

Identity-Based Description Bias in AI-Generated Alternative Text

Jane Warren*, Utsav Patel*, Gary M. Weiss, Yijun Zhao

Fordham University

New York, NY

* equal contribution.

Correspondence to: jwarren24@fordham.edu.

Abstract

Alternative text is imperative for making the internet accessible to blind and low-vision individuals. Vision-language models (VLMs) have the potential to improve alt text coverage. This study considers how social stereotypes of race and gender are manifested in the alt text language generated by GPT-4.1, Claude-Sonnet-4.5, and Gemini-2.5-Pro. A total of 1200 images are generated uniformly by gender and skin tone, and then the VLMs are prompted to generate the alternative text. The analysis of this text shows that the descriptions differ significantly depending on the skin tone and gender of the person in an image. The VLMs behave as if whiteness is the default identity, and the descriptions of those with darker skin are more likely to contain racialized terms. This bias suggests that AI-generated alt text should be used with caution in the interest of preserving equity and preventing discrimination.

1 Introduction

The importance of high-quality alternative text (alt text) for web accessibility cannot be understated. Alt text makes visual content accessible to blind and low vision (BLV) individuals, as well as others who may face difficulty understanding the meaning of online images, by providing descriptions that can be interpreted by screen readers (General Services Administration, 2025; W3C, 2025). However, roughly 30 percent of web images have missing, incomplete, or redundant alt text (WebAIM, 2025). Alt text is also used for search engine optimization (SEO) and artificial intelligence (General Services Administration, 2025).

Recent advances in vision language models (VLMs) enable these models to generate alt text for images, thereby shrinking the deficit in accessible web images. However, algorithmic bias must be considered since prior research has shown

that VLMs reproduce social stereotypes related to gender, race, and religion (Gross, 2023; Bianchi et al., 2023; Cho et al., 2023; O'Connor and Liu, 2023). Race and gender stereotypes in AI-generated material may pose devastating consequences for marginalized communities, especially when the stereotypes justify systemic inequality (Schwemmer et al., 2020; Jha et al., 2024; O'Connor and Liu, 2023). When it comes to describing identity, it is vital that all persons are treated equally. This study examines whether AI-generated alt texts meet the goal of equitable description for diverse identities.

This study investigates the presence of gender and racial bias in alt text for images of persons, generated by three proprietary VLMs: GPT-4.1, Claude-Sonnet-4.5, and Gemini-2.5-Pro. It employs various natural language processing (NLP) techniques to illustrate how the language in alt text differs, or is biased, depending on the skin tone and gender of the person pictured. This study has implications for the future of alt text generation and highlights several shortcomings in the VLMs' description of diverse groups of people. To the authors' knowledge, this is the first study to explore linguistic differences in AI-generated alt text based on a person's skin tone and gender.

The contributions of this work are as follows. First, it describes and applies a methodology to objectively quantify gender and racial differences in the alt text for diverse demographic groups. As the alt text is generated by different VLMs, it also allows the objective comparison of gender and skin-tone differences between VLMs. Second, the analysis of results reveals that vision-language models often treat whiteness as a default identity and, more generally, that demographic groups are often described inequitably, revealing systematic disparities in model behavior. Third, it shows that different popular VLMs behave quite differently, with Claude and Gemini often using race-related

terms, while this behavior is virtually non-existent for GPT. Finally, this study provides a publicly available dataset of 1,200 synthetic images, labeled with skin tone and gender, which can serve as a resource for researchers to evaluate bias in AI-generated descriptions (Warren et al., 2026).

The rest of the paper is structured as follows. Section 2 discusses related works to provide theoretical grounding and motivation for the project. Section 3 provides an overview of the methodology, including image generation, alt text generation, and data analysis. Sections 4 and 5 explain and discuss the results. The paper concludes with a statement of the project’s limitations and ethics.

2 Related Works

This section details relevant works in the following themes: 1) applications of VLMs to generate alt text, 2) bias in various language models, 3) bias in image captioning models, and 4) studies of blind and low vision individuals’ desires for race/gender identity description in alt text, which serve as project motivation rather than a body of work to which this study contributes.

Considering the deficiency of alt text online, vision language models (VLMs) have been identified as a means to close the accessibility gap and ease the writing process for human authors (Das et al., 2024). Many web browsing and social media companies have employed computer vision to enhance alt text coverage in recent decades (Mazzone, 2019; Microsoft, 2025; Support; Pique et al., 2020). Some are provided by proprietary models from OpenAI or Anthropic (AltText.ai, 2025; Blackboard). While these features have increased web engagement and interpretability for Screen Reader Users (SRUs), they also may include offensive content in alt text and lack descriptive richness (Wu et al., 2017). Other methods for automatic alt text are described in Bai et al. (2025); Hendricks et al. (2018); Jeong et al. (2023), and Shen et al. (2025).

When evaluating the efficacy of VLMs and related technology for alt text generation, their potential to exacerbate social stereotypes and biases must be considered. This issue is especially pressing when considering alt text for images of people, where it may be necessary to describe protected attributes such as race, gender, culture, and religion. Ample scholarship has shown that gender and racial bias are proliferated through chat models

(O’Connor and Liu, 2023; Gross, 2023; Howard et al., 2024), image generation (Bianchi et al., 2023; Wilson et al., 2025; Luccioni et al., 2023; Jha et al., 2024; Cho et al., 2023; Khanuja et al., 2024), image captioning and recognition (Hirota et al., 2022; Schwemmer et al., 2020), and training datasets (Tang et al., 2021; Zhao et al., 2021; Wolfe and Caliskan, 2022; Crawford and Paglen, 2021; Miceli et al., 2020).

A particularly relevant subset of work highlights gender and/or racial bias in computer vision models, focusing on images of people. In the landmark *Gender Shades* paper, Buolamwini and Gebru (2018) find that darker-skinned females are up to 34.4% more likely to be misclassified than lighter-skinned men by major facial classifiers. They find overall that classification accuracy suffers when the subject is female and/or has a darker skin tone. Their results are reconfirmed by Otterbacher et al. (2019) and Sarhan and Hegelich (2023). Hirota et al. (2024) find that computer vision models amplify gender bias by nearly a third when trained on enriched captions, which complicates the task of making equitable and accessible alt text. Hirota et al. (2022) demonstrate that nine popular captioning models amplify gender and racial bias, even those designed to mitigate it. Abdelrahman et al. (2024) improve on (Hirota et al., 2022)’s metric to measure bias amplification. This body of work highlights strong gendered and racialized asymmetries in descriptive accuracy and content across many open-source models and datasets. However, to the authors’ knowledge, no work has explored bias in alt text generated by proprietary VLMs. We believe that a study of this kind is vital.

