A Study of Implicit Ranking Unfairness in Large Language Models

Chen Xu[1], **Wenjie Wang**[2], **Yuxin Li**[1], **Liang Pang**[3], **Jun Xu**[1*], **Tat-Seng Chua**[2]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]NExT++ Research Center, National University of Singapore
[3]Institute of Computing Technology, Chinese Academy of Sciences
{xc_chen, yuxin_li, junxu}@ruc.edu.cn
{wangwenjie, dcscts}@nus.edu.sg, pangliang@ict.ac.cn

Abstract

Content Warning: This paper contains examples of misgendering and erasure that could be offensive and potentially triggering.

Recently, Large Language Models (LLMs) have demonstrated a superior ability to serve as ranking models. However, concerns have arisen as LLMs will exhibit discriminatory ranking behaviors based on users' sensitive attributes (*e.g.,* gender). Worse still, in this paper, we identify a subtler form of discrimination in LLMs, termed *implicit ranking unfairness*, where LLMs exhibit discriminatory ranking patterns based solely on non-sensitive user profiles, such as user names. Such implicit unfairness is more widespread but less noticeable, threatening the ethical foundation. To comprehensively explore such unfairness, our analysis will focus on three research aspects: (1) We propose an evaluation method to investigate the severity of implicit ranking unfairness. (2) We uncover the reasons for causing such unfairness. (3) To mitigate such unfairness effectively, we utilize a pair-wise regression method to conduct fair-aware data augmentation for LLM fine-tuning. The experiment demonstrates that our method outperforms existing approaches in ranking fairness, achieving this with only a small reduction in accuracy. Lastly, we emphasize the need for the community to identify and mitigate the implicit unfairness, aiming to avert the potential deterioration in the reinforced human-LLMs ecosystem deterioration.

1 Introduction

Large language models (LLMs), represented by ChatGPT (Wu et al., 2023b) have empowered ranking tasks (Wu et al., 2023a), which is important in filtering overload information to users (Liu et al., 2009). However, ensuring that LLMs do not pose ethical risks becomes crucial. Recently, various evaluation methods have been introduced to assess the degree of discrimination in LLMs (Kasneci et al., 2023; Chang et al., 2023; Dai et al., 2024), showing that LLMs frequently exhibit pronounced ranking discriminatory behaviors against explicit sensitive attributes, such as gender (Zhang et al., 2023b; Tamkin et al., 2023).

Although a massive amount of work focuses on addressing unfairness when explicitly using sensitive attributes in ranking tasks (Dai et al., 2024), our investigation reveals the persistence of *implicit ranking unfairness*: LLMs even generate substantial discriminatory ranking behaviors when using non-sensitive yet personalized user profiles (*e.g.,* user names). These profiles are commonly used as identifiers of gender and race since humans often draw stereotypical conclusions based on them (Smith and Williams, 2021; Romanov et al., 2019; De-Arteaga et al., 2019). *Implicit ranking unfairness* in LLMs highlights new and more urgent risks towards LLMs-based ranking application (*e.g.,* recommendation) because (1) such unfairness is often inconspicuous because it only depends on non-sensitive user profiles; and (2) such unfairness is more widespread since these non-sensitive user profiles can be easily acquired and used by existing platforms, such as user names or email addresses. To comprehensively analyze the problem, in this paper, we will focus on three research aspects regarding implicit ranking unfairness in LLMs.

Firstly, we propose an evaluation method to investigate how serious the implicit ranking unfairness is in existing LLMs. Specifically, following the practice in (Zhang et al., 2023b), we design a ranking task prompt template (Figure 1). Then we give substantial empirical evidence to confirm the existence of implicit ranking unfairness. Finally, we find that the degree of implicit ranking unfairness is nearly 2-4 times more serious than explicit unfairness, and the unfairness is caused by collaborative information. Empirical evidence is in Section 4).

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7957–7970
November 12-16, 2024

Secondly, since this implicit unfairness is more severe and more hidden, we aim to investigate the reasons behind its occurrence. Specifically, we identify that the LLMs can probe sensitive attributes exclusively from these personalized and non-sensitive user profiles. Then we also show that the word embeddings of certain non-sensitive user profiles are more closely aligned with the sensitive attribute. Such phenomena contribute to the collection of unfair datasets during the pre-training phases (see evidence in Section 5).

Finally, we aim to propose a method to mitigate such implicit ranking unfairness. Previous research proposed to mitigate user unfairness either by employing privacy policies that hide sensitive attributes (Xiao et al., 2023; Brown et al., 2022; Kandpal et al., 2022), utilizing certain prompts to instruct LLMs to disregard sensitive attributes (Hua et al., 2023) or add counterfactual sample to enhance fairness (Ghanbarzadeh et al., 2023). However, they show limited effectiveness in mitigating implicit ranking unfairness (See Section 6).

In this paper, we propose a fair-aware data argumentation method to mitigate such unfairness. Specifically, we incorporate counterfactual samples that contain certain implicit attributes to help the model produce fair ranking results. Due to the massive and noisy characteristic of the non-sensitive features, we employ a pair-wise regression method to choose hard and informational non-sensitive features to conduct data argumentation. The experiments demonstrate that our method outperforms the existing methods on two ranking datasets.

Major Contributions: (1) We uncover that the LLMs-based ranking system demonstrates substantial implicit unfairness. (2) We analyze the reasons for causing such implicit unfairness. (3) We propose a new fair-aware data argumentation method to mitigate the implicit ranking unfairness effectively. Our code is available at `https://github.com/XuChen0427/Implicit_Rank_Unfairness/`.

2 Preliminary

In this section, we will formulate the LLMs-based ranking tasks and the implicit ranking unfairness concept formally.

2.1 LLMs-based Ranking Tasks

In LLMs-based ranking applications (Bao et al., 2023b,a), let $\mathcal{U}$ be the user set. A user $u \in \mathcal{U}$ will have non-sensitive features v_u (*e.g.,* user names) and sensitive features $s_u \in \mathcal{S}$ (*e.g.,* user gender). In our work, we define the set $\mathcal{S}$ to represent sensitive attribute types such as gender, race, or continent, and s_u is selected from options [Male, Female], [White, Black, Asian], or [Asian, Africa, Americas, Europe, Oceania]. When a user u engages ranking systems, a personalized prompt p_u will be used to instruct LLMs to conduct ranking. Given the prompt p_u and optional user features v_u and s_u, the LLMs-based ranking model will output a ranking list $L_K(u) = \{i_1, i_2, \cdots, i_K\}$, where K is the fixed ranking size and i_j is the j-th given item.

