Supplementary information: Fairness Matters: A look at LLM-generated group recommendations

Anonymous Author(s)

ABSTRACT

This document includes additional methodology details, Tables with results and charts. The companion repository can be found at: https://bit.ly/3WNYAwx.

ACM Reference Format:
Anonymous Author(s). 2024. *Supplementary information:* Fairness Matters: A look at LLM-generated group recommendations. In . ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 TASK AND METHOD

1.1 LLM-generated recommendations

Sensitive attributes. We explored the combinations of sensitive attributes presented in Table 1.

neutral	
gender-only	man, non-binary, woman
race-only	Afro-American, Asian, white
combinations	Afro-American-man, Afro-American-non-binary, Afro-American-woman Asian-man, Asian-non-binary, Asian-woman white-man, white-non-binary, white-woman

Table 1: Combinations of sensitive attributes

Using LLMs, recommendations can be framed as prompt-based tasks, integrating user information and watching history into personalized prompts [2]. Unlike [8], 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:

1) **Group watching history**
2) **User individual watching history**.
3) **User sensitive attribute**.
4) **Movies to recommend**.
5) **Task to solve**.

To ease processing, we required responses to be in JSON format. After several prompt iterations, we instructed the LLM to refrain from adding any extra information to the response and to base recommendations solely on the provided movie list. Additionally, it was clarified that the movies to recommend were presented in alphabetical order, with no indication of group preferences.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

The system prompt was defined as follows:

> *You are a helpful movie recommender system. Your task is to recommend movies to a group of people based on their watching history.*
> *You will receive:*
> - The group preferences.
> - The individual user preferences.
> - User information (if available).
> - The set of movies to recommend.
>
> *Your task is to use all the provided information to generate a list of recommended movies. You have access to all the information you need.*

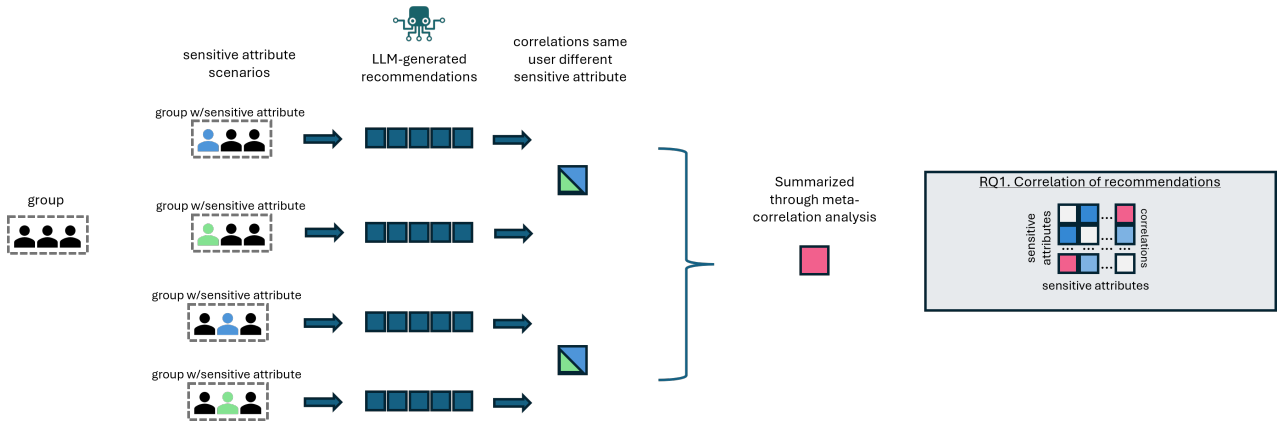
This is the prompt describing the task used for the evaluation, including the modification for the neutral recommendations.

