Supplementary Materials of Stairway to Fairness: Connecting Group and Individual Fairness

Theresia Veronika Rampisela[*]
University of Copenhagen
Copenhagen, Denmark
thra@di.ku.dk

Maria Maistro
University of Copenhagen
Copenhagen, Denmark
mm@di.ku.dk

Tuukka Ruotsalo
University of Copenhagen
Copenhagen, Denmark
LUT University
Lahti, Finland
tr@di.ku.dk

Falk Scholer
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

Christina Lioma
University of Copenhagen
Copenhagen, Denmark
c.lioma@di.ku.dk

ACM Reference Format:
Theresia Veronika Rampisela, Maria Maistro, Tuukka Ruotsalo, Falk Scholer, and Christina Lioma. 2025. Supplementary Materials of Stairway to Fairness: Connecting Group and Individual Fairness. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3705328.3748031

A Details on Experimental Setup

A.1 User grouping

For JobRec, we explain in more detail how we group users based on the study major. The study major is in free text format, so we first performed exploratory data analysis to get a better overview of the its distribution. More than 80% test users with a high school (HS) degree have no specific study major, so we do not use major to divide users with a HS degree into smaller groups. Further, an author manually grouped the 235 unique majors of the non-HS test users into six fields of study, taking inspiration from the grouping in Xu et al. [25]. During this process, we also remove users with generic or erroneous majors. We group the train/val users' majors as for the test users when possible, otherwise we perform fuzzy string matching between the annotated (manually grouped) and unannotated majors with `rapidfuzz` [3]. We map the unannotated major to the same group as the most similar annotated major, if similarity ≥ 0.75, otherwise we map it to an 'Others' category.

We summarise the statistics of user sensitive attributes (after preprocessing) in Tab. 1. We present the size of intersectional groups for each dataset in Tabs. 2–4.

A.2 Prompt templates and examples

We provide the templates and examples for the non-sensitive (NS) prompts in Tab. 5 and for the sensitive (S) prompts in Tab. 6. In the S-prompt, we replace "a user" in the NS-prompt with a description containing their sensitive attribute, e.g., "a user with a bachelor's degree, majoring in biology, has a total of 10 years of experience'. We use the original user attribute prior to re-grouping as they are more informative and precise (e.g., the user's major is included as 'biology', not 'STEM'). While this may not be the most natural way a user may interact with LLMs (e.g., explicitly disclosing their gender), it is a way to ensure that the LLMs have information on the users' sensitive attribute.

A.3 Measure formulations and technical details

To enhance reproducibility, we provide the formulations of all Fair measures that we compute. A Fair measure can be used to compute individual fairness, (between-)group fairness, and/or within-group fairness. We denote the individual, between-group, and within-group fairness versions of a measure as $\cdot_{\text{ind}}$, $\cdot_{\text{b-group}}$, and $\cdot_{\text{w-group}}$, respectively. When applicable, we provide the technical details (e.g., parameters) that we used. ↑/↓ means the higher/lower the better

A.3.1 Average scores of the worst 25% groups (Min). ↑Min [24] measures fairness between groups. To compute $\text{Min}_{\text{b-group}}$, the Eff score is first calculated per user, and then averaged per user group. The group mean Eff score is then sorted, and the first quartile value is set as a threshold. $\text{Min}_{\text{b-group}}$ averages the mean group Eff score of groups with an Eff score below/equal to the threshold. The range of Min follows the Eff score range.

A.3.2 Range. ↓Range can be used to quantify fairness between groups. It is computed as the range of mean group Eff scores [18].

A.3.3 Standard Deviation (SD). ↓SD[1] can be used to measure between-group fairness [18, 26], within-group fairness, and individual fairness [20]. $\text{SD}_{\text{b-group}}$ is the (population) standard deviation of the group mean Eff scores and SD_{ind} is the (population) standard deviation of users' Eff scores. We compute $\text{SD}_{\text{w-group}}$ as follows:

$$SD_{\text{w-group}} = \sum_{j=1}^{N'} s_j \cdot SD(g_j) \tag{1}$$

[*] Part of this work was done while visiting ADM+S at RMIT University.

[1] We also include prior work that computes Variance under the same family of measure, as Variance is the square of SD.

RecSys '25, Prague, Czech Republic
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1364-4/2025/09
https://doi.org/10.1145/3705328.3748031

where g_j is the list of group j users' effectiveness scores and N' is the number of user groups. The range of SD depends on the Eff score range. The weight of group j is s_j, the ratio of the group's total Eff scores to the total across all groups.

