
VISUAL CUES OF GENDER AND RACE ARE ASSOCIATED WITH STEREOTYPING IN VISION-LANGUAGE MODELS

Messi H.J. Lee, Soyeon Jeon, Jacob M. Montgomery

Washington University in St. Louis
St. Louis, Missouri 63130

hojunlee@wustl.edu, j.soyeon@wustl.edu, jacob.montgomery@wustl.edu

Calvin K. Lai

Rutgers University
New Brunswick, NJ 08901
calvin.lai@rutgers.edu

March 10, 2025

ABSTRACT

Current research on bias in Vision Language Models (VLMs) has important limitations: it is focused exclusively on trait associations while ignoring other forms of stereotyping, it examines specific contexts where biases are expected to appear, and it conceptualizes social categories like race and gender as binary, ignoring the multifaceted nature of these identities. Using standardized facial images that vary in prototypicality, we test four VLMs for both trait associations and homogeneity bias in open-ended contexts. We find that VLMs consistently generate more uniform stories for women compared to men, with people who are more gender prototypical in appearance being represented more uniformly. By contrast, VLMs represent White Americans more uniformly than Black Americans. Unlike with gender prototypicality, race prototypicality was not related to stronger uniformity. In terms of trait associations, we find limited evidence of stereotyping—Black Americans were consistently linked with basketball across all models, while other racial associations (i.e., art, healthcare, appearance) varied by specific VLM. These findings demonstrate that VLM stereotyping manifests in ways that go beyond simple group membership, suggesting that conventional bias mitigation strategies may be insufficient to address VLM stereotyping and that homogeneity bias persists even when trait associations are less apparent in model outputs.

1 Introduction

Large Language Models (LLMs) that are trained on vast amounts of text have demonstrated remarkable capabilities in natural language understanding (e.g., sentiment analysis, text classification), reasoning, and natural language generation (e.g., translation, question-answering). LLMs exhibit strong in-context learning (ICL) ability, quickly adapting to new tasks with few examples [Brown et al., 2020, Wei et al., 2022, Dong et al., 2023], leading to impressive performance across a variety of downstream tasks. Building on LLMs, Vision-Language Models (VLMs) introduce a new sensory modality by integrating visual and text information. Trained on large datasets of image-text pairs, VLMs learn the relationship between the two modalities in a shared embedding space [Radford et al., 2021, Li et al., 2023, Wang et al., 2022, Yu et al., 2022]. However, as VLMs bridge two modalities, they may not only reproduce existing biases present in each modality but also introduce new ones unique to this dual modality [OpenAI, 2023].

Despite growing attention to bias in VLMs, past work has important limitations. Studies focus exclusively on trait associations while ignoring other forms of stereotyping [e.g., Sheng et al., 2019, Lucy and Bamman, 2021]. They also examine specific contexts where biases are expected to appear (e.g., occupations) conceptualize social categories like race and gender as discrete, ignoring nuances in how those social categories are expressed in real life. In this paper,

we address these limitations. First, we examine both trait associations and homogeneity bias [Lee et al., 2024]—the tendency to represent certain groups as more similar to each other than others—a form of bias understudied in VLMs. Second, we explore stereotyping in an open-ended context where models generate stories without specific instructions related to known stereotype-prone domains. Third, we investigate how racial and gender prototypicality—the degree to which an individual’s physical features are representative of the stereotypical characteristics of their group—influences these biases, moving beyond binary conceptualizations of identity.

We find that all four VLMs tested in this work exhibit gender homogeneity bias, generating more uniform stories for women than for men, with uniformity increasing as gender prototypicality increases. We also find that VLMs tend to represent the majority racial group—White Americans—as more uniform than Black Americans. However, we do not detect effects of racial prototypicality on racial homogeneity bias. Additionally, the only consistent differences we find in trait associations relate to race, where Black Americans are disproportionately and consistently associated with basketball, a form of positive stereotyping about Blackness and athleticism that can also have unintended negative consequences. These findings underscore the complex ways visual cues shape bias expression in VLMs, presenting unique challenges for bias mitigation and fair representation in AI systems.

