Understanding Intrinsic Socioeconomic Biases in Large Language Models

Mina Arzaghi[1,2], **Florian Carichon**[2], **Golnoosh Farnadi**[1,3]
[1]MILA - Quebec AI Institute, 6666 Saint-Urbain Street, #200 Montreal, QC, H2S 3H1
[2] HEC Montreal, 3000 Chem. de la Côte-Sainte-Catherine, Montréal, QC H3T 2A7
[3] McGill University, 845 Sherbrooke St W, Montreal, Quebec H3A 0G4
mina.arzaghi@mila.quebec, florian.carichon@hec.ca, farnadig@mila.quebec

Abstract

Large Language Models (LLMs) are increasingly integrated into critical decision-making processes, such as loan approvals and visa applications, where inherent biases can lead to discriminatory outcomes. In this paper, we examine the nuanced relationship between demographic attributes and socioeconomic biases in LLMs, a crucial yet understudied area of fairness in LLMs. We introduce a novel dataset of one million English sentences to systematically quantify socioeconomic biases across various demographic groups. Our findings reveal pervasive socioeconomic biases in both established models such as GPT-2 and state-of-the-art models like Llama 2 and Falcon. We demonstrate that these biases are significantly amplified when considering intersectionality, with LLMs exhibiting a remarkable capacity to extract multiple demographic attributes from names and then correlate them with specific socioeconomic biases. This research highlights the urgent necessity for proactive and robust bias mitigation techniques to safeguard against discriminatory outcomes when deploying these powerful models in critical real-world applications.**Warning: This paper discusses and contains content that can be offensive or upsetting.**

Introduction

In recent years, Large Language Models (LLMs) have been increasingly integrated into various fields such as healthcare (Hadi et al. 2023), insurance (Dimri et al. 2019), employment (Qin et al. 2018), and credit scoring (Wu et al. 2023). While the evaluation of social biases in LLMs (Nadeem, Bethke, and Reddy 2021; Kaneko and Bollegala 2022) and their potential harms (Blodgett et al. 2020; Beukeboom and Burgers 2019) has been extensively studied, the rapid development and integration of these models into critical decision-making areas necessitate the ongoing need to assess and address their inherent biases.

Previous research by Shen et al. (2022) has demonstrated how seemingly innocuous details like names or subtle language cues can significantly influence the outcomes of language-based recommender systems, skewing the price of restaurant recommendations based on perceived gender, race, or sexual orientation. This work, however, did not study the root causes of this bias. Motivated by this gap, we investigate how LLMs inherently exhibit biases related to socioeconomic status, particularly when considering different demographic groups. Concurrent with our research, Singh et al. (2024) explored LLMs' understanding and empathy towards underprivileged populations in extreme situations. While both studies offer valuable insights, they have not fully addressed how demographic factors influence and amplify harmful socioeconomic biases in LLMs. This paper aims to fill this research gap and contribute to a more nuanced understanding of bias and fairness in LLMs.

In this study, we uncover intrinsic socioeconomic biases in state-of-the-art LLMs such as Llama 2 and Falcon, as well as the widely-used models such as BERT and GPT-2. We focus on evaluating these biases across different demographic attributes -including birth-assigned gender, marital status, race, and religion - and exploring how these biases manifest and vary across models. Furthermore, we assess the dynamics between these biases by exploring intersectionality among gender, race, and marital status. Moreover, we evaluate the socioeconomic bias of LLMs towards names, which is an identification of the individuals used in most applications, such as credit scoring. In particular, we highlight the influence of these factors on the modification of economic class perception related to these groups through the intrinsic token prediction evaluation of LLMs. More precisely, our contributions include:

1. Creating a novel evaluation dataset of 1M English sentences with socioeconomic context prompts.[1]
2. Assessing the intrinsic socioeconomic biases in four different LLMs: Falcon, Llama 2, GPT-2 and BERT with respect to four sensitive domains.
3. Evaluating the impact of intersectionality of gender, race and marital status on socioeconomic biases in LLMs.
4. Assessing the capacity of LLMs to extract race and gender information from names and their socioeconomic biases associated with these names.

Related Works

Our paper contributes to the extensive literature on intrinsic bias evaluation in LLMs.

[1] Our dataset and code are publicly available at https://github.com/MinaArzaghi/Understanding_Intrinsic_Socioeconomic_Biases_in_Large_Language_Models.

Audit & Evaluation of LLMs: Bias in LLMs is categorized into two main types (Delobelle et al. 2022; Goldfarb-Tarrant et al. 2020): *intrinsic bias* and *extrinsic bias*. *Intrinsic bias* (Caliskan, Bryson, and Narayanan 2017; Kaneko and Bollegala 2022) refers to biases inherent in the embedding or representation of language models. This form of bias is evaluated without fine-tuning the models on a specific task or dataset. In contrast, *extrinsic bias* involves assessing the fairness of system outputs for the downstream tasks. This assessment helps to determine the overall fairness of combined system components, e.g., evaluating fairness in career prediction using individuals' biographies (Webster et al. 2020; Zhao et al. 2020).

Shen et al. (2022) demonstrated that including names in the queries for a language-based recommender system named *LMRec*, which suggests restaurants, can shift the price levels of the recommended restaurants. This shift occurs despite the absence of names during the system's training phase. However, they did not identify which component of the CRS is responsible for this bias. In our work, we demonstrate how Large Language Models (LLMs) can extract race and gender from names—a finding supported by other studies (Haim, Salinas, and Nyarko 2024; Meltzer, Lambourne, and Grandi 2024)—and exhibit biases toward names and various demographic groups. Specifically, our study evaluates the intrinsic biases in four LLMs: BERT, GPT-2, Falcon, and Llama 2. We concentrate on the perceived socioeconomic status as influenced by individuals' demographic information. More precisely, we investigate how the embeddings in LLMs reflect biases toward individuals' financial status, influenced by gender, marital status, and a combination of these.

Dataset Generation to Audit LLMs: The task of generating datasets for evaluating bias in LLMs involves two main approaches: unmasking tokens within sentences (Nadeem, Bethke, and Reddy 2021; Zhao et al. 2018) and selecting sentences based on given contexts (Nangia et al. 2020; Kiritchenko and Mohammad 2018). In the first approach, LLMs are tasked with filling in a masked token, considering the sentence's context. The second approach requires the LLM to choose a sentence that aligns with a given context. Zhao et al. (2018) introduced the WinoBias dataset, comprising 3,160 sentences designed for co-reference resolution tasks to identify gender bias in associating genders with occupations. Based on their evaluation method, an unbiased model should be able to link gendered pronouns to an anti-stereotypical occupation as accurately as it does to stereo-typically linked occupations. *StereoSet*, a dataset introduced by Nadeem, Bethke, and Reddy (2021), includes 16,995 crowd-sourced instances. This dataset focuses on stereotypes related to gender, race, religion, and profession, where LLMs are evaluated based on their ability to select contextually relevant tokens.

We contribute to this field by introducing a new dataset of 1M masked tokens specifically designed to assess socioeconomic bias in LLMs. This fill-in-the-blank task presents options reflecting poor and wealthy statuses and neutral as potential fills for the masked token. While previous works like WinoBias and StereoSet offer valuable resources in the bias evaluation field, our dataset focuses exclusively on socioeconomic biases.