> `{intersection}` (1)
> *These are the "individual user preferences":* `{users_history}` (2)
> `{stereotype}` (3, optional)
> *Movies to recommend:* `{to_recommend}` (4)
> *Your task is:* (5)
> *1. Using the "group preferences" and the "individual user preferences", pick 10 movies from "movies to recommend" and sort them based on how well they would satisfy the group as a whole. Position 1 should be the movie that best satisfies the group. Please, use only the movies in the list, do not add any additional movie. Do not change the given movies. Do not change the given titles.*
> *2. Return your answer in a JSON format including the key 'movies' and the list with the ranked movies.*
> *All the information you need is available in this conversation. Note that "movies to recommend" is alphabetically sorted, and that order those not reflect the group preferences. Use only "movies to recommend". Do not add any extra movie.*
> *Your JSON answer:*

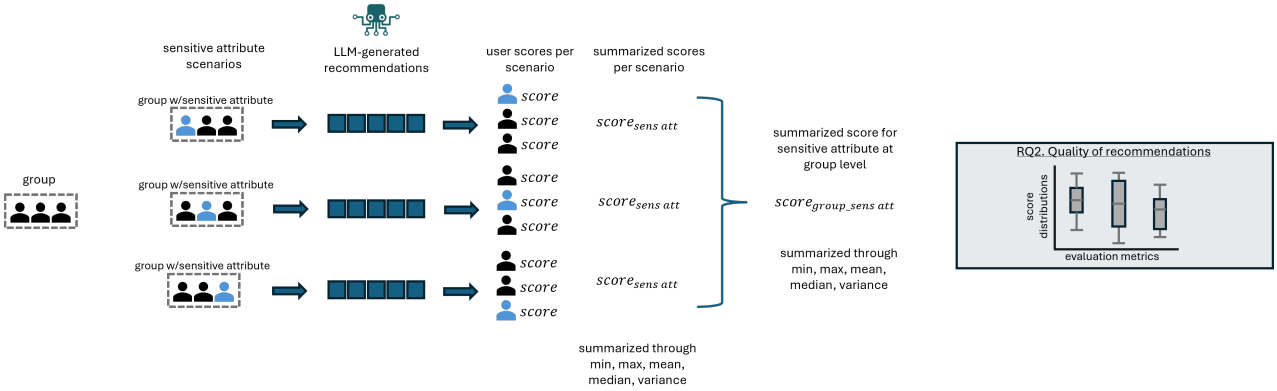
1.2 Evaluating recommendations

We generated recommendations for each group in both neutral and sensitive-aware 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. The number of recommendation lists increases with group size.

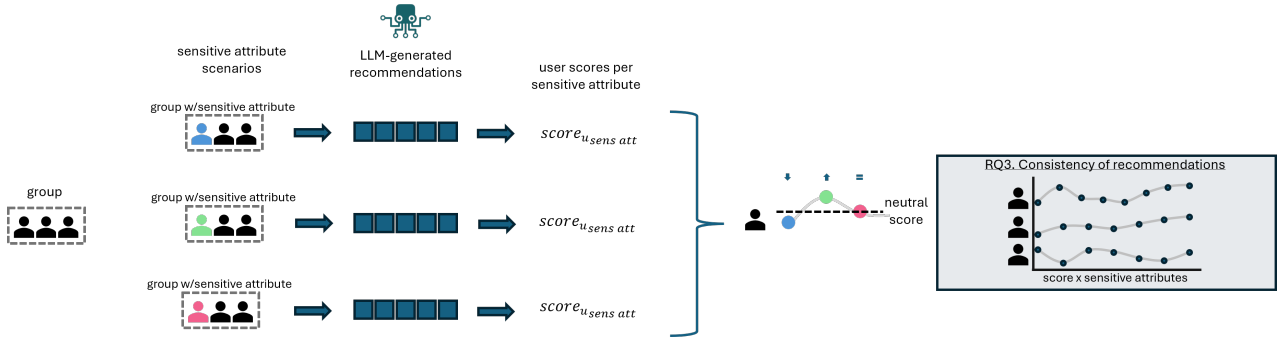
Correlation of recommendations. Based on a meta-analysis of correlations [5, 6], Figure 1 shows the median group correlations between