1.1 Two Forms of Stereotyping

Trait association refers to the belief that certain groups are differentially associated with specific traits or occupations (e.g., women as nurses and Asians as engineers). To systematically understand the associations involved in perceiving individuals based on group membership, social psychologists have proposed models such as the Stereotype Content Model [Fiske et al., 2002] and the ABC model of stereotype content [Koch et al., 2016], which highlight distinct dimensions of stereotypes.

Research on Natural Language Processing (NLP) systems has predominantly focused on trait associations. Studies on word embedding models [Garg et al., 2018, Caliskan et al., 2017], sentence encoders [Nadeem et al., 2021, May et al., 2019], and text generative models [Sheng et al., 2019, Lucy and Bamman, 2021] have extensively documented gender and racial trait associations, establishing this as the primary lens through which algorithmic bias is understood.

However, homogeneity bias is another critical form of stereotyping that remains relatively understudied in the VLM literature. This bias stems from a part of the stereotyping literature in social psychology, which focuses on perceived variability and documents how members of different groups are perceived heterogeneously or homogeneously [Linville et al., 1986]. Historically, research on perceived variability consistently found that individuals tend to perceive their outgroup as more homogeneous than their ingroup [Judd et al., 1991, Mullen and Hu, 1989, Linville et al., 1989]. Subsequent research has examined perceived variability in terms of group status and power, finding both culturally dominant and subordinate groups see culturally subordinate groups as more homogeneous than culturally dominant groups [Guinote et al., 2002, Lorenzi-Cioldi et al., 1998, Fiske, 1993].

In the context of language models, Lee et al. [2024] and Cheng et al. [2023a] have documented homogeneity bias in LLMs, demonstrating that these models portray marginalized groups as more homogeneous than their dominant group counterparts. This bias risks neglecting the diversity and richness of minority group identities, reinforcing prejudice and discrimination. While homogeneity bias has been examined in relation to skin tone [Lee and Jeon, 2024], its broader implications for racial and gender representations in VLMs remain largely unexplored.

1.2 Bias Assessment in Open-Ended Contexts

Another limitation of existing approaches is their focus on artificial contexts designed specifically to elicit biases. Current methods examine scenarios where biases are expected to appear: asking models to generate images for specific occupations [e.g., doctors, software engineers; Bianchi et al., 2023, Naik and Nushi, 2023, Sun et al., 2023, Sami et al., 2023] and analyzing demographic distributions in the resulting images. Alternative studies provide question-image pairs or captioning templates to quantify VLM completions that reflect societal stereotypes [Zhou et al., 2022, Ruggeri and Nozza, 2023]. While these approaches enable direct bias assessment, they lack ecological validity since VLMs are rarely deployed in such constrained contexts. In real-world applications, VLMs are used in open-ended conversational scenarios where users request information or ask questions with visual inputs. To develop a more comprehensive understanding of bias in VLMs, work should examine whether models express stereotypes during these naturalistic interactions, complementing the insights gained from more controlled experimental settings.

1.3 Prototypicality and Stereotyping

Previous research on AI bias has primarily focused on group identities that are operationalized dichotomously, using categorical labels (e.g., man, woman), names (e.g., Emily, Greg), or images that represent group categories. This

dichotomous approach, while informative, undermines the nuanced nature of social categories, which often exist on a continuum rather than as discrete categories. In reality, individuals vary in prototypicality: the degree to which an individual’s features are representative of the stereotypical characteristics of their group. People can appear more or less prototypically female, male, Black, or White, and these gradations may significantly influence how VLMs express stereotypes. The vision modality of VLMs provides a unique opportunity to examine these effects in ways that text-only approaches cannot capture.

Psychological research has extensively documented that prototypicality is linked to greater stereotyping [e.g., Ma et al., 2018, Livingston and Brewer, 2002, Blair et al., 2002, Maddox and Gray, 2002, Anderson and Cromwell, 1977]. For example, Maddox and Gray [2002] found that participants listed more Black-stereotypic traits for darker-skinned than lighter-skinned Black individuals, suggesting that more prototypical faces evoke stronger category judgments, leading to stronger trait associations and less perceived variability. While this relationship between prototypicality and stereotyping is well-established in psychology, we know little about how visual cues like skin tone, facial features, and hairstyle shape VLM stereotyping.

