More of the Same: Persistent Representational Harms Under Increased Representation

JENNIFER MICKEL, Independent Researcher, USA
MARIA DE-ARTEAGA, The University of Texas at Austin, USA
LEQI LIU, The University of Texas at Austin, USA
KEVIN TIAN, The University of Texas at Austin, USA

To recognize and mitigate the harms of generative AI systems, it is crucial to consider who is represented in the outputs of generative AI systems and how people are represented. A critical gap emerges when naively improving who is represented, as this does not imply bias mitigation efforts have been applied to address how people are represented. We critically examined this by investigating gender representation in occupation across state-of-the-art large language models. We first show evidence suggesting that over time there have been interventions to models altering the resulting gender distribution, and we find that women are more represented than men when models are prompted to generate biographies or personas. We then demonstrate that representational biases persist in how different genders are represented by examining statistically significant word differences across genders. This results in a proliferation of representational harms, stereotypes, and neoliberalism ideals that, despite existing interventions to increase female representation, reinforce existing systems of oppression.

1 INTRODUCTION

The existence of social biases and representational harms in the outputs of language models and generative AI systems is well documented [23, 88, 110]. As a result, a number of benchmarks and evaluations have been created to measure social biases. These methods rely on templates specifying social groups [16], sentences containing specific stereotypes [71, 72], or observing the marked words that result from prompting specific social groups [23]. Although these methods give insights into the nature of existing social biases and representational harms, they do not provide insight into whether these social biases proliferate when social groups are not prompted, which is troubling as most usage of generative AI does not specify a social group as such representational harms may proliferate and current benchmarks and evaluations would not capture this. In this work, we study representational harms in the output of language models when demographics are not explicitly prompted.

Representational harms are multi-dimensional [24]. In particular, both *who* is represented and *how* they are represented matters. This can pose challenges for bias measurement and mitigation efforts. While a number of bias mitigation methods have been developed to address harmful social biases, spanning pre-processing approaches [31, 95, 108], prompting techniques [1, 34, 38, 64, 100, 107], in-training approaches [45, 111], intra-processing approaches [48], and post-training approaches [30], usage of commercial systems regularly reveals failure modes. For example, a version of Gemini prompted Google to apologize after the commercial model generated historically inaccurate images, such as images perceived to be Black, Asian, or Indigenous to the prompt ‘a portrait of a Founding Father of America’ [54, 81]. This may have resulted from bias mitigation interventions aimed at addressing *who* is represented but not accounting for *how* they are represented and the context, illustrating that mitigating social biases is complex and requires a nuanced approach. A gap emerges in that addressing *who* is represented does not imply that harms in *how* people are represented have been addressed.

Authors’ addresses: Jennifer Mickel, jamickel@utexas.edu, Independent Researcher, USA; Maria De-Arteaga, dearteaga@mccombs.utexas.edu, The University of Texas at Austin, Austin, Texas, USA; Leqi Liu, leqiliu@utexas.edu, The University of Texas at Austin, Austin, Texas, USA; Kevin Tian, kjtian@cs.utexas.edu, The University of Texas at Austin, Austin, Texas, USA.

We critically examine this gap by investigating gender bias in state-of-the-art models. We accomplish this by generating personas and biographies of various occupations without specifying gender. This allows us to investigate who is represented within these generations and analyze how people are represented, by examining the statistically significant differences in how men and women are described within these generations. In Section 5.1, we examine this and find that, on average, across all models and occupations, the representation of women is greater than men which to our knowledge has not been explicitly observed in prior work.[1] This indicates that companies may have utilized bias interventions to address who is represented within occupation as previous work has demonstrated gender bias in occupation representation, where male-dominated occupations are more likely to be associated with men and female-dominated occupations are more associated with women, by observing word embeddings [88, 110] and utilizing or building upon existing bias benchmarks to analyze gender biases in occupation [12, 56, 64]. This indicates that model developers may have used bias mitigation methods to address these past gender biases as our findings indicate that the likelihood that a doctor is described with female pronouns has changed over time with recent models more likely to describe a doctor with female pronouns.

After establishing that there has been a change in *who* is represented, we investigate how men and women are represented differently in Section 5.2 by investigating the difference in how women and men are described within these personas and biographies generated without specifying gender. By not specifying gender, we can examine how differences in gender persist in contexts where gender is not specified. To do this analysis, we develop the Subset Similarity Metric and the Subset Representational Bias Score. As we do not specify gender in the prompt, we utilize the Gender Association algorithm we develop to associate gender with each generation. We compare these generations associated with gender to generations specifying gender using the Subset Similarity Metric to calibrate the comparison utilizing the statistically significant words for each occupation, gender, model triple using a calibrated version we develop of the Marked Personas method developed by Cheng et al. [23]. We then utilize the Subset Representational Bias Score to directly observe the difference in how women and men are represented. We find glaring statistically significant differences between women and men, indicating that stereotypes and representational harms associated with gender persist when gender is not specified in the prompt. We also observe the percent change in Subset Representational Bias Scores from GPT-3.5 to GPT-4o-mini, and find on average that the biases between associated gender and specified gender have strengthened (i.e. generations associated with women are more similar to generations of specified women in GPT-4o-mini than GPT-3.5), indicating that on average gender biases have amplified from GPT-3.5 to GPT-4o-mini.

We analyze the statistically significant words in Section 5.3 by identifying trends and examining the context in which these words are used. This analysis reveals the perpetuation of stereotypes and the reinforcement of systems of oppression through neoliberal narratives and the meritocracy myth within the generated outputs. This shifts the responsibility for addressing systemic oppression onto individuals rather than challenging and dismantling oppressive systems. Additionally, we observe that the representational harms identified by Cheng et al. [23] persist in generations associated with gender, even when gender was not explicitly prompted. With the increased representation of women, these representational harms proliferate as the representational harms are primarily associated with women and the representation of women has increased. We discuss the implications of these harms and provide recommendations to model developers, researchers, and practitioners in Section 6. The dataset of generated personas and biographies as well as the code to reproduce our results and use the methods and metrics we propose is located at https://github.com/jennm/more-of-the-same.

[1] Based on personal communication, we believe this was known by the developers of commercial large language models but has not been formally published.

2 BACKGROUND

Our work on understanding who is represented and how people are represented is grounded in the representational harm literature. A number of representational harm taxonomies have been developed [24, 52, 92]. Representational harms refer to how system outputs shape people's understandings, beliefs, and attitudes toward specific social groups, thereby influencing those groups' societal standing [52]. These harms have been categorized into various types, including social stereotyping, reification of social groups or essentialist categories, inaccurate or skewed representations, demeaning or derogatory language, denial of self-identification, (hyper)attention, exposure or erasure, discrimination, hate or violence, outsider gaze, and hierarchies or marginalization [24, 52, 92]. While these categories provide valuable insights, they are not exhaustive [24]. Notably, most categories focus on how people are represented, with the exception of the (hyper)attention, exposure, or erasure category, which addresses who is represented. Erasure and lack of representation negatively affect communities and individuals who are not represented [36, 43] as does harmful depictions of individuals and communities [36, 99].

Previous research has extensively investigated gender bias in occupations within word embeddings and language models such as GPT-2 and GPT-3. Rudinger et al. [88] and Zhao et al. [110] identified occupational gender biases in word embeddings. Kirk et al. [55] demonstrated that GPT-2 associates more occupations with male pronouns than female pronouns. Similarly, Brown et al. [12] found that GPT-3 more frequently associates women with participant roles compared to men. Mattern et al. [64] showed that GPT-3 is more likely to associate men with male-dominated occupations and women with female-dominated ones. Kotek et al. [56] further revealed that occupational gender biases—where occupations associated with men are more strongly linked to men and those associated with women are more strongly linked to women—persist across four publicly available large language models as of 2023. Importantly, these analyses were conducted by explicitly providing gender or pronoun options, specifying gender, pronouns, or names (which carry gender associations) in the prompt [55, 56, 64] or by utilizing existing bias benchmark datasets [12]. This reliance on specifying gender or gender options in prompts or templates highlights a critical gap in understanding: the presence of gender bias in occupational associations when gender is not explicitly mentioned remains largely unexplored. We seek to address this gap as gender biases can emerge in generations from prompts not specifying gender, yet our understanding of gender biases in these contexts is limited. This is crucial to understand as we think this is a more realistic depiction of how gender biases proliferate in natural settings as the majority of users do not specify gender in the prompt.

To better understand representation in AI systems, various bias benchmarks and evaluations have been developed to measure social biases that contribute to representational harms. These evaluations typically rely on templates specifying gender and occupation [88, 110], sentences containing specific stereotypes [71, 72], or the analysis of marked words generated when prompting specific social groups [23]. However, all of these evaluations require the explicit specification of gender, despite the fact that gender biases can also emerge in outputs where gender is not specified in the prompt. Some evaluations focus on who is represented, such as gender and occupation benchmarks [88, 110], while others measure the presence of stereotypes using crowdsourced templates [71, 72]. Marked Personas [23] provides insights into how people are represented by identifying statistically significant words that differentiate social groups, but it requires analysis of these words and does not enable aggregate analysis across context. To our knowledge, no existing evaluation framework allows for the analysis of how groups are represented without explicitly specifying the group in the prompt. This gap is critical, as generative AI is frequently used in scenarios where users do not explicitly mention gender or other demographic groups, making it essential to analyze implicit biases in such contexts.

A range of bias mitigation methods have been developed to address representational harms in AI systems. These methods can be applied at various stages, including during training (in-training), modifying inference behavior (intra-processing), or through pre-processing inputs and post-processing outputs [37]. Key approaches include pre-processing techniques [31, 95, 108], prompting strategies (a specialized form of pre-processing) [1, 34, 38, 64, 100, 107], in-training methods [45, 111], such as reinforcement learning with human feedback (RLHF) [77], intra-processing techniques [48], and post-processing approaches [30]. While many of these methods focus on addressing *who* is represented in AI outputs [31, 108], fewer address *how* individuals and groups are portrayed in ways that preserve experiences tied to identity [30]. Several bias mitigation strategies aim to reduce differences between groups [45, 48, 111], but this approach can unintentionally erase critical historical context about systemic inequality without effectively addressing it. This erasure risks perpetuating inequality by ignoring its structural foundations [28]. RLHF [77] offers potential for improving how people are represented; however, the perspectives of the crowd workers providing feedback heavily influence the process. Their viewpoints, biases, and beliefs are embedded within the model, potentially limiting the effectiveness of this approach [20]. Prompting techniques may adjust how individuals are represented in a single generation [1, 34, 38, 64, 100, 107], but they fail to fundamentally alter the underlying biases encoded in the model itself. Given these limitations, it is essential to carefully design and implement bias mitigation interventions to ensure they effectively address representational harms without inadvertently exacerbating them.

