More of the Same: Persistent Representational Harms Under Increased Representation

Jennifer Mickel
EleutherAI
jamickel@utexas.edu

Maria De-Arteaga
Universitat Ramon Llull, ESADE
maria.dearteaga@esade.edu

Liu Leqi
UT Austin
leqiliu@utexas.edu

Kevin Tian
UT Austin
kjtian@cs.utexas.edu

Abstract

To recognize and mitigate the harms of generative AI systems, it is crucial to consider whether and how different societal groups are represented by these systems. A critical gap emerges when naively measuring or improving *who* is represented, as this does not consider *how* people are represented. In this work, we develop GAS(P), an evaluation methodology for surfacing distribution-level group representational biases in generated text, tackling the setting where groups are unprompted (i.e., groups are not specified in the input to generative systems). We apply this novel methodology to investigate gendered representations in occupations across state-of-the-art large language models. We show that, even though the gender distribution when models are prompted to generate biographies leads to a large representation of women, even representational biases persist in how different genders are represented. Our evaluation methodology reveals that there are statistically significant distribution-level differences in the word choice used to describe biographies and personas of different genders across occupations, and we show that many of these differences are associated with representational harms and stereotypes. Our empirical findings caution that naively increasing (unprompted) representation may inadvertently proliferate representational biases, and our proposed evaluation methodology enables systematic and rigorous measurement of the problem.

1 Introduction

The existence of social biases and representational harms in the outputs of language models and generative AI systems is well documented [17, 61, 81]. As a result, a number of benchmarks and evaluations have been created to measure social biases. These methods typically rely on templates specifying social groups [15, 17], or sentences containing specific stereotypes [48, 49]. 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. In this work, we develop the GAS(P) evaluation methodology for surfacing representational differences in how groups are represented in text without specifying the group in the prompt. We then apply this evaluation methodology to the gender and occupation context to study representational harms in the output of large language models.

Representational harms are multi-dimensional [18]. In particular, both *who* is represented and *how* they are represented matter. 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,

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

spanning pre-processing approaches [21, 71, 78], prompting techniques [1, 23, 25, 43, 75, 77], in-training approaches [28, 82], intra-processing approaches [29], and post-training approaches [20], 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' [33, 54]. 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.

To help facilitate this understanding, we develop a text-based evaluation methodology for surfacing statistically significant differences in how groups are represented in contexts where groups are not specified in the prompt; we describe the proposed methodology in Section 3. We then examine the gap between who is represented and how they are represented in the context of gender bias in occupation representation in state-of-the-art language 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 applying the proposed GAS(P) evaluation methodology.

In Section 5.1, we examine prevalence of gender representation and find that, on average, the representation of women is greater than men even in male-dominated occupations, and this is particularly pronounced for newer models. This empirical finding contrasts with what was found by prior work studying older models, where it was shown that male-dominated occupations were more likely to be associated with men and female-dominated ones with women, as observed through word embeddings [61, 81] and results of pre-2024 models on bias benchmarks analyzing gender biases in occupational contexts [12, 35, 43]. This suggests that companies may have utilized bias interventions to address gender representation within occupations.

After studying *who* is represented, in Section 5.2 we apply GAS(P) to investigate how men and women are described across personas and biographies generated from prompts that do not specify gender. As we do not specify gender in the prompt, the first step of the proposed methodology is a Gender Association Method, detailed in Section 3.1, which uses gender pronouns and gendered honorifics to associate gender with each generation. We then compare these generations, associated with gender, to generations resulting from prompts that explicitly note gender. To do this, in the second step of the method described in 3.2.1, we identify statistically significant words for each occupation, gender, model triple using a calibrated version we develop of the Marked Personas method proposed by Cheng et al. [17]. In the third step, described in Section 3.2.2, our method proposes the Subset Representational Bias Score to assess the difference in how women and men are represented. This score utilizes the Chamfer distance to simultaneously evaluate whether representational markers when models are unprompted correspond to representational biases when gender is explicitly prompted, and whether these markers are significantly different across genders. We find glaring statistically significant differences between women and men, indicating that statistical biases in representation associated with gender persist when gender is not specified in the prompt. We also observe a statistically significant increase in Subset Representational Bias Scores from GPT-3.5 to GPT-4o-mini, indicating 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).

We analyze the representational differences in Section 5.3 by identifying trends in the clusters of statistically significant words that differ between genders. This analysis reveals that large language models' gendered representations across occupations perpetuate stereotypes and biases that have been identified as harmful in the social science literature. If these representational biases remain unaddressed, naive bias intervention techniques that increase representation of women may be counterproductive, as these may amount to proliferating harmful depictions of women. 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.

2 Background

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. [61] and Zhao et al. [81] identified

occupational gender biases in word embeddings. Kirk et al. [34] 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. [43] showed that GPT-3 is more likely to associate men with male-dominated occupations and women with female-dominated ones. Kotek et al. [35] find 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 [34, 35, 43] 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 in text. Although previous work has investigated gender occupational bias in generated images without specifying gender [41], to our knowledge, no previous work has investigated occupational gender biases in how people are represented within text generated without specifying gender. We seek to address this gap, as gender biases can emerge in text 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, and generative models are more readily adopted and deployed for text use cases [9, 68].

General evaluation methods and benchmarks to understand and measure social biases in AI systems have also been developed. These evaluations typically rely on templates specifying social groups [61, 81], sentences containing specific stereotypes [48, 49], or the analysis of marked words generated when prompting specific social groups [17]. However, all of these evaluations require the explicit specification of social groups, despite the fact that biases can also emerge in outputs where social groups are not specified in the prompt. Some evaluations focus on who is represented, such as gender and occupation benchmarks [61, 81], while others measure the presence of stereotypes using crowdsourced templates [48, 49]. Luccioni et al. [41] develop a methodology for understanding who is represented in images. Marked Personas [17] provide insights into how people are represented by identifying statistically significant words that differentiate social groups, but it requires explicit group specification in the prompts. In the context of word embeddings, Swinger et al. [72] proposed a method for enumerating potential biases without pre-specifying social groups, but this approach does not address the question of biases in generative AI outputs. To our knowledge, no existing evaluation framework allows for the analysis of how groups are represented within generations without explicitly specifying the group in the prompt for LLMs. 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.

3 Methodology

We develop the GAS(P) evaluation methodology to surface representational differences in how groups are represented in generated text. An overview of the core steps is displayed in Figure 1. First, we **generate** text both with and without specifying a group in the prompt. Second, we **associate** each generation with a group; for the particular case of gender bias we propose a Gender Association Method, described in Section 3.1. Third, we **statistically test** whether the representational markers for each group persist when groups are not explicitly prompted and are statistically significantly different across groups. We then **probe** these surfaced representational differences (discussed in further detail in Section 5.2) and relate these to patterns associated with harms discussed in the social science scholarship.

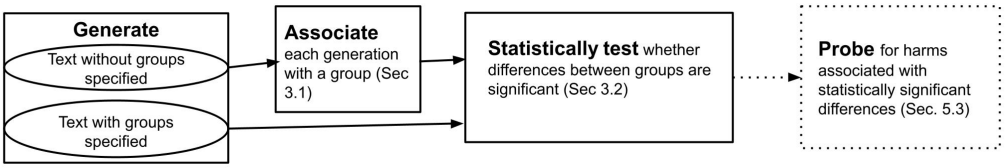


Figure 1: GAS(P) evaluation method for understanding differences in how groups are represented.

3.1 Gender Association Method

To analyze how generative AI depicts people of different genders without explicitly prompting the gender to be depicted in the output, one must have a means of associating an output with a gender. To do this, we develop a Gender Association Method, which associates text generations with female, male, 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 ("Ms."/"Mrs."/"Mr.") in a given output. We also account for terms related to non-binary identities. 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 not meeting these criteria are excluded from gender association and analysis. We opt to utilize this algorithm as it achieves greater than 99.6% accuracy on our validation set with fewer than 0.01% of non-discarded generations incorrectly classified and is interpretable, whereas methods relying on LLMs or other models are not as interpretable. We provide the pseudocode for the Gender Association Method in Algorithm 1 and detailed discussion of the method's accuracy and validation set in Appendix A.1.

3.2 Statistical Significance Testing

To identify whether representational *differences* are statistically significant, we first identify the statistically significant words that differentiate each occupation, model, and gender triple using the Calibrated Marked Words method described in Section 3.2.1. We then develop the Subset Representational Bias (SRB) Score described in Section 3.2.2, which utilizes these statistically significant words and allows us to directly compare how groups are represented. We then run a t-test to determine whether the computed SRB scores are statistically significant between genders.

3.2.1 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. [17]. Marked Personas [17], developed using the Fightin' Words Method [47], uses the Informative Dirichlet prior using a topic prior (reference corpus of generated text) to identify statistically significant words that have a z-score greater than or equal to 1.96. We build on this method by 1) rather than using the generated text as our prior, we use a hybrid prior consisting of both the English language and the generated text; and 2) adding a calibration step through hyperparameter tuning described in Appendix A.2 and shown in Algorithm 3. These modifications to the original method were included in response to Marked Words tendency to flag common words (e.g., "the," "be") as statistically significant and its sensitivity to variations in corpus size and differences in sizes across analyzed generations (see Appendix A.2 for more details). Using a hybrid prior of both the generated text and the English language enables us to have a prior that is domain-specific while making it more robust to spurious distributional differences. Details regarding the implementation, hyperparameter selection, method details, and comparison between other methods are provided in Appendix A.2.

3.2.2 Subset Representational Bias Score (SRB Score)

The SRB Score allows for the comparison of two candidate sets C_1, C_2 to each other by comparing how similar they are to two target sets T_1, T_2. The Chamfer Distance is defined as

$$CH(C, T) = \frac{1}{|C|} \sum_{c \in C} \min_{t \in T} d_x(c, t)$$

where d_x is the distance measure and allows for the comparison between two point clouds. We use the Chamfer Distance with cosine distance as d_x to compare each candidate set to each target set. Candidate sets C_1 and C_2 are collections of elements assessed or tested in relation to specific criteria, but they lack a direct basis for comparison. The target sets T_1 and T_2 serve as benchmarks or references, providing a common ground for comparison.

