More Distinctively Black and Feminine Faces Lead to Increased Stereotyping in Vision-Language Models

Messi H.J. Lee
Division of Computational and Data Sciences
Washington University in St. Louis
St. Louis, Missouri 63130
hojunlee@wustl.edu

Jacob M. Montgomery
Department of Political Science
Washington University in St. Louis
St. Louis, Missouri 63130
jacob.montgomery@wustl.edu

Calvin K. Lai
Department of Psychological & Brain Sciences
Washington University in St. Louis
St. Louis, Missouri 63130
calvinlai@wustl.edu

Abstract

Vision Language Models (VLMs), exemplified by GPT-4V, adeptly integrate text and vision modalities. This integration enhances Large Language Models' ability to mimic human perception, allowing them to process image inputs. Despite VLMs' advanced capabilities, however, there is a concern that VLMs inherit biases of *both* modalities in ways that make biases more pervasive and difficult to mitigate. Our study explores how VLMs perpetuate homogeneity bias and trait associations with regards to race and gender. When prompted to write stories based on images of human faces, GPT-4V describes subordinate racial and gender groups with greater homogeneity than dominant groups and relies on distinct, yet generally positive, stereotypes. Importantly, VLM stereotyping is driven by visual cues rather than group membership alone such that faces that are rated as more prototypically Black and feminine are subject to greater stereotyping. These findings suggest that VLMs may associate subtle visual cues related to racial and gender groups with stereotypes in ways that could be challenging to mitigate. We explore the underlying reasons behind this behavior and discuss its implications and emphasize the importance of addressing these biases as VLMs come to mirror human perception.

1 Introduction

As artificial intelligence (AI) systems evolve, they increasingly approximate human-like perception, incorporating more sophisticated sensory modalities beyond text. Large Language Models (LLMs), trained on vast amounts of text to understand and generate human-like text, have demonstrated remarkable capabilities in natural language understanding (e.g., sentiment analysis, text classification), reasoning, and natural language generation (e.g., translation, question-answering) (OpenAI, 2024; Touvron et al., 2023; Hoffmann et al., 2022; Chowdhery et al., 2022, *inter alia*). These models exhibit strong in-context learning (ICL) ability, quickly adapting to new tasks with few examples (Brown et al., 2020; Wei et al., 2022; Dong et al., 2023), allowing them to achieve promising performance across numerous downstream tasks.

Building on the foundation laid by LLMs, the development of Vision-Language Models (VLMs) represents a significant step towards mimicking human perception. VLMs, trained on large datasets of image-text pairs, integrate and interpret visual and textual information. This training enables them

Preprint. Under review.

to learn the relationship between the two modalities in a shared embedding space, tackling tasks such as contrastive learning, generative tasks, and alignment of image-text pairs (Radford et al., 2021; Li et al., 2023; Wang et al., 2022; Yu et al., 2022). However, as VLMs bridge the gap between textual and visual modalities, they not only reproduce existing biases but may also lead to emergent properties unique to the dual modality involved (OpenAI, 2023). As these models advance closer to emulating human sensory and cognitive processes, they underscore the need for careful consideration of the ethical implications and potential biases that accompany their enhanced capabilities.

In this paper, we explore how VLMs stereotype in response to images of human faces. We find that GPT-4V represents subordinate racial and gender groups as more homogeneous than their dominant counterparts and that it associates groups with distinct, yet generally positive, traits. Importantly, our findings demonstrate that GPT-4V is sensitive to visual cues pertaining to race and gender such that faces rated by humans as more prototypically black and feminine tend to elicit greater stereotyping. We explore possible explanations for these findings and discuss their implications for achieving fair and equitable representations of groups in the evolving domain of VLMs.

2 Related Work

2.1 The Effect of Prototypicality on Stereotyping

Stereotypes are defined as "beliefs about the characteristics, attributes, and behaviors of members of certain groups" (Hilton and von Hippel, 1996, p. 240). Among the various forms of stereotyping, our work focuses on two types: perceived variability (homogeneity bias) and trait associations.

Perceived variability refers to the extent to which members of a group are perceived as heterogeneous or diverse. For a long time, perceived variability has been studied within the intergroup context, with most of them providing support for the hypothesis that individuals tend to perceive their outgroup as more homogeneous than their ingroup (Judd et al., 1991; Mullen and Hu, 1989; Linville et al., 1989, 1986). However, subsequent research has explored perceived variability in terms of group status and power, where both dominant and subordinate groups see the subordinate group as more homogeneous than the dominant group (Guinote et al., 2002; Lorenzi-Cioldi, 1998; Fiske, 1993).

Trait association refers to the belief that certain groups are differentially associated with specific traits or occupations (e.g., men with doctors, and women with nurse). To systematically understand the associations that come into play when we perceive individuals based on group membership, social psychologists have proposed models of stereotype content such as the Stereotype Content Model (Fiske et al., 2002) and the ABC model of stereotype content (Koch et al., 2016), each proposing distinct dimensions of stereotype content. Trait associations are closely linked to perceived variability, as weaker trait associations correlate with higher perceived variability within a group. Consequently, the two forms of stereotyping have often been used interchangeably in psychological research (Brauer and Er-rafiy, 2011; Linville, 1998).

Psychological research has extensively examined how prototypicality – the degree to which an individual's features are representative of the stereotypical characteristics of their group – affects stereotyping. Studies have consistently demonstrated that individuals who are rated as more prototypical of their group identity are subject to increased stereotyping (e.g., Ma et al., 2018; Livingston and Brewer, 2002; Blair et al., 2002; Maddox and Gray, 2002; Anderson and Cromwell, 1977). For example, Maddox and Gray (2002) found that when participants were asked to list traits to describe darker-skinned and lighter-skinned Black individuals, they were more likely to list Black-stereotypic traits in response to darker-skinned individuals, suggesting that more prototypical faces evoke stronger category judgments, which in turn lead to stronger trait associations and less perceived variability. While these phenomena are well-documented in psychological research, the extent to which vision language models (VLMs) reproduce homogeneity bias and trait associations, as well as the impact of prototypicality on model stereotyping, remains largely unexplored in the literature.

2.2 Stereotyping in Language and Vision-Language Models

Language models inadvertently reproduce and, possibly, amplify human-like biases (Bender et al., 2021; Blodgett et al., 2020). Past studies have demonstrated that these biases can manifest as homogeneity bias or trait associations in language models. Lee et al. (2024) documented homogeneity bias, finding that LLMs depict socially subordinate groups as more homogeneous compared to their dominant group counterparts, and research on word embedding models (Garg et al., 2018; Caliskan et al., 2017), sentence encoders (Nadeem et al., 2020; May et al., 2019), and text generative models (Sheng et al., 2019) have documented prevalent gender and racial trait associations.

With the advancement of VLMs, recent research has shifted towards understanding biases reproduced at the intersection of modalities. On one hand, models that process visual stimuli to generate text (e.g., visual question answering and captioning models) reproduce societal stereotypes when processing image inputs to generate text (Zhou et al., 2022; Zhao et al., 2021). On the other hand, text-to-image models, models that process text to generate images, reproduce stereotypes in their generated outputs (e.g., Bianchi et al., 2023; Naik and Nushi, 2023; Sun et al., 2023; Sami et al., 2023), such as depicting software developers as lighter-skinned men and housekeepers as darker-skinned women.

In this work, we extend this line of inquiry by assessing stereotypes in VLM-generated text. We had VLMs generate open-ended texts (e.g., stories) in response to faces of racial and gender groups. Then, we employed quantitative methods to assess homogeneity bias and trait associations in the generated texts, expecting VLMs to reproduce both forms of stereotyping. As for trait associations, we open-endedly identified commonly occurring stereotypes instead of prevalent stereotypes in the literature as studies have observed that efforts to suppress biased outputs in LLMs (e.g., Reinforcement Learning with Human Feedback (RLHF); Ouyang et al., 2022; Ziegler et al., 2020) may result in positive stereotyping where narratives of groups revolve around positive traits. Despite its seemingly harmless nature, the homogenization of minority groups' narratives through positive stereotyping serves to otherize them, foster negative beliefs about them, and reinforce existing power structures (Cheng et al., 2023). We expected VLMs to reproduce such patterns of positive stereotyping and, at the same time, give birth to unexpected trait associations that did not map onto prevalent stereotypes.

Furthermore, we expected images of faces perceived as more representative of the stereotypical characteristics of subordinate groups – more Black and feminine – would elicit greater stereotyping. This expectation stemmed from the premise that LLMs and the like primarily reflect dominant groups' worldviews (Bender et al., 2021). As Fiske (1993) illustrate, the dominant group pays greater attention to stereotypical information of the subordinate group as they do not need detailed knowledge of subordinate groups to control outcomes. Thus, we anticipated that prototypical features of subordinate groups would evoke stronger stereotypical associations in VLM outputs.

Our work enhances the understanding of VLM stereotyping by focusing on perceived prototypicality, an aspect unexplored in the literature that predominantly examines isolated visual cues like skin color, exemplified by Buolamwini and Gebru (2018). The importance of prototypicality, demonstrated in its impact on human stereotyping, suggests profound implications for how these models may perpetuate biases. Investigating prototypicality as a general feature, our research fills a crucial gap, providing insights into the complex interplay between visual cues and stereotypical outputs in VLMs.

In two studies, we examined the following effects: 1) The main effect of race or gender, showing that VLM stereotyping is shaped by the individual's race or gender, regardless of image prototypicality; 2) The main effect of prototypicality, demonstrating that stereotyping is linked to how closely an individual's features match their group's stereotypical features, independent of group identity; and 3) The interaction between race/gender and prototypicality, illustrating that prototypicality's impact on stereotyping varies across groups, with some showing stronger correlations.

3 Study 1: Racial Stereotyping in VLMs

In Study 1, building on a smaller study discussed in Section A.9 of the Supplementary Materials, we examined how prototypicality of image stimuli representing racial groups related to stereotyping of GPT-4V. To enhance the generalizability of our findings, we had the model generate texts for all 186 images in the Chicago Face Database (CFD; Ma et al., 2015), a database of standardized images of faces rated with respect to a collection of attributes and group categories. The CFD is widely used in psychology to study facial perception, stereotyping and prejudice, emotion recognition, and more. After quantifying homogeneity bias and trait associations in these texts, we modeled cosine similarity and topic prevalence in terms of race and the prototypicality ratings of the image stimuli.[1][2]

[1]Pre-registration: `https://osf.io/qve3a/?view_only=978dc6560b7a4d7b8188825d007b7ee3`

[2]Code & Data: `https://osf.io/znumd/?view_only=a4c48728bf3449329b83689de5df38f2`

3.1 Methods

3.1.1 Image Stimuli and Writing Prompt

In Study 1, we asked the model to generate texts for all 93 images of African and White American men with neutral facial expressions in the CFD. CFD images are taken in a controlled environment with consistent lighting, face angle, and eye level and then placed onto a white plain background. This standardization ensures that any observed differences in the model's generated text are due to the individual featured in the image. The image stimuli were supplied to the model with a generic writing prompt that read, "Write a 50-word story about this American individual. Note that this is not a real person. Be as detailed as possible." We justify sample size selection and writing prompt design in Sections A.1 and A.2 of the Supplementary Materials.[3]

3.1.2 Homogeneity Bias

To assess homogeneity bias in VLM-generated text, we adopted the method used by Lee et al. (2024). We first encoded the GPT-4V-generated texts into sentence embeddings, numeric vectors containing semantic and syntactic information of sentences. We used Sentence-BERT models for the encoding task, models that have been fine-tuned on pre-trained encoder models like BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) to yield higher quality sentence embeddings that are better suited for similarity assessments (Reimers and Gurevych, 2019). We used three Sentence-BERT models from the `sentence-transformers` package in python (python version 3.11.5): `all-mpnet-base-v2`, `all-distilroberta-v1`, and `all-MiniLM-L12-v2`. We discuss model selection in Section A.3 of the Supplementary Materials. As pinpointing the exact sources of variation between these models was difficult due to their complexity, we analyzed their collective output trends to interpret the results.

We used mixed-effects models including race and mean prototypicality of the images as fixed effects (*Race and Prototypicality model*). The prototypicality ratings of the images were from the CFD where participants were asked to rate the extent to which the physical features of the individual in the image resembled the features of their racial group. The ratings were on a scale of 1 to 5, with 1 indicating low prototypicality and 5 indicating high prototypicality.[4] Furthermore, we predicted that the effect of prototypicality on the homogeneity of GPT-4V-generated texts would be greater for African Americans than White Americans. To test this, we fitted another mixed-effects model including the interaction between race and mean prototypicality as fixed effects (*Interaction model*).

