# How Trustworthy is AI?
# A Deep Dive into the Bias in LLM-Based Recommendations

**Anonymous ACL submission**


## Abstract

Large Language Model (LLM)-based recommendation systems provide more comprehensive recommendations than traditional systems by deeply analyzing content and user behavior. However, these systems often exhibit biases, favoring mainstream content while marginalizing non-traditional options due to skewed training data. This study investigates the intricate relationship between bias and LLM-based recommendation systems, with a focus on music, song, and book recommendations across diverse demographic and cultural groups. Through a comprehensive analysis, this paper evaluates the impact of bias on recommendation outcomes and assesses various strategies, such as prompt engineering and hyperparameter optimization, for bias mitigation. Our findings indicate that neither prompt engineering nor hyperparameter optimization are particularly effective in mitigating biases, highlighting the need for further research in this area.


## 1 Introduction

Consider an LLM-based music recommendation system that enhances user experience by leveraging the advanced capabilities of large language models. Traditional algorithms typically rely on user listening history and genre preferences. In contrast, an LLM-based system delves deeper into musical content and user behavior. For example, a user who frequently listens to progressive and alternative rock would benefit from recommendations generated through a comprehensive analysis of genres like psychedelic rock. By considering lyrical themes, musical styles, and emotional tones, the system can suggest tracks from emerging artists in related rock genres, showcasing the nuanced and highly personalized recommendations LLMs can provide.

However, such a personalized recommendation system has drawbacks. Users from Western countries may predominantly receive recommendations for mainstream Western genres like pop or rock, while underrepresented genres, such as traditional indigenous music or world music, receive limited exposure. This bias stems from training data skewed towards popular Western music. Thus, bias in recommendation systems has emerged as a critical concern, impacting fairness, diversity, and societal equity. While bias in traditional systems has been extensively studied (Mansoury et al., 2020; Abdollahpouri et al., 2021, 2019; Kordzadeh and Ghasemaghaei, 2022), integrating LLMs introduces new challenges. Due to their massive scale and ability to learn intricate patterns from vast datasets, LLMs can amplify existing biases, leading to skewed recommendations that perpetuate societal inequalities.

Recent studies have critically examined the performance and fairness of LLM-based recommendation systems. Wan et al. (Wan et al., 2023) and Plaza-del-Arco et al. (Plaza-del Arco et al., 2024) analyzed gender biases in reference letters and emotion attribution, revealing significant gendered stereotypes. Naous et al. (Naous et al., 2023) highlighted cultural biases in multilingual LLMs, while Zhang et al. (Zhang et al., 2023) found that music and movie recommendations can perpetuate existing biases. Xu et al. (Xu et al., 2023a) studied implicit user unfairness, and Sah et al. (Sah et al., 2024) explored personality profiling to enhance fairness. However, these studies often focus on specific biases or contexts, underscoring the need for a comprehensive approach to address the multifaceted nature of biases in LLM-based recommendation systems.

This paper aims to address the limitations of previous studies by exploring the intricate relationship between bias and LLM-based recommendation systems, shedding light on the underlying mechanisms that contribute to bias propagation and its implications for users and society at large. Furthermore, we investigate various techniques to evaluate their effectiveness for bias mitigations.

The rest of the paper is organized as follows: Section 2 provides an overview of LLM-based recommendation systems and our problem formulation. Section 3 describes the synthesis of our experimental data using LLMs. Section 4 delivers an in-depth analysis of the inherent biases of LLMs, offering both qualitative and quantitative insights. Section 5 analyzes the performance of two different techniques with a focus on bias mitigation. Finally, Section 6 discusses the implications and concludes with insights for practitioners and researchers.

# 2 Background and Problem Formulation

## 2.1 Related Works

Research on social biases in NLP models distinguishes between allocational and representational harms (Blodgett et al., 2020; Wang et al., 2022a). Studies focus on evaluating and mitigating biases in Natural Language Understanding (Dev et al., 2021; Bordia and Bowman, 2019) and Generation tasks (Sheng et al., 2021, 2020; Dinan et al., 2019). Metrics like the Odds Ratio (OR) (Szumilas, 2010) measure gender biases in items with large frequency differences (Sun and Peng, 2021). Controlling NLG model biases has been explored (Cao et al., 2022; Gupta et al., 2022), but applicability to closed API-based LLMs is uncertain. Emphasizing social and technical aspects is crucial for understanding bias sources (Wang et al., 2022b; Ovalle et al., 2023). Social science research highlights the detrimental effects of gender biases in professional documents, underscoring the need for grounded bias definitions and metrics (Khan et al., 2023).

Significant work has also analyzed cultural bias in language models (LMs). Recent studies have explored cultural alignment by examining encoded moral knowledge and cultural variations in moral judgments (Hämmerl et al., 2022; Xu et al., 2023b; Ramezani and Xu, 2023). LMs often reflect the moral values of specific societies and political ideologies, such as American values and liberalism (Abdulhai et al., 2023; Johnson et al., 2022). Research has also investigated LMs' understanding of cross-cultural differences in values and beliefs, and their opinions on political and global topics (Cao et al., 2023; Arora et al., 2022; Feng et al., 2023). Cultural surveys and questions probing culture-related commonsense knowledge show LMs tend to align with Western values across multiple languages (Wang et al., 2023; Masoud et al., 2023). Additionally, studies have examined LMs' knowledge of geo-diverse facts, cultural norms, culinary customs, and social norm reasoning (Nguyen et al., 2023; Palta and Rudinger, 2023; Huang and Yang, 2023).