Socioeconomic Bias in LLMs: Investigating socioeconomic biases in LLMs remains a relatively unexplored area of research. A notable exception is the work of Singh et al. (2024), which focused on the ability of LLMs to demonstrate empathy towards socioeconomically disadvantaged individuals in challenging situations. Using a dataset of 3,000 samples and a question-answering approach. In this work, they used terms such as ”homeless” to represent the socioeconomic situation. Our work contributes to this field by showing that perceived socioeconomic status is influenced by individuals' demographic information. Our work demonstrates that biases are not solely directed toward the economically disadvantaged; even groups considered affluent are subject to specific perceptions and labeling by language models. This comprehensive evaluation offers critical insights into the implicit biases present in current linguistic technologies.

Dataset Creation

Our study aimed to assess socioeconomic biases within LLMs, focusing on four crucial demographic domains: Birth-Assigned Gender, Marital Status, Race, and Religion. We proposed a novel dataset of 1M English sentences to evaluate the impact of demographic attributes on the socioeconomic status assigned to groups by LLMs. Our dataset generation process consists of three phases.

Phase 1. Terms Generation and Selection: The first phase of creating our dataset involved generating pairs of target terms for each demographic domain along with pairs of terms reflecting financial status. In this regard, Belmont University (2024) helped us identify various religious terms. For marital status, we were inspired by the terms defined by Statistics Canada (2023). Terms for the gender and race domains were manually curated, drawing inspiration from other works in bias evaluation. Our research involved evaluating the inherent socioeconomic biases of language models towards names and assessing the embedded information in names, utilizing the list of names proposed by Shen et al. (2022). For intersectionality terms, we combined various demographic attributes (for example, 'Muslim fathers' for gender and religion). We also manually curated a list of neutral terms—terms that do not belong to any social group, such as 'those people.' These terms were used for comparison purposes. Finally, we utilized WordNet (Miller 1995) to address financial status terms. After entering a word and its definition, the system generates an extensive list of related phrases and synonyms. Each term was comprehensively reviewed using the WordNet Online (Princeton University 2010) interface to ensure its relevance and appropriateness. We removed terms that were inaccurate in representing financial status. We provide examples of these terms in Table 6 in Appendix[2].

[2] The appendices for this paper can be found in the complete version available on ArXiv.

Phase 2. Template Sentence Generation: Once the terms have been generated, we need to generate sentence templates with [MASK] and [TARGET] tokens that will be used for our fill-in-the-blank task. We employed OpenAI (2022) to generate these template sentences with the following constraints:

- **Positioning of Socioeconomic Terms:** Our sentence construction focused on the precise placement of two categories of terms: socioeconomic status terms (e.g., 'wealthy'), denoted by the [MASK] token, and domain-related terms (e.g., 'men'), identified by the [TARGET] token. To maintain optimal sentence structure, we avoided placing the [MASK] token at the beginning of the sentence or before the [TARGET] token. This condition allowed us to precisely evaluate the impact of the [TARGET] term on predicting the [MASK] term and assess socioeconomic biases in language models.
- **Financial Context:** Each sentence must be situated within a financial framework to correspond with the study's emphasis on socioeconomic prejudice. For example, having text segments like "in terms of financial stability" in the sentence provides this context.
- **Single Sensitive Context:** Templates were restricted to include only one sensitive feature. More precisely, we had a [TARGET] token in each sentence that was replaced by domain-related terms. In other words, we ensured the generated sentences did not include any demographic information except the [TARGET] token.

Due to the complex nature of the mentioned constraints, there were times when ChatGPT couldn't adhere to all the restrictions. However, the entire process was interactive, and we made iterative modifications to improve the outputs. Ultimately, all the generated templates were checked and refined by the authors, resulting in a list of 50 template sentences.

Phase 3. Template Robustification and Augmentation: After acquiring 50 templates, we performed data augmentation by introducing controlled operations to perturb our templates. Previous research has shown that LLMs are highly influenced by the prompt or template formulation (Kwon and Mihindukulasooriya 2022), especially in tasks that involve predicting masks (Mishra et al. 2023). As the intrinsic evaluations proposed in this article are based on this principle, increasing the robustness of the templates we use is crucial to ensure that the biases created during their generation do not impact our conclusions (Raffel et al. 2020). To this end, we applied four types of perturbation inspired from Raffel et al. (2020). The first perturbations focused on lexical changes; we altered the templates' adverbs from 'often' to five other negative and positive adverbs. Additionally, we added quantifiers to the target token (such as 'all [TARGET]'). The second type was structural perturbation, where we shortened templates and reorganized the phrases within the sentences. The third type was grammatical; we modified the templates to include singular and plural targets and switched from passive to active voice. Additionally, we adjusted the templates to incorporate various verb forms in past and future tenses. Finally, as our fourth type of perturbation, we made heavier changes to the semantics. In this step, we generated two paraphrased templates using a T5 model pre-trained on HuggingFace (Romero 2021).

From our 50 original templates, we ultimately obtained 843 different templates to perform our analyses. In Table 7 of Appendix A, we provide examples of the template sentences in a structured breakdown based on the perturbation strategy. With these controlled increases, we hope to increase our analyses' robustness. Finally, by replacing the [TARGET] token with domain-related terms, we arrived at 956,805 templates with [MASK] token. This [MASK] token was replaced with 18 financial statuses during inference time.

Evaluation

Experimental Setup

Large Language Models (LLMs) We evaluate four LLMs in our study: three autoregressive language models—Falcon, Llama 2, and GPT-2—and one bidirectional transformer model, BERT. Below, we offer a brief overview of each model.

Falcon (Penedo et al. 2023) is an open-source autoregressive large language model introduced by the Technology Innovation Institute (TII). As of 2023, Falcon models are ranked highly on the OpenLLM leaderboard according to Face (2024e). Our study utilized the Falcon-7B model from Face (2024b), which contains 7 billion parameters, in an inference mode. This means we did not train it on new data and only generated predictions based on the pre-trained knowledge of the LLM.

Llama 2 (Touvron et al. 2023), introduced by Meta in 2023, is a new version of the Llama, an optimized autoregressive transformer. Like the other autoregressive models, it predicts the next word in a sentence based on the previous words. In our work, we used the 7B parameter version of Llama 2 from Face (2024d) that, according to Meta, outperforms the Falcon 7B model in several tasks, including commonsense reasoning. The 7B Llama model offers double the context length (4K vs 2K) and has been trained on 2.0 trillion tokens, twice as many as its previous version.

GPT-2 (Radford et al. 2019), developed by OpenAI in 2019, is a Transformer-based autoregressive model trained on the WebText dataset (Gokaslan and Cohen 2019), which contains data extracted from 8 million web pages. Our project uses the small version from Face (2024c), which has 124 million parameters.

BERT (Devlin et al. 2018) developed by Google in 2018. Utilizing WordPiece embeddings (Wu et al. 2016), it is trained on a vocabulary of 30K tokens and focuses on two main tasks: fill-in-the-mask and next-sentence prediction. In our project, we utilized BERT Base, which has 110 million parameters, specifically for the task of fill-in-the-mask from Face (2024a).