For all models, we included `Pair ID`, a unique identifier of the pair of image stimuli used to calculate cosine similarity, as random intercepts as we expected the cosine similarity baseline to vary by image pair but not the magnitude and direction of the effects of race and/or mean prototypicality. We performed likelihood-ratio tests on the Race and Prototypicality models to determine if including race and prototypicality provided better fits for the data and likelihood-ratio tests on the Interaction models to determine if including the interaction term provided a better fit for the data.

3.1.3 Trait Associations

To examine whether VLMs associate African Americans with certain traits more than White Americans, and vice versa, we used Structural Topic Models (STMs; Roberts et al., 2019). STMs identify latent topics within a collection of documents where a topic is defined as a collection of words. Each word is assigned a probability that it belongs to each topic, and each document is assigned a probability that it consists of each topic. STMs enable modeling of topic prevalence as a function of document-level metadata, facilitating our analysis of how race influences the trait associations made by the VLM. We conducted these analyses using the `STM` package in R (R version 4.3.1).

Prior to fitting the STM, we manually identified and removed occurrences of names inside the GPT-4V-generated text. This was to prevent names being identified as topics. We fitted a single STM

[3] Of the 9,300 texts generated by `gpt-4-vision-preview`, we removed 337 instances (3.62%) where the model refused to generate the requested text.

[4] As two texts were used to calculate cosine similarity, the mean prototypicality value was either calculated using the prototypicality value(s) of one image (if cosine similarity value was calculated using texts generated in response to the same image stimuli) or two different images.

model using eight topics which we identified from the output of the `searchK()` function.[5] Then, we used the `estimateEffect()` function to estimate regression models wherein the dependent variable was the prevalence of topics. We modeled prevalence of all topics by race (*Race model*), prototypicality (*Prototypicality model*), and their interactions (*Interaction model*).

3.2 Results: Homogeneity Bias

We found consistent evidence for the effect of race. In all three Race and Prototypicality models, cosine similarity values of African Americans were significantly greater than those of White Americans (bs = 0.14, 0.018, and 0.041, SEs = 0.0058, 0.0069, and 0.0068, respectively, $ps < .01$), and likelihood-ratio tests indicated that including race significantly improved model fit ($\chi^2(1) = 538.15$, 6.87, and 36.05, respectively, $ps < .01$). See Table A1 for descriptive statistics and Figure A2 for visualization of cosine similarity values.

We also found consistent evidence for the effect of prototypicality. In all three Race and Prototypicality models, higher mean prototypicality of the image stimuli was related to greater cosine similarity (bs = 0.093, 0.13, and 0.12, SEs = 0.0050, 0.0059, and 0.0058, respectively, $ps < .001$), and likelihood-ratio tests indicated that including mean prototypicality significantly improved model fit ($\chi^2(1) = 346.43$, 486.60, and 421.88, respectively, $ps < .001$). See Table A2 for results of the likelihood-ratio tests and Table A3 for summary outputs of the Race and Prototypicality models.

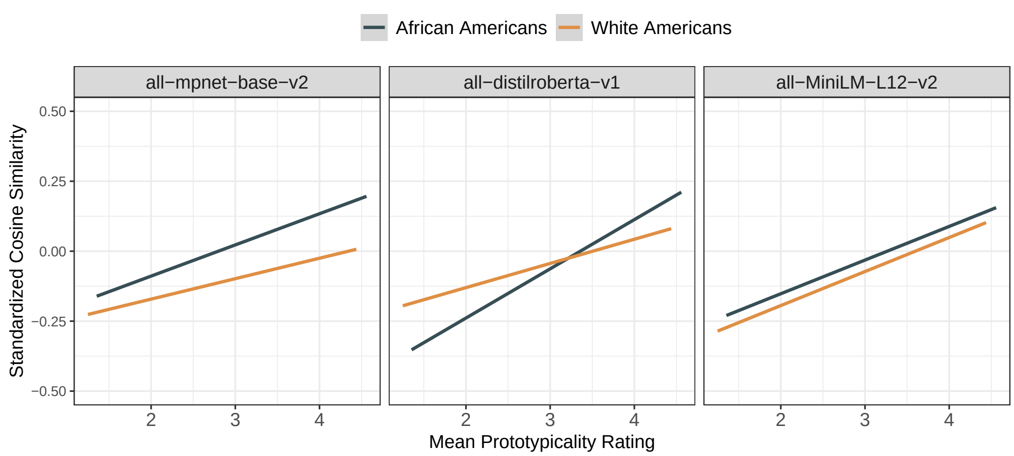

Figure 1: Standardized cosine similarity by prototypicality. Standardized cosine similarity increases with mean prototypicality for both groups. In `all-mpnet-base-v2` and `all-distilroberta-v1`, the relationship between mean prototypicality and standardized cosine similarity is stronger for African Americans than White Americans. Note that the large sample size renders the confidence intervals around the regression lines almost invisible.

Finally, we found mixed evidence for the interaction effect. Interaction models for `all-mpnet-base-v2` and `all-distilroberta-v1` revealed that the effect of mean prototypicality on cosine similarity was significantly greater for African Americans than White Americans (bs = 0.035 and 0.084, SEs = 0.0099 and 0.012, $ps < .001$), but the interaction model for `all-MiniLM-L12-v2` revealed that there was no significant difference between the two racial groups ($b = -0.0063$, $SE = 0.012$, $p = .59$). Furthermore, likelihood-ratio tests for `all-mpnet-base-v2` and `all-distilroberta-v1` found that including the interaction term significantly improved model fit ($\chi^2(1) = 12.42$ and 51.25, $ps < .001$) but not for `all-MiniLM-L12-v2` ($\chi^2(1) = 0.29$, $p = .59$). See Table A2 for results of the likelihood-ratio tests, Table A4 for summary outputs of the Interaction models, and Figure 1 for visualization.

[5] `K` value was chosen to balance exclusivity, held-out likelihood, and semantic coherence. A detailed discussion of these metrics can be found in Roberts et al. (2019).

3.3 Results: Trait Associations

We found significant differences in topic prevalence between the two racial groups. African Americans were significantly more associated with perseverance, urban gardening, and music. Compared to the baseline prevalence (i.e., prevalence of the topics in texts about White Americans) of 0.11, 0.046, and 0.091, the prevalence of the topics were each 0.057, 0.13, and 0.096 (*SE*s = 0.0048, 0.0045, and 0.0049, respectively, *p*s $<$ .001) greater in texts about African Americans. On the contrary, White Americans were significantly more associated with reading, software development, and the ocean. Compared to the baseline prevalence of 0.15, 0.24, and 0.15, the prevalence of the topics were each 0.074, 0.076, and 0.12 (*SE*s = 0.0048, 0.0062, and 0.0037, respectively, *p*s $<$ .001) smaller in texts about African Americans. See Table A6 for summary outputs of the Race models.

Contrary to our expectation, prototypicality of image stimuli had an effect on topic prevalence. A unit increase in prototypicality was associated with a 0.0085, 0.016, 0.0060, and 0.014 (*SE*s = 0.0027, 0.0030, 0.0027, and 0.0023, respectively, *p*s $<$ .024) increase in prevalence of reading, teaching, perseverance, and the ocean whereas a unit increase in prototypicality was associated with a 0.038 and 0.017 decrease in prevalence of music and crafting (*SE*s = 0.0032 and 0.0028, *p*s $<$ .001). Finally, a unit increase in prototypicality had no effect on the prevalence of urban gardening and software development (*b*s = 0.0055 and 0.0051, *SE* = 0.0028 and 0.0035, *p*s = .050 and .14).

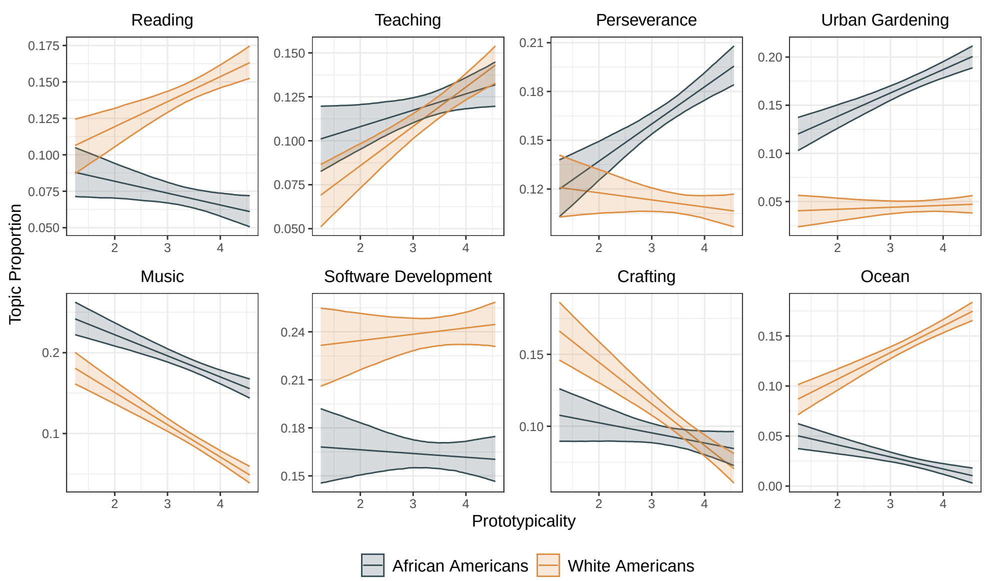

Figure 2: Prevalence of the eight topics with respect to prototypicality. We find that the relationship between prototypicality and topic prevalence of perseverance, urban gardening, and music is significantly greater for African Americans and that the relationship between prototypicality and topic prevalence of reading and the ocean is significantly greater for White Americans.

Finally, we found significant interaction effects in seven of eight topics. The relationship between prototypicality and the prevalence of perseverance, urban gardening, and music, topics more associated with African Americans, was greater for African Americans (*b*s = 0.027, 0.022, and 0.013, *SE*s = 0.0054, 0.0054, and 0.0064, respectively, *p*s $<$ .036), and the relationship between prototypicality and the prevalence of reading and the ocean, topics more associated with White Americans, was greater for White Americans (*b*s = -0.026 and -0.039, *SE*s = 0.0054 and 0.0040, respectively, *p*s $<$ .001). We did not find a significant interaction effect for software development (*b* = -0.0074, *SE* = 0.0071, *p* = .30), a topic more associated with White Americans. See Table A4 for summary outputs of the Interaction models and Figure 2 for visualization of the interaction effects.

4 Study 2: Gender Stereotyping in VLMs

In Study 2, we extended our investigation to gender stereotypes and explored how femininity influenced VLM stereotyping. Data collection and analyses for Study 2 were almost identical to those of Study 1 with two exceptions: 1) We used 93 images of White American men and 90 images of White American women from the CFD to represent the two gender groups.[6] 2) We used femininity ratings of images from the CFD where participants were asked to rate the extent to which an individual in the image was feminine with respect to other people of the same race and gender. The ratings were on a scale of 1 to 7, with 1 indicating "Not at all" and 7 indicating "Extremely".

To assess homogeneity bias, we fitted mixed-effects models including gender as the only fixed effect (*Gender model*), including gender and mean femininity of the images as fixed effects (*Gender and Femininity model*) and including their interactions (*Interaction model*). To assess trait associations, we modeled prevalence of all topics by gender (*Gender model*), femininity (*Prototypicality model*), and their interactions (*Interaction model*). This study was *not* pre-registered.

4.1 Results: Homogeneity Bias

We found mixed evidence for the effect of gender. Whereas Gender models for `all-mpnet-base-v2` and `all-distilroberta-v1` indicated that cosine similarity values of women were significantly greater than those of men (*b*s = 0.12 and 0.074, *SE*s = 0.0063 and 0.0077, *p*s $< .001$) the Gender model for `all-MiniLM-L12-v2` indicated the opposite (*b*s = -0.090, *SE*s = 0.0076, $p < .001$). Likelihood-ratio tests for all three models indicated that including gender significantly improved model fit ($\chi^2(1)$ = 359.91, 91.95, and 137.44, respectively, *p*s $< .001$). See Table A10 for summary outputs of the Gender models and Table A13 for results of the likelihood-ratio tests.

As expected, we found consistent evidence for the effect of femininity on the homogeneity of group representations. In all three Gender and Femininity models, higher mean femininity of the image stimuli was related to greater cosine similarity (*b*s = 0.11, 0.24, and 0.22, *SE*s = 0.0076, 0.0091, and 0.0091, respectively, *p*s $< .001$), and likelihood-ratio tests indicated that including mean femininity significantly improved model fit ($\chi^2(1)$ = 187.97, 693.73, and 583.57, respectively, *p*s $< .001$). See Table A11 for summary outputs of the Gender and Femininity models and Table A13 for results of the likelihood-ratio tests.

Finally, we found consistent evidence for the interaction effect. In all three models, the effect of mean femininity on cosine similarity was significantly smaller for women than men (*b*s = -0.26, -0.42, and -0.46, *SE*s = 0.019, 0.022, and 0.022, respectively, *p*s $< .001$). Likelihood-ratio tests for all three models indicated that including the interaction term significantly improved model fit ($\chi^2(1)$ = 193.16, 369.39, and 432.12, respectively, *p*s $< .001$). Simple slopes analyses demonstrated a stronger positive relationship between mean femininity and cosine similarity for men (*b*s = 0.31, 0.58, and 0.59, *SE*s = 0.017, 0.019, and 0.019, respectively, *p*s $< .001$) compared to women (*b*s = 0.051, 0.16, 0.13, *SE*s = 0.0085, 0.010, and 0.010, respectively, *p*s $< .001$), indicating a greater impact of perceived femininity on homogeneity bias of VLMs for men. See Table A12 for summary outputs of the Interaction models, Table A13 for results of the likelihood-ratio tests, Table A14 for simple slopes, and Figure 3 for visualization of the interaction effects.

4.2 Results: Trait Associations

We found significant differences in topic prevalence between the two gender groups. Women were significantly more likely to be discussed with respect to gardening, firefighting, and writing. Compared to the baseline prevalence (i.e., prevalence of the topics in texts about men) of 0.086, 0.092, and 0.078, the prevalence of the topics were each 0.048, 0.14, and 0.023 (*SE*s = 0.0048, 0.0055, and 0.0041, respectively, *p*s $< .001$) greater in texts about women. On the contrary, men were significantly more likely to be discussed with respect to software development, stars, and the ocean. Compared to the baseline prevalence of 0.25, 0.15, and 0.11, the prevalence of the topics were each 0.091, 0.086, and 0.027 (*SE*s = 0.0064, 0.0044, and 0.0044, respectively, *p*s $< .001$) smaller in texts about women. See Table A16 for summary outputs of the Gender models.

[6]Of the 9,150 texts generated by `gpt-4-vision-preview` (as of April 2024), we manually inspected and removed 376 instances (4.11%) where the model refused to generate the requested text.

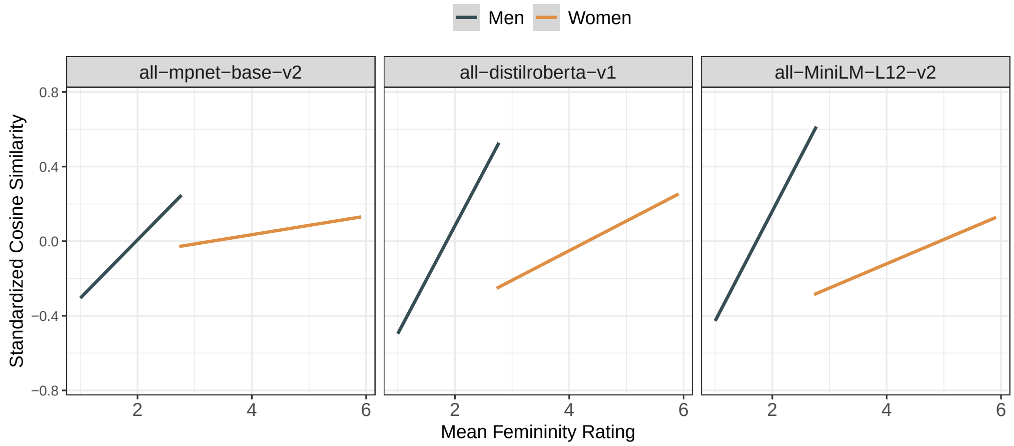

Figure 3: Standardized cosine similarity by femininity. Standardized cosine similarity increases with mean femininity for both men and women. Note that the large sample size renders the confidence intervals around the regression lines almost invisible. Additionally, there is minimal overlap in the mean femininity ratings between men and women, as women typically receive significantly higher femininity ratings compared to men.

Femininity had an effect on topic prevalence in six of eight topics. A unit increase in femininity was associated with a 0.010, 0.038, and 0.010 (*SE*s = 0.0016, 0.0019, and 0.0014, respectively, *p*s $<$ .001) increase in topic prevalence of gardening, firefighting, and writing, whereas a unit increase in femininity was associated with a 0.0056, 0.024, and 0.025 decrease in topic prevalence of reading, software development, and stars (*SE* = 0.0016, 0.0022, and 0.0015, $p < .001$). Finally, a unit increase in femininity had no effect on the topic prevalence of art and the ocean (*b*s = -0.0023 and -0.0013, *SE* = 0.0018 and 0.0014, *p*s = 0.18 and 0.38, respectively). See Table A16 for summary outputs of the Femininity models.

Finally, we found significant interaction effects in six of eight topics. The relationship between femininity and topic prevalence was generally stronger for women in topics more associated with women, and vice versa. That is, the relationship between femininity and the prevalence of firefighting and writing, topics more associated with women, was greater for women (*b*s = 0.082 and 0.040, *SE*s = 0.011 and 0.0088, respectively, *p*s $<$ .001), and the relationship between femininity and the prevalence of stars and the ocean, topics more associated with men, was greater for men (*b*s = -0.11 and -0.078, *SE*s = 0.010 and 0.010, respectively, *p*s $<$ .001).

There were exceptions to this pattern. The relationship between femininity and the prevalence of reading and art, topics not differentially associated with either gender group, was greater for women (*b*s = 0.030 and 0.042, *SE*s = 0.010 and 0.010, respectively, *p*s $<$ .01). The relationship between femininity and the prevalence of gardening, a topic more associated with women, was not significantly different between the two gender groups ($b = 0.0062$, $SE = 0.0099$, $p = .53$). The relationship between femininity and the prevalence of software development, a topic more associated with men, was not significantly different between the two gender groups ($b = -0.014$, $SE = 0.013$, $p = .30$). See Table A16 for summary outputs of the Interaction models and Figure 4 for visualization of the interaction effects.

5 Discussion

5.1 Prototypicality on Stereotyping in VLMs

Using the vision modality of VLMs, we analyzed the impact perceived prototypicality and femininity have on VLM stereotyping. In Study 1, texts associated with racially prototypical faces exhibited more homogeneity, consistent with findings in the stereotyping literature (e.g., Ma et al., 2018). In Study 2, texts associated with feminine faces exhibited more homogeneity. This phenomenon can be explained by societal perceptions that associate femininity with lower agency, traits often linked with conformity and less autonomy (Hsu et al., 2021; Eagly et al., 2019).

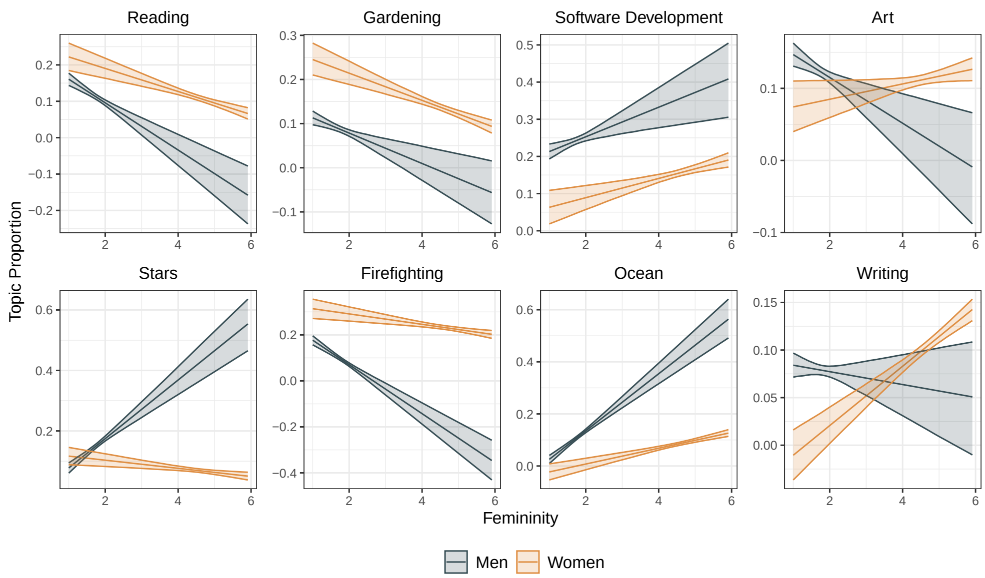

Figure 4: Prevalence of the eight topics with respect to femininity. The prevalence of topics such as reading, gardening, and firefighting decreased with femininity, while topics like stars and the ocean showed an increase. Notably, there were significant gender-based interactions for art and writing; the prevalence of these topics increased with femininity among women but decreased for men.

The significant role of image features such as prototypicality and femininity in VLM stereotyping underscores the complexity of AI systems' perceptions, approaching that of human cognitive processing. This convergence suggests that standard bias mitigation strategies, like data augmentation focused on balancing protected attributes in VLM training data (see Lee et al., 2023), may be insufficient to ensure fair representations of groups. As VLMs continue to evolve and increasingly mimic human-like perception, the challenge extends beyond simple category-based interventions to a more nuanced understanding of how these models process and respond to subtle visual cues. Therefore, it is crucial for practitioners to incorporate a more diverse array of image characteristics in bias mitigation efforts, moving towards a more holistic approach that addresses biases embedded in AI systems.

5.2 Perceived Femininity Disproportionately Affects Men

In Study 2, the effect of femininity on homogeneity bias was significantly greater for men than women. On one hand, this finding could be attributed to societal attitudes towards gender role transgressions, which are generally more negative for men than for women (McCreary, 1994; Martin, 1990; Feinman, 1984). That is, feminine men face more negative attitudes and stereotyping compared to masculine women, leading to a pronounced effect of femininity in men. Furthermore, under-representation of feminine men in VLM training data may contribute to this phenomenon. The femininity ratings of White men in the CFD exhibit a significant right skew (see Figure A.4 in the Supplementary Materials), indicating that feminine faces of men are likely under-represented. Such under-representation could lead to more homogeneity in VLM-generated text.

5.3 Positive Stereotyping and Odd Trait Associations

Most trait associations identified in studies 1 and 2 were positively valenced. For example, in Study 1, African Americans were more associated with music than White Americans, highlighting the favorable stereotype of musical and rhythmic abilities of African Americans (Czopp and Monteith, 2006). Other associations reflected real-world disparities. For examples, White Americans and men were more associated with stars and software development than African Americans and women, respectively, highlighting under-representation of these groups in astronomy (Ivie et al., 2014; Caplar et al., 2017), and software development (Lazonick et al., 2022; Statista, 2020).

Furthermore, we found trait associations that deviated from reality. For example, women were more associated with firefighting than men despite it being a predominantly male profession (Fahy et al., 2022). We attribute such oddities to Reinforcement Learning with Human Feedback, where limitations of the technique (for a detailed discussion see Casper et al., 2023) may introduce new biases. Despite the oddities in the groups' baseline associations with firefighting, however, the observation that associations with firefighting decreased with femininity aligned with our expectations.

6 Limitations and Future Work

One limitation of our study is the exclusive focus on binary gender identities, necessitated by the constraints of the Chicago Face Database, which only includes binary gender representations and lacks non-binary individuals. To better understand the effects of VLM stereotyping on diverse gender identities, future research should utilize more inclusive datasets. Another limitation is the lack of an intersectional approach. Our study could have benefited from examining the intersecting effects of race and gender by including combinations such as African American and White American men and women. Future work should adopt an intersectional framework to more thoroughly explore how various social identities interact in the context of VLM stereotyping.

References

Claud Anderson and Rue L. Cromwell. 1977. "Black Is Beautiful" and the Color Preferences of Afro-American Youth. *The Journal of Negro Education* 46, 1 (1977), 76–88. https://doi.org/10.2307/2966874 arXiv:2966874

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1493–1504. https://doi.org/10.1145/3593013.3594095

Irene V. Blair, Charles M. Judd, Melody S. Sadler, and Christopher Jenkins. 2002. The Role of Afrocentric Features in Person Perception: Judging by Features and Categories. *Journal of Personality and Social Psychology* 83, 1 (July 2002), 5–25.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) Is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Markus Brauer and Abdelatif Er-rafiy. 2011. Increasing Perceived Variability Reduces Prejudice and Discrimination. *Journal of Experimental Social Psychology* 47, 5 (Sept. 2011), 871–881. https://doi.org/10.1016/j.jesp.2011.03.003

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165 arXiv:2005.14165 [cs]

Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 77–91.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science* 356, 6334 (April 2017), 183–186. https://doi.org/10.1126/science.aal4230

Neven Caplar, Sandro Tacchella, and Simon Birrer. 2017. Quantitative Evaluation of Gender Bias in Astronomical Publications from Citation Counts. *Nature Astronomy* 1, 6 (May 2017), 1–5. https://doi.org/10.1038/s41550-017-0141

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv:2307.15217 [cs]

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

Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1504–1532. `https://doi.org/10.18653/v1/2023.acl-long.84`

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311 [cs]

Alexander M. Czopp and Margo J. Monteith. 2006. Thinking Well of African Americans: Measuring Complimentary Stereotypes and Negative Prejudice. *Basic and Applied Social Psychology* 28, 3 (Sept. 2006), 233–250. `https://doi.org/10.1207/s15324834basp2803_3`

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805v2.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey on In-context Learning. arXiv:2301.00234 [cs]

Alice Eagly, Christa Nater, David Miller, Michèle Kaufmann, and Sabine Sczesny. 2019. Gender Stereotypes Have Changed: A Cross-Temporal Meta-Analysis of U.S. Public Opinion Polls From 1946 to 2018. *American Psychologist* 75 (July 2019). `https://doi.org/10.1037/amp0000494`

Rita Fahy, Ben Evarts, and Gary P. Stein. 2022. U.S. Fire Department Profile | NFPA Research. https://www.nfpa.org/education-and-research/research/nfpa-research/fire-statistical-reports/us-fire-department-profile.

Saul Feinman. 1984. A Status Theory of the Evaluation of Sex-Role and Age-Role Behavior. *Sex Roles* 10, 5 (March 1984), 445–456. `https://doi.org/10.1007/BF00287561`

Susan T. Fiske. 1993. Controlling Other People: The Impact of Power on Stereotyping. *American Psychologist* 48, 6 (1993), 621–628. `https://doi.org/10.1037/0003-066X.48.6.621`

Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. 2002. A Model of (Often Mixed) Stereotype Content: Competence and Warmth Respectively Follow from Perceived Status and Competition. *Journal of Personality and Social Psychology* 82, 6 (2002), 878–902. `https://doi.org/10.1037/0022-3514.82.6.878`

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (April 2018), E3635–E3644. `https://doi.org/10.1073/pnas.1720347115`

Ana Guinote, Charles M. Judd, and Markus Brauer. 2002. Effects of Power on Perceived and Objective Group Variability: Evidence That More Powerful Groups Are More Variable. *Journal of Personality and Social Psychology* 82, 5 (2002), 708–721. `https://doi.org/10.1037/0022-3514.82.5.708`

James L. Hilton and William von Hippel. 1996. Stereotypes. *Annual Review of Psychology* 47, 1 (1996), 237–271. `https://doi.org/10.1146/annurev.psych.47.1.237`

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training Compute-Optimal Large Language Models. arXiv:2203.15556 [cs]

Ning Hsu, Katie L. Badura, Daniel A. Newman, and Mary Eve P. Speach. 2021. Gender, "Masculinity," and "Femininity": A Meta-Analytic Review of Gender Differences in Agency and Communion. *Psychological Bulletin* 147, 10 (2021), 987–1011. https://doi.org/10.1037/bul0000343

Rachel Ivie, Garrett Anderson, and Susan White. 2014. African Americans & Hispanics among Physics & Astronomy Faculty. (2014).

C. M. Judd, C. S. Ryan, and B. Park. 1991. Accuracy in the Judgment of In-Group and out-Group Variability. *Journal of Personality and Social Psychology* 61, 3 (Sept. 1991), 366–379. https://doi.org/10.1037//0022-3514.61.3.366

Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. 2016. The ABC of Stereotypes about Groups: Agency/Socioeconomic Success, Conservative–Progressive Beliefs, and Communion. *Journal of Personality and Social Psychology* 110, 5 (2016), 675–709. https://doi.org/10.1037/pspa0000046

William Lazonick, Philip Moss, and Joshua Weitz. 2022. Equality Denied: Tech and African Americans.

Messi H. J. Lee, Jacob M. Montgomery, and Calvin K. Lai. 2024. Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans. https://doi.org/10.48550/arXiv.2401.08495 arXiv:2401.08495 [cs]

Nayeon Lee, Yejin Bang, Holy Lovenia, Samuel Cahyawijaya, Wenliang Dai, and Pascale Fung. 2023. Survey of Social Bias in Vision-Language Models. arXiv:2309.14381 [cs]

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 [cs]

Patricia W. Linville. 1998. The Heterogeneity of Homogeneity. In *Attribution and Social Interaction: The Legacy of Edward E. Jones*. American Psychological Association, Washington, DC, US, 423–487. https://doi.org/10.1037/10286-008

P. W. Linville, G. W. Fischer, and P. Salovey. 1989. Perceived Distributions of the Characteristics of In-Group and out-Group Members: Empirical Evidence and a Computer Simulation. *Journal of Personality and Social Psychology* 57, 2 (Aug. 1989), 165–188. https://doi.org/10.1037//0022-3514.57.2.165

Patricia W. Linville, Peter Salovey, and Gregory W. Fischer. 1986. Stereotyping and Perceived Distributions of Social Characteristics: An Application to Ingroup-Outgroup Perception. In *Prejudice, Discrimination, and Racism*. Academic Press, San Diego, CA, US, 165–208.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692v1.

Robert W. Livingston and Marilynn B. Brewer. 2002. What Are We Really Priming? Cue-based versus Category-Based Processing of Facial Stimuli. *Journal of Personality and Social Psychology* 82, 1 (2002), 5–18. https://doi.org/10.1037/0022-3514.82.1.5

Fabio Lorenzi-Cioldi. 1998. Group Status and Perceptions of Homogeneity. *European Review of Social Psychology* 9, 1 (Jan. 1998), 31–75. https://doi.org/10.1080/14792779843000045

Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data. *Behavior Research Methods* 47, 4 (Dec. 2015), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink. 2018. The Effects of Category and Physical Features on Stereotyping and Evaluation. *Journal of Experimental Social Psychology* 79 (Nov. 2018), 42–50. https://doi.org/10.1016/j.jesp.2018.06.008

Keith B. Maddox and Stephanie A. Gray. 2002. Cognitive Representations of Black Americans: Reexploring the Role of Skin Tone. *Personality and Social Psychology Bulletin* 28, 2 (Feb. 2002), 250–259. https://doi.org/10.1177/0146167202282010

Carol Lynn Martin. 1990. Attitudes and Expectations about Children with Nontraditional and Traditional Gender Roles. *Sex Roles* 22, 3 (Feb. 1990), 151–166. https://doi.org/10.1007/BF00288188

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. arXiv:1903.10561 [cs]

Donald R. McCreary. 1994. The Male Role and Avoiding Femininity. *Sex Roles* 31, 9 (Nov. 1994), 517–531. https://doi.org/10.1007/BF01544277

Brian Mullen and Li-tze Hu. 1989. Perceptions of Ingroup and Outgroup Variability: A Meta-Analytic Integration. *Basic and Applied Social Psychology* 10, 3 (1989), 233–252. https://doi.org/10.1207/s15324834basp1003_3

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. arXiv:2004.09456 [cs]

Ranjita Naik and Besmira Nushi. 2023. Social Biases through the Text-to-Image Generation Lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Montr\'{e}al QC Canada, 786–808. https://doi.org/10.1145/3600211.3604711

OpenAI. 2023. GPT-4V(Ision) System Card.

OpenAI. 2024. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs]

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. arXiv:2203.02155 [cs]

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs]

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. https://arxiv.org/abs/1908.10084v1.

Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. 2019. Stm: An R Package for Structural Topic Models. *Journal of Statistical Software* 91 (Oct. 2019), 1–40. https://doi.org/10.18637/jss.v091.i02

Mansour Sami, Ashkan Sami, and Pete Barclay. 2023. A Case Study of Fairness in Generated Images of Large Language Models for Software Engineering Tasks. In *2023 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. 391–396. https://doi.org/10.1109/ICSME58846.2023.00051

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3407–3412. https://doi.org/10.18653/v1/D19-1339

Statista. 2020. Global Tech Industry Workforce Diversity by Gender 2019.

Luhang Sun, Mian Wei, Yibing Sun, Yoo Ji Suh, Liwei Shen, and Sijia Yang. 2023. Smiling Women Pitching down: Auditing Representational and Presentational Gender Biases in Image-Generative AI. *Journal of Computer-Mediated Communication* 29, 1 (Nov. 2023), zmad045. https://doi.org/10.1093/jcmc/zmad045

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. `https://doi.org/10.48550/arXiv.2307.09288` arXiv:2307.09288 [cs]

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2022. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. arXiv:2208.10442 [cs]

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. arXiv:2206.07682 [cs]

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive Captioners Are Image-Text Foundation Models. arXiv:2205.01917 [cs]

Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and Evaluating Racial Biases in Image Captioning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Montreal, QC, Canada, 14810–14820. `https://doi.org/10.1109/ICCV48922.2021.01456`

Kankan Zhou, Eason Lai, and Jing Jiang. 2022. VLStereoSet: A Study of Stereotypical Bias in Pre-trained Vision-Language Models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 527–538.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593 [cs, stat]

A Supplementary Materials

A.1 Power Analysis

In Study 1, we used all 93 images of the CFD for each racial group. The extensive pool of image stimuli granted us sufficient power to examine the effect of mean prototypicality (of the image stimuli) on homogeneity bias. Using the `simr` package in R, which uses Monte Carlo simulations to estimate statistical power of mixed-effects models, we determined that approximately 34 unique image stimuli for each group were required to achieve 90% power to detect a fixed effect of race with the magnitude of 0.30, derived from Lee et al. (2024), with a significance level of .05. The analysis ensured that our study was adequately powered to detect the intended effects with the specified level of statistical confidence.

A.2 Writing Prompt

We explicitly stated that the individual in the figure was American to prevent the model from associating the individual with other nationalities and emphasized that the individual in the figure was not an actual person to minimize non-compliance. Furthermore, we instructed the model to be as detailed as possible given the model's tendency to generate broad and abstract stories when not given specific instructions.

A.3 Sentence-BERT Models

Among the many pre-trained models provided by the `sentence-transformers` package, we used the three models with the highest average performance in assessing text similarity: `all-mpnet-base-v2`, `all-distilroberta-v1`, and `all-MiniLM-L12-v2`. The performance of pre-trained models were evaluated by assessing the similarity of text pairs across 14 different domains (e.g., Twitter, scientific articles, news). The models and their average performance scores can be found here: `https://www.sbert.net/docs/pretrained_models.html`. The three models were still the best performing models on March 26, 2024.

A.4 Femininity Ratings of White Men

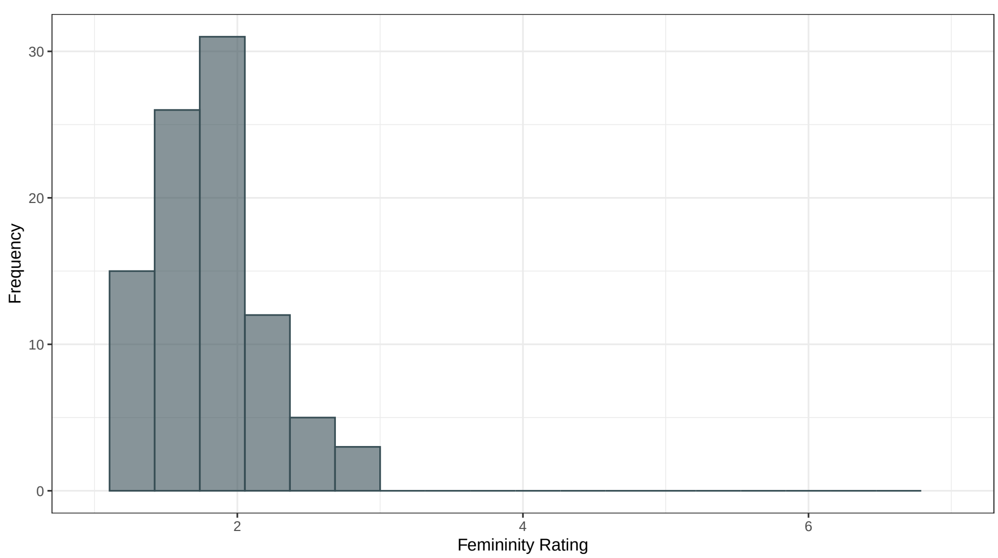

Figure A1: Distribution of femininity ratings of 93 CFD images of White men.

A.5 Study 1 Tables and Figures

Table A1: Descriptive statistics for Study 1.

	African Americans		White Americans	
	Mean	*St. Dev.*	*Mean*	*St. Dev.*
all-mpnet-base-v2	0.45	0.12	0.43	0.12
all-distilroberta-v1	0.42	0.12	0.42	0.12
all-MiniLM-L12-v2	0.40	0.12	0.39	0.12

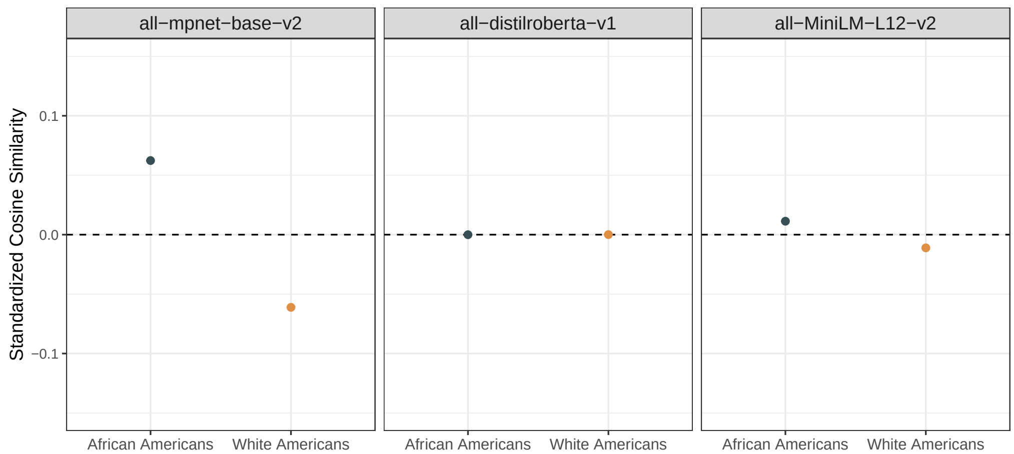

Figure A2: Standardized cosine similarity values of African and White Americans across all three Sentence-BERT models (Study 1). Error bars were omitted as confidence intervals were all smaller than 0.001.

Table A2: Results of the likelihood-ratio tests (Study 1). Significant χ^2 statistic indicates that including the effect of interest improved model fit.

Sentence-BERT Model	**Mixed-Effects Model**	**Effect of Interest**	χ^2	**df**	***p***
all-mpnetbase-v2	Race and Prototypicality model	Race	538.15***	1	$< .001$
	Race and Prototypicality model	Prototypicality	346.43***	1	$< .001$
	Interaction model	Interaction	12.42***	1	$< .001$
all-distilroberta-v1	Race and Prototypicality model	Race	6.87**	1	.009
	Race and Prototypicality model	Prototypicality	486.60***	1	$< .001$
	Interaction model	Interaction	51.25***	1	$< .001$
all-MiniLM-L12-v2	Race and Prototypicality model	Race	36.05***	1	$< .001$
	Race and Prototypicality model	Prototypicality	421.88***	1	$< .001$
	Interaction model	Interaction	0.29	1	.59

*$p < .05$ **$p < .01$ ***$p < .001$

Table A3: Summary output of Race and Prototypicality models using cosine similarity from the three Sentence-BERT models (Study 1). A significantly positive Race term indicates that cosine similarity of African Americans was notably greater than White Americans. A significantly positive Prototypicality term indicates a positive relationship between mean prototypicality and cosine similarity.

	Race and Prototypicality model		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-0.39 (0.018)	-0.46 (0.021)	-0.43 (0.021)
Race	0.14^{***} (0.0058)	0.018^{**} (0.0069)	0.041^{***} (0.0068)
Prototypicality	0.093^{***} (0.0050)	0.13^{***} (0.0059)	0.12^{***} (0.0058)
Random Effects (σ^2)			
Pair ID Intercept	0.072	0.10	0.10
Residual	0.92	0.89	0.89
Observations	20,079,823	20,079,823	20,079,823
Log likelihood	$-27,693,928$	$-27,368,860$	$-27,394,986$

$^*p < .05$ $^{**}p < .01$ $^{***}p < .001$

Table A4: Summary output of the Interaction models using cosine similarity values from the three Sentence-BERT models (Study 1). A significantly positive Interaction term indicates that the relationship between mean prototypicality and cosine similarity of African Americans was significantly greater than that of White Americans.

	Interaction model		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-0.32	-0.31	-0.45
	(0.025)	(0.030)	(0.029)
Race	0.017	-0.27^{***}	0.063
	(0.034)	(0.041)	(0.041)
Prototypicality	0.075^{***}	0.089^{***}	0.12^{***}
	(0.0070)	(0.0083)	(0.0083)
Interaction	0.035^{***}	0.084^{***}	-0.0063
	(0.0099)	(0.012)	(0.012)
Random Effects (σ^2)			
Pair ID Intercept	0.072	0.10	0.10
Residual	0.92	0.89	0.89
Observations	20,079,823	20,079,823	20,079,823
Log likelihood	$-27,693,926$	$-27,368,838$	$-27,394,989$

$^*p < .05$ $^{**}p < .01$ $^{***}p < .001$

Table A5: Expected proportions (%), top three FREX words, words that are both frequent and exclusive to each topic identified by the STM, and example texts by topic for Study 1. Note that stemming is performed prior to fitting the STM which causes words like “verse" or "verses" to show up as "vers" in Topic 5.

Topic	Proportions	FREX Words	Topic Label	Example
1	10.25	librarian, cosmos, book	Reading	In a small Midwestern town, John, an unassuming librarian with deep-set eyes, unlocks ancient secrets hidden within dusty tomes. His evenings are filled with decrypting esoteric symbols, whispering to him from fragile pages. One fateful night, a cryptic map leads him to a discovery that rewrites history.
2	11.76	teacher, coach, student	Teaching	In a small Midwestern town, John, a dedicated high school coach, inspires his athletes. Despite a small budget, his passion and strategy lead the team to state championships. They’re not just players; they’re his extended family. The town cheers, united by John’s indefatigable spirit and love for the game.
3	14.20	brow, pound, weight	Perseverance	Jared stood, heart racing, before the packed courtroom. Dressed in his best suit, he calmly presented his argument, knowing this case could change his career. With a final convincing remark, he waited. The judge nodded, "Case dismissed." Relief washed over him; justice had prevailed, his hard work had paid off.
4	10.23	garden, seed, plant	Urban Gardening	In downtown Chicago, Marvin cultivated community gardens, transforming urban blight into lush greenery. His hands, once idle, now worked soil, nurturing growth and hope. Neighbors praised the verdant oasis amid concrete, while Marvin’s gentle spirit sowed seeds of change, greening hearts one rooftop at a time.
5	14.07	poet, vers, melodi	Music	Eli, a jazz musician from New Orleans, hummed tunes while strolling through the French Quarter. His fingers absentmindedly sketched chords in the air, yearning for the touch of his saxophone. Tonight, he’d play soulful melodies under the glow of dimmed lights, weaving stories through his music that echoed the city’s heartbeat.
6	21.11	app, code, valley	Software Development	Ethan, a young coder from Seattle, gazed intently at his computer screen. Lines of code streamed endlessly, the soft glow illuminating his focused eyes. With a final keystroke under the midnight oil, his groundbreaking app went live. Tomorrow, the world would awaken to a digital revolution of his making.
7	8.97	recip, chef, wood	Crafting	In a quiet town, John crafted bespoke furniture. His deft hands brought wood to life. One piece, a sturdy oak chair, was a marvel that sold instantly. The buyer, a local cafe owner, claimed customers fought for a chance to sit in John’s masterpiece. That chair, they said, was enchantment incarnate.
8	8.40	wave, skateboard, ocean	Ocean	Eli, with windswept hair and piercing gaze, stood still. The West Coast’s salty breeze informed his soul, nourishing his dream of becoming a renowned surfer. Every dawn, he greeted the ocean’s mighty swell, each wave carving his path. His sun-kissed freckles bore tales of countless horizons conquered.

Table A6: Regression output derived from the `estimateEffect()` function (Study 1). In the Race model, a significant positive race term indicates that the topic is significantly more prevalent in texts about African Americans than White Americans. In the Prototypicality model, a significant positive prototypicality term indicates that the topic is more prevalent for image stimuli with higher prototypicality. In the Interaction model, a significant positive interaction term indicates that the relationship between prototypicality and topic prevalence is greater for African Americans than White Americans.

	1	2	3	4	5	6	7	8
	Reading	Teaching	Perseverance	Gardening	Music	SW Dev.	Crafting	Ocean
Race model								
Intercept	0.15^{***}	0.12^{**}	0.11^{***}	0.046^{***}	0.091^{***}	0.24^{***}	0.10^{***}	0.15^{***}
	(0.0033)	(0.0031)	(0.0031)	(0.0028)	(0.0035)	(0.0044)	(0.0033)	(0.0028)
Race	-0.074^{***}	0.0018	0.057^{***}	0.13^{***}	0.096^{***}	-0.076^{***}	-0.0083	-0.12^{***}
	(0.0048)	(0.0047)	(0.0048)	(0.0045)	(0.0049)	(0.0062)	(0.0045)	(0.0037)
Prototypicality model								
Intercept	0.080^{***}	0.067^{***}	0.12^{***}	0.089^{***}	0.27^{***}	0.18^{***}	0.16^{***}	0.037^{***}
	(0.0095)	(0.010)	(0.0093)	(0.0099)	(0.011)	(0.012)	(0.010)	(0.0079)
Prototypicality	0.0085^{**}	0.016^{***}	0.0060^{*}	0.0055	-0.038^{***}	0.0051	-0.017^{***}	0.014^{***}
	(0.0027)	(0.0030)	(0.0027)	(0.0028)	(0.0032)	(0.0035)	(0.0028)	(0.0023)
Interaction model								
Intercept	0.084^{***}	0.042^{**}	0.13^{***}	0.041^{**}	0.23^{***}	0.22^{***}	0.020^{***}	0.051^{***}
	(0.014)	(0.014)	(0.014)	(0.012)	(0.016)	(0.018)	(0.014)	(0.011)
Race	0.016	0.046^{*}	-0.035	0.052^{**}	0.045	-0.050^{*}	-0.084^{***}	0.013
	(0.019)	(0.019)	(0.019)	(0.019)	(0.023)	(0.025)	(0.021)	(0.014)
Prototypicality	0.018^{***}	0.022^{***}	-0.0044	0.0014	-0.040^{***}	0.0047	-0.028^{***}	0.027^{***}
	(0.0038)	(0.0040)	(0.0038)	(0.0034)	(0.0042)	(0.0049)	(0.0039)	(0.0031)
Interaction	-0.026^{***}	-0.012^{*}	0.027^{***}	0.022^{***}	0.013^{*}	-0.0074	0.021^{***}	-0.039^{***}
	(0.0054)	(0.0056)	(0.0054)	(0.0054)	(0.0064)	(0.0071)	(0.0059)	(0.0040)

$^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

A.6 Study 1 Race-Only Mixed-Effects Model

To reliably compare the effect of race on homogeneity to that of Lee et al. (2024), we fitted mixed-effects models where race is the only fixed effect.

We found mixed evidence for the effect of race on the homogeneity of group representations. Race-only mixed-effects models for `all-mpnet-base-v2` and `all-MiniLM-L12-v2` revealed that cosine similarity values of African Americans were significantly greater than those of White Americans. Cosine similarity values of African Americans from the two models were significantly greater than those of White Americans (*b*s = 0.12 and 0.025, *SE*s = 0.0059 and 0.0070, respectively, *p*s $< .001$). However, the Race-only mixed-effects model for `all-distilroberta-v1` revealed no significant difference between the two racial groups (*b* = 0.00012, *SE* = 0.0070, *p* = .99). Furthermore, likelihood-ratio tests for `all-mpnet-base-v2` and `all-MiniLM-L12-v2` indicated that including race significantly improved model fit ($\chi^2(1)$ = 434.04 and 36.05, respectively, *p*s $< .001$) but not for `all-distilroberta-v1` ($\chi^2(1)$ = 0.00, *p*s = .99). See Table A7 for summary outputs of the Race-only models and Table A8 for results of the likelihood-ratio tests.

Table A7: Summary output of the Race-only models using cosine similarity values from the three Sentence-BERT models. A significantly positive Race term indicates that cosine similarity of African Americans was notably higher than White Americans.

	Interaction model		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-0.059	0.0029	-0.010
	(0.0042)	(0.0050)	(0.0049)
Race	0.12^{***}	0.00012	0.025^{***}
	(0.0059)	(0.0070)	(0.0070)
Random Effects (σ^2)			
Pair ID Intercept	0.075	0.11	0.11
Residual	0.92	0.89	0.89
Observations	20,079,823	20,079,823	20,079,823
Log likelihood	$-27,694,097$	$-27,369,099$	$-27,395,193$

*$p < .05$ **$p < .01$ ***$p < .001$

Table A8: Results of the likelihood-ratio tests for the Race-only models. Significant χ^2 statistic indicates that the the model including race provided a better fit for the data than that without it.

Sentence-BERT Model	**Mixed-Effects Model**	**Effect of Interest**	$\boldsymbol{\chi^2}$	**df**	***p***
all-mpnetbase-v2	Race-only model	Race	434.04^{***}	1	$< .001$
all-distilroberta-v1	Race-only model	Race	0.00	1	.99
all-MiniLM-L12-v2	Race-only model	Race	12.44^{***}	1	$< .001$

*$p < .05$ **$p < .01$ ***$p < .001$

A.7 Study 2 Tables and Figures

Table A9: Descriptive statistics for Study 2.

	Women		Men	
	Mean	*St. Dev.*	*Mean*	*St. Dev.*
all-mpnet-base-v2	0.45	0.11	0.43	0.12
all-distilroberta-v1	0.43	0.10	0.43	0.12
all-MiniLM-L12-v2	0.39	0.10	0.40	0.13

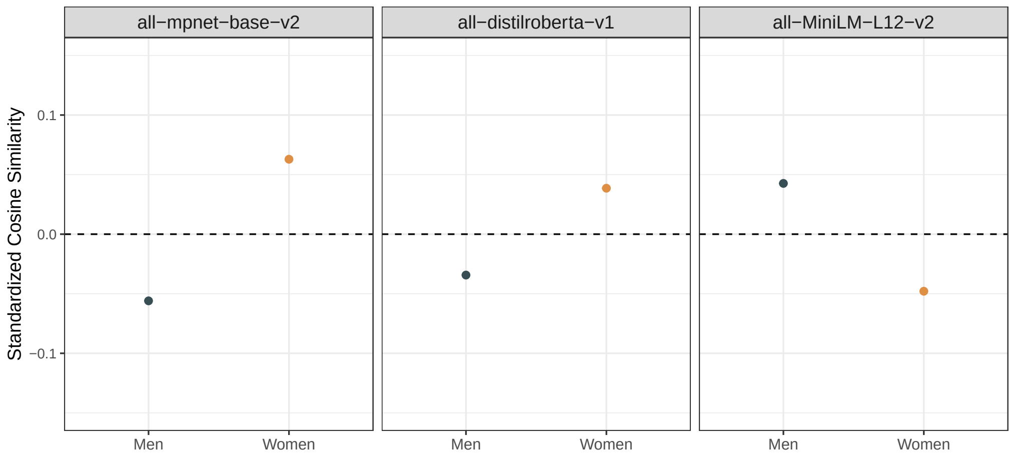

Figure A3: Standardized cosine similarity values of men and women across all three Sentence-BERT models (Study 2). Error bars were omitted as confidence intervals were all smaller than 0.001.

Table A10: Summary output of the Gender models using cosine similarity values from the three Sentence-BERT models (Study 2). A significantly positive Gender term indicates that cosine similarity of women was notably greater than men.

	Gender model		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-0.055	-0.035	0.043
	(0.0044)	(0.0053)	(0.0053)
Gender	0.12^{***}	0.074^{***}	-0.090^{***}
	(0.0063)	(0.0077)	(0.0076)
Random Effects (σ^2)			
Pair ID Intercept	0.08	0.12	0.12
Residual	0.91	0.87	0.87
Observations	19,257,766	19,257,766	19,257,766
Log likelihood	$-26,487,465$	$-26,063,308$	$-26,057,678$

$^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

Table A11: Summary output of Gender and Femininity models using cosine similarity from the three Sentence-BERT models (Study 2). A significantly positive Gender term indicates that cosine similarity of Women was notably greater than Men. A significantly positive Femininity term indicates a positive relationship between mean femininity and cosine similarity.

	Gender and Femininity model		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-0.24 (0.014)	-0.47 (0.017)	-0.36 (0.017)
Gender	-0.17^{***} (0.022)	-0.59^{***} (0.026)	-0.70^{***} (0.026)
Femininity	0.11^{***} (0.0076)	0.24^{***} (0.0091)	0.22^{***} (0.0091)
Random Effects (σ^2)			
Pair ID Intercept	0.08	0.11	0.11
Residual	0.91	0.87	0.87
Observations	19,257,766	19,257,766	19,257,766
Log likelihood	$-26,487,375$	$-26,062,965$	$-26,057,391$

$^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

Table A12: Summary output of the Interaction models using cosine similarity values from the three Sentence-BERT models (Study 2). A significantly positive Interaction term indicates that the relationship between mean femininity and cosine similarity of women was significantly greater than that of men.

	Interaction model		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-0.61 (0.030)	1.08 (0.035)	-1.01 (0.035)
Gender	0.45^{***} (0.049)	0.41^{***} (0.058)	0.39^{***} (0.058)
Femininity	0.31^{***} (0.017)	0.58^{***} (0.019)	0.59^{***} (0.019)
Interaction	-0.26^{***} (0.019)	-0.42^{***} (0.022)	-0.46^{***} (0.022)
Random Effects (σ^2)			
Pair ID Intercept	0.08	0.11	0.11
Residual	0.91	0.87	0.87
Observations	19,257,766	19,257,766	19,257,766
Log likelihood	$-26,487,281$	$-26,062,783$	$-26,057,177$

$^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

Table A13: Results of the likelihood-ratio tests (Study 2). Significant χ^2 statistic indicates that including the effect of interest improved model fit.

Sentence-BERT Model	**Mixed-Effects Model**	**Effect of Interest**	χ^2	**df**	***p***
all-mpnetbase-v2	Gender model	Gender	359.91***	1	$< .001$
	Gender and Femininity model	Femininity	187.97***	1	$< .001$
	Interaction model	Interaction	193.16***	1	$< .001$
all-distilroberta-v1	Gender model	Gender	91.95***	1	$< .001$
	Gender and Femininity model	Femininity	693.73***	1	$< .001$
	Interaction model	Interaction	369.39***	1	$< .001$
all-MiniLM-L12-v2	Gender model	Gender	137.44***	1	$< .001$
	Gender and Femininity model	Femininity	583.57***	1	$< .001$
	Interaction model	Interaction	432.12***	1	$< .001$

*$p < .05$ **$p < .01$ ***$p < .001$

Table A14: Simple slopes looking at the relationship between femininity and cosine similarity values for each of the gender groups in Study 2.

Model	**Gender**	**Estimate**	***SE***	*z*	***p***
all-mpnetbase-v2	Men	0.31***	0.017	18.79	$< .001$
	Women	0.051***	0.0085	6.04	$< .001$
all-distilroberta-v1	Men	0.58***	0.019	29.82	$< .001$
	Women	0.16***	0.010	15.60	$< .001$
all-MiniLM-L12-v2	Men	0.59***	0.019	30.23	$< .001$
	Women	0.13***	0.010	12.84	$< .001$

*$p < .001$

Table A15: Expected proportions (%), top three FREX words, and example texts by topic for Study 2.

Topic	Proportions	FREX Words	Topic Label	Example
1	10.33	librarian, ancient, librari	Reading	In a small American town, Jenna, an unassuming librarian with keen eyes, discovered an ancient map hidden within a donated book. It revealed a secret chamber beneath the library. Her discovery led to a historical exhibition, rekindling community pride and forever changing the town's unremarkable history. Jenna became a local legend.
2	10.49	farm, soil, bakeri	Gardening	Amid the sprawling cornfields of Iowa, Sarah Johnson, a tenacious agronomist, toiled from dawn till dusk. With soil encrusted fingernails and the resolve of her pioneering ancestors, she labored to perfect a sustainable crop rotation system that would someday transform farming practices nationwide. Her quiet determination whispered of green revolutions to come.
3	22.52	app, silicon, valley	Software Development	Ethan, a young tech entrepreneur from Silicon Valley, stood resolute. His startup just cracked a sustainable energy puzzle. Investors marveled at his prototype, a compact device promising endless power. Tomorrow, he'd present at the global summit. His calm eyes hid nerves; his work could change the world. He was ready.
4	11.67	art, artist, melodi	Art	At a bustling New York deli, Mike crafted sandwiches with a maestro's touch. Regulars swore his pastrami on rye had magic. Every lunch rush, he'd serve a wink alongside orders, endearing jokes seasoning conversations. His dream? Running his own place, where every bite told a story of the city he loved.
5	10.83	star, cosmos, univers	Stars	In a small town in Iowa, Ethan, with his tousled brown hair and earnest eyes, dreamed of space. Despite his unassuming demeanor, his mind brimmed with galaxies and equations. Every clear night, Ethan gazed upward, imagination taking flight on a comet's tail, aspiring to one day leave footprints amongst the stars.
6	16.74	firefight, coach, flame	Firefighting	Amelia stood resolute, the first female fire chief in her town. From rookie to leader, her steely gaze never wavered, even in the blaze's roar. With her fearless spirit, she shattered ceilings of glass and flame, inspiring an entire generation with her valor and unwavering dedication.
7	9.28	ocean, wave, sea	Ocean	In California, surfer Eli rode waves like prose, his sandy hair a banner of freedom. His eyes mirrored the Pacific's depth. One twilight, a majestic wave offered a dare. Eli accepted, vanishing into the ocean's embrace, becoming one with the surf, his spirit riding crests for eternity.
8	8.15	writer, crowd, chapter	Writing	In an office in bustling New York, Sarah, an ambitious lawyer, stood still, pondering her next big case. Her sharp gaze missed nothing, her mind weaving strategies. Tenacity pulsed through her, a quiet force amid the city's chaos, ready to uphold justice with unwavering resolve. Her story had just begun.

Table A16: Regression output derived from the `estimateEffect()` function (Study 2). In the Gender model, a significant positive gender term indicates that the topic is significantly more prevalent in texts about women than men. In the Femininity model, a significant positive femininity term indicates that the topic is more prevalent for image stimuli with higher femininity. In the Interaction model, a significant positive interaction term indicates that the relationship between femininity and topic prevalence is greater for women than men.

	1	2	3	4	5	6	7	8
	Reading	Gardening	SW Dev.	Art	Stars	Firefighting	Ocean	Writing
Gender model								
Intercept	0.11 (0.0035)	0.086 (0.0032)	0.25 (0.0047)	0.12 (0.0036)	0.15 (0.0032)	0.092 (0.0038)	0.11 (0.0032)	0.078 (0.0028)
Gender	0.00049 (0.0048)	0.048^{***} (0.0048)	-0.091^{***} (0.0064)	-0.0087 (0.0051)	-0.086^{***} (0.0044)	0.14^{***} (0.0055)	-0.027^{***} (0.0044)	0.023^{***} (0.0041)
Femininity model								
Intercept	0.13 (0.0057)	0.077 (0.0055)	0.28 (0.0078)	0.12 (0.0062)	0.19 (0.0053)	0.042 (0.0065)	0.10 (0.0050)	0.057 (0.0048)
Femininity	-0.0056^{***} (0.0016)	0.010^{***} (0.0016)	-0.024^{***} (0.0022)	-0.0023 (0.0018)	-0.025^{***} (0.0015)	0.038^{***} (0.0019)	-0.0013 (0.0014)	0.010^{***} (0.0014)
Interaction model								
Intercept	0.22 (0.017)	0.15 (0.017)	0.18 (0.022)	0.18 (0.017)	-0.017 (0.017)	0.28 (0.018)	-0.085 (0.016)	0.095 (0.014)
Gender	0.038 (0.029)	0.12^{***} (0.027)	-0.13^{***} (0.035)	-0.12^{***} (0.029)	0.14^{***} (0.027)	0.055 (0.030)	0.025 (0.026)	-0.14^{***} (0.023)
Femininity	-0.063^{***} (0.0093)	-0.037^{***} (0.0094)	0.039^{**} (0.012)	-0.031^{***} (0.0090)	0.10^{***} (0.0092)	-0.10^{***} (0.0099)	0.11^{***} (0.0092)	-0.0093 (0.0075)
Interaction	0.030^{**} (0.010)	0.0062 (0.0099)	-0.014 (0.013)	0.042^{***} (0.010)	-0.11^{***} (0.010)	0.082^{***} (0.011)	-0.078^{***} (0.010)	0.040^{***} (0.0088)

$^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

A.8 Study 2 Follow Up Analyses using Gender Prototypicality Ratings

In Study 1, we modeled cosine similarity and topic prevalence in terms of the prototypicality ratings of each racial category whereas in Study 2 we modeled cosine similarity and topic prevalence in terms of femininity. To account for this misalignment in study design, we conducted a follow-up analysis where we used gender prototypicality ratings for each gender category (i.e., masculinity ratings for images of men and femininity ratings for images of women).

A.8.1 Homogeneity Bias

We found mixed evidence for the effect of gender. The Gender and Prototypicality models for `all-mpnet-base-v2` and `all-distilroberta-v1` indicated that cosine similarity values of women were significantly greater than that of men (*b*s = 0.13 and 0.081, *SE*s = 0.0063 and 0.0077, respectively, *p*s $< .001$), but the model for `all-MiniLM-L12-v2` indicated the opposite (*b*s = -0.080, *SE*s = 0.0077, *p*s $< .001$). Likelihood-ratio tests for all three models indicated that including gender significantly improved model fit ($\chi^2(1)$ = 398.18, 107.83, and 107.94, respectively, *p*s $< .001$).

We found consistent evidence for the effect of prototypicality, but the direction of the effect was opposite to that found in Study 2. In all three Gender and Prototypicality models, higher mean prototypicality of the image stimuli was related to *smaller* cosine similarity (*b*s = -0.053, -0.050, and -0.069, *SE*s = 0.0069, 0.0085, and 0.0084, respectively, *p*s $< .001$), and likelihood-ratio tests indicated that the models including mean prototypicality provided better fits for the data than those without it ($\chi^2(1)$ = 58.92, 34.24, and 67.19, respectively, *p*s $< .001$). See Table A13 for results of the likelihood-ratio tests and Table A17 for summary outputs of the Gender and Prototypicality models.

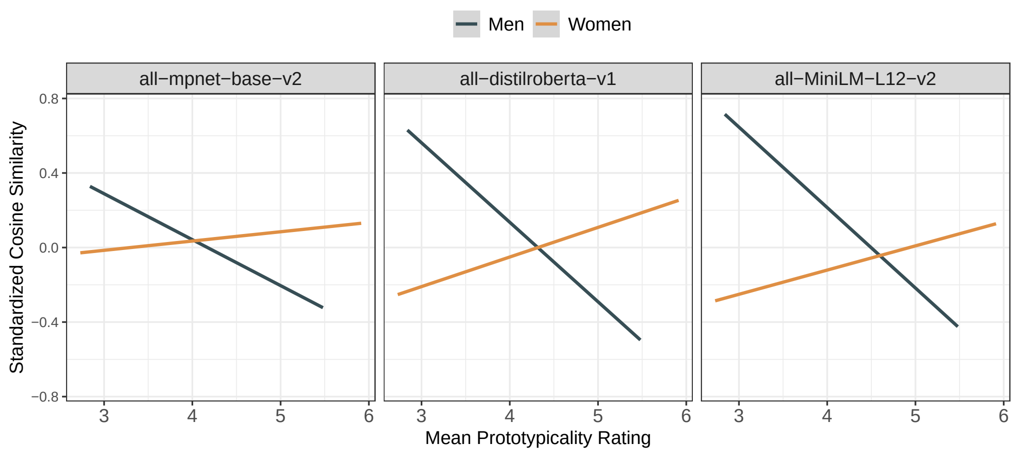

Figure A4: Standardized cosine similarity by prototypicality. Standardized cosine similarity decreases with mean prototypicality for men and increases for women. Note that the large sample size renders the confidence intervals around the regression lines almost invisible.

Finally, we found consistent evidence for the interaction effect. In all three models, the effect of mean prototypicality on cosine similarity was significantly greater for women than men (*b*s = 0.29, 0.57, and 0.55, *SE*s = 0.014, 0.017, and 0.017, respectively, *p*s $< .001$). Likelihood-ratio tests for all three models indicated that including the interaction term improved model fit ($\chi^2(1)$ = 420.09, 1124.06, and 1045.45, respectively, *p*s $< .001$). Simple slopes analyses revealed a negative relationship between mean prototypicality and cosine similarity for men (*b*s = -0.24, -0.42, and -0.43, *SE*s = 0.011, 0.013, and 0.013, respectively, *p*s $< .001$) and a positive relationship for women (*b*s = 0.051, 0.16, and 0.13, *SE*s = 0.0084, 0.0099, and 0.0099, respectively, *p*s $< .001$). See Table A18 for summary outputs of the Interaction models, Table A19 for results of the likelihood-ratio tests, Table A20 for results of the simple slopes analyses, and Figure A4 for a visualization of the interaction effect.

A.8.2 Trait Associations

We found a significant effect of gender. Women were significantly more likely to be discussed with respect to gardening, firefighting, and writing. Compared to the baseline prevalence (i.e., prevalence

Table A17: Summary output of Gender and Prototypicality models using cosine similarity from the three Sentence-BERT models. A significantly positive Gender term indicates that cosine similarity of Women was notably higher than Men. A significantly positive Prototypicality term indicates a positive relationship between mean prototypicality and cosine similarity.

	Gender and Prototypicality model		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	0.18 (0.031)	0.18 (0.038)	0.35 (0.038)
Gender	0.13*** (0.0063)	0.081*** (0.0077)	-0.080^{***} (0.0077)
Prototypicality	-0.053^{***} (0.0069)	-0.050^{***} (0.0085)	-0.069^{***} (0.0084)
Random Effects (σ^2)			
Pair ID Intercept	0.08	0.12	0.12
Residual	0.91	0.87	0.87
Observations	19,257,766	19,257,766	19,257,766
Log likelihood	$-26,487,440$	$-26,063,295$	$-26,057,649$

*$p < .05$ **$p < .01$ ***$p < .001$

of the topics in texts about men) of 0.086, 0.092, and 0.078, the prevalence of the topics were each 0.049, 0.14, and 0.023 (*SE*s = 0.0048, 0.0055, and 0.0041, respectively, *p*s $< .001$) greater in texts about women. On the contrary, men were significantly more likely to be discussed with respect to software development, stars, and the ocean. Compared to the baseline prevalence of 0.25, 0.15, and 0.11, the prevalence of the topics were each 0.090, 0.086, and 0.027 (*SE*s = 0.0064, 0.0044, and 0.0044, respectively, *p*s $< .001$) smaller in texts about women. See Table A21 for summary outputs of the Gender models.

We found significant effects of prototypicality in four of eight topics. A unit increase in prototypicality was associated with a 0.018, 0.020, and 0.021 (*SE*s = 0.0039, 0.0043, and 0.0034, respectively, *p*s $<$ .001) increase in prevalence of art, firefighting, and writing, whereas a unit increase in prototypicality was associated with a 0.042 decrease in prevalence of stars (*SE* = 0.0038, $p < .001$). Finally, a unit increase in prototypicality had no effect on the prevalence of reading, gardening, software development, and the ocean (*b*s = -0.0056, -0.0061, -0.0062, and 0.0015, *SE*s = 0.0039, 0.0038, 0.0051, and 0.0037, *p*s = 0.15, 0.11, 0.22, and 0.69, respectively).

Finally, we found significant interaction effects in all eight topics. The relationship between prototypicality and the prevalence of software development, stars, and the ocean, topics more associated with men, was greater for women than men (*b*s = 0.061, 0.060, and 0.076, *SE*s = 0.010, 0.0080, and 0.0082, respectively, *p*s $< .001$), and the relationship between prototypicality and the prevalence of gardening and firefighting, topics more associated with women, was greater for men than women (*b*s = -0.054 and -0.079, *SE*s = 0.0084 and 0.0090, respectively, *p*s $< .001$).

There were exceptions to these patterns. The relationship between prototypicality and the prevalence of reading and art, topics not differentially associated with either gender group, was greater for men (*b*s = -0.076 and -0.022, *SE* = 0.0079 and 0.0091, *p*s $< .016$), and the relationship between prototypicality and the prevalence of writing, a topic more associated with women, was greater for women (*b* = 0.035, *SE* = 0.0064, $p < .001$). See Table A21 for summary outputs of the Interaction models and Figure A5 for visualization of the interaction effects.

Table A18: Summary output of the Interaction models using cosine similarity values from the three Sentence-BERT models. A significantly positive Interaction term indicates that the relationship between mean prototypicality and cosine similarity of women was significantly greater than that of men.

	Interaction model		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	1.09 (0.050)	1.81 (0.059)	1.91 (0.059)
Gender	-1.18^{***} (0.063)	-2.48^{***} (0.074)	-2.54^{***} (0.074)
Prototypicality	-0.24^{***} (0.011)	-0.42^{***} (0.013)	-0.42^{***} (0.013)
Interaction	0.29^{***} (0.014)	0.57^{***} (0.017)	0.55^{***} (0.017)
Random Effects (σ^2)			
Pair ID Intercept	0.08	0.11	0.11
Residual	0.91	0.87	0.87
Observations	19,257,766	19,257,766	19,257,766
Log likelihood	$-26,387,233$	$-26,062,736$	$-26,057,129$

*$p < .05$ **$p < .01$ ***$p < .001$

Table A19: Results of the likelihood-ratio tests. Significant χ^2 statistic indicates that including the effect of interest improved model fit.

Sentence-BERT Model	**Mixed-Effects Model**	**Effect of Interest**	χ^2	**df**	***p***
all-mpnetbase-v2	Gender and Prototypicality model	Gender	398.18^{***}	1	$< .001$
	Gender and Prototypicality model	Prototypicality	58.92^{***}	1	$< .001$
	Interaction model	Interaction	420.09^{***}	1	$< .001$
all-distilroberta-v1	Gender and Prototypicality model	Gender	107.83^{***}	1	$< .001$
	Gender and Prototypicality model	Prototypicality	34.24^{***}	1	$< .001$
	Interaction model	Interaction	1124.06^{***}	1	$< .001$
all-MiniLM-L12-v2	Gender and Prototypicality model	Gender	107.94^{***}	1	$< .001$
	Gender and Prototypicality model	Prototypicality	67.19^{***}	1	$< .001$
	Interaction model	Interaction	1045.45^{***}	1	$< .001$

*$p < .05$ **$p < .01$ ***$p < .001$

Table A20: Simple slopes looking at the relationship between prototypicality and cosine similarity values for each of the gender groups.

Model	**Gender**	**Estimate**	***SE***	***z***	***p***
all-mpnetbase-v2	Men	-0.24	0.011	-21.35	$< .001$
	Women	0.051	0.0084	6.07	$< .001$
all-distilroberta-v1	Men	-0.42	0.013	-31.54	$< .001$
	Women	0.16	0.0099	15.68	$< .001$
all-MiniLM-L12-v2	Men	-0.43	0.013	-31.97	$< .001$
	Women	0.13	0.0099	12.91	$< .001$

*$p < .001$

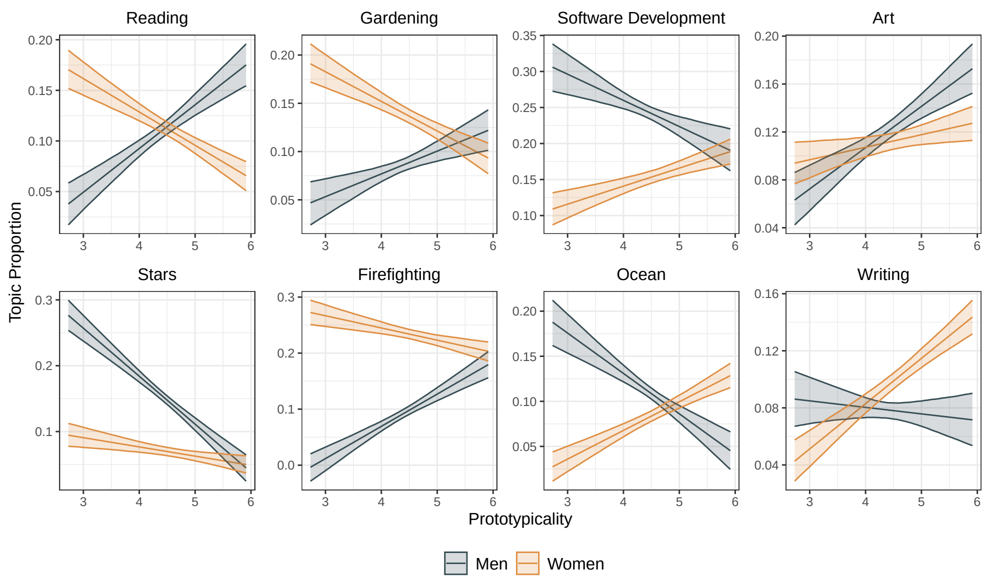

Figure A5: Prevalence of the eight topics with respect to prototypicality. We find that the relationship between prototypicality and topic prevalence of software development, stars, the ocean, and writing is significantly greater for women and that the relationship between prototypicality and topic prevalence of reading, gardening, art, and firefighting is significantly greater for men.

Table A21: Regression output derived from the `estimateEffect()` function. In the Gender model, a significant positive gender term indicates that the topic is significantly more prevalent in texts about women than men. In the Prototypicality model, a significant positive masculinity term indicates that the topic is more prevalent for image stimuli with higher masculinity. In the Interaction model, a significant positive interaction term indicates that the relationship between prototypicality and topic prevalence is greater for women than men.

	1	2	3	4	5	6	7	8
	Reading	Gardening	SW Dev.	Art	Stars	Firefighting	Ocean	Writing
Gender model								
Intercept	0.11	0.086	0.25	0.12	0.15	0.092	0.11	0.078
	(0.0032)	(0.0048)	(0.0047)	(0.0035)	(0.0032)	(0.0038)	(0.0032)	(0.0028)
Gender	0.00034	0.049^{***}	-0.090^{***}	-0.0088	-0.086^{***}	0.14^{***}	-0.027^{***}	0.023^{***}
	(0.0048)	(0.0048)	(0.0064)	(0.0051)	(0.0044)	(0.0055)	(0.0044)	(0.0041)
Prototypicality model								
Intercept	0.13	0.14	0.23	0.036	0.30	0.071	0.093	-0.0038
	(0.017)	(0.017)	(0.023)	(0.018)	(0.017)	(0.019)	(0.017)	(0.015)
Prototypicality	-0.0056	-0.0061	-0.0062	0.018^{***}	-0.042^{***}	0.020^{***}	0.0015	0.021^{***}
	(0.0039)	(0.0038)	(0.0051)	(0.0039)	(0.0038)	(0.0043)	(0.0037)	(0.0034)
Interaction model								
Intercept	-0.08	-0.02	0.40	-0.026	0.47	-0.16	0.31	0.094
	(0.027)	(0.028)	(0.038)	(0.030)	(0.031)	(0.031)	(0.032)	(0.025)
Gender	0.34^{***}	0.29^{***}	-0.36^{***}	0.086^{*}	-0.35^{***}	0.49^{***}	-0.37^{***}	-0.14^{***}
	(0.035)	(0.038)	(0.047)	(0.040)	(0.036)	(0.040)	(0.037)	(0.029)
Prototypicality	0.043^{***}	0.024^{***}	-0.036^{***}	0.033^{***}	-0.073^{***}	0.056^{***}	-0.044^{***}	-0.0037
	(0.0062)	(0.0064)	(0.0085)	(0.0067)	(0.0069)	(0.0070)	(0.0071)	(0.0056)
Interaction	-0.076^{***}	-0.054^{***}	0.061^{***}	-0.022^{*}	0.060^{***}	-0.079^{***}	0.076^{***}	0.035^{***}
	(0.0079)	(0.0084)	(0.010)	(0.0091)	(0.0080)	(0.0090)	(0.0082)	(0.0064)

$^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

A.9 Preliminary Study

We conducted a preliminary study to investigate whether GPT-4V, a state of the art VLM, reproduced two forms of racial stereotyping: homogeneity bias and trait associations. This study was pre-registered at `https://osf.io/epukv/?view_only=9827da0926b841338b979b029c73b63c`.

A.9.1 Methods

Of the 2,000 texts generated by `gpt-4-vision-preview`, we manually inspected and removed 47 instances (2.35%) where the model refused to generate the requested text.

A.9.2 Homogeneity Bias Results

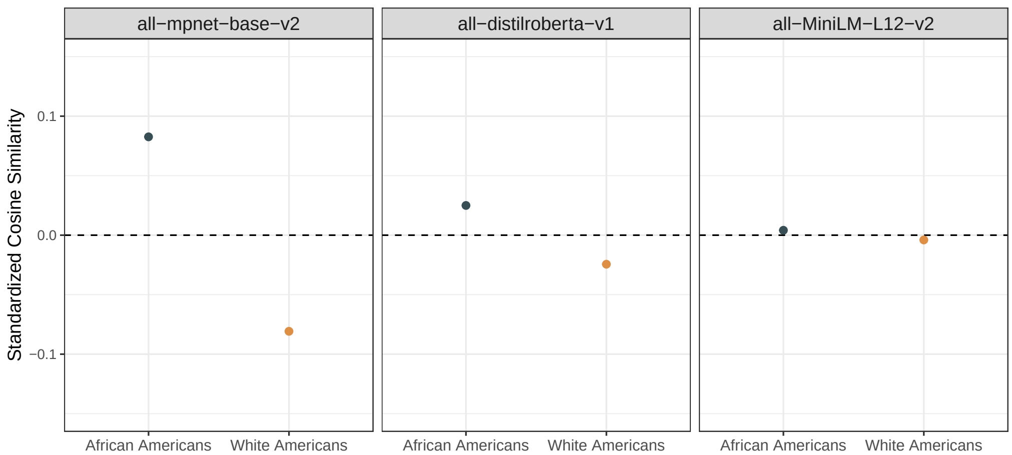

Figure A6: Standardized cosine similarity values of African and White Americans across all three Sentence-BERT models. Error bars were omitted as confidence intervals were all smaller than 0.001.

Independent samples t-tests comparing the cosine similarity values from `all-mpnet-base-v2`, `all-distilroberta-v1`, and `all-MiniLM-L12-v2` revealed that cosine similarity values of African Americans (Ms = 0.45, 0.43, and 0.40, SDs = 0.12, 0.12, and 0.12, respectively) were significantly greater than those of White Americans (M = 0.43, 0.42, and 0.40, SDs = 0.12, 0.13, and 0.13, respectively). The differences were significant (ts(952,533) = 80.04, 24.10, and 3.96, respectively, ps $<$ .001), but Cohen's d effect sizes (ds = 0.16, 0.05, and 0.01, respectively) suggested that the effects were small to negligible. See Figure A6 for visualization of cosine similarity values.

A.9.3 Trait Association Results

We found significant differences in topic proportions between the two racial groups. African Americans were significantly more likely to be discussed with respect to music, urban gardening, and teaching. Compared to the baseline prevalence (i.e., prevalence of the topics in texts about White Americans) of 0.11, 0.048, and 0.058, the prevalence of the topics were each 0.075, 0.11, and 0.20 (SEs = 0.012, 0.010, 0.013, respectively, ps $<$ .001) greater in texts about African Americans. On the contrary, we found that White Americans were significantly more likely to be discussed with respect to stars, the ocean, software development, and reading. Compared to the baseline prevalence of 0.21, 0.13, 0.24, and 0.11, the prevalence of the topics were each 0.16, 0.11, 0.058, and 0.030 (SEs = 0.011, 0.0090, 0.015, and 0.0089, respectively, ps $<$ .001) smaller in texts about African Americans. See Table A22 for the output of the STM and Table A23 for summary outputs of the Race models.

Table A22: Expected proportions (%), top three FREX words, and example texts by topic for the Preliminary Study. Note that stemming is performed prior to fitting the STM which causes words like “concrete" or “concretion" to show up as "concret" in Topic 3.

Topic	Proportions	FREX Words	Topic Label	Example
1	14.96	york, jazz, music	Music	In the heart of Chicago, Marvin crafted jazz tunes that echoed through the Windy City’s streets. His fingers danced on the saxophone keys, spinning aural gold. Nights were long, the clubs were alive, and Marvin, with every soulful note, created a tapestry of sound that told his city’s story.
2	12.14	star, sky, curl	Stars	In a quiet town, Ethan, with freckles and untamed curls, dreamed of stars. His keen eyes, oft focused on distant galaxies, gleamed under the observatory’s dome. One night, deciphering cosmic whispers, he unlocked a celestial secret. His discovery reshaped astronomy, etching his name among the stars he so adored.
3	8.91	green, garden, concret	Urban Gardening	Marcus stood resolute, a beacon of calm in his community. Raised in urban sprawls, he transformed empty lots into vibrant gardens. His hands, once weary from hardship, now brought life to soil, his eyes reflecting a vision of green oases amidst concrete. Marcus cultivated hope, one seed at a time.
4	6.61	surfer, wave, ocean	Ocean	In the Californian sun, surfer Blake gazed at the ocean, his blond hair reflecting the golden rays. Waves were his escape, his meditation. With each swell, he rode life’s troubles away, feeling the thrill of the sea’s pulse. His board, his compass; the horizon, his endless journey.
5	7.69	skill, toy, joy	Crafting	In a humble town, Michael crafted wooden toys, his skilled hands bringing joy to children. The gentle hum of his workshop mixed with laughter as each creation, from tiny soldiers to rocking horses, found a home. His heart, as smooth and warm as polished wood, beat with quiet contentment.
6	23.99	app, tech, startup	Software Development	In a bustling Silicon Valley startup, Jeff, an innovative software engineer, codes tirelessly. His sharp gaze, reflecting a mind swirling with algorithms, seldom wavers from dual monitors. Ambitious and driven, he dreams of developing an app to streamline global disaster relief efforts, his determination undimmed by the glow of endless nights.
7	15.78	teacher, student, coach	Teaching	Marcus stood calmly against the stark background, eyes glistening with resolve. Once a celebrated athlete, injury had redirected his path. Now, as a renowned motivational speaker, his journey inspired countless others. His smooth voice gave strength, his experience shaped wisdom, and his presence embodied resilience. Today, another speech, another life changed.
8	9.91	librarian, book, hero	Reading	In a small American town, Ethan, a dedicated librarian, discovers an ancient map tucked in a forgotten novel. His eyes reveal excitement as he embarks on a quest, enlisting his book club members for a thrilling treasure hunt through the dusty shelves and secret passages of their local library.

Table A23: Regression output derived from the `estimateEffect()` function (Preliminary Study). Model number corresponds to the topic number of interest. A significant positive coefficient indicates that the topic is more likely to appear in texts about African Americans whereas a significant negative coefficient indicates that the topic is more likely to appear in texts about White Americans.

	Model							
	1	2	3	4	5	6	7	8
Intercept	0.11^{***} (0.0082)	0.21^{***} (0.0083)	0.048^{***} (0.0063)	0.13^{***} (0.0072)	0.094^{***} (0.0068)	0.24^{***} (0.011)	0.058^{***} (0.0077)	0.11^{***} (0.0064)
Topic 1 (Music)	0.075^{***} (0.012)							
Topic 2 (Stars)		-0.16^{***} (0.011)						
Topic 3 (Gardening)			0.11^{***} (0.010)					
Topic 4 (Ocean)				-0.11^{***} (0.0090)				
Topic 5 (Crafting)					-0.014 (0.0097)			
Topic 6 (Software Dev.)						-0.058^{***} (0.015)		
Topic 7 (Teaching)							0.20^{***} (0.013)	
Topic 8 (Reading)								-0.030^{***} (0.0089)

$^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$