## 2.2 Problem Formulation

Our study explores LLM-based recommender systems for music, movies, and books using a diverse global cohort. By inputting user information and categorizing recommendations by genre, we aim to assess content distribution and identify demographic and cultural biases. ***Our objectives are to understand recommendation variations across different contexts and evaluate techniques for bias mitigation.***

**Demographic Bias:** Analyzing demographic bias in LLM-based recommendation systems uncovers substantial issues arising from historical disparities and cultural consumption patterns. These systems often rely on biased training data, leading to recommendations that disproportionately favor certain demographics while neglecting others. For instance, mainstream music genres popular among specific age groups or cultural backgrounds are overrepresented, marginalizing less popular styles. Similarly, in books and movies, demographic bias perpetuates dominant cultural narratives, limiting exposure to works from underrepresented communities.

**Cultural Bias:** Examining cultural bias in LLM-based recommendation systems reveals significant issues rooted in entrenched cultural norms. These systems frequently prioritize mainstream content, thereby overlooking diverse and alternative cultural expressions, perpetuating cultural homogeneity and marginalizing underrepresented voices. For instance, LLM algorithms may tend to recommend commercially successful Western pop music over traditional folk music from other cultures, thereby limiting exposure to diverse musical traditions. Such cultural bias hinders cross-cultural understanding, exacerbates inequalities, and diminishes the richness of human cultural experiences.

# 3 Data Synthesis and Acquisition

## 3.1 Prompt Design

In this study, we investigate three distinct scenarios involving the recommendation of *songs*, *movies*, and *books* tailored to individuals from diverse demographic and cultural backgrounds. Utilizing a LLM-based recommendation system, specifically

GPT-3.5, we aim to uncover potential biases by incorporating relevant demographic (or cultural) information into the prompt generation process.

#### 3.1.1 Context-Less Generation (CLG)

For CLG, we employ a straightforward prompt to generate recommendations without incorporating additional contextual information. For analyzing demographic bias, we include demographic information in the prompt. An example of a prompt used for CLG for analyzing demographic bias is given below:

> **Ashley** is a **40-year-old female chef**. Can you recommend 25 **movies** for her?

Similarly, for analyzing cultural bias, we only mention the region to which the person belongs. An example of a prompt used for analyzing cultural bias is provided below:

> Can you recommend 25 movies for **Mateo**, who is from the **South America region**?

#### 3.1.2 Context-Based Generation (CBG)

We extend the CLG approach to develop prompts for CBG. Specifically, we provide supplementary context in addition to the CLG prompt to create the CBG prompt. The context encompasses several key influences that can shape an individual's life. Specifically, we address the following questions:

- Did the person grow up in an **affluent** family or an **impoverished** family?
- Are they **introverted** or **extroverted** by nature?
- Do they currently live in a **rural** or **metropolitan** area?

Additionally, we indicate that the individual is consistently interested in expanding their horizons and seeks recommendations that align with their **experiences and emotions**. The additional context of CBG covers this information. A sample CBG prompt is shown below:

> Ashley is a 40-year-old female chef. Can you recommend 25 movies for her? She was raised in an **affluent** family and is **introvert** in nature. Currently, she resides in a **rural** region. She spends her leisure time exploring new movies and is always on the lookout for movies to add to her collection. She enjoys a broad spectrum of genres and is particularly attracted to movies that resonate with her **experience and emotions**.

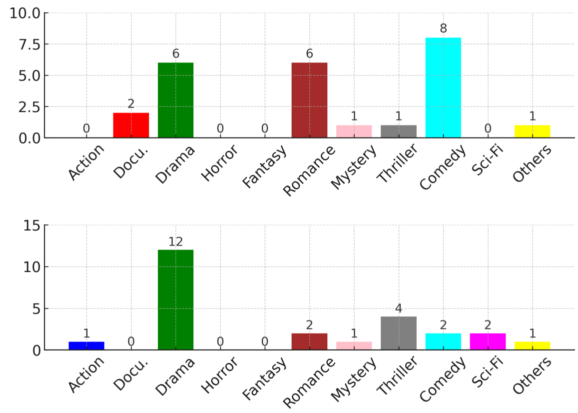



Figure 1: Genre distribution for the recommended 25 movies for Ashley, a 40-year old female chef (top) and Thomas, a 50-year old male writer (bottom)

### 3.2 Methodology for Genre Classification

Following the prompt design and generation phase, we retrieve and classify the recommendations provided by GPT-3.5 into different genres. Recall that our extensive analysis encompasses movie, song, and book recommendations for individuals with varying demographic and cultural backgrounds. For genre classification, we have considered the top ten prevalent genres suggested by ChatGPT. If a suggested movie does not fit within any of these predefined genres, it is categorized under "others."

