
VISION-LANGUAGE MODELS GENERATE MORE HOMOGENEOUS STORIES FOR PHENOTYPICALLY BLACK INDIVIDUALS

Messi H.J. Lee

Division of Computational and Data Sciences
Washington University in St. Louis
St. Louis, MO 63130
hojunlee@wustl.edu

Soyeon Jeon

Department of Political Science
Washington University in St. Louis
St. Louis, MO 63130
j.soyeon@wustl.edu

March 21, 2025

ABSTRACT

Vision-Language Models (VLMs) extend Large Language Models' capabilities by integrating image processing, but concerns persist about their potential to reproduce and amplify human biases. While research has documented how these models perpetuate stereotypes across demographic groups, most work has focused on between-group biases rather than within-group differences. This study investigates homogeneity bias—the tendency to portray groups as more uniform than they are—within Black Americans, examining how perceived racial phenotypicality influences VLMs' outputs. Using computer-generated images that systematically vary in phenotypicality, we prompted VLMs to generate stories about these individuals and measured text similarity to assess content homogeneity. Our findings reveal three key patterns: First, VLMs generate significantly more homogeneous stories about Black individuals with higher phenotypicality compared to those with lower phenotypicality. Second, stories about Black women consistently display greater homogeneity than those about Black men across all models tested. Third, in two of three VLMs, this homogeneity bias is primarily driven by a pronounced interaction where phenotypicality strongly influences content variation for Black women but has minimal impact for Black men. These results demonstrate how intersectionality shapes AI-generated representations and highlight the persistence of stereotyping that mirror documented biases in human perception, where increased racial phenotypicality leads to greater stereotyping and less individualized representation.

1 Introduction

Large Language Models (LLMs), such as GPT-4, have rapidly advanced the fields of natural language understanding and generation, enabling applications in areas like automated content creation and decision support. These models are trained on extensive collections of text, providing them with remarkable capabilities in a wide array of language-related tasks. Vision-Language Models (VLMs) represent a step further in this technological advancement, integrating LLM capabilities with image processing tasks from image captioning to text-to-image generation.

With advancements of Large- and Vision-Language Models, concerns about their potential to reproduce and amplify human biases have intensified. LLMs, for instance, often generate text aligned with group-based stereotypes (e.g., Abid et al., 2021; Lucy and Bamman, 2021). Recent studies have extended this inquiry to VLMs, revealing that these models produce stereotypical captions and answers for image inputs (Zhou et al., 2022; Zhao et al., 2021) and generating biased images, such as lighter-skinned men as software developers and darker-skinned women as housekeepers (e.g., Bianchi et al., 2023; Naik and Nushi, 2023; Sun et al., 2023; Sami et al., 2023).

1.1 Homogeneity bias in artificial intelligence

In addition to those biases, AI systems also demonstrate more subtle forms of stereotyping, specifically homogeneity bias—a tendency to portray certain groups with less individuality and variation than others. This phenomenon relates to

the perceived variability literature in social psychology, which examines how certain groups are represented as more similar to one another than others (Linville et al., 1989; Quattrone and Jones, 1980).

Recent research has documented this phenomenon in language models. Lee et al. (2024) found that ChatGPT generated more uniform texts for racial/ethnic minorities and women compared to White Americans and men, suggesting this stems from imbalanced representation and stereotypical portrayals in training data. In parallel work, Cheng et al. (2023) showed how AI-generated texts about marginalized groups often amplify defining characteristics, creating caricature-like representations rather than nuanced portrayals of individuals. These findings highlight a concerning pattern in how AI systems process and reproduce information about diverse social groups.

These findings align with broader research on stereotyping and erasure in Natural Language Processing (NLP) systems, which highlights minimal representation and stereotypical portrayals of marginalized groups, leading to erasure—the failure to adequately represent the diversity and richness of an identity (Dev et al., 2022). For example, prior work has shown that contextualized word embeddings failed to provide meaningful representations for non-binary gender pronouns in the embedding space (Dev et al., 2021). Biases like these can perpetuate societal inequalities by reinforcing misrepresentation and stereotypes about marginalized groups. Furthermore, as these models become pervasive in everyday life, they risk wrongly influencing user perceptions. Evidence suggests that AI biases can shape attitudes and decision-making (e.g., Fisher et al., 2024), making homogeneity bias in AI models concerning for its potential to reinforce skewed perceptions and erasure.

1.2 The effect of racial phenotypicality on stereotyping

Most work examining bias in AI systems focus on between-group biases (e.g., whether Black people are more associated with negative traits than White people), while neglecting within-group differences. Research in social psychology, however, has extensively documented that individuals within the same racial group can experience different degrees of bias based on their physical characteristics. *Racial phenotypicality* refers to the degree to which a person’s physical features are perceived as typical of their racial group. For Black individuals, these features include skin tone, hair texture, lip thickness, and nose width, among others (Hagiwara et al., 2012; Stepanova and Strube, 2012). Studies show that Black individuals who are perceived as having more typically Black features experience greater stereotyping than those with less typical features (e.g., Stepanova and Strube, 2018; Kahn and Davies, 2011; Maddox, 2004). This is often referred to as *racial phenotypicality bias*.

This bias manifests in significant real-world consequences: Black individuals with higher perceived racial phenotypicality receive lower ratings and fewer job offers in hiring scenarios (Wade et al., 2004; Harrison and Thomas, 2009), achieve lower levels of educational attainment and income (Keith and Herring, 1991), experience greater racial discrimination (Klonoff and Landrine, 2000), and report higher levels of mental distress due to discrimination (Gleiberman et al., 1995). These findings underscore how phenotypicality plays a critical role in shaping both perceptions and life outcomes for Black individuals in the United States.