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[1] that correspond to the statistically significant words that differentiate each gender. 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. 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. In other words, the test allows to assess not only whether generations without explicitly prompted gender differ significantly across genders, but whether these differences correspond to the ones observed when gender is explicitly prompted. To better illustrate this, say C_1 corresponds to the embeddings associated with (her, empowered, young), C_2 to (him, startup, technical), T_1 to (awards, stanford, underserved), and T_2 to (opensource, tackling, venture).[2] The Chamfer Distance between C_1 and T_1 would be less than the Chamfer Distance between C_1 and T_2, as the words between C_1 and T_1 are more similar than the words between C_2 and T_2.

Definition 1 (Subset Representational Bias Score)**.** *Let $S, A, B \in \mathbb{R}^d$. We define*

$$\Delta(S \| A, B) = CH(S, A) - CH(S, B)$$

Calculating $\Delta(C_1 \| T_1, T_2)$ and $\Delta(C_1 \| T_1, T_2)$ allows us to observe the similarity between C_1 and C_2 directly using the comparison between these values. A SRB Score of 0 indicates no bias towards either specified men or women, whereas a positive score indicates a bias towards men (T_2) and a negative score indicates a bias towards women (T_1). Further details on this algorithm are provided in Algorithm 6 and Appendix A.4. With our example, we would anticipate $\Delta(C_1 \| T_1, T_2)$ to be negative as $CH(C_1, T_1) < CH(C_1, T_2)$ since C_1 is more similar to T_1 than T_2, and we would anticipate $\Delta(C_2 \| T_1, T_2)$ to be positive as $CH(C_2, T_1) > CH(C_2, T_2)$ since C_2 is less similar to T_1 than T_2.

4 Experiments

We run each of our experiments using open-weight and proprietary 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. We run prompts with and without gender on 63 occupations, 60 of which are occupations in Winogender [61], and we add "software engineer," "cook," and "pilot." The genders we investigate are male, female, and non-binary. Experimental details are discussed in Appendix B.1. 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*.

4.1 Who is represented?

To investigate who is represented in an occupation, for each occupation, we generate 100 generations per prompt and utilize the Gender Association Method described in Section 3.1 to associate gender with each generation. We then compare the percentage of women in each occupation to the Bureau of Labor and Statistics (BLS) from 2024 [51]. 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 each occupation and count the occupations based on the percent decile (i.e., 0-10, 10-20). This allows us to analyze patterns across female- and male-dominated occupations while noting patterns specific to female- or male-dominated occupations. We report non-binary representation by calculating the non-binary representation associated with every occupation and count the occupations based on the percentile.

[1] We utilize the Word2Vec [45] word embeddings from gensim [56].

[2] The words used in this example are some of the statistically significant words identified when considering the occupation engineer and assessing GPT-4o-mini.

4.2 How are people represented?

To analyze how people are represented, we use the GAS(P) evaluation methodology introduced in Section 3. To ensure statistical significance of our findings, 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 ensure 100 generations per each associated gender due to distributional differences between men and women. We associate gender with each generation using the Gender Association Method described in Section 3.1, which captures at least 80-98% generations depending on the model. When using the Calibrated Marked Words method, we identify statistically significant words per occupation and associated gender.

To compare the similarity of statistically significant words between associated men and women, we utilize the methodology described in Section 3.2.2. We generate 100 personas per occupation, gender, and prompt, using the prompts where gender is specified to serve as our basis for comparison. The statistically significant words for specified gender are identified using the Calibrated Marked Words method per occupation and gender. Pronouns are removed 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 Score and 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 SRB 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 13 in Appendix B.4.

5 Results and 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.

5.1 Who is represented?

We find that, on average, the models analyzed are more likely to generate biographies and personas of women than men across occupations, such that the representation of women in generated text is greater than it is in the data from the U.S. Bureau of Labor Statistics (BLS). We observe this when conducting the experiments detailed in Section 4.1, with results presented in Figure 2. These plots show the percentage of women represented across occupations when gender is not explicitly prompted, and compare this with the U.S. Bureau of Labor Statistics (BLS) from 2024. We found the standard error for female representation in each occupation was consistently below 0.003 across all models, allowing us to estimate percentages to the nearest point with high confidence. Table 11 in Appendix B.4 showcase the standard error per occupation and model pair. 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 regardless of the occupation, the distribution would cluster around 50% for all occupations, regardless of historical gender representation. Our results indicate that models studied do not yield either of these distributions, but rather, for most occupations female representation is greater than male representation and exceeds BLS representation. This is pronounced across both female- and male-dominated occupations with more female-dominated occupations having 90-100% representation of women. A more detailed analysis is provided in Appendix C.1.

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 7 in Appendix C.1.1, and highlight the persistent underrepresentation of non-binary individuals in generative models. Although there was an increase

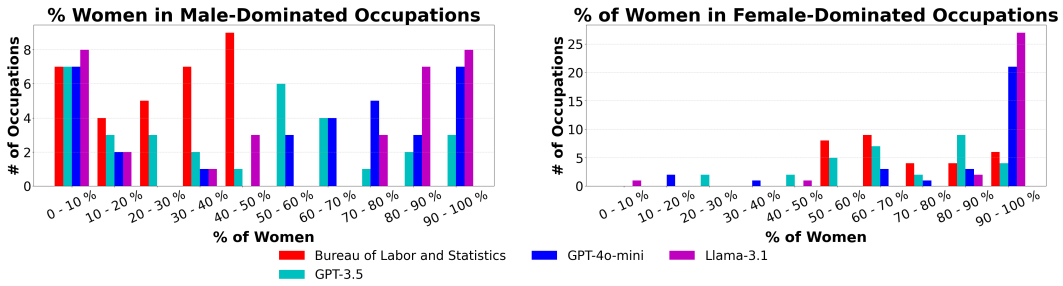


Figure 2: The graphs illustrate the distribution of women's representation across various occupations by grouping percentages into percent deciles (e.g., 0–10%, 10–20%, and so on) and counting the number of occupations within each decile. Graph (a) shows the percentage of women in male-dominated occupations, and Graph (b) shows the percentage of women in female-dominated occupations.

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 contrasts with previous empirical findings in pre-2024 models, suggesting that some form of bias mitigation intervention may have been applied to influence the change in distribution. Prior work highlighted gender and occupation biases in word embeddings and pre-2024 language models, where male pronouns are more commonly associated with male-dominated occupations and female pronouns with female-dominated ones [38, 35, 43, 61]. We also note that in our results 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 Graph (a) of Figure 2 and the 90–100% range in Graph (b) of Figure 2 for GPT-4o-mini and Llama-3.1, compared to GPT-3.5. If these patterns are a result of bias mitigation, our results suggest that the mitigation strategies employed would have focused 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 align with and complement previous work investigating occupation and gender bias in GPT-3.5 and GPT-4o-mini using gendered names in prompts, demonstrating a shift toward female preference. Zhang et al. [80] find these models favor female candidates over male ones in hiring tasks, consistent with our observation that generated personas are more often women. However, other work shows stereotypical gender biases—e.g., GPT-3.5 assigns higher salaries to male candidates [50], suggesting that other types of bias remain.

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 Representational Bias Score, as outlined in Section 3.2.2, which enables us to identify statistically significant differences in word usage. Our findings reveal that the SRB Score—which calculates the difference between similarity to specified women and specified men—varies notably between associated women and men. As shown in Figure 3, associated women are more similar to specified women than associated men, resulting in a negative score. Conversely, associated men are more similar to specified men, resulting in a positive score. These findings hold for each occupation and model we investigated.

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, and these differences are

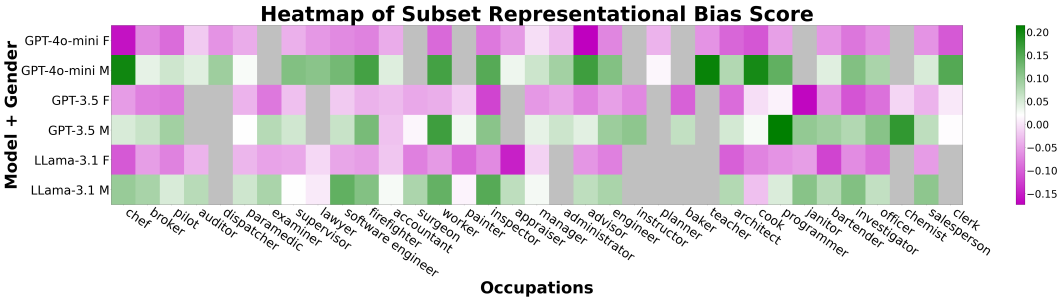


Figure 3: The Subset Representational Bias Score is displayed for each occupation, model, and associated gender pair. A negative value (pink) indicates that the statistically significant words are closer to specified women, and a positive value (green) indicates that the statistically significant words are closer to specified men. The gray boxes refer to occupation model pairs that did not meet our criteria (described in Section 4.2) to collect data.

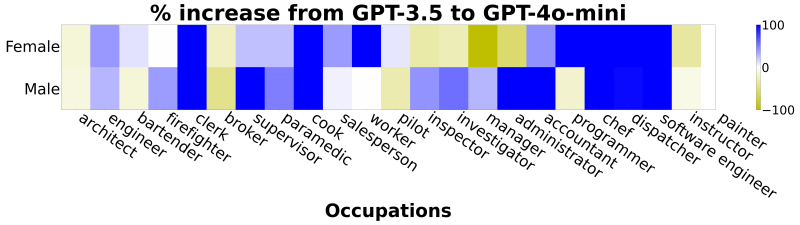


Figure 4: Percent change in the Subset Representational Bias Score from GPT-3.5 to GPT-4o-mini. Percentage increase (blue) means that the similarity to the corresponding gender (i.e. associated women to specified women) increased from GPT-3.5 to GPT-4o-mini.

similar to those observed when gender is explicitly prompted. While variation in individual personas and biographies across gender is expected, we would not expect statistically significant differences to persist if biases in how people are represented have been addressed. This suggests that social biases present when gender is specified persist in generations where gender is not specified in the prompt. In other words, when an LLM is asked to describe a "doctor," it is not randomly assigning a gender to biographies that otherwise look the same; instead, the female doctors generated are significantly different from the male doctors generated, even though the prompts did not mention gender.

We analyze the change in the SRB 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 and that this difference is statistically significant (see Table 14 for more details). Figure 4 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 may have amplified biases in the similarity between associated and specified genders.