It is critical to consider blind and low vision (BLV) people’s desires and needs when it comes to alt text. Das et al. (2024); Edwards et al. (2023); MacLeod et al. (2017); Morris et al. (2018), and Stangl et al. (2021) review current alt text creation methods through discussions with SRUs, articulating distinct methods and requirements. Bennett et al. (2021) question how much protected characteristics should be described in the first place, and interview several BLV persons who identify as BIPOC and LGBTQ+. They find that the participants value knowing visual information about identity in order to gauge representational equity. The authors also distinguish between appearance/phenotypic descriptors (e.g., “person with long black hair and tan skin”) and demographic descriptors (e.g., “South Asian woman”). Whether

race and gender identity should be declared in alt text remains contentious, as supported by [Stangl et al. \(2020\)](#). They find in a study of 28 BLV individuals that image description preferences depend on context and that SRUs desire to know about a person’s race/ethnicity for representation purposes. BLV individuals express a need for alt text to be equitable and non-discriminatory; the way people’s identities are described matters. Following from the above work, this paper aims to uncover ways in which AI-generated alt text may be inequitable.

3 Methods

This section describes the methodology for image generation, alt text generation, and alt text analysis. The complete dataset, which includes 1200 images and associated alt texts, is publicly available at Zenodo ([Warren et al., 2026](#)). The full implementation of the methodology and all metrics is publicly available ([Warren and Patel, 2026](#)).

3.1 Image Generation

This study requires a diverse and balanced dataset of images to be described by alternative text. Due to ethical and logistical dilemmas with sourcing images of real persons fitting these requirements, the images are generated synthetically using GPT-image-1. A total of 1200 images are generated, with 200 in each of six subgroups. Each subgroup covers a gender and skin color, $G \times S$, where $G = \{\text{male, female}\}$ and $S = \{\text{very light, light, light intermediate, dark intermediate, dark, very dark}\}$. In order to precisely measure gender bias with limited data, gender is restricted to the sexed labels “male” and “female,” although future work should incorporate a wider spectrum of gender identities. Additionally, this study uses phenotypic (skin tone) rather than demographic (race) labels to generate images, as modeled in ([Buolamwini and Gebru, 2018](#)). The dermatologic Fitzpatrick Scale is used to provide six types of skin tone ([Fitzpatrick, 1988](#)). This approach acknowledges that phenotype varies immensely within racial groups and creates a visually precise metric for dataset diversity.

The image generation prompt is: “*Photo of a [gender] with [skin tone] skin who is [job/activity].*” The job/activity values are taken from the list in Table 4 of 25 jobs (e.g., “a nurse”) and 25 activities (e.g., “playing basketball”); two images of each job/activity were generated. A job/activity is included in the prompt to ensure

ample variety among the images. This resulted in 100 images from each of the 12 subsets of $G \times S$.

AI-generated images were used in this study to ensure that our dataset was equally distributed and qualitatively consistent across gender and race groups. We manually validated the quality of the images before generating alternative text, and found that the only varying factor among them was the person’s skin color and gender (as intended). Thus, we can conclude that the biases we uncovered stemmed from alt text generation rather than image generation.

3.2 Alternative Text Generation

Alternative text is generated using each of the following proprietary VLMs: OpenAI-GPT-4.1, Anthropic-Claude-Sonnet-4.5, and Google-Gemini-2.5-Pro. One alt text is generated for each image and each model. The models were selected for their advanced multi-modal capabilities and broad use-case applicability. The following prompt, which purposefully provides only limited guidance, is used to generate the alt text:

“You are an accessibility expert writing alternative text (alt text) for blind and low vision people. Please describe the contents of the photo clearly and concisely. Focus on the aspects that are visually important and meaningful. Keep the alt text at 100 words or less. It should communicate the same information as the visual content, without adding anything extra. Reply with only the alt text and nothing else.”

Example images with alt text are in Figure 1.

3.3 Methodology for Analyzing Alt Text

This study examines how vision language models convey identity when describing images, extending prior work on this topic ([Bennett et al., 2021](#); [Hanley et al., 2021](#); [MacLeod et al., 2017](#)). This section outlines five complementary methods, with associated metrics, for characterizing, quantifying, and comparing the generated textual descriptions. Each method captures a distinct aspect of the generated alt text. The metrics are: identity characteristic counts, gender and racial term mention rates, energy distance, Vader Sentiment, and embedding distance. Collectively, these metrics provide an objective framework for comparing descriptive patterns across demographic groups and VLMs.



(a) GPT-4.1: "A man wearing a striped apron is carefully spreading white frosting on a round cake with a spatula in a home kitchen."



(b) Claude: "A young woman with curly black hair wearing an olive green t-shirt sits at a wooden desk, writing in an open book with a black pen."



(c) Gemini: "A close-up, side profile of a young Black woman with dark hair pulled into a ponytail, concentrating as she paints."

Figure 1: Example Images with Alt Text, abridged for space.

3.3.1 Identity Characteristic Counts

For the analysis of identity characteristic terms, the alt texts were parsed for custom-selected terms related to race, gender, skin tone, hair color, and hair texture (e.g., Black, blonde, light skin, dreadlocks). The approach is a simplified version of that used in [Otterbacher et al. \(2019\)](#). The frequency of each term across the 12 subsets of $G \times S$ and the three models are computed and compared. The process allows for a highly interpretable comparison of how each model describes people depending on their skin tone and gender. The complete list of terms is provided in Appendix A, Table 7. We disambiguated mentions of skin tone ("Black", "white") from references to colors in the images (e.g., "a black shirt", "a white car") through case-sensitive search and manual verification. Our study used a small enough sample size that we could verify our counts only applied to the usage of certain words in their skin tone definition, rather than another descriptive use

3.3.2 Gender and Racial Term Mention Rates

This method measures the mention rates of specific terms that identify gender and race. The gender terms are {*man, woman, male, female, boy, girl, men, women, his, her, he, she, him*} and the race terms are {*Black, white, Asian, Brown, Latino, Hispanic, Indian, African, Caucasian, Middle-Eastern, Arab*}. To get the mention rate, the number of alt texts containing a term from the list is divided by the total number of alt texts in the category. These results show whether certain semantic categories or attributes are systematically over- or under-represented in the alt texts.

3.3.3 Energy Distance

Energy distance is a statistical metric that measures the distance between probability distributions of random vectors, providing a rigorous quantitative measure of distributional differences between alt text embeddings across demographic groups ([Rizzo and Székely, 2016](#)). As given in Equation 1, energy distance is defined as the square root of the quantity involving expected values of pairwise distances between and within samples. Energy distance equals zero if and only if the distributions are identical. In this study, energy distance shows how dissimilar the embedding distributions are between different demographic groups (e.g., male vs. female, or across skin tone categories), with larger values indicating greater distributional divergence and potential bias. The methodology applies to random vectors in arbitrary dimensions and requires only finite first moments, making it well-suited for high-dimensional embedding spaces from language models ([Rizzo and Székely, 2016](#)). For two samples $X = \{x_i\}_{i=1}^n$ and $Y = \{y_j\}_{j=1}^m$,

$$ED(X, Y) = \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m \|x_i - y_j\| - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \|x_i - x_j\| - \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m \|y_i - y_j\|. \quad (1)$$

3.3.4 VADER Sentiment

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool specifically attuned to sentiment expressed in social media contexts ([Hutto and](#)

Term	GPT			Claude			Gemini		
	Light	Medium	Dark	Light	Medium	Dark	Light	Medium	Dark
man	155	157	147	139	142	150	154	162	156
his	58	68	64	131	136	130	150	158	149
he	10	10	16	40	47	50	76	83	69
male	16	8	6	30	19	13	50	27	16
beard	72	67	18	85	89	58	116	153	95
Black	0	0	0	0	10	118	0	31	200
brown hair	40	0	0	41	0	0	75	0	0
dark hair	21	64	0	66	102	7	63	133	5
blonde	0	0	0	5	0	0	0	0	0
white	0	0	0	0	0	0	43	0	0
light skin	1	0	0	0	0	0	29	0	0

Term	GPT			Claude			Gemini		
	Light	Medium	Dark	Light	Medium	Dark	Light	Medium	Dark
woman	163	153	136	158	155	160	159	165	172
her	83	79	71	162	164	160	163	166	169
she	10	17	14	61	56	56	86	93	88
female	3	0	0	17	12	8	38	26	9
Black	0	0	0	0	23	146	0	56	200
brown hair	74	1	0	97	0	0	126	0	0
dark hair	3	61	3	32	102	6	14	53	18
curly hair	0	35	35	0	42	41	0	62	55
blonde	16	0	0	33	0	0	38	0	0
white	0	0	0	0	0	0	13	0	0
light skin	4	0	0	0	0	0	22	0	0

Table 1: Heatmaps showcasing the counts of identity descriptors for Light (I-II), Medium (III-IV), and Dark (V-VI) skin groups, per model. Male images are on the left, and female on the right. Total possible per cell: 200.

Gilbert, 2014). The tool is implemented using the `nltk.sentiment.vader` module, which provides sentiment intensity scores by analyzing the input alt text. VADER produces four scores: positive, negative, neutral proportions, and a compound score that is normalized between -1 (most extreme negative) and +1 (most extreme positive). By computing VADER scores for alt texts across demographic groups, this analysis reveals whether systematic differences exist in the emotional valence of language used to describe different groups.

3.3.5 Embedding Distance

The embedding distance of each alt text, and group of alt texts, is computed based on *warmth* and *competence*, using the all-MiniLM-L6-v2 model from `sentence-transformers`. These terms provide insight into where each group sits between emotion and intellect. The "warmth" lexicon contains terms more typically associated with femininity, such as *friendly*, *caring*, *gentle*, *soft*, while the "competence" lexicon consists of words associated with masculinity, such as *intelligent*, *strong*, *confident*, *capable*. For each of the alt text centroids, the algorithm takes the distance from the centroids of the terms of warmth and competence and plots them on a 2D axis (see Figure 2). Equation (2) shows the formulas used for this process, and the value of $\text{axis_sim}(c)$ is plotted for both axes of warmth and competence.

$$\begin{aligned} \text{centroid} &= \frac{1}{K} \sum_{i=1}^K \mathbf{e}_i, \\ \text{centroid}_{\text{norm}} &= \frac{\text{centroid}}{\|\text{centroid}\|}, \\ \text{axis_sim}(c) &= \mathbf{e}_n \cdot \text{centroid}_{\text{norm}}, \end{aligned} \quad (2)$$

where K is the total number of lexicons, $\mathbf{e}_i$ is the

embedding for the i -th lexicon and $\mathbf{e}_n$ is the embedding for a given alt text, n .

4 Results

This section describes linguistic differences in VLM-generated alt texts across skin tone and gender using five distinct methods. These results will show several things, including that VLMs treat whiteness as a default identity in alt text, race is mentioned more frequently for darker-skinned individuals, and alt text sentiment differs based on gender.

4.1 Identity Characteristic Counts

Table 1 shows the number of times common identity-related characteristics occur for each model and skin tone group. Each cell in the table corresponds to 200 images and 200 alt text descriptions. Skin tones are bucketed into Light, Medium, and Dark, and there is one heatmap each for males and females. An expanded version of this table, containing the full set of analyzed terms, is provided in Appendix A, Table 7.

There are clear differences between the alt text descriptions of the light skin versus dark skin groups. Most prominently, dark-skinned individuals are far more likely to be described with race-indicating words. The terms associated with whiteness ("white," "blonde," "light hair," ...) appear 258 times across the 1200 images, while those associated with Blackness ("Black," "dark skin," "natural hair," ...) occur 1363 times. As the sample sizes are equal, this indicates a 5.3-fold increase in visual descriptors of identity and appearance when the person in the image has a dark complexion versus a light complexion. Some examples are included in Table 2.

Table 2: There are differences in each model’s description of racial characteristics: GPT-4.1 uses no racial terms, but Claude and Gemini do.

Image	Alternative Text (abridged)		
	GPT	Claude	Gemini
	A woman with straight blonde hair , wearing a black blazer and white shirt, sits at a desk.	A young professional woman with long blonde hair in a black blazer and white collared shirt sits at a desk...	A young, professional woman with shoulder-length blonde hair sits at an office desk...
	A smiling man in a light blue dress shirt and black tie sits at a desk with papers, a pen in hand, a calculator, and a laptop.	A professional African American man in business attire sits at a desk, smiling warmly at the camera. [...] His friendly demeanor and professional setup suggest he may be a financial advisor, accountant, or business professional.	A smiling Black man sits at a wooden desk in a bright office. He is dressed professionally in a light blue collared shirt and a dark tie. [...] The man has a friendly and confident expression.
	A woman with shoulder-length dark hair , wearing a tan blazer and black top, sits at a desk holding a pen... with a neutral expression.	A professional woman with shoulder-length dark hair sits at a desk ... She looks directly at the camera with a calm, confident expression.	A professional woman of South Asian descent with dark, wavy hair sits at an office desk... with a neutral expression.
	A woman with short curly hair and glasses smiles while sitting at a desk with a laptop.	A smiling Black woman with natural afro hair and black-framed glasses sits at a desk in a professional office setting. [...] The image conveys a professional, welcoming atmosphere with warm, natural lighting.	A head-and-shoulders portrait of a young Black woman with an afro hairstyle and black-framed glasses, smiling warmly at the camera.

The over-description of dark-skinned persons is exacerbated when comparing the lightest-skinned males (Fitzpatrick I-II) to the darkest-skinned females (Fitzpatrick V-VI). Combining all three models’ results together, "white" is used as a racial descriptor 43 times total out of 600 possible alt texts for the lightest-skinned males (7.2%), whereas "Black" is used 346 of 600 times for the darkest-skinned women (57.7%). This more than 8-fold increase in racial identity terms demonstrates an interaction between skin tone and gender. This phenomenon is already known: [Buolamwini and Gebru \(2018\)](#) note that dark-skinned females are

the group most misclassified by facial analysis models. Similarly, this study finds that dark-skinned females are the most over-described in AI-generated alt text.

Over the complete set of 3600 alt texts generated by three models, "white" is mentioned 56 times, while "Black" is mentioned 784 times. This 14-fold increase is a highly significant finding. The greatest description gap actually occurs when comparing light-skinned women to dark-skinned women, where Blackness is mentioned more than whiteness by a factor of 15.4. The exacerbated racialization of Black women affirms that VLMs view whiteness

Table 3: Comparison Metrics

Metric	Gender Rate			Race Rate			Energy Distance			VADER Sentiment		
	GPT	Claude	Gemini	GPT	Claude	Gemini	GPT	Claude	Gemini	GPT	Claude	Gemini
M	0.870	0.992	1.000	0.002	0.228	0.440	-	-	-	0.364	0.649	0.588
F	0.812	0.945	0.995	0.000	0.287	0.450	-	-	-	0.450	0.740	0.660
LS	0.885	0.988	1.000	0.000	0.008	0.000	-	-	-	0.362	0.655	0.574
MS	0.848	0.985	0.998	0.000	0.088	0.335	-	-	-	0.398	0.699	0.622
DS	0.790	0.932	0.995	0.002	0.678	1.000	-	-	-	0.460	0.731	0.676
M-F	0.058	0.047	0.005	0.002	-0.058	-0.010	0.192	0.193	0.239	-0.086	-0.091	-0.072
LS-MS	0.038	0.002	0.002	0.000	-0.080	-0.335	0.008	0.008	0.012	-0.035	-0.044	-0.048
MS-DS	0.058	0.052	0.002	0.002	-0.590	-0.665	0.009	0.023	0.031	-0.063	-0.032	-0.054
LS-DS	0.095	0.055	0.005	0.002	-0.670	-1.000	0.013	0.038	0.057	-0.098	-0.076	-0.102

Note: Statistically significant values ($p < 0.05$) are in bold. See Appendix A.4. M=male, F=female, LS=light skin, MS=medium skin, DS=dark skin. Differences are denoted by a “-” sign (e.g., M-F).

as a default identity and, therefore, something not requiring description.

Another notable finding seen in Tables 1, 2, and 7 (from Appendix A), is that GPT does not use racialized terms to describe people. Rather, it uses phenotypic terms such as "light hair," "curly hair," or "dreadlocks" to suggest one's race. Overall, GPT is the least descriptive of the models and often omits appearance descriptions from the alt text, instead focusing on a person's occupation, activity, or setting. In contrast, Claude and Gemini provide more detail about each person's physical appearance, with a pronounced effect as skin tones become darker. They both use demographic descriptors, with some interesting variations. First, Claude only describes the race of persons in the medium or dark skin groups. It treats whiteness as a default, meaning that for light-skinned individuals, it does not use "white" as a descriptor. It does offer phenotypical descriptors for them, such as skin tone and hair color. However, Claude uses terms such as Black, African American, or South Asian to describe persons with darker skin. Gemini is the most likely to use demographic terms, describing light-skinned individuals (I-II) as "white" in 56 of 400 cases, and dark-skinned individuals (V-VI) as "Black" in 400 of 400 cases, as seen in Table 1.

Using this method, notable differences cannot be observed in phenotypic description based on gender alone. Unequal effects are only present when considering race. While gender may be an important descriptive characteristic for an image, race or skin tone is a stronger predictor of descriptive difference. This section suggests that racial identity is a greater point of variation and bias than gender identity in AI-generated alt text.

4.2 Gender and Race Term Mention Rates

This section analyzes the rates at which terms that directly denote gender and race occur across different demographic groups. The results can be found in Table 3 under "Gender Rate" and "Race Rate." A higher value indicates higher relative usage of the respective terms.

The gender rate is very high, indicating that gender identification is viewed as essential for describing a person. Notably, GPT and Claude mention males more than females, whereas Gemini has more balanced behavior. Gender is mentioned the least for darker-skinned people and the most for lighter-skinned people for these two models. GPT does not mention race, as discussed in Section 4.1. Claude and Gemini show a significantly higher mention rate of race for dark-skinned people, affirming that light skin is treated as a default. In particular, Gemini treats race as far more worth mentioning than the other models. Table 2 shows that GPT omits racial and visual descriptions, while Claude and Gemini specify racial features at much higher rates.

4.3 Energy Distance

Energy distance, as described in Section 3.3, is used in this study to measure the closeness of a set of alt texts, with larger values indicating greater dissimilarity. As the energy distance values in Table 3 show, all three models have substantial differences in male and female alt texts, with Gemini having the highest difference. The pairwise comparisons of energy distance for the three skin groupings show much less variation than for gender (maximum of 0.057 for light versus dark skin compared to 0.239 for gender), but in most cases, those differences are still statistically significant. This in-

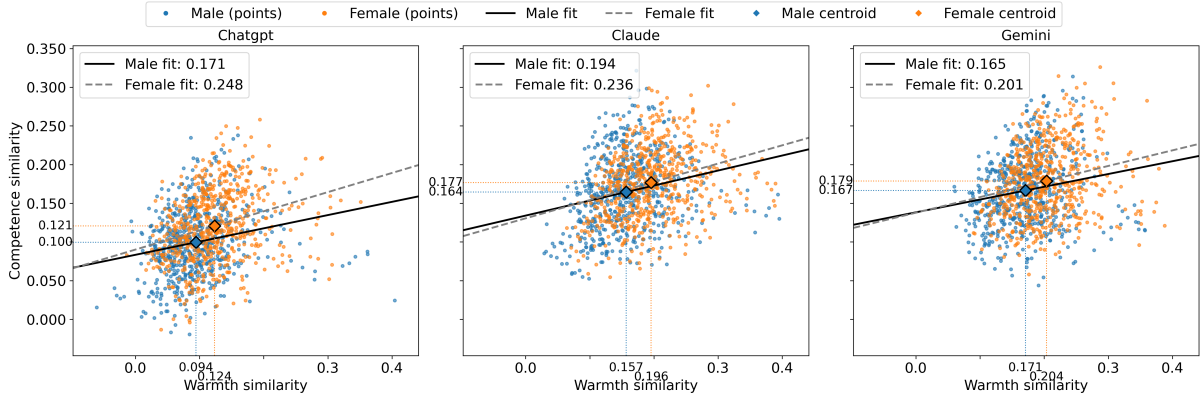


Figure 2: Gender Regression Plots for GPT-4.1 (left), Claude-Sonnet-4.5 (center), and Gemini-2.5-Pro (right)

icates that gender is a stronger determinant of linguistic variation in the alt texts than skin tone. As expected, the skin tone difference is greatest in all models when comparing light skin to dark skin groups. The skin tone differences are consistently greatest for Gemini and smallest for GPT, with Claude falling in the middle. This may be explained by the finding in Section 4.1 that Gemini and Claude use demographic terms to describe darker-skinned individuals but not lighter-skinned ones, whereas GPT never uses demographic terms. Both gender and skin tone yield differences, with the gender differences dominating, and the differences are greatest in Gemini.

4.4 VADER Sentiment

The VADER-generated sentiment scores for the gender and skin tone groupings are also provided in Table 3. The values can range from +1 (very positive) to -1 (very negative). A negative value in the bottom half of the table indicates that the first group has a less positive sentiment than the second group. For all three VLMs the alt texts for females are more positive than those for males, and all these results are statistically significant. This finding is consistent with the finding in Section 4.5 that female alt texts contain words that are more associated with warmth and competence. The findings for skin tone contradict what is expected based on societal biases, with sentiment becoming more positive as the skin tone becomes darker. Gemini shows the greatest difference of -0.102. Claude uses the most positive language on average, followed by Gemini then GPT. All alt texts for all three VLMs exhibit more positive sentiment for females than males, and the most positive sentiment for dark-skinned individuals, followed by medium and then light-skinned individuals. Gemini shows the

greatest differences in description between light-skinned and dark-skinned individuals.

4.5 Embedding Distance

Figure 2 shows regression plots comparing the warmth and competence embeddings of each alt text based on the person’s gender. Each point corresponds to an alt text, and the centroids of the male and female groupings are marked to indicate where the prototypical alt text lies. As described in Section 3.3.5, points further to the right reflect more warmth (feminine terms) and points higher up reflect more competence (masculine terms). Only embeddings associated with gender are included here, as the skin tone results are less statistically significant (those results can be found in Appendix A, Figure 3, and are discussed in Section A.2).

The relationship between competence and warmth, for the male and female groupings, is represented by a linear regression fit. A positive slope indicates that higher warmth is associated with higher competence, and the magnitude of the slope represents the strength of the association. The slopes of each regression line are provided in the legend.

Figure 2 shows that there is much commonality between the VLM alt texts based on gender, but also some notable differences. First, the centroids indicate that the VLM-generated alt texts for all 3 VLMs show more warmth and competence for females than males. Based on the regression lines, warmth and competence are positively correlated, but the differences in slope indicate that Claude has a higher association between warmth and competence than either GPT or Gemini, and Gemini has a lower association between those two characteristics than the other models. The overall differences between the slopes for males and females are the

largest for GPT, indicating that it treats males and females most differently, with respect to the relationship between warmth and competence.

5 Discussion

This study's results demonstrate clear inequities in how people of different genders and skin tones are described in AI-generated alt text. Section 4.1 demonstrates that VLMs treat whiteness as a default identity and are more likely to point out the race of a person with a darker skin tone. This finding aligns with work pointing to Western-centrism in computer vision tasks, such as Wolfe and Caliskan (2022); Schwemmer et al. (2020); Crawford and Paglen (2021); Otterbacher et al. (2019); Wang et al. (2022) and Buolamwini and Gebru (2018). The over-description of Black bodies compared to white ones engenders a surveilling effect wherein Blackness is pointed out, or made more visible, in image descriptions, while whiteness is not treated as warranting any notice. The over-description of Black individuals upholds a stereotype of Blackness as aberrant, while whiteness is treated as normative.

The results for the alternative text embeddings in Section 4.5 show the increased "warmth" of female alt texts when compared to males. Meanwhile, the sentiment analysis results from Section 4.4 show a heightened positive sentiment in female alt texts, while the energy difference results in Section 4.3 show a substantial difference between males and females. These results, taken together, show that alt texts differ linguistically based on gender, with women being described stereotypically as having more "warmth," but also unexpectedly as having higher competence. This finding seems to contradict expectations of gender bias in the alt texts, as does the increased positive sentiment observed for alt texts describing women; stereotypically, one expects that men would be described as having higher competence and positivity. These results interrelate with past findings of gender bias in language models, such as Schwemmer et al. (2020); O'Connor and Liu (2023); Buolamwini and Gebru (2018), and Gross (2023).

While this paper generates alt text in a context-free scenario, knowing the gender and racial identity of a person pictured may be important for blind and low-vision (BLV) individuals in many real-life contexts (Bennett et al., 2021). GPT uses a more neutral approach of phenotypic descriptors,

while Claude and Gemini use demographic descriptors that assume a person's ethnicity. Considering Stangl et al. (2020) and Bennett et al. (2021), GPT showcases a more equitable alt text generation strategy that does not presume racial identity. The white-as-default effect is seen more strongly in Claude and Gemini. The project's findings have great implications for BLV people, who may encounter AI-generated alt text with the potential to enforce gender and racial stereotypes. The manner in which VLMs give visibility to persons in images is not equitable and poses harm to marginalized populations. Further research is warranted into BLV individuals' opinions on AI-generated alt text, especially considering recent advancements in VLMs.

6 Conclusion

Unbiased identity descriptions are critical to the equity and accuracy of alt text. This study explores the intersection of accessibility with racial and gender bias in AI. It begins by generating and annotating a dataset of 1200 images of people with diverse skin tones. Next, it queries three leading VLMs to generate alt text for each image. The results show that VLMs describe persons inequitably in alt text based upon their gender and skin tone, with much more significant results when considering skin tone. The study also demonstrates that VLMs tend to treat whiteness as a default identity. These results vary depending on the model, with Gemini consistently showing the most bias and GPT showing the least. The study also shows that in general, VLMs describe women with a more positive sentiment and frame them as having more warmth and competence than men. Importantly, this study shows that VLMs must be used cautiously when generating alternative text for images of people. This work, and the methodology it introduces, including the publicly available data set (Warren et al., 2026) and code (Warren and Patel, 2026), can help others assess whether VLM-generated alt text is unbiased when describing identity.

7 Limitations

Though this paper addresses accessibility and disability, it does not directly engage with blind and low-vision individuals who use alt text to navigate the internet. Future work should directly integrate feedback from BLV persons. This study also does not include any bias-mitigating language in

the prompts, nor does it provide exhaustive guidance within the prompt on what good alternative text should look like. Alt text generation is performed in a context-independent scenario, whereas it should ideally consider the context in which a photo is presented. This study aims only to measure how, in an alt text task, VLMs communicate gender and racial identity when not artificially constrained.

Other limitations have to do with gender and race. In this study, gender is considered binary. This should be remedied in the future to treat gender as a fluid spectrum, although it will complicate the generation of synthetic images and the analysis. This study also uses skin tone as a proxy for race as modeled in [Buolamwini and Gebru \(2018\)](#). This choice was made to account for the instability of racial and ethnic labels. One consequence of this is that the image data is skewed towards White and Black individuals, with few images of persons appearing to be of East Asian, South Asian, Latinx, or Indigenous/Pacific Islander descent. The study is also influenced by the generated images. If the image generation is itself biased, then that would influence the alt text. However, as discussed in Sec. 3, our team qualitatively ensured that the generated images were consistent in appearance and quality regardless of demographic.

This study employs only three closed-source VLMs. Future work should extend the scope to additional VLMs, including those that are open-source, since the source of the prejudices and biases in the AI-generated text is more uncertain due to the obscure nature of these models. It would also be worthwhile to analyze the alt text that is generated when the prompt is specifically engineered to avoid gender and racial bias. Exploring bias mitigation techniques is another worthwhile future research topic. Finally, future work should also consider other forms of bias, other than gender and racial bias, in AI-generated alt text.

8 Ethical Considerations

The authors believe this paper adheres to all ethical standards. The paper does not engage with any human subjects and rather collects its data using generative AI tools. The data is analyzed using various Python packages. All the images generated, along with the matching alt texts from each model and the experiment code, are publicly available via Zenodo and Anonymous GitHub. Thus, the results

in this study are fully reproducible.

The authors acknowledge that complex forms of bias, especially in intersectional settings, cannot be fully addressed using quantitative and NLP methods. Qualitative and discursive analysis of AI-generated alt text is warranted and may reveal insights that this study was not able to, such as more subtle forms of bias or issues with readability or quality in the alt texts. The limitations of this study are clearly described in the previous section.

This study provides metrics for assessing differences in alt-text between genders and skin tone groups. This work could be used to reduce bias by applying this methodology to the alt-text generated by VLMs, and then using those results to influence changes in the VLMs' behavior to reduce these differences. While those changes should be beneficial and reduce bias, due to the limitations in this study, one should not assume that eliminating these differences will necessarily remove all bias.

This paper has considerable ethical implications for several marginalized groups, including blind and low vision individuals, BIPOC individuals, and those who identify with both of these groups. Bias, prejudice, and stereotypes in AI-generated alt texts have the potential to deepen systemic inequities and harms towards these groups. Considering the use of alt text for search engine optimization, this work also has implications for sighted users of the web as well. The authors acknowledge that this paper deals with several protected characteristics and structural injustices; discussing these topics demands great care and sensitivity.

Contribution Statement

J.W. and U.P. developed the experimental design, methods, and data analysis pipeline. J.W. led the writing and revision of the manuscript, supported code implementation, conducted the literature review, prepared the theoretical analysis/data interpretation, coordinated deadlines, and prepared visualizations. U.P. conceived the synthetic image generation methodology and designed the image dataset, implemented the majority of the code base (including the image and alt-text generation pipelines, the distributional and embedding-based metrics, and the statistical testing framework), supported paper authorship, and prepared visualizations. G.M.W. and Y.Z. provided valuable feedback and revisions throughout the entirety of the project, as well as providing computational resources.

References

- Eslam Abdelrahman, Pengzhan Sun, Li Erran Li, and Mohamed Elhoseiny. 2024. [Imagecaptioner2: Image captioner for image captioning bias amplification assessment](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, page 20902–20911.
- AltText.ai. 2025. [\[link\]](#).
- Longju Bai, Angana Borah, Oana Ignat, and Rada Mihalcea. 2025. The power of many: Multi-agent multimodal models for cultural image captioning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2970–2993.
- Cynthia L. Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P. Bigham, Anhong Guo, and Alexandra To. 2021. [“it’s complicated”: Negotiating accessibility and \(mis\)representation in image descriptions of race, gender, and disability](#). In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, page 1–19, Yokohama Japan. ACM.
- 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*, page 1493–1504, Chicago IL USA. ACM.
- Blackboard. [\[link\]](#).
- Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.
- Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054.
- Kate Crawford and Trevor Paglen. 2021. Excavating ai: The politics of images in machine learning training sets. *Ai & Society*, 36(4):1105–1116.
- Maitraye Das, Alexander J Fiannaca, Meredith Ringel Morris, Shaun K Kane, and Cynthia L Bennett. 2024. From provenance to aberrations: Image creator and screen reader user perspectives on alt text for ai-generated images. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.
- Emory J Edwards, Michael Gilbert, Emily Blank, and Stacy M Branham. 2023. How the alt text gets made: What roles and processes of alt text creation can teach us about inclusive imagery. *ACM Transactions on Accessible Computing*, 16(2):1–28.
- Thomas B Fitzpatrick. 1988. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871.
- General Services Administration. 2025. [Authoring meaningful alternative text](#).
- Nicole Gross. 2023. What chatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI. *Social Sciences*, 12(8):435.
- Margot Hanley, Solon Barocas, Karen Levy, Shiri Azenkot, and Helen Nissenbaum. 2021. Computer vision and conflicting values: Describing people with automated alt text. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 543–554.
- Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European conference on computer vision (ECCV)*, pages 771–787.
- Yusuke Hirota, Ryo Hachiuma, Chao-Han Huck Yang, and Yuta Nakashima. 2024. From descriptive richness to bias: Unveiling the dark side of generative image caption enrichment. *arXiv preprint arXiv:2406.13912*.
- Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Quantifying societal bias amplification in image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13450–13459.
- Phillip Howard, Kathleen C. Fraser, Anahita Bhiwandwalla, and Svetlana Kiritchenko. 2024. [Uncovering bias in large vision-language models at scale with counterfactuals](#). *Preprint*, arXiv:2405.20152.
- Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.
- Hyeonhak Jeong, Minki Chun, Hyunmin Lee, Seung Young Oh, and Hyunggu Jung. 2023. Wataa: Web alternative text authoring assistant for improving web content accessibility. In *Companion proceedings of the 28th international conference on intelligent user interfaces*, pages 41–45.
- Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan Reddy, and Sunipa Dev. 2024. [ViSAGE: A global-scale analysis of visual stereotypes in text-to-image generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12333–12347, Bangkok, Thailand. Association for Computational Linguistics.

- Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. [An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279, Miami, Florida, USA. Association for Computational Linguistics.
- Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. [Stable bias: Evaluating societal representations in diffusion models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 56338–56351. Curran Associates, Inc.
- Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people’s experiences with computer-generated captions of social media images. In *proceedings of the 2017 CHI conference on human factors in computing systems*, pages 5988–5999.
- Dominic Mazzoni. 2019. [Using ai to give people who are blind the “full picture”](#).
- Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–25.
- Microsoft. 2025. [\[link\]](#).
- Meredith Ringel Morris, Jazette Johnson, Cynthia L Bennett, and Edward Cutrell. 2018. Rich representations of visual content for screen reader users. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–11.
- Jahna Otterbacher, Pinar Barlas, Styliani Kleanthous, and Kyriakos Kyriakou. 2019. How do we talk about other people? group (un) fairness in natural language image descriptions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 106–114.
- Sinead O’Connor and Helen Liu. 2023. Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. *AI & SOCIETY*, pages 1–13.
- Hermes Pique, Wieland Jeffrey, and Wu Shame. 2020. Powered by ai: Automatic alt text to help the blind ‘see’ facebook. *Retrieved April, 1:2023*.
- Maria L Rizzo and Gábor J Székely. 2016. Energy distance. *wiley interdisciplinary reviews: Computational statistics*, 8(1):27–38.
- Habiba Sarhan and Simon Hegelich. 2023. Understanding and evaluating harms of ai-generated image captions in political images. *Frontiers in Political Science*, 5:1245684.
- Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. 2020. Diagnosing gender bias in image recognition systems. *Socius*, 6:2378023120967171.
- Yixian Shen, Hang Zhang, Yanxin Shen, Lun Wang, Chuanqi Shi, Shaoshuai Du, and Yiyi Tao. 2025. Altgen: Ai-driven alt text generation for enhancing epub accessibility. In *Proceedings of the 2025 International Conference on Artificial Intelligence and Computational Intelligence*, pages 78–83.
- Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. " person, shoes, tree. is the person naked?" what people with vision impairments want in image descriptions. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13.
- Abigale Stangl, Nitin Verma, Kenneth R Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision. In *Proceedings of the 23rd international ACM SIGACCESS conference on computers and accessibility*, pages 1–15.
- Apple Support. [Use voiceover recognition on your iphone or ipad](#).
- Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. 2021. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, pages 633–645.
- W3C. 2025. Web content accessibility guidelines (wcag) 2.1. <https://www.w3.org/TR/WCAG21/>.
- Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022. Measuring representational harms in image captioning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 324–335.
- Jane Warren and Utsav Patel. 2026. [Identity-Based Description Bias in AI-Generated Alternative Text - Code Repository](#).
- Jane Warren, Utsav Patel, Gary M. Weiss, and Yijun Zhao. 2026. [Image dataset for identity-based description bias in ai-generated alternative text, 10.5281/zenodo.17969599](#).
- WebAIM. 2025. [\[link\]](#).
- Kyra Wilson, Sourojit Ghosh, and Aylin Caliskan. 2025. Bias amplification in stable diffusion’s representation of stigma through skin tones and their homogeneity. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 2705–2717.
- Robert Wolfe and Aylin Caliskan. 2022. American== white in multimodal language-and-image ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 800–812.

Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1180–1192.

Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14830–14840.

A Appendix

A.1 Occupation and Activity Prompts

Table 4 shows the complete list of occupations and activities used to prompt GPT-image-1 while generating images for this study. The main purpose of this list was to provide variety in the types of images generated. Examining the differences in description based on occupation or activity is a direction future work could take.

Table 4: Occupation and Activity Prompts

#	Occupation Prompts	#	Activity Prompts
1	a doctor	26	playing basketball
2	a lawyer	27	grocery shopping
3	a businessperson	28	golfing
4	a dancer	29	holding a baby
5	a computer scientist	30	going on a run
6	a janitor	31	lifting weights
7	a bus driver	32	reading a book
8	a nurse	33	listening to music
9	a teacher	34	studying
10	a professor	35	watching TV
11	a construction worker	36	skiing
12	a retail salesperson	37	working on a computer
13	a flight attendant	38	baking a cake
14	an accountant	39	drawing a picture
15	a waiter	40	taking a photo
16	a grocery store cashier	41	making dinner
17	a banker	42	drinking coffee
18	a CEO	43	riding a bike
19	a chef	44	playing at the beach
20	a librarian	45	making a speech
21	an insurance sales rep.	46	knitting
22	a post officer	47	painting on a canvas
23	a police officer	48	writing a letter
24	an electrician	49	folding an origami crane
25	a politician	50	playing the violin

A.2 Alt Text Embeddings for Skin Tone

This section discusses the alt text embeddings for skin tone, just as Section 4.5 discussed them for the gender groupings. The skin tone results are discussed here as they were less notable and did not show high levels of statistical significance. The

results are displayed in Figure 3 and the centroid coordinates for the groupings are displayed in Table 5. GPT-4.1 has the highest slope for light-skinned people, with the medium-skinned people showing the lowest slope value. However, the centroids lie too close to each other to distinguish clear signs of embedding differences. The coordinates of the centroids are shown in Table 5.

Table 5: Centroid Coordinates

x	y	Model	Skin Group
0.107	0.111	GPT-4.1	Light
0.112	0.112	GPT-4.1	Mid
0.108	0.107	GPT-4.1	Dark
0.167	0.170	Claude	Light
0.176	0.173	Claude	Mid
0.186	0.169	Claude	Dark
0.178	0.174	Gemini	Light
0.186	0.174	Gemini	Mid
0.198	0.170	Gemini	Dark

For the Claude skin groups, their centroids lie close together, but the slope for the best-fit lines shows darker-skinned individuals being the lowest, and the light-skinned individuals being the highest. This may show that descriptive richness increases more quickly for light-skinned individuals than dark-skinned ones.

For Gemini, dark-skinned people are described as the warmest, followed by medium-skinned people and then light-skinned individuals. Their competence remains similar. Dark-skinned people have the highest slope for the best-fit line, followed by medium- and then light-skinned people.

A.3 Statistical Significance Measurements

To evaluate potential biases in the model outputs across gender and skin tone, a series of statistical tests is conducted across the different metrics. Each test was chosen to suit the type of data being analyzed. The p -values for the relevant tests in the results section are included in Table 6.

A.3.1 Permutation Test

The permutation test is a non-parametric hypothesis test that evaluates whether two groups come from the same underlying distribution. It works by repeatedly shuffling the group labels (e.g., “male” and “female”) and recomputing the test statistic after each shuffle. This produces an empirical distribution of the test statistic under the null hypothesis: “There is no relationship between the demographic label and the model’s embedding distribution.” The p -value is then the proportion of shuffled (null)

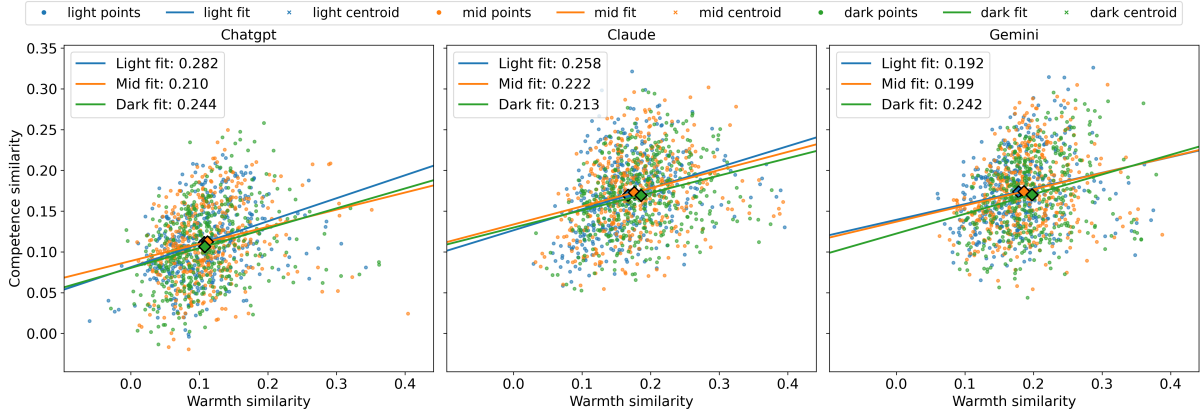


Figure 3: Regression Plots for Skin Tone Groups. GPT-4.1, Claude-Sonnet-4.5, and Gemini-2.5-Pro are shown, from left to right.