(a) Example of correlation computation



(b) Example of group score computation across sensitive attributes

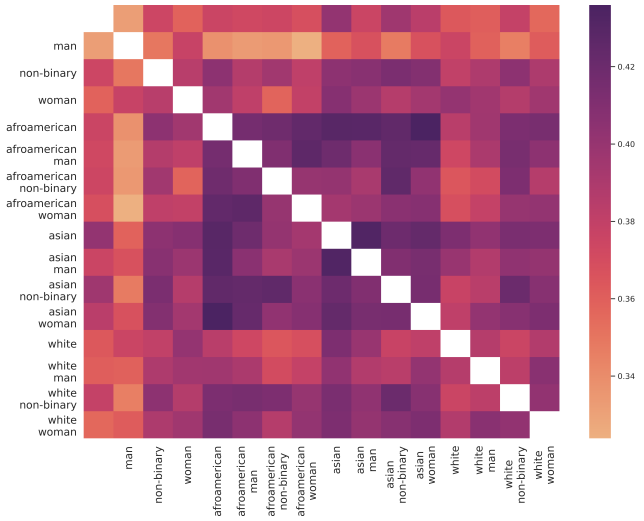


(c) Example of user score building across sensitive attributes

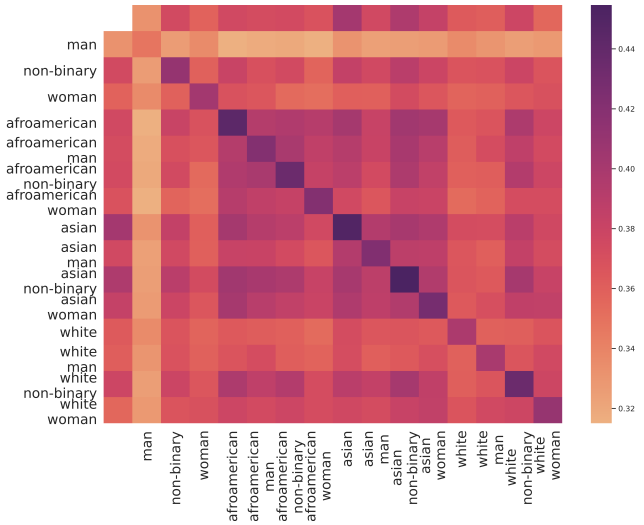
recommendations. To ensure comparability, we compared rankings for the same user identified with different sensitive attributes (e.g., recommendations when user *A* identified as an *Afro-Am-woman* versus an *Asian-man*), ensuring the only change in LLMs' input was such attribute. Figure **??**(a) exemplifies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

Quality of recommendations. 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. Figure **??**(b) exemplifies this summarization. Finally, the group-level scores are summarized to compute the score across groups.

Consistency of recommendations. Figure **??**(c) exemplifies how scores were prepared to analyze the consistency of results for users in a group across sensitive attributes to determine if attributes contributed to improved personalization or led to unfairness.



(a) Correlation for different attributes with the same user identifying with it



(b) Correlations for same attribute, regardless of the user identifying with it (diagonal)

Figure 1: Correlations between group recommendations across sensitive attributes

2 ANALYSIS

2.1 RQ1. Correlation of recommendations

Based on a meta-analysis of correlations [5, 6], Figure 1 shows the median group correlations between recommendations for groups of 5 members. To ensure comparability across sensitive attributes, for any pair of sensitive attributes, we compared the rankings obtained when the same user was identified with the sensitive attributes (e.g., we compared the group recommendations when user A identified as an *Afro-American-woman* with the recommendations when user A identified as an *Asian-man*), ensuring that the only change in the recommender's input was the identified attribute.

