thus, fairness is a critical criterion for recommendations due to their significant social impact [5], as they could perpetuate existing stereotypes or economic disparities, among others.

While fairness regarding sensitive attributes in traditional recommendation systems has been extensively studied, fairness in LLM-generated recommendations has received comparatively less attention [14, 29]. In this work, we present a study of *how LLM-generated group recommendations are affected by users' sensitive attributes*. To this end, we covered group definition, sensitive attribute combinations and evaluation. Using textual prompts, we generated group recommendations, and evaluated them across different sensitive attributes. Our study, instantiated in a movie recommendation task, demonstrates that the presence of sensitive attributes not only alters recommendations, but also results in unique effects for each combination of LLM and attribute.

The contributions of this study can be summarized as follow: to the best of our knowledge, this is one of the first studies related to fairness towards sensitive attributes in group recommendation settings involving LLM-based recommendations, the introduction of an evaluation framework for group fairness of LLM-based recommendations based on alignment, quality and consistency aspects of recommendations, and the evaluation of the intersectionality of sensitive attributes. This study aims at advancing the understanding of fairness considerations in group recommenders, laying the groundwork for future research.

2 RELATED WORK

Fairness and bias have long been areas of interest, yet research on fairness in LLMs for recommendation is still preliminary [9],

RecSys '24, October 14–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0505-2/24/10
https://doi.org/10.1145/3640457.3688182

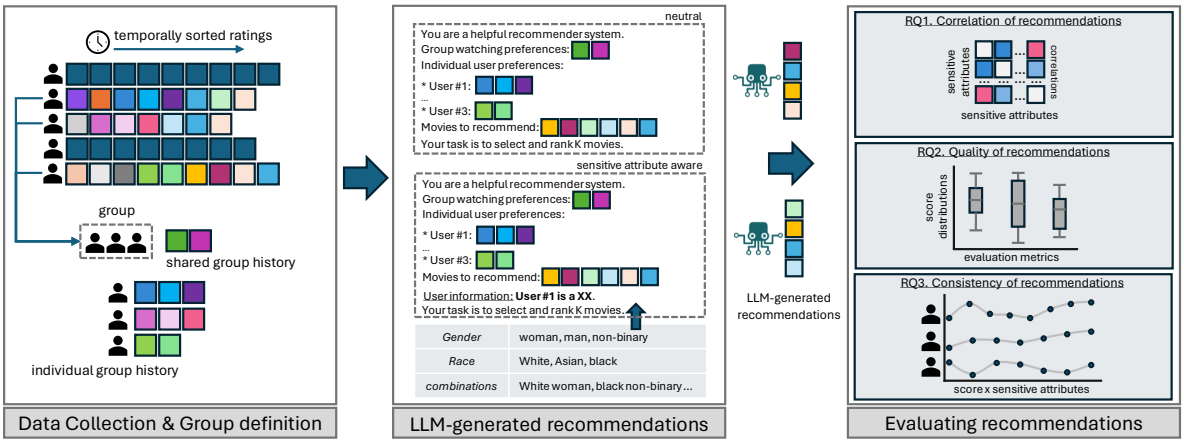

Figure 1: Schematic steps of the proposed framework

in particular in relation to group recommendation tasks. Fairness in group recommendation tasks has typically focused on balancing user utilities to minimize user dissatisfaction [16, 27] or on assessing users' fairness perception [2, 22]. However, this often overlooks user-sensitive attributes. Research on standard user-item recommendation has indicated that sensitive attributes can significantly influence recommendations, even when excluded from the model, as other characteristics may encode such information [8, 29]. We argue that sensitive attributes can also significantly influence group recommendations. On one hand, they can lead to uneven performance distributions, thus fostering discrimination and limiting diversity [3, 8]. On the other hand, these differences may represent an improved adaptation or personalization to user interests [4], correcting inherent biases towards attributes and stereotypes derived from their training data [29].

Related to this study, we find the works of Zhang et al. [29], and Deldjoo and Di Noia [4], who assessed user fairness in recommendations tasks. Zhang et al. [29] evaluated the fairness of zero-shot LLM recommendations using Jaccard Similarity to compare rankings with and without 8 independent sensitive attributes. Evaluation, based on GPT, revealed unfairness. The authors argued that traditional fairness metrics are inadequate for evaluating LLM recommendations, as requiring model predictions restricts LLMs outputs to fit into a fixed candidate set. However, ignoring model prediction quality assumes all differences are unfair, neglecting potential personalization effects. Deldjoo and Di Noia [4] expanded on Zhang et al. [29] by evaluating the fairness of recommendations with combinations of two sensitive attributes. Unlike Zhang et al. [29], they included a list of movies watched by the user in the prompt, but no fixed list of potential recommendations. Instead, the authors processed the obtained lists to identify movie titles matching those in their data, although the overlap between recommendations and the dataset was not detailed.

3 TASK AND METHOD

Unlike prior research, we investigate *fairness in group recommendations through a user-sensitive lens*, considering both user preferences and sensitive attributes. Usually, fairness is analyzed either from an individual or group perspective. Particularly, group fairness requires that identified groups are treated equally, while individual fairness requires equal treatment of similar individuals. In group recommendations, we extend this concept to "*individual fairness in a group context*", in which there should not be any prejudice or favoritism towards individuals with specific sensitive attributes, involving not only a uniform distribution of utility across groups, but also that recommendations also reflect the interests and preferences of all users. To this end, we propose the framework in Figure 1, addressing group formation, recommendation generation in both sensitive-aware and -unaware (neutral) scenarios, and recommendation evaluation. We exercise the framework with 3 research questions:

- **RQ1**: *Do sensitive attributes impact recommendations?* Rather than focusing on the similarity of recommendation lists, we examine the correlation between recommendations aware and unaware of sensitive attributes.
- **RQ2**: *How are recommendations affected by sensitive attributes?* We evaluate if recommendation quality varies based on sensitive attributes.
- **RQ3**: *Do users in a group receive consistent treatment across sensitive attributes?* We explore whether LLMs tend to favour specific attributes, regardless of the user identified with them.

3.1 Data collection & Group definition

Our evaluation used the *MovieLens* dataset[1], which includes users, movies, ratings, genres, and timestamps. No sensitive attributes are included. To mitigate the risk of user interests shifting due to extensive watching histories, we only considered ratings created after 1st January 2016. We selected users who rated more than 30 movies, resulting in 29, 283 users and 56, 719 movies. Given the temporal nature of ratings, after groups are created, users' watching histories were partitioned into train/test sets based on timestamps, with the first 80% of ratings as training and the remaining 20% as test. Movies were considered liked if rated 4 or higher.

Since the collected data lacked group membership information, we created synthetic groups[2]. Following prior works [16, 27], we simulated groups of 2 to 8 members with similar watching histories

[1] https://grouplens.org/datasets/movielens/25m/
[2] Code can be found in the companion repository at: https://github.com/tommantonela/fairness-group-llm-recommendations.

as a proxy for similar interests. As in previous works [1, 16], we computed user similarity using the Pearson Correlation Coefficient, ensuring each pair's similarity exceeded a set threshold. Following [1, 28] (among others), similarities were computed based on the full data collection. The average group similarity was 0.23, with each pair of users rating at least 9 common movies (median 22), ensuring reliability [1]. Other group creation strategies, like considering divergent or random interests, could also be considered.

3.2 LLM-generated recommendations

Sensitive attributes. We chose two sensitive attributes: gender (*man*, *non-binary* and *woman*) and race (*Afro-American*, *Asian*, *white*). Most fairness literature focuses on a single protected group per user, missing the importance of intersectionality, by which biases can be amplified in subgroups combining multiple categories, especially in historically underrepresented groups [11]. Therefore, we explored combinations of these attributes. As the dataset does not include sensitive attribute information, sensitive attributes were assigned to users in a controlled manner, with each group member associated once to each sensitive attribute combination.

Making recommendations. Using LLMs, recommendations can be framed as prompt-based tasks, integrating user information and watching history into personalized prompts [10]. Unlike [29], we defined a structured prompt to capture group preferences, individual user preferences, and sensitive attributes. LLMs were tasked with acting as helpful recommender systems. The prompts were structured as follows[3]:

1) **Group watching history**. A list of movies liked by all members of the group, providing insight into common group interests.

2) **User individual watching history**. For each user, a list of titles exclusively liked by them, showcasing their specific interests.

3) **User sensitive attribute**. A description of a user identified with a sensitive attribute (e.g., "User XX is a white woman"). This part was omitted for neutral recommendations.

4) **Movies to recommend**. Unlike [4, 29], we provided the LLM with a set of movies to choose from. This set included all movies in the group test set liked by all members, at least 5[4] movies exclusively liked by each member (movies liked by one group member that were either disliked or not watched by any of the other group members), and 5 movies disliked by each member (i.e., rated below 4 by a member, and not liked by any other member in either training/test). Although this may constrain recommendations [29], it prevents suggesting movies users might not know about (e.g., premiered after users' ratings) and allows for better preference control towards specific users or sensitive attributes. Movies were sorted in alphabetical order.

5) **Task to solve**. The LLM was tasked with ranking k movies from "movies to recommend" that would satisfy the entire group. No additional information (e.g., explanations) was requested.

3.3 Evaluating recommendations

We generated recommendations for each group in both sensitive-aware and neutral (or sensitive-unaware) scenarios. In the neutral scenario, users are not associated with any sensitive attribute, while in the sensitive-aware scenario, one user at a time is assigned one of the 15 possible sensitive attribute combinations. For example, for groups with 2 members, 31 different recommendation lists are generated. We focused on three evaluation aspects, outlined below.

Correlation of recommendations. We evaluated the alignment between neutral recommendations and those including sensitive attributes by computing their rank correlations using Kendall's Tau. For the comparison, we considered the specific rank of each movie in the various recommendation rankings. Low correlations could suggest unfairness, indicating recommendations that might reflect stereotypical assumptions about preferences.

Quality of recommendations. Our goal was to determine if incorporating sensitive attributes improved recommendation quality. To this end, we analyzed the distribution of recommendation quality across sensitive attributes for each group. We focused on relevance metrics: precision and nDCG. Group-level results were derived by summarizing user results using metrics such as minimum, mean, median, variance, or maximum scores [19, 20, 23]. Scores were defined for each sensitive attribute at two levels: within each group and across groups. Each sensitive attribute generated multiple group configurations, each with its own score, which were summarized to obtain the group-level score[5]. Finally, the group-level scores were summarized to compute the score across groups.

We compared scores for each sensitive attribute across user groups to determine statistical significance, evaluating whether any attribute consistently led to lower or higher recommendation quality, potentially indicating unfairness. Paired samples tests (with $\alpha = 0.01$) were used, including corrections for multiple tests.

Consistency of recommendations. We assessed the consistency of results for users in a group across sensitive attributes to determine if attributes contributed to improved personalization or led to unfairness, and whether users received similar treatment when they or another group member identified with a sensitive attribute.

4 ANALYSIS

We evaluated 3 general-domain LLMs: `Mistral:7b`, `GPT-3.5-turbo` and `Gemma:2b-instruct`[6]. Given the experiment complexity, we ran the full experiment set over 1000 groups with `Gemma` and a sample with `Mistral` and `GPT`[7]. While defining the "best LLM-based group recommender" is not the main focus of this study, we believe it is important to contextualize the quality of recommendations (i.e., check whether LLM-based recommendations are at an "acceptable" level). To this end, we considered *ImplicitMF* [13], a high-performing matrix factorization method, unaware of sensitive attributes, and adapted to group recommendation through preference aggregation [17].

4.1 RQ1. Correlation of recommendations

Based on a meta-analysis of correlations [24, 25], Figure 2 shows the median group correlations between recommendations. To ensure

[3]The full prompt (not in Figure 1) can be found in the companion repository.

[4]A larger number of movies could be considered. The limit was selected to avoid surpassing LLMs' token limits.

[5]See the supplementary information for a calculation example.

[6]`Llama` models were excluded due to excessive execution time on available hardware.

[7]Due to space restrictions, we report only the `Gemma` evaluations for 1000 groups of size 5, with additional results and analyses available in the companion repository.

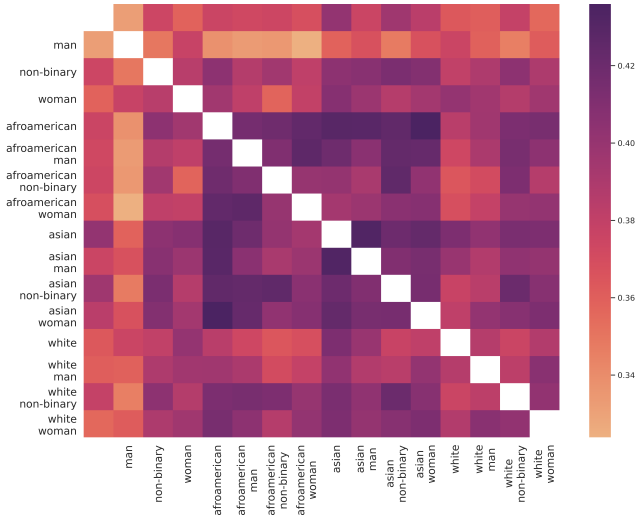

Figure 2: Correlations between group recommendations across sensitive attributes

comparability, we compared rankings for the same user identified with different sensitive attributes (e.g., recommendations when user *A* identified as an *Afro-Am-woman* versus an *Asian-man*), ensuring the only change in LLMs' input was such attribute.

Sensitive attributes generally showed lower correlations with neutral recommendations than with other attributes, a trend consistent across different LLMs. *Asian-any*[8] and *Afro-Am-any* attributes had higher correlations with each other. This suggests a moderating effect of cross attributes [4, 11]. `Mistral` and `GPT` exhibited similar correlation patterns, with *Afro-Am-any* showing distinctively lower correlations with all other attributes.

We also analyzed the correlations for the same sensitive attribute across groups. Correlations were not perfect, with the highest scores for *Afro-Am-only*, *Asian-only*, and *non-binary-only*, and the lowest for *man-only*. This suggests that LLMs do not react uniformly to sensitive attributes for every user but still consider individual interests when making recommendations.

In summary, correlations indicated that sensitive attributes influence LLMs' recommendations. Correlations varied according to the considered attribute and LLM.

4.2 RQ2. Quality of recommendations

Figure 3 shows the distribution of maximum variances (which can indicate unfairness [20, 27]) for precision[9]. `Gemma` achieved similar results to the best *ImplicitMF* results (obtained using additive aggregation), demonstrating its suitability for the task.

Most differences in Figure 3 are statistically significant, with larger effects for gender- and race-only attributes. *White-any* and *Asian-any* achieved significantly higher results compared to *Afro-Am-any*. This could indicate not only a difference in how LLMs treat sensitive attributes but also unfair treatment of group members.

The distribution of mean precision/nDCG scores showed that neutral recommendations had significantly higher minimum scores compared to sensitive-aware recommendations. However, for maximum mean scores, sensitive-aware recommendations outperformed neutral ones. This suggests that incorporating sensitive attributes can enhance the overall quality of recommendations. However, the increased variance indicates that sensitive attributes can disrupt the balance of neutral recommendations, potentially leading to unfairness by not balancing group preferences.

Different interactions between sensitive attributes were observed. Generally, *Asian-any* and *Afro-Am-any* outperformed *white-any*. This indicates that LLMs respond differently to attributes, with race playing a significant role. Despite the limited set of movies to recommend, stereotyped assumptions about interests linked to sensitive attributes were partially reflected. For instance, `Mistral`'s responses included comments on how recommended movies aligned with the lives and experiences associated with specific attributes. The lower score distributions and higher variances for *white-any* could relate to the LLM consciously avoiding prioritizing a commonly over-represented attribute. Further studies with a larger list of options or allowing the LLM to choose recommendations freely are needed to assess this phenomenon effectively.

Significant differences were observed among the three selected LLMs. `Mistral` and `GPT` achieved significantly lower results than `Gemma`, with differences up to 96% and 180% for neutral precision, respectively. The largest differences were for recommendations involving *Afro-Am-any* and *Asian-any*. For example, with `GPT`, *white-any* scores were higher than *Asian-man/woman* and *Afro-Am-man/woman* attributes, and *Afro-Am-man* and *non-binary-any* had larger variances. For `Mistral`, only a few significant differences were observed for the mean and maximum score distributions.

In summary, results confirm that sensitive attributes affect recommendation quality, highlighting LLMs' distinct preferences for certain attributes and the challenge of balancing assumptions about these attributes and group interests.

4.3 RQ3. Consistency of recommendations

To assess the effect of attributes at the user level, we compared the precision and nDCG scores of users in a group identifying with sensitive attributes. We observed that these users generally had lower scores than the maximum mean scores for such attribute, indicating that they did not always receive the best recommendations.

Table 1 summarizes how users' precision scores changed compared to neutral recommendations in two scenarios[10]: i) identifying with the sensitive attribute, and ii) another group member identifying with the attribute. While group results suggested that including sensitive attributes could improve performance, this trend was not predominant at the individual level. On average, scores improved for 8% of users in a group, with at least 50% of groups seeing no improvement. On average, 6% to 9% of users experienced worse scores due to identifying with a sensitive attribute. *Men-only*, *white-only* (usually the most represented attributes), and *Afro-Am-any* (usually among the least represented attributes) had the highest proportion of users with worse scores. In the second scenario, 13%

[8]We will use "*any*" as a replacement for attribute combinations. For example, in this case, *Asian-any* replaces *Asian-man*, *Asian-woman* and *Asian-non-binary*.
[9]Results summarizing precision and nDCG across groups for each sensitive attribute can be found in the supplementary material in the repository.
[10]Results for nDCG can be found in the supplementary material in the repository.

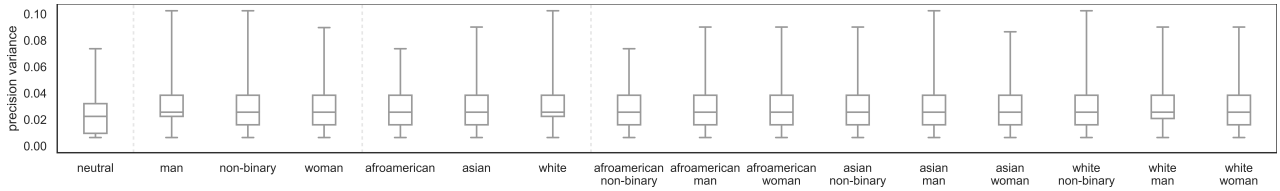

Figure 3: Distribution of maximum variance scores across sensitive attributes $k = 5$

	user w/Sens Attr		other w/Sens Attr	
	incr	decr	incr	decr
man	0.088±0.132	0.115±0.128	0.143±0.167	0.203±0.164
non-binary	0.085±0.135	0.109±0.132	0.133±0.169	0.18±0.155
woman	0.087±0.134	0.104±0.14	0.133±0.164	0.186±0.161
Afro-Am	0.089±0.14	0.106±0.124	0.134±0.168	0.165±0.155
Asian	0.089±0.137	0.106±0.13	0.139±0.175	0.166±0.15
white	0.086±0.136	0.11±0.13	0.142±0.173	0.178±0.151
Afro-Am-man	0.09±0.142	0.112±0.13	0.132±0.168	0.181±0.159
Afro-Am-non-bin	0.082±0.133	0.118±0.132	0.133±0.168	0.175±0.157
Afro-Am-woman	0.085±0.136	0.123±0.144	0.132±0.163	0.176±0.156
Asian-man	0.085±0.138	0.116±0.134	0.138±0.172	0.178±0.153
Asian-non-bin	0.083±0.136	0.115±0.132	0.13±0.167	0.17±0.155
Asian-woman	0.085±0.138	0.11±0.136	0.134±0.169	0.175±0.155
white-man	0.086±0.134	0.115±0.137	0.137±0.164	0.194±0.163
white-non-binary	0.086±0.139	0.108±0.13	0.138±0.169	0.182±0.157
white-woman	0.083±0.13	0.114±0.131	0.138±0.171	0.183±0.155

Table 1: Variation of user precision across sensitive attributes

to 14% of users improved their scores, whereas 12% to 17% saw their scores decrease. Thus, score reductions were more frequent when another group member identified with a sensitive attribute. Finally, on average, for 11% and 23% of groups at least one user identifying with a sensitive attribute saw their score changed and became the user with the minimum or maximum score in the group.

For GPT, the proportion of users whose scores improved in the first scenario ranged, on average, between 9% and 14%, while 7% to 15% saw their scores decrease. Following the overall group trends, *white-man* and *white-only* attributes improved the scores of a larger proportion of users, while *Asian-man* for the smallest. Similarly, *white-man* and *white-non-binary* showed the smallest proportions of users with worse scores, while *Afro-Am-any* the largest. As seen with Gemma, in the second scenario, the proportions of users both improving and worsening their scores increased, with *Afro-Am-any* showing the largest proportions in both cases.

Overall, observations suggest unfair user treatment. This unfairness is reflected in that users identifying with an attribute often experienced worse scores. Moreover, an even larger proportion of users experienced worse scores when another group member identified with an attribute. Since identifying with a sensitive attribute does not consistently lead to better outcomes, it appears that these differences result from attempts to match users with stereotyped assumptions of their interests. When these assumptions do not align with user's actual interests, it worsens their results.

In summary, while including sensitive attributes can improve overall group recommendations, they may lead to an inconsistent and unfair treatment by worsening recommendations for users identifying with those attributes, while also negatively affecting other group members.

5 CONCLUSIONS

Our study highlights the different preferences of LLMs for specific attributes and the challenge of balancing assumptions over sensitive attributes and the group's overall interests, resulting in unfair treatment.

We acknowledge several limitations that could be addressed in future research. Firstly, the examination was limited to a specific set of LLMs and parameters, such as temperature, decoding strategy and a defined prompt. These factors can affect the generalizability of findings, as prompt formulation influences response toxicity and inherent biases, particularly regarding gender and race [7], while the decoding strategy can influence the fluency and coherence of responses. Secondly, the study focused solely on the commonly used *MovieLens* collection [16, 26, 27]. While this choice highlighted the framework's applicability, the framework has potential for diverse applications, and thus evaluating other domains would help assess the generalizability of the results. For example, evaluation could include Zhang et al. [30]'s LastFM dataset, which includes social relations between users and would allow considering real user groups/communities. Thirdly, including additional sensitive attributes, as examined in [29], group sizes [6] and group creation strategies could enable a more comprehensive fairness analysis.

ETHICAL STATEMENT

References to sensitive attributes are included solely to study fairness. We acknowledge that we have not explored the whole set of variations within the chosen categories. Although LLMs can improve recommendation accuracy and user experience, they also raise concerns about the reinforcement and amplification of existing biases, data privacy (users must be informed about how their data is collected, stored and used), transparency and informed consent (users should give explicit consent before sharing their data, and should have the option to opt-out or share partial data), legal compliance (any sharing of sensitive data should comply with regulatory frameworks, such as GDPR).

Privacy concerns should also be considered in terms of group formation [18], as users in tightly coupled homogeneous groups usually perceive lower privacy concerns compared to those in loose heterogeneous groups, leading to more disclosure. Nonetheless, involving an LLM might alter users' perceptions, and thus change their willingness to share information. Ensuring transparency, accountability, and fairness in LLM-based recommendation systems is essential to mitigate these ethical aspects and maintain user trust in a healthy environment.

REFERENCES

[1] Linas Baltrunas, Tadas Makcinskas, and Francesco Ricci. 2010. Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the*

fourth ACM conference on Recommender systems. 119–126.
[2] Francesco Barile, Tim Draws, Oana Inel, Alisa Rieger, Shabnam Najafian, Amir Ebrahimi Fard, Rishav Hada, and Nava Tintarev. 2024. Evaluating explainable social choice-based aggregation strategies for group recommendation. *User Modeling and User-Adapted Interaction* 34, 1 (2024), 1–58.
[3] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.
[4] Yashar Deldjoo and Tommaso Di Noia. 2024. CFaiRLLM: Consumer Fairness Evaluation in Large-Language Model Recommender System. *arXiv preprint arXiv:2403.05668* (2024).
[5] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2024. Fairness in recommender systems: research landscape and future directions. *User Modeling and User-Adapted Interaction* 34, 1 (2024), 59–108.
[6] Amra Delic, Julia Neidhardt, Thuy Ngoc Nguyen, and Francesco Ricci. 2018. An observational user study for group recommender systems in the tourism domain. *Information Technology & Tourism* 19 (2018), 87–116.
[7] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv:2304.05335* (2023).
[8] Michael D Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, et al. 2022. Fairness in information access systems. *Foundations and Trends® in Information Retrieval* 16, 1-2 (2022), 1–177.
[9] Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046* (2023).
[10] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-REC: Towards Interactive and Explainable LLMs-Augmented Recommender System. *arXiv preprint arXiv:2303.14524* (2023).
[11] Avijit Ghosh, Lea Genuit, and Mary Reagan. 2021. Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*. PMLR, 22–34.
[12] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*. Springer, 364–381.
[13] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*. Ieee, 263–272.
[14] Wenyue Hua, Yingqiang Ge, Shuyuan Xu, Jianchao Ji, and Yongfeng Zhang. 2023. Up5: Unbiased foundation model for fairness-aware recommendation. *arXiv preprint arXiv:2305.12090* (2023).
[15] Meng Jiang, Keqin Bao, Jizhi Zhang, Wenjie Wang, Zhengyi Yang, Fuli Feng, and Xiangnan He. 2024. Item-side Fairness of Large Language Model-based Recommendation System. In *Proceedings of the ACM on Web Conference 2024*. 4717–4726.
[16] Mesut Kaya, Derek Bridge, and Nava Tintarev. 2020. Ensuring fairness in group recommendations by rank-sensitive balancing of relevance. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 101–110.
[17] Judith Masthoff. 2010. Group recommender systems: Combining individual models. In *Recommender systems handbook*. Springer, 677–702.
[18] Shabnam Najafian, Geoff Musick, Bart Knijnenburg, and Nava Tintarev. 2023. How do people make decisions in disclosing personal information in tourism group recommendations in competitive versus cooperative conditions? *User Modeling and User-Adapted Interaction* (2023), 1–33.
[19] Ladislav Peska and Ladislav Malecek. 2021. Coupled or Decoupled Evaluation for Group Recommendation Methods?. In *Perspectives@ RecSys*.
[20] Dimitris Sacharidis. 2019. Top-n group recommendations with fairness. In *Proceedings of the 34th ACM/SIGAPP symposium on applied computing*. 1663–1670.
[21] Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. 2018. Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity. In *Proceedings of the 2018 World Wide Web Conference*. 923–932.
[22] Thi Ngoc Trang Tran, Müslüm Atas, Alexander Felfernig, Viet Man Le, Ralph Samer, and Martin Stettinger. 2019. Towards social choice-based explanations in group recommender systems. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. 13–21.
[23] Christoph Trattner, Alan Said, Ludovico Boratto, and Alexander Felfernig. 2023. Evaluating group recommender systems. In *Group recommender systems: an introduction*. Springer, 63–75.
[24] Robbie CM van Aert. 2023. Meta-analyzing partial correlation coefficients using Fisher's z transformation. *Research Synthesis Methods* 14, 5 (2023), 768–773.
[25] David A Walker. 2003. JMASM9: converting Kendall's tau for correlational or meta-analytic analyses. *Journal of Modern Applied Statistical Methods* 2 (2003), 525–530.
[26] Wen Wang, Wei Zhang, Jun Rao, Zhijie Qiu, Bo Zhang, Leyu Lin, and Hongyuan Zha. 2020. Group-aware long-and short-term graph representation learning for sequential group recommendation. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 1449–1458.
[27] Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. 2017. Fairness-aware group recommendation with pareto-efficiency. In *Proceedings of the eleventh ACM conference on recommender systems*. 107–115.
[28] Quan Yuan, Gao Cong, and Chin-Yew Lin. 2014. COM: a generative model for group recommendation. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 163–172.
[29] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 993–999.
[30] Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip S Yu. 2015. Cosnet: Connecting heterogeneous social networks with local and global consistency. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1485–1494.