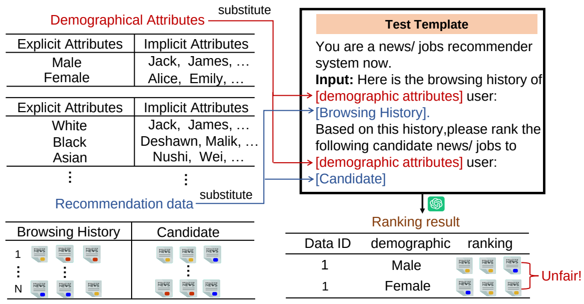


Figure 1: Overall workflow of our evaluation. The ranking list outputs by LLMs should be the same when replacing different sensitive attributes in prompts.

Previous work shows LLMs' powerful ability to serve as an information retriever (Dai et al., 2023; Bao et al., 2023b). Figure 1 shows the overall LLMs-based ranking workflow. Specifically, the full picture for utilizing LLMs in the context of personalized ranking usually consists following steps:

(1) *Prompt designing.* Firstly, we will design the ranking prompt template: "You are a ranking system now. Here is the browsing history of [demographic attributes] user: [Browsing History]. Based on this history, please rank the following candidates to [demographic attributes] user: [Candidate]".

(2) *Personalized information replacement.* For each user, we will collect user's demographic attributes if available (e.g. user names, user genders), his/her browsing history, and the ranking candidates. Then we will replace the collected user [demographic attributes], [Browsing History] and [Candidate] to replace the placeholders in the ranking prompt template.

(3) *Ranking list generation.* Then, we will feed the template into the LLMs, and the LLMs will provide a ranked list of candidate indices.

Table 1: Statics of different user names, where $|\mathcal{N}_s|$ denotes the number of user names belonging to the demographic group s.

s	Gender		Race		
	Male	Female	White	Black	Asian
$\lvert\mathcal{N}_s\rvert$	1068	1040	1175	256	463
s	Continent				
	Asia	Americas	Africa	Europe	Oceania
$\lvert\mathcal{N}_s\rvert$	463	374	136	1075	60

2.2 Implicit Ranking Unfairness

We consider the measurement as counterfactual fairness in individual-level (Wu et al., 2019; Li et al., 2023), *i.e.,* the ranking list $L_K(u)$ outputs by LLMs should be the same in the counterfactual world as in the real world. For example, if we modify a user's sensitive attribute from "male" (real world) to "female" (counterfactual world) while keeping all other characteristics constant (*e.g.,* browsing histories), the ranking list should remain unchanged. Formally, given the same personalized prompt p_u and features v_u, s_u of the user, the general ranking model $f : L_K(u) = f(p_u, v_u, s_u)$ is counterfactually fair if for any $s', s \in \mathcal{S}$:

$$P(L_K(u)|s_u = s) = P(L_K(u)|s_u = s'), \quad (1)$$

where $P(L_K(u))$ is the distribution of $L_K(u)$.

Previous works (Zhang et al., 2023b) have found that when we explicitly take the sensitive feature s_u as input user features, recommender model f often does not meet the criteria outlined in the Equation (1). Formally, we can define:

Explicit ranking unfairness: $L_K(u) = f(p_u, v_u, s_u)$, which do not satisfy Equation (1).

However, we discover that even if we mask s_u as an input in the LLMs-based ranking model f, it still yields significantly discriminatory output distributions when categorized based on different sensitive attributes s_u. Formally, we can define:

Implicit ranking unfairness: $L_K(u) = f(p_u, v_u)$, and $L_K(u)$ do not satisfy the Equation (1). Because non-sensitive attribute v_u may have a strong correlation with sensitive attribute s_u learned in the pre-training phase of LLMs.

3 Evaluation Settings

In this section, we will describe our evaluation settings including the datasets and some details.

3.1 Non-sensitive Attribute Selection

Specifically, we collect first names by choosing the most popular first names in 2014 from 229 countries (regions) across different genders, races, nationalities groups . The detailed statistic information is in Table 1. Note that a name does not necessarily have gender, race, and continent attributes simultaneously and according to our statistics, no names exist for different genders and different races.

3.2 Discrimination Measurement

Following (Gallegos et al., 2023), we utilize the metric U-Metric to measure the discrimination degree under the previous evaluation settings:

$$U(\mathcal{S}) = \sum_{s\in\mathcal{S}} |\text{Metric(s)} - \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\text{Metric(s)}|/|\mathcal{S}|,$$

where Metric(s) is the evaluation metric under s group, which can be either NDCG@K $= \frac{1}{N}\sum_{j=1}^{N} \frac{\sum_{k=1}^{K}(2^{r_k}-1)/(\log_2(j+1))}{(2^{\text{rank}_j}-1)/(\log_2(\text{rank}_j+1))}$, or other ranking metric such as MRR (Dai et al., 2023), where rank_j is the rank of the first correct answer in the ranking list $L_K(s,j)$ for user u within the top K recommendations, and r_k is a relevance score of the item with the k-th rank, which is 1 if it is a positive sample otherwise 0.

3.3 Other Settings

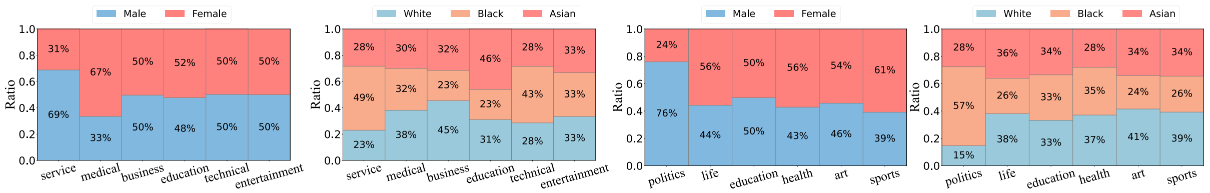
In this section, we will describe our evaluation settings including the datasets and some details.

Dataset. We utilize the two common-used ranking datasets: **MIND** (Wu et al., 2020) collected user news click behaviors on the Microsoft platform, which comprises 15,777,377 impression logs from a total of 1 million users; **CareerBuilder** is collected based on their previous online job applications, and work history. The data covers the records of 321,235 users applying for 365,668 jobs from April 1 to June 26, 2012.

Following the practice in (Dai et al., 2023; Zhang et al., 2023c), we also apply the filter criteria where both the impression list and history list are required to have more than 5 items each and sample 300 data uniformly to evaluate the LLMs in every trial.

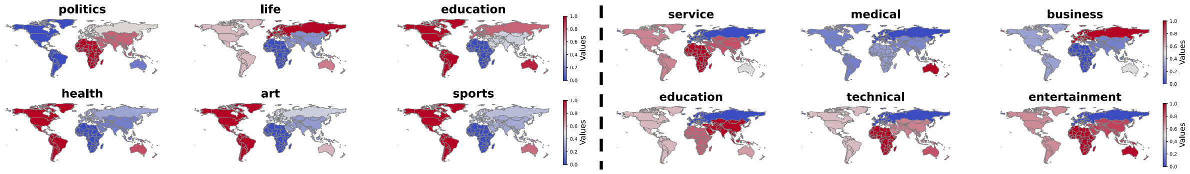
LLM Settings. In all the experiments, we utilize the ChatGPT series (gpt-3.5-turbo-xxx) and Llama2 (Touvron et al., 2023). The numbers "xxx" refer to the release or revision dates. In all LLMs, we set the maximum generated token number to

https://forebears.io/forenames/most-popular
https://platform.openai.com/



(a) Job Gender (b) Job Race (c) News Gender (d) News Race

Figure 2: The discriminatory behaviors (*i.e.,* topic distribution $P(L_K(s))$) against certain topics of LLMs under job and news domain for user names belonging to different Gender and Race groups.



(a) news (b) jobs

Figure 3: The discriminatory ranking behaviors (*i.e.,* topic distribution $P(L_K(s))$) against certain topics of LLMs under job and news domain for user names belonging to different Continent groups. A deeper red color indicates that LLMs are more likely to assign this type of news or jobs to users in the continent, while a deeper blue color suggests that LLMs are less likely to assign this type of news or jobs to users in the continent.

2048, the nucleus sampling ratio is 1, the temperature is 0.2, the penalty for frequency is 0.0, and the penalty for presence is 0.0.

4 Implicit Unfairness of LLMs

In this section, we aim to evaluate the implicit unfairness. Note that we average the different ChatGPT versions and Llama 2 (Touvron et al., 2023) results to conduct the analysis.

4.1 Existence of Implicit Unfairness

Specifically, we design N topic sentences, where several keywords of certain topics are formed into a topic sentence. The detailed topic sentence construction can be seen in Appendix B. Suppose $T_1, T_2, \cdots, T_N$ denotes the constructed topic sentence, where N denotes the topic number. The topic distribution $P(L_K(s))$ of group s is defined as $[S_1, S_2, \cdots, S_N] = \textbf{Softmax}([Z_1, Z_2, \cdots, Z_N])$, where $Z_j = \sum_{n \in \mathcal{N}_s} \sum_{i \in L_K(n)} e(T_j)^\top e(i)$. Note that we obtain the embeddings $e(i)$ by utilizing LLMs (Llama-2), extracting the hidden states of the last token, and averaging the word-level tokens to derive the final sentence embeddings.

Gender Discrimination. From the sub-figures in Figures 2(a) and 2(c), we can observe that LLMs tend to provide noticeably different responses for different genders. For example, in news recommendations, ChatGPT will deliver more political news to male users while giving more life, health, art, and sports related news to female users. In the context of job recommendations, ChatGPT tends to suggest a higher number of service-related positions to male users and an increased number of medical-related jobs to female users.

Race Discrimination. From the sub-figures in Figure 2(b) and 2(d), we find that LLMs also give different category ratios for different races. For example, LLMs will deliver more political but less art news to black users. As for job recommendations, LLMs tend to recommend more service-related but less educational jobs to black users. Meanwhile, LLMs are likely to give more business and educational jobs to white and Asian users, respectively.

Continent Discrimination. From Figure 3 we can observe that LLMs reveal stereotype bias at the geographical level. Similarly, LLMs will deliver more political news to African users while more education, health, art, and sports-related news to users in America. In the realm of job recommendations, there is a tendency for LLMs to suggest a greater number of service-oriented positions to African users, whereas it leans toward proposing more educational jobs to Asian users.

Influences for Other Attributes. We also examine whether LLMs can exhibit implicit ranking unfairness when email addresses are used as non-sensitive features. Specifically, we choose the con-

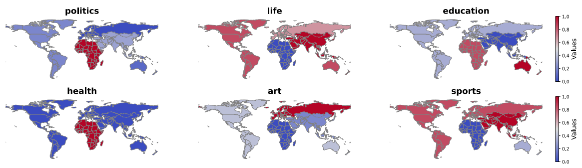
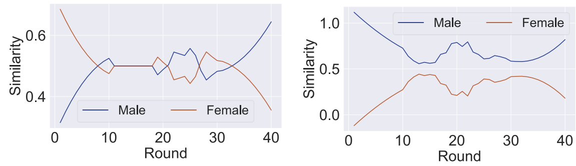


Figure 4: The discriminatory ranking behaviors against certain topics of LLMs under the news domain for user emails. A deeper red/blue color indicates that LLMs are more/less likely to assign this type of news.



(a) Education News (b) Art News

Figure 5: Similarity curves of different gender groups *w.r.t.* interaction rounds. Higher similarity denotes the LLMs will deliver more items related to topics to users.

tinental top 10 university email domain address .

From Figure 4, we can observe a similar discriminatory ranking pattern compared to the implicit ranking fairness when utilizing user names (see Figure 3). For example, LLMs will deliver more political and healthy news to users whose email domain addresses are African universities and more life and sports news to users whose email domain addresses are America's universities. The experiments also verified different non-sensitive features can all cause serious implicit user unfairness.

Implicit Unfairness During Conversation. Next, to investigate the implicit unfairness degree during the conversation process, following the practice in (Zhang et al., 2023a), we will give a simulation interactive process between the user and ranking models every round. For each round, the LLMs will give a ranking list L_K with size K according to a user's browsing history. Next, the user will select an item whose is in the first position of L_K, to serve as their browsing history for the next interaction round, since previous research has indicated that users tend to view items in higher positions (Craswell et al., 2008). The similarity is computed as **Softmax**$([S_i(\text{male}), S_i(\text{female})])$, where $S_i(\text{gender})$ is the $i-$th topic similarity under gender names.

From Figure 5 (a) and (b), we can observe that in the long term, LLMs exhibit a higher tendency to recommend unipolar news. For example, it tends to recommend more art and education news to male users than female users gradually, causing information bubbles for male and female groups.

The experiment confirmed that implicit ranking unfairness in LLMs-based ranking models may lead to more reinforced unipolar ranking results, which pose a threat to diversity and potentially trap different user groups within information bubbles.

4.2 Implicit Ranking Unfairness Degree

In this section, our objective is to investigate how is implicit ranking unfairness compared with explicit unfairness and unfairness caused by the collaborative filtering information.

Comparsion with Explicit Unfairness. In Figure 7, we compare the discrimination degrees (U-NDCG@K) under three demographic types with the explicit and implicit ranking unfairness utilizing different versions of ChatGPT and Llama2.

From Figure 7, we discern that in the evaluation at the Continent level, both the explicit and implicit ranking unfairness exhibit similar averaged discrimination measurements. However, when comparing the Gender and Race levels, we find that explicit unfairness is often lower than the implicit fairness degree by about 2-4 times. These experiments also confirm that when utilizing common demographic terms such as "Male" and "White", LLMs are more likely to cause implicit fairness.

Influence of Collaborative Filtering. Previous research indicates that collaborative filtering information utilized in ranking during pre-training may also contribute to unfairness (Yao and Huang, 2017). Therefore, we aim to conduct a simulation to investigate the unfairness degree raised by collaborative filtering (CF) information. We choose DCN (Wang et al., 2017) and GRU4Rec (Tan et al., 2016) as two commonly used ranking models for learning CF information.

Specifically, owing to the privacy policy, the dataset does not include any sensitive attributes of users. Therefore, for every user, we utilized the point-wise probing described in Section 5 to predict the sensitive attributes of a user. Specifically, for at time t, we utilized the historical clicked item sequence $[i^{t-H}, i^{t-H+1}, \cdots, i^{t-1}]$ to simulate, i.e. $\hat{s}_u = \arg\max_{s\in S} \sum_{h=1}^{H} \left(\hat{z}_s^{\text{point}}(i^{t-h})/\hat{z}\right)$, where H is the pre-defined maximum history length. Given the simulated sensitive attribute as the user context, trained a ranking model based on this con-

https://www.usnews.com/education/best-global-universities/

Table 2: Testing accuracy for probing using ChatGPT and Llama2 on news and job recommendation tasks.

demographic		gender	race	continent
news	ChatGPT	0.667	0.659	**0.510**
	Llama2	**0.833**	**0.777**	0.466
jobs	ChatGPT	0.552	0.645	0.505
	Llama2	**0.916**	**0.666**	**0.533**
random		0.500	0.333	0.200

text. In the inference phase, we mixed the data both in real-world and counterfactual world (Wu et al., 2019; Kusner et al., 2017), i.e. keeping other features constant, we replaced the user-sensitive attributes to assess the performance variation among different groups, considering this difference as a measure of unfairness degree.

From the reported results in Table 3, we can see the degree of implicit ranking unfairness in LLMs significantly outperforms all of the unfairness learned with CF information. The experiment verifies that implicit ranking unfairness does not rely on much on collaborative information but contributes to the correlation between non-sensitive attributes and sensitive attributes.

5 Implicit Ranking Unfairness Traceback

In this section, our objective is to investigate why the implicit ranking unfairness exist.

5.1 Inferring Sensitive Attribute Ability

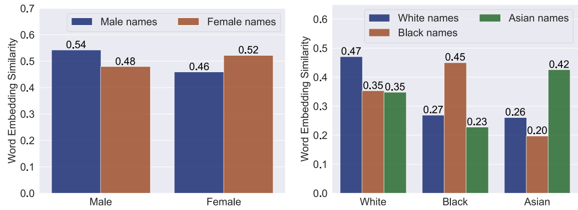
Firstly, we utilize the probing technique (Vulić et al., 2020; Gurnee and Tegmark, 2023) under two most-performing LLMs ChatGPT (Roumeliotis and Tselikas, 2023) and Llama-2 7B (Touvron et al., 2023) to investigate whether LLMs can inference the sensitive attribute from the non-sensitive attribute in terms of their wide world knowledge.

To validate the effectiveness of pair-wise regression, we also compare the commonly used point-wise probing (Gurnee and Tegmark, 2023) to predict the appropriate demographic attribute utilizing non-sensitive attributes:

$$l^{\text{point}} = \mathbb{E}_j \left[\sum_{s \in \mathcal{S}} \sum_{n \in \mathcal{N}_s} \sum_{i \in L_K(n,j)} \textbf{CE}\left(z, \hat{z}^{\text{point}}(i)\right) \right], \quad (2)$$

where $\hat{z}^{\text{point}}(i) = \textbf{MLP}\left(e(i); \theta^{\text{point}}\right)$, θ is the parameters of MLP, $\textbf{CE}(\cdot)$ denotes the cross entropy loss and the function $e(\cdot)$ represents the embedding function, as described in Section 4.1.

Note that the predicted $\hat{z}$ is s-dimensional vector, which measures the distribution of the sensitive attribute. For example, when measuring gender fairness, $\hat{z} = [0.8, 0.2]$ indicates that the sample/-pair has an 80% likelihood of being male and a 20% likelihood of being female, aiming to map the embedding to the sensitive label space as a probing technique (Gurnee and Tegmark, 2023) does. For the label z, we can get from the dataset about the real sensitive attribute of a demographic/name.



(a) Gender (b) Race

Figure 6: Word embeddings similarities between user names and sensitive attribute words.

From Table 2, we can observe that probing ability on ChatGPT and Llama2 are reliable, as they consistently outperform random probing with a substantial margin. The experiment also verifies that different LLMs both have the ability to inference sensitive attributes from the non-sensitive attribute in terms of their wide world knowledge.

5.2 Word Embedding Similarities.

Secondly, we aim to investigate whether LLMs learn a close embedding between popular names and their sensitive attributes to determine if LLMs capture their relationships at a more fine-grained level. Since we cannot get embeddings from black-box LLMs ChatGPT, we only utilize the white-box LLM Llama 2 to conduct the experiments. We extract the word embeddings from the embedding table and average the sub-word embeddings.

We compute the distance of two embeddings based on cosine similarities $\cos(\cdot)$. Formally, the similarities between the sensitive attribute s and all non-sensitive attributes $[\mathcal{N}_s]_{s \in \mathcal{S}}$ are: $\textbf{Softmax}([\cos(\boldsymbol{e}_s, \sum_{n \in \mathcal{N}_{s'}} \boldsymbol{e}_n / |\mathcal{N}_{s'}|)]_{s' \in \mathcal{S}})$, where $\boldsymbol{e}_s, \boldsymbol{e}_n$ denote the word embeddings of sensitive attribute s and non-sensitive attribute n.

From Figure 6, it is evident that at the word level, non-sensitive attributes such as user names exhibit a significant correlation with sensitive attributes. This suggests that during the pre-training phase, LLMs can effectively learn and exploit these correlations, resulting in unfair ranking outcomes.

6 Implicit Ranking Unfairness Mitigation

In this section, we propose a fair data Augmentationation method to mitigate implicit ranking un-

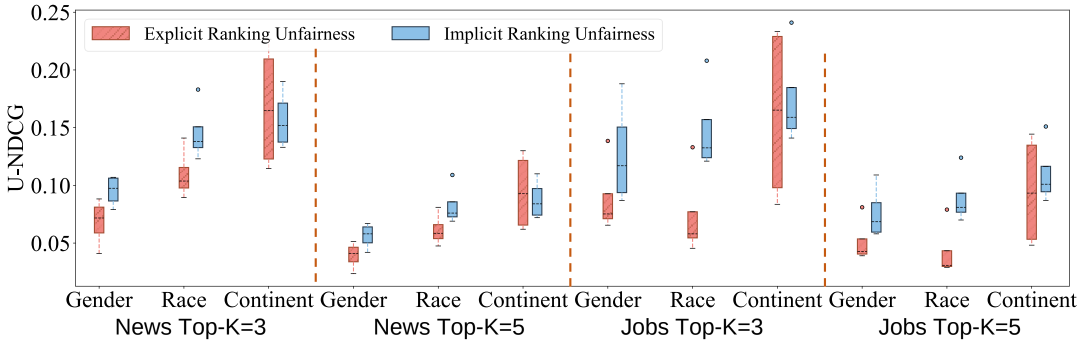


Figure 7: Comparing the averaged discrimination degrees (U-NDCG@3 and U-NDCG@5) of different versions of ChatGPT and Llama 2 under three demographic types (Gender, Race, and Continent) for news and job domain.

Table 3: Unfairness degree compared ranking models learned collaborative information from and the implicit ranking unfairness of different versions of ChatGPT. The metric is U-NDCG@5. "Improv." denotes the percentage of ChatGPT's implicit user unfairness exceeding the highest degree of unfairness brought from collaborative information.

Models	DCN	GRU4Rec	**ChatGPT**	**Improv.**
		News		
Gender	0.104	0.016	**0.203**	95.1%
Race	0.158	0.231	**0.319**	38.1%
Continent	0.324	0.158	**0.711**	119.4%
		Jobs		
Gender	0.08	0.137	**0.220**	60.6%
Race	0.043	0.110	**0.479**	335%
Continent	0.139	0.115	**0.798**	474.1%

fairness. We employ the 2SLS procedure (Kmenta, 2010) to remove the noise in non-sensitive attributes. After that, we can conduct data augmentation effectively by utilizing the top-N different feature sets that exhibit the most serious unfair behaviors in ranking.

6.1 Stage-1.

In the first stage, we utilize pair-wise regression to train a RankNet model (Burges et al., 2005), which aims to select user names that can be easily inferred from their demographic information.

In the ranking tasks, we take into account the order of the generated text within the ranking list. Ranking task implies a higher position in the ranking list L_K signifies greater importance for the associated item (Craswell et al., 2008). Therefore, we aim to investigate how LLM can infer demographic attributes through the patterns of ranking orders. Similarly, we also formulate this problem as a multi-classification task, where the class number corresponds to the demographic size $|\mathcal{S}|$.

Then, every item pair (i_j^n, i_m^n) is constructed from the ranking list $L_K(n, l)$, which takes n as a proxy for the demographic attribute in the prompts (Figure 1), where $i_j^n, i_m^n \in L_K(n, l)$ is the item in the j-th and m-th position of the ranking list, respectively with $m > j$. The pair reveals the ranking patterns in the ranking list.

Given the training data, we train the pair-wise regression network using the RankNet (Burges et al., 2005) with the loss function as

$$l^{\text{pair}} = \mathbb{E}_l \left[\sum_{s\in\mathcal{S}} \sum_{n\in\mathcal{N}_s} \sum_{j=1}^{K-1} \sum_{m=j+1}^{K} \mathbf{CE}\left(z, \hat{z}^{\text{pair}}(i_j^n, i_m^n)\right) \right], \tag{3}$$

where the loss is a expectation among different sample i and $z \in \mathbb{R}^{|\mathcal{S}|}$ is the one-hot encoding representation of true demographic label s, and $\hat{z}^{\text{pair}}(i_j, i_m) \in \mathbb{R}^{|\mathcal{S}|}$ is computed through RankNet:

$$\hat{z}^{\text{pair}}(i_j, i_m) = \mathbf{MLP}\left(e(i_j) \| e(i_m); \theta^{\text{pair}}\right),$$

where $\|$ is the concat function for two vectors and θ^{pair} is the parameter of MLP network and $e(i)$ can be obtained by averaging the hidden embeddings of Llama2 to encode the textual item i as a vector.

6.2 Stage-2

In the second stage, after deciding the parameters of RankNet, we will decide the $\mathcal{N}_s'$ for all sensitive group s to conduct data Augmentationation. Specifically, we will replace each non-sensitive attribute $n \in \mathcal{N}_s'$ to the "[demographic]" placeholder in Figure 1. In this way, one ranking sample can be augmented into $\sum_{s\in\mathcal{S}} |\mathcal{N}_s'|$ samples and feed these samples into instruction tuning phases of LLMs-based ranking tasks (Bao et al., 2023a). Specifically, we will choose the N non-sensitive attributes n of each sensitive group s. $\mathcal{N}_s'$ is defined as:

$$\mathcal{N}_s' = \underset{\mathcal{L}\in\mathcal{N}_s, |\mathcal{L}|=N}{\arg\max} \sum_{n\in\mathcal{L}} \mathbb{E}_{j<m}[\mathbf{CE}\left(z, \hat{z}^{\text{pair}}(i_j^n, i_m^n)\right)], \tag{4}$$

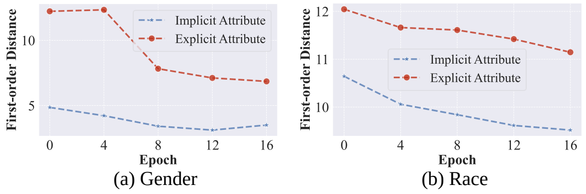


Figure 8: The first-order distance between embeddings of implicit attributes (such as user names) and embeddings of explicit attributes is measured during the tuning epochs of our method on News datasets.

where the $\hat{z}, z$ has the same meanings described in Section 5.1.

6.3 Discussion

In terms of computational costs, the complex pairwise regression and data processing occur in stage-1, which does not involve LLMs and requires small computational resources. In stage-2, to mitigate excessive computational costs, we introduce the hyper-parameter N, which specifies the number of useful user names utilized per sample to control the augmented size. Compared to other data Augmentationation method (Ghanbarzadeh et al., 2023), our time complex will reduce from $|D|^2$ to $|D|N$, where $|D|$ is the data size and $N \ll |D|$.

7 Experiments

In this section, we will conduct experiments to show the effectiveness of our methods.

7.1 Settings

The dataset and evaluation details are the same as Section 3.3. Due to the constraint of ChatGPT-series API, we only utilize Lora (Hu et al., 2021) techniques to conduct instruction tuning for ranking tasks (Bao et al., 2023a) on Llama2 by employing different fairness strategies. The experiments were conducted under four NVIDIA A5000.

For the baseline, we compare four common-used types of methods to mitigate unfairness in LLMs: (1) **Self-Align**: following the practice in Sun et al. (2023), we utilize ChatGPT-3.5 (stronger LLM) to generate more reliable and fair responses to user's queries and fine-tune the original Llama2 with the high-quality self-aligned responses. (2) **Re-Weight**: following (Jiang et al., 2024), during the tuning phase, we set the weight to be inversely proportional to the popularity of the item. (3) **Data-Augmentation** (Ghanbarzadeh et al., 2023): we replace the "[Demographic]" placeholder with explicit sensitive attribute as illustrated in Section 3.3. (4) **Prompt-Tuning** (Chisca et al., 2024): we utilize the prompt-tuning techniques to learn a fairness prompt to decrease the unfairness behaviors.

7.2 Experimental Results

In our experimental results, we mainly compare the unfairness of the most common sensitive attributes: gender and race. For the continent, we also observe a similar tendency.

In Table 4, it becomes evident that our method significantly outperforms the baselines across all datasets and sensitive attributes, encompassing different top-K ranking sizes. The experiments conclusively demonstrate that our method can mitigate the implicit ranking unfairness effectively.

Meanwhile, in Table 4, we observe that our model maintains a high level of effectiveness, with only a small drop in accuracy (around 1-2%) compared to the best-performing baselines, while achieving significantly better fairness with an improvement of nearly 10-20% for most cases. These findings indicate that our model strikes an excellent trade-off between fairness and effectiveness, addressing your concerns effectively.

7.3 Experimental Analysis

In this section, we will analyze why our method can mitigate implicit ranking unfairness. In Figure 8, we use TSNE(Van der Maaten and Hinton, 2008) to reduce the dimensionality of the vectors and calculate the distances between them to assess whether the large model reduces the distance between different groups of sensitive attributes in the ranking task.

From Figure 8, we can observe that using implicit attributes for data augmentation not only reduces the embedding distances between different implicit attributes but also brings embeddings of explicit attributes (such as "Male, Female") closer together. In this way, the LLM-based ranking model will find it difficult to infer demographic attributes from user names, thereby effectively achieving ranking fairness.

As for other experiments, Appendix C shows a case study to show how the interaction of names of users and their demographic features leads to unfairness in the context of ranking.

Meanwhile, to further investigate the intrinsic bias of LLMs, we will conduct experiments to analyze the intrinsic bias of LLMs by removing the history information/names in the ranking prompt to

Table 4: Unfairness degree (U-NDCG) and ranking accuracy degree (NDCG) compared between different models. "Improv." denotes the percentage of implicit ranking unfairness exceeding the highest degree of implicit unfairness of baselines. Bold numbers mean the improvements over the best baseline are statistically significant (t-tests and p-value < 0.05).

model/domain	**News**				**Jobs**			
	gender		race		gender		race	
	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5
Unfairness Degree (U-NDCG)								
Self-Align	0.0671	0.0379	0.0848	0.0471	0.0814	0.0464	0.1069	0.0627
Re-Weight	0.0751	0.0412	0.0807	0.0475	0.0536	0.0297	0.0501	0.0267
Data-Augmentation	0.0886	0.0498	0.0620	0.0363	0.0471	0.0264	0.0434	0.0235
Prompt-Tuning	0.0504	0.0276	0.0534	0.0297	0.0580	0.0344	0.0805	0.0459
Ours	**0.0424***	**0.0219***	**0.0526***	**0.0287***	**0.0406***	**0.0226***	**0.0356***	**0.0190***
Improv.	15.8%	20.6%	1.5%	3.4%	13.8%	14.4%	18.0%	19.1%
Accuracy Degree (NDCG)								
Self-Align	0.4485	0.6022	**0.4597**	**0.5593**	0.4603	0.6097	0.4476	0.5454
Re-Weight	**0.4540**	**0.6110**	0.4535	0.5580	0.4985	0.6413	0.5741	0.6686
Data-Augmentation	0.4434	0.6016	0.4489	0.5575	**0.5006**	**0.6442**	**0.5944**	**0.6785**
Prompt-Tuning	0.4320	0.5957	0.4139	0.5272	0.4915	0.6254	0.4427	0.5401
Ours	0.4439	0.5960	0.4395	0.5505	0.4882	0.6372	0.5896	0.6749
Improv.	-2.21%	-2.45%	-4.38%	-1.57%	-2.48%	-1.09%	-0.80%	-0.54%

compare the ranking performances with and without browsing history/names. The experiments are shown in Appendix D.

8 Related Work

Recently, researchers have discovered that LLMs can exhibit discriminatory behaviors (Gallegos et al., 2023). In previous discrimination evaluation settings, researchers often measure stereotype sentence pairs that only differ in the sensitive attribute. For example, they often adapt terms "Male" and "Female" (Nangia et al., 2020; Delobelle et al., 2022; Gallegos et al., 2023) and for Race, they often substitute terms "Black", "White" and "Asian" (Zhang et al., 2023b; Tamkin et al., 2023). Among the allocational harms, previous studies found that LLMs often exhibit discrimination against certain groups. For example, Salinas et al. (2023); de Vassimon Manela et al. (2021); McGee (2023); Thakur et al. (2023); Bolukbasi et al. (2016) discovered that LLMs will generate discriminatory content for disadvantaged gender. (Zhang et al., 2023b) show recommendation outcomes may discriminate against certain groups, see also (Rozado, 2023; Hutchinson et al., 2020). In our research, we mainly utilize the counterfactual fairness concept to measure the *implicit ranking unfairness* of LLMs-based recommendation.

There are some works that try to mitigate unfairness problems in LLMs. For example, RLHF (Ouyang et al., 2022) and RLAIF (Bai et al., 2022) try to utilize reinforcement learning to align LLMs with human values. Generally, to address the imbalance in the original dataset against certain groups, some work (Ghanbarzadeh et al., 2023; Zhang et al., 2023b; Lu et al., 2020) create matched pairs (*e.g.,* male or female) to ensure a more equitable dataset and other methods (Dixon et al., 2018; Sun et al., 2022) add non-toxic examples for groups. Other approaches (Orgad and Belinkov, 2022; Deldjoo and di Noia, 2024) suggest the use of down-weighting samples containing social group or discriminated information as a re-sampling strategy. While some method proposes to utilize the prompt-tuning method to learn a fair-aware prompt (Hua et al., 2023; Chisca et al., 2024). Moreover, other studies (Raffel et al., 2020; Ngo et al., 2021) propose to filter out and remove discriminated or taxonomic content from datasets.

9 Conclusion

In conclusion, our findings show that LLMs exhibit serious implicit ranking unfairness. This implies that, even when sensitive attributes are not explicitly provided, LLMs can still exhibit discriminatory ranking behaviors. Regarding the root causes, we find that LLMs' capability to deduce sensitive attributes from non-sensitive attributes contributes to intrinsic discriminatory knowledge. Finally, we propose to mitigate such unfairness effectively by utilizing fair-aware data augmentation. We emphasize the necessity of identifying and moderating implicit ranking unfairness in existing LLMs.

Limitations

Finally, in our paper, we mainly utilize ChatGPT, and Llama2 as our evaluation LLMs and only test the discrimination behaviors against demographic information in recommendation tasks. Meanwhile, we currently only select user names and user emails as the implicit attribute. However, different LLMs and different discrimination behaviors may exhibit different forms of implicit unfairness. This paper serves as a valuable illustration to the community, emphasizing the importance of careful consideration when assessing the discrimination behaviors in LLMs.

Ethics Statement

This study is a retrospective analysis conducted on publicly available datasets with research-oriented licenses, involving neither human participants nor the requirement for informed consent. All results generated by LLMs are utilized for offline analysis by the authors and remain invisible to real-world users, ensuring no actual social impact. User profiles used in the experiments, including names, genders, races, and nationalities, are simulated, and all user identities have been completely anonymized. The primary objective of this study is to enhance the fairness of LLMs, aligning with the principles of responsible and ethical usage.

Acknowledgements

This work was funded by the National Key R&D Program of China (2023YFA1008704), the National Natural Science Foundation of China (No. 62377044, 62276248), the Youth Innovation Promotion Association CAS (No. 2023111), Beijing Key Laboratory of Big Data Management and Analysis Methods, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Public Computing Cloud, funds for building world-class universities (disciplines) of Renmin University of China. Supported by the Outstanding Innovative Talents Cultivation Funded Programs 2024 of Renmin University of China.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yancheng Luo, Fuli Feng, Xiangnaan He, and Qi Tian. 2023a. A bi-step grounding paradigm for large language models in recommendation systems. *arXiv preprint arXiv:2308.08434*.

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023b. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. *arXiv preprint arXiv:2305.00447*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2280–2292.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Andrei-Victor Chisca, Andrei-Cristian Rad, and Camelia Lemnaru. 2024. Prompting fairness: Learning prompts for debiasing large language models. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 52–62, St. Julian's, Malta. Association for Computational Linguistics.

Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*, pages 87–94.

Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering chatgpt's capabilities in recommender systems. *Recsys*.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6437–6447, New York, NY, USA. Association for Computing Machinery.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra

Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the ACL: Main Volume*, pages 2232–2242.

Yashar Deldjoo and Tommaso di Noia. 2024. Cfair-llm: Consumer fairness evaluation in large-language model recommender system.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *NAACL 2022: the 2022 Conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pages 1693–1706.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2023. Bias and fairness in large language models: A survey.

Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. 2023. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5448–5458.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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*.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 5491–5501.

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*, pages 4717–4726.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Jan Kmenta. 2010. Mostly harmless econometrics: An empiricist's companion.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. 2023. Fairness in recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–48.

Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday*, pages 189–202.

Robert W McGee. 2023. Gender discrimination arguments and non sequiturs: A chatgpt essay. *Available at SSRN 4413432*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Helen Ngo, Cooper Raterink, João GM Araújo, Ivan Zhang, Carol Chen, Adrien Morisot, and Nicholas Frosst. 2021. Mitigating harm in language models with conditional-likelihood filtration. *arXiv preprint arXiv:2108.07790*.

Hadas Orgad and Yonatan Belinkov. 2022. Blind: Bias removal with no demographics. *arXiv preprint arXiv:2212.10563*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a name? reducing bias in bios without access to protected attributes. *arXiv preprint arXiv:1904.05233*.

Konstantinos I Roumeliotis and Nikolaos D Tselikas. 2023. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192.

David Rozado. 2023. The political biases of chatgpt. *Social Sciences*, 12(3):148.

Abel Salinas, Parth Vipul Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. 2023. The unequal opportunities of large language models: Revealing demographic bias through job recommendations. *arXiv preprint arXiv:2308.02053*.

Eric Michael Smith and Adina Williams. 2021. Hi, my name is martha: Using names to measure and mitigate bias in generative dialogue models.

Hao Sun, Zhexin Zhang, Fei Mi, Yasheng Wang, Wei Liu, Jianwei Cui, Bin Wang, Qun Liu, and Minlie Huang. 2022. Moraldial: A framework to train and evaluate moral dialogue systems via moral discussions. *arXiv preprint arXiv:2212.10720*.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*.

Yong Kiam Tan, Xinxing Xu, and Yong Liu. 2016. Improved recurrent neural networks for session-based recommendations. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 17–22.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. *arXiv preprint arXiv:2306.04597*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. *arXiv preprint arXiv:2010.05731*.

Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*, pages 1–7.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *ACL*, pages 3597–3606.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2023a. A survey on large language models for recommendation.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023b. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the twenty-eighth international joint conference on Artificial Intelligence*.

Yijia Xiao, Yiqiao Jin, Yushi Bai, Yue Wu, Xianjun Yang, Xiao Luo, Wenchao Yu, Xujiang Zhao, Yanchi Liu, Haifeng Chen, et al. 2023. Large language models can be good privacy protection learners. *arXiv preprint arXiv:2310.02469*.

Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems*, 30.

An Zhang, Leheng Sheng, Yuxin Chen, Hao Li, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2023a. On generative agents in recommendation. *arXiv preprint arXiv:2310.10108*.

Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023b. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. *arXiv preprint arXiv:2305.07609*.

Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023c. Recommendation as instruction following: A large language model empowered recommendation approach. *arXiv preprint arXiv:2305.07001*.

Table 5: Detailed symbol definitions.

symbol	explain
u	user
i	item
v_u	user u's non-sensitive attribute (e.g. names)
$s_u \in \mathcal{S}$	user u's sensitive attribute (e.g. male)
p_u	personalized prompt for each user u
K	ranking size
$L_K(u)$	ranking list to user
$\mathcal{N}_s, s \in \mathcal{S}$	user names set belong to attribute s
$e(\cdot)$	function to map text to embedding
T_i	i-th topic sentence
S_i	similarity score of i-th topic
$\hat{z}$	predicted attribute similarity distribution
(i_j, i_n)	item pair from ranking list

Appendix

A Symbol Descriptions

Given that some symbol definitions may be complex, we provide the notations in Table 5 to help readers better understand them.

B Topic Sentences

The topic sentences T_i consist of the keywords of certain topics. In the recommendation task, we choose six topics for comparison using the following prompt: *Please give me 20 keywords related to the topic news/jobs*.

After filtering some neural words, the news keywords associated with the topics are listed below.

- Politic: violence, elections, government, legislation, political, diplomacy, corruption, democracy, voting, legislation, trump, hurricane
- Life: dog, technology, travel, food, finance, environment, weather, transportation, relationships, family, career, hobbies, events, shopping
- Education: school, university, teacher, student, curriculum, exams, educational, scholarships, literacy, academic
- Health: health, fitness, pandemic, vaccine, medication, fat, sleep, nutrition, exercise, diet, death
- Art: art, venice, gallery, artist, exhibition, painting, sculpture, museum, culture, fashion, entertainment, auction, design
- Sports: sports, football, game, team, coach, basketball, baseball, swimming, athletics, exercise

After filtering some neural words, the job keywords are listed as follows.

- Service: Customer, Service, Sales, Associate, Receptionist, Waiter, Waitress, Hotel, Concierge, Flight, Attendant, Cook, Housekeeper, Lifeguard
- Health&Medical: Health, Medical, doctors, nurses, surgeons, medical, technicians, pharmacists, healthcare
- Business&Finance: Business, Finance, management, finance, accounting, marketing, entrepreneurship, administrater
- Education&Teaching: Education, Teaching, teachers, professors, tutors, school, librarians, educational, counselors
- Engineering&Technical: information, technology, computer, science, programming, software, development, network, administration, data, analysis, electrical
- Arts&Entertainment: arts, media, entertainment, actors, musicians, writers, filmmakers, designers, photographers, artists

C Case Study

In this section, we give a case study how user names interact with the sensitive attribute. We test three items (news) as the ranking [Candidate]:

- A: Some believe Mason Rudolph, hit in the head with his own helmet, isn't getting enough blame.
- B: Taylor Swift Rep Hits Back at Big Machine, Claims She's Actually Owed 7.9 Million in Unpaid Royalties.
- C: This is it, this is the luckiest break in the history of golf.

Then, we test gender discrimination by utilizing the male name *Jack* and the female name *Sophie*. The ranking results are "A,B,C" nad "B,A,C", respectively.

By analyzing the word embedding by Llama2, we find that the embedding of word *Jack* is more close to Male: the word *Jack* has 56% similarity with *Male* and 44% with *Female*). While *Sophie* is more close to *Female*: about 54% similarity with *Male* and 46% with Male. Therefore, from the

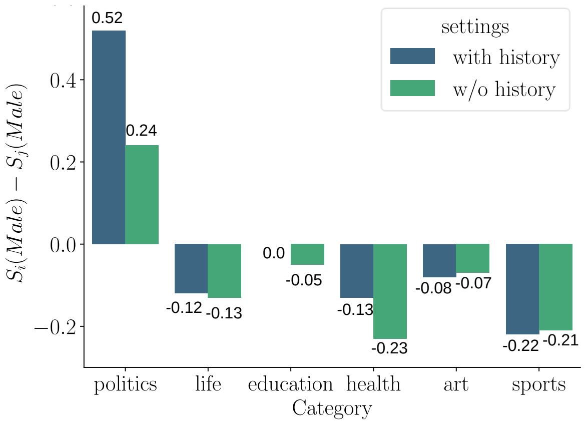


Figure 9: Topic similarities comparison with and without using the browsing history.

ranking result, we can observe user named Jack is more likely to rank male-related news (A) in a higher position, and the user named Sophie is more likely to rank female-related news (B) in a higher position. Such intrinsic bias embedded in LLMs leads to the implicit ranking unfairness behaviors of LLMs.

D Intrinsic Bias of LLMs

In this section, to further investigate the bias is only from the intrinsic bias of LLMs but not from browsing history, we will conduct experiments to analyze the intrinsic bias of LLMs by removing the history information/names in the ranking prompt to compare the ranking performances with and without browsing history/names.

D.1 Removing History

Firstly, we conducted the experiment only based on user names by utilizing the following prompt: "Please give five news/jobs ranking list to the user named [names]". Then, we analyze the topic distribution with and without using the browsing history to investigate whether their discrimination patterns are similar, assessing the extent of the browsing history's influence. The experiments were conducted under the MIND dataset and gender-based sensitive attributes.

Following the settings in Section 3.3, we compare the topic similarity gap between Male and Female: $S_i(Male) - S_j(Male)$, where $S_i(\text{gender})$ is the $i-$th topic similarity under gender names. We report the distribution value on genders in Figure 9. From the data in Figure 9, the two data groups have a Pearson correlation coefficient of 0.966, indicating a significant positive correlation.

Table 6: Ranking accuracy for removing user names. The ranking metric is NDCG@3 and "w/o names" denotes removing user names only contains user historical behaviors.

domain	w/o names	male names	female names
news	0.460	0.463	0.659
jobs	0.495	0.500	0.502

The experiments demonstrate that even when we remove the history and only utilize user names, the models still exhibit similar patterns of discriminatory ranking, confirming the intrinsic bias present in LLMs.

D.2 Removing User Names

In this section, we aim to remove the user names in the prompts to observe the bias from the user's historical browsing history. We compare the ranking performances (NDCG@3) with and without user names in Table 6. The experiments were conducted under both the news and job dataset and gender-based sensitive attributes.

From the table, we can observe that compared to the situation without names as input, the inclusion of users' names significantly amplifies the disparities in treatment among different groups, thereby verifying the intrinsic bias present in LLMs.