A.3.4 *Mean Absolute Difference (MAD).*

↓MAD [12] measures between-group fairness as the mean absolute pairwise difference of the groups' mean Eff scores. It is computed follows:

$$MAD = \frac{1}{N'(N'-1)} \sum_{j=1}^{N'-1} \sum_{j'=j+1}^{N'} \left|\bar{g_j} - \bar{g_{j'}}\right| \tag{2}$$

where $\bar{g_j}$ is the group j's mean Eff scores. The range of MAD follows the Eff score range.

A.3.5 *Gini Index (Gini).*

↓Gini [14] is a general-purpose inequality measure that has been used to measure between-group fairness [13], within-group fairness [11, 19], and individual fairness [17]. Given N non-negative values, $x_1, x_2, \ldots, x_N$, the general-purpose Gini is computed as follows:

$$Gini = \frac{\sum\limits_{j=1}^{N} (2j - N - 1)x_j}{N \sum\limits_{j=1}^{N} x_j} = \frac{\sum\limits_{j=1}^{N} (2j - N - 1)x_j}{N^2 \bar{x_j}} \tag{3}$$

Gini$_{\text{b-group}}$ is the Gini of the group mean Eff scores and Gini$_{\text{ind}}$ is the Gini of users' Eff scores. We compute Gini$_{\text{w-group}}$ as follows:

$$Gini_{\text{w-group}} = \sum_{j=1}^{N'} s_j \cdot Gini(g_j) \tag{4}$$

The range of Gini is [0,1].

A.3.6 *Coefficient of Variation (CV).*

↓CV is a general-purpose statistical measure of data dispersion, which has been used to measure between-group fairness [24, 27]. The general-purpose CV is computed as the ratio of the (population) standard deviation to the mean. We compute CV$_{\text{b-group}}$ as follows:

$$CV_{\text{b-group}} = \frac{SD(\bar{g_1}, \bar{g_2}, \ldots, \bar{g_{N'}})}{mean(\bar{g_1}, \bar{g_2}, \ldots, \bar{g_{N'}})} \tag{5}$$

The range of CV is $[0, \infty)$.

A.3.7 *F-statistic (FStat).*

↓FStat is originally a statistical test that measures the ratio of between-group variance to within-group variance. It has been used to quantify between-group fairness for N' groups and N' users [23], as follows:

$$V = \frac{1}{N} \sum_{j=1}^{N'} |g_j| \cdot (\bar{g_j} - \bar{x}_{\text{eff}})^2 \tag{6}$$

$$U = \frac{1}{N} \sum_{j=1}^{N'} \sum_{\ell=1}^{|g_j|} (x_{j,\ell} - \bar{g_j})^2 \tag{7}$$

$$FStat_{\text{b-group}} = \frac{V/(N'-1)}{U/(N-N')} \tag{8}$$

where $|g_j|$ is the number of users in group j, $\bar{x}_{\text{eff}}$ is the mean Eff scores of all users, and $x_{j,\ell}$ is the Eff score of the ℓ-th user in the j-th group. The range of FStat is $[0, \infty)$.

A.3.8 *Kullback-Leibler Divergence (KL).*

↓KL can be customised to measure various fairness criteria; we select the version where between-group fairness means that effectiveness should be proportional to group size, as it is a popular choice [1]. KL is computed as follows:

$$KL_{\text{b-group}}(p||q) = \sum_{j=1}^{N'} p_j \log_2 \frac{p_j}{q_j} \tag{9}$$

$$p_j = \frac{\bar{g_j}}{\sum_{j=1}^{N'} \bar{g_j}} \tag{10}$$

$$q_j = \frac{|g_j|}{\sum_{j=1}^{N'} |g_j|} \tag{11}$$

The range of KL is $[0, \infty)$.

A.3.9 *Generalised Cross Entropy (GCE).*

↓ GCE can be used to quantify various fairness criteria [8, 9]; we choose the version that requires equal utility (effectiveness) across groups. We compute GCE as follows:[2]

$$GCE_{\text{b-group}} = -\frac{1}{B(1-B)} \sum_{j=1}^{N'} p_{ref}^{B} \cdot \left(\frac{\hat{p_j}}{\sum_{j=1}^{N'} \hat{p_j}} \right)^{(1-B)} - 1 \tag{12}$$

$$\hat{p_j} = \lambda p_j + (1-\lambda)c \tag{13}$$

where B is a parameter, $p_{ref} = 1/N'$, p_j is as per Eq. (10) and $\hat{p_j}$ is the smoothened value of p_j. We set $\beta = 2$, and perform smoothing with $\lambda = 0.95$ and $c = 10^{-4}$. The range of GCE is $[0, \infty)$.