	min		mean		max	
	prec	nDCG	prec	nDCG	prec	nDCG
ImplicitMF	0.204	0.460	0.204	0.460	0.204	0.460
neutral	0.167	0.377	0.167	0.377	0.167	0.377
man	0.144	0.334	0.165	0.372	0.183	0.406
non-binary	0.149	0.347	0.167	0.381	0.183	0.408
woman	0.147	0.342	0.166	0.377	0.182	0.407
Afro-Am	0.151	0.355	0.168	0.385	0.184	0.411
Asian	0.154	0.354	0.171	0.386	0.186	0.412
white	0.15	0.345	0.168	0.38	0.184	0.41
Afro-Am man	0.149	0.35	0.167	0.383	0.183	0.409
Afro-Am non-binary	0.149	0.352	0.167	0.382	0.182	0.409
Afro-Am woman	0.148	0.351	0.167	0.382	0.183	0.409
Asian man	0.151	0.35	0.169	0.383	0.184	0.41
Asian non-binary	0.151	0.356	0.169	0.385	0.183	0.409
Asian woman	0.15	0.351	0.169	0.383	0.184	0.41
white man	0.145	0.341	0.165	0.376	0.183	0.406
white non-binary	0.149	0.35	0.167	0.382	0.183	0.408
white woman	0.146	0.345	0.165	0.379	0.182	0.408

Table 2: Summarized mean group recommendation results across sensitive attributes $k = 5$

In general, sensitive attributes showed lower correlations with neutral recommendations than with other attributes. This trend was consistent across the different group sizes and LLMs. *Afro-American-any*[1] and *Asian-any* attributes had higher correlations with each other. This suggests a particular moderating effect of cross attributes [1, 3].

`Mistral` and `GPT` exhibited similar correlation patterns. *Afro-American-any* and *Asian-any* had high correlations with each other, with *Afro-American-any* showing distinctively lower correlations with all other attributes.

Finally, we also analyzed the correlation between the recommendations obtained for the same sensitive attribute across groups (the diagonal in Figure 1(b)). Correlations were not perfect, with the maximum scores observed for *Afro-American-only*, *Asian-only* and *non-binary-only*, while *man-only* had the lowest scores. This implies that LLMs do not react to the sensitive attributes in the same way for every user (as correlations should have been perfect), but still consider users' interests for making the recommendations.

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

2.2 RQ2. Quality of recommendations

For each sensitive attribute, Table 2 presents the summarization of precision and nDCG across groups, while Figure 2 shows the distribution of maximum variances (which can indicate the presence of unfairness [4, 7]. *ImplicitMF* results were obtained using the additive aggregation method. As the table shows, `Gemma`'s results are similar to those of *ImplicitMF*, demonstrating its suitability for the task.

Most differences in Figure 2 are statistically significant, with larger effects for *gender-only* and *race-only* attributes. Generally,

[1]We will use "*any*" as a replacement for all possible attribute combinations. For example, in this case, *Afro-American-any* replaces *Afro-American-man*, *Afro-American-woman* and *Afro-American-non-binary*.