We propose two competing hypotheses about how prototypicality might influence VLM stereotyping. On one hand, VLMs may reproduce human patterns of prototypicality-based stereotyping, with higher prototypicality generally associated with increased stereotyping. This prediction is consistent with recent work by Lee and Jeon [2024], which found that VLMs represent darker-skinned Black individuals more homogeneously than lighter-skinned Black individuals, suggesting that skin tone, a significant contributor to perceived racial prototypicality [Strom et al., 2012, Wilkins et al., 2010], is linked to greater VLM stereotyping. Alternatively, VLMs may fail to accurately process prototypicality due to limitations in their training data and differences in how they process features within images. Training datasets may not adequately represent or label the nuanced physical characteristics that humans use to determine prototypicality, leading models’ perceptions to diverge from humans. Past research has demonstrated that VLMs extract and prioritize other visual features differently than humans do already [Geirhos et al., 2022, 2020, Baker et al., 2018].

1.4 Our Work

Our research aims to advance the understanding of VLM stereotyping by addressing three critical limitations in the current literature. First, we examine not only trait associations but also homogeneity bias—a form of stereotyping that remains understudied in VLMs. Second, we explore these biases in open-ended contexts rather than in domains where stereotypes are expected to appear. Third, we investigate how racial and gender prototypicality—the degree to which an individual’s features are representative of stereotypical group characteristics—influences VLM stereotyping. By examining how variations in visual cues relate to VLM stereotyping, we provide insights into whether these models replicate established patterns from social cognition where prototypicality significantly influences stereotyping.

We used four VLMs to generate open-ended stories in response to images of faces from four groups—Black men, Black women, White men, and White women—with varying racial and gender prototypicality. To assess homogeneity bias, we adopted the measurement strategy used by Lee et al. [2024], which involves measuring pairwise similarity between sentence embeddings of the generated texts. In addition, we used structural topic models [STMs; Roberts et al., 2019] to identify and compare commonly occurring traits and attributes in the VLM-generated stories.

Our research centers on two hypotheses and one research question. First, we hypothesized that subordinate racial and gender groups are subject to greater VLM stereotyping, independent of prototypicality. Second, we posit that VLM stereotyping is linked to how closely an individual’s facial features align with their group’s stereotypical characteristics, regardless of group identity. Additionally, we examine whether the relationship between VLM stereotyping and prototypicality matters more for some groups compared to others.

2 Method

2.1 Facial Stimuli

We conducted our experiment using the Racially Diverse Affective Expression (RADIATE) face stimulus set, a standardized database that features diverse racial groups and facial expressions [Conley et al., 2018, Tottenham et al., 2009]. From RADIATE, we randomly sampled 40 images total, consisting of 10 images from each of four intersectional groups: Black men, Black women, White men, and White women. All selected images had neutral facial expressions with closed mouths and were standardized for background, lighting, and pose (see Figure 1 for sample faces).¹

¹While RADIATE includes faces with various expressions from smiling to sad, we selected only neutral expressions to control for potential confounding effects of facial expression.



Figure 1: Sample RADIATE facial stimuli showing the lowest-rated (left) and highest-rated (right) faces for racial prototypicality within each demographic category: Black men, Black women, White men, and White women.

While RADIATE provides standardized facial stimuli, it does not include ratings for racial and gender prototypicality. Therefore, we collected these ratings through a survey on Prolific with 415 participants, gathering prototypicality ratings for each social category: how Black or White the faces appear for their respective racial groups, and how female or male the faces appear for their respective gender groups.² Each participant rated 10 randomly assigned faces using two questions: “How much does this person look Black/White?” for racial prototypicality and “How much does this person look male/female?” for gender prototypicality, both on 7-point Likert scales (1 = Not at all to 7 = Extremely). This design aimed to collect approximately 100 ratings per face. Detailed participant demographics are provided in Section S3 of the Supplementary Materials.