A.3.10 *Atkinson Index.*

We adopt Atkinson Index (Atk), a measure of income inequality from the economics domain [2], to quantify group/individual fairness for RS users based on recommendation effectiveness disparity.

General formulation. Given N non-negative values, $x_1, x_2, \ldots, x_N$, the general-purpose Atk is defined as follows:[3]

$$Atk(x_1, x_2, \ldots, x_N) = 1 - \frac{1}{\bar{x}} f(x_1, x_2, \ldots, x_N) \tag{14}$$

$$f(x_1, x_2, \ldots, x_N) = \left(\sum_{j=1}^{N} \frac{w_j}{\sum_{j=1}^{N} w_j} x_j^{1-\varepsilon} \right)^{\frac{1}{1-\varepsilon}} \tag{15}$$

where $\bar{x}$ is the mean of the values, w_j is the weight of the j-th score, and $\varepsilon \geq 0$ is the inequality aversion parameter.[4] Atk with a low ε is more sensitive to disparity among higher scores than among lower scores; we compute all Atk with $\varepsilon = 0.5$ [7]. Atk has a [0,1]-range.

Measuring user fairness with Atk. To measure individual fairness for N users with Atk (referred to as Atk$_{\text{ind}}$), we input the effectiveness score (e.g., NDCG) per user to Eq. (14) and set $w_j = 1, \forall j \in \{1, \ldots, N\}$ to weigh each user equally. As a user's effectiveness score could be 0, we choose $\varepsilon < 1$ to avoid division by zero in $x_j^{1-\varepsilon}$ (Eq. (15)). Atk$_{\text{b-group}}$ is computed as follows:

$$Atk_{\text{b-group}} = Atk\left(f(g_1), f(g_2), \ldots, f(g_{N'})\right) \tag{16}$$

Essentially, Atk$_{\text{b-group}}$ measures between-group fairness while accounting for the disparity within each group via f. As f may be unstable for small groups, we reduce the score contributions from

[2] Note that this is the nonnegative version of the measure in [9].

[3] Atk with $\varepsilon = 1$ is defined differently from Eq. (14).

[4] If $x_j = 0, \forall j \in \{1, 2, \ldots, N\}$, we define ↓Atk = 0.

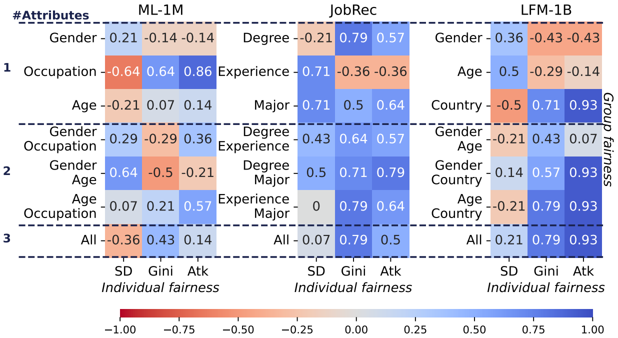

Figure 1: Agreement (Kendall's τ) between the same family of measure in ranking LLMRecs for NDCG-based group fairness (y-axis) and individual fairness (x-axis). Group fairness is computed for each combination of users' sensitive attributes.

small groups by using the group size as w_j to weigh $f(g_j)$. To compute $f(g_j)$ itself, we weigh each user in the group equally.

Decomposition of Atk. Atk is subgroup-decomposable [22], i.e., it can be broken down into between- and within-group components without residual terms [5], establishing a coherent relationship between individual fairness and group fairness [10]. Specifically, it detects if individual unfairness occurs mainly due to between- or within-group effectiveness disparity. Atk_{ind} can be decomposed as follows [4–6]:

$$\underbrace{1 - Atk_{\text{ind}}}_{\text{individual fairness}} = \underbrace{(1 - Atk_{\text{b-group}})}_{\text{between-group fairness}} \cdot \underbrace{(1 - Atk_{\text{w-group}})}_{\text{within-group fairness}} \tag{17}$$

$$Atk_{\text{w-group}} = \sum_{j=1}^{N'} s_j \cdot Atk(g_j) \tag{18}$$

B Extended Results

B.1 Evaluation of all LLMRecs

We present Eff and P-based Fair scores for LLMRecs in Tab. 7.

B.2 Agreement of group and individual fairness measures

We show in Fig. 1 the agreement of group-individual fairness between the same family of measures (e.g., $\text{SD}_{\text{b-group}}$ vs SD_{ind}) for all possible groupings.

B.3 Fairness decomposability

We show in Fig. 2 the individual, between-group, and within-group unfairness of GLM-4-9B (NS-prompt) for all ways of grouping users in ML-1M.

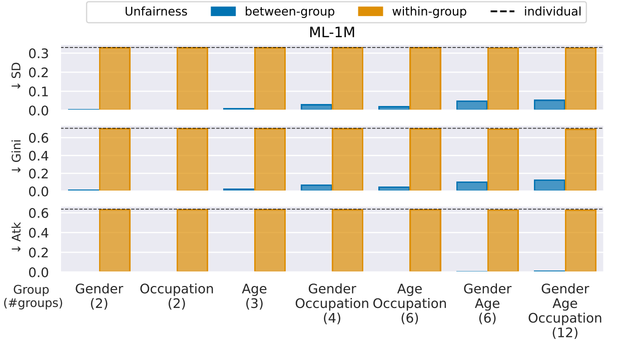

Figure 2: NDCG-based individual, between- and within-group unfairness of GLM-4-9B (NS-prompt) for all ways of grouping users in ML-1M.

References

[1] Enrique Amigó, Yashar Deldjoo, Stefano Mizzaro, and Alejandro Bellogín. 2023. A unifying and general account of fairness measurement in recommender systems. *Information Processing & Management* 60, 1 (1 2023), 103115. https://doi.org/10.1016/J.IPM.2022.103115

[2] Anthony B Atkinson. 1970. On the measurement of inequality. *Journal of Economic Theory* 2, 3 (1970), 244–263. https://doi.org/10.1016/0022-0531(70)90039-6

[3] Max Bachmann. 2024. *rapidfuzz/RapidFuzz: Release 3.8.1.* https://doi.org/10.5281/zenodo.10938887

[4] Charles Blackorby, Walter Bossert, and David Donaldson. 1999. *Income Inequality Measurement: The Normative Approach.* Springer Netherlands, Dordrecht, 133–161. https://doi.org/10.1007/978-94-011-4413-1_4

[5] Francois Bourguignon. 1979. Decomposable Income Inequality Measures. *Econometrica* 47, 4 (1979), 901–920. http://www.jstor.org/stable/1914138

[6] Meltem Dayioğlu and Cem Başlevent. 2006. Imputed Rents and Regional Income Inequality in Turkey: A Subgroup Decomposition of the Atkinson Index. *Regional Studies* 40, 8 (2006), 889–905. https://doi.org/10.1080/00343400600984395

[7] Fernando G. De Maio. 2007. Income inequality measures. *Journal of Epidemiology and Community Health* 61, 10 (10 2007), 849. https://doi.org/10.1136/JECH.2006.052969

[8] Yashar Deldjoo, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogín, and Tommaso Di Noia. 2019. Recommender Systems Fairness Evaluation via Generalized Cross Entropy. In *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019).* CEUR-WS.

[9] Yashar Deldjoo, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogín, and Tommaso Di Noia. 2021. A flexible framework for evaluating user and item fairness in recommender systems. *User Modeling and User-Adapted Interaction* 31 (2021), 457–511. https://doi.org/10.1007/s11257-020-09285-1

[10] Guido Erreygers, Roselinde Kessels, Linkun Chen, and Philip Clarke. 2018. Subgroup Decomposability of Income-Related Inequality of Health, with an Application to Australia. *Economic Record* 94, 304 (2018), 39–50. https://doi.org/10.1111/1475-4932.12373

[11] Andres Ferraro, Michael D. Ekstrand, and Christine Bauer. 2024. It's Not You, It's Me: The Impact of Choice Models and Ranking Strategies on Gender Imbalance in Music Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems* (Bari, Italy) *(RecSys '24).* Association for Computing Machinery, New York, NY, USA, 884–889. https://doi.org/10.1145/3640457.3688163

[12] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, and Gerard de Melo. 2020. Fairness-Aware Explainable Recommendation over Knowledge Graphs. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20).* Association for Computing Machinery, New York, NY, USA, 69–78. https://doi.org/10.1145/3397271.3401051

[13] Avijit Ghosh, Tomo Lazovich, Kristian Lum, and Christo Wilson. 2024. Reducing Population-level Inequality Can Improve Demographic Group Fairness: a Twitter Case Study. arXiv:2409.08135 [cs.SI] https://arxiv.org/abs/2409.08135

[14] Corrado Gini. 1912. *Variabilità e Mutabilità.* C. Cuppini, Bologna. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche.

[15] Ben Hamner, Road Warrior, and Wojciech Krupa. 2012. Job Recommendation Challenge. https://kaggle.com/competitions/job-recommendation.

[16] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages. https://doi.org/10.1145/2827872

[17] Jurek Leonhardt, Avishek Anand, and Megha Khosla. 2018. User Fairness in Recommender Systems. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18).* International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 101–102. https://doi.org/10.1145/3184558.3186949

[18] Zhiqiang Liu, Xiaoxiao Xu, Jiaqi Yu, Han Xu, Lantao Hu, Han Li, and Kun Gai. 2024. A Self-Adaptive Fairness Constraint Framework for Industrial Recommender System. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 4726–4733. https://doi.org/10.1145/3627673.3680099

[19] Eliana Pastor and Francesco Bonchi. 2024. Intersectional fair ranking via subgroup divergence. *Data Min. Knowl. Discov.* 38, 4 (May 2024), 2186–2222. https://doi.org/10.1007/s10618-024-01029-8

[20] Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. 2019. Fighting Fire with Fire: Using Antidote Data to Improve Polarization and Fairness of Recommender Systems. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 231–239. https://doi.org/10.1145/3289600.3291002

[21] Markus Schedl. 2016. The LFM-1b Dataset for Music Retrieval and Recommendation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR '16)*. Association for Computing Machinery, New York, NY, USA, 103–110. https://doi.org/10.1145/2911996.2912004

[22] Anthony F. Shorrocks. 1984. Inequality Decomposition by Population Subgroups. *Econometrica* 52, 6 (1984), 1369–1385. http://www.jstor.org/stable/1913511

[23] Mengting Wan, Jianmo Ni, Rishabh Misra, and Julian McAuley. 2020. Addressing Marketing Bias in Product Recommendations. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston, TX, USA) *(WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 618–626. https://doi.org/10.1145/3336191.3371855

[24] Yifan Wang, Peijie Sun, Weizhi Ma, Min Zhang, Yuan Zhang, Peng Jiang, and Shaoping Ma. 2024. Intersectional Two-sided Fairness in Recommendation. In *Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 3609–3620. https://doi.org/10.1145/3589334.3645518

[25] Chen Xu, Wenjie Wang, Yuxin Li, Liang Pang, Jun Xu, and Tat-Seng Chua. 2024. A Study of Implicit Ranking Unfairness in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 7957–7970. https://aclanthology.org/2024.findings-emnlp.467

[26] 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* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 993–999. https://doi.org/10.1145/3604915.3608860

[27] Ziwei Zhu, Jianling Wang, and James Caverlee. 2020. Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 449–458. https://doi.org/10.1145/3397271.3401177

Table 1: Statistics of user sensitive attributes after preprocessing.

	ML-1M [16]	JobRec [15]	LFM-1B [21]
sensitive attr. #1	gender (2): M (457); F (163)	degree (3): University (254); High School (157); College (112)	gender (2): M (12,826); F (3,785)
sensitive attr. #2	age (3): 25–49 years (437); 18–24 years (110); ≥50 years (73)	years of experience (3): >10 (219); >5–10 (200); ≤5 (104)	age (3): 18–24 years (9,821); 25–49 years (6,590); ≥50 years (200)

sensitive attr. #3	occupation (2):
	working (478); non-working (142)
		major (6):
		Business, Management, Finance (154); Social Science, Humanities, Education (103); STEM (41); Health & Medical (29); Arts, Creative, Entertainment (22); Others (17);
			country/continent (5):
			Europe (9,976); America & Antarctica (5,337); Asia (863); Oceania (376); Africa (59)

Table 2: Size of intersectional groups (ML-1M).

gender	age	occupation	#user
M	18–24 years	non-working	44
		working	37
	25–49 years	non-working	41
		working	279
	≥50 years	non-working	8
		working	48
F	18–24 years	non-working	21
		working	8
	25–49 years	non-working	26
		working	91
	≥50 years	non-working	2
		working	15

Table 3: Size of intersectional groups (JobRec). Combinations resulting in group size of 0 are excluded.

degree	experience (years)	major	#user
High School	≤5	-	33
	>5−10	-	54
	>10	-	70
College	≤5	Business, Management, Finance	7
		Social Science, Humanities, Education	3
		STEM	4
		Health & Medical	8
		Arts, Creative, Entertainment	2
		Others	2
	>5−10	Business, Management, Finance	13
		Social Science, Humanities, Education	6
		STEM	8
		Health & Medical	10
		Arts, Creative, Entertainment	3
		Others	4
	>10	Business, Management, Finance	14
		Social Science, Humanities, Education	12
		STEM	3
		Health & Medical	3
		Arts, Creative, Entertainment	4
		Others	6
University	≤5	Business, Management, Finance	18
		Social Science, Humanities, Education	16
		STEM	7
		Health & Medical	3
		Arts, Creative, Entertainment	1
	>5−10	Business, Management, Finance	47
		Social Science, Humanities, Education	38
		STEM	9
		Arts, Creative, Entertainment	8
	>10	Business, Management, Finance	55
		Social Science, Humanities, Education	28
		STEM	10
		Health & Medical	5
		Arts, Creative, Entertainment	4
		Others	5

Table 4: Size of intersectional groups (LFM-1B). Combinations resulting in group size of 0 are excluded.

gender	age	continent	#user
F	18–24 years	Europe	1635
		America & Antarctica	901
		Asia	144
		Oceania	46
		Africa	8
	25–49 years	Europe	551
		America & Antarctica	392
		Asia	59
		Oceania	19
		Africa	1
	≥50 years	Europe	15
		America & Antarctica	11
		Asia	2
		Africa	1
M	18–24 years	Europe	4260
		America & Antarctica	2251
		Asia	371
		Oceania	180
		Africa	25
	25–49 years	Europe	3418
		America & Antarctica	1723
		Asia	276
		Oceania	128
		Africa	23
	≥50 years	Europe	97
		America & Antarctica	59
		Asia	11
		Oceania	3
		Africa	1

Table 5: Non-sensitive (NS) prompt templates and examples.

Dataset: ML-1M
Prompt template: “You are a movie recommender. If a user has watched the following movies, listed chronologically from earliest to latest, and rated them positively: <*items in train*>, then you should recommend the following movies: <*items in val*>. Now that the user has watched the recommended movies, you should recommend 10 other movies from the year between 1919 and 2000 (inclusive) that the user is most likely to watch next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend movies that have been watched by the user. Do not output other words."
Prompt example: “You are a movie recommender. If a user has watched the following movies, listed chronologically from earliest to latest, and rated them positively: American Beauty, Magnolia, Boogie Nights, The Tao of Steve, Dark City, Fight Club, then you should recommend the following movies: The Evening Star, The MatchMaker, West Side Story, Dirty Dancing, The Prince of Egypt, Grease, Fargo, Good Will Hunting, Playing by Heart, The Man Without a Face. Now that the user has watched the recommended movies, you should recommend 10 other movies from the year between 1919 and 2000 (inclusive) that the user is most likely to watch next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend movies that have been watched by the user. Do not output other words."
Dataset: JobRec
Prompt template: “You are a job title recommender. If a user has applied to job positions with the following job titles, listed chronologically from earliest to latest: <*items in train*>, you should recommend the following job titles: <*items in val*>. Now that the user has applied to positions with those job titles, you should recommend 10 other job titles that the user is most likely to apply for next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not output other words.”
Prompt example: “You are a job title recommender. If a user has applied to job positions with the following job titles, listed chronologically from earliest to latest: Billing/Collections Process Coordinator, Administrative Assistant, Office Manager/Bookkeeper, Bookkeeper, you should recommend the following job titles: Accounts Payable Clerk, Assistant Director of Human Resources. Now that the user has applied to positions with those job titles, you should recommend 10 other job titles that the user is most likely to apply for next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not output other words.”
Dataset: LFM-1B
Prompt template: “You are a music artist recommender. If a user has listened to the following artists, listed chronologically from earliest to latest: <*items in train*> and considering that the user listened to the artist <*playcount of items in train*> times respectively, then you should recommend the following artists: <*items in val*>. Now that the user has listened to the recommended artists, you should recommend 10 other artists, who has released songs prior to 2017, that the user is most likely to listen to next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend artists that the user has listened to. Do not output other words.”
Prompt example: “You are a music artist recommender. If a user has listened to the following artists, listed chronologically from earliest to latest: Lee Van Dowski, The Oohlas, XTC, Brenmar, Mountain, Bladerunner, Efdemin, Tom Encore, Droid Sector, Xhin, and considering that the user listened to the artist 2, 1, 2, 4, 12, 1, 12, 1, 2, 3 times respectively, then you should recommend the following artists: VNV Nation, Space, Ja Rule, Nneka, Parks, Enzyme X, David Bisbal, Primer 55, A-Sides, Cydonia. Now that the user has listened to the recommended artists, you should recommend 10 other artists, who has released songs prior to 2017, that the user is most likely to listen to next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend artists that the user has listened to. Do not output other words.”

Table 6: Sensitive (S) prompt templates and examples.

Dataset: ML-1M
Prompt template: "You are a movie recommender. If a <*gender*> user whose age is <*age range*> years old and whose occupation is <*occupation*> has watched the following movies, listed chronologically from earliest to latest, and rated them positively: <*items in train*>, then you should recommend the following movies: <*items in val*>. Now that the user has watched the recommended movies, you should recommend 10 other movies from the year between 1919 and 2000 (inclusive) that the user is most likely to watch next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend movies that have been watched by the user. Do not output other words." **Prompt example:** "You are a movie recommender. If a female user whose age is 18-24 years old and whose occupation is writer has watched the following movies, listed chronologically from earliest to latest, and rated them positively: American Beauty, Magnolia, Boogie Nights, The Tao of Steve, Dark City, Fight Club, then you should recommend the following movies: The Evening Star, The MatchMaker, West Side Story, Dirty Dancing, The Prince of Egypt, Grease, Fargo, Good Will Hunting, Playing by Heart, The Man Without a Face. Now that the user has watched the recommended movies, you should recommend 10 other movies from the year between 1919 and 2000 (inclusive) that the user is most likely to watch next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend movies that have been watched by the user. Do not output other words."
Dataset: JobRec
Prompt template: "You are a job title recommender. If a user with a <*degree type*> degree, majoring in <*major*>, has a total of <*years of experience*> years of experience has applied to job positions with the following job titles, listed chronologically from earliest to latest: <*items in train*>, then you should recommend the following job titles: <*items in val*>. Now that the user has applied to positions with those job titles, you should recommend 10 other job titles that the user is most likely to apply for next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not output other words." **Prompt example:** "You are a job title recommender. If a user with a Bachelor's degree, majoring in Accounting and Information Systems, has a total of 24 years of experience has applied to job positions with the following job titles, listed chronologically from earliest to latest: Billing/Collections Process Coordinator, Administrative Assistant, Office Manager/Bookkeeper, Bookkeeper, then you should recommend the following job titles: Accounts Payable Clerk, Assistant Director of Human Resources. Now that the user has applied to positions with those job titles, you should recommend 10 other job titles that the user is most likely to apply for next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not output other words."
Dataset: LFM-1B
Prompt template: ‘You are a music artist recommender. If a <*gender*> user whose age is <*age*> years old and lives in <*country*> has listened to the following artists, listed chronologically from earliest to latest: <*items in train*>, and considering that the user listened to the artist <*playcount of items in train*> times respectively, then you should recommend the following artists: <*items in val*>. Now that the user has listened to the recommended artists, you should recommend 10 other artists, who has released songs prior to 2017, that the user is most likely to listen to next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend artists that the user has listened to. Do not output other words." **Prompt example:** "You are a music artist recommender. If a male user whose age is 30 years old and lives in Russia has listened to the following artists, listed chronologically from earliest to latest: Lee Van Dowski, The Oohlas, XTC, Brenmar, Mountain, Bladerunner, Efdemin, Tom Encore, Droid Sector, Xhin, and considering that the user listened to the artist 2, 1, 2, 4, 12, 1, 12, 1, 2, 3 times respectively, then you should recommend the following artists: VNV Nation, Space, Ja Rule, Nneka, Parks, Enzyme X, David Bisbal, Primer 55, A-Sides, Cydonia. Now that the user has listened to the recommended artists, you should recommend 10 other artists, who has released songs prior to 2017, that the user is most likely to listen to next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend artists that the user has listened to. Do not output other words."

Table 7: Effectiveness (Eff) and fairness (Fair) scores at $k = 10$ for intersectional groups (Grp.) and individuals (Ind.) of LLMRecs with non-sensitive (NS) and sensitive (S) prompts. All Fair scores are computed with P. All measures range in [0,1], except the Grp. measures below the grey lines. The best Eff/Fair scores are bolded. Darker green marks scores closer to the best Eff/Fair per measure. ↑ /↓ means the higher/lower the better.