Our setup differs for BERT and three other models due to their distinct structures. For autoregressive models, we calculate the probability of entire sentences. To ensure that sentence length does not impact the total probability, we normalize the probability by the sentence length. Conversely, for BERT, we calculate the probability of the [MASK] token, comparing probabilities in fill-in-the-blank scenarios.

Evaluation Metric This section presents the measures used to quantitatively evaluate the LLMs, focusing on their linguistic coherence and impartiality towards different socioeconomic levels. We introduce three main metrics: the Language Model Coherence Score (LMCS), the Poverty Association Ratio (PAR), and the EquiLexi Score. These metrics are inspired by the methodologies of Nadeem, Bethke, and Reddy (2021), but our evaluation differs significantly from theirs. While they compare stereotypical associations over anti-stereotypical ones, thereby considering an LLM biased if it shows a high association with stereotypical pairs, our work considers an LLM biased if it shows a high association with either extreme side (poor or rich groups of terms). In other words, a non-biased LLM assigns the same probability to the [MASK] token being replaced by terms representing either the poor or rich.

Language Model Coherence Score (LMCS): Given a target term, let $P(\text{relevant}|\text{target})$ be the probability that the language model selects a contextually relevant word, and $P(\text{irrelevant}|\text{target})$ be the probability of choosing an irrelevant word. LMCS is defined as:

$$\text{LMCS} = \frac{P(\text{relevant}|\text{target})}{P(\text{relevant}|\text{target}) + P(\text{irrelevant}|\text{target})} \quad (1)$$

For example, in the sentence 'In the property market, Asian people are often seen as [MASK],' a relevant term to fill the mask could be 'wealthy,' while an irrelevant term would be 'apple.' The LMCS value ranges between 0 and 1. A score of 1 indicates that the model always selects contextually relevant words, while 0 indicates a consistent selection of irrelevant words. Values between 0 and 1 represent varying degrees of contextual coherence.

Poverty Association Ratio (PAR): PAR measures the model's tendency to associate a target term with poverty. It is defined as:

$$\text{PAR} = \frac{P(\text{poor}|\text{target})}{P(\text{poor}|\text{target}) + P(\text{rich}|\text{target})} \quad (2)$$

In Equation 2, we use the terms 'poor' and 'rich', but in practice, we sum up the probabilities of 18 terms (9 words in each group). This ratio indicates the likelihood of associating a target term with poverty versus wealth. It ranges from 0 (strong association with wealth) to 1 (strong association with poverty). Although a consistent preference for terms representing 'rich' over 'poor' also indicates bias, we have chosen this aspect of the ratio for comparative purposes. It should be noted that either side of the equation can be used, as both effectively highlight potential discrimination.

EquiLexi Score (ELS): This score combines LMCS and PAR to suggest equity in both language coherence and socioeconomic representation:

$$\text{EquiLexi Score} = \text{LMCS} \times \frac{\min(\text{PAR}, 1 - \text{PAR})}{0.5} \quad (3)$$

It ranges from 0 to 1, with higher values indicating better performance in terms of both linguistic accuracy and socioeconomic fairness.

Baseline Models In our study, we use three theoretical baseline models inspired by Nadeem, Bethke, and Reddy (2021). Moreover, we have established a Neutral Level to benchmark the performance and biases of LLMs.

IdealLM: If we consider an ideal theoretical language model, it would exhibit several key characteristics. Firstly, it would have a Language Model Coherence Score (LMCS) of 1, indicating perfect coherence. Secondly, it would demonstrate a complete absence of socioeconomic bias, with a Poverty Association Ratio (PAR) of 0.5, reflecting a perfect balance between associations with 'poor' and 'rich'.

FullBiasLM: Another theoretical model, FullBiasLM, represents the lower end of fairness. This model clearly prefers one socioeconomic term over another, with a Poverty Association Ratio (PAR) skewed entirely towards either 'poor' or 'rich' (1 or 0).

RandomLM: Serves as another theoretical baseline model in our study, selecting associations purely arbitrarily. This indicates that while the model does not exhibit a strong bias towards any particular socioeconomic term, it also fails to make contextually logical choices consistently.

Neutral Level: In our study, we establish a 'Neutral Level' baseline for each LLM by replacing the [TARGET] token with neutral terms like 'people'. This approach allows us to evaluate whether the model's tendency to associate sentences with 'poor' or 'rich' is due to inherent biases rather than the demographic content of the tokens.

Experimental Results

RQ1. Intrinsic Socioeconomic Bias Evaluation: To what extent do LLMs exhibit intrinsic biases related to socioeconomic status across different sensitive attributes?

To address our first research question, we evaluated the intrinsic socioeconomic biases present in Falcon, Llama 2, GPT-2, and BERT, comparing these LLMs against baseline models, as shown in Table 1. At an aggregated level, the results demonstrate that all LLMs, except BERT, scored high in LMCS, approaching our IdealLM benchmark. However, their ELS are lower, falling below the IdealLM and, in most cases, even below RandomLM. BERT exhibits the highest ELS, while Falcon has the lowest. Despite BERT's relative superiority among LLMs in terms of ELS, its score remains below IdealLM. Autoregressive language models have PARs higher than 0.5, whereas BERT has PARs lower than 0.5. Without comparing them to the Neutral Level, we might interpret that the former LMs tend to associate demographic group terms with poverty, while the latter tends to associate them with wealth. However, when comparing them to the Neutral Level, we see that, in general, all language models have PARs higher than the Neutral Level, indicating a bias towards poverty. However, these aggregate PAR scores may not fully capture all biases, particularly in overlapping demographic groups (e.g. female terms vs male terms in gender). Consequently, we will delve into each domain more closely in the following sections to better understand these biases.

Birth-Assigned Gender Our study indicates that terms associated with the female gender are more frequently

LLM	**Falcon**			**Llama 2**			**GPT-2**			**BERT**		
Demographic Group	*ELS*	*LMCS*	*PAR*	*ELS*	*LMCS*	*PAR*	*ELS*	*LMCS*	*PAR*	*ELS*	*LMCS*	*PAR*
Birth-Assigned Gender	0.400	0.998	0.601	0.413	0.998	0.598	0.520	0.999	0.566	**0.717**	0.816	0.463
Marital Status	0.346	0.999	0.640	0.350	0.999	0.662	0.518	0.999	0.609	**0.781**	0.888	0.463
Race	0.302	0.999	0.617	0.310	0.999	0.654	0.487	0.999	0.647	**0.764**	0.877	0.460
Religion	0.375	0.999	0.599	0.395	0.999	0.603	0.505	0.999	0.606	**0.739**	0.841	0.463
Aggregated	0.389	0.998	0.628	0.398	0.998	0.633	0.508	0.999	0.5625	**0.760**	0.877	0.458
Neutral Level	0.396	0.999	0.596	0.413	0.999	0.602	0.510	0.999	0.598	**0.740**	0.862	0.457
IdealLM	1.000	1.000	0.500	1.000	1.000	0.500	1.000	1.000	0.500	1.000	1.000	0.500
FullBiasLM	0.000	0-1	0 or 1	0.000	0-1	0 or 1	0.000	0-1	0 or 1	0.000	0-1	0 or 1
RandomLM	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500

Table 1: Comparing Demographic Domains for LLMs: In this table, all LLMs are compared with baseline models with respect to the EquiLexi Score (ELS), Language Model Coherence Score (LMCS), and Poverty Association Ratio (PAR) across all four demographic domains. Higher ELS and LMCS indicate better performance and a smaller difference between PAR and the Neutral Level suggests less socioeconomic bias. In each domain, the highest ELS is shown in bold and the lowest is underlined, respectively.