2.2 VLM Data Collection

VLMs generated 50 stories per image in response to the following prompt: “Write a 50-word fictional story capturing this American person’s typical day.”³ We selected this prompt after evaluating over 50 variations that requested stories about individuals in images. This particular wording consistently generated stories across all models without triggering noncompliances. We tested one proprietary model—GPT-4 Turbo—and three open-source models: BLIP-3 [*xgen-mm-phi3-mini-instruct-r-v1*; Xue et al., 2024], Ovis1.6 [*Ovis1.6-Ovis1.62-9B* Lu et al., 2024], and Llama-3.2 [*Llama-3.2-11B-Vision-Instruct*; Grattafiori et al., 2024]. This combination of proprietary and open-source models was selected to enhance the generalizability of our findings. Default parameters were used for all models.

2.3 Homogeneity Bias

To assess homogeneity bias in VLM-generated text, we adopted the method used by Lee et al. [2024]. We first encoded the VLM-generated texts into sentence embeddings, numeric vectors containing semantic and syntactic information of sentences. We used Sentence-BERT models for the encoding task, which have been fine-tuned on pre-trained models like BERT [Devlin et al., 2018] and RoBERTa [Liu et al., 2019] to produce high-quality sentence embeddings optimized for similarity assessments [Reimers and Gurevych, 2019]. We used three Sentence-BERT models from the *sentence-transformers* package in python version 3.11.5: *all-mpnet-base-v2*, *all-distilroberta-v1*, and *all-MiniLM-L12-v2*. We discuss model selection in Section S5 of the Supplementary Materials. As pinpointing the exact sources of variation between these encoder models was difficult due to their lack of interpretability, we analyzed their overall patterns to interpret the results; that is, we only reported findings that were consistent across the majority of encoder models (i.e., at least 2 out of 3), and patterns observed in only one encoder model were not included in our interpretations.

²We made these ratings publicly available: *URL blinded for anonymity during review*

³Prompt development and sample size determination are detailed in Supplementary Materials, Sections S4 and S1, respectively).

We fitted two types of mixed-effects models [Pinheiro and Bates, 2000, Bates et al., 2014] to examine the relationship between race, prototypicality, and text homogeneity in VLM-generated outputs. The first model included race/gender and mean prototypicality of the images as fixed effects (*Race/Gender and Prototypicality models*), which we used to assess the effects of race/gender and prototypicality. In the second model, we added an interaction term between race/gender and mean prototypicality to test whether the relationship between prototypicality and text homogeneity varied by racial/gender group (*Race/Gender Interaction models*).

In all models, we included *Pair ID*, a unique identifier of the image stimuli pair used to calculate cosine similarity (e.g., the Pair ID value for a cosine similarity measurement derived from texts generated for “BM01_NC” and “BM02_NC” was “BM01_NC-BM02_NC”) as random intercepts as we expected cosine similarity baselines to vary by image pair. We performed likelihood-ratio tests (LRTs) on the Race/Gender and Prototypicality models to determine if race/gender and prototypicality improved model fit and LRTs on the Race/Gender Interaction models to determine if the interaction term improved fit.

2.4 Trait Associations

To examine whether VLMs associate Black Americans with certain traits more than White Americans, and vice versa, we used Structural Topic Models [STMs; Roberts et al., 2019]. STMs discover latent topics within a collection of documents, where each topic is characterized by a distribution over words. They estimate the proportion of each topic within a document and the probability of each word belonging to a topic. By modeling topic prevalence as a function of document-level metadata, STMs allow us to analyze how race influences the trait associations made by VLMs. We conducted these analyses using the *STM* package in R Version 4.4.0.

Prior to fitting the STMs, we removed names from the VLM-generated texts to prevent them from being identified as topics. We fitted two separate STMs: one where topic prevalence was predicted by race, racial prototypicality, and VLM, and another where topic prevalence was predicted by gender, gender prototypicality, and VLM (i.e., which VLM was used for data collection).⁴ For each STM, we first determined the optimal number of topics by balancing exclusivity, held-out likelihood, and semantic coherence.⁵ Using a combination of topic-associated keywords and representative texts, we identified cohesive and interpretable topics. Finally, we used the STM to compare topic prevalence across covariates such as race/gender, prototypicality, and their interactions. We present these effects for each individual VLM in Tables S15 and S8.