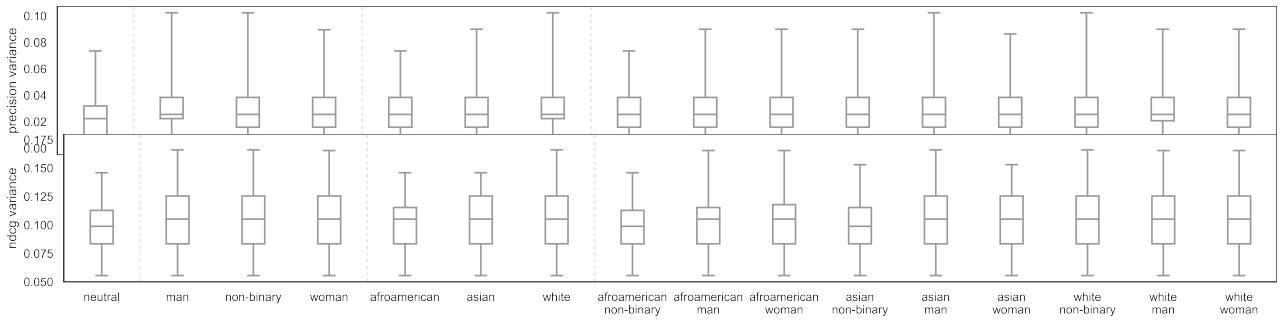


Figure 2: Distribution of maximum variance scores across sensitive attributes $k = 5$

	min		mean		max	
	prec	nDCG	prec	nDCG	prec	nDCG
ImplicitMF	0.204	0.460	0.204	0.460	0.204	0.460
neutral	0.363	0.693	0.363	0.693	0.363	0.693
man	0.31	0.66	0.358	0.691	0.406	0.721
non-binary	0.314	0.662	0.357	0.69	0.398	0.716
woman	0.312	0.661	0.356	0.69	0.399	0.716
Afro-Am	0.315	0.664	0.355	0.689	0.395	0.714
Asian	0.326	0.671	0.364	0.695	0.402	0.718
white	0.319	0.666	0.361	0.693	0.403	0.719
Afro-Am man	0.314	0.663	0.354	0.689	0.394	0.713
Afro-Am non-binary	0.312	0.662	0.355	0.689	0.392	0.712
Afro-Am woman	0.308	0.659	0.353	0.688	0.397	0.715
Asian man	0.319	0.667	0.359	0.692	0.398	0.715
Asian non-binary	0.316	0.664	0.356	0.69	0.393	0.713
Asian woman	0.316	0.664	0.358	0.691	0.396	0.715
white man	0.306	0.657	0.353	0.688	0.399	0.716
white non-binary	0.311	0.661	0.355	0.689	0.396	0.714
white woman	0.305	0.657	0.352	0.687	0.395	0.714

Table 3: Summarized maximum group recommendation results across sensitive attributes $k = 5$

white-any and *Asian-any* achieved significantly higher results compared to *Afro-American-any*. These differences suggest higher variability in results for White-any and *Asian-any* groups, which could indicate not only a difference in how LLMs treat the sensitive attributes, but also an unfair treatment to users in a group.

The distribution of mean precision/nDCG scores (summarized in Table 2) showed that neutral recommendations achieved significantly higher minimum scores than sensitive-aware recommendations. However, for maximum mean scores, neutral recommendations performed significantly lower than sensitive-aware ones. This implies that the overall quality of recommendations could be improved by introducing sensitive attributes. However, as the observations for the variance, this suggests that introducing sensitive attributes disrupts the balance of neutral recommendations, causing LLMs to tailor recommendations to the interests of the user identified with the sensitive attribute rather than balancing group preferences, potentially leading to unfairness. Only a few statistically significant differences favoring neutral recommendations were observed for the mean scores.

Different interactions between sensitive attributes were observed. In most cases, *Asian-any* and *Afro-American-any* outperformed White-any, and *Asian-any* outperformed *Afro-American-any*. These results reinforce that the LLM responds differently to the attributes, and that race might carry more importance. Despite being limited by the constrained set of movies to recommend from, stereotyped assumptions about interests linked to sensitive attributes were still reflected, at least partially, in the recommendations. For instance, `Mistral`'s responses (although not required by the prompt) included comments on how each recommended movie aligned with the lives and experiences of specific sensitive attributes.

The tendency for *white-any* to have lower score distributions and higher variances could be related to two opposing phenomena. First, an "unconscious" assumption of the LLM that the absence of a sensitive attribute implicitly aligns with a *white* one, leading to neutral recommendations resembling white-aware recommendations more than those for other attributes. Second, a "conscious" effort of the LLM to avoid prioritizing a commonly over-represented stereotype. Further studies in a less constrained scenario (e.g., providing a larger list of options or allowing the LLM to choose recommendations freely) are needed to effectively assess this phenomenon.