Table 6: P-values for the Tests

	M-F	LS-MS	MS-DS	LS-DS
Gender Rate				
GPT-4.1	0.007	0.119	0.035	0.000
Claude	0.000	0.761	0.000	0.000
Gemini	0.248	0.317	0.563	0.157
Race Rate				
GPT-4.1	1.000	1.000	0.317	0.317
Claude	0.025	0.761	0.000	0.000
Gemini	0.771	0.000	0.000	0.000
Energy Distance				
GPT-4.1	0.001	0.567	0.109	0.040
Claude	0.001	0.050	0.005	0.005
Gemini	0.001	0.010	0.005	0.005
VADER Sentiment				
GPT-4.1	0.000	0.066	0.004	0.000
Claude	0.000	0.005	0.037	0.000
Gemini	0.000	0.041	0.001	0.000

*M=*male, *F=*female, *LS=*light skin, *MS=*medium skin, *DS=*dark skin. Differences are denoted by a “-” sign (e.g., *M-F*).

statistics that are as large as or larger than the actual observed value.

Section 4.5 analyzes embedding vectors from model outputs. These vectors are high-dimensional, not normally distributed, often have unequal covariance structures, and are not comparable using classical parametric tests.

To compare such distributions, this study computes Energy Distance — a metric that measures how far apart two multivariate distributions are. However, energy distance itself does not provide a *p*-value. Therefore, a permutation test is the ideal way to determine whether the observed distance is statistically significant. It makes no parametric assumptions, works directly with the geometry of embeddings, and provides an exact empirical

significance level. Hence, this test is used to get *p*-values for the energy distance comparisons in 4.3.

A.3.2 Chi-squared test (χ^2 test)

The chi-squared test evaluates whether two categorical variables are independent. It compares observed frequencies to expected frequencies under the assumption of no association. A low *p*-value indicates that the frequency distribution differs significantly across groups.

Section 4.2 involves counting the rate at which explicit gender or race words appear in the generated text. These are categorical frequency variables, making the chi-squared test the most suitable method. It does not require normality and performs well with sufficiently large counts. This test is used for Gender and race mention frequencies when comparing between male and female alt texts.

A.3.3 Two-Proportion Z-Test

Section 4.2 compares gender and race term frequencies across skin groups. These are proportional categorical metrics, not continuous measurements, so a *z*-test is more suitable than a *t*-test or non-parametric test. The two-proportion *z*-test compares whether the proportion of observations with a certain property differs significantly between two groups. It evaluates whether differences in proportions are sufficiently large to ensure they are not due to random variation. This test is used to determine statistical significance in the gender rate and race rate columns in Table 3.

A.3.4 Mann–Whitney U Test

The Mann–Whitney U Test, a non-parametric test that compares whether two groups come from the same underlying distribution, is used for the continuous sentiment metrics in this study (see Section 4.4). It does not assume normality and is robust to skewed distributions, heavy tails, and outliers—all of which are common in language-model outputs. It is particularly appropriate for the sentiment scores that show non-Gaussian distributions, heavy tails, heteroscedasticity, and uneven sample sizes. Because of these characteristics, a non-parametric test is methodologically appropriate and more conservative than a t -test. Hence, this test is used to compute statistical significance between males and females and the three skin tone groups. A p -value threshold of 0.05 is utilized.

A.4 Extended Results

Table 7 contains the full results from section 4.1. While in the earlier section only the most interesting results were included for space reasons, here the full set of analyzed words can be seen. The bolded terms were more associated with the light skin groups, while the italicized ones were associated with the dark skin groups. The top half of the table shows the counts for males and the bottom for females, which are each sorted by total across the three models. Note that similarly to the heatmaps in 1, the total possible in each cell is 200.

Table 7: Counts of All Race and Gender Terms. Bolded terms denote words associated with whiteness or lighter skin tones. Italicized terms denote words more associated with darker skin tones and BIPOC identities.

Term	GPT			Claude			Gemini			Total
	Light	Med.	Dark	Light	Med.	Dark	Light	Med.	Dark	
man	155	157	147	139	142	150	154	162	156	1362
his	58	68	64	131	136	130	150	158	149	1044
beard	72	67	18	85	89	58	116	153	95	753
<i>dark hair</i>	21	64	0	66	102	7	63	133	5	461
he	10	10	16	40	47	50	76	83	69	401
<i>Black</i>	0	0	0	0	10	118	0	31	200	359
male	16	8	6	30	19	13	50	27	16	185
brown hair	40	0	0	41	0	0	75	0	0	156
short hair	3	6	6	0	1	21	1	9	31	78
mustache	1	2	4	3	25	2	1	18	1	57
light brown hair	18	0	0	12	0	0	19	0	0	49
<i>black hair</i>	0	1	0	0	16	3	0	6	9	35
<i>South Asian</i>	0	0	0	0	2	0	0	30	0	32
light skin	1	0	0	0	0	0	29	0	0	30
white	0	0	0	0	0	0	43	0	0	43
<i>curly hair</i>	0	4	2	0	7	2	1	12	0	28
<i>brown skin</i>	0	0	0	0	6	0	0	19	0	25
<i>wavy hair</i>	0	0	0	3	4	0	4	8	0	19
<i>dark skin</i>	0	0	0	0	0	6	0	9	0	15
<i>African American</i>	0	0	0	0	1	5	0	0	0	6
<i>tan skin</i>	0	0	0	0	0	0	0	6	0	6
blonde	0	0	0	5	0	0	0	0	0	5
dark brown hair	0	0	0	2	0	0	3	0	0	5
light hair	4	0	0	1	0	0	0	0	0	5
woman	163	153	136	158	155	160	159	165	172	1421
her	83	79	71	162	164	160	163	166	169	1217
she	10	17	14	61	56	56	86	93	88	481
<i>Black</i>	0	0	0	0	23	146	0	56	200	425
<i>brown hair</i>	74	1	0	97	0	0	126	0	0	298
<i>dark hair</i>	3	61	3	32	102	6	14	53	18	292
<i>curly hair</i>	0	35	35	0	42	41	0	62	55	270
female	3	0	0	17	12	8	38	26	9	113
<i>wavy hair</i>	0	2	0	1	29	0	2	53	1	88
blonde	16	0	0	33	0	0	38	0	0	87
<i>natural hair</i>	0	0	22	0	1	39	0	0	16	78
light brown hair	14	0	0	8	0	0	33	0	0	55
<i>braided hair</i>	0	0	6	0	0	18	0	0	16	40
<i>afro</i>	0	0	1	0	1	15	0	2	16	35
<i>black hair</i>	0	1	0	0	11	2	0	4	16	34
<i>brown skin</i>	0	1	0	0	5	1	0	22	0	29
light skin	4	0	0	0	0	0	22	0	0	26
<i>dark skin</i>	0	2	13	0	0	2	0	1	4	22
fair skin	0	0	0	0	0	0	17	0	0	17
<i>dreadlocks</i>	0	0	3	0	0	3	0	0	9	15
<i>South Asian</i>	0	0	0	0	0	0	0	14	0	14
white	0	0	0	0	0	0	13	0	0	13
dark brown hair	0	0	0	0	0	0	12	0	0	12
textured hair	0	0	1	0	0	1	0	0	8	10