3 Results

3.1 Homogeneity Bias (Gender)

We found a significant gender effect in all four VLMs, with cosine similarity values higher for women than for men ($bs \geq 0.19$, $ps \leq .001$; see Figure 2). LRTs indicated that including gender significantly improved model fit in all models ($\chi^2(1)s \geq 21.84$, $ps < .001$). See Tables S2 and S3 for summary outputs of the Gender and Prototypicality models and Table S6 for LRT results.

We found a significant gender prototypicality effect in all four VLMs where higher mean prototypicality of the images was related to greater cosine similarity ($bs \geq 0.18$, $ps < .05$). LRTs indicated that including racial prototypicality significantly improved model fit in all models ($\chi^2(1)s \geq 7.66$, $ps < .01$). See Tables S2 and S3 for summary outputs of the Femininity models and Table S6 for LRT results.

Finally, we found limited evidence of an interaction between gender and prototypicality. In GPT-4 Turbo and Llama-3.2, the association between gender prototypicality and homogeneity was stronger for women than men ($bs \geq .39$, $ps < .01$; see Figure S2). In BLIP-3, there was not significant interaction effect whereas in Ovis1.6, the interaction effect was observed in the opposite direction ($bs \leq -0.29$, $ps < .05$). LRTs indicated that including the interaction term significantly improved model fit ($\chi^2(1)s \geq 41.58$, $ps < .001$). See Tables S4 and S5 for summary outputs of the Interaction models and Table S6 for LRT results.

3.2 Homogeneity Bias (Race)

We found a significant race effect in GPT-4 Turbo, BLIP-3, and Llama-3.2 but in the opposite direction of what we expected. Cosine similarity values were smaller for Black Americans than for White Americans ($bs \leq -0.15$, $ps \leq .01$;

⁴We included the VLM term to account for variations in effects across different models.

⁵A detailed discussion of identifying the optimal number of topics in STMs can be found in Weston et al. [2023].



Figure 2: Standardized cosine similarity of the two gender groups calculated using *all-mpnet-base-v2*.

see Figure 3). We did not find a significant effect of race in Ovis1.6. LRTs indicated that including race significantly improved model fit in models where a race effect was significant ($\chi^2(1)s \geq 7.66, ps < .01$).

We found a significant racial prototypicality effect in one of four VLMs—Llama-3.2—where higher mean prototypicality of the images was related to smaller cosine similarity ($bs \leq -0.10, ps < .05$). LRTs indicated that including prototypicality significantly improved model fit in models where prototypicality effect was significant ($\chi^2(1)s \geq 3.91, ps < .05$). See Tables S9 and S10 for summary outputs of the Race and Prototypicality models and Table S13 for LRT results.



Figure 3: Standardized cosine similarity of the two racial groups calculated using *all-mpnet-base-v2*.

Finally, we did not find evidence of an interaction effect between race and racial prototypicality (see Figure S4). See Tables S11 and S12 for summary outputs of the Interaction models and Table S13 for LRT results.

3.3 Trait Associations (Gender)

We fitted an STM with 8 topics. Table S7 presents each topic's proportion, its FREX words, and representative example text. From these, we identified four topics with cohesive and interpretable keywords that aligned with their example texts. The same four topics were identified in this STM.

We did not find consistent differences in topic prevalence across VLMs, though individual models showed some gender differences. In GPT-4 Turbo, women were significantly more associated with basketball than men, and men were significantly more associated with art than women ($bs = 0.014$ and -0.029 , $95\% CIs = [0.00050, 0.027]$ and $[-0.046,$

-0.012]). In Ovis1.6, women were significantly more associated with appearance than men ($b = 0.021$, 95% CI = [0.0034, 0.038]). In BLIP-3, women were significantly more associated with healthcare than men ($b = 0.041$, 95% CI = [0.017, 0.066]). Finally, in Llama-3.2, men were significantly more associated with appearance than women ($b = -0.022$, 95% CI = [-0.039, -0.0042]). See Table S8 and Figure S3. Finally, we did not find any significant interaction effects (see Table S8).

3.4 Trait Associations (Race)

We fitted an STM with 8 topics. Table S14 presents each topic’s proportion, its FREX words (words that are both frequent and exclusive to the topic of interest), and representative example text. From these, we identified four topics with cohesive and interpretable keywords that aligned with their example texts. The four topics were: basketball, art, healthcare, and appearance.