#### 3.2.1 Genre Distribution Comparison

In Fig. 1, we present the distribution of suggested movies for Ashley, the 40-year-old female chef and Thomas, the 50-year-old male writer, showcasing how the recommendations align with various genres. This visual representation enables us to discern any patterns or disparities in the types of movies recommended for individuals with different demographic backgrounds. ***For example, there is a hint that GPT-3.5 may suggest more romantic movies to the females compared to males.***

#### 3.2.2 KL-Divergence Analysis

In this section, we provide an example to quantitatively measure the divergence in genre preferences and recommendations across various socioeconomic backgrounds, specifically occupations. We analyze how the LLM-based recommendation system suggests movies from different genres to individuals from different occupations. Kullback-Leibler Divergence (KLD) (Kullback and Leibler, 1951) is an ideal metric for such analysis as it quantifies how one probability distribution diverges from another. *A higher KLD value indicates that*

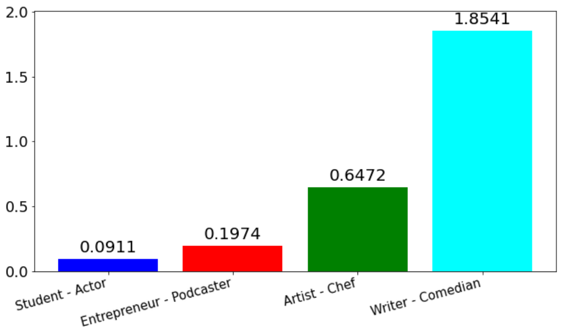



Figure 2: KL divergence between LLM-recommended movie genres for different occupation pairs.

*the two distributions being compared are less similar, suggesting a more pronounced bias or divergence between them.*

Fig. 2 demonstrates a corresponding comparison of KLD values for the genre distribution among different pairs of occupations. For example, the LLM-based movie recommendations exhibit greater divergence between writers and comedians compared to entrepreneurs and podcasters. ***This disparity arises because the LLM-based system recommends significantly more comedy movies to "comedians", whereas this preference is less pronounced for "writers."***

## 4 Bias in LLM Recommendations

This section examines the demographic and cultural biases in LLM recommendations, comparing how these biases manifest in context-less generation (CLG) and context-based generation (CBG) prompts. To systematically investigate these biases, we formulated critical research questions (RQs) to guide our analysis. These RQs help us understand the extent and nature of biases in LLM outputs. By addressing these questions, we aim to uncover underlying bias patterns and assess how context influences LLM recommendations.

### 4.1 Context-less generation (CLG)

To explore potential biases in LLM-based recommendation systems, we begin by analyzing recommendations generated in context-less generation (CLG). We focus on whether and how LLMs' recommendations for books, songs, and movies show demographic and cultural biases, guided by a specific research question.

**RQ1**: Do certain genres of books, movies, or songs receive more frequent recommendations within the CLG?

To investigate this, we analyze the number of books, songs, and movies recommended from various genres within the context-less generation (CLG) framework. We identified several significant instances of bias. We define a metric, normalized fraction, $F_a$, representing the fraction of recommendations from genre $a$ among the analyzed cases. Figures 3a-3c illustrate demographic biases in LLM-based recommendations, highlighting gender, age, and occupation biases.

In Fig. 3a, we observe gender bias in movie recommendations. ***It is evident that the system suggests more romantic movies to females and more thriller and sci-fi movies to males.*** Similarly, Fig. 3b shows age bias in song recommendations, ***with fewer hip-hop and more blues songs suggested as age increases.***

Lastly, Fig. 3c reveals occupation bias in book recommendations. ***Writers receive more fiction book suggestions than comedians or students, while comedians get more biographies***. This might be because biographies provide material for comedians to create relatable stories, while fiction helps writers develop novel ideas.

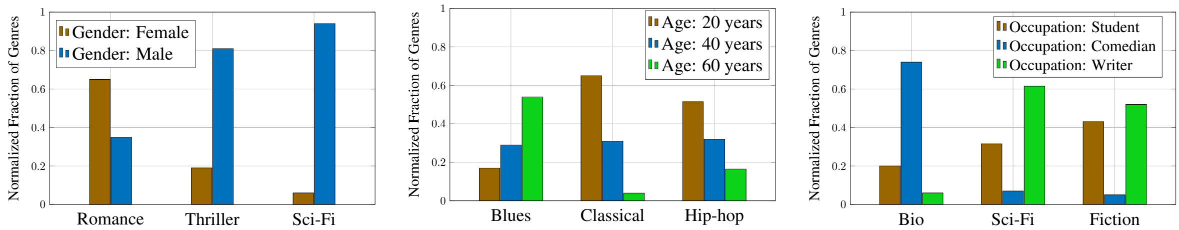
Furthermore, Fig. 4 shows cultural bias in LLM-based recommendations. ***North Americans receive more sci-fi movie suggestions compared to Western Europeans or South Asians***. Conversely, ***Western Europeans get more romantic book recommendations than the other groups***. This indicates significant cultural bias in the recommendation system within CLG.

Next, we state the following research question to address the impact of the bias developed by intersecting identities (e.g., occupation and gender).

**RQ2**: Do intersecting identities, (e.g., occupation and gender combined) have an additional impact on the recommendations produced by the LLM within CLG?

To address this, we analyzed the number of recommendations for various genres across different scenarios, observing how biases change with multiple identities. We found significant shifts in overall recommendation patterns when specific identities were added.

Fig. 5 illustrates the movie recommender system's bias. Generally, it suggests more romantic movies to females than males, with a normalized ratio of $0.65 : 0.35$. However, male dancers receive slightly more romantic recommendations than fe-




(a) Gender bias in movie recommendations

(b) Age bias in song recommendations

(c) Occupation bias in book recommendations

Figure 3: Demographic Bias in the LLM-based recommendation system (for movies, songs and books) within CLG

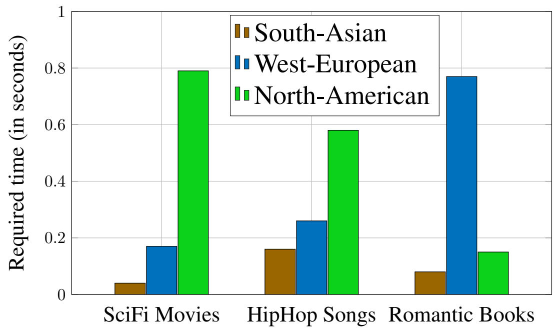



Figure 4: Cultural bias in movies, songs and books recommendations

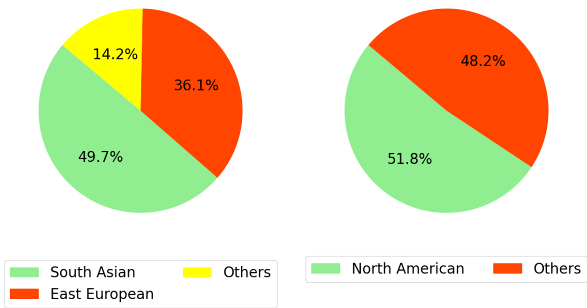



Figure 6: (a) Classical music is highly suggested to South-Asians and East-Europeans people, and (b) SciFi movies are highly suggested to North-Americans

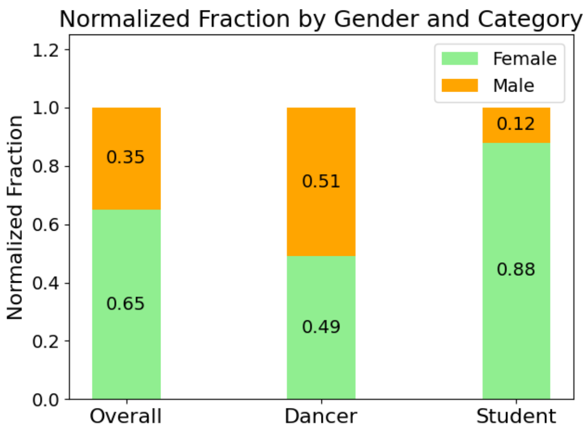



Figure 5: Impact of Bias for intersecting identities.

male dancers (0.51 : 0.49). Conversely, female students receive significantly more romantic recommendations than male students (0.88 : 0.12). This shows that occupation further impacts gender bias in LLM-based recommendations.

To delve further, we pose the following research question and address it with careful analysis.

> **RQ3**: Do certain groups tend to receive recommendations (by the LLM within CLG) that are more stereotypical or less diverse compared to others?

In order to address this, we observe the numbers (of movies, songs or books) of recommended genres in different scenarios, and analyze if there are any particular stereotypes within different groups.

We present two examples of cultural bias in recommendation systems. First, song recommendations show a disparity: ***users from South Asia and Eastern Europe receive more classical music than those from other regions***, as shown in Figure 6a. Second, movie recommendations reveal that ***North American users are disproportionately suggested science fiction movies***, as depicted in Figure 6b.

These findings reveal cultural stereotypes in LLM-based recommendation systems, as shown by biased content suggestions for for users from different backgrounds. This suggests the algorithms perpetuate cultural biases rather than providing balanced recommendations.

### 4.2 Context-based generation (CBG)

We now analyze LLM-based recommendations within CBG (context-based generations) and investigate the ***impact of context*** compared to CLG. To explore this systematically, we state the following research problems and address them with examples.

> **RQ4**: How does the bias in recommendations vary between CLG and CBG?

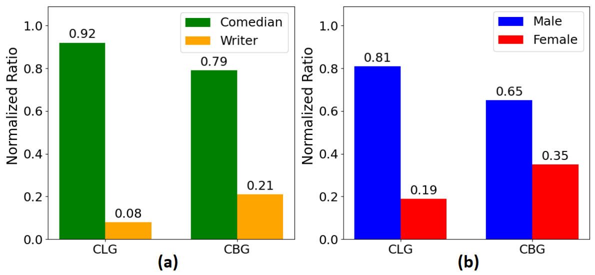



Figure 7: Variation between CLG and CBG

| Question | Metric Values | | |
|---|---|---|---|
| | SPD | DI | EOD |
| FQ1 | 0.211 | 1.633 | -0.106 |
| | 0.256 | $\infty$ | 0.667 |
| FQ2 | 0.333 | $\infty$ | 0.667 |
| | 0.182 | $\infty$ | 0.750 |
| FQ3 | 1 | $\infty$ | 1 |
| | 0.941 | 0.059 | 0.059 |
| FQ4 | -1 | 0 | 1 |

Table 1: Fairness Metrics Values.

We observe the number of genres recommended (movies, songs, books) within CBG, similar to CLG cases. First, we explore occupation bias in recommending biographic books. In CLG, comedians receive more biographic book suggestions than writers (ratio $0.92 : 0.08$). However, with the presence of different contexts in CBG, this ratio reduces to $0.79 : 0.21$, as shown in Fig. 7a.

Another example in Fig. 3a shows that in CLG, LLM-based recommendations predominantly suggest thriller movies to males. However, with different contexts, more thriller movies are recommended to females. Fig. 7b depicts this change in the normalized ratio of thriller movie recommendations to males and females.

**RQ5**: Do the recommendations exhibit bias depending on the context in CBG?

To investigate this, we analyze the numbers of recommendations in different scenario of varying contexts, and observe some interesting events. For example, the LLM-based system ***suggests blues or classical songs more to introverts and HipHop songs more to extroverts***, indicating an obvious bias, as shown in Fig. 8a.

In addition, as we observe in Fig. 8b, ***SciFi movies are significantly more recommended to affluent people compared to the impoverished ones***, whereas ***dramas are more recommended to the impoverished people***. Furthermore, from Fig. 8c, we notice that ***HipHop songs are more recommended to the metro area people, while country songs are more recommended to the rural area people***. These results indicate a considerable bias of the LLM-based recommendation system depending on the context within CBG.

## 4.3 Fairness Measures

This section analyzes three fairness measures: Statistical Parity Difference (SPD), Disparate Impact (DI), and Equal Opportunity Difference (EOD), to quantify bias in LLM-based recommendations.

### 4.3.1 Metrics Definitions

Let us consider a dataset $D = (X, Y, Z)$, where $X$ represents the training data, $Y$ denotes the binary classification labels, and $Z$ is the sensitive attribute such as ethnicity. Additionally, predicted label is indicated by $\hat{Y}$.

**Statistical Parity Difference (SPD)** assesses whether the probability of receiving a favorable outcome ($\hat{Y} = 1$) is the same for different groups. Mathematically, it is defined as follows:

$$SPD = P(\hat{Y} = 1 \mid Z = Q) - P(\hat{Y} = 1 \mid Z = \bar{Q}). \quad (1)$$

An SPD of zero indicates complete fairness, meaning that the model does not favor one group over another in terms of favorable outcomes.

**Disparate Impact (DI)** measures the ratio of favorable outcome probabilities between groups. It is expressed as follows:

$$DI = \frac{P(\hat{Y} = 1 \mid Z = Q)}{P(\hat{Y} = 1 \mid Z = \bar{Q})}. \quad (2)$$

A DI of one signifies complete fairness, indicating that both groups have an equal proportion of favorable outcomes.

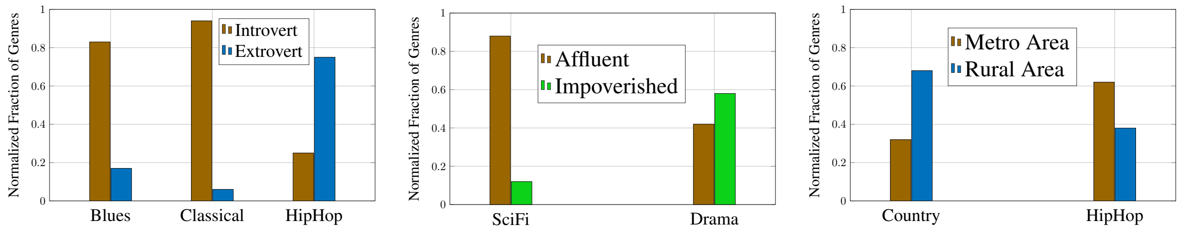
**Equal Opportunity Difference (EOD)** evaluates whether the probability of receiving a favorable outcome given the true positive label ($Y = 1$) is the same for different groups. An EOD of zero suggests complete fairness. It is calculated as follows:

$$\begin{aligned} EOD = {} & P(\hat{Y} = 1 \mid Z = Q, Y = 1) \\ & - P(\hat{Y} = 1 \mid Z = \bar{Q}, Y = 1). \end{aligned} \quad (3)$$

### 4.3.2 Fairness Questions of Interest

We shall now address several fairness-related questions (FQs) and utilize these metrics to evaluate the bias present in the recommendations. The metric values are presented in Table 1.

**FQ1**: (a) Do LLM-based recommendations suggest more romantic movies to females compared to males? (b) Conversely, do they recommend more Sci-Fi movies to males?




(a) Bias in song recommendations depending on personality

(b) Bias in movie recommendations depending on financial condition

(c) Bias in song recommendations depending on living area

Figure 8: Bias in the LLM-based recommendation system within CBG depending on the context

We answer this question by analyzing how likely women are to receive the average number of romantic movie suggestions compared to men and how likely men are to receive the average number of Sci-Fi movie suggestions compared to women.

In the first segment, an SPD of 0.211 indicates that females receive romantic movie recommendations 21.1% more frequently than males. The DI of 1.633 further shows that females are 1.633 times more likely to receive an average amount of romantic movie recommendations compared to males. However, an EOD of $-0.106$ reveals that despite the higher recommendation rate for females, the true positive rate is lower, suggesting less accurate or relevant recommendations for females. In the second segment, high values of both SPD and EOD for science fiction movie recommendations indicate that these recommendations are more frequent and accurate for males. ***Notably, the DI being infinite highlights that no female is receiving an average amount of science fiction movie recommendations, underscoring a significant gender disparity in the recommendation system.***

> **FQ2**: (a) Do LLM-based recommendations suggest more hip-hop songs to younger individuals compared to older ones? (b) Conversely, do they recommend more blues songs to older individuals?

Similar to FQ1, we shall answer this question by evaluating how likely younger individuals are to receive the average number of hip-hop song suggestions compared to older individuals, and similarly for blues songs with older individuals. By examining the fairness metric values of FQ2 from Table 1, we observe significant disparities across different age groups in terms of music genre preferences. Blues music demonstrates a noticeable bias in favor of older individuals, indicated by an SPD of 0.18 and an EOD of 0.75. Conversely, rap music exhibits a strong preference for younger listeners, as reflected by an SPD of 0.33 and an EOD of 0.67. ***In both instances, the DI is infinite, signifying a substantial bias.***

> **FQ3**: (a) Do LLM-based recommendations suggest more non-fiction books to chefs compared to writers? (b) Conversely, do they recommend more fiction to writers?