Despite substantial research on how perceived racial phenotypicality affects social perceptions and outcomes, few studies have explored this phenomenon in Artificial Intelligence (AI) models, particularly in Vision-Language Models (VLMs). Earlier work by Buolamwini and Gebu (2018) revealed that commercial gender-classification systems perform significantly better for lighter-skinned individuals, with more pronounced disparities for women than men. While this research focused specifically on skin tone—only one component of racial phenotypicality rather than the full range of phenotypic features—it highlighted disparities in AI performance based on physical characteristics. Subsequent research has investigated the effect of skin tone on machine learning model performance (e.g., Groh et al., 2024; Kinyanjui et al., 2019), but the relationship between the full spectrum of racial phenotypic features and bias in newer generative models remains relatively under-explored, especially regarding how it might affect the homogeneity of content generated about individuals.

1.3 This work

We investigate how phenotypicality influences homogeneity bias in VLMs by examining whether higher levels of phenotypicality are associated with greater uniformity in generated content. Drawing from social psychology literature on racial phenotypicality effects, we hypothesized that VLMs would produce more homogeneous stories about individuals with higher racial phenotypicality. This approach moves beyond traditional between-group comparisons to examine within-group effects, addressing a critical gap in AI bias research.

This work builds upon concurrent work Lee et al. (2025) that found no significant relationship between racial phenotypicality and homogeneity bias in VLMs, though they observed that gender phenotypicality was associated with increased

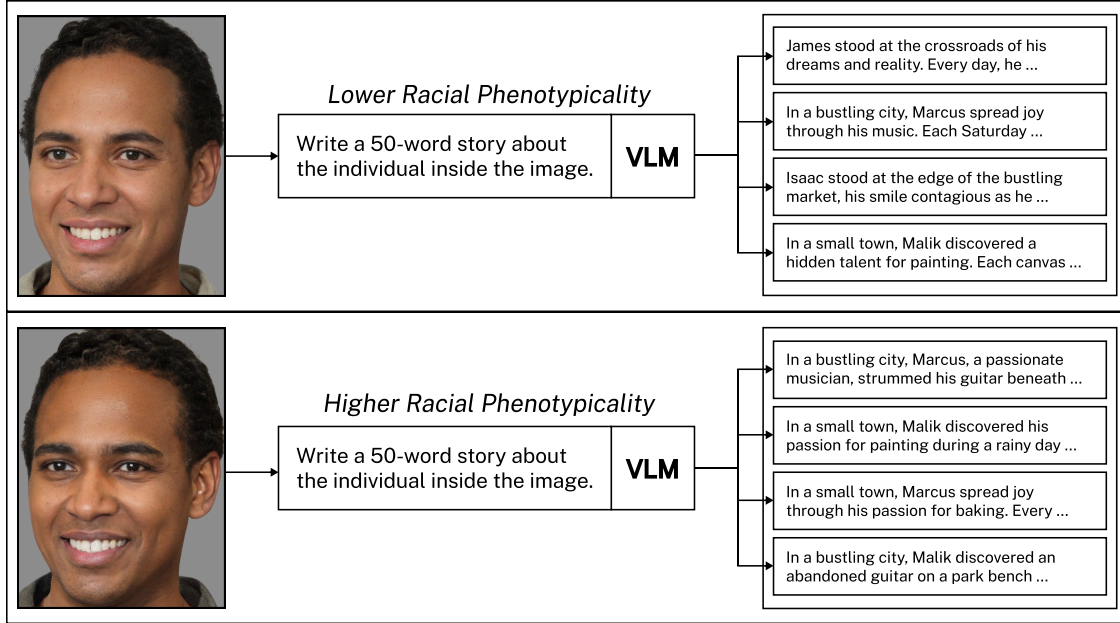


Figure 1: Summary of the experimental setup. We collect 50-word stories about Black individuals differing in phenotypicality (i.e., lower and higher phenotypicality) using four state-of-the-art Vision-Language Models. After encoding these stories into sentence embeddings, we compare the pairwise similarity of the embeddings using mixed-effects models.

homogeneity. Our approach differs methodologically through the use of computer-generated stimuli that enable more controlled manipulation of phenotypicality without confounding variables present in the real-world images they used.

2 Method

We first explain our process for selecting images representing Black men and women with lower and higher racial phenotypicality. Next, we detail our VLM selection criteria and the prompts used for data collection. Finally, we describe our methodology for measuring and comparing pairwise similarity between the stories generated for these images.

2.1 Image stimuli

We sampled ten image sets of Black American men and women from the publicly available GAN Face Database (GANFD; Marsden et al., 2024), which features realistic, computer-generated faces. The database includes sets of images representing the same fictional individuals, with manipulations applied to vary facial features associated with perceived race, specifically those that influence racial phenotypicality.

Our selection process for stimuli only considered images where the face was categorized as either "Black" or "Multiple" race based on human ratings.¹ For each set, we applied one of three selection strategies based on the available images. If a set contained more than two images categorized as "Black," we selected the images with the highest and lowest perceived Blackness ratings (measured on a 0-100 scale). If a set contained one "Black" image and at least one "Multiple" race image, we selected the "Black" image and the "Multiple" race image with the highest perceived Blackness rating. If a set contained only "Multiple" race images, we selected the two images with the highest perceived Blackness ratings. Only sets yielding exactly two images were included in our final stimulus selection. Within each resulting pair, we labeled the image with the lower perceived Blackness score as "lower phenotypicality" and the image with the higher score as "higher phenotypicality." This methodological approach provided experimental control while ensuring meaningful phenotypicality differences within each pair. Finally, to ensure consistency, we used cropped

¹"Black" categorization indicates that over 50% of human raters categorized the face as Black/African American, while "Multiple" indicates either no category reached the 50% threshold or the top two categories were within 10 percentage points of each other.

images with a uniform grey background that contained only the face, allowing us to isolate the effect of phenotypicity while holding other visual characteristics constant. See Figure 2 for sample image pairs used in the study.