The only consistent difference in topic prevalence between racial groups across all VLMs was that Black Americans were significantly more associated with basketball than White Americans ($b_s \geq 0.027$; see Figure 4). Other racial differences in topic associations varied by model: in GPT-4 Turbo, White Americans were significantly more associated with art than Black Americans ($b = -0.020$, 95% CI = [-0.033, -0.0074]); in Ovis1.6, White Americans were significantly more associated with healthcare than Black Americans ($b = -0.053$, 95% CI = [-0.074, -0.032]); and in BLIP-3, Black Americans were significantly more associated with appearance than White Americans ($b = 0.057$, 95% CI = [0.037, 0.078]). See Table S15 and Figure S5.

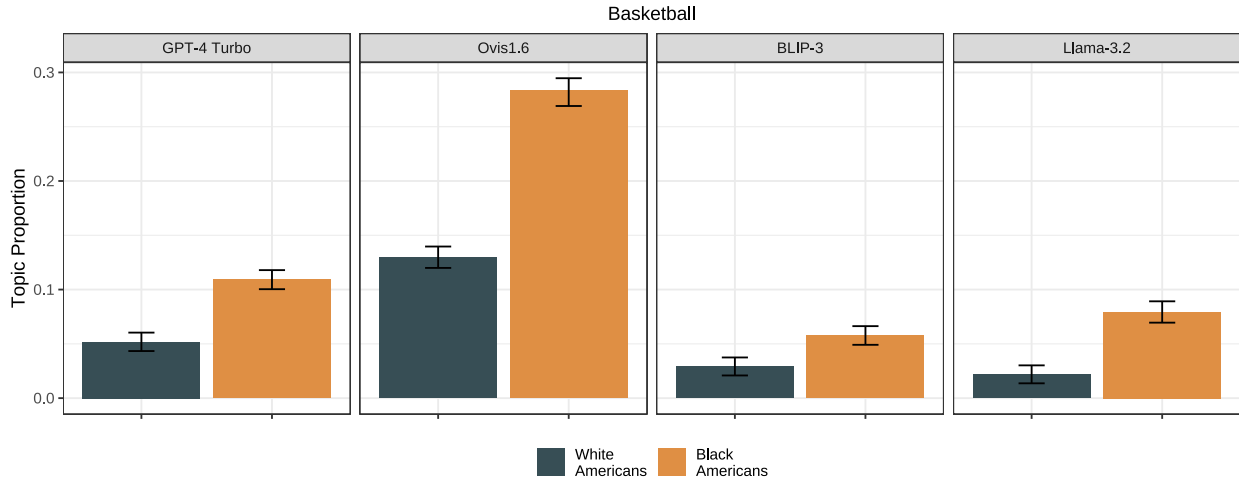


Figure 4: Prevalence of basketball of racial groups. In all four VLMs, Black Americans were significantly more associated with basketball than White Americans. Error bars indicate 95% confidence intervals. Visualization for other topics can be found in Figure S5 of the Supplementary Materials.

In GPT-4 Turbo and Ovis1.6, we found that racial prototypicality had a significantly stronger effect on basketball for Black Americans compared to White Americans (see Table S15). As faces were rated more prototypically Black, the prevalence of basketball as a topic increased in prevalence more sharply than it did for White faces rated as more prototypically White. All other interaction effects were not significant.

4 Discussion

4.1 Prototypicality and Vision-Language Model Stereotyping

Consistent with psychological research documenting that prototypicality is linked to greater stereotyping in humans, we found that VLMs show similar patterns that are especially pronounced in gender homogeneity bias. More prototypically female faces were subject to greater homogeneity bias. This pattern may be attributed to societal perceptions that associate femininity with lower agency and traits often linked with conformity and less autonomy. These perceptions may be reflected in the training data of VLMs, thereby being reproduced in model outputs [Hsu et al., 2021, Eagly et al., 2019]. The finding that VLM stereotyping may be linked to visual features like gender prototypicality suggests that conventional bias mitigation strategies such as data augmentation focused on balancing protected attributes (e.g., race

and gender) may be insufficient [see Lee et al., 2023]. As VLMs evolve to increasingly mimic human-like perception, the challenge extends beyond simple category-based interventions. A more nuanced understanding of how these models process and respond to visual cues is needed. Consequently, practitioners should incorporate a wider range of image characteristics in their bias mitigation efforts.