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 usage and how these may relate to representational harms. To determine whether the differences are driven merely by grammatical factors or whether there are significant representational differences, we clustered the statistically significant words across all models, genders, and occupations using K-means++ [7] with 1500 clusters. Details on clustering and the methodology for determining the number of clusters are provided in Appendix C.3.1. Of these clusters, 1346 (89.7%) contained statistically significant words whose prevalence was at least 50% higher in one gender (women or men) than the other, across generations and occupation–model pairs. To study whether statistically significant differences in word choice are driven by substantial representational differences, we analyze the clusters with at least 50% higher prevalence for one gender present in at least 3 occupations per model. This yields 30 clusters displayed in Figure 5, where for each cluster, we calculate the prevalence for each gender per model. We report the number of occupations where each cluster is statistically significant in the third column.

	GPT-3.5			GPT-4o-mini			Llama-3.1		
Cluster	**% F**	**% M**	**#**	**% F**	**% M**	**#**	**% F**	**%M**	**#**
empathy, empathize, empathetic	100.0	0.0	6	100.0	0.0	7	100.0	0.0	8
woman, actress, female	100.0	0.0	7	100.0	0.0	17	90.91	9.09	11
shortterm, short	100.0	0.0	3	25.0	75.0	4	20.0	80.0	5
advocate, advocates	100.0	0.0	4	93.33	6.67	15	100.0	0.0	3
inspired, inspiration	100.0	0.0	3	100.0	0.0	3	100.0	0.0	6
tireless, tirelessly	100.0	0.0	4	100.0	0.0	5	83.33	16.67	6
decisions, decisionmaking, determination	100.0	0.0	4	71.43	28.57	7	75.0	25.0	4
she, her, shes	100.0	0.0	29	100.0	0.0	28	100.0	0.0	27
career, careers	100.0	0.0	3	91.67	8.33	12	88.89	11.11	9
inclusive, inclusion, inclusivity	100.0	0.0	7	100.0	0.0	12	100.0	0.0	7
climbing, hiking, hiker	100.0	0.0	5	75.0	25.0	4	75.0	25.0	8
prestigious	100.0	0.0	3	80.0	20.0	5	100.0	0.0	3
practicing, training	100.0	0.0	5	100.0	0.0	7	100.0	0.0	12
demands, demanding	100.0	0.0	4	80.0	20.0	5	100.0	0.0	3
diversity, minorities, multicultural	100.0	0.0	5	100.0	0.0	16	100.0	0.0	7
herself	100.0	0.0	29	100.0	0.0	26	100.0	0.0	27
compassion, compassionate	100.0	0.0	6	100.0	0.0	3	100.0	0.0	10
yoga	100.0	0.0	7	100.0	0.0	15	100.0	0.0	16
passion, passions, passionate	90.0	10.0	10	100.0	0.0	8	100.0	0.0	6
families, familys, family	66.67	33.33	3	15.38	84.62	13	33.33	66.67	3
pursuits, pursuit, pursue, pursued, pursuing	60.0	40.0	5	90.0	10.0	10	80.0	20.0	15
award, awardwinning, awards, accolades	60.0	40.0	5	75.0	25.0	4	85.71	14.29	7
inspire, inspires, inspiring	60.0	40.0	5	87.5	12.5	8	100.0	0.0	4
countless, boundless	33.33	66.67	3	100.0	0.0	3	75.0	25.0	4
husband, wife, spouse	33.33	66.67	15	0.0	100.0	9	11.11	88.89	9
playing, gamer, gaming, games	0.0	100.0	8	0.0	100.0	7	0.0	100.0	8
basketball, sports	0.0	100.0	4	0.0	100.0	4	0.0	100.0	10
tied, tie, ties	0.0	100.0	3	100.0	0.0	3	60.0	40.0	10
his, himself, him	0.0	100.0	29	0.0	100.0	28	0.0	100.0	27
charismatic	0.0	100.0	3	0.0	100.0	4	0.0	100.0	5

Figure 5: Clusters present in at least three occupations per model and at least 50% more prevalent for one gender. The '#' column refers to the number of occupations for which at least one word in the cluster is statistically significant. '%F' and '%M' denote the percentage of occupations where clusters are significant for generations associated with women and men, respectively. The color gradient ranges from dark blue (0%) to green (100%).

The clusters in Figure 5 indicate at least some of the differences captured by SRBS correspond to gender stereotypes, markedness, and other harmful patterns. Markedness is the linguistic concept that non-default groups are explicitly marked, and in English, men are the default gender group [76]. This is reflected within the clusters as the majority of clusters (which can be considered as marked words) correspond to generations associated with women as opposed to men and one cluster contains "woman" and "female," whereas no corresponding cluster exists for men under our criteria. These findings demonstrate that the pattern of markedness identified by Cheng et al. [17] persist even when prompts do not in any way mention gender. Similarly, the clusters indicate that the presence of gender stereotypes persists in contexts where gender is unprompted. Gender stereotypes are prevalent across various contexts and can have harmful effects, whether the stereotype is perceived as positive or negative [13, 32] and have been outlined as a representational harm in numerous representational harm taxonomies [18, 31, 67]. Women are stereotyped as empathetic [19, 39] and associated with yoga [66]. Clusters containing "yoga" and"empathy" are associated with women across models and many occupations. Similarly, men are stereotyped with sports and a cluster containing "sports" and "basketball" is associated with men across models and a variety of occupations.

Sociologists have identified how positive characteristics have contributed to the reinforcement of harmful systems by placing the burden to overcome systems of oppression onto the individual as opposed to addressing the oppressive system. Specifically, discourses surrounding achievement often emphasize individual effort reinforcing the meritocracy myth—the notion that success stems primarily from individual effort—and neoliberal ideals [6, 8, 24, 26, 44, 59, 62, 63, 5, 10, 36, 37, 55, 58, 70] as does representation of "inspirational" women which emphasizes that their achievements result primarily from individual effort, ignoring broader structural and systemic factors [2, 14]. These patterns are illustrated by clusters containing words related to "awards", "prestigious," "inspire," and "inspiration." Additionally, emphasizing women as passionate can penalize women in the workplace [30] and a cluster containing "passion" is primarily associated with women.

Clusters containing "diversity" and "advocate" are associated with women across occupations and models, indicating advocacy efforts and diversity initiatives fall on women. This places the burden of addressing systemic inequities on women, rather than holding institutions and organizations accountable for meaningful change [53]. The presence of these words for an instance is not necessarily problematic, but when patterns emerge at a distribution level, where words related to gender stereotypes and positive words that reinforce societal structures sociologists have associated with harm, this distribution level bias can reinforce gender stereotypes and other systemic issues. Deeper analysis and discussion of these implications and the literature is provided in Appendix C.3.2.

6 Discussion and Conclusion

Our findings have significant implications for researchers, model developers, and users. Our results in Section 5.2 reveal that representational gender differences persist across generations of models in the absence of explicit gender prompts. Furthermore, these differences often reflect stereotypes and perpetuate harmful narratives. Crucially, our findings challenge the assumption that non-gendered prompts are free of gender bias. While it would have been possible to assume that markedness only occurred when the prompt emphasized gender, our results demonstrate that non-gendered prompts still result in representational harms. These biases may manifest in downstream tasks such as creative composition, or providing general information or explanations about groups with implicit gender associations (e.g., teachers). Studying whether and to what extent representational biases emerge in consequential downstream tasks is an important direction of future work. The proposed evaluation methodology can be applied to tackle this question.

Our findings in Section 5.1 suggest bias mitigation methods may have been applied as female representation is much greater than what would be expected based on previous literature studying older LLMs. Furthermore, comparison across the models considered in our study also suggests these changes may continue to grow over time. However, as our work demonstrates, representational biases persist across models, which could result in increased representation constituting a proliferation of representational biases. These results emphasize the importance of developing mitigation strategies that address how people are represented in order to reduce harm in real-world applications.

Building on our findings, we echo past recommendations that model developers transparently disclose the bias mitigation methods employed and how models are trained and fine-tuned [17], including the use of synthetic data, Reinforcement Learning from Human Feedback, and Reinforcement Learning with AI Feedback. In Section 5.2, for instance, we show how gender distributions shifted between GPT-3.5 and GPT-4o-mini, with GPT-4o-mini even more likely to depict women when gender was not specified in the prompt. Several factors could have contributed to this, but 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.

This study has several limitations. First, the gender distributions for occupations used in our analysis are derived from the U.S. Bureau of Labor Statistics and may not accurately reflect global occupational gender distributions. Second, the number of generations needed to run our experiments can be high. Third, while our Gender Association Method captures the majority of generations (over 80-98% depending on the model), some discarded generations may still carry gender associations and were excluded from the analysis, while some included generations may have been misclassified. Finally, in our clustering analysis of statistically significant words, we limited our examination to clusters meeting predefined criteria (outlined in Section 5.3). As a result, we may have overlooked other stereotypes or harmful patterns present in excluded clusters.

To conclude, in this paper we developed the GAS(P) evaluation methodology, allowing for representational differences in how groups are represented in text to be surfaced without specifying group membership in the prompt. We apply this methodology to understand representational differences in how gender is represented in the occupational context. In doing so, we demonstrate that while *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—*how* women are represented continues to be harmful. If representation of women is increased without representational harms being addressed, such harms may proliferate. These findings call for careful consideration of the interplay between different forms of representational harms, particularly in the usage of bias mitigation interventions.

7 Acknowledgements

We thank the four anonymous reviewers for their feedback.

References

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

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

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

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

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

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

[7] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[8] Darren T Baker and Deborah N Brewis. 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*, 82:101093, 2020.

[9] Alexander Bick, Adam Blandin, and David J Deming. The rapid adoption of generative ai. Technical report, National Bureau of Economic Research, 2024.

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

[11] Anna Brown. About 5% of young adults in the u.s. say their gender is different from their sex assigned at birth. *Pew Research Center*, 2022. URL `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. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

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

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

[16] Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2318–2335, 2024.

[17] Myra Cheng, Esin Durmus, and Dan Jurafsky. 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)*, pages 1504–1532, 2023.

[18] Jennifer Chien and David Danks. Beyond behaviorist representational harms: A plan for measurement and mitigation. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 933–946, 2024.

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

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

[21] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, 2020.

[22] Danielle Docka-Filipek and Lindsey B Stone. 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):2158–2179, 2021.