LLM	Neutral Term	Female	Male	Gap F vs M
Falcon	0.596	0.677	0.527	**0.150**
Llama 2	0.602	0.672	0.528	0.144
GPT-2	0.598	0.584	0.550	0.034
BERT	0.457	0.462	0.464	0.002

Table 2: Gender PAR Comparison Between Female and Male Terms across Different LLMs. The highest gap between the two genders is seen in Falcon, and the lowest in BERT, indicating that Falcon is more biased toward gender.

linked to poverty across all autoregressive LLMs compared to their male counterparts, as shown in Table 2. Specifically, Falcon exhibits the largest disparity. Conversely, BERT demonstrates a minimal gap of 0.002, suggesting lower socioeconomic bias related to gender compared to other models. Furthermore, the PAR of neutral terms consistently positions between the extremes of female and male terms. This pattern suggests that sentences containing gender-related terms are more likely to align with terms denoting wealth or poverty. For instance, in neutral contexts such as ”In the property market, these people are often seen as [MASK],” replacing ’these people’ with ’women’ increases the likelihood of [MASK] being filled with poverty-associated words, whereas replacing it with ’men’ tends to decrease this probability.

A pairwise comparison of terms, as shown in Figure 1, offered even more clarification. These bar-plot comparisons show that in Falcon, Llama 2, and GPT-2, terms related to women are more likely to be associated with poverty than terms linked to men. Conversely, BERT has a uniform PAR for both genders and is near the Neutral Level, indicating less bias towards the gender domain.

Marital Status Our analysis of marital status consistently shows that ’Married’ has the lowest PAR, while ’Separated’ and ’Divorced’ rank highest in Falcon, Llama 2, and GPT-2. However, in BERT, all marital statuses have the same PAR, around the Neutral Level. This led us to conclude that there is no evidence of socioeconomic bias in BERT towards marital status. In Falcon, Llama 2, and GPT-2, the Neutral Level consistently highlights the discrimination between ’married’ and other marital statuses by being positioned between them. This ordering across the four language models is depicted in Figure 2, which provides a clear understanding of how various marital statuses are positioned with respect to PAR.

Racial Identities A closer look at race in Figure 3 reveals that BERT treats different races more uniformly around the Neutral Level. In other words, replacing neutral terms like ’these people’ with any terms from the race domain does not impact the PAR. GPT-2 shows minimal variation, with the lowest PAR for ’white’, close to the Neutral Level. In contrast, Llama 2 and Falcon exhibit noticeable biases, disproportionately associating poverty with ’Indigenous’, ’Latino’, and ’Black’ terms while attributing it less to ’White’, which also shows a bias toward another extreme.

Religions Observations from Table 1 indicate a smaller gap between the PAR of each language model and its neutral level, which could be interpreted as a lack of bias in the religion domain. However, a closer examination in Figure 4 reveals discrimination among different religions. Generally, Muslims exhibit a higher PAR and Jews a lower PAR than the neutral level in Falcon and Llama 2, indicating a socioeconomic bias for these religions. The gap of between these two races in Falcon and Llama 2 could explain why the aggregated PAR appears balanced, resulting from the overlapping impacts of these terms. In GPT-2, Hindus have the highest PAR, followed by Muslims, with Christians and Jews showing no clear evidence of socioeconomic bias in our analysis. The behavior of BERT remains consistent with its performance in other domains.

Our analysis of the first research question reveals that advanced language models generally exhibit higher biases toward demographic groups, contrary to expectations given their cleaned datasets. On the other hand, BERT demon-

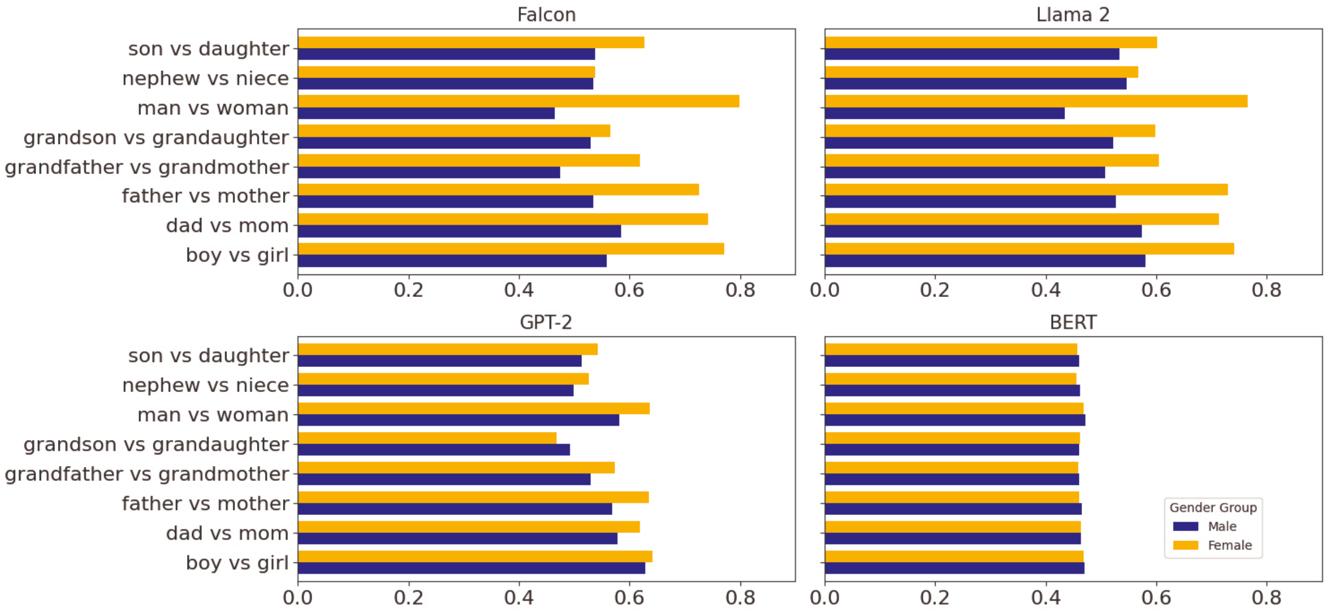

Figure 1: Pairwise PAR Comparison of Gender Terms Across Models. Female terms consistently exhibit higher PAR scores than male terms.

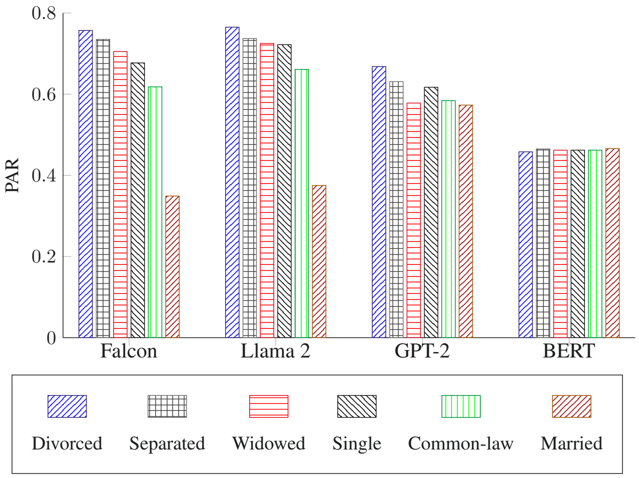

Figure 2: Comparison of PAR for Marital Status across Language Models. For Falcon and Llama 2, a significant gap is observed between Married and other marital statuses, while for GPT-2 and BERT, the levels are comparatively uniform.

strates a smaller gap in PAR across different groups, which may indicate a fairer language model. However, since the LMCS (Language Model Consistency Score) for BERT is significantly lower than autoregressive language models, this could suggest a limitation in its capacity to select relevant words rather than an inherent fairness.

RQ2. Compound Bias Analysis: Does the intersection of multiple sensitive attributes exacerbate the biases present in LLMs?

To answer the second research question, we conducted further research beyond our initial focus on individual sensitive attributes. In real-life situations, people often have multiple sensitive attributes, a condition called intersectionality. This phenomenon has the potential to magnify biases in LLMs. We investigated the intersectionality between race, gender, and marital status. In all three domains, in general, intersectionality amplified socioeconomic bias more than individual demographic terms for Falcon and Llama 2 to some extent in GPT-2. However, this was not the case with BERT. Comparing the PAR of the combination of terms to the one of each term alone provided further insight. According to Figure 5, in Llama 2, for instance, we constant the highest PAR belongs to 'Indigenous mothers', which has a deviation of 0.09 from the term 'mother' individually. This means the 'Indigenous' terms shift the PAR of combination compared to the PAR of 'mother'. On the other hand, the disparities become worse when we study the lower extremes of a sensitive attribute in combination with a higher extreme. For example, when adding race terms like 'Indigenous' to the term 'man', PAR increases by 0.26 compared to PAR of 'man'. Similarly, combining 'white' with 'granddaughter' increases PAR with respect to 'white' and reduces it with respect to 'granddaughter'. The heat-maps for other LLMs are provided in Appendix. In general, similar trends have been observed in autoregressive LLMs. In BERT, the variation is not significantly pronounced since the variation in the isolated studies between terms in race and gender was not too significant. Regarding marital status and gender, we have observed similar trends as those seen with race and gender in autoregressive LLMs. For example, in Falcon, combining the 'Widowed' status with 'mother' term results in an increased PAR compared to 'mother' term in-

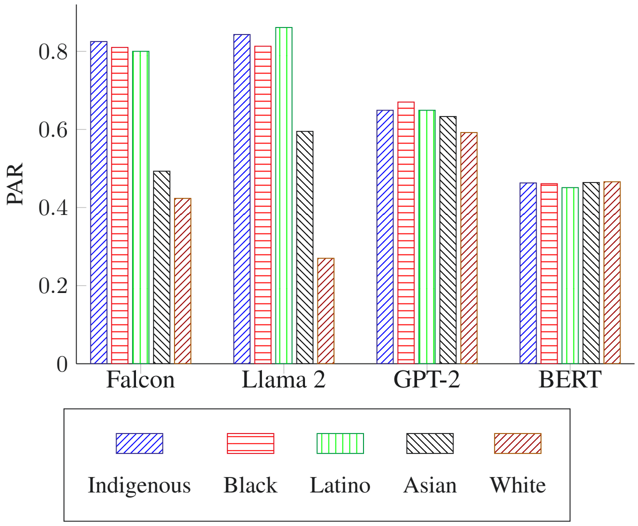

Figure 3: Comparison of Race PAR across LLMs: The differences in PAR among different races are extremely pronounced in Falcon and Llama 2. In contrast, BERT shows PARs around the neutral level for all races, indicating no evidence of socioeconomic bias in this model.

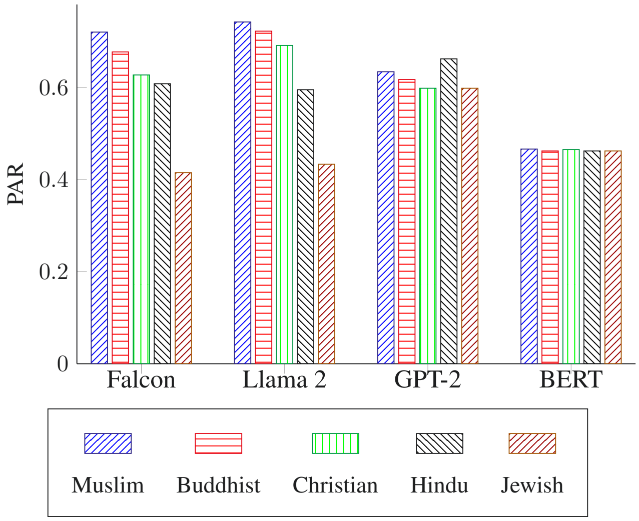

Figure 4: Comparison of PAR for Religion across LLMs: ’Muslim’ is highly associated with poverty, while ’Jewish’ shows the lowest PAR, indicating wealth across all autoregressive language models.

dividually, as well as to ’indigenous’ terms (see Figure 6). These trends remain consistent for Llama 2, and GPT-2 but not for BERT. We have taken another step towards intersectional analysis by studying triple-level intersectionality, which combines marital status, race, and gender. For this phase of analysis, we adopted a method where we isolated one domain and compared the impact of its combination with compositions from the other two domains. We examined all 768 combined terms, but for simplicity, we present some of the extreme cases in Tables 3. In Table 3, we compare the PAR of combinations like ’man’ and ’woman’ with terms having the highest and lowest PAR from marital status and race. For instance, in Falcon, alternating between ’common-law white’ and ’widowed indigenous’ in combination with ’man’ increases the PAR by 0.502. For instance for Falcon, in Figures 6 we observed that combining ’man’ with ’widowed’ increased the PAR by 0.12. More over and combining ’Indigenous’ with ’man’ increased the PAR by 0.21 with respect to the term ’man’ (Figure 9 of Appendix). Now, combining these three terms increased it by 0.27. This is one of several examples of bias amplification by intersectionality. This impact remains consistent with other terms. As another example, replacing ’married boy’ with ’divorced girl’ in combination with ’white’ showed an increase of 0.293. In Falcon, the term ’widowed Indigenous woman’ has a PAR of 0.876, showing increases of 0.16 compared to ’widowed,’ 0.04 compared to ’Indigenous,’ and 0.07 compared to ’woman’. While this analysis is more complex than the previous ones due to the interaction of the three components, which can shift the dynamics towards lower or higher PAR, we observed consistent trends. For instance, combining terms with high PAR typically increases PAR, at least compared with one of the included terms. Conversely, combinations of terms with low PAR decrease the composite PAR. Furthermore, our analysis shows that combinations of two or three terms from different domains, each having different levels of PAR, can bring the PAR close to the Neutral Level, which is the lowest socioeconomic bias level, such as ’separated Asian man’ or ’married Arab girl’ (see Table 8 and 9 of Appendix for additional details).

Overall, our findings indicate that intersectionality amplifies biases in LLMs. The highest PAR of 0.885 in our entire study is for ”Divorced Indigenous Mother” in Llama 2, while the lowest PAR of 0.208 is related to ”White Man” in Falcon. This means that in our study, a ”Divorced Indigenous Mother” is tagged as poor 88.5% of the time by Llama 2, whereas a ”White Man” is tagged as rich 79.2% of the time by Falcon. These results underscore how intersectionality can worsen biases, highlighting the need for more comprehensive bias mitigation strategies in LLMs.(For more details, refer to Table 10 in the Appendix.)

RQ3. How do LLMs discern sensitive attributes from names, and what is the impact of compound names embodying multiple attributes on socioeconomic biases in these models?

Names are fundamental identifiers that carry embedded information about attributes like gender and race shown by Haim, Salinas, and Nyarko (2024) and Meltzer, Lambourne, and Grandi (2024), which may influence the operational biases of LLMs in contexts such as recruitment where such attributes should not affect decisions. Our research involved examining the ability of these four LLMs to predict gender and race from a list of names. For this purpose, we employed a zero-shot learning approach to predict gender and race based on the list of names we got from Shen et al. (2022). The results, as detailed in Table 4, show a clear proficiency in extracting such sensitive information with preci-

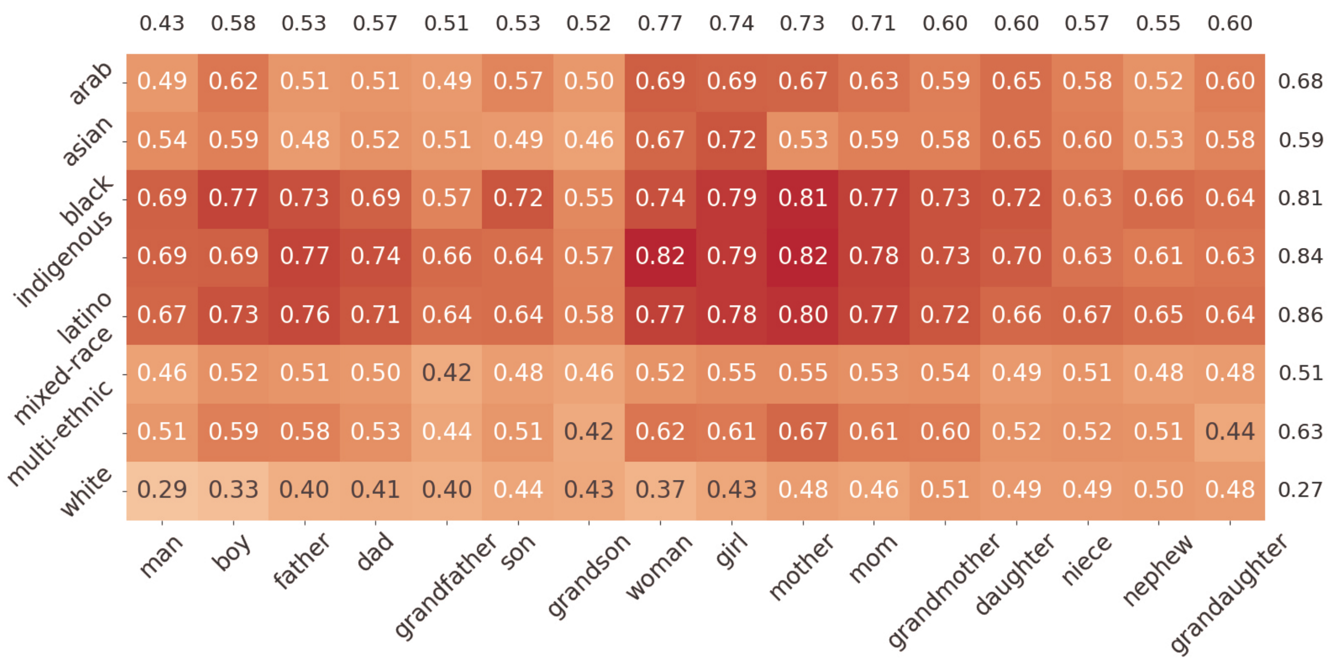

Figure 5: This heat-map shows the intersectionality impact of race and gender on PAR in Llama 2. It compares composite PAR values with individual PAR for each domain. The values inside the heat-map display the PAR of intersectionality, with values at the top and right side showing individual PAR of gender and race.

sion. For instance, Llama 2 accurately predicted gender in all instances, and Falcon showed similarly high accuracy. The challenge was greater with race prediction due to overlaps across races, where GPT-2 and BERT showed an accuracy of 50%, which could be interpreted as them choosing races at random.

Furthermore, we explored how names influence the socioeconomic biases inherent in LLMs. Using the same list of names, we assessed the models' responses when these names were used as a target domain. In Table 5, these results are compared to the intersectionality between race and gender, where the results are aggregated to the same groups as names (for races other than 'White', they are summed up in 'non-white', and gender terms are grouped into 'female' and 'male'). Although the PAR values for names differ from intersectionality, the general trends remain consistent. This could be explained by the fact that names may include other demographic information such as religion, nationality, and even age, while our intersectionality contains terms from race and gender. As we can observe, the general trends are the same in Falcon and GPT-2 when comparing names to race and gender intersectionality. For Llama 2, the general trend also stays the same, except that white female names have a lower PAR than the male group. In BERT, however, the trend is not the same at all.

In this study, we assessed the socioeconomic biases embedded in LLMs towards demographic groups such as gender, marital status, race, religion, and combinations thereof, including race, gender and marital status, and biases associated with names. Our results indicate that autoregressive models not only exhibit biases across these domains but also that such biases are intensified by intersectionality. Specifically, models like Falcon and Llama 2 display pronounced socioeconomic biases compared to GPT-2. Although BERT shows negligible bias in our tests, its LMCS is lower than other LLMs. This means BERT is less effective than other models in picking relevant words, so the low bias of BERT may result from its lower capacity. Falcon and Llama 2, designed with instruction-based architectures, appear to reflect training data more accurately and demonstrate enhanced reasoning capabilities. We hypothesize that this design aspect is a key factor in why Falcon and Llama 2 exhibit more pronounced socioeconomic biases than GPT-2 and BERT. To test this hypothesis, we conducted a limited experiment in which we input extreme terms from each demographic domain into simple sentences, allowing the LLMs to complete them. Specifically, we used prompts structured around extreme cases with the highest and lowest PAR identified within each domain. For example, we provided the prompt '[TARGET1] are often rich and [TARGET2] are poor because' and repeated this exercise with five different seeds. One of the generated outputs from this experiment is displayed in Appendix. In this experiment, we observed that Llama 2 and Falcon generated text with the reasoning behind the phenomena, whereas GPT-2 produced sentences largely unrelated to the context. On the other hand, BERT was unable to generate coherent sentences, which is not surprising given that text generation is not its primary function. While these preliminary results validate our hypothesis, a more thorough future investigation is needed.

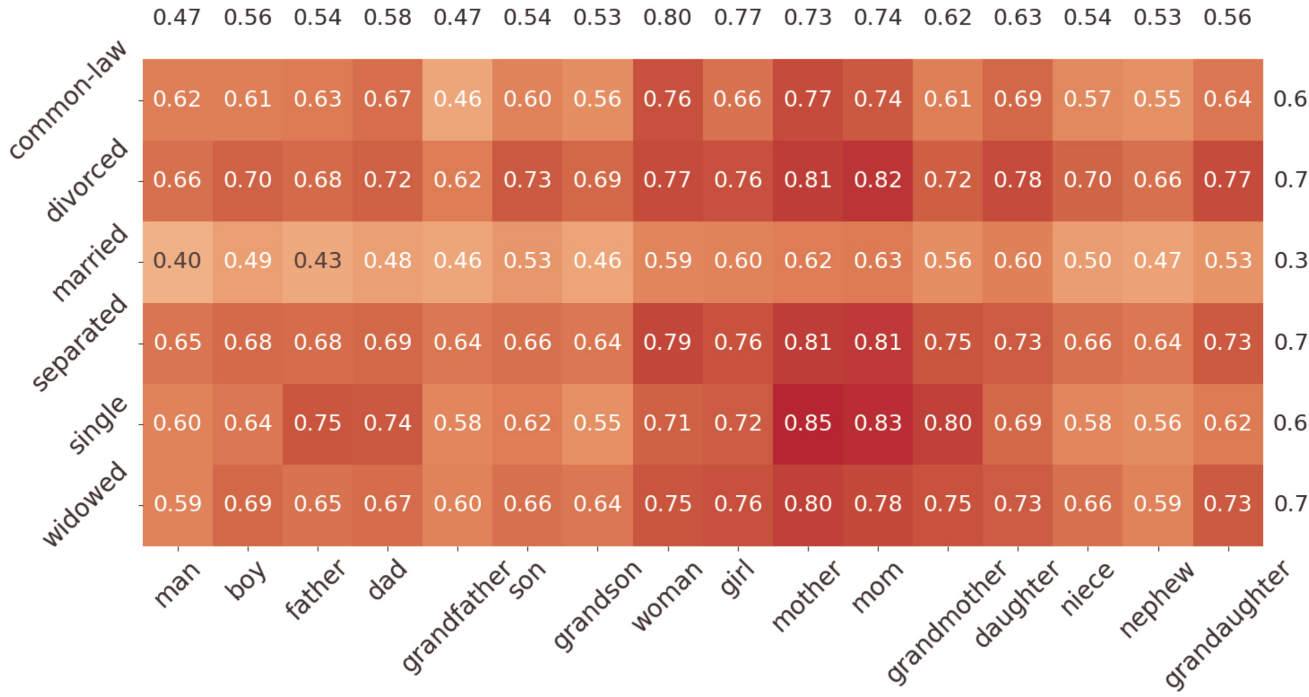

Figure 6: This heat-map shows how combination of marital status and gender terms affect PAR in Falcon.

Discussion

In our study, we compared two types of Large Language Models, BERT and autoregressive models, such as Falcon, Llama 2, and GPT-2, which differ in structure and the methods used to achieve results. The development of these models has increasingly focused on optimizing their capabilities to understand and generate human-like text. However, our comparative analysis leads us to hypothesize that the efficacy of these models in bias mitigation and reasoning tasks is heavily dependent on the nature and quality of their training data. This highlights the potential impact of data diversity on model behavior and performance, although further evidence is needed to confirm these observations.

Impact of Data Quality and Composition: BERT's lower intrinsic socioeconomic biases could be primarily due to its reliance on cleaner, more structured data sources, namely BookCorpus(Gokaslan and Cohen 2019) and English Wikipedia. These sources provide a well-curated foundation that reduces the likelihood of learning and propagating biases found in less controlled environments. In contrast, GPT-2 are trained on extensive but unfiltered datasets like WebText(Gokaslan and Cohen 2019), which capture a more comprehensive array of internet discourse and its inherent biases. This is evident from the model documentation provided by OpenAI, highlighting concerns regarding the propagation of stereotypes and the need for careful deployment in bias-sensitive applications. While data cleaning strategies such as RefinedWeb(Penedo et al. 2023) for Falcon might help obtain better results for reducing direct biases (He et al. 2023), they are not enough to compensate the capability of new models to reflect these biases in other aspects where complex link exist. This could explain why BERT shows less socioeconomic bias towards different social groups, which needs further investigation in future research.

Training Data Diversity and Model Generalization: The diversity in training datasets, as seen with Falcon and Llama 2, prepares models to handle a broader range of linguistic styles and contexts. This is beneficial for generalization across different tasks. However, the challenge remains in balancing this diversity with the need to control for quality and bias, which is critical for ensuring that models do not amplify harmful stereotypes or misinformation, as shown in this work.

Instruction-Based Design and Enhanced Reasoning: Llama 2 and Falcon, designed to be more instruction-sensitive, demonstrate an advanced capacity for detailed reasoning. Through their diversified training that likely includes conversational and contextual data, these models excel in generating nuanced responses. This capability is crucial for applications requiring detailed explanatory outputs and where understanding the context of queries is essential. However, the detailed reasoning ability does not necessarily result in fairness or lesser bias. Their higher socioeconomic bias highlighted this compared to GPT-2 and BERT.

Conclusion

Our analysis of well-known Large Language Models (LLMs) such as GPT-2, BERT, Llama 2, and FALCON has revealed socioeconomic biases affecting various de-

LLMs / Target Term	*diff*	*PAR*	*PAR_Gender*
Falcon			
Neutral level		0.596	
widowed indigenous **woman**		0.867	0.799
married white **woman**		0.311	0.799
diff	**0.556**		
widowed indigenous **man**		0.740	0.465
common-law white **man**		0.238	0.465
diff	0.502		
Llama 2			
Neutral level		0.602	
divorced indigenous **woman**		0.863	0.765
single white **woman**		0.468	0.765
diff	0.395		
divorced indigenous **man**		0.829	0.434
married white **man**		0.384	0.434
diff	**0.445**		

Table 3: Examples of Triple Intersectionality PAR Variation for Gender. We compare the terms 'woman' and 'man', which represent the highest and lowest PAR values in terms of gender, by combining them with composite terms of marital status and race. Here we can observe in Falcon and Llama 2 how other race and marital status combinations could increase or decrease PAR for the terms 'man' and 'woman'.

Domain	Falcon	Llam 2	GPT-2	BERT
Gender	0.989	**1.000**	0.705	0.841
Race	0.737	**0.830**	0.500	0.500

Table 4: Comparison of LLMs' Accuracy in Identifying Gender and Race from Name

mographic groups. Specifically, we observed how demographic information like gender, marital status, race and religion could impact the perceived financial status of individuals from different groups. Our detailed analysis demonstrated that biases, while evident in high-level categories like male and female, become more pronounced when examining details such as 'daughter' and 'father.' Furthermore, we showed how intersectionality can alter the dynamics of socioeconomic bias. Moreover, our assessments revealed not only that state-of-the-art LLMs are proficient in accurately extracting gender and race from names but also how they discriminate based on names alone. These findings highlight an urgent need for bias mitigation in LLMs before their deployment in sensitive domains.

Limitations & Future Directions In this work, we used English due to its simplicity and widespread use. In the future, we plan to extend our research to include other languages. Moreover, we included ChatGPT, which is based on a language model, for the dataset generation pipeline. Despite our alignment and refinement, there is still a possibility that some level of bias from the underlying language model in ChatGPT could have influenced our dataset. Furthermore, considering the high sensitivity of LLMs to input and their text conditioning behavior, we recognize that even a slight change in the prompt may lead to different outcomes. Although our thorough analysis and presentation of aggregated results can minimize such errors, a lack of prompt robustness may still affect our findings.

LLM	Domain	WM	WF	NWM	NWF
Falcon	Name	0.436	0.438	0.478	**0.524**
	Intersect.	0.341	0.404	0.603	**0.677**
Llama 2	Name	0.429	0.413	0.463	**0.510**
	Intersect.	0.394	0.462	0.589	**0.665**
GPT-2	Name	0.452	0.458	0.462	**0.508**
	Intersect.	0.553	0.586	0.602	**0.630**
BERT	Name	0.441	0.441	**0.438**	0.433
	Intersect.	0.456	0.459	0.457	**0.460**

Table 5: Comparing Poverty Association Ratio (PAR) for Names and Intersectionality (Intersect.) Between Race and Gender Across LLMs.

In our study, we used the same name list proposed by Shen et al. (2022). However, this list lacks a comprehensive range of names from different races, which could have provided a more extensive comparison. A more diverse name list would allow future research to assess the impact of factors like nationality, religion, and age on socioeconomic bias. In addition, our findings indicate that intersectionality amplifies socioeconomic biases. In future research, one could incorporate more dimensions and assess how the dynamics of bias change by adding multiple intersecting demographic attributes. This direction could yield deeper insights into the complex nature of bias in LLMs and how different identity aspects interact to compound biases.

Finally, the implications of the socioeconomic biases examined in our paper are profound, and they can reach beyond LLMs themselves. Future research is needed to show how these biases, embedded in LLMs, could impact critical aspects of individuals' lives when applied to downstream tasks. This influence may span from suggesting lower-paying job opportunities to affecting insurance rates, influencing bank loan approvals, and even contributing to visa rejections based on assumptions about financial capabilities. Future studies are needed to understand how inherent bias affects various tasks and find mitigation strategies to reduce these biases in different LLM models and minimize potential harms.

Acknowledgements

Funding support for project activities has been partially provided by Canada CIFAR AI Chair, Google award, and NSERC Discovery Grants program. We also express our gratitude to Compute Canada for their support in providing facilities for our evaluations.

References

Belmont University. 2024. World Religions. https://belmont.libguides.com/worldreligions. Accessed: 2024-01-21.

Beukeboom, C. J.; and Burgers, C. 2019. How stereotypes are shared through language: a review and introduction of the aocial categories and stereotypes communication (SCSC) framework. *Review of Communication Research*, 7: 1–37.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Delobelle, P.; Tokpo, E. K.; Calders, T.; and Berendt, B. 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*, 1693–1706.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dimri, A.; Yerramilli, S.; Lee, P.; Afra, S.; and Jakubowski, A. 2019. Enhancing Claims Handling Processes with Insurance Based Language Models. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 1750–1755.

Face, H. 2024a. BERT. https://huggingface.co/docs/transformers/main/model_doc/bert. Accessed: 2024.

Face, H. 2024b. falcon-7b. https://huggingface.co/tiiuae/falcon-7b. Accessed: 2024.

Face, H. 2024c. GPT-2. https://huggingface.co/openai-community/gpt2. Accessed: 2024.

Face, H. 2024d. Lama-2-7B. https://huggingface.co/docs/transformers/main/model_doc/llama2#llama2. Accessed: 2024.

Face, H. 2024e. OpenLLM Leaderboard. https://huggingface.co/open-llm-leaderboard. Accessed: 2024.

Gokaslan, A.; and Cohen, V. 2019. OpenWebText Corpus. http://Skylion007.github.io/OpenWebTextCorpus. Accessed: 2024.

Goldfarb-Tarrant, S.; Marchant, R.; Sánchez, R. M.; Pandya, M.; and Lopez, A. 2020. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*.

Hadi, M. U.; Qureshi, R.; Shah, A.; Irfan, M.; Zafar, A.; Shaikh, M. B.; Akhtar, N.; Wu, J.; Mirjalili, S.; et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Haim, A.; Salinas, A.; and Nyarko, J. 2024. What's in a Name? Auditing Large Language Models for Race and Gender Bias. *arXiv preprint arXiv:2402.14875*.

He, J.; Lin, N.; Shen, M.; Zhou, D.; and Yang, A. 2023. Exploring Bias Evaluation Techniques for Quantifying Large Language Model Biases. In *2023 International Conference on Asian Language Processing (IALP)*, 265–270. IEEE.

Kaneko, M.; and Bollegala, D. 2022. Unmasking the mask–evaluating social biases in masked language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11954–11962.

Kiritchenko, S.; and Mohammad, S. M. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Kwon, B. C.; and Mihindukulasooriya, N. 2022. An empirical study on pseudo-log-likelihood bias measures for masked language models using paraphrased sentences. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, 74–79.

Meltzer, P.; Lambourne, J. G.; and Grandi, D. 2024. What's in a Name? Evaluating Assembly-Part Semantic Knowledge in Language Models Through User-Provided Names in Computer Aided Design Files. *Journal of Computing and Information Science in Engineering*, 24(1): 011002.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

Mishra, A.; Soni, U.; Arunkumar, A.; Huang, J.; Kwon, B. C.; and Bryan, C. 2023. Promptaid: Prompt exploration, perturbation, testing and iteration using visual analytics for large language models. *arXiv preprint arXiv:2304.01964*.

Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

OpenAI. 2022. Title. ChatGPT,OpenAI. Accessed: 2023-11-01.

Penedo, G.; Malartic, Q.; Hesslow, D.; Cojocaru, R.; Cappelli, A.; Alobeidli, H.; Pannier, B.; Almazrouei, E.; and Launay, J. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Princeton University. 2010. WordNet Online lexical database. https://wordnet.princeton.edu/. Accessed: 2024-01-21.

Qin, C.; Zhu, H.; Xu, T.; Zhu, C.; Jiang, L.; Chen, E.; and Xiong, H. 2018. Enhancing person-job fit for talent recruitment: An ability-aware neural network approach. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 25–34.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Romero, M. 2021. T5 Small Fine-tuned for Paraphrasing on Quora Dataset. https://huggingface.co/mrm8488/t5-small-finetuned-quora-for-paraphrasing. Accessed: 2023-05-06.

Shen, T.; Li, J.; Bouadjenek, M. R.; Mai, Z.; and Sanner, S. 2022. Unintended bias in language model-driven conversational recommendation. *arXiv preprint arXiv:2201.06224*.

Singh, S.; Keshari, S.; Jain, V.; and Chadha, A. 2024. Born With a Silver Spoon? Investigating Socioeconomic Bias in Large Language Models. *arXiv preprint arXiv:2403.14633*.

Statistics Canada. 2023. Classification of legal marital status. https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=1314707&CVD=1314707&CLV=0&MLV=1&D=1. Accessed: 2024-01-21.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Webster, K.; Wang, X.; Tenney, I.; Beutel, A.; Pitler, E.; Pavlick, E.; Chen, J.; Chi, E.; and Petrov, S. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wu, Z.; Dong, Y.; Li, Y.; and Shi, B. 2023. Unleashing the power of text for credit default prediction: Comparing human-generated and AI-generated texts. *Available at SSRN 4601317*.

Zhao, J.; Mukherjee, S.; Hosseini, S.; Chang, K.-W.; and Awadallah, A. H. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. *arXiv preprint arXiv:2005.00699*.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.