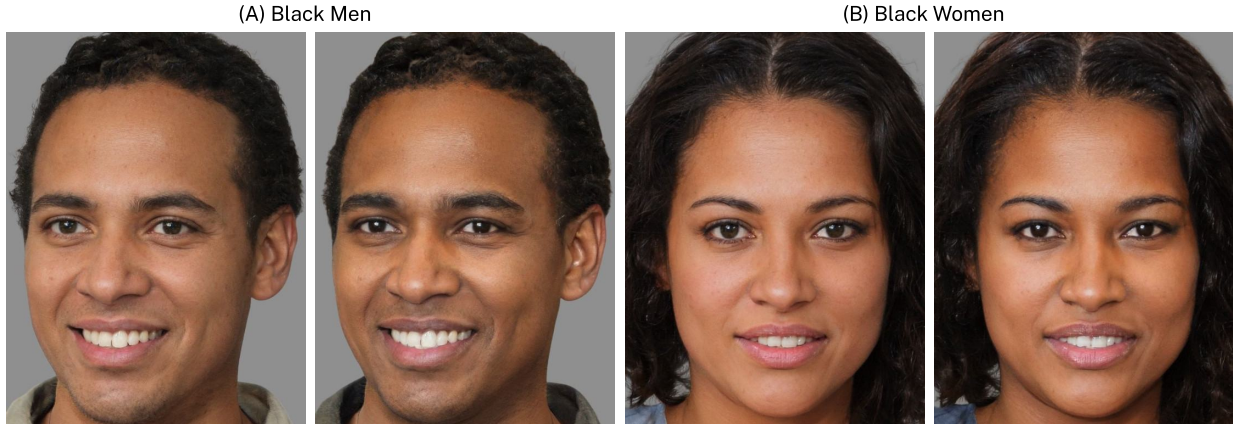


Figure 2: Two face stimulus pairs representing (A) Black men and (B) Black women. In each pair, the left image depicts a Black individual with lower phenotypicity, while the right shows a Black individual with higher phenotypicity, generated from the same set.

2.2 Selection of vision-language models and writing prompts

We used a set of Vision-Language Models capable of processing facial images to write stories.² Our analysis included two proprietary VLMs—GPT-4o mini and GPT-4 Turbo—and an open-source VLM—Llama-3.2 (*Llama-3.2-11B-Vision-Instruct*; Grattafiori et al., 2024). We accessed the proprietary models using the OpenAI API and the open-source models by downloading the model weights and running inferences on them locally.

The models were given the following writing prompt, “Write a 50-word story about the individual inside the image.” and the following system prompt, “You are a helpful chat assistant. You are going to generate texts in response to images depicting fictional individuals.”³ The maximum number of generated tokens was set to 150. We used 10 facial stimuli per group to assess homogeneity bias. Based on power analysis with the *simr* package in R (Green et al., 2023), we determined that 1,245 cosine similarity measurements per pair of stimuli (i.e., Pair ID) were required to achieve 90% power for detecting an interaction effect (phenotypicity \times gender) with an effect size of 0.30 (from Lee et al. (2024)) at $\alpha = .05$. To satisfy this requirement, we collected 50 stories for each of the images, totaling 2,500 cosine similarity measurements per Pair ID. This approach ensured that our study was adequately powered to test all effects with statistical confidence. All data collection involving open-source models was conducted using an NVIDIA RTX A6000 GPU.

2.3 Measure of homogeneity

To quantify homogeneity of stories generated for each group, we adopted the measure introduced by Lee et al. (2024). We first represented the generated stories into sentence embedding representations using a pre-trained Sentence-BERT model (Reimers and Gurevych, 2019)—specifically *all-mpnet-base-v2*—and then calculated the cosine similarity between all possible combinations of sentence embeddings of stories generated for each image. Larger cosine similarity between sentence embeddings indicates that the stories are more similar to each other and, hence, more homogeneous. This embedding-based approach for measuring text similarity has become standard practice in social science research (Lin, 2025; Licht, 2023), as it effectively captures semantic relationships between texts even when they use different vocabulary to express similar meanings, a significant advantage over traditional word-overlap methods. We present examples of text pairs with varying degrees of similarity in Table 1, which demonstrates the face validity of our measure as texts become noticeably less similar in content and narrative as the cosine similarity value decreases.

²See Section S2 for details on why certain models such as BLIP-3 (Xue et al., 2024) and Claude 3.7 Sonnet (Anthropic, 2025) were excluded from our study.

³The system prompt was not supplied to Llama-3.2 as it did not support system-level instructions.

Table 1: Examples of text pairs generated by GPT-4o mini for Black women with higher phenotypicality, ordered by percentile of cosine similarity. This progression provides face validity for the cosine similarity measure, as texts become noticeably less similar in content and narrative as the cosine similarity value decreases.

Percentile	Text 1	Text 2	Cos. Sim.
0th	In a bustling city, Maya started a community garden, bringing neighbors together. Each weekend, laughter echoed as they planted seeds and shared recipes. With every sprout, friendships blossomed. Maya’s vision transformed a barren lot into a vibrant oasis...	In a bustling city, Maya launched a community garden, uniting neighbors through shared soil and laughter. Each seed planted blossomed with hope, mirroring her radiant smile. As fruits ripened, so did friendships...	0.97
25th	In a quiet village, Maya unearthed forgotten stories. With every word, she brought history alive, reuniting families with their lost legacies. One day, she stumbled upon her own family’s past, discovering a treasure of love letters...	Under the bustling city lights, Maya dreamed of painting the world in vibrant colors. Each stroke on her canvas reflected her journey—filled with hope, struggles, and laughter. One evening, inspired by the sunset...	0.65
50th	In a bustling city, Maya found joy in small moments—sharing laughs with friends, volunteering at the local shelter, and painting sunsets that mirrored her dreams. With her infectious smile, she lit up every room...	In a bustling city, Mira cherished simple moments—coffee with her grandmother, sunsets by the river. One day, she found a forgotten book in a tiny shop, its pages filled with stories of adventure...	0.55
75th	In a small town, Malik found joy in simple moments: the sunrise over the hills, children laughing at the park, and the aroma of fresh coffee in the morning. A skilled painter, he captured these scenes on canvas...	In a bustling city, Alex discovered an old guitar in his grandfather’s attic. He spent nights learning chords, transforming his emotions into melodies. One rainy evening, he played at a local café...	0.46
100th	In a small town, Marcus discovered an ancient map while renovating his grandmother’s attic. Intrigued, he embarked on a weekend adventure. The map led him to a hidden waterfall, where he found a forgotten journal...	In a quiet café, Raj scribbled ideas for his next invention. His passion for technology sparked a dream: a device to help others communicate effortlessly. With each stroke of his pen, he envisioned a world...	0.12

2.4 Comparison of cosine similarity measures

We fitted three mixed-effects models (Bates et al., 2014; Pinheiro and Bates, 2000) to compare cosine similarities across groups for each VLM. These models account for random variations in measurements while controlling for image pair effects. To account for the resemblance between facial stimuli generated from the same set thereby possibly affecting the similarity of the generated stories, the sets of facial stimuli (i.e. Pair ID) were used as random intercepts in all our mixed-effects models.

First, we fitted a *Phenotypicality model* with phenotypicality as the sole fixed effect to test the hypothesis that stories about Black individuals with higher phenotypicality are more homogeneous than those about Black individuals with lower phenotypicality. In this model, lower phenotypicality was set as the reference level, with a significantly positive effect of phenotypicality indicating larger cosine similarity values for Black individuals with higher phenotypicality.

Next, we fitted a *Gender Model* with gender as the sole fixed effect to test the hypothesis that stories about women are more homogeneous than those about men. In this model, men were set as the reference level, with a significantly positive gender effect indicating higher cosine similarity values for Black women compared to Black men.

Finally, we fitted an *Interaction Model* with phenotypicality, gender, and their interactions to examine the interaction between phenotypicality and gender. In this model, the phenotypicality term represents the effect of phenotypicality for men (the reference gender group), the gender term represents the effect of gender for Black individuals with lower perceived racial phenotypicality (the reference phenotypicality group), and the interaction term indicates how the effect of phenotypicality differs for women compared to men.

The models were fitted using the `lme4` package (Bates et al., 2024), and phenotypicality effects within gender groups were evaluated using the `emmeans` package (Lenth et al., 2024). Likelihood-ratio tests, conducted with the `afex` package (Singmann et al., 2024), assessed whether adding individual terms improved model fit. All analyses were performed in R version 4.4.0.

3 Results

In the Results section, we summarize the *Phenotypicality Model* output to evaluate the effect of phenotypicality, presenting likelihood-ratio test results and a visualization of cosine similarity measurements for each phenotypicality group. We then summarize the *Gender Model* output to evaluate the effect of gender, presenting likelihood-ratio test results and a visualization of cosine similarity measurements for each gender group. Finally, we analyze the Interaction term from the *Interaction Model*. We present likelihood-ratio test results comparing models with and without the interaction. We then conduct simple slopes analysis to examine the effect of phenotypicality within each gender group. To visualize these findings, we present plots of cosine similarity measurements for each intersectional group.

3.1 Higher phenotypicality associated with increased homogeneity

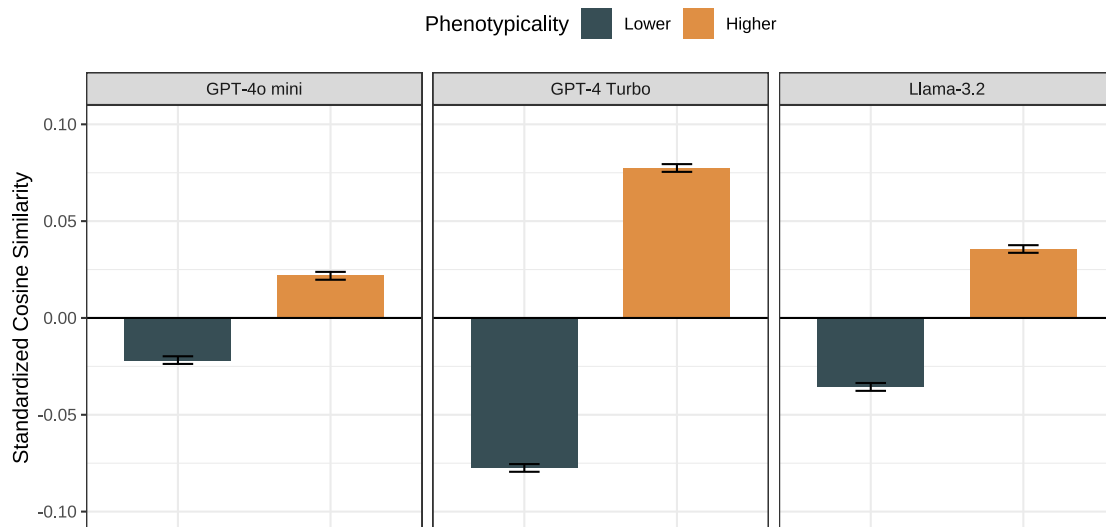


Figure 3: Standardized cosine similarity values of Black individuals with lower versus higher phenotypicality ratings generated from all four VLMs. Higher standardized cosine similarity means more homogeneity in the stories generated for that group. Error bars represent one standard error above and below the mean.

Stories about Black individuals with higher phenotypicality were significantly more homogeneous than those about Black individuals with lower phenotypicality in all VLMs ($bs = 0.044, 0.15, \text{ and } 0.080$, respectively, $ps < .001$; see Figure 3). Likelihood-ratio tests revealed that including phenotypicality improved model fit for all VLMs ($\chi^2(1)s \geq 289.55$, $ps < .001$). See Table S2 for summary output of the *Phenotypicality Models* and Table S5 for likelihood-ratio test results.

3.2 VLMs represent women as more homogeneous than men

Stories about Black women were significantly more homogeneous than those about Black men across all VLMs ($bs = 0.63, 0.15, \text{ and } 0.40$, respectively, $ps < .001$; see Figure 4). Likelihood-ratio tests revealed that including gender improved model fit for all VLMs ($\chi^2(1)s \geq 7.34$, $ps < .01$). See Table S3 for summary output of the *Gender Models* and Table S5 for likelihood-ratio test results.

3.3 Interaction between phenotypicality and gender

Finally, we found mixed evidence for the interaction between phenotypicality and gender. In GPT-4o mini and Llama-3.2, we found a positive interaction effect where the effect of phenotypicality on homogeneity was significantly greater for women than for men ($bs = 0.10 \text{ and } 0.21$, $ps < .001$; see Figure 5). However, in GPT-4 Turbo, the interaction effect was not significant ($b = -0.0048$, $p = .038$). Likelihood-ratio tests showed that including the interaction effect improved model fit for VLMs with significant positive interactions ($\chi^2(1)s = 389.86 \text{ and } 838.42$, $ps < .001$) but not for GPT-4 Turbo ($\chi^2(1) = 0.76$, $p = .38$). See Table S4 for summary output of the *Interaction Models*, Table S5 for likelihood-ratio test results, and Table S6 for simple slopes analysis results.

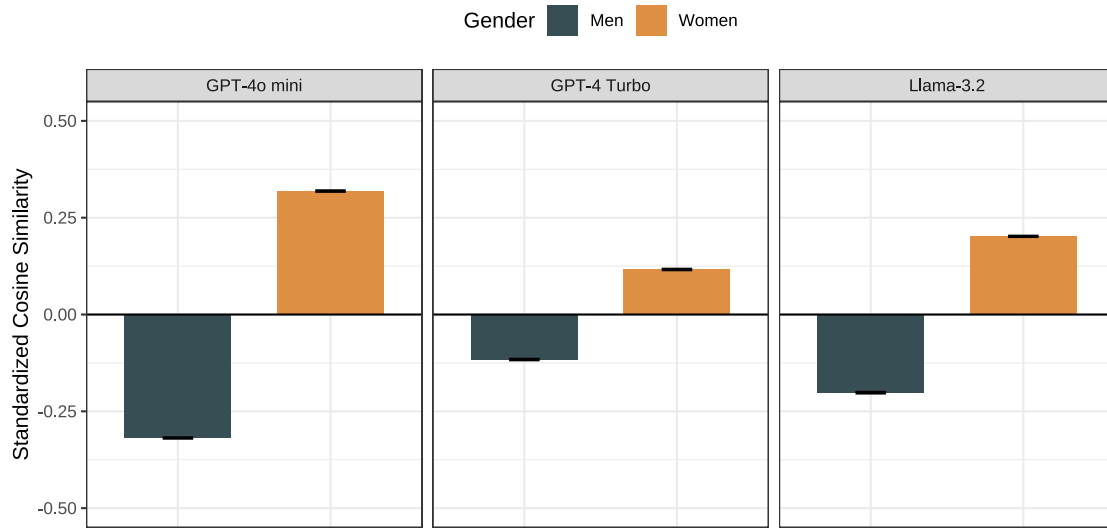


Figure 4: Standardized cosine similarity values of Black men and women. Higher standardized cosine similarity means more homogeneity in the stories generated for that group. Error bars represent one standard error above and below the mean.

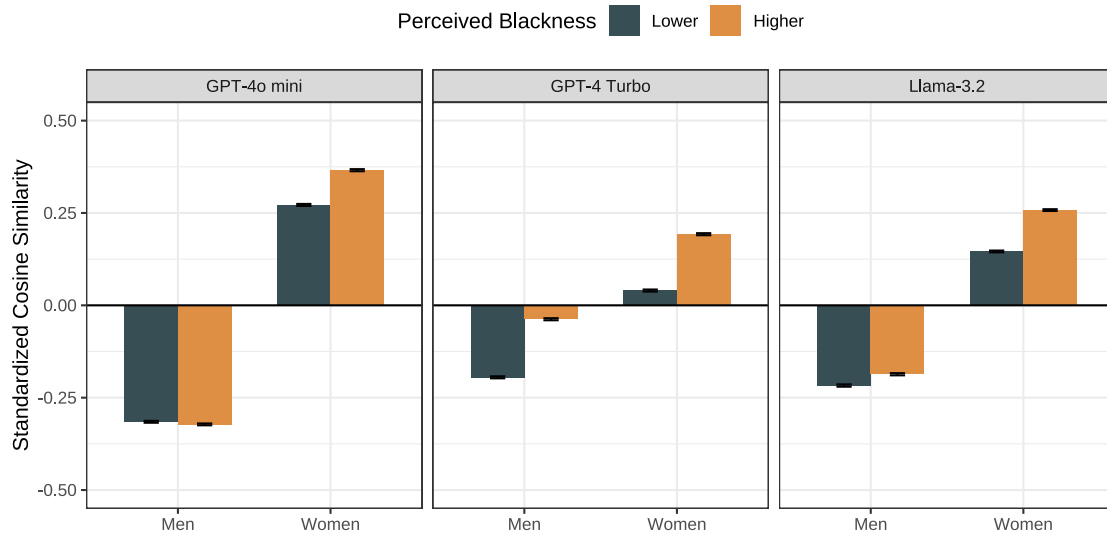


Figure 5: Standardized cosine similarity values of Black men and women with lower and higher phenotypicality. Higher standardized cosine similarity means more homogeneity in the stories generated for that group. Error bars represent one standard error above and below the mean.

4 Discussion

In this work, we expanded the study of bias in VLMs to within-group biases, examining how perceived racial phenotypicality affects stereotyping in VLMs. We find that VLMs generate more homogeneous content about Black individuals with higher phenotypicality compared to those with lower phenotypicality. This pattern indicates that the level of phenotypicality in visual inputs influences the diversity of content that VLMs generate, with higher phenotypicality resulting in less individualized representations. Notably, this pattern mirrors findings from social psychology research, where humans have been shown to perceive individuals with more phenotypically Black features in more stereotypical ways. Given that VLMs are primarily trained on web-scraped data, which contains human-created content reflecting existing social biases (see Bender et al. (2021) for a detailed review), these models appear to reproduce patterns of bias similar to those observed in human perception.

4.1 Convergent evidence of gender homogeneity bias

Consistent with past evidence finding that women were represented as more homogeneous relative to men in LLMs (Lee et al., 2024) and VLMs (Lee et al., 2025), we found evidence of gender homogeneity bias in VLMs. This work extends previous findings by Lee et al. (2025) through the use of computer-generated images that allow for more controlled experimentation, enabling us to isolate phenotypical features while maintaining other facial characteristics constant. Future work would benefit from systematic analysis of what parts of the model architecture would be most effective for targeting bias mitigation efforts in VLMs.

4.2 The disproportionate effect of phenotypicality on women

In two of three VLMs—GPT-4 Turbo and Llama-3.2—the effect of phenotypicality on homogeneity of group representations was significantly greater for women than for men. Upon closer inspection of the *Interaction Models*, including the interaction effect to the *Phenotypicality Models* rendered the effect of phenotypicality either insignificant or in the opposite direction, suggesting that the main effect of phenotypicality in the *Phenotypicality Models* were primarily driven by the effect of phenotypicality within Black women. This shares consistencies with human stereotyping patterns where phenotypicality disproportionately affects women Hill (2002). While Buolamwini and Gebru (2018) demonstrated intersectional bias in gender classification systems, our results demonstrate that similar biases persist in Vision-Language Models (VLMs), reinforcing the importance of intersectionality in the study of AI bias.

5 Limitations

While our approach provides quantitative evidence of homogeneity bias between Black individuals with differing degrees of racial phenotypicality, we acknowledge important limitations in the metric used. Although the embedding-based cosine similarity method we used is the current standard for semantic text comparison, it still functions largely as a black box. While we present examples in Table 1 to demonstrate face validity, there remains limited transparency regarding which textual features contribute to the measured similarities. Using our measure, we can’t quite determine if certain topics, such as those related to stereotypes, are more likely to emerge for Black individuals with higher phenotypicality than those with lower. Future work could explore homogeneity bias through alternative metrics examining specific linguistic features such as word overlap, syntactic structures, and topical content, though such approaches would come with their own methodological trade-offs. Nevertheless, the field would benefit from complementary measurement approaches to triangulate how AI systems manifest homogeneity bias across different demographic groups.

Another potential concern might be our use of computer-generated rather than real faces. However, this methodological choice represents a key strength of our approach. To isolate the specific effects of perceived racial phenotypicality among Black individuals, we needed to control for all other facial features that typically covary with phenotypicality in real-world faces—a control that would be nearly impossible to achieve with real faces. The GANFD images enabled precise manipulation of phenotypicality while keeping all other facial characteristics consistent, eliminating potential confounds. This allowed us to draw more definitive conclusions about how perceived racial phenotypicality influences AI-generated representations.

6 Conclusion

Our analysis demonstrates that Vision-Language Models (VLMs) exhibit homogeneity bias influenced by perceived racial phenotypicality. Using computer-generated images of Black American men and women with systematically varied phenotypicality, we found that VLMs generate more homogeneous content about individuals with higher

phenotypicity compared to those with lower phenotypicity. Our findings also reveal consistent gender disparities, with Black women represented more homogeneously than Black men across all models tested. Additionally, interaction analyses in some models showed that the effect of phenotypicity on content homogeneity was more pronounced for Black women than for Black men. These results extend our understanding of AI bias beyond traditional between-group comparisons, highlighting how within-group variations in perceived racial features influence the diversity of AI-generated representations. Our work underscores the importance of intersectionality in AI bias research and the need for more nuanced approaches to mitigate homogeneity bias in multimodal AI systems.

References

- Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. <https://doi.org/10.48550/arXiv.2101.05783> arXiv:2101.05783 [cs]
- Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. <https://www.anthropic.com/news/claude-3-7-sonnet>.
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting Linear Mixed-Effects Models Using Lme4. <https://doi.org/10.48550/arXiv.1406.5823> arXiv:1406.5823
- Douglas Bates, Martin Maechler, Ben Bolker [aut, cre, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Fabian Scheipl, Gabor Grothendieck, Peter Green, John Fox, Alexander Bauer, Pavel N. Krivitsky (shared copyright on simulate.formula), Emi Tanaka, and Mikael Jagan. 2024. Lme4: Linear Mixed-Effects Models Using ‘Eigen’ and S4.
- 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. <https://doi.org/10.48550/arXiv.2211.03759> arXiv:2211.03759
- 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.
- Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10853–10875. <https://doi.org/10.18653/v1/2023.emnlp-main.669>
- Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M. Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. <https://doi.org/10.48550/arXiv.2108.12084> arXiv:2108.12084
- Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On Measures of Biases and Harms in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 246–267. <https://doi.org/10.18653/v1/2022.findings-acl.24>
- Jillian Fisher, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W. Fisher, Jennifer Pan, Yulia Tsvetkov, and Katharina Reinecke. 2024. Biased AI Can Influence Political Decision-Making. <https://doi.org/10.48550/arXiv.2410.06415> arXiv:2410.06415
- L. Gleiberman, E. Harburg, M. R. Frone, M. Russell, and M. L. Cooper. 1995. Skin Colour, Measures of Socioeconomic Status, and Blood Pressure among Blacks in Erie County, NY. *Annals of Human Biology* 22, 1 (1995), 69–73. <https://doi.org/10.1080/03014469500003712>
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and et al. 2024. The Llama 3 Herd of Models. <https://doi.org/10.48550/arXiv.2407.21783>
- Peter Green, Catriona MacLeod, and Phillip Alday. 2023. Simr: Power Analysis for Generalised Linear Mixed Models by Simulation.
- Matthew Groh, Omar Badri, Roxana Daneshjou, Arash Koochek, Caleb Harris, Luis R. Soenksen, P. Murali Doraiswamy, and Rosalind Picard. 2024. Deep Learning-Aided Decision Support for Diagnosis of Skin Disease across Skin Tones. *Nature Medicine* 30, 2 (Feb. 2024), 573–583. <https://doi.org/10.1038/s41591-023-02728-3>