3 METHODOLOGY

The pipeline of our methodology is as follows. First, we generate personas and biographies. We then utilize the Gender Association Method (described in Section 3.1) to associate gender with each generation as we do not specify gender in the prompt. To understand how men and women are represented in generations, we utilize the Calibrated Marked Words method (described in Section 3.2) to identify the statistically significant words that differentiate associated female generations from associated male generations. The Subset Similarity Metric and Subset Representational Bias Score (described in Section 3.3) utilizes these statistically significant words to analyze how similar men and women are described.

3.1 Gender Association Method

In order to analyze how generative AI depicts people of different genders without explicitly prompting the gender that should be depicted in the output, one must have means to associate an output to a gender. To do this, we propose a Gender Association Method, which associates generations with male, female, and non-binary gender identities. To determine gender associations for each generation, we analyze the frequency of female, male, and neutral pronouns, as well as gendered honorifics like “Ms.,” “Mrs.,” and “Mr.” in a given output. We also account for terms related to non-binary identities, such as “non-binary”, “nonbinary”, or “they/them.” Generations are associated with a non-binary identity if non-binary related terms are present and neutral pronouns outnumber both male and female pronouns. Generations are associated with a female identity if they have more female pronouns than both male and neutral pronouns, or if non-binary related terms are absent and female pronouns outnumber male pronouns. Similarly, generations are associated with a male identity if they have more male pronouns than both female and neutral pronouns, or if non-binary related terms are absent and male pronouns outnumber female pronouns. Generations that do not meet any of these criteria are excluded from gender association and analysis. The pseudocode for the Gender Association Method is provided in Algorithm 5, and the method’s accuracy can be found in Appendix B.1. We opt to utilize this algorithm as

Algorithm 1 Gender Association Method.

```
Input: text (generation text lowercase); counts (word counts generated content from generative AI system)
Output: Associated gender with generation
$b_{\text{non-binary presence}} \leftarrow$ “nonbinary" is in text or “non-binary" is in text or “they/them" is in text
$b_{\text{ms presence}} \leftarrow$ counts[“ms"] and “ms." is in text
$c_{\text{female}} \leftarrow$ counts[“she"] + counts[“her"] + counts[“hers"] + counts[“herself"] + counts[“female"] + $b_{\text{ms presence}}$ + counts[“mrs"]
$c_{\text{male}} \leftarrow$ counts[“he"] + counts[“his"] + counts[“male"] + counts[“him"] + counts[“himself"] + counts[“mr"]
$c_{\text{neutral}} \leftarrow$ counts[“they"] + counts[“their"]
$g \leftarrow$ None
if $b_{\text{non-binary presence}}$ and $(c_{\text{neutral}} > c_{\text{male}} + c_{\text{female}})$ then
    $g \leftarrow$N
else if not $b_{\text{non-binary presence}}$ and $c_{\text{male}} > c_{\text{female}}$ or $c_{\text{male}} > c_{\text{female}} + c_{\text{neutral}}$ then
    $g \leftarrow$M
else if not $b_{\text{non-binary presence}}$ and $c_{\text{female}} > c_{\text{male}}$ or $c_{\text{female}} > c_{\text{male}} + c_{\text{neutral}}$ then
    $g \leftarrow$F
end if
return $g$
```

it achieves greater than 99.6% accuracy on each gender in our validation set[2] and is interpretable, whereas methods relying on LLM calls or other models are not as interpretable.

3.2 Calibrated Marked Words

To identify the statistically significant words that differentiate generations from men and women, we develop the Calibrated Marked Words method, inspired by the Marked Personas method introduced by Cheng et al. [23]. Marked Personas [23], developed using the Fightin' Words Method [70], uses the log-odds probability with a prior (reference corpus of generated text) to identify statistically significant words. We build on this method by 1) adding a calibration step to ensure common words that appear in English (i.e., “the”, “a”, “an”, etc.) and the twenty[3] most common words shared across all generations do not appear as statistically significant through hyperparameter tuning described in Appendix B.2 and 2) rather than using the generated text as our prior, we use a hybrid prior consisting of both the English language[4] and the generated text. We selected the hybrid prior because we observed that the quality of statistically significant words was highest for the hybrid prior as opposed to either the topic or the English language prior. We find that our Calibrated Marked Words method removes common words and results in higher quality statistically significant words. Appendix B.2 contains this method and the qualitative difference between the resulting words from Calibrated Marked Words and Marked Words.

3.3 Subset Similarity Metric and Subset Representational Bias Score

The Subset Representational Bias Score allows for the comparison of two candidate sets S_{C_1}, S_{C_2} to each other by comparing how similar they are to two target sets S_{T_1}, S_{T_2}. The Subset Similarity Metric allows for the comparison of each candidate set to each target set. Candidate sets S_{C_1} and S_{C_2} are collections of elements assessed or tested in

[2] The construction of our validation set differs from our test set. The validation set consists of generations where we specify gender in the prompt, whereas our test set consists of generations where we do not specify gender. This is described in Appendix B.1.

[3] This is a hyperparameter we selected for our method.

[4] We us the Brown corpus from nltk [62].

relation to specific criteria, but they lack a direct basis for comparison. On the other hand, the target sets S_{T_1} and S_{T_2} serve as benchmarks or references, providing a common ground for comparison. For example, in our case, C_1 refers to associated female, C_2 refers to associated male, T_1 refers to specified female and T_2 refers to specified male.

We are interested in understanding how similar the two associated gender sets are to each other by comparing their similarity to the two specified gender sets which consist of word embeddings[5] that correspond to the statistically significant words that differentiate each gender. Here the *associated gender* sets refer to the word embeddings associated with generations where gender is not prompted, whereas the *specified gender* sets refer to the word embeddings associated with generations where gender is prompted.

Definition 3.1 (Subset similarity metric). Let A, B be two sets of vectors in $\mathbb{R}^d$ and let $\|\cdot\|$ be the Euclidean norm. We define

$$\mathrm{d_{sub}}(A\|B) := \frac{1}{|A|}\sum_{v\in A}\min_{u\in B}\left(1-\frac{\langle u,v\rangle}{\|u\|\|v\|}\right).$$

This approach enables us to calculate the average cosine distance between each element in A and its most similar counterpart in B. This means that $\mathrm{d_{sub}}(A\|B)$ always belongs to $[0,2]$. $\mathrm{d_{sub}}(A\|B) = 0$ indicates that every element in A is located in B whereas $\mathrm{d_{sub}}(A\|B) = 2$ indicates that all of the elements of A are very different from B.

Definition 3.2 (Subset Representational Bias Score). Let S, A, B be three sets of vectors in $\mathbb{R}^d$. We define

$$\Delta(S\|A,B) = \mathrm{d_{sub}}(S\|A) - \mathrm{d_{sub}}(S\|B)$$

We calculate $\mathrm{d_{sub}}(S_{C_1}\|S_{T_1})$ to understand the similarity between S_{C_1} and S_{T_1} and $\mathrm{d_{sub}}(S_{C_1}\|S_{T_2})$ to understand the similarity between S_{C_1} and S_{T_2}. We repeat this process for S_{C_2} to calculate $\mathrm{d_{sub}}(S_{C_2}\|S_{T_1})$ and $\mathrm{d_{sub}}(S_{C_2}\|S_{T_2})$. In our case, where C_1 refers to associated women and T_1 refers to specified women, $\mathrm{d_{sub}}(S_{C_1}\|S_{T_1})$ entails comparing each statistically significant word for inferred women to every significant word for specified women, selecting the word with the smallest cosine distance. This process is repeated for each significant word for associated women, and the average cosine distance serves as the similarity metric between associated and specified women. We can now calculate $\Delta(S_{C_1}\|S_{T_1}, S_{T_2})$ and $\Delta(S_{C_1}\|S_{T_1}, S_{T_2})$ allowing us to observe the similarity between S_{C_1} and S_{C_2} directly using the comparison between these values. A Subset Representational Bias Score of 0 indicates no bias towards either specified men or women. Further details on this algorithm are provided in Algorithm 5, Appendix B.1.

4 EXPERIMENTS

We run each of our experiments using various open-source and closed-source state-of-the-art models. Specifically, we test Meta's Llama-3.1-70b model and OpenAI's GPT-3.5 and GPT-4o-mini models.[6] We run prompts with and without gender. The genders we investigate are male, female, and non-binary. We test 63 occupations.[7] The prompts used are 'Generate a persona of a [OCCUPATION]' and 'Describe a [OCCUPATION] as if you are writing a biography,' and when we prompt for gender, we use 'Generate a persona of a [OCCUPATION] who is a [woman/man/non-binary person]' and 'Describe a [OCCUPATION] who is a [woman/man/non-binary person] as if you are writing a biography'. When describing the generations, we refer to generations resulting from the prompts without gender as *associated gender*, and we refer to the prompt resulting from genders with specified gender as *specified gender.*

[5] We utilize the Word2Vec [68] word embeddings from gensim [83].
[6] Specifically, gpt-3.5-turbo-0125 and gpt-4o-mini-2024-07-18.
[7] 60 of which are from the Winogender dataset [88], and we add "software engineer", "cook", and "pilot."

4.1 Who is represented?

To investigate who is represented in an occupation, for each occupation, we generate 100 generations per prompt. We then utilize the Gender Association Algorithm described in Section 3.1 to associate gender to each generation. We then compare the percentage of women in each occupation to the Bureau of Labor and Statistics (BLS) [76]. To observe the differences between the BLS and the models, we divide the occupations based on whether the occupation is female or male-dominated according to the BLS. We calculate the percentage of women associated with every occupation and count the occupations based on the percent decile (i.e., 0-10, 10-20, etc.). This allows us to analyze patterns across female and male-dominated occupations while also noting patterns specific to either female or male-dominated occupations. We also report non-binary representation by calculating the non-binary representation associated with every occupation and count the occupations based on the percentile (i.e., 0, 0-1, 1-2, etc.).

4.2 How are people represented?

To analyze the similarity between the generations associated with gender, we first identify the statistically significant words between generations associated with gender. To ensure statistical significance, we generate personas until we have at least 100 personas per occupation, associated gender, and prompt. We require that at least 10% of instances be associated with each gender for an occupation to be considered due to computational limitations. We do not consider non-binary gender in this analysis as generations associated with non-binary constitute less than 10% of generations. On average 1000 generations per occupation and prompt are needed to have 100 generations per associated gender. We associate gender with each generation using the Gender Association Method described in Section 3.1. When using the Calibrated Marked Words method, we identify statistically significant words per occupation and associated gender to ensure occupational words shared across gender are not identified. For example, “doctor” might be equally present in generations associated with women and generations associated with men for the “doctor” occupation, but if “doctor” is statistically significant for female engineers, it might be flagged as statistically significant for women if the generations compared are not from the same occupation.

To compare the similarity of statistically significant words between associated men and women, we utilize the Subset Similarity Metric as we cannot directly compare generations associated with men and women. Thus, we also generate 100 personas per occupation, gender, and prompt using the prompts where gender is specified to serve as our basis for comparison. This results in 600 ($100 \times 3 \times 2$) generations per occupation and 37800 (600×63) total generations. The statistically significant words for specified gender are identified using the Calibrated Marked Words method per occupation and gender. Prior to comparison using the Subset Similarity Metric, we remove pronouns from the statistically significant words as differences in pronouns are expected. Our candidate sets are the word embeddings for the statistically significant words for associated men (S_{AM}) and women (S_{AF}), and our target sets are the word embeddings for the statistically significant words for specified men (S_M) and women (S_F).

We then utilize the Subset Representational Bias Scores to understand if there is a statistically significant difference in how associated men and associated women are described. We compare the $\Delta(S_{AF}||S_F, S_M)$ for associated women and the $\Delta(S_{AM}||S_F, S_M)$ for associated men which is between -2 and 2. If $\Delta(S_{AF}||S_F, S_M)$ is equivalent to $\Delta(S_{AM}||S_F, S_M)$, this implies that there is no gendered difference in the statistically significant words for associated men and women. We find that the differences between $\Delta(S_{AF}||S_F, S_M)$ and $\Delta(S_{AM}||S_F, S_M)$ are statistically significant, as we compute the p-scores per model between the average Subset Representational Bias Score for each occupation between associated men and women. Each p-score was less than 0.05, and the exact p-scores are provided in Table 5 in Appendix C.1.

5 ANALYSIS

We first analyze who is represented within occupations by observing the gender distribution. We then compare how generations associated with men and women are described across occupations and models. Finally, we look at the statistically significant words and analyze how stereotypes, representational harms, and neoliberal ideals are reinforced. Throughout our analysis we use associated gender and specified gender. *Associated gender* refers to generations resulting from prompts where gender is not explicitly prompted, and *specified gender* refers to generations resulting from prompts where gender is explicitly prompted.

5.1 Who is represented?

Previous research has highlighted biases linking gender to occupation, where male pronouns are more commonly associated with male-dominated occupations and female pronouns with female-dominated ones [56, 60, 64, 88]. To examine whether these biases persist in downstream tasks where gender is not explicitly specified, and analyze how these associations have evolved as models are updated, we conducted the experiment detailed in Section 4.1, with results presented in Figure 1. These plots show the percentage of women represented across occupations when gender is not explicitly prompted, and compares this with the U.S. Bureau of Labor Statistics (BLS). In a model that accurately reflects real-world labor distributions, we would expect gender representation to align more closely with BLS data. If the models were designed to equally represent men and women, the distribution would cluster around 50% for all occupations, regardless of historical gender representation. The results reveal a clear trend: the models are more likely to generate biographies of women than men, and this is true on average across occupations, such that the representation of women is greater than it is in the data from the U.S. Bureau of Labor Statistics (BLS).

Kirk et al. [55] demonstrated that GPT-2 associates more occupations with male pronouns than with female pronouns, Mattern et al. [64] showed that GPT-3 is more likely to associate men with male-dominated occupations and women with female-dominated ones, and Kotek et al. [56] showed that in four publicly available large language models in 2023 occupations associated with men are more strongly linked to men. The results in Figure 1 indicate a departure in how models associate gender with occupation from these previous results. As our results show, GPT-3.5, GPT-4o-mini, and Llama-3.1 are more likely to generate biographies of women than with men, and this extends even to male-dominated occupations, where the majority are still primarily associated with women. On average, across occupations, the percentage of women exceeds that of men, and this trend seems to exacerbate for more recent models. For instance, among male-dominated occupations, GPT-3.5 was much more likely to depict a small percentage of women, whereas GPT-4o-mini was more likely to depict majority women. Interestingly, the increase in female representation is pronounced across both male- and female-dominated occupations. However, this shift is not observed in traditionally male-dominated blue-collar occupations, such as technician, plumber, janitor, and carpenter, where female representation remains largely unchanged. While there are slight variations in gender association percentages based on the model and prompt used, the overall trend of increased female representation persists across all prompts, models, and occupations tested.

We also examine non-binary representation across occupations and find that non-binary representation is 0% for all occupations in both GPT-3.5 and Llama-3.1 and for the majority of occupations (35 out of 63) in GPT-4o-mini. In the U.S., approximately 1.6% of the population identifies as non-binary [11], and our analysis shows that only 12 occupations surpass this representation benchmark in GPT-4o-mini. These results are presented in Figure 6 in Appendix A, and they highlight the persistent underrepresentation of non-binary individuals in generative models. Although there

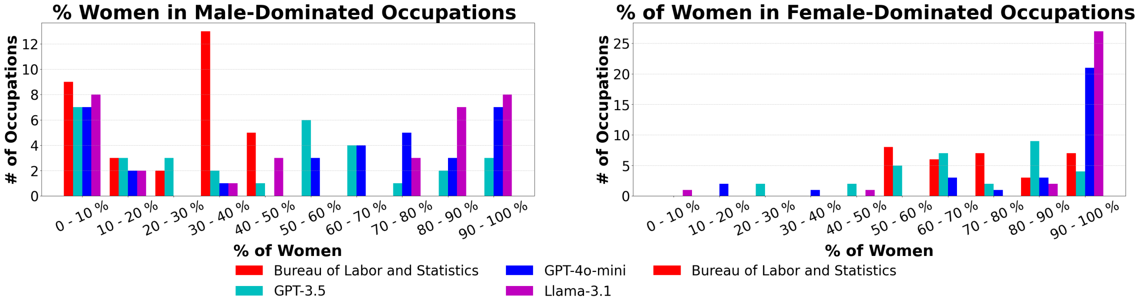


Fig. 1. The graphs illustrate the distribution of women's representation across various occupations by grouping percentages into percent deciles (e.g., 0–10%, 10–20%, and so on) and counting the number of occupations within each decile. Graph (a) shows the percentage of women in male-dominated occupations, and Graph (b) shows the percentage of women in female-dominated occupations.

was an increase in non-binary representation for some occupations in GPT-4o-mini, an increase in representation does not necessarily translate into accurate or non-stereotypical descriptions of non-binary individuals. Unfortunately, the limited data on non-binary representation prevents a more detailed analysis of how non-binary individuals are characterized within these generations.

The large representation of women, which departs from what one would expect based on empirical findings in prior models, suggests that some form of bias mitigation intervention may have been applied to influence the change in distribution. However, this mitigation appears to focus on increasing female representation rather than ensuring men and women are equally represented, as we do not observe a corresponding decrease in female representation within female-dominated occupations. While increasing female representation in male-dominated occupations can help challenge gender stereotypes, failing to address representation imbalances between men and women or further increasing female representation in female-dominated occupations risks reinforcing existing stereotypes associated with these roles. Our findings indicate that female representation in female-dominated occupations has increased significantly, as the number of occupations falling within the 70–100% representation range is substantially higher than what the BLS reports for those categories. This suggests that bias amplification is occurring, leading to an overrepresentation of women in traditionally female-dominated occupations [56]. Furthermore, we note that GPT-4o-mini and Llama-3.1 exhibit a higher percentage of women compared to GPT-3.5. This is particularly evident in the increased number of occupations falling within the 70–100% representation range in Figure 1a and the 90–100% range in Figure 1b for GPT-4o-mini and Llama-3.1, compared to GPT-3.5.

5.2 How are people represented?

The representation of women across occupations, especially across male-dominated occupations, may address some concerns of visibility, insofar as not being represented would constitute a representational harm. However, this does not entail that women and men are described similarly or that stereotypes and other representational harms have been eliminated in model generations. To explore these disparities, we employ the Subset Representation Bias Score, as outlined in Section 3.3, to analyze statistically significant differences in word usage. Our findings reveal that the Subset Representation Bias Score—which calculates the difference between similarity to specified women and specified men—varies notably between associated women and men. As shown in Figure 2, associated women are more similar to

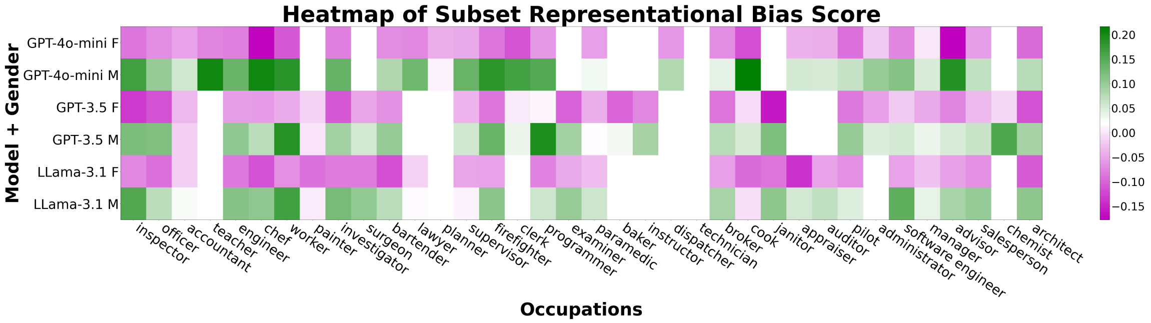


Fig. 2. The Subset Representation Bias Score is displayed for each occupation, model, and associated gender pair. A negative value (pink) indicates that the statistically significant words are closer to specified women, and a positive value (green) indicates that the statistically significant words are closer to specified men. The white boxes refer to occupation model pairs that did not meet our criteria (described in Section 4.2) to collect data.

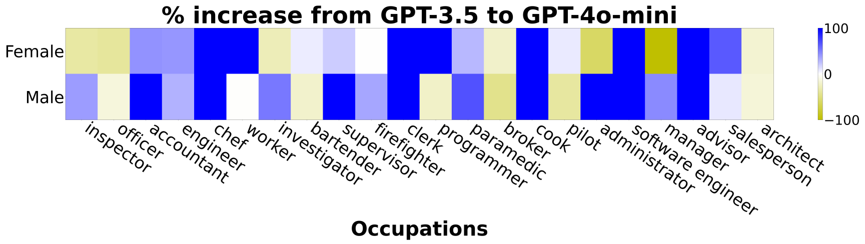


Fig. 3. Percent change in the Subset Representational Bias Score from GPT-3.5 to GPT-4o-mini. Percentage increase (blue) means that the similarity to the corresponding gender (i.e. associated women to specified women) increased from GPT-3.5 to GPT-4o-mini.

specified women than associated men, resulting in a negative score. Conversely, associated men are more similar to specified men, resulting in a positive score.[8]

The statistically significant differences between the scores of men and women reveal that personas and biographies of men and women are described and treated differently. While some variation in individual personas and biographies across gender is expected, we would not anticipate statistically significant differences to persist if biases in how people are represented have been adequately addressed. This suggests that social biases present when gender is specified persist in generations where gender is not specified in the prompt.

We analyze the change in the Subset Representation Bias Score from GPT-3.5 to GPT-4o-mini and observe that, on average, the similarity between statistically significant words associated with associated gender and specified gender is higher in GPT-4o-mini compared to GPT-3.5. Figure 3 illustrates the percentage change between the two models, with occupations such as "software engineer," "cook," and "chef" showing a 100% increase in similarity across both genders. This finding indicates that the transition from GPT-3.5 to GPT-4o-mini has amplified biases in the similarity between associated and specified genders.

[8] A positive score is associated with men because $\mathrm{d}_{\mathrm{sub}}(S_{AM} \| S_F)$ would be closer to 2 as statistically significant words for associated men and specified women are not very similar. $\mathrm{d}_{\mathrm{sub}}(S_{AM} \| S_F)$ would be closer to 0 as statistically significant words for associated men and specified men would be similar. As $\Delta(S_{AM} \| S_F, s_M) = \mathrm{d}_{\mathrm{sub}}(S_{AM} \| S_F) - \mathrm{d}_{\mathrm{sub}}(S_{AM} \| S_M)$, a subset representative bias score that is positive indicates that associated men are more similar to specified men than women. A negative subset representative bias score for associated women indicates that associated women are more similar to specified women than men.

5.3 What are the implications of how people are represented?

Understanding the implications of how people are represented requires understanding the statistically significant differences in language choice as well as the types of representational harms and their implications. For instance, it could be unproblematic if the statistical differences captured in Figure 2 were driven by gendered nouns, such as women being more often described as “wife" and men being more often described as “husband." To determine whether the differences are driven merely by grammatical factors, or whether there are signficant representational differences, we qualitatively examined the statistically significant words and developed categories grounded on terms associated with gender stereotypes and other representational harms studied in the social science literature.

In Section 5.3.1 and Section 5.3.2, we examine trends of word choices across occupations that relate to stereotypes and reinforcing systems of oppression by shifting responsibility to the individual rather than addressing structural issues. These trends indicate that the difference in scores between associated women and men implies that the stereotypes and representational harms observed in generations where gender is explicitly prompted persist even when gender is not referenced in the prompt. This also implies that the high prevalence of women depicted across occupations could risk a proliferation of these stereotypes. If models are increasingly more likely to portray women when gender is unspecified in the prompt (as discussed in Section 5.1), but *how* women are portrayed is still subject to the representational harms characterized in prior literature, these harmful representations are increasingly produced unpromptedly.

5.3.1 Stereotypes and Representational Harms. Stereotypes as a representational harm have been outlined by numerous representational harm taxonomies [24, 52, 92]. Gender stereotypes are prevalent across various contexts and can have harmful effects, whether the stereotype is perceived as positive or negative [14, 53]. Stereotype-related words identified in our analysis include personality traits, sports, societal expectations, and negative stigmas. These categories are grounded in the social science literature and reflect the U.S. cultural context. Research highlights that women are often stereotyped as caring and nurturing [33, 49], empathetic [25, 61], and family-oriented [73, 75]. Brough et al. [10] discuss how women are stereotyped as eco-friendly and environmentally conscious. Men are frequently stereotyped as being linked to sports, including basketball, soccer, and sports generally [21] and golf [66]. Meanwhile, women are associated with activities such as yoga [91]. Additionally, societal emphasis on women’s appearance and its harmful effects are well-documented [13]. Regarding negative stigmas, men face significant societal stigma surrounding mental health, as explored by Chatmon [22] and Pattyn et al. [78]. We determined the words that would be associated with these categories by qualitatively analyzing the words and validating that the context the words were used in positioned them to be part of the categories. For example, we validated that “tie" was typically used to describe an article of clothing as opposed to “tie ideas together." The words associated with these categories are provided Appendix D.1.

The prevalence of these stereotypes is illustrated in Figure 4. Figure 4a shows that words related to care, empathy, mental health, nurturing, and the environment are predominantly associated with women. This finding suggests that stereotypes of women as caring, empathetic, and nurturing are perpetuated across occupations in generations associated with women. The absence of mental health related words in male-associated generations suggests a continuation of the status quo, wherein men’s mental health is more stigmatized. Similarly, the higher prevalence of environmental language for women reinforces the stereotype that women are more concerned about environmental issues, more likely to engage in environmental advocacy, and more inclined to make eco-friendly purchases [10]. This reinforcement may discourage men from participating in these activities, further entrenching gendered expectations around environmental responsibility. Figure 4b highlights that gender stereotypes persist in sports-related words where terms associated with women are statistically significant only for women, while those associated with men are significant only for men. This

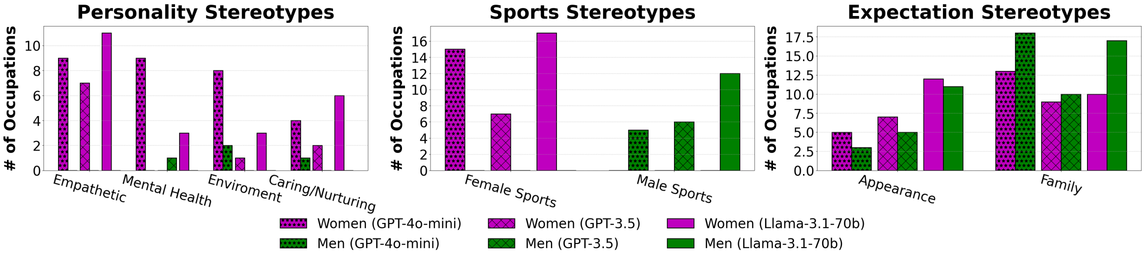


Fig. 4. Graphs (a) (personality stereotypes), (b) (sports stereotypes), and (c) (expectation stereotypes) demonstrate the number of occupations with statistically significant words in the categories provided on the x-axis. Pink refers to female and green refers to male. The star hatch refers to GPT-4o-mini, the cross hatch refers to GPT-3.5, and no hatch refers to Llama-3.1.

is evident as no bars of the other gender are present indicating there were 0 occupations with statistically significant words for that category. This finding indicates gendered sports stereotypes proliferate.

Figure 4c demonstrates that appearance-related words are more frequently associated with women than men. This trend reflects and reinforces societal norms that place disproportionate emphasis on women's appearance compared to men's. The continued focus on women's appearance in generative AI outputs is problematic, as it perpetuates norms where discussions of women's looks often overshadow their achievements and contributions [8, 17, 18, 96, 101]. Moreover, this focus has historically been used as a means to trivialize women's accomplishments [65], further entrenching harmful gender stereotypes [8, 17, 29, 39]. Interestingly, while most stereotypes are reinforced in these outputs, the association of women with family roles is challenged. Familial words are statistically more significant for men than for women, suggesting a bias intervention may have been utilized to counteract this stereotype. The patterns illustrated in Figure 4 reveal that many gender stereotypes are reinforced in generative outputs. This reinforcement has significant implications for downstream tasks and user interactions, as these stereotypes will likely persist and impact users. Furthermore, if synthetic data from these models is used to train new models, these stereotypes may become more ingrained, amplifying their presence in subsequent models [102, 106].

5.3.2 *Maintaining the Status Quo–Reinforcing Systems of Oppression.*

Cheng et al. [23] identified that the myth of resilience, associated with words such as “resilience" and “strength," has been utilized to normalize the environments that lead to poverty, inequality, and other social issues rather than challenging these structures that require “strength" and “resilience" [58, 87, 103]. Furthermore, Cheng et al. [23] connect the myth of resilience to greater harm for minority women and harms associated with the Strong Black Woman stereotype [105]. Education scholars have discussed how resilience is often framed within neoliberal discourses as a form of empowerment for marginalized individuals, but this framing shifts responsibility from systems to address systemic inequities onto individuals, ultimately reinforcing these structures and harming marginalized communities [26, 104]. Similarly, Clay [26] and Joseph [51] discuss how political discourses of resilience contribute to neoliberalism and place the burden on individuals to overcome systemic structures. Thus, using resilience-related words to describe underrepresented groups can contribute to neoliberal ideals that harm these groups. Neoliberalism is the theory that the well-being of communities and individuals is best advanced by free markets that allow for the optimal distribution of resources within society [19, 47]. Policies developed using this theory have been associated with poorer collective health and social well-being [19, 27, 40, 41, 46] and can lead to greater poverty [50, 74]. Thus, in our analysis of representational harms, we identify the presence of statistically significant words which social scientists have noted contribute to neoliberal narratives that can reinforce systems of oppression.

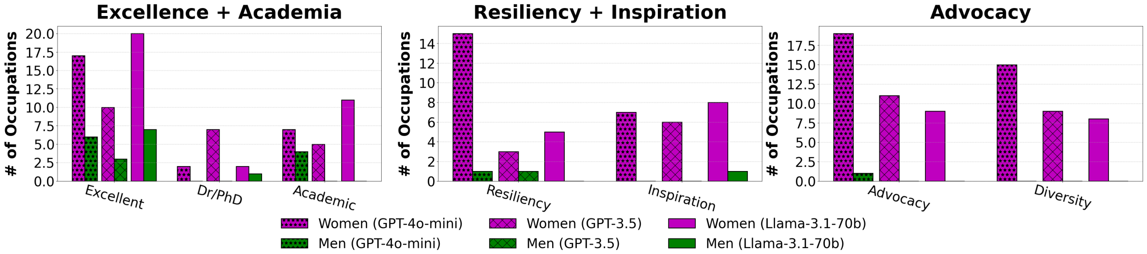


Fig. 5. Graphs (a) (excellence + academia), (b) (resiliency + inspiration), and (c) (advocacy) demonstrate the number of occupations with statistically significant words in the categories provided on the x-axis. Pink refers to female and green refers to male. The star hatch refers to GPT-4o-mini, the cross-hatch refers to GPT-3.5, and no hatch refers to Llama-3.1.

Our categories for this analysis are excellence, academia, resilience, and inspiration. We also include the categories of advocacy and diversity as they are associated with gendered expectations. Table 6 in Appendix D.1 provides a detailed breakdown of the specific words related to each category. Sociology researchers note that the discourse surrounding excellence and academia often emphasizes individual effort and achievement reinforcing neoliberal ideals [6, 7, 35, 42, 67, 86, 89, 90]. Researchers also discuss how discussion of excellence and academia support the meritocracy myth—the notion that success stems primarily from individual effort [5, 9, 57, 59, 82, 85, 94]. Adamson & Kelan [2] and Byrne et al. [15] discuss how media representation of “inspirational" women emphasize that their achievements result primarily from individual effort, ignoring broader structural and systemic factors. Such narratives contribute to the perpetuation of systemic inequalities by obscuring the societal and institutional barriers that many face.

Figure 5a illustrates that words related to excellence, academia, and doctorate degrees are predominantly statistically significant for women across all models. Figure 5b shows that words related to resilience are predominantly associated with women across various occupations. This association perpetuates the notion that individuals can overcome systemic oppression solely through personal resilience, thereby shifting the focus from dismantling oppressive systems to placing the burden on individuals. Similarly, Figure 5b also reveals that words related to inspiration are also more frequently associated with women across all occupations. The manner in which “inspiration" is used in the female personas echos media discussion of inspirational women which emphasizes that their achievement resulting from individual effort ignoring broader structural and systemic factors [2, 15]. Figure 5c shows a significantly higher prevalence of words related to advocacy and diversity for women compared to men across all occupations and models. Women and other underrepresented groups often feel, or are pressured, to represent their communities, engage in diversity initiatives, and mentor junior colleagues or students from underrepresented backgrounds [3, 4, 32, 44, 63, 69, 79, 80, 84, 93, 97, 109]. The overrepresentation of advocacy- and mentorship-related words associated with women reinforces the expectation that women bear greater responsibility for advancing diversity and inclusion than men. This places the burden of addressing systemic inequities on women, rather than holding institutions and organizations accountable for meaningful change [80]. Achieving equity in the prevalence of advocacy-related language for both men and women would signal that advocacy and mentorship are collective responsibilities, not burdens to be disproportionately shouldered by marginalized groups. Recommendations for improving diversity, equity, and inclusion in workplaces and universities emphasize the importance of involving stakeholders from all groups and levels of the organization [98].

6 IMPLICATIONS

Our findings have significant implications for researchers, model developers, and users. Our results in Section 5.2 reveal that gender differences persist across generations of models in the absence of explicitly prompting gender. Furthermore, these differences often reflect stereotypes and perpetuate harmful narratives. These results suggest harms stemming from gender representations can propagate and proliferate into downstream tasks such as creative composition, explanation and reasoning, general information retrieval, and persona generation.

Our findings in Section 5.1 indicate bias mitigation methods may have been applied as female representation is much greater than what would be expected based on previous literature studying older models of LLMs. Furthermore, comparison across the models considered in our study also suggests that these changes may continue to exacerbate over time. Crucially, our findings also challenge the assumption that non-gendered prompts are free of gender bias. As we showcase, non-gendered prompts still result in representational harms. These biases may manifest in downstream tasks such as writing a story about a doctor, providing general information or explanations about groups with implicit gender associations (e.g., teachers), or generating personas like students. This highlights the need for robust bias evaluation methods to identify biases in contexts where gender and other social groups are not explicitly specified. Additionally, it emphasizes the importance of developing mitigation strategies that address representational biases in who is represented and how people are represented to reduce harm in real-world applications.

Building on our findings, we echo Cheng et al.'s [23] recommendation that model developers transparently disclose the bias mitigation methods employed, including the use of synthetic data, Reinforcement Learning from Human Feedback (RLHF), and Reinforcement Learning with AI Feedback (RLAIF). These methods may unintentionally reinforce existing biases or introduce new ones. For example, in Section 5.2 we show how gender distributions shifted between GPT-3.5 and GPT-4o-mini, with GPT-4o-mini being even more likely to depict women when gender was not specified in the prompt. Several factors could contribute to this. As OpenAI has not disclosed specific details about how GPT-4o-mini was trained, we cannot confirm the exact cause of this effect. Thus, transparency in these processes is essential for anticipating and addressing unintended consequences.

Although our analysis does not directly include race or other sociodemographic axes (like religion or economic status), it is crucial to understand how these axes of identity are present in model outputs when not specifically prompted. We urge researchers to develop new methods to allow for this understanding. Our development of the Subset Similarity Metric and Subset Representational Bias Score aids in this by helping practitioners and researchers understand the difference between groups as these metrics aided us in understanding differences between how men and women are described when gender is not prompted. Usage of these metrics requires consideration of context as expected behavior and representational harms are context-specific. We also advocate for the development of evaluations that assess social biases in contexts where social groups are not explicitly prompted. When considering depictions of different demographic groups without explicit prompts, extending beyond gender poses challenges; while gender associations can be directly drawn from the pronouns used in the biography, the same is not true for other demographics.

7 CONCLUSION

In this paper, we demonstrate that *who* is represented within occupation has departures from previous analysis of gender in occupation–women comprise the majority of personas and biographies across state-of-the-art models– but *how* women are represented continues to be harmful. Specifically, we demonstrate statistical differences between women and men continue when models are not prompted with gender, and the representational harms associated with women persist in these generations. Furthermore, we qualitatively investigated these statistically significant differences in word choice between men and women and found that stereotypes and harmful narratives, such as neoliberal ideals embedded in the meritocracy myth, are perpetuated. With the increased representation of women, this implies that these representational harms proliferate, calling for careful consideration of the interplay between different forms of representational harms, particularly in the usage of bias mitigation interventions.

8 ADVERSE IMPACT STATEMENT

Our work should not be misinterpreted to mean that model developers should avoid using bias mitigation methods for fear of unintentionally worsening representational harms. Rather, the use of bias mitigation methods requires careful consideration and nuance of their impact. The metrics we propose should also not be used to "game" treating groups equally regardless of context. Rather, the metrics are a tool to provide insight into how groups are treated differently, and its interpretation requires consideration of context as we have illustrated in our analysis.

Furthermore, this paper highlights the importance of transparency in model development, underscoring the need to disclose how the inclusion or absence of bias mitigation methods influences observed model behavior. Such transparency fosters a deeper understanding and more informed efforts to address representational harms effectively.

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 298–306.

[2] Maria Adamson and Elisabeth K Kelan. 2019. 'Female heroes': celebrity executives as postfeminist role models. *British Journal of Management* 30, 4 (2019), 981–996.

[3] Antonette A Ajayi, Fatima Rodriguez, and Vinicio de Jesus Perez. 2021. Prioritizing equity and diversity in academic medicine faculty recruitment and retention. In *JAMA Health Forum*, Vol. 2. American Medical Association, e212426–e212426.

[4] Özlem Altan-Olcay and Suzanne Bergeron. 2024. Care in times of the pandemic: Rethinking meanings of work in the university. *Gender, Work & Organization* 31, 4 (2024), 1544–1559.

[5] Lorriz Anne Alvarado. 2010. Dispelling the meritocracy myth: Lessons for higher education and student affairs educators. *The Vermont Connection* 31, 1 (2010), 2.

[6] Martin Benedict Andrew. 2024. Just get them over the line': Neoliberalism and the execution of 'excellence. *Journal of Applied Learning and Teaching* 7, 1 (2024).

[7] Darren T Baker and Deborah N Brewis. 2020. The melancholic subject: A study of self-blame as a gendered and neoliberal psychic response to loss of the 'perfect worker'. *Accounting, Organizations and Society* 82 (2020), 101093.

[8] Judith Baxter. 2018. *Women leaders and gender stereotyping in the UK Press: A poststructuralist approach*. Springer.

[9] Mary Blair-Loy and Erin A Cech. 2022. *Misconceiving merit: Paradoxes of excellence and devotion in academic science and engineering*. University of Chicago Press.

[10] Aaron R Brough, James EB Wilkie, Jingjing Ma, Mathew S Isaac, and David Gal. 2016. Is eco-friendly unmanly? The green-feminine stereotype and its effect on sustainable consumption. *Journal of consumer research* 43, 4 (2016), 567–582.

[11] Anna Brown. 2022. About 5% of young adults in the U.S. say their gender is different from their sex assigned at birth. *Pew Research Center* (2022). https://www.pewresearch.org/short-reads/2022/06/07/about-5-of-young-adults-in-the-u-s-say-their-gender-is-different-from-their-sex-assigned-at-birth/

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[13] Vanessa M Buote, Anne E Wilson, Erin J Strahan, Stephanie B Gazzola, and Fiona Papps. 2011. Setting the bar: Divergent sociocultural norms for women's and men's ideal appearance in real-world contexts. *Body image* 8, 4 (2011), 322–334.

[14] Melissa Burkley and Hart Blanton. 2009. The positive (and negative) consequences of endorsing negative self-stereotypes. *Self and Identity* 8, 2-3 (2009), 286–299.

[15] Janice Byrne, Salma Fattoum, and Maria Cristina Diaz Garcia. 2019. Role models and women entrepreneurs: Entrepreneurial superwoman has her say. *Journal of Small Business Management* 57, 1 (2019), 154–184.

[16] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[17] Deborah Cameron and Sylvia Shaw. 2016. *Gender, power and political speech: Women and language in the 2015 UK General Election*. Springer.

[18] Donatella Campus. 2013. *Women political leaders and the media.* Springer.

[19] Kiffer G Card and Kirk J Hepburn. 2023. Is neoliberalism killing us? A cross sectional study of the impact of neoliberal beliefs on health and social wellbeing in the midst of the COVID-19 pandemic. *International Journal of Social Determinants of Health and Health Services* 53, 3 (2023), 363–373.

[20] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217* (2023).

[21] Aïna Chalabaev, Philippe Sarrazin, Paul Fontayne, Julie Boiché, and Corentin Clément-Guillotin. 2013. The influence of sex stereotypes and gender roles on participation and performance in sport and exercise: Review and future directions. *Psychology of sport and exercise* 14, 2 (2013), 136–144.

[22] Benita N Chatmon. 2020. Males and mental health stigma. , 1557988320949322 pages.

[23] Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 1504–1532.

[24] Jennifer Chien and David Danks. 2024. Beyond Behaviorist Representational Harms: A Plan for Measurement and Mitigation. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency.* 933–946.

[25] Leonardo Christov-Moore, Elizabeth A Simpson, Gino Coudé, Kristina Grigaityte, Marco Iacoboni, and Pier Francesco Ferrari. 2014. Empathy: Gender effects in brain and behavior. *Neuroscience & biobehavioral reviews* 46 (2014), 604–627.

[26] Kevin L Clay. 2019. "Despite the odds": Unpacking the politics of Black resilience neoliberalism. *American Educational Research Journal* 56, 1 (2019), 75–110.

[27] Chik Collins, Gerry McCartney, and Lisa Garnham. 2015. Neoliberalism and health inequalities. *Health inequalities: Critical perspectives* 124 (2015).

[28] Jenny L Davis, Apryl Williams, and Michael W Yang. 2021. Algorithmic reparation. *Big Data & Society* 8, 2 (2021), 20539517211044808.

[29] Carolin Debray, Stephanie Schnurr, Joelle Loew, and Sophie Reissner-Roubicek. 2024. An 'attractive alternative way of wielding power'? Revealing hidden gender ideologies in the portrayal of women Heads of State during the COVID-19 pandemic. *Critical Discourse Studies* 21, 1 (2024), 52–75.

[30] Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. 2023. Queer people are people first: Deconstructing sexual identity stereotypes in large language models. *arXiv preprint arXiv:2307.00101* (2023).

[31] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 8173–8188.

[32] Danielle Docka-Filipek and Lindsey B Stone. 2021. Twice a "housewife": On academic precarity,"hysterical" women, faculty mental health, and service as gendered care work for the "university family" in pandemic times. *Gender, Work & Organization* 28, 6 (2021), 2158–2179.

[33] Alice H Eagly and Valerie J Steffen. 1984. Gender stereotypes stem from the distribution of women and men into social roles. *Journal of personality and social psychology* 46, 4 (1984), 735.

[34] Zahra Fatemi, Chen Xing, Wenhao Liu, and Caimming Xiong. 2023. Improving Gender Fairness of Pre-Trained Language Models without Catastrophic Forgetting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* 1249–1262.

[35] Zeena Feldman and Marisol Sandoval. 2018. Metric power and the academic self: Neoliberalism, knowledge and resistance in the British university. *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society* 16, 1 (2018), 214–233.

[36] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. " I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.* 205–216.

[37] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics* (2024), 1–79.

[38] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, and Franck Dernoncourt. 2024. Self-Debiasing Large Language Models: Zero-Shot Recognition and Reduction of Stereotypes. *arXiv preprint arXiv:2402.01981* (2024).

[39] Iñaki Garcia-Blanco and Karin Wahl-Jorgensen. 2012. The discursive construction of women politicians in the European press. *Feminist Media Studies* 12, 3 (2012), 422–441.

[40] Lisa M Garnham. 2017. Public health implications of 4 decades of neoliberal policy: a qualitative case study from post-industrial west central Scotland. *Journal of Public Health* 39, 4 (2017), 668–677.

[41] Kathomi Gatwiri, Julians Amboko, and Darius Okolla. 2019. The implications of Neoliberalism on African economies, health outcomes and wellbeing: a conceptual argument. *Social Theory & Health* 18, 1 (2019), 86.

[42] Natalia Gerodetti and Martha McNaught-Davis. 2017. Feminisation of success or successful femininities? Disentangling 'new femininities' under neoliberal conditions. *European Journal of Women's Studies* 24, 4 (2017), 351–365.

[43] Sourojit Ghosh, Nina Lutz, and Aylin Caliskan. 2024. "I Don't See Myself Represented Here at All": User Experiences of Stable Diffusion Outputs Containing Representational Harms across Gender Identities and Nationalities. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 463–475.

[44] Cassandra M Guarino and Victor MH Borden. 2017. Faculty service loads and gender: Are women taking care of the academic family? *Research in higher education* 58 (2017), 672–694.

[45] Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1012–1023.

[46] Christopher Hartmann. 2016. Postneoliberal public health care reforms: neoliberalism, social medicine, and persistent health inequalities in Latin America. *American Journal of Public Health* 106, 12 (2016), 2145–2151.

[47] David Harvey. 2007. *A brief history of neoliberalism*. Oxford University Press, USA.

[48] Lukas Hauzenberger, Shahed Masoudian, Deepak Kumar, Markus Schedl, and Navid Rekabsaz. 2023. Modular and On-demand Bias Mitigation with Attribute-Removal Subnetworks. In *Findings of the Association for Computational Linguistics: ACL 2023*. 6192–6214.

[49] David Hesmondhalgh and Sarah Baker. 2015. Sex, gender and work segregation in the cultural industries. *The sociological review* 63, 1_suppl (2015), 23–36.

[50] Evelyne Huber and Fred Solt. 2004. Successes and failures of neoliberalism. *Latin American Research Review* 39, 3 (2004), 150–164.

[51] Jonathan Joseph. 2013. Resilience as embedded neoliberalism: a governmentality approach. *Resilience* 1, 1 (2013), 38–52.

[52] Jared Katzman, Angelina Wang, Morgan Scheuerman, Su Lin Blodgett, Kristen Laird, Hanna Wallach, and Solon Barocas. 2023. Taxonomizing and measuring representational harms: A look at image tagging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14277–14285.

[53] Aaron C Kay, Martin V Day, Mark P Zanna, and A David Nussbaum. 2013. The insidious (and ironic) effects of positive stereotypes. *Journal of Experimental Social Psychology* 49, 2 (2013), 287–291.

[54] Arjun Kharpal. 2024. Google pauses Gemini AI image generator after it created inaccurate historical pictures. *CNBC News* (2024). https://www.cnbc.com/2024/02/22/google-pauses-gemini-ai-image-generator-after-inaccuracies.html

[55] Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems* 34 (2021), 2611–2624.

[56] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*. 12–24.

[57] Anne Lawton. 2000. The meritocracy myth and the illusion of equal employment opportunity. *Minn. L. Rev.* 85 (2000), 587.

[58] Kelly Yu-Hsin Liao, Meifen Wei, and Mengxi Yin. 2020. The misunderstood schema of the strong Black woman: Exploring its mental health consequences and coping responses among African American women. *Psychology of Women Quarterly* 44, 1 (2020), 84–104.

[59] Amy Liu. 2011. Unraveling the myth of meritocracy within the context of US higher education. *Higher education* 62 (2011), 383–397.

[60] Yiran Liu, Ke Yang, Zehan Qi, Xiao Liu, Yang Yu, and ChengXiang Zhai. 2024. Bias and Volatility: A Statistical Framework for Evaluating Large Language Model's Stereotypes and the Associated Generation Inconsistency. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[61] Charlotte S Löffler and Tobias Greitemeyer. 2023. Are women the more empathetic gender? The effects of gender role expectations. *Current Psychology* 42, 1 (2023), 220–231.

[62] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. 63–70.

[63] Jessica L Malisch, Breanna N Harris, Shanen M Sherrer, Kristy A Lewis, Stephanie L Shepherd, Pumtiwitt C McCarthy, Jessica L Spott, Elizabeth P Karam, Naima Moustaid-Moussa, Jessica McCrory Calarco, et al. 2020. In the wake of COVID-19, academia needs new solutions to ensure gender equity. *Proceedings of the National Academy of Sciences* 117, 27 (2020), 15378–15381.

[64] Justus Mattern, Zhijing Jin, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing. *ArXiv* (2022), 2212–10678.

[65] Sharon Mavin, Patricia Bryans, and Rosie Cunningham. 2010. Fed-up with Blair's babes, Gordon's gals, Cameron's cuties, Nick's nymphets: Challenging gendered media representations of women political leaders. *Gender in Management: An International Journal* 25, 7 (2010), 550–569.

[66] Lee McGinnis, Julia McQuillan, and Constance L Chapple. 2005. I just want to play: Women, sexism, and persistence in golf. *Journal of Sport and Social Issues* 29, 3 (2005), 313–337.

[67] Angela McRobbie. 2015. Notes on the perfect: Competitive femininity in neoliberal times. *Australian feminist studies* 30, 83 (2015), 3–20.

[68] Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* 3781 (2013).

[69] Joya Misra, Jennifer Hickes Lundquist, Elissa Holmes, Stephanie Agiomavritis, et al. 2011. The ivory ceiling of service work. *Academe* 97, 1 (2011), 22–26.

[70] Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16, 4 (2008), 372–403.

[71] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).

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

[73] Clotilde Napp. 2024. Gender Stereotypes about career and family are stronger in more economically developed countries and can explain the Gender Equality Paradox. *Personality and Social Psychology Bulletin* (2024), 01461672241286084.

[74] Vicente Navarro. 2007. Neoliberalism as a class ideology; or, the political causes of the growth of inequalities. *International Journal of Health Services* 37, 1 (2007), 47–62.

[75] Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, research, and practice* 6, 1 (2002), 101.

[76] U.S. Bureau of Labor Statistics. 2024. Labor Force Statistics from the Current Population Survey. In *Occupational Outlook Handbook*. U.S. Department of Labor. https://www.bls.gov/cps/cpsaat11.htm

[77] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[78] Elise Pattyn, Mieke Verhaeghe, and Piet Bracke. 2015. The gender gap in mental health service use. *Social psychiatry and psychiatric epidemiology* 50 (2015), 1089–1095.

[79] Kamaria B Porter, Julie R Posselt, Kimberly Reyes, Kelly E Slay, and Aurora Kamimura. 2018. Burdens and benefits of diversity work: Emotion management in STEM doctoral students. *Studies in Graduate and Postdoctoral Education* 9, 2 (2018), 127–143.

[80] Karen Pyke. 2015. Faculty gender inequity and the "just say no to service" fairy tale. In *Disrupting the culture of silence*. Routledge, 83–95.

[81] Prabhakar Raghavan. 2024. Gemini image generation got it wrong. We'll do better. *URL https://blog. google/products/gemini/gemini-image-generation-issue* (2024).

[82] Saleem Razack, Torsten Risør, Brian Hodges, and Yvonne Steinert. 2020. Beyond the cultural myth of medical meritocracy. *Medical education* 54, 1 (2020), 46–53.

[83] Radim Rehurek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. 45.

[84] Rebecca A Reid. 2021. Retaining women faculty: The problem of invisible labor. *PS: Political Science & Politics* 54, 3 (2021), 504–506.

[85] Deborah L Rhode. 1996. Myths of meritocracy. *Fordham L. Rev.* 65 (1996), 585.

[86] Jessica Ringrose. 2007. Successful girls? Complicating post-feminist, neoliberal discourses of educational achievement and gender equality. *Gender and education* 19, 4 (2007), 471–489.

[87] Catherine Rottenberg. 2014. The rise of neoliberal feminism. *Cultural studies* 28, 3 (2014), 418–437.

[88] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 8–14.

[89] Daniel B Saunders. 2015. Resisting Excellence: Challenging Neoliberal Ideology in Postsecondary Education. *Journal for Critical Education Policy Studies (JCEPS)* 13, 2 (2015).

[90] Daniel B Saunders and Gerardo Blanco Ramírez. 2017. Against 'teaching excellence': Ideology, commodification, and enabling the neoliberalization of postsecondary education. *Teaching in Higher Education* 22, 4 (2017), 396–407.

[91] Alison Shaw and Esra S Kaytaz. 2021. Yoga bodies, yoga minds: contextualising the health discourses and practices of modern postural yoga. *Anthropology & Medicine* 28, 3 (2021), 279–296.

[92] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.

[93] Laurel Smith-Doerr, Ethel L Mickey, and Ember Skye W Kane-Lee. 2023. Deciding together as faculty: narratives of unanticipated consequences in gendered and racialized departmental service, promotion, and voting. *Journal of Organizational Sociology* 1, 2 (2023), 171–198.

[94] Fernanda Staniscuaski. 2023. The science meritocracy myth devalues women. *Science* 379, 6639 (2023), 1308–1308.

[95] Hao Sun, Zhexin Zhang, Fei Mi, Yasheng Wang, Wei Liu, Jianwei Cui, Bin Wang, Qun Liu, and Minlie Huang. 2023. MoralDial: A Framework to Train and Evaluate Moral Dialogue Systems via Moral Discussions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2213–2230.

[96] Irmgard Tischner, Helen Malson, and Kelly Fey. 2021. Leading ladies: discursive constructions of women leaders in the UK media. *Feminist Media Studies* 21, 3 (2021), 460–476.

[97] Briana Toole. 2019. From standpoint epistemology to epistemic oppression. *Hypatia* 34, 4 (2019), 598–618.

[98] Natsumi Ueda, Adrianna Kezar, Elizabeth Holcombe, Darsella Vigil, and Jordan Harper. 2024. Emotional Labor: Institutional Responsibility and Strategies to Offer Emotional Support for Leaders Engaging in Diversity, Equity, and Inclusion Work. *AERA Open* 10 (2024), 23328584241296092.

[99] Faye-Marie Vassel, Evan Shieh, Cassidy R Sugimoto, and Thema Monroe-White. 2024. The Psychosocial Impacts of Generative AI Harms. In *Proceedings of the AAAI Symposium Series*, Vol. 3. 440–447.

[100] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality Bias in Text Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 116–122.

[101] Clare Walsh. 2015. Media capital or media deficit? Representations of women in leadership roles in old and new media. *Feminist Media Studies* 15, 6 (2015), 1025–1034.

[102] Ze Wang, Zekun Wu, Jeremy Zhang, Navya Jain, Xin Guan, and Adriano Koshiyama. 2024. Bias Amplification: Language Models as Increasingly Biased Media. *arXiv preprint arXiv:2410.15234* (2024).

[103] Natalie N Watson and Carla D Hunter. 2016. "I had to be strong" tensions in the strong Black woman schema. *Journal of Black Psychology* 42, 5 (2016), 424–452.

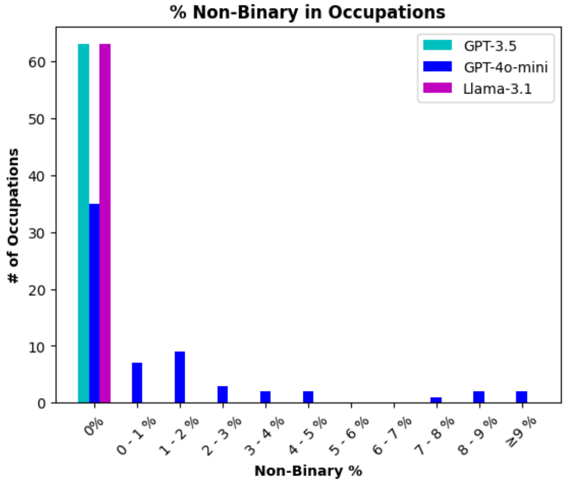


Fig. 6. Percent of generations associated with non-binary per occupation based on model.

[104] David Webster and Nicola Rivers. 2019. Resisting resilience: disrupting discourses of self-efficacy. *Pedagogy, Culture & Society* 27, 4 (2019), 523–535.

[105] Cheryl L Woods-Giscombé. 2010. Superwoman schema: African American women's views on stress, strength, and health. *Qualitative health research* 20, 5 (2010), 668–683.

[106] Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. 2024. Fairness feedback loops: training on synthetic data amplifies bias. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency.* 2113–2147.

[107] Ke Yang, Charles Yu, Yi R Fung, Manling Li, and Heng Ji. 2023. Adept: A debiasing prompt framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 10780–10788.

[108] Liu Yu, Yuzhou Mao, Jin Wu, and Fan Zhou. 2023. Mixup-based unified framework to overcome gender bias resurgence. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1755–1759.

[109] Ruth Enid Zambrana, Adia Harvey Wingfield, Lisa M Lapeyrouse, Brianne A Davila, Tangere L Hoagland, and Robert Burciaga Valdez. 2017. Blatant, subtle, and insidious: URM faculty perceptions of discriminatory practices in predominantly White institutions. *Sociological Inquiry* 87, 2 (2017), 207–232.

[110] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876* (2018).

[111] Chujie Zheng, Pei Ke, Zheng Zhang, and Minlie Huang. 2023. Click: Controllable Text Generation with Sequence Likelihood Contrastive Learning. In *Findings of the Association for Computational Linguistics: ACL 2023.* 1022–1040.

A NON-BINARY REPRESENTATION

Figure 6 demonstrates non-binary representation across occupations and models.

B METHODOLOGY

B.1 Gender Association Method

We test our Gender Association Method on generations where we specify gender and our method's performance is reported in Table 1. Of the generations analyzed where gender is not specified, the percent of generations where gender is associated per model is displayed in Table 2.

Gender	Correct %	Incorrect%	Not Captured%
Female	99.9180	0.0080	0.0740
Male	99.8463	0.0053	0.1483
Non-binary	99.6693	0.0037	0.3280

Table 1. Percentage of generations across GPT-3.5, GPT-4o-mini, and Llama-3.1 for which gender is correctly and incorrectly identified using the Gender Association Method as well as the percentage of generations that are not captured. In practice, not-captured generations are dropped and not used in analysis.

	GPT-3.5	GPT-4o-mini	Llama-3.1-70b
% Captured	80.460	94.310	98.167

Table 2. Percent of generations for which gender can be associated using the Gender Association Method per model.

Algorithm 2 Marked Personas method from Cheng et al. [23]

```
Input: $W$ (set of calibration words), $T_G$ (word counts of generations concerning group), $T_U$ (word counts for generations concerning unmarked group), $P$ (word counts of the prior)
Output: $\delta$ the z-scores of each word
Initialize $n_G \leftarrow \sum_{w \in T_G} T_G[w]$
Initialize $n_U \leftarrow \sum_{w \in T_U} T_U[w]$
Initialize $n_P \leftarrow \sum_{w \in P} P[w]$
for $w \in P$ do
    $l_1 \leftarrow \frac{T_G[w]+P[w]}{(n_U+n_P)-(T_G[w]+P[w])}$
    $l_2 \leftarrow \frac{T_U[w]+P[w]}{(n_U+n_P)-(T_U[w]+P[w])}$
    $\sigma^2 \leftarrow \frac{1}{T_G[w]+P[w]} + \frac{1}{T_U[w]+P[w]}$
    $ll_1 \leftarrow \log l_1$
    $ll_2 \leftarrow \log l_2$
    $\delta[w] \leftarrow \frac{ll_1 - ll_2}{\sigma}$
end for
return $\delta$
```

B.2 Calibrated Marked Words

The Calibrated Marked Words algorithm is described in Algorithm 4. The values of the hyperparameters, C_{English} and C_{topic} were determined using binary search and observing the resulting marked words when each prior (English and topic) where used individually. We began with the min value of 0 and 1 and the English prior and continued to binary search until the marked words returned by the English prior did not contain common words. We repeated this process for the topic prior and then observed the results of the hybrid prior using the hyperparameter p which constitutes the mixture between the topic and English priors. We then selected $p = 0.25$ as we wanted to place greater weight on the topic prior while still reaping the benefits of the English prior. We observed that the C_{English} and C_{topic} we had selected resulted in calibrated marked words that did not contain common words.

The qualitative difference between Marked Personas [23] and our Calibrated Marked Words method is demonstrated in Tables 3 and 4 by examining the difference in statistically significant words for female, male, and non-binary software engineers. Table 3 demonstrates the words captured by Marked Personas [23] and not by our Calibrated Marked Words method. Table 4 demonstrates the words captured by our Calibrated Marked Words method and not by Cheng et al.'s [23] Marked Personas.

Algorithm 3 Calculation of regularizing terms.

Input: W (set of calibration words), G_1 (word counts of generations concerning group 1), G_2 (word counts for generations concerning group 2 (unmarked group)), P_{topic} (word counts of the topic prior), P_{English} (word counts of the English prior), p (hyperparameter), C_{English} (hyperparameter), C_{topic} (hyperparameter)

Output: Return hybrid prior and regularizing terms r_1, r_2 where r_1 is the regularizing term for G_1 and r_2 is the regularizing term for G_2

```
Initialize $P \leftarrow \text{map}()$
$C \leftarrow p \cdot C_{\text{topic}} + (1 - p) \cdot C_{\text{English}}$
for $w \in P_{\text{topic}}$ do
    $P[w] \leftarrow p \cdot P_{\text{topic}}[w] + (1 - p) \cdot P_{\text{English}}$
end for
Initialize $w_p \leftarrow 0$
Initialize $w_{g_1} \leftarrow 0$
Initialize $w_{g_2} \leftarrow 0$
for $w \in W$ do
    $w_p \leftarrow w_p + P[w]$
    $w_{g_1} \leftarrow w_{g_1} + G_1[w]$
    $w_{g_2} \leftarrow w_{g_2} + G_2[w]$
end for
$r_1 \leftarrow C \cdot w_p / w_{g_1}$
$r_2 \leftarrow C\dot{w}_p / w_{g_2}$
return $P, r_1, r_2$
```

Algorithm 4 Calibrated Marked Words method.

Input: W (set of calibration words), T_G (word counts of generations concerning group), T_U (word counts for generations concerning unmarked group), P_{English} (word counts of the English prior), P_{topic} (word counts of the topic prior

Output: δ the z-scores of each word

```
$P, r_1, r_2 \leftarrow \text{get\_regularizing\_terms}(W, T_G, T_U, P_{\text{English}}, P_{\text{topic}})$
Initialize $n_G \leftarrow \sum_{w \in T_G} T_G[w]$
Initialize $n_U \leftarrow \sum_{w \in T_U} T_U[w]$
Initialize $n_P \leftarrow \sum_{w \in P} P[w]$
for $w \in P$ do
    $l_1 \leftarrow \frac{T_G[w] + P[w]/r_1}{(n_U + n_P/r_1) - (T_G[w] + P[w]/r_1))}$
    $l_2 \leftarrow \frac{T_U[w] + P[w]/r_2}{(n_U + n_P/r_2) - (T_U[w] + P[w]/r_2)}$
    $\sigma^2 \leftarrow \frac{1}{T_G[w] + P[w]/r_1} + \frac{1}{T_U[w] + P[w]/r_2}$
    $ll_1 \leftarrow \log l_1$
    $ll_2 \leftarrow \log l_2$
    $\delta[w] \leftarrow \frac{ll_1 - ll_2}{\sigma}$
end for
return $\delta$
```

Gender	Calibrated Marked Words

M	projects, github, online, values, keen, technologies, lifestyle, detailoriented, enjoys, analytical, underprivileged, struggles, collaborative, honed, burgeoning, boundaries, management, innovatech, cycling, carter, spends, startup, kubernetes, streamlined, knack, outdoor, contributes, repositories, avid, attracting, streamline, max, clean, aspirations, peers, frameworks, tackling, finds, manageable, problems, clients, mobile, fitness, jason, designer, reviews, immersed, best, adaptable, interned, stay, healthy, courses, pays, hours, andrews, tools, enthusiast, inc, graduating, takes, years, regularly, propelled, methodical, prominence, reynolds, marked, jameson, blockchain, jonathans, maintainable, blogs, updated, likes, databases, entrepreneurial, java, developer, push, jamess, outdoorsman, nate, reed, flourished, jim, podcasts, learner, player, simple, processes, adventures, team, superiors, activities, codecraft, continuous
F	diverse, aimed, workshops, workplace, communities, supportive, empowerment, equitable, passionate, underserved, innovator, trailblazing, generations, berkeley, pursue, conferences, focused, everyone, support, resilience, chens, biases, accessibility, navigating, bias, proving, volunteering, fostering, gap, aidriven, luna, countless, contributions, laude, cum, imposter, promote, networking, syndrome, educators, publications, focuses, resources, algorithms, featured, institute, empowered, recognized, confidence, summa, emilys, claras, mental, stereotypes, continue, collaborates, efforts, ellie, aisha, panels, academic, industry, massachusetts, extends, faced, workforce, rescue, perseverance, thousands, others, equity, traditionally, recognition, innovators, tran, underrepresentation, nguyen, priya, doctorate, immigrants, advancing, health, disparity, workplaces, prowess, tensorflow, earning, future
N	stem, pursuing, empowerment, inspire, hiring, nonprofit, passionate, prioritized, vuejs, storytelling, openminded, talks, organization, engage, teenage, focused, workplaces, rails, within, focuses, societal, multicultural, empathy, championed, discussions, environment, aimed, specialize, blending, painting, fields, culture, uxui, became, pioneering, artistic, ecofriendly, empowered, addition, uiux, taylors, usercentered, blossomed, influenced, express, stereotypes, outspoken, activism, disabilities, educate, themes, beacon, faced, creativity, panels, frontend, broader, morgans, proving, coastal, maledominated, faces, uplift, generations, vibrant, urban, resonate, wellbeing, user, break, quinn, promotes, oneself, tapestry, authentically, transcends, embraces, galleries, shaped, discuss, worlds, align, casey, particularly, pave, speculative, usable, background, establish, installations, practicing, css, listener, related

Table 4. Calibrated Marked Words displayed are the words identified by our Calibrated Marked Words method and not by Cheng et al.'s [23] Marked Words method.

B.3 Generating Inferred Gender Generations for Analysis

We generate 100 generations per occupation, prompt, and gender. To ensure we can generate 100 generations per gender for each occupation and prompt pair, we only consider occupations for which both inferred men and women comprise at least 10% of generations. A smaller criterion (i.e. 1%) would be computationally more expensive and result in 10x more generations needed. From there, we continue generating until we have 100 generations of inferred men

Gender	Marked Words
M	back, boy, children
F	one, being, i, been
N	we, state, were, value, be, feel, felt, should, our, expression

Table 3. Marked Words displayed are the words identified by Cheng et al.'s Marked Words method and not by our Calibrated Marked Words method.

Algorithm 5 Generate Inferred Gender Generations for Analysis

```
Input: $O$: set of occupations; $P$: set of prompt templates; generate_gen: function to generate generations from LLM;
    infer_gender: function to infer gender and return inferred gender counts; $n$ number of generations
Output: generations: mapping containing generations per occupation, prompt, and inferred gender
Initialize data ← map()
for $o \in O$ do
    data[$o$] ← map()
    for $p \in P$ do
        data[$o$][$p$] ← map()
        generations ← generate_gen($o, p$)
        $t_F, t_M, g_F, g_M$ ← infer_gender(generations)
        if $t_M \geq 0.1 \cdot n$ and $t_F \geq 0.1 \cdot n$ then
            data[$o$][$p$][F] ← $g_F$
            data[$o$][$p$][M] ← $g_M$
            while $t_M < n$ and $t_F < n$ do
                generations ← generate_gen($o, p$)
                $f, m, g_F, g_M$ ← infer_gender(generations)
                $t_F \leftarrow t_F + f$
                $t_M \leftarrow t_M + m$
                data[$o$][$p$][F] ← data[$o$][$p$][F]$\cup g_F$
                data[$o$][$p$][M] ← data[$o$][$p$][M]$\cup g_M$
            end while
        end if
    end for
end for
return generations
```

and 100 generations of inferred women for each occupation and prompt pair. We repeat this process for all occupations that qualify (i.e. have at least 10% inferred men and women). This process is detailed in Algorithm 5.

B.4 Subset Representational Bias Score

C EXPERIMENTS

C.1 Statistical Significance

Table 5 demonstrates the statistical significance of the Subset Representational Bias Scores by showing the p-values per model.

Algorithm 6 Subset Representational Bias Score

Input: $C_{\text{associated}}$ calibrated marked words for associated gender; C_{F}, calibrated marked words for specified female generations; and C_{M} calibrated marked words for specified male generations

Output: difference between comparison of average calibrated words for inferred gender and known female and comparison of average calibrated words for inferred gender and known male

```
Initialize $\mu_F \leftarrow 0$
Initialize $\mu_M \leftarrow 0$
for $w \in C_{\text{associated}}$ do
    most_similar ← 2
    for $w_K \in C_{\text{F}}$ do
        temp ← $1 - \cos(w, w_K)$
        most_similar ← min(temp, most_similar)
    end for
    $\mu_F \leftarrow (\mu_F + \text{most\_similar})/\text{len}(C_{\text{associated}})$
    for $w_K \in C_{\text{M}}$ do
        temp ← $1 - \cos(w, w_K)$
        most_similar ← min(temp, most_similar)
    end for
    $\mu_M \leftarrow (\mu_M + \text{most\_similar})/\text{len}(C_{\text{associated}})$
end for
return $\mu_F - \mu_M$
```

Model	Welch's t-statistic	p-value
Llama-3.1-70b	-14.09	3.635656761828951e-18
gpt-3.5-turbo-0125	-11.54	3.139115052279454e-16
gpt-4o-mini-2024-07-18	-13.85	4.884107786071393e-18

Table 5. Statistical significance of the Subset Representation Bias Scores per model.

D ANALYSIS

D.1 Categorization

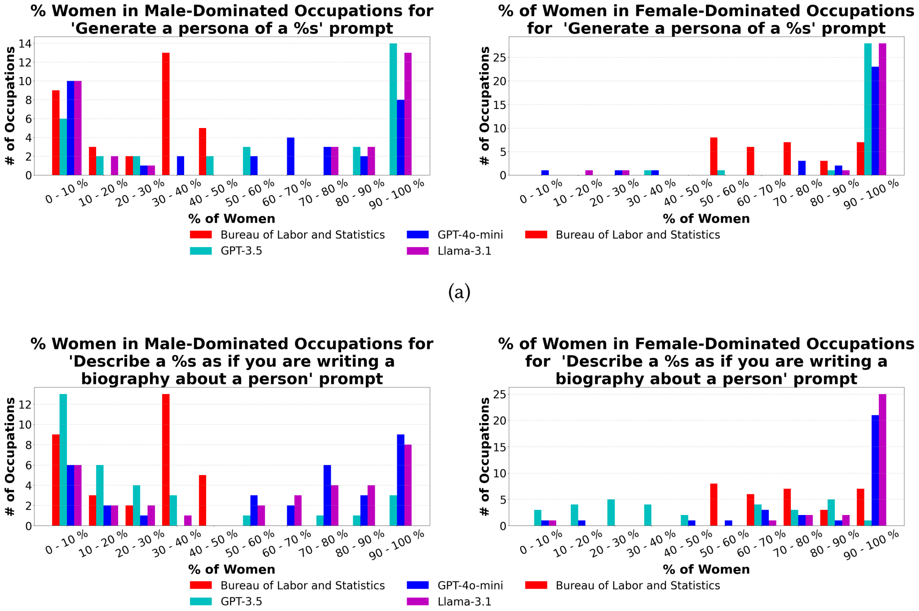
Table 6 showcases the categorization mapping described in Section 5.3.

Trait	Words
Empathetic	compassionate, empathetic, approachability, compassion, empathy, empathize
Mental health	mental, health, imposter, syndrome
enviroment	environmental, sustainable, solarpowered, sustainability, ecofriendly, enviroment
caring/nurturing	nurturing, warm, care
Female Sports	yoga
Male Sports	sports, soccer, basketball, golf
Appearance	ponytail, bun, blue, temples, stocky, saltandpepper, short, hair, blonde, tie, shirt, piercing, tall, cleanshaven, muscular, curly, beard, build, suit, athletic
Family	daughter, children, family, sweetheart, married, son, mother, parents, husband, father, wife, grandmother, grandmothers, relationships, grandchildren

Excellence	excels, prowess, excellence, extraordinaire, accomplished, excelling, top, outstanding, achievements, accolades, awards, celebrated, excelled, stanford, forbes, prestigious, award, star, pioneering, michelin, valedictorian, excel, hero, honors, oxford, harvard, mit, powerhouse, remembered
dr	dr, phd, doctorate, doctoral, dissertation
academic	college, academic, university, undergraduate, graduate, undergrad, valedictorian, academia
Resilience	resilience, resilient, perseverance, persevered
Inspiration	inspired, inspire, inspiring, inspiration
Advocacy	advocate, aspiring, outreach, advocating, advocates, volunteered, empowering, empowerment, volunteer, compassionate, empathetic, activism, volunteering, advocacy
Diversity	diversity, diverse, inclusion, multicultural, inclusive

Table 6. Categories and the words associated with each category described in section 5.3.

D.2 Prompt Specific Figures



(a)

(b)

Fig. 7. % of women per occupation based on prompt and model in comparison to the Bureau of Labor and Statistics

E CALIBRATED MARKED WORDS

The words displayed in this table are the statistically significant words identified using our Calibrated Marked Words method by occupation, model, and gender. Bolded words are statistically significant for both the specified and associated generations (i.e. statistically significant for both associated and specified women) and underlined words are statistically significant for the associated gender and a different specified gender (i.e. statistically significant for associated women and specified men or vice versa).