When analyzing the distribution of maximum scores (summarized in Table 3), we observed that for minimum and mean scores, neutral recommendations achieved significantly higher results than all sensitive-aware recommendations. However, for maximum scores, neutral recommendations performed significantly lower than *gender-only* and *race-only* attributes. This suggests that introducing sensitive attributes may reduce the scores of at least one group member (either the one identifying with the sensitive attribute or another member), potentially resulting in unfairness. Overall, both *Asian-any* and *white-any* groups achieved significantly higher results than *Afro-American-any* groups.

When varying group sizes, no statistically significant differences were observed for smaller groups. However, as group size increased, differences in precision and nDCG favored smaller groups. These differences were mainly in the distributions of variance and maximum scores. The same tendencies regarding differences with neutral scores were noted, but interactions between sensitive attributes varied. For instance, *non-binary-any* generally achieved significantly higher score distributions than any other gender attributes, highlighting the importance of this gender for recommendations as group size increased.

When setting $k = 10$, significant differences appeared across combinations of LLMs and group sizes. While $k = 5$ showed most differences in variances and maximum scores, $k = 10$ also revealed differences in mean score distributions. This suggests that in a

	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 4: Variation of user scores across sensitive attributes

	mean		max	
	prec.	nDCG	prec.	nDCG
man	0.167	0.38	0.361	0.693
non-binary	0.17	0.386	0.364	0.695
woman	0.17	0.385	0.363	0.694
Afro-Am	0.171	0.39	0.363	0.694
Asian	0.173	0.39	0.368	0.697
white	0.17	0.385	0.365	0.695
Afro-Am-man	0.171	0.389	0.362	0.694
Afro-Am-non-binary	0.168	0.383	0.358	0.691
Afro-Am-woman	0.169	0.385	0.36	0.692
Asian-man	0.171	0.386	0.363	0.694
Asian-non-binary	0.17	0.387	0.363	0.694
Asian-woman	0.171	0.386	0.365	0.695
white-man	0.168	0.382	0.358	0.69
white-non-binary	0.169	0.385	0.36	0.693
white-woman	0.167	0.382	0.357	0.69

Table 5: Summarized scores obtained by the users identifying with a sensitive attribute $k = 5$

constrained recommendation scenario, where the movie choices are fixed and not all of them can satisfy the entire group, increasing the number of recommendations leads to lower overall results.

Significant differences were observed among the three selected LLMs. `Mistral` achieved significantly lower results than `Gemma` (with differences up to 96% and 75% when considering neutral precision and nDCG, respectively), and `GPT` achieved significantly lower results than `Mistral` (with differences up to 180% and 134% when considering neutral precision and nDCG, respectively, compared to `Gemma`). In both cases, on average, the largest differences were observed for recommendations involving *Afro-American-any* and *Asian-any*. These differences also manifested in the interactions between the sensitive attributes. For example, with `GPT`, we observed that *white-any* tended to be higher than *Asian-man/woman* and *Afro-American-man/woman* attributes, and that *Afro-American-man* and *non-binary-any* had larger variances. In the case of `Mistral`, while we observed the same tendencies regarding neutral and sensitive-aware recommendations, statistically significant differences among sensitive attributes were scarce for mean and maximum scores distributions. This highlights how LLMs treat (and prioritize) sensitive attributes differently.

> 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.