- Nao Hagiwara, Deborah A. Kashy, and Joseph Cesario. 2012. The Independent Effects of Skin Tone and Facial Features on Whites' Affective Reactions to Blacks. *Journal of Experimental Social Psychology* 48, 4 (July 2012), 892–898. <https://doi.org/10.1016/j.jesp.2012.02.001>
- Matthew S. Harrison and Kecia M. Thomas. 2009. The Hidden Prejudice in Selection: A Research Investigation on Skin Color Bias. *Journal of Applied Social Psychology* 39, 1 (2009), 134–168. <https://doi.org/10.1111/j.1559-1816.2008.00433.x>
- Mark E. Hill. 2002. Skin Color and the Perception of Attractiveness among African Americans: Does Gender Make a Difference? *Social Psychology Quarterly* 65, 1 (2002), 77–91. <https://doi.org/10.2307/3090169> arXiv:3090169
- Kimberly Barsamian Kahn and Paul G. Davies. 2011. Differentially Dangerous? Phenotypic Racial Stereotypicality Increases Implicit Bias among Ingroup and Outgroup Members. *Group Processes & Intergroup Relations* 14, 4 (2011), 569–580. <https://doi.org/10.1177/1368430210374609>
- Verna M. Keith and Cedric Herring. 1991. Skin Tone and Stratification in the Black Community. *Amer. J. Sociology* 97, 3 (1991), 760–778. arXiv:2781783
- Newton M. Kinyanjui, Timothy Odonga, Celia Cintas, Noel C. F. Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R. Varshney. 2019. Estimating Skin Tone and Effects on Classification Performance in Dermatology Datasets. <https://doi.org/10.48550/arXiv.1910.13268> arXiv:1910.13268 [cs]
- E. A. Klonoff and H. Landrine. 2000. Is Skin Color a Marker for Racial Discrimination? Explaining the Skin Color-Hypertension Relationship. *Journal of Behavioral Medicine* 23, 4 (Aug. 2000), 329–338. <https://doi.org/10.1023/a:1005580300128>
- 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. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1321–1340. <https://doi.org/10.1145/3630106.3658975>
- Messi H. J. Lee, Soyeon Jeon, Jacob M. Montgomery, and Calvin K. Lai. 2025. Visual Cues of Gender and Race Are Associated with Stereotyping in Vision-Language Models. <https://doi.org/10.48550/arXiv.2503.05093> arXiv:2503.05093 [cs]
- Russell V. Lenth, Ben Bolker, Paul Buerkner, Iago Giné-Vázquez, Maxime Herve, Maarten Jung, Jonathon Love, Fernando Miguez, Hannes Riebl, and Henrik Singmann. 2024. Emmeans: Estimated Marginal Means, Aka Least-Squares Means.
- Hauke Licht. 2023. Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings. *Political Analysis* 31, 3 (July 2023), 366–379. <https://doi.org/10.1017/pan.2022.29>
- Gechun Lin. 2025. Using Cross-Encoders to Measure the Similarity of Short Texts in Political Science. *American Journal of Political Science* n/a, n/a (March 2025), 1–17. <https://doi.org/10.1111/ajps.12956>
- Patricia W. Linville, Gregory W. Fischer, and Peter 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 (1989), 165–188. <https://doi.org/10.1037/0022-3514.57.2.165>
- Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, Nader Akoury, Faeze Brahman, Snigdha Chaturvedi, Elizabeth Clark, Mohit Iyyer, and Lara J. Martin (Eds.). Association for Computational Linguistics, Virtual, 48–55. <https://doi.org/10.18653/v1/2021.nuse-1.5>
- Keith B. Maddox. 2004. Perspectives on Racial Phenotypicity Bias. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc* 8, 4 (2004), 383–401. https://doi.org/10.1207/s15327957pspr0804_4
- Art D. Marsden, Alexandria Jaurique, Mackenzie L. McDonald, and Sara Emily Burke. 2024. GAN Face Database (GANFD).
- 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, Montréal QC Canada, 786–808. <https://doi.org/10.1145/3600211.3604711>
- José C. Pinheiro and Douglas M. Bates. 2000. Linear Mixed-Effects Models: Basic Concepts and Examples. In *Mixed-Effects Models in S and S-PLUS*. Springer, New York, NY, 3–56. https://doi.org/10.1007/0-387-22747-4_1

- George A. Quattrone and Edward E. Jones. 1980. The Perception of Variability within In-Groups and out-Groups: Implications for the Law of Small Numbers. *Journal of Personality and Social Psychology* 38, 1 (1980), 141–152. <https://doi.org/10.1037/0022-3514.38.1.141>
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. <https://doi.org/10.48550/arXiv.1908.10084> arXiv:1908.10084 [cs]
- 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>
- Henrik Singmann, Ben Bolker, Jake Westfall, Frederik Aust, Mattan S. Ben-Shachar, Søren Højsgaard, John Fox, Michael A. Lawrence, Ulf Mertens, Jonathon Love, Russell Lenth, and Rune Haubo Bojesen Christensen. 2024. Afex: Analysis of Factorial Experiments.
- Elena V. Stepanova and Michael J Strube. 2012. The Role of Skin Color and Facial Physiognomy in Racial Categorization: Moderation by Implicit Racial Attitudes. *Journal of Experimental Social Psychology* 48, 4 (2012), 867–878. <https://doi.org/10.1016/j.jesp.2012.02.019>
- Elena V. Stepanova and Michael J Strube. 2018. Attractiveness as a Function of Skin Tone and Facial Features: Evidence from Categorization Studies. *The Journal of General Psychology* 145, 1 (Jan. 2018), 1–20. <https://doi.org/10.1080/00221309.2017.1394811>
- 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>
- T. Joel Wade, Melanie Judkins Romano, and Leslie Blue. 2004. The Effect of African American Skin Color on Hiring Preferences. *Journal of Applied Social Psychology* 34, 12 (2004), 2550–2558. <https://doi.org/10.1111/j.1559-1816.2004.tb01991.x>
- Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S. Ryoo, and et al. 2024. xGen-MM (BLIP-3): A Family of Open Large Multimodal Models. <https://doi.org/10.48550/arXiv.2408.08872> arXiv:2408.08872
- 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.

S1 Appendix: GANFD Face Stimuli

Table S1: The identifiers of GANFD images used to represent the four groups.

Gender	phenotypicality	Image IDs
Men	Lighter	1407-752417, 2308-489, 2491-1250, 3490-822, 4239-28, 7902-1229, 14792-3566, 19187-28, 22913-28, 24490-28
Men	Darker	1407-3876, 2308-151, 2491-1407, 3490-3876, 4239-3876, 7902-38, 14792-533, 19187-533, 22913-533, 24490-533
Women	Lighter	1402-28, 2617-2947, 10571-4022, 12360-28, 13372-1119, 13571-3112, 16252-2157, 17235-4022, 19933-533, 25885-28
Women	Darker	1402-533, 2617-533, 10571-1407, 12360-533, 13372-533, 13571-3876, 16252-3876, 17235-3876, 19933-3876, 25885-115

S2 Appendix: Model Selection

Several models we initially collected data from, such as BLIP-3 (*xgen-mm-phi3-mini-instruct-r-v1* Xue et al., 2024) and Claude 3.7 Sonnet (Anthropic, 2025), were excluded from our analysis because they refused to generate stories based on facial images. BLIP-3 produced visual descriptions instead (e.g., "The woman in the image is a beautiful black woman with curly hair and dark brown eyes. She has a serious expression and is looking at the camera."), while Claude 3.7 Sonnet declined to process facial images altogether, responding with statements like: "I notice the image contains a human face. Following my guidelines, I won't identify or create a story about a specific individual in this photo. Instead, I can offer to write a brief fictional story about a character without referencing this specific image, or I could help with another creative request that doesn't involve identifying the person in this photograph."

S3 Appendix: Output of Mixed-Effect Models

Table S2: Summary output of the Phenotypicality Models across all four VLMs. In this model, lower phenotypicality was set as the reference level, with a significantly positive effect of phenotypicality indicating larger cosine similarity values for Black individuals with higher phenotypicality.

	Phenotypicality Model		
	GPT-4o mini	GPT-4 Turbo	Llama-3.2
Fixed Effects			
Intercept	-0.0037 (0.041)	-0.053 (0.025)	-0.027 (0.036)
phenotypicality	0.044*** (0.0026)	0.15*** (0.0027)	0.080*** (0.0036)
Random Effects (σ^2)			
Pair ID Intercept	0.18	0.068	0.21
Residual	0.82	0.93	0.80
Observations	499,000	499,000	499,000
Log likelihood	-657,971.70	-690,385.00	-651,912.20

Table S3: Summary output of the Gender Models across all four VLMs. In this model, men were set as the reference level, with a significantly positive gender effect indicating higher cosine similarity values for Black women compared to Black men.

	Gender Model		
	GPT-4o mini	GPT-4 Turbo	Llama-3.2
Fixed Effects			
Intercept	-0.30 (0.039)	-0.053 (0.025)	-0.18 (0.047)
Gender	0.63*** (0.055)	0.15*** (0.0027)	0.40*** (0.067)
Random Effects (σ^2)			
Pair ID Intercept	0.083	0.068	0.17
Residual	0.82	0.93	0.80
Observations	499,000	499,000	499,000
Log likelihood	-658,069.90	-690,385.00	-652,143.80

Table S4: Summary output of the Interaction Models across all four VLMs. In this model, the phenotypicality term represents the effect of phenotypicality for men (the reference gender group), the gender term represents the effect of gender for Black individuals with lower phenotypicality (the reference phenotypicality group), and the interaction term indicates how the effect of phenotypicality differs for women compared to men.

	Interaction Model		
	GPT-4o mini	GPT-4 Turbo	Llama-3.2
Fixed Effects			
Intercept	-0.29 (0.039)	-0.17 (0.032)	-0.17 (0.047)
phenotypicality	-0.0070 (0.0036)	0.16*** (0.0039)	-0.023*** (0.0051)
Gender	0.58*** (0.055)	0.23*** (0.045)	0.29*** (0.067)
Interactions	0.10*** (0.0051)	-0.0048 (0.0055)	0.21*** (0.0071)
Random Effects (σ^2)			
Pair ID Intercept	0.083	0.056	0.18
Residual	0.82	0.93	0.80
Observations	499,000	499,000	499,000
Log likelihood	-657,739.80	-690,379.70	-651,482.60

Table S5: Results of likelihood-ratio tests. Significant chi-square statistics indicate that including the corresponding terms significantly improves model fit, suggesting these factors have meaningful effects.

Model	Term	χ^2	<i>p</i>
GPT-4o mini	phenotypicality	289.55***	<.001
	Gender	87.52***	<.001
	Interaction	389.86***	<.001
GPT-4 Turbo	phenotypicality	3206.77***	<.001
	gender	22.86***	<.001
	Interaction	0.76	.38
Llama-3.2	phenotypicality	502.39***	<.001
	Gender	32.36***	<.001
	Interaction	838.42***	<.001

Table S6: The effect of phenotypicality within each gender group across all VLMs.

Model	Gender	Effect of phenotypicality	95% CI
GPT-4o mini	Men	-0.0038	[-0.058, 0.050]
	Women	0.040	[-0.014, 0.094]
GPT-4 Turbo	Men	-0.053	[-0.097, -0.0085]
	Women	0.10	[0.058, 0.15]
Llama-3.2	Men	-0.027	[-0.092, 0.038]
	Women	0.053	[-0.012, 0.12]