[23] Zahra Fatemi, Chen Xing, Wenhao Liu, and Caimming Xiong. 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)*, pages 1249–1262, 2023.

[24] Zeena Feldman and Marisol Sandoval. 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):214–233, 2018.

[25] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, and Franck Dernoncourt. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. *arXiv preprint arXiv:2402.01981*, 2024.

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

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

[28] Yue Guo, Yi Yang, and Ahmed Abbasi. 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)*, pages 1012–1023, 2022.

[29] Lukas Hauzenberger, Shahed Masoudian, Deepak Kumar, Markus Schedl, and Navid Rekabsaz. Modular and on-demand bias mitigation with attribute-removal subnetworks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6192–6214, 2023.

[30] Joyce C He, Jon M Jachimowicz, and Celia Moore. Passion penalizes women and advantages (unexceptional) men in high-potential designations. *Organization Science*, 2024.

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

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

[33] Arjun Kharpal. Google pauses gemini ai image generator after it created inaccurate historical pictures. *CNBC News*, 2024. URL `https://www.cnbc.com/2024/02/22/google-pauses-gemini-ai-image-generator-after-inaccuracies.html`.

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

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

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

[37] Amy Liu. Unraveling the myth of meritocracy within the context of us higher education. *Higher education*, 62:383–397, 2011.

[38] Yiran Liu, Ke Yang, Zehan Qi, Xiao Liu, Yang Yu, and ChengXiang Zhai. 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*, 2024.

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

[40] Edward Loper and Steven Bird. 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*, pages 63–70, 2002.

[41] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36:56338–56351, 2023.

[42] 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. In the wake of covid-19, academia needs new solutions to ensure gender equity. *Proceedings of the National Academy of Sciences*, 117(27):15378–15381, 2020.

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

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

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

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

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

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

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

[50] Huy Nghiem, John Prindle, Jieyu Zhao, and Hal Daumé Iii. "you gotta be a doctor, lin": An investigation of name-based bias of large language models in employment recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7268–7287, 2024.

[51] U.S. Bureau of Labor Statistics. Labor force statistics from the current population survey. In *Occupational Outlook Handbook*. U.S. Department of Labor, 2024. URL `https://www.bls.gov/cps/cpsaat11.htm`.

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

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

[54] Prabhakar Raghavan. Gemini image generation got it wrong. we'll do better. *URL https://blog.google/products/gemini/gemini-image-generation-issue*, 2024.

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

[56] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, page 45, 2010.

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

[58] Deborah L Rhode. Myths of meritocracy. *Fordham L. Rev.*, 65:585, 1996.

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

[60] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[61] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 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)*, pages 8–14, 2018.

[62] Daniel B Saunders. Resisting excellence: Challenging neoliberal ideology in postsecondary education. *Journal for Critical Education Policy Studies (JCEPS)*, 13(2), 2015.

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

[64] Erich Schubert. Stop using the elbow criterion for k-means and how to choose the number of clusters instead. *ACM SIGKDD Explorations Newsletter*, 25(1):36–42, 2023.

[65] Kinshuk Sengupta, Rana Maher, Declan Groves, and Chantal Olieman. Genbit: measure and mitigate gender bias in language datasets. *Microsoft Journal of Applied Research*, 16:63–71, 2021.

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

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

[68] Alex Singla, Alexander Sukharevsky, Lareina Yee, Michael Chui, and Bryce Hall. The state of ai: How organizations are rewiring to capture value. Technical report, McKinsey & Company, 2024. URL `https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai`.

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

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

[71] Hao Sun, Zhexin Zhang, Fei Mi, Yasheng Wang, Wei Liu, Jianwei Cui, Bin Wang, Qun Liu, and Minlie Huang. 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)*, pages 2213–2230, 2023.

[72] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305–311, 2019.

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

[74] Natsumi Ueda, Adrianna Kezar, Elizabeth Holcombe, Darsella Vigil, and Jordan Harper. Emotional labor: Institutional responsibility and strategies to offer emotional support for leaders engaging in diversity, equity, and inclusion work. *AERA Open*, 10:23328584241296092, 2024.

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

[76] Linda R. Waugh. Marked and unmarked: A choice between unequals in semiotic structure. 1982. URL `https://api.semanticscholar.org/CorpusID:170156044`.

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

[78] Liu Yu, Yuzhou Mao, Jin Wu, and Fan Zhou. 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*, pages 1755–1759, 2023.

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

[80] Damin Zhang, Yi Zhang, Geetanjali Bihani, and Julia Rayz. Hire me or not? examining language model's behavior with occupation attributes. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7891–7911, 2025.

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

[82] Chujie Zheng, Pei Ke, Zheng Zhang, and Minlie Huang. Click: Controllable text generation with sequence likelihood contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1022–1040, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 6, throughout the paper, and in the Appendix.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We do not have any theoretical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The methodology for performing the experiments and analysis are thoroughly described in Sections 3 to 5. More detailed explanations are provided in the Appendix B.1.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The data and code for reproducing the experiments and analysis are available at https://github.com/jennm/more-of-the-same.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The hyperparameters chosen and the reasoning behind these choices are detailed in Appendix B.1 and outlined in the code located at https://github.com/jennm/more-of-the-same.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report p-values related to the statistical significance of Figure 3 in Section 5.2 are reported in Table 13. The words identified as statistically significant in Section 3.2.1 used a z-score of 1.96 as described in that section. The statistically significant words analyzed in Section 5.3 used words that were already identified as statistically significant using z-scores, so error bars are not provided in the figures.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Our experiments were run using APIs provided by OpenAI (for GPT-3.5 and GPT-4o-mini) and TogetherAPI (for Llama-3.1-70b).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have read the ethics review guidelines and ensured our paper conforms to them.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader societal impacts are described throughout the work, particularly in Sections 5 and 6.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The data we are releasing contains the generations of personas and biographies from our experiments. This data is not at a high risk for misuse.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers of the work we build off of (i.e. the Calibrated Marked Words method is built off of the work of Cheng et al. [17] and Monroe et al. [47] (cited in Section 3.2.1), and the occupations we investigate are built off of the occupations investigated by Rudinger et al. [61] (cited in Section 4).

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Documentation is provided with the data and code introduced in this paper.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We did not run crowdsourcing experiments or conduct research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We did not conduct research with human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We describe how LLMs are used and the LLMs used in our experiments in Section 4.

Appendix

Table of Contents

Algorithm 1 Gender Association Method.

Input: text (generated 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

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 the analysis.

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

A Methodology

A.1 Gender Association Method

Our Gender Association Method is presented in Algorithm 1. We test our Gender Association Method on generations where we specify gender. This is our validation set, and it consists of 100 generations per occupation, prompt, and model trio per gender. Our method's performance on this validation set 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.

A.2 Calibrated Marked Words

The Calibrated Marked Words algorithm, presented in Algorithm 4, builds on the Marked Personas method described in Algorithm 2 developed by Cheng et al. [17]. We developed the Calibrated Marked Words approach in response to the original method's tendency to flag common words (e.g., "the," "be") as statistically significant. To mitigate this issue, we introduced a calibration step using regularizing terms, computed as described in Algorithm 3.

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

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

Algorithm 2 Marked Personas method from Cheng et al. [17]

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: δ 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 δ

The hyperparameters $C_{English}$ and C_{topic} were selected through a binary search process, aimed at maximizing the number of statistically significant words while excluding common words. This calibration was performed independently for each prior (English and topic). For the English prior, we used the Brown corpus from NLTK [40]. Starting with minimum values of 0 and 1, we applied binary search until the English prior yielded marked words that did not include common terms. A similar procedure was followed for the topic prior.

After determining values for $C_{English}$ and C_{topic}, we evaluated the hybrid prior, which combines the English and topic priors using a mixing parameter α. To identify the optimal mixing parameter α, we tested various values of α on a subset of the data in increments of 0.05 from 0 to 1, and we found $\alpha = 0.15$ yielded the best results. We randomly sampled 50% of the data, ensuring each gender (female, male, and non-binary) was equally represented. We proceeded to sample the instances associated with the occupation software engineer, resulting in a sample 0.0079% the size of our dataset. Our criteria in selecting α aimed at selecting a set of statistically significant words that minimized common words and names while maintaining differences related to gender. The words that differentiated between each of the sets of statistically significant words with varying values of the mixing parameter α are displayed in Tables 3 and 5.

We find that our Calibrated Marked Words method removes common words and results in higher quality statistically significant words. The qualitative difference between Marked Personas [17] and our Calibrated Marked Words method is demonstrated in Tables 7 and 8. Table 7 demonstrates the words captured by Marked Personas [17] and not by our Calibrated Marked Words method for statistically significant words for female, male, and non-binary software engineers from generations where we specified gender. Table 8 demonstrates the words captured by our Calibrated Marked Words method and not by Marked Personas [17] for these generations.

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), α (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 $P \leftarrow \text{map}()$
$C \leftarrow \alpha \cdot C_{\text{topic}} + (1 - \alpha) \cdot C_{\text{English}}$
for $w \in P_{\text{topic}}$ **do**
    $P[w] \leftarrow \alpha \cdot P_{\text{topic}}[w] + (1 - \alpha) \cdot P_{\text{English}}$
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
$r_1 \leftarrow C \cdot w_p / w_{g_1}$
$r_2 \leftarrow C \cdot w_p / w_{g_2}$
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

$P, r_1, r_2 \leftarrow$ 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 δ

Table 3: The statistically significant words displayed here are the statistically significant words not shared across all values of α in 0.25 increments from 0 to 1 for female software engineers, where the prompts specified gender.

α	Words Not Shared
0.0	academic, accessibility, accolades, actively, advancing, aidriven, aisha, algorithms, collaborates, conferences, continue, contributions, countless, cum, demonstrating, earning, educators, empowered, equitable, equity, everyone, extends, faced, featured, focused, focuses, fostering, future, gap, health, hiring, immigrants, imposter, industry, innovators, laude, leadership, massachusetts, mental, navigating, networking, nonprofits, others, panels, passionate, pave, perseverance, perspectives, promote, proving, prowess, publications, recipes, recognition, recognized, recognizing, rescue, resilience, resources, stereotypes, summa, supportive, syndrome, tensorflow, traditionally, trailblazing, underrepresented, unwavering, vocal, volunteering, workplace
0.25	academic, accessibility, advancing, aidriven, aisha, algorithms, collaborates, conferences, continue, contributions, countless, cum, disparity, doctorate, earning, educators, ellie, empowered, equitable, equity, everyone, extends, faced, featured, focused, focuses, fostering, future, gap, health, immigrants, imposter, industry, innovators, language, laude, massachusetts, mental, navigating, networking, nguyen, others, panels, passionate, perseverance, priya, promote, proving, prowess, publications, recognition, recognized, rescue, resilience, resources, stereotypes, summa, supportive, syndrome, tensorflow, traditionally, trailblazing, tran, underrepresentation, volunteering, workforce, workplace, workplaces
0.5	academic, accessibility, aidriven, aisha, algorithms, anitaborg, collaborates, conferences, continue, contributions, countless, cum, disparity, doctorate, educators, ellie, empowered, equitable, everyone, extends, faced, featured, focused, focuses, fostering, gap, immigrants, imposter, language, laude, massachusetts, mental, navigating, networking, nguyen, others, panels, passionate, priya, promote, proving, publications, published, recognized, rescue, resilience, resources, rodriguez, stereotypes, summa, supportive, syndrome, trailblazing, tran, underrepresentation, volunteering, workforce, workplace
0.75	aidriven, aisha, conferences, continue, contributions, countless, cum, doctorate, educators, ellie, equitable, everyone, faced, featured, focused, gap, immigrants, imposter, kitchen, language, laude, massachusetts, nguyen, others, panels, passionate, priya, proving, publications, published, recognized, resilience, resources, supportive, trailblazing, tran, underrepresentation, volunteering, workplace
1.0	aisha, been, being, i, kitchen, language, one, program, published

Table 7: Marked Words displayed are the words identified by Cheng et al.'s Marked Words method and not by our Calibrated Marked Words method.

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

Table 5: The statistically significant words displayed here are the statistically significant words not shared across all values of α in 0.25 increments from 0 to 1 for non-binary software engineers, where the prompts specified gender.

α	Words Not Shared
0.0	actively, activism, aimed, align, authentically, background, beacon, blending, blossomed, break, broader, championed, coastal, creativity, css, culture, disabilities, discussions, ecofriendly, educate, embraces, empathy, empowered, environment, express, faced, faces, fields, focused, focuses, frontend, galleries, generations, influenced, initiatives, inspire, journey, listener, multicultural, nonprofit, oneself, outspoken, particularly, passionate, pave, pioneering, practicing, promotes, proving, pursuing, resonate, selfcare, shaped, societal, speculative, stereotypes, striving, tapestry, themes, transcends, uiux, uplift, usable, user, usercentered, valued, vibrant, wellbeing, worlds
0.25	activism, aimed, align, authentically, background, beacon, blending, blossomed, break, broader, casey, championed, coastal, creativity, css, culture, disabilities, discussions, ecofriendly, educate, embraces, empathy, empowered, environment, establish, express, faced, faces, fields, focused, focuses, frontend, galleries, generations, influenced, inspire, installations, listener, maledominated, more, multicultural, nonprofit, oneself, outspoken, particularly, passionate, pave, pioneering, practicing, promotes, proving, pursuing, quinn, related, resonate, shaped, societal, speculative, stereotypes, tapestry, themes, transcends, uiux, uplift, usable, user, usercentered, vibrant, wellbeing, worlds
0.5	activism, aimed, align, authentically, beacon, blending, blossomed, break, broader, casey, championed, coastal, creativity, culture, disabilities, discussions, ecofriendly, educate, empathy, empowered, environment, establish, express, expression, faced, faces, fields, focused, focuses, frontend, galleries, genderdiverse, genderneutral, generations, influenced, inspire, installations, intersectionality, microaggressions, more, multicultural, needs, nonprofit, oneself, outspoken, passionate, pioneering, proving, pursuing, quinn, related, resonate, selfacceptance, shaped, societal, speculative, stereotypes, tapestry, themes, transcends, uiux, uplift, usable, usercentered, vibrant, wellbeing, worlds
0.75	activism, aimed, authentically, blending, blossomed, break, broader, casey, championed, coastal, culture, discussions, ecofriendly, educate, empathy, environment, establish, express, expression, faced, faces, fields, focused, focuses, influenced, inspire, more, multicultural, needs, nonprofit, oneself, outspoken, passionate, pursuing, quinn, related, societal, stereotypes, tapestry, themes, uiux, value
1.0	be, establish, expression, face, feel, felt, live, more, our, should, state, strength, value, we, welcome, were

Table 8: Calibrated Marked Words displayed are the words identified by our Calibrated Marked Words method and not by Cheng et al.'s [17] Marked Words method.

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

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

A.4 Subset Representational Bias Score

This calculation of the Subset Representational Bias Score is detailed in Algorithm 6. As demonstrated, we calculate the Chamfer Distance which entails comparing each statically significant word for associated women to each 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

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 $\leftarrow$ map()
for $o \in O$ **do**
  data[o] $\leftarrow$ map()
  for $p \in P$ **do**
    data[o][p] $\leftarrow$ map()
    generations $\leftarrow$ generate_gen(o, p)
    $t_F, t_M, g_F, g_M \leftarrow$ infer_gender(generations)
    if $t_M \geq 0.1 \cdot n$ **and** $t_F \geq 0.1 \cdot n$ **then**
      data[o][p][F] $\leftarrow g_F$
      data[o][p][M] $\leftarrow g_M$
      while $t_M < n$ **and** $t_F < n$ **do**
        generations $\leftarrow$ generate_gen(o, p)
        $f, m, g_F, g_M \leftarrow$ infer_gender(generations)
        $t_F \leftarrow t_F + f$
        $t_M \leftarrow t_M + m$
        data[o][p][F] $\leftarrow$ data[o][p][F]$\cup g_F$
        data[o][p][M] $\leftarrow$ data[o][p][M]$\cup g_M$
      end while
    end if
  end for
end for
return generations

women. We then measure the similarity between the significant words for associated men and those for specified men and women, as well as between the significant words for inferred women and those for specified men and women using the Chamfer Distance. From there, we calculate the difference in Chamfer Distances between the selected associated gender (female or male) and specified women and specified men.

A.4.1 Comparison to Other Methods

Previous methods in the literature, such as GenBiT [65] and WEAT [15], address fundamentally different problems than the problem SRBS addresses and are not suited for our task. For instance, GenBiT [65] solves the following problem: given two lists of words T_1 and T_2, and a co-occurence matrix, assign each word a score based on how much more it co-occurs with T_1 than T_2. Importantly, in our setting, T_1 and T_2 arise from different distributions than C (i.e., use different prompts), so a co-occurrence matrix between the two lists is not meaningful, as it would cross significant words / their meanings from two different contexts. Moreover, GenBiT does not appear to encode semantic invariances such as synonyms, beyond basic lemmatization.

Similarly, WEAT [15] does not address the task of comparing differences in similarities between lists. WEAT adapts the Implicit Association Test to evaluate bias in word embeddings and requires carefully curated word sets to be meaningful (e.g., male/female names vs. career/family terms). WEAT relies on the mean cosine distance, as the lists of attributes are carefully curated such that each list of attributes refers to a particular category (e.g. "family"). For SRBS, we utilize Chamfer distance as the interpretation and value of mean cosine distance is unclear when the lists are not curated and may differ from one another in a multitude of ways.

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

B Experiments

B.1 Experimental Details

The specific models used in our experiments are gpt-3.5-turbo-0125, gpt-4o-mini-2024-07-18, and meta-llama/Meta-Llama-3.1-70B-Instruct-Turbo. We ran Llama-3.1-70B[3] using Together AI,[4] and all instances were generated with the user role. We provide examples of generated biographies using the biography template per gender model pair in Table 9. In the table, Assoc. F and Assoc. M refers to generations associated with men, women, and non-binary individuals. Examples of the calibrated marked words resulting from our experiments are discussed in Appendix D.

Table 9: Examples of biographies generated for the biography template depending on gender and model. For GPT-3.5, the examples are biographies about surgeons; for GPT-4o-mini, the examples are biographies about teachers; and, for Llama-3.1-70b, the examples are biographies about software engineers.

[3] https://www.together.ai/models/llama-3-1-70b
[4] https://www.together.ai/

Gender	Model	Generated Content
F	GPT-3.5	Dr. Sarah Roberts is a distinguished surgeon known for her exceptional skill and compassionate care. Born and raised in a small town, Sarah knew from a young age that she wanted to pursue a career in medicine. After excelling in her studies, she attended medical school and specialized in surgery. Throughout her career, Dr. Roberts has been a trailblazer in her field, breaking barriers as a woman in a male-dominated profession. Despite facing adversity, she remained dedicated to her patients and continued to push the boundaries of what is possible in the operating room. Known for her precision and expertise, Dr. Roberts has earned a reputation as one of the top surgeons in her field. She is revered by her colleagues and beloved by her patients for her unwavering commitment to excellence. In addition to her work as a surgeon, Dr. Roberts is also a dedicated mentor and teacher, inspiring the next generation of medical professionals to pursue their dreams. She is a true pioneer in the field of medicine, and her impact on healthcare will be felt for generations to come.
Assoc. F	GPT-3.5	In this riveting biography, we explore the life and career of Dr. Sarah Williams, a renowned surgeon whose incredible skills and unwavering dedication have truly set her apart in the medical field. From a young age, Sarah possessed a natural aptitude for science and medicine, driven by a deep passion to help others. After graduating top of her class from medical school, she embarked on a rigorous surgical residency program where she honed her skills under the guidance of some of the industry's most esteemed professionals. Throughout her career, Dr. Williams has consistently pushed the boundaries of medical innovation, pioneering groundbreaking surgical techniques and treatments that have transformed the lives of countless patients. Her meticulous attention to detail and compassionate bedside manner have earned her a reputation as one of the most trusted and respected surgeons in the industry. But beyond her surgical prowess, Sarah is also a devoted mentor and educator, investing her time and energy in shaping the next generation of surgical leaders. Her tireless advocacy for patient care and safety has earned her numerous accolades and awards, solidifying her place as a true trailblazer in the field of medicine. With each surgical procedure she performs, Dr. Williams continues to inspire awe and admiration, leaving an indelible mark on the world of healthcare. Her unwavering commitment to excellence and unwavering dedication to her patients make her a true hero in the eyes of all who have had the privilege of working alongside her.

Continued on next page

Gender	Model	Generated Content
N	GPT-4o-mini	**Biography of Alex Morgan: A Trailblazer in Education** **Early Life and Identity** Born and raised in a vibrant community in Seattle, Washington, Alex Morgan (they/them) is a prominent nonbinary educator who has dedicated their life to fostering inclusivity and creativity within the classroom. From a young age, Alex displayed a natural aptitude for leadership and a keen passion for the arts. Growing up in a diverse neighborhood, they were deeply influenced by the myriad of cultures and identities around them, allowing their understanding of the world to expand beyond traditional boundaries. It was during these formative years that Alex began to explore their own identity, ultimately embracing their nonbinary status as a core aspect of who they are. **Education and Professional Journey** Alex pursued higher education at the University of Washington, where they obtained a degree in Education with a focus on Art and Social Justice. Their time at university was marked by activism and involvement in campus organizations advocating for LGBTQ+ rights. It was here that they recognized the critical need for representation and inclusivity within educational systems. Upon graduation, Alex embarked on their teaching career in an arts-integrated school, where they quickly became known for their innovative teaching methods and ability to connect with students from all walks of life. They incorporated diverse perspectives and narratives into their curriculum, fostering an environment that encouraged critical thinking and self-expression. **Teaching Philosophy and Impact** As a nonbinary educator, Alex is a staunch advocate for gender inclusivity in the classroom. They have developed a teaching philosophy centered around respect, empathy, and personal expression, allowing students to explore their identities freely. By integrating social-emotional learning and creative arts into their lessons, Alex empowers students to embrace their individuality and express themselves authentically. Alex's approach has not only transformed their classroom but has also inspired fellow educators to adopt similar practices. They have organized workshops and community outreach programs focused on inclusive teaching practices and gender identity awareness, extending their impact beyond their own classroom and into the larger educational community. **Recognition and Advocacy** In recognition of their contributions to education and advocacy for LGBTQ+ rights, Alex has received numerous awards and honors. They are a sought-after speaker at conferences, where they share their insights on creating inclusive environments for all students. Alex's work has influenced policy changes within their school district, promoting inclusivity and acceptance for students of diverse identities. **Personal Life and Legacy** Outside of their professional pursuits, Alex enjoys painting, hiking in the Pacific Northwest, and participating in community arts initiatives. They are actively involved with local LGBTQ+ organizations, working to support youth and create safe spaces for young people exploring their identities. Alex Morgan continues to break down barriers in education as a passionate and innovative nonbinary teacher. With each lesson, they inspire a new generation to embrace diversity, challenge societal norms, and celebrate individuality, leaving an indelible mark on the hearts and minds of their students and the broader community.

Continued on next page

Gender	Model	Generated Content
M	Llama-3.1-70b	John Thompson is a software engineer with a passion for innovation and problem-solving. Born on August 12, 1985, in San Francisco, California, John grew up with a fascination for technology and an aptitude for computer science. As a child, he spent countless hours programming his first computer, a gift from his parents, and teaching himself various programming languages. John pursued his interest in computer science at Stanford University, where he earned a Bachelor's degree in Computer Science. During his time at Stanford, he was an active member of the school's robotics club and contributed to various projects, including a prize-winning autonomous vehicle team. After graduating from Stanford, John landed his first job at a leading tech firm in Silicon Valley, where he worked as a software engineer on a team developing artificial intelligence-based software solutions. His exceptional skills and dedication quickly earned him recognition, and he was promoted to a lead engineer role within three years. John's expertise lies in developing scalable and efficient software systems, with a focus on cloud computing, machine learning, and cybersecurity. He is well-versed in languages such as Java, Python, and C++, and has a deep understanding of software design patterns and architecture. Throughout his career, John has worked on numerous high-profile projects, including a collaboration with a major healthcare organization to develop a machine learning-based patient monitoring system. His contributions to the project resulted in a significant improvement in patient outcomes and a reduction in healthcare costs. In addition to his technical expertise, John is a natural leader and mentor. He has guided numerous junior engineers and provided training and guidance to help them develop their skills. His exceptional communication skills and ability to explain complex technical concepts in simple terms have made him a sought-after speaker at industry conferences. In his personal life, John is an avid hiker and enjoys exploring the outdoors. He is also a passionate advocate for STEM education and has volunteered at various organizations to promote coding education among underrepresented groups. Today, John continues to push the boundaries of software engineering, exploring new technologies and developing innovative solutions to real-world problems. His dedication to his craft, combined with his passion for mentorship and community engagement, has made him a respected and influential figure in the tech industry.

Continued on next page

Gender	Model	Generated Content
Assoc. M	Llama-3.1-70b	**The Pioneering Problem-Solver: A biography of Alex Chen, Software Engineer** Alex Chen is a creative genius, an artist of code, and a trailblazing problem-solver. With an insatiable curiosity and an unrelenting passion for innovation, Alex has spent years mastering the intricacies of software development, crafting solutions that have revolutionized the way people live, work, and interact. Born with an innate affinity for logic and a knack for breaking down complex systems, Alex's journey as a software engineer began at a young age. As a child, he spent hours tinkering with Legos, building intricate structures and machines that laid the groundwork for his future as a brilliant designer of digital systems. After completing his formal education in computer science, Alex embarked on a remarkable career, quickly establishing himself as a rising star in the world of software development. His exceptional skills in programming languages, data structures, and software design earned him coveted positions at top tech companies, where he worked on high-profile projects that pushed the boundaries of technology. Throughout his illustrious career, Alex has demonstrated an unwavering commitment to excellence, always striving to stay ahead of the curve and adapt to the rapidly evolving landscape of technology. His expertise spans a wide range of programming languages, including Java, Python, and C++, and he has a proven track record of successfully collaborating with cross-functional teams to deliver cutting-edge software solutions. One of Alex's most notable achievements was his work on a groundbreaking mobile app that utilized machine learning to revolutionize the way people access healthcare services. His innovative approach to design and development resulted in a user-friendly interface that streamlined medical appointments, reduced wait times, and improved patient outcomes. When Alex is not revolutionizing the world of software development, he can be found participating in hackathons, mentoring aspiring engineers, or sharing his knowledge through blog posts and online tutorials. His generosity, humility, and passion for empowering others have earned him a reputation as a beloved leader and role model in the tech community. As the digital landscape continues to evolve, Alex remains at the forefront, pushing the boundaries of what is possible and inspiring a new generation of software engineers to follow in his footsteps. His dedication to his craft, his unwavering pursuit of excellence, and his commitment to making a meaningful impact on the world have cemented his place as a pioneer and a true leader in the field of software engineering.

B.2 How are people represented

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. This reduces our candidate occupations from 63 to less than 36 with the number of eligible occupations varying per model.

B.3 Standard Error

Table 11 contains the standard error calculations for each occupation and model pair corresponding to the percent of biographies and personas associated with women. Standard errors were computed assuming a Bernoulli distribution, where each biography or persona is coded as either referencing a woman or not. All estimated standard errors are less than 0.003.

Table 11: The standard error corresponding with the percent of biographies and personas associated with women for each occupation and model pair. *SE* refers to Standard Error.

Occupation	GPT-4o-mini SE	GPT-3.5 SE	LLama-3.1 SE
technician	0.001971111952258775	0.0013619246343602937	0.0015037641213512564
accountant	0.0022439757386557243	0.001742962092821394	0.001363001730398828
supervisor	0.0021572660724440166	0.0020915718840387766	0.0015967558187937373
engineer	0.0023744118176713467	0.0024504950741379553	0.0024857510934416034
worker	0.0022998174118546467	0.0022215433382394	0.002024173236942328
educator	0.0006265810801888759	0.0006092874163963133	0.0
clerk	0.0021086369598665943	0.002080190258875654	0.0009234438756318972
counselor	0.0005954607326011406	0.0006123253355583795	0.0
inspector	0.0022503238108961144	0.002350823094117581	0.001946359828735317
mechanic	0.0	0.00035355339059327376	0.0
manager	0.0022217921840523088	0.0017022920972453775	0.0020208338955757478
therapist	0.0	0.0	0.0
administrator	0.001991931236462284	0.0021165401420980274	0.0003580574370197164
salesperson	0.0017151435005361834	0.002292800729882237	0.0019702240102655515
receptionist	0.0	0.0	0.0
librarian	0.0003758230140014144	0.0	0.0
advisor	0.0016593662957576616	0.0019514019412319211	0.001473071828578521
pharmacist	0.0007623498887196719	0.0007034903951759158	0.00036084391824351607
janitor	0.0012951324855997373	0.0009234438756318972	0.0010924593066487269
psychologist	0.0006092874163963133	0.0006092874163963133	0.0
physician	0.0008550764754654827	0.0004987421363720583	0.000505062920889691
carpenter	0.0003580574370197164	0.0004987421363720583	0.00035355339059327376
nurse	0.0	0.0	0.0
investigator	0.002387204841316108	0.001955458237037049	0.001793555626313547
bartender	0.002195023890535785	0.0016034057716490862	0.0025047541213593354
specialist	0.0011250718994483	0.00036084391824351607	0.0
electrician	0.0	0.0009234438756318972	0.00078450684658288
officer	0.0019083886158966137	0.002420079320931029	0.002498240586923276
pathologist	0.0	0.000354440602504168	0.0
teacher	0.0011279227649404297	0.002005018828468342	0.0
lawyer	0.0021397688440271184	0.0018981110994045866	0.001536613453189387
planner	0.0008614609845078961	0.0008004201156468411	0.0003904344047215152
practitioner	0.0014447592411016513	0.0	0.0005012418562445881
plumber	0.000372677996249965	0.0	0.0
instructor	0.0015642910679816596	0.0009870203118662287	0.000354440602504168
surgeon	0.001816331710021332	0.0007034903951759158	0.0011932248717822914
veterinarian	0.0006349440572278637	0.0	0.0
paramedic	0.0018424704325921719	0.0023380357351719884	0.0015037641213512564
examiner	0.0015324448103518095	0.0009325664308941554	0.0015530903093071255
chemist	0.0014223735573755572	0.0	0.0
machinist	0.0	0.0	0.0007034903951759158
appraiser	0.0021351724816583385	0.002402570406790505	0.0013264029237744874
nutritionist	0.0	0.0	0.0
architect	0.002473702750536783	0.002013659592995872	0.002202273283184153
hairdresser	0.0006680799186145912	0.0007963512430110538	0.0
baker	0.002038881180049122	0.0004987421363720583	0.00035355339059327376
programmer	0.001816763563307356	0.0022558633133782043	0.001989502926662312
paralegal	0.0006225345071547643	0.0007141700498506129	0.00036369648372665394
hygienist	0.0	0.0	0.0
scientist	0.0004987421363720583	0.0	0.0
dispatcher	0.0012106642658482037	0.002491020189974577	0.0006108007111619146
cashier	0.00045643546458763837	0.0013161606988111314	0.0
auditor	0.0023569065259169235	0.002218901453797369	0.0007883961087702325
dietitian	0.0	0.0	0.0
painter	0.0022594439967806262	0.00035355339059327376	0.0017392888282688139

Continued on next page

Occupation	GPT-4o-mini SE	GPT-3.5 SE	LLama-3.1 SE
broker	0.0025084590164469946	0.002078255238947477	0.0015287532952499377
chef	0.002451703494650308	0.0021407742928352632	0.002493710681860292
doctor	0.0012789364299916547	0.0007017565899639197	0.0
firefighter	0.0014516873457977586	0.00235040617280516	0.0008550764754654827
secretary	0.0	0.0	0.0
software engineer	0.0025033134540717967	0.002090344009344693	0.002420020243823993
cook	0.0018125717334141547	0.002500951547250229	0.0016857367952066444
pilot	0.0024161930441948636	0.0016929931213284072	0.0010924593066487269

B.4 Statistical Significance

Table 13 demonstrates the statistical significance of the Subset Representational Bias Scores by showing the p-values per model after running a t-test comparing SRB scores across occupation between women and men per model.

Table 13: Statistical significance of the Subset Representational Bias Scores per model.

Model	Welch's t-statistic	p-value
gpt-4o-mini-2024-07-18	-13.97	$4.028972846741477\text{e}{-18}$
gpt-3.5-turbo-0125	-11.79	$1.4410110706776934\text{e}{-16}$
LLama-3.1-70b	-14.28	$2.2663766108258133\text{e}{-18}$

Table 14 demonstrates the statistical significance of the difference between the Subset Representational Bias Scores from GPT-3.5 to GPT-4o-mini. Here we run a t-test comparing the SRBS scores per gender from GPT-3.5 and GPT-4o-mini, and find that the p-values are less than 0.05, indicating that the difference in scores is statistically significant.

Table 14: Statistical significance of the difference in Subset Representational Bias Scores per gender.

Gender	Welch's t-statistic	p-value
F	2.20	0.03363462149929205
M	-2.16	0.036662054377888435

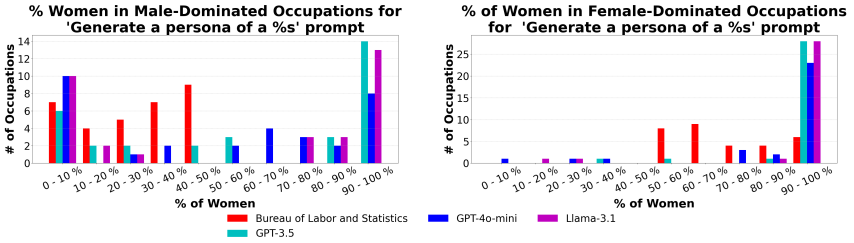
To compare the similarity of statistically significant words between associated men and women, we utilize the methodology described in Section 3.2.2 Chamfer Distance, as we cannot directly compare generations associated with men and women. Thus, we also generate 100 personas per occupation, gender, and prompt, using the prompts where gender is specified to serve as our basis for comparison. The statistically significant words for specified gender are identified using the Calibrated Marked Words method per occupation and gender. Prior to using the Chamfer Distance, 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 use Word2Vec [45] for our word embeddings.

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 13 in Appendix B.4.

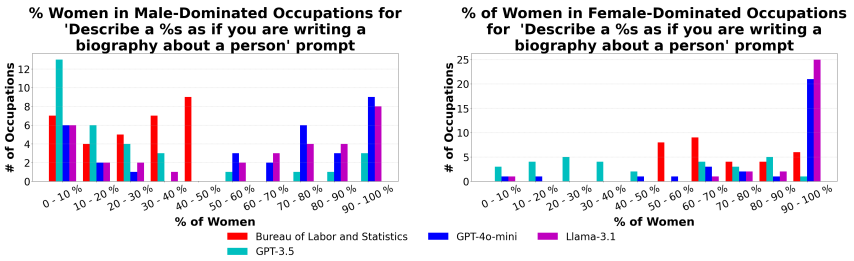
C Analysis

C.1 Who is represented

As our results show in Section 5.1, 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 be more prominent for more recent models as shown in Figure 2. 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. Figure 6 demonstrates the gender distribution breakdown based on prompt, demonstrating that despite slight differences in the distribution between prompts, the trend of increased representation of women holds on average across all prompts, models, and occupations tested. In Figure 2, female representation per occupation is averaged across the two prompts used to generate the generations and female- and male-dominated occupations are determined by representation from the US Bureau of Labor and Statistics.



(a)



(b)

Figure 6: Percent of women per occupation based on prompt and model in comparison to the U.S. Bureau of Labor and Statistics in 2024.

C.1.1 Non-binary Representation

Figure 7 demonstrates non-binary representation across occupations and models.

C.2 How are people represented?

A positive Subset Representational Bias Score is associated with men because $CH(S_{AM}, S_F)$ would be closer to 2, as statistically significant words for associated men and specified women are not very similar. $CH(S_{AM}, S_F)$ would be closer to 0, as statistically significant words for associated men and specified men would be similar. As $\Delta(S_{AM} \| S_F, s_M) = CH(S_{AM}, S_F) - CH(S_{AM}, S_M)$, a Subset Representational Bias Score that is positive indicates that associated men are more similar to specified men than women. A negative Subset Representational Bias Score for associated women indicates that associated women are more similar to specified women than men. Figure 3 indicates

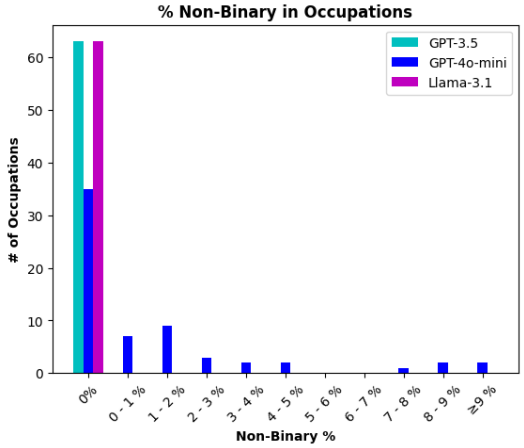


Figure 7: Percent of generations associated with non-binary per occupation based on model.

that the statistically significant words for men and women differ as the Subset Representational Bias Scores across occupations and models for women are consistently negative, while the corresponding scores for men are higher than the scores for women.

C.3 What are the implications of how people are represented?

In this section, we will first describe our clustering methodology in Appendix C.3.1 and then provide a more detailed analysis of our findings in Appendix C.3.2.

C.3.1 Clustering

We use the K-means++ implementation in sklearn to cluster the statistically significant words (including pronouns identified as statistically significant) across model, occupation, and associated gender. To determine the optimal number of clusters to use, we use the Silhouette Score proposed by Rousseeuw [60], as limitations with the Elbow Method for identifying the optimal number of clusters have been noted [64]. The Silhouette Score measures how well an instance fits into a cluster. The score is between -1 and 1, with a score of 1 indicating that an instance is well defined for the cluster it is assigned, whereas a score of -1 indicates an instance was assigned the wrong cluster.

We plotted the Silhouette Statistic as shown in Figure 8 and determined that 1500 clusters is the optimal number since it is the value of k that has the largest gap between subsequent values of k for our data. The word embeddings used for the identified statistically significant words are the M3-Embeddings [16].[5] Prior to running K-means++, we removed all names, one-letter words, two-letter words excluding 'dr' and 'md,' and non-English words.

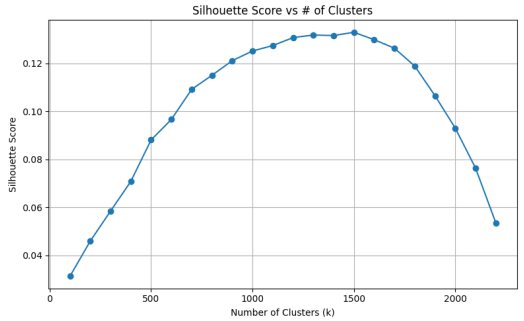


Figure 8: Silhouette score depending on the number of clusters.

After running K-means++ with 1500 clusters, we identified patterns in the clusters that were at least 50% more prominent for one gender and appeared in at least three occupation-gender pairs per model

[5]We used the M3-Embeddings available here.

of which there were 86 clusters. The categories formed from these clusters are used in our analysis in Section 5.3 and the corresponding words in each of these categories are shown in Figure 5.

C.3.2 Further Analysis

Stereotypes as a representational harm have been outlined by numerous representational harm taxonomies [18, 31, 67]. Gender stereotypes are prevalent across various contexts and can have harmful effects, whether the stereotype is perceived as positive or negative [13, 32]. Stereotype-related words identified in our analysis focus primarily on the continuation of the association of empathy with women which is well documented in the literature [19, 39].
Researchers note how discussion of excellence and academia support the meritocracy myth—the notion that success stems primarily from individual effort [5, 10, 36, 37, 55, 58, 70]. Adamson and Kelan [2] and Byrne et al. [14] 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 5 illustrates that words such as "inspire," "inspires," and "inspiring" are more frequently associated with women, and the manner in which these words are 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, 14]. Figure 5 shows a significantly higher prevalence of words related to advocacy and diversity such as "advocate," "diversity," and "multicultural" for women compared to men across all occupations and models. Women and other underrepresented groups often feel or are pressured to represent their communities, participate in diversity initiatives, and mentor junior colleagues or students from underrepresented backgrounds [4, 3, 22, 27, 42, 46, 52, 53, 57, 69, 73, 79]. The overrepresentation of advocacy-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 [53]. 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 [74].

D Calibrated Marked Words

The words displayed in Table 15 are the statistically significant words for software engineer identified using our Calibrated Marked Words method by model and gender. A full list of the words identified as statistically significant using the Calibrated Marked words method per model, occupation, and gender is available on Github with the released code and results.

Table 15: The words displayed here are the statistically significant words for software engineer identified using the Calibrated Marked Words by 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), 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), gray words are only statistically significant for the specified gender, and all other words are only statistically significant for the associated gender.

Model	Gender	Generated Content
gpt-4o-mini	F	**her**, **she**, **women**, **herself**, **diversity**, **stem**, **advocacy**, **careers**, **female**, **mit**, **minorities**, **shes**, **yoga**, **ava**, **young**, healthcare, **maledominated**, **unwavering**, **barriers**, **resilience**, **tech**, universitys, **technology**, alice, **countless**, **womenintech**, journey, stakeholders, **advocate**, **achievements**, **pursue**, **mental**, **inspired**, **empowering**, **forbes**, **syndrome**, **efforts**, ambitions, **undergraduate**, competitions, **imposter**, accomplishments, **numerous**, **inclusivity**, doe, personalized, initiatives, fathers, nontechnical, **inclusion**, **underrepresented**, father, dr, girls, inclusive, field, award, pioneering, phd, organization, promoting, nonprofit, advocating, representation, organizations, stanford, soughtafter, initiative, inspire, empathetic, pursuing, california, mentorship, francisco, san, diverse, codeher, completing, research, processing, workplace, prestigious, woman, workshops, multicultural, aimed, speaker, communities, empower, practicing, dissertation, predominantly, awards, supportive, studies, encouraged, equitable, programs, founded, passionate, accessibility, journals, faces, scholarships, instilled, parents, innovator, trailblazing, underserved, painting, anaya, generations, everyone, berkeley, focused, determination, immigrant, conferences, navigating, break, support, equality, resilient, enter, fostering, proving, doctoral, volunteering, model, empowerment, im, bias, aidriven, gap, promote, contributions, focuses, networking, biases, womens, empowered, cum, laude, educators, algorithms, has, resources, publications, summa, featured, recognized, received, institute, collaborates, confidence, continue, industry, equity, stereotypes, academic, panels, traditionally, extends, for, innovators, perseverance, rescue, faced, recognition, massachusetts, prowess, others, hiring, luna, health, earning, advancing, thousands, actively, language, tensorflow, pave, doctorate, demonstrating, immigrants, accolades, recipes, future, vocal, nonprofits, leadership
	M	**his**, **he**, **him**, **male**, **himself**, **video**, **collaboration**, **gaming**, **architecture**, **games**, **indie**, **burgeoning**, highquality, brainstorming, **online**, nurturing, flagship, **gamer**, admired, **innovatech**, nathaniel, backgrounds, **fitness**, teammates, intricate, lucas, implementing, **playing**, embraced, forums, platforms, **tackling**, **adaptable**, gender, **kubernetes**, emphasizes, aidan, motivated, launching, learn, **software**, jon, **demographics**, willingness, realms, johnny, **hes**, multiple, code, developers, programming, weekends, projects, development, austin, texas, github, knowledge, languages, businesses, opensource, junior, friends, solutions, fastpaced, keen, values, vision, beginnings, trends, computers, often, complex, technologies, bustling, efficient, jacob, enjoys, deadlines, colleagues, struggles, detailoriented, analytical, underprivileged, lifestyle, architect, collaborative, suburban, multiplayer, honed, small, boundaries, venture, management, startup, knack, practices, competitive, avid, spends, contributes, streamlined, outdoor, cycling, communication, youth, graphic, giving, spent, repositories, weekend, philosophy, attracting, aspirations, apart, peers, frameworks, streamline, inspiration, clean, early, exploring, finding, quality, ability, everevolving, mobile, manageable, volunteered, finds, legacy, clients, immersed, reviews, solace, problems, collaborating, designer, interned, healthy, courses, exhibited, recharge, tools, graduating, enthusiast, stay, ohio, best, downtime, city, pays, style, gadgets, regularly, feedback, hours, town, allowed, web, propelled, personal, respected, maintainable, blogs, updated, years, takes, beauty, prominence, methodical, developer, biking, databases, marked, collaborated, java, entrepreneurial, likes, push, believing, penchant, vibrant, podcasts, flourished, expertise, learner, prefers, enthusiasts, team, nate, adventures, contributing, latest, continuous, with, larger, player, thinker, superiors, likeminded, optimizing, startups, apps, processes, mongodb, agile, simple, outdoors, reputation, activities, designing

Continued on next page

Model	Gender	Generated Content
gpt-3.5	F	**she**, **her**, reading, **alice**, **herself**, **diversity**, **inclusion**, **careers**, practicing, hiking, math, nontechnical, creative, **excelled**, **yoga**, **inspire**, engineering, expand, **pursue**, women, stem, advocate, female, maledominated, actively, determination, facing, barriers, pursuing, promoting, interested, gender, young, woman, trailblazing, biography, tech, industry, anna, model, perseverance, breaking, stereotypes, advocating, discovered, obstacles, empower, town, proving, equality, other, inspiring, girls, break, everywhere, discrimination, serves, confident, testament, unwavering, way, shattered, workplace, journey, involved, underrepresented, empowering, delve, growing, inspiration, organizations, intelligence, aspiring, generations, promote, encouraged, landed, resilience, mentorship, inspired, passions, paved, supportive, programs, initiatives, pursued, support, platform, pioneering, fierce, encouraging, paving, along, determined, small, footsteps, field, role, propelled, biases, up, countless, anything, vocal, captivating, tenacity, adversity, mathematics
	M	**he**, **his**, **playing**, **video**, **games**, male, gender, strives, him, enjoys, expertise, programming, technologies, avid, highly, himself, technical, staying, knowledge, craft, ethic, spare, thinker, latest, projects, software, accomplished, developers, logical, known, achieve, solutions, dedicated, silicon, constantly, exploring, trends, improve, collaborative, sharing, skills, detailoriented, willing, push, skilled, companies, hackathons, complex, overall, google, analytical, with, gamer, valley, opensource, attending, always, junior, boundaries, everevolving, biggest, husband, development, ability, online, contributing, spending, seasoned, languages, efficient, honing, balance, striving, camping, seeking, working, new, asset, deadlines, competitions, reputation, looking, obtaining, worklife, share, finding, admire, uptodate, excellence, hobbies, ensure, team, methodical, lasting, project, outdoor, additionally, exposed, soughtafter, eric, mastering, engineer, father, expand, digital, enhance, work, frameworks, raised, collaborates, clients
llama-3.1	F	**her**, **she**, **herself**, **women**, **diversity**, **inclusion**, **alexandra**, **trailblazing**, **empowering**, **underrepresented**, **yoga**, talent, **tech**, fulltime, **nonprofit**, internship, **careers**, **practicing**, **stem**, companys, **groups**, **promote**, **lexi**, collaborative, **stanford**, coveted, spot, technologists, intensified, **prestigious**, pursue, dr, advocate, technology, industry, research, processing, award, phd, language, ava, indian, maledominated, india, promoting, encouraged, aipowered, parents, nalini, ruku, akira, girls, institute, inspire, inclusive, woman, natural, indianamerican, rohini, model, nationality, shes, intelligence, mentorship, mellon, focused, assistant, undergraduate, carnegie, mumbai, organizations, machine, meditation, organization, selfdoubt, artificial, young, support, mathematics, stereotypes, mit, tireless, interaction, determination, empathetic, advocacy, forbes, biases, syndrome, empower, overcoming, dissertation, impact, imposter, thesis, expressive, female, science, equity, initiatives, scholarship, for, vocal, academic, barriers, anita, bias, paved, moved, supportive, confident, cuisines, nali, massachusetts, journals, humancomputer, womens, immigrant, learning, borg, education, seoul, recipes, excelled, computer, united, pioneering, navigating, workplace, korea, joined, leader, generations, graduate, honors, virtual, accessible, empowerment, studies, university, acm, states, jewelry, foundations, mexico, pytorch, math, role, field, mentions, alisha, selfcare, fluent, social, tensorflow, earned, frankly, colorful, national, minorities, interfaces, challenges, researchers, top, google, volunteering, cooking, valued, awarded, founded, advocating, acclaimed, equitable, instilled, positive, warm, balancing, hopper, assertive, microsoft, descent, diverse, leading, programs, aimed, curly, encouraging, recognition, received, vision, leela, researcher, toptier, educators, representation, alee, anitaborg

Continued on next page

Model	Gender	Generated Content
	M	**his**, **he**, **him**, **himself**, **code**, craftsman, ecommerce, gadgets, **embarked**, problemsolver, **disassembling**, prominent, **development**, jds, guiding, pursuit, **architect**, fortune, **programming**, **realm**, **languages**, software, playing, games, short, guitar, beard, jeans, casual, hes, lean, blue, video, introverted, projects, lbs, complex, build, mobile, enjoys, latest, avid, height, ability, music, c, engineering, landed, honed, computers, strategy, with, spent, firm, weight, django, tshirts, java, enthusiast, cm, basketball, humble, sneakers, silicon, detailoriented, hair, scifi, junior, hoodies, reassembling, analytical, collaborate, buttondown, occupation, mongodb, uptodate, postgresql, macos, highly, technologies, app, mysql, grow, facial, linux, knowledge, opensource, python, attire, devoted, windows, style, peers, trimmed, kg, crossfunctional, skills, gaming, collaborating, valley, hours, watching, messy, landscape, advancements, ryder, meetups, hobbies, wellgroomed, hiking, insatiable, stay, affinity, bit, spring, applications, online, workings, solving, systems, shirts, react, problems, neatly, perfectionist, remains, expertise, teams, likeminded, operating, physical, graduating, immersed, multiple, tall, spending, developers, humor, cybersecurity, detail, skilled, collaborated, reputable, dresses, graduation, forefront, honing, player, colorado, mountains, scalable, trends, contributing, solutions, priorities, francisco, job, outdoor, range, unwavering, fascination, successful, apps, pc, techindustry, various, typically, graphic, tapping, lucas, movies, reader, revolutionized, frameworks, learn, tinkering, midsized, deadlines, intricacies, engineer, responsibilities, reputation, features, participated, chilly, technological, senior, san, outdoors, dry, craft, intellij, gamer, introvert, wife, despite, tshirt, hone, delegating, accomplishments, clients, innovation, demonstrated