4.2 Positive Stereotyping in Vision-Language Models

When we asked VLMs to write open-ended stories about racial and gender groups, our Structural Topic Models (STMs) revealed only one consistent difference across all models, while other racial associations varied by model. GPT-4 Turbo associated White Americans more with art, Ovis1.6 associated White Americans more with healthcare, and BLIP-3 associated Black Americans more with appearance. The basketball association, while this association partly reflects the disproportionate representation of Black Americans in basketball [Lapchick et al., 2023], also aligns with stereotypes about Black Americans’ athleticism. Such associations, even when seemingly positive, represent a form of positive stereotyping [Czopp and Monteith, 2006]—a pattern previously observed both in humans and language models [Cheng et al., 2023b]. Positive stereotyping can have negative consequences, leading to the depersonalization of minority groups, reinforcing feelings of being reduced to group membership rather than being seen as individuals. It can also potentially limit the possibilities that members of these groups envision for themselves [Czopp and Monteith, 2006].

Our findings contrast with Lucy and Bamman [2021]’s study finding that GPT-3 tended to write stories about politics, war, sports, and crime for men. It remains unclear whether these tendencies have been mitigated in the more recent VLMs, but our results suggest that such trait associations may be context-dependent. Certain topics like politics and war likely didn’t emerge in our stories because they don’t typically feature in narratives about an individual’s "typical day." This highlights how prompt design significantly impacts which trait associations are detected, and suggests that a comprehensive collection of prompts are needed to thoroughly assess trait associations in language model outputs.

4.3 Homogeneity Bias and Trait Associations

Significant gender homogeneity bias was observed across all four models examined, yet no consistent gender effects were found among trait associations. This finding suggests that homogeneity bias, while a form of stereotyping, may operate independently of trait associations and cannot be explained by topic prevalence alone. Moreover, it indicates that while post-training steps to align language models with human values may prevent models from associating certain groups with stereotypic attributes, these measures may still fail to respect the diversity and richness of group identities, particularly those of minority groups. This discrepancy highlights the critical need for increased attention to homogeneity bias as a distinct phenomenon in language models, beyond traditional trait-based stereotypes. Understanding and addressing homogeneity bias could be key to developing fairer AI systems that better represent diversity.

5 Limitations and Future Work

5.1 Representation and Coverage of the Face Database

Our study has several important limitations regarding representation within standardized facial stimuli, which were necessary to systematically isolate the effects of social categories and their associated visual features. While RADIATE contains faces of diverse racial backgrounds, our analysis was constrained to examining two racial groups due to insufficient sample sizes for other racial categories. Similarly, the dataset only includes binary gender identities. This reflects broader challenges in the field, as many standardized facial databases suffer from limited representation [Conley et al., 2018, Buolamwini and Gebre, 2018, Kärkkäinen and Joo, 2019]. To advance this research meaningfully, we need more inclusive standardized face databases that better represent gender diversity and a wider range of racial and ethnic groups, while preserving the control that make these stimuli valuable for studying stereotyping and prejudice.

Additionally, our use of 40 images to represent all groups constrained the range of prototypicality ratings we could examine. The limited sample size may fail to capture the full spectrum of visual cues and their variations in each group category, potentially obscuring the more nuanced relationship between prototypicality and VLM stereotyping. Future work should leverage more expansive face databases to document how prototypicality relates to VLM stereotyping.

5.2 Prompt Variations

Previous research has shown that language model behaviors are influenced by prompt variations, suggesting bias evaluations should incorporate diverse prompts [Hida et al., 2024]. However, this approach presents unique challenges for vision-language models (VLMs), particularly when processing images of people. Many VLMs are designed to

avoid or reject prompts describing human images, likely as a safeguard against misuse. During prompt design, it was observed that even minor variations often resulted in noncompliance across models, making comparison across VLMs impossible. Consequently, despite recognizing the value of prompt variations, this study maintained a consistent prompt to ensure comparability. Future research should explore methods to study VLM stereotyping while addressing the challenges of prompt variation in this area.