When comparing non-fiction book preferences between chefs and writers, both SPD and EOD are 1.0 (refer to Table 1, indicating a perfect preference for chefs in the non-fiction genre. An ***infinite DI*** further exacerbates this bias. In contrast, the bias towards writers for the fiction genre is less pronounced, as indicated by the smaller values of DI and EOD. However, writers still receive high recommendations for the fiction genre, as evidenced by the high SPD.

> **FQ4**: Do LLM-based recommendations suggest more Mystery movies to North Americans compared to South Asians?

From the metric values presented in Table 1, it is evident that individuals in North America have a significantly lower probability of receiving a mystery movie suggestion compared to individuals residing in South America. Furthermore, individuals in North America are considerably less likely to be accurately identified as interested in the Mystery movie genre.

#### 4.3.3 Discussion on DI = $\infty$

As seen in Table 1, several instances show DI $= \infty$. To address this, we ask: "Do LLM-based recommendations suggest more Sci-Fi movies to males compared to females?"

To compute the DI metric, a threshold was established by calculating the mean number of Sci-Fi movie recommendations for all users (including both males and females). Closer analysis revealed a significant imbalance: only 17 Sci-Fi movies were recommended to females, compared to 258 for males. This higher number of recommendations for males skewed the mean (and therefore, the threshold) upward. Consequently, ***no female user was recommended at least the average number of Sci-Fi movies***, resulting in a DI = $\infty$. While this is an extreme case, it highlights the strong stereotypes present in LLM-based recommendations.

# 5 Evaluating Bias Mitigation Strategies

This section examines the performance of two specific techniques—prompt engineering and hyperparameter optimization—focusing on their effectiveness in mitigating bias.

## 5.1 Prompt Engineering

Prompt engineering can be employed to craft prompts that ensure LLMs produce fair, unbiased, and high-quality responses by meticulously considering phrasing, context, and inclusivity. In our approach, we appended the following additional instruction to each of our prompt to ensure the fair and robust recommendations: "***Ensure that the recommendations are inclusive of various demographic and cultural groups.***"

## 5.2 Hyperparameter Optimization

The optimized hyperparameters (max tokens and temperature) were selected to minimize the sum of KL divergence between different demographic or cultural groups. Max tokens ensure responses are focused and contextually complete, while a lower temperature reduces randomness, making the model adhere to probable responses. Details of the optimized parameters are in Table 2.

| Parameter | Books | Songs | Movies |
|---|---|---|---|
| Max Tokens | 75 | 100 | 75 |
| Temperature | 0.8 | 0.7 | 0.8 |

Table 2: Optimized Hyperparameter Values.

## 5.3 Comparative Analysis

This section will discuss the performance of these techniques in mitigating demographic bias, focusing on FQ1 to FQ3. The results of the fairness metric values for Prompt Engineering are shown in Table 3, and Table 4 shows the results for the Hyperparameter Optimization technique.

| Question | SPD | DI | EOD |
|---|---|---|---|
| **FQ1 (a)** | -0.089 | 0.778 | -0.234 |
| **FQ2 (a)** | 0.364 | $\infty$ | 0.526 |
| **FQ3 (a)** | 1.0 | $\infty$ | 1.0 |

Table 3: Fairness Metric Values for Prompt Engineering.

| Question | SPD | DI | EOD |
|---|---|---|---|
| **FQ1 (a)** | 0.133 | 2.333 | 0.035 |
| **FQ2 (a)** | 0.091 | $\infty$ | 0.50 |
| **FQ3 (a)** | 0.524 | 3.882 | 0.206 |

Table 4: Fairness Metric Values for Hyperparameter Optimization.

From Table 3, it is evident that the Prompt Engineering technique consistently reduces bias for FQ1 (a), demonstrating improvements in both the SPD and EOD, and shows a slight improvement in FQ2 (a), while exhibiting no change for FQ3 (a). Conversely, the Hyperparameter Optimization technique, as shown in Table 4, achieves significant reductions in bias for FQ1 (a) and FQ3 (a), particularly in the EOD of FQ3 (a). However, it introduces a concerning increase in bias for DI of FQ1 (a) and leaves the infinite DI in FQ2 (a) unchanged. ***Therefore, while Prompt Engineering demonstrates more stable effectiveness, Hyperparameter Optimization offers substantial bias reduction potential but with greater variability and risk of increasing bias in certain areas. Nonetheless, neither approach achieves significant bias reduction across all fairness measures.***

# 6 Conclusion and Future Work

In this paper, we identified and highlighted various demographic and cultural biases in LLM-based recommendations. By formulating and answering several research questions, we gained insights into how these biases persists in LLM. We quantified the biases using fairness metrics and illustrated our findings through detailed visualizations. Despite exploring Prompt Engineering and Hyperparameter Optimization as mitigation approaches, neither method consistently addressed all fairness metrics. This underscores the complexity of mitigating bias in LLMs and suggests a more nuanced approach may be necessary. Future research should develop and test strategies to ensure AI systems are equitable across diverse demographic and cultural contexts.

## Limitations

While our work has addressed several recent potential issues, we want to mention that our work has several limitations that warrant consideration. These are briefly described below.

**Limited Dataset**: We used a limited range of demographic and cultural information, such as focusing on binary gender groups. This may not comprehensively represent the diversity of real-world populations. Future studies should address fairness for minority groups, including non-binary individuals and various racial, ethnic, and socio-economic backgrounds.

**Specific Recommendation System**: Our analysis was centered on GPT-3.5 due to its widespread accessibility and popularity. Even though ChatGPT has approximately 180.5 million users globally (Topics, 2024; Sage, 2024), this focus limits the applicability of our findings to other language models, particularly those with multimodal capabilities and advanced architectures.

**Limited Contexts**: Our Context-Based Generation (CBG) analysis was limited to specific contexts like individual nature, current residence, and upbringing. Including factors such as educational background, professional experiences, and social influences could provide a more comprehensive understanding.

**Limited Analysis**: We developed five research questions and four fairness questions, but many other relevant questions remain unexplored. Future research should address additional aspects of fairness, such as intersectional biases and the impact of AI on marginalized communities, to provide a more comprehensive understanding of AI fairness.

**Mitigation Techniques**: We explored Prompt Engineering and Hyperparameter Optimization to mitigate biases. However, these approaches did not comprehensively address the biases. More nuanced methods may be necessary for effective mitigation.

## Ethical Considerations

This study investigates biases in LLM-based recommendation systems, focusing on music, song, and book recommendations across diverse demographic and cultural groups using GPT-3.5. Our findings reveal that such models can inadvertently reinforce existing biases, disproportionately affecting marginalized communities. Despite evaluating bias mitigation techniques like prompt engineering and hyperparameter optimization, we found them insufficient, highlighting the need for more effective solutions. While this study does not involve real user data, thus avoiding direct privacy concerns, it emphasizes the importance of transparency and accountability in AI systems. We advocate for the development of fairer, more inclusive AI technologies and adhere to ethical standards that promote responsible AI use, contributing to the broader discourse on ethical AI practices.

## References


Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing popularity bias in recommender systems with personalized re-ranking. *arXiv preprint arXiv:1901.07555*.

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. 2021. User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM conference on user modeling, adaptation and personalization*, pages 119–129.

Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2023. Moral foundations of large language models. *arXiv preprint arXiv:2310.15337*.

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2022. Probing pre-trained language models for cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. *arXiv preprint arXiv:2203.13928*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, et al. 2021. On measures of biases and harms in nlp. *arXiv preprint arXiv:2108.03362*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2019. Queens are powerful too: Mitigating gender bias in dialogue generation. *arXiv preprint arXiv:1911.03842*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. 2022. Mitigating gender bias in distilled language models via counterfactual role reversal. *arXiv preprint arXiv:2203.12574*.

Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovickỳ, Constantin A Rothkopf, Alexander Fraser, and Kristian Kersting. 2022. Speaking multiple languages affects the moral bias of language models. *arXiv preprint arXiv:2211.07733*.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Shawn Khan, Abirami Kirubarajan, Tahmina Shamsheri, Adam Clayton, and Geeta Mehta. 2023. Gender bias in reference letters for residency and academic medicine: a systematic review. *Postgraduate medical journal*, 99(1170):272–278.

Nima Kordzadeh and Maryam Ghasemaghaei. 2022. Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3):388–409.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2145–2148.

Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.

Tarek Naous, Michael J Ryan, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pages 1907–1917.

Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. 2023. Factoring the matrix of domination: A critical review and reimagination of intersectionality in ai fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 496–511.

Shramay Palta and Rachel Rudinger. 2023. Fork: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. *arXiv preprint arXiv:2403.03121*.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. *arXiv preprint arXiv:2306.01857*.

Demand Sage. 2024. Chatgpt statistics (june 2024) — users, growth & revenue. https://www.demandsage.com/chatgpt-statistics/. Accessed: 2024-06-15.

Chandan Kumar Sah, Dr Lian Xiaoli, and Muhammad Mirajul Islam. 2024. Unveiling bias in fairness evaluations of large language models: A critical literature review of music and movie recommendation systems. *arXiv preprint arXiv:2401.04057*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2020. " nice try, kiddo": Investigating ad hominems in dialogue responses. *arXiv preprint arXiv:2010.12820*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*.

Jiao Sun and Nanyun Peng. 2021. Men are elected, women are married: Events gender bias on wikipedia. *arXiv preprint arXiv:2106.01601*.

Magdalena Szumilas. 2010. Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, 19(3):227.

Exploding Topics. 2024. Number of chatgpt users (jun 2024). https://explodingtopics.com/blog/chatgpt-users. Accessed: 2024-06-15.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022a. Measuring representational harms in image captioning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 324–335.

Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. 2022b. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 336–349.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R Lyu. 2023. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *arXiv preprint arXiv:2310.12481*.

Chen Xu, Wenjie Wang, Yuxin Li, Liang Pang, Jun Xu, and Tat-Seng Chua. 2023a. Do llms implicitly exhibit user discrimination in recommendation? an empirical study. *arXiv preprint arXiv:2311.07054*.

Chunpu Xu, Steffi Chern, Ethan Chern, Ge Zhang, Zekun Wang, Ruibo Liu, Jing Li, Jie Fu, and Pengfei Liu. 2023b. Align on the fly: Adapting chatbot behavior to established norms. *arXiv preprint arXiv:2312.15907*.

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*, pages 993–999.

# A Details of Demographic and Cultural Information

## A.1 Demographic Information Descriptors

The descriptors for demographic information are similar to those used by Wan et al. (Wan et al., 2023). We have employed their demographic descriptors, as detailed in Table 5, to generate the prompts for our work on analyzing demographic bias.

| **Demo_Feature** | **Descriptor Items** |
|---|---|
| Female Names | [Kelly, Jessica, Ashley, Emily, Alice] |
| Male Names | [Joseph, Ronald, Bob, John, Thomas] |
| Occupations | [Student, Entrepreneur, Actor, Artist, Chef, Comedian, Dancer, Model, Musician, Podcaster, Athlete, Writer] |
| Ages | [20, 30, 40, 50, 60] |

Table 5: Descriptors for Demographic Bias Analysis

## A.2 Cultural Information Descriptors

For generating the descriptors for cultural bias analysis, we employed our own approach by first creating a list of regions and then asking ChatGPT to provide a list of the most prominent names for each region. We subsequently concatenated these names to compile our final list. The details are provided in Table 6.

| **Cultural Features** | **Descriptor Items** |
|---|---|
| General Names | [Li Wei, Kim Yoo-jung, Sato Yuki, Aarav, Muhammad, Fahim, Nur Aisyah, Nguyen Van Anh, Putu Ayu, Luca, Emma, Sofia, Aleksandr, Jan, Anna, Liam, Olivia, Santiago, Sofia, Mateo, Maria, Oliver, Charlotte, Mia, Mohamed, Youssef, Ahmed, Amina, Grace, John] |
| Regions | [East Asia, Southeast Asia, South Asia, Western Europe, Eastern Europe, Oceania, North America, North Africa, South America, Sub-Saharan Africa] |

Table 6: Descriptors for Cultural Bias Analysis

# B Top 10 Genre List

The details of the top ten genres, as recommended by ChatGPT are provided in Table 7. If a suggested movie does not fit within any of these predefined genres, it is categorized under "others."

Subsequently, we used the following prompt to assign the genre for each of the recommendations:

> Based on the following genres: {list_of_top_10_genres}, what is the most likely genre for {specific_recommendation}? Please respond only with the most likely genre name.

Even though we explicitly instructed the model

| Topic | Top Ten Genres |
|---|---|
| Books | Mystery, Thriller, Romance, Horror, Science Fiction (Sci-Fi), Fantasy, Biography, Fiction, Historical Fiction, Non-Fiction |
| Movies | Action, Documentary, Drama, Horror, Fantasy, Romance, Mystery, Thriller, Comedy, Science Fiction (Sci-Fi) |
| Songs | Rock, R&B, Country, Jazz, Blues, Reggae, Classical, Pop, Hip Hop, EDM (Electronic Dance Music) |

Table 7: Top Ten Genres Recommended by ChatGPT

to provide the most likely genre name from a specified list, there were numerous instances where the responses included genre names not present in the list. These cases were categorized as "Others."

# C Details of Overall Recommendations for Each Genre

In this section, we present the details of overall genre recommendations for each of the selected topics, namely books, movies, and songs.

## C.1 Genre Recommendation for Books

In our study, we analyzed the genre recommendations for books to understand the distribution of demographic bias and culture bias across different genres. The details of these genre recommendations are provided in Table 8.

| Genre | Demo Bias | Culture Bias |
|---|---|---|
| Non-fiction | 6793 | 274 |
| Biography | 2717 | 329 |
| Fiction | 1127 | 2248 |
| Hist. Fiction | 1042 | 2361 |
| Romance | 539 | 433 |
| Mystery | 387 | 653 |
| Sci-Fi | 252 | 225 |
| Fantasy | 207 | 392 |
| Thriller | 104 | 67 |
| Horror | 35 | 42 |
| Other | 1797 | 476 |

Table 8: Overall recommendation of different genres of Books

## C.2 Genre Recommendation for Movies

In addition to books, we also analyzed the genre recommendations for movies to investigate the distribution of demographic bias and culture bias. The details of these genre recommendations for movies are provided in Table 9.

| Genre | Demo Bias | Culture Bias |
|---|---|---|
| Drama | 7060 | 3756 |
| Romance | 2957 | 658 |
| Comedy | 2458 | 301 |
| Thriller | 664 | 410 |
| Documentary | 439 | 80 |
| Action | 278 | 526 |
| Sci-Fi | 275 | 218 |
| Fantasy | 169 | 237 |
| Mystery | 133 | 216 |
| Horror | 86 | 287 |
| Other | 481 | 811 |

Table 9: Overall recommendation of different genres of Movies

## C.3 Genre Recommendation for Songs

Similarly, we analyzed the genre recommendations for songs to examine the distribution of demo bias and culture bias. The details of these genre recommendations for songs are provided in Table 10.

| Genre | Demo Bias | Culture Bias |
|---|---|---|
| Pop | 6092 | 3341 |
| Rock | 3674 | 485 |
| R&B | 1398 | 256 |
| Hip Hop | 804 | 382 |
| Jazz | 346 | 126 |
| Country | 275 | 111 |
| EDM | 213 | 180 |
| Classical | 161 | 155 |
| Blues | 140 | 56 |
| Reggae | 60 | 451 |
| Other | 1837 | 1957 |

Table 10: Overall recommendation of different genres of Songs