2.3 RQ3. Consistency of recommendations

To assess the effect of sensitive attributes at the user level, we first compared the scores of the users in a group identifying with sensitive attributes (summarized in Table 5). These trends mirrored those observed for the overall group scores, with a tendency to favor *Asian-any* over *white-any* and *Afro-American-any*. `GPT` and `Mistral` also reflected the group trends, with *white-any* leading to significantly higher scores than other sensitive attributes. These differences were more pronounced than for `Gemma`, resulting in average differences of 12.8% in precision and 9.2% in nDCG when comparing *white-any* with *Afro-American-any*. Comparing with the max distribution of scores (summarized in Table 3), we observed that users identifying with a sensitive attribute generally had lower scores than the maximum scores for that attribute. This indicates that these users did not always receive the best recommendations.

Table 4 summarizes how users' scores changed (when compared to the neutral recommendations) in two scenarios i) they identified with the sensitive attribute, ii) other user in the group identified with the attribute. Although overall group results suggested that including sensitive attributes could improve performance, a closer look at the individual user level reveals that this trend was not predominant. As the Table shows, scores improved for an average of 8% of users in a group, while at least 50% of the groups saw no score improvement for any user (under any sensitive attribute). On average, between 6% and 9% of users in a group experienced worse scores due to identifying with a sensitive attribute. *Men-only*, *white-only* (usually the most represented sensitive attributes) and *Afro-American-any* (usually among the least represented sensitive attributes) had the largest proportion of users with worse scores, while *Asian-only* had the lowest. Only a few differences were statistically significant, such as *Asian-any* compared to *Afro-American-any*. Regarding the second scenario, a higher proportion of users improved their scores when another user identified with the sensitive attribute, ranging between 13% to 14%, whereas 12% to 17% users decreased their scores. As a result, users experienced score reductions more frequently when another group member, rather than themselves, identified with the 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, respectively.

When considering $k = 10$, the rates of users increasing/decreasing their scores doubled, with *Man-only* and *Afro-American-any* showing the highest improvement and *Afro-American-any* showing the highest decrease.

For `GPT`, the proportion of users in a group whose scores improved in the first scenario ranged, on average, between 9% and 14%, while 7% to 15% of users in a group 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-American-any* the largest. Differences favouring *white-only* and *white-man* were statistically significant in almost all cases. As seen with *Gemma*, in the second scenario, the proportions of users both improving and worsening their scores increased. In this scenario, 19% to 23% of users in a group increased their scores, with *Afro-American-any* showing the largest proportions, while 13% to 18% of users worsened their scores, with *Afro-American-any* and *Asian-any* showing the largest proportions.

Overall, these observations suggest unfair treatment of users. This unfairness not only reflects in the fact that users identifying with a sensitive attribute often experience worse scores, but also in that even a larger proportion of users experience worse scores when other group member identifies with a sensitive attribute. Since identifying with a sensitive attribute does not consistently lead to better outcomes, it appears these differences result from attempts to match users with stereotyped assumptions of their interests. When these assumptions do not align with the user's actual interests, it worsens their results, and affects those of the other group members.

In summary, although including sensitive attributes could globally improve group recommendations, their presence might lead to unfair user treatment, as it worsen recommendations for users identifying with the attributes.

ETHICAL STATEMENT

References to sensitive attributes are included solely to study fairness. We understand 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 fairness and biases. Ensuring transparency, accountability, and fairness in LLM-based recommendation systems is essential to mitigate these ethical challenges and maintain user trust in a healthy environment.

REFERENCES

[1] Yashar Deldjoo and Tommaso Di Noia. 2024. CFaiRLLM: Consumer Fairness Evaluation in Large-Language Model Recommender System. *arXiv preprint arXiv:2403.05668* (2024).

[2] 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).

[3] 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.

[4] Dimitris Sacharidis. 2019. Top-n group recommendations with fairness. In *Proceedings of the 34th ACM/SIGAPP symposium on applied computing*. 1663–1670.

[5] Robbie CM van Aert. 2023. Meta-analyzing partial correlation coefficients using Fisher's z transformation. *Research Synthesis Methods* 14, 5 (2023), 768–773.

[6] 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.

[7] 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.

[8] 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.