5.3 Non-linear Effects of Prototypicality on VLM Stereotyping

A notable limitation of our analyses is their focus on linear relationships between prototypicality and both forms of stereotyping. Previous research has found evidence of quadratic relationships with prototypicality, where, for example, targets with average prototypicality elicit the most stereotyping [Ma et al., 2018]. If VLMs indeed reproduce human patterns of stereotyping, methods like linear mixed-effects models may fail to fully capture the relationship between prototypicality and stereotyping. In fact, regarding racial stereotyping, we found that in three of four VLMs—Ovis1.6, BLIP-3, and Llama-3.2—the homogeneity of VLM-generated stories decreased at the extreme range of prototypicality, suggesting a possible non-linear relationship that could explain the non-significant linear relationships we observed with racial prototypicality. However, non-linear effects of prototypicality on stereotyping remain relatively unexplored in social psychology. Cross-disciplinary efforts examining this relationship through the lenses of both social cognition and AI could be mutually beneficial, with insights from each field informing the other.

6 Conclusion

Our study reveals complex patterns in how VLMs process visual cues related to race and gender. While these models exhibit gender homogeneity bias that intensifies with gender prototypicality, their handling of race and racial prototypicality shows an unexpected reversal—depicting White Americans more homogeneously than Black Americans. This pattern, coupled with the persistence of positive stereotyping in language model outputs, suggests that current bias mitigation strategies may be creating unintended consequences. The dissociation between homogeneity bias and trait associations indicates these are distinct phenomena requiring separate consideration. Our findings highlight that as VLMs advance, traditional approaches to ensuring fair representation may be insufficient. Future work should focus on developing more nuanced bias mitigation strategies that account for both protected attributes like race and gender and visual features, while adequately representing the rich diversity within different social group categories.

7 Ethics Statement

This study was approved by the Institutional Review Board (IRB) at Washington University in St. Louis (IRB ID: 202501082). We utilized open-source VLMs, which were downloaded and run locally on personal devices, along with secure API endpoints for proprietary models that do not share information with third parties (e.g., OpenAI). This approach ensured full compliance with ethical standards and privacy concerns associated with submitting images of human faces to Artificial Intelligence (AI) models.

References

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models, October 2022.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A Survey on In-context Learning, June 2023.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, June 2023.

- Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks, August 2022.
- Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive Captioners are Image-Text Foundation Models, June 2022.
- OpenAI. GPT-4V(ision) System Card. 2023.
- Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The Woman Worked as a Babysitter: On Biases in Language Generation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1339.
- Li Lucy and David Bamman. Gender and Representation Bias in GPT-3 Generated Stories. In Nader Akoury, Faeze Brahman, Snigdha Chaturvedi, Elizabeth Clark, Mohit Iyyer, and Lara J. Martin, editors, *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nuse-1.5.
- Messi H.J. Lee, Jacob M. Montgomery, and Calvin K. Lai. 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*, pages 1321–1340, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658975.
- Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6):878–902, 2002. ISSN 1939-1315. doi: 10.1037/0022-3514.82.6.878.
- Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology*, 110(5):675–709, 2016. ISSN 1939-1315. doi: 10.1037/pspa0000046.
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, April 2018. doi: 10.1073/pnas.1720347115.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April 2017. doi: 10.1126/science.aal4230.
- Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416.
- Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On Measuring Social Biases in Sentence Encoders, March 2019.
- Patricia W. Linville, Peter Salovey, and Gregory W. Fischer. Stereotyping and perceived distributions of social characteristics: An application to ingroup-outgroup perception. In *Prejudice, Discrimination, and Racism*, pages 165–208. Academic Press, San Diego, CA, US, 1986. ISBN 978-0-12-221425-7.
- C. M. Judd, C. S. Ryan, and B. Park. Accuracy in the judgment of in-group and out-group variability. *Journal of Personality and Social Psychology*, 61(3):366–379, September 1991. ISSN 0022-3514. doi: 10.1037//0022-3514.61.3.366.
- Brian Mullen and Li-tze Hu. Perceptions of ingroