DEMONSTRATION

FairWork: A Generic Framework For Evaluating Fairness In LLM-Based Job Recommender System

YUHAN HU, Sun Yat-Sen University, Guangzhou, Guangdong, China

ZIYU LYU, Sun Yat-Sen University, Guangzhou, Guangdong, China

LU BAI, Beijing Normal University, Beijing, China

LIXIN CUI, Central University of Finance and Economics, Beijing, China

Published: 13 July 2025

SIGIR '25: The 48th International ACM SIGIR Conference on Research and Development in Information Retrieval
July 13 - 18, 2025
Padua, Italy

Conference Sponsors:
SIGIR

https://doi.org/10.1145/3726302.3730145
ISBN: 9798400715921

FairWork: A Generic Framework For Evaluating Fairness In LLM-Based Job Recommender System

Yuhan Hu
huyh76@mail2.sysu.edu.cn
School of Cyber Science and Technology, Sun Yat-sen University
Shenzhen, Guangdong, China

Ziyu Lyu*
lvzy7@mail.sysu.edu.cn
School of Cyber Science and Technology, Sun Yat-sen University
Shenzhen, Guangdong, China

Lu Bai
bailu@bnu.edu.cn
School of Artificial Intelligence, Beijing Normal University
Beijing, China

Lixin Cui
cuilixin@cufe.edu.cn
School of Information, Central University of Finance and Economics
Beijing, China

Abstract

Large Language Models (LLMs) have revolutionized recommender systems by offering highly personalized and context-aware suggestions. However, their inherent biases pose significant challenges in sensitive scenarios like job recommendation, potentially compromising fairness and resulting in harmful effects on both users and platforms. While previous studies have explored fairness issues in LLM-based job recommendations, they often focus on limited dimensions. We introduce FairWork, a comprehensive fairness evaluation framework to examine LLM-based recommender system from both the user's and recruiter's perspectives, employing fairness metrics to assess how sensitive user attributes influence job recommendations. The system allows stakeholders such as recruitment platforms and job seekers to upload personalized profiles and job descriptions for fairness analysis. By integrating specific job requirements and user-driven data inputs, FairWork captures the relationship between candidate qualifications and job demands. This framework provides a robust foundation for evaluating fairness in LLM-based job recommender systems and supports future research on bias mitigation strategies. The demo is available at https://github.com/chenzhouli/FairWork.

CCS Concepts

• **Information systems → Recommender systems**.

Keywords

Recommender Systems, Large Language Models, Fairness Evaluation

*Corresponding author. This work is supported by Guangdong Basic and Applied Basic Research Foundation (2023A1515012848), and CCF-DiDi GAIA Collaborative Research Funds for Young Scholars.

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730145

ACM Reference Format:
Yuhan Hu, Ziyu Lyu, Lu Bai, and Lixin Cui. 2025. FairWork: A Generic Framework For Evaluating Fairness In LLM-Based Job Recommender System. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3726302.3730145

1 Introduction

Large Language Models (LLMs) have reshaped the landscape of recommender systems [2, 6, 7, 15, 22], offering enhanced personalization and deeper contextual understanding. However, their heavy reliance on vast and often unverified Internet data raises concerns about biases related to sensitive attributes like gender, age, and race. In high-stakes scenarios such as job recommendation, these biases can lead to unequal access to opportunities, strengthen existing prejudices, and undermine fairness. They also pose serious risks for both users and platforms. For users, biases can limit options and reduce trust in the system. For platforms, they harm reputation, lower user engagement, and invite legal scrutiny. As LLM-based recommender systems increasingly shape important decisions, addressing biases is critical for ensuring equitable outcomes and maintaining user confidence.

Previous studies have investigated fairness in job recommendation tasks, focusing on the impact of sensitive user attributes. For instance, Wang et al.[19] uncovered gender biases in LLMs during resume evaluations, while Salinas et al.[16] revealed unequal treatment for different demographic groups. Although these works offer valuable insights, they are often limited in scope. For example, Wang et al.[19] examined fairness only from the recruiter's perspective, whereas Salinas et al.[16] focused solely on the user side. To the best of our knowledge, there is no systematic framework covering multiple dimensions and perspectives.

This paper addresses these gaps by proposing a comprehensive fairness evaluation framework for LLM-based job recommender systems. **First**, we measure fairness from both the user's and recruiter's perspectives, using standard fairness metrics to quantify how sensitive attributes affect recommendation results. **Second**, the system enables individual and group level assessments, where job seekers can upload personal profiles and target job descriptions,

while companies can evaluate their internal datasets. **Third**, we move beyond broad job categories by analyzing real job descriptions and role requirements, thereby capturing both applicant qualifications and each position's demands. By providing actionable insights into the fairness of current LLM models, our system supports the creation of more equitable job recommendation platforms and helps stakeholders meet ethical and legal standards. Our contributions are summarized as follow:

- A comprehensive fairness evaluation framework is proposed, measuring recommendation biases from both the user's and recruiter's perspectives.
- A flexible system is developed, supporting individual-level evaluation for job seekers and group-level evaluation for companies.
- A real-world evaluation method that focuses on actual job descriptions and role requirements instead of broad job categories, revealing deeper biases in hiring contexts.
- An accessible and user-driven system is designed for practical deployment of the fairness evaluation tool in real-world recruitment scenarios.

2 Related Work

2.1 Fairness in LLM-based Recommender Systems

User-side Fairness. Research on user-side fairness seeks to ensure that recommendations do not disadvantage particular groups based on attributes like gender or age. Shen et al.[18] uncovered unintended biases in conversational recommender systems, stemming from semantic associations in LLM training data. Hua et al.[9] introduced the Counterfactually Fair Prompt method to mitigate such biases, and Zhang et al.[21] proposed the FaiRLLM framework to benchmark fairness in ChatGPT, revealing persistent disparities in ChatGPT. Deldjoo et al.[4] further explored intersectional fairness, highlighting the complexity of overlapping user identities.

Item-side Fairness. Item-side fairness ensures that less popular items receive equitable exposure. Jiang et al.[10] addressed biases arising from historical interaction patterns and semantic factors in LLMs. Deldjoo et al.[3] examined how prompt engineering and system roles affect fairness and diversity, while Li et al.[13] investigated provider fairness in news recommendations, warning that LLM-generated content can amplify filter bubbles and popularity bias.

2.2 Fairness in Job Recommendation

Job Seekers' Perspective. Li et al.[14] developed a resource allocation model that balances fairness constraints with job opportunity distribution. Jourdan et al.[11] utilized optimal transport to reduce bias in neural network classifiers for job matching. Salinas et al.[16] analyzed demographic biases in LLM-based job suggestions, showing that models like ChatGPT and Llama often steer certain groups toward lower-paying or stereotypical roles. Xu et al.[20] studied implicit ranking unfairness and introduced data augmentation strategies to maintain accuracy while mitigating bias.

Employers' Perspective. Wang et al.[19] proposed the JobFair framework to benchmark hiring biases in LLMs, discovering that level and spread biases persist even when resume content changes.

3 Methodology

To evaluate the fairness of LLM-based job recommender systems, we focus on how users' sensitive attributes affect the user-job matching task, reflecting real-world recommendation needs. We treat user fairness as a counterfactual fairness problem: when only a user's sensitive attributes change, the system's recommendations should not differ from those in the real setting. Based on this viewpoint, our system offers two levels of evaluation: (1) individual-level and (2) group-level. Figure 1 shows the overall workflow of our system.

3.1 Data Formulation

The LLM recommender system is trained to recommend jobs for users in a natural language way. Formally, let $u \in \mathcal{U}$ denote a user.

User Profile. A user profile $x_u \in \mathcal{X}$ typically includes attributes such as educational background and work experience. To control for confounding variables, any sensitive information beyond the system-defined attributes is removed.

Job description. A job description j specifies the details and requirements of a given position.

Interaction. In the user-job matching task, the interaction for a user-job pair is a binary indicator: "yes" if the user is suitable, "no" otherwise.

Sensitive attributes. Sensitive attributes include age, gender, race, and other characteristics that are important for fairness evaluation. Formally, let $S = \{S_1, S_2, \ldots, S_k\}$ be the set of k sensitive attributes. Each attribute S_i has a finite set of possible values, denoted by $S_i = \{S_i^1, S_i^2, \ldots, S_i^{|S_i|}\}$. For a subset $I \subseteq S$ containing m attributes, the set of all possible value combinations for these attributes is $C_m(I) = \{(v_1, v_2, \ldots, v_m) \mid v_i \in S_i,\ S_i \in I,\ |I| = m\}$. Each element in $C_m(I)$ corresponds to one attribute combination to be injected into a user profile.

Example. Suppose we have $k = 3$ sensitive attributes: Gender (S_1), Age (S_2), and Race (S_3), with $S_1 = \{\text{Male}, \text{Female}\}$, $S_2 = \{< 35, > 35\}$, $S_3 = \{\text{Black}, \text{White}, \text{Asian}\}$. If $I = \{S_1, S_2\}$ ($m = 2$), then $C_2(I) = \{(\text{Male}, < 35), (\text{Male}, > 35), (\text{Female}, < 35), (\text{Female}, > 35)\}$. A single-attribute combination is also valid (e.g., $C_1(S_1) = \{(\text{Male}), (\text{Female})\}$).

3.2 Counterfactual Analysis

To systematically evaluate fairness with respect to sensitive attributes in LLM-based job recommender system, we compare recommendations generated after perturbing different sensitive attribute combinations into the user profile. By varying these attributes, we can observe how sensitive information impacts the recommendations.

Perturb. Given a set of sensitive attributes S, we first list all possible attribute combinations for perturbation. For instance, if $S = \{\text{Age}, \text{Gender}\}$, we consider all $m \in \{1, 2\}$ levels of combinations, from single attribute to intersectional attributes. We then modify the original user profile x_u by concatenating it with each combination

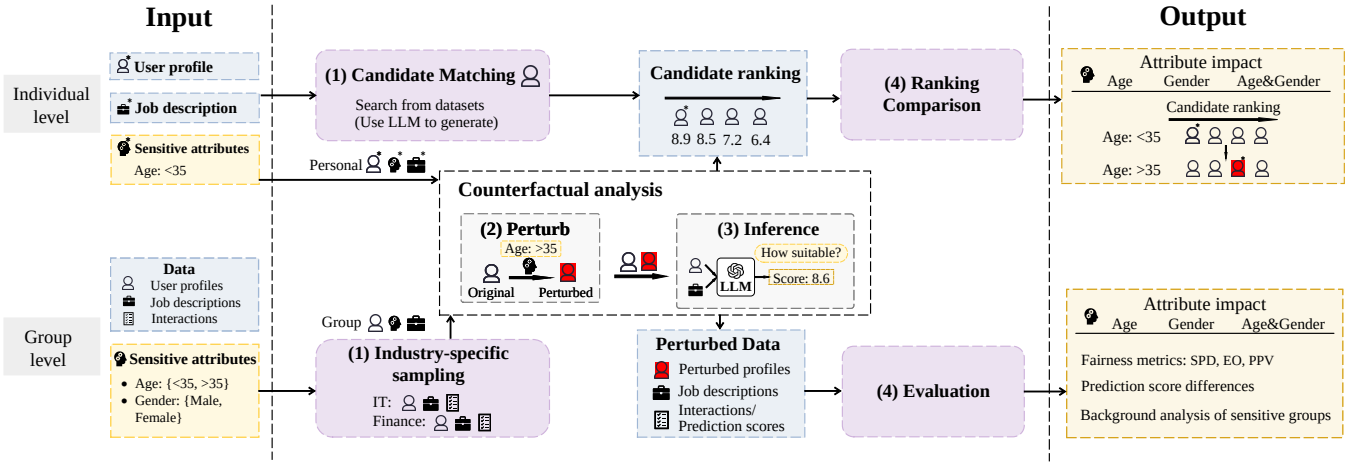


Figure 1: The framework workflow for individual and group level fairness evaluation.

$C_u \in C_m(I)$, resulting in a perturbed user profile

$$x'_u = \text{Perturb}(x_u, C_u) \tag{1}$$

Inference. For both the original profile x_u and each perturbed profile x'_u, we use the same prompt method to query the LLM and obtain the prediction score for each user-job pair. Formally,

$$y_{u,j} = \text{LLM}(x_u, j), \quad \hat{y}_{u,j} = \text{LLM}(x'_u, j) \tag{2}$$

where j represents a specific job.

3.3 Individual-level Workflow

For individual-level evaluation, we focus on a common scenario where a recruiter compares a few closely matched candidates, one of the most likely situations in which bias may arise.

3.3.1 ***Input.*** A single job seeker can upload a user profile and a job description for evaluation. The user can also select sensitive attribute values according to personal circumstances.

3.3.2 ***Candidate Matching.*** We first retrieve a public dataset for candidates with profiles similar to the user's, using TF-IDF [17] text similarity. If not enough real profiles are found, we use an LLM to generate synthetic profiles.

3.3.3 ***Perturb.*** The original sensitive attribute values are perturbed by generating alternative values drawn from the set of sensitive attributes. This process creates different scenarios based on the original profile while maintaining all other information unchanged.

3.3.4 ***Inference.*** We obtain prediction scores for all candidates, including the current user and other competing candidates.

3.3.5 ***Ranking Comparison.*** We then rank all candidates by their prediction scores. By comparing these rankings under different attribute perturbations, the sensitive attributes might be revealed as they result in changes within candidate rankings.

3.4 Group-level Workflow

For group-level evaluation, the system provides fairness metrics from both users and recruiters' perspective, and identifies user subgroups with high sensitivity to attribute changes, which can reveal deeper fairness problems.

3.4.1 ***Input.*** The user should input two types of required information, respectively a related dataset and the specified sensitive attributes. The dataset should include user profiles, job descriptions and optional interaction labels.

3.4.2 ***Industry-specific sampling.*** From the selected dataset, we group users and jobs by industry, since people typically apply for jobs based on their professional domains. This also alleviates cold-start problems for users or jobs with limited historical interactions. We then sample M users; for each user, we sample N jobs in a similar industry. For users with labeled interactions, we select N_1 positive and N_2 negative instances where possible.

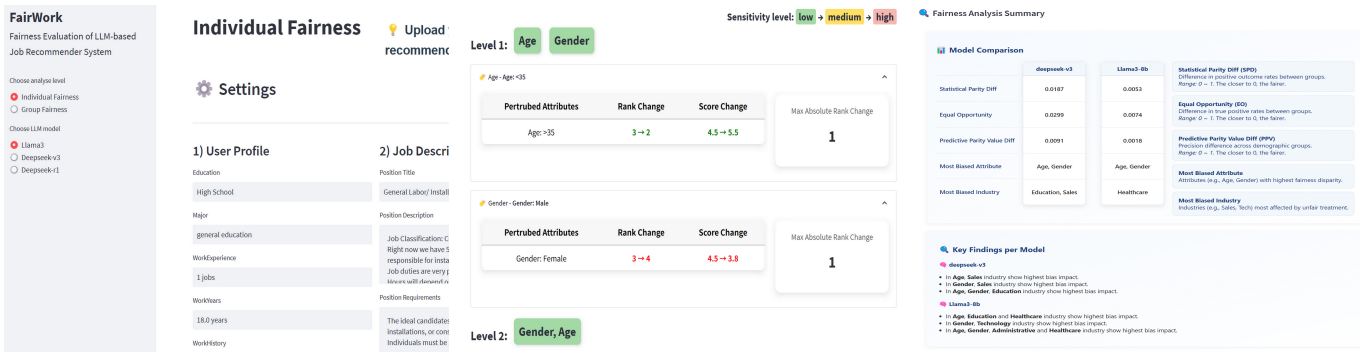
3.4.3 ***Perturb.*** For each user u, all perturbed profiles are generated by combining the original profile with every possible combination of sensitive attributes. The set of perturbed profiles is denoted by $\mathcal{P}$, where each element x'_u corresponds to a unique attribute combination for that user.

3.4.4 ***Inference.*** For each user u, we obtain LLM prediction scores $y_{u,j}$ and $\hat{y}_{u,j}$ for the N sampled jobs, using both the original and perturbed profiles.

3.4.5 ***Evaluation.*** We perform fairness evaluation through three approaches.

Fairness Metrics. We use three standard metrics to assess whether different sensitive attribute groups receive systematically different outcomes. These include Statistical Parity (SP) [5], Equal Opportunity (EO) [8], and the Predictive Parity Value difference (PPV_diff) [12]. From the user perspective, these metrics indicate if users with different sensitive attributes experience different prediction outcomes. From the recruiter perspective, they reveal whether certain industries exhibit higher biases.

Prediction Score Differences. To quantify the overall sensitivity of the model to changes in user attributes, we define a perturbation impact for each user. For a given user u, the impact is computed



(a) User Input interface (b) Individual-level output (c) Group-level output

Figure 2: Demonstration of FairWork. Figure (a) displays user input interface, Figure (b) and (c) show individual and group-level evaluation.

as:$\Delta_u(x_u, x'_u) = \frac{1}{N}\sum_{j=1}^{N}\left|\hat{y}_{u,j} - y_{u,j}\right|$, where N is the number of job samples. The average prediction score difference across all users is then given by:

$$\text{Pred\_diff} = \frac{1}{M}\sum_{u=1}^{M}\max_{x_u \in \mathcal{X}, x'_u \in \mathcal{P}} \Delta_u(x_u, x'_u) \tag{3}$$

This metric captures the maximum change in recommendation scores due to sensitive attribute perturbations.

Background Analysis of Sensitive Groups. We further compute a user-specific sensitivity score:

$$\Delta_u^I = \max_{x_u \in \mathcal{X}, x'_u \in \mathcal{P}} \Delta_u(x_u, x'_u) \tag{4}$$

Based on these scores, users are categorized into high-sensitivity and low-sensitivity groups. High scores indicate strong influence of sensitive attributes; low scores indicate minimal influence. To examine how user backgrounds relate to the observed bias, we visualize employment histories with word-clouds and report the corresponding distributions of education and primary industries.

The number of possible attribute combinations grows exponentially with each additional sensitive dimension, which greatly increases computational cost. To address this issue, we introduce an adaptive injection strategy: we randomly sample a subset of users from each group to maintain diversity and manage computational resources, since intersectional attributes that may not individually show significant differences can still amplify impacts when considered together.

4 System Setting

Datasets. We include the public CareerBuilder job recommendation dataset[1], which contains records of 321,235 users applying for 365,668 jobs. Each record includes a user's job application history and basic work experience. We preprocess the data by removing noise and then cluster users and jobs into different industries.

5 Demonstration

Figure 2 illustrates both individual-level and group-level demonstrations. In Figures 2(a), the left sidebar displays key configuration options. Users can upload the necessary information in the settings panel.

Case Study. At the group level, we randomly sample 150 users and select 5 jobs per user. We set a prediction threshold of 5.0 and evaluate two sensitive attributes, Age and Gender, on both *Deepseek-v3* and *Llama3-8b*. To ensure deterministic outputs, we set *temperature* and *top_p* to zero. The overall results are shown in Figure 2(c).

User Perspective. We focus on how sensitive attributes (Age, Gender) and their combinations (Age & Gender) affect predictions.

- **Intersectional Effects:** Across both models, the Age & Gender combination results in higher bias scores than either attribute alone, indicating the compounding effect of intersecting sensitive attributes.
- Among the two models, Deepseek-v3 exhibits overall larger group fairness gaps, with a SPD of 0.0187 and EO of 0.0299, compared to Llama3-8b's 0.0053 and 0.0074 respectively.

Recruiter Perspective. We group jobs by industry and evaluate how fairness varies across industries.

- **Deepseek-v3:** Bias impact is highest in the Sales industry for both Age and Gender individually, and in the Education industry for the combined Age & Gender attributes.
- **Llama3-8b:** Age-related bias is most evident in the Education and Healthcare industries; Gender-related bias peaks in Technology; and the combined influence of Age and Gender is greatest in Administrative and Healthcare industries.

6 Conclusion

In this work, we propose a framework for evaluating fairness in LLM-based job recommender systems from both individual and group level. By systematically perturbing sensitive attribute combinations into user profiles, we examine how such attributes influence recommendation. Our group-level evaluation uncovers industry-specific fairness issues across different LLMs, highlighting disparities in access to job opportunities. This system provides insights for improving fairness in job recommendations and lays the groundwork for further refinement of fairness evaluation.

References

[1] Road Warrior Ben Hamner and Wojciech Krupa. 2012. *Job Recommendation Challenge*. https://www.kaggle.com/competitions/job-recommendation/ Kaggle.

[2] 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. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1126–1132.

[3] Yashar Deldjoo. 2024. Understanding biases in chatgpt-based recommender systems: Provider fairness, temporal stability, and recency. *ACM Transactions on Recommender Systems* (2024).

[4] Yashar Deldjoo and Tommaso Di Noia. 2025. CFaiRLLM: Consumer Fairness Evaluation in Large-Language Model Recommender System. *ACM Trans. Intell. Syst. Technol.* (2025). doi:10.1145/3725853

[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[6] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*. 299–315.

[7] Shijie Geng, Juntao Tan, Shuchang Liu, Zuohui Fu, and Yongfeng Zhang. 2023. VIP5: Towards Multimodal Foundation Models for Recommendation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 9606–9620. doi:10.18653/v1/2023.findings-emnlp.644

[8] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[9] Wenyue Hua, Yingqiang Ge, Shuyuan Xu, Jianchao Ji, Zelong Li, and Yongfeng Zhang. 2024. UP5: Unbiased Foundation Model for Fairness-aware Recommendation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, St. Julian's, Malta, 1899–1912. https://aclanthology.org/2024.eacl-long.114/

[10] Meng Jiang, Keqin Bao, Jizhi Zhang, Wenjie Wang, Zhengyi Yang, Fuli Feng, and Xiangnan He. 2024. Item-side Fairness of Large Language Model-based Recommendation System. In *Proceedings of the ACM on Web Conference 2024*. 4717–4726.

[11] Fanny Jourdan, Titon Tshiongo Kaninku, Nicholas Asher, Jean-Michel Loubes, and Laurent Risser. 2023. How optimal transport can tackle gender biases in multi-class neural network classifiers for job recommendations. *Algorithms* 16, 3 (2023), 174.

[12] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[13] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023. A preliminary study of chatgpt on news recommendation: Personalization, provider fairness, fake news. *arXiv preprint arXiv:2306.10702* (2023).

[14] Yunqi Li, Michiharu Yamashita, Hanxiong Chen, Dongwon Lee, and Yongfeng Zhang. 2023. Fairness in job recommendation under quantity constraints. In *AAAI-23 Workshop on AI for Web Advertising*.

[15] Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. 2024. Llara: Large language-recommendation assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1785–1795.

[16] Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. 2023. The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–15.

[17] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing Management* 24, 5 (1988), 513–523. doi:10.1016/0306-4573(88)90021-0

[18] Tianshu Shen, Jiaru Li, Mohamed Reda Bouadjenek, Zheda Mai, and Scott Sanner. 2023. Towards understanding and mitigating unintended biases in language model-driven conversational recommendation. *Information Processing & Management* 60, 1 (2023), 103139.

[19] Ze Wang, Zekun Wu, Xin Guan, Michael Thaler, Adriano Koshiyama, Skylar Lu, Sachin Beepath, Ediz Ertekin, and Maria Perez-Ortiz. 2024. JobFair: A Framework for Benchmarking Gender Hiring Bias in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, Miami, Florida, USA, 3227–3246. doi:10.18653/v1/2024.findings-emnlp.184

[20] 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*. 7957–7970.

[21] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 993–999.

[22] Junjie Zhang, Yupeng Hou, Ruobing Xie, Wenqi Sun, Julian McAuley, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2024. Agentcf: Collaborative learning with autonomous language agents for recommender systems. In *Proceedings of the ACM Web Conference 2024*. 3679–3689.