	LLMRec	GLM-4-9B		Llama-3.1-8B		Ministral-8B		Qwen2.5-7B	
	prompt type	NS	S	NS	S	NS	S	NS	S
					ML-1M				
Eff	↑ HR	**0.377**	0.358	0.260	0.269	0.342	0.340	0.363	0.371
	↑ MRR	**0.189**	0.174	0.101	0.113	0.159	0.173	0.165	0.180
	↑ P	**0.082**	0.074	0.046	0.051	0.067	0.067	0.074	0.080
Fair (Grp.)	↑ Min	0.036	**0.049**	0.030	0.034	0.022	0.040	0.024	0.035
	↓ Range	**0.075**	0.088	**0.075**	0.088	0.089	0.125	0.175	0.125
	↓ SD	0.024	0.023	**0.019**	0.020	0.027	0.028	0.044	0.033
	↓ MAD	0.028	0.028	**0.022**	0.023	0.031	0.031	0.051	0.040
	↓ Gini	0.169	**0.166**	0.201	0.188	0.229	0.196	0.301	0.228
	↓ Atk	**0.018**	0.019	0.031	0.021	0.032	0.037	0.057	0.041
	↓ CV	0.323	**0.299**	0.388	0.358	0.444	0.391	0.577	0.416
	↓ FStat	0.504	0.450	0.472	**0.369**	0.625	0.580	1.586	0.973
	↓ KL	0.836	1.211	1.328	1.349	**0.793**	1.292	1.380	1.150
	↓ GCE	0.106	**0.051**	0.068	0.140	659.834	659.704	0.387	0.273
Fair (Ind.)	↓ SD	0.141	0.129	**0.096**	0.109	0.124	0.123	0.136	0.141
	↓ Gini	**0.752**	0.760	0.818	0.819	0.777	0.774	0.764	0.761
	↓ Atk	**0.658**	0.674	0.760	0.755	0.693	0.692	0.675	0.667
					JobRec				
Eff	↑ HR	0.054	0.033	0.044	0.021	0.057	0.044	**0.065**	0.057
	↑ MRR	0.037	0.023	0.017	0.009	**0.048**	0.033	0.046	0.045
	↑ P	0.006	0.003	0.005	0.002	0.006	0.004	**0.007**	**0.007**
Fair (Grp.)	↑ Min	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
	↓ Range	0.050	**0.009**	0.050	0.050	0.033	0.021	0.033	0.033
	↓ SD	0.010	**0.003**	0.010	0.008	0.008	0.005	0.009	0.008
	↓ MAD	0.008	**0.003**	0.008	0.004	0.007	0.005	0.009	0.007
	↓ Gini	0.800	0.798	0.828	0.911	0.742	0.792	**0.726**	0.770
	↓ Atk	0.511	0.473	0.642	0.759	0.500	0.509	0.467	**0.456**
	↓ CV	2.082	1.866	2.265	3.634	1.631	1.853	**1.510**	1.774
	↓ FStat	0.850	**0.433**	1.194	1.009	0.873	0.697	0.928	0.643
	↓ KL	2.715	**1.319**	3.205	4.682	1.724	1.826	2.012	1.934
	↓ GCE	1685.894	1979.079	1685.949	2052.428	1612.570	1759.184	**1539.280**	1759.169
Fair (Ind.)	↓ SD	0.024	0.018	0.022	**0.014**	0.026	0.021	0.028	0.028
	↓ Gini	0.948	0.967	0.958	0.979	0.946	0.956	**0.940**	0.948
	↓ Atk	0.947	0.967	0.956	0.979	0.943	0.956	**0.936**	0.944
					LFM-1B				
Eff	↑ HR	0.658	**0.661**	0.609	0.618	0.451	0.317	0.375	0.317
	↑ MRR	**0.409**	0.408	0.347	0.357	0.266	0.174	0.199	0.160
	↑ P	0.203	**0.206**	0.172	0.181	0.118	0.080	0.106	0.085
Fair (Grp.)	↑ Min	0.096	**0.101**	0.100	0.098	0.036	0.018	0.038	0.018
	↓ Range	0.313	0.300	0.500	0.600	0.600	0.200	0.200	**0.136**
	↓ SD	0.076	0.068	0.088	0.094	0.102	0.042	0.047	**0.040**
	↓ MAD	0.079	0.073	0.081	0.078	0.080	**0.046**	0.051	**0.046**
	↓ Gini	0.204	**0.192**	0.225	0.223	0.341	0.318	0.250	0.314
	↓ Atk	0.004	**0.003**	0.005	0.005	0.008	0.015	0.007	0.022
	↓ CV	0.406	**0.367**	0.509	0.552	0.899	0.601	0.484	0.565
	↓ FStat	2.932	2.707	3.357	3.827	2.865	3.399	**1.721**	4.534
	↓ KL	2.795	2.744	3.546	3.291	3.510	2.726	2.740	**2.194**
	↓ GCE	338.854	225.918	**113.016**	**113.016**	451.861	451.868	451.815	564.777
Fair (Ind.)	↓ SD	0.226	0.228	0.204	0.210	0.183	**0.157**	0.187	0.170
	↓ Gini	0.586	**0.583**	0.610	0.603	0.718	0.803	0.769	0.807
	↓ Atk	0.415	**0.411**	0.454	0.445	0.599	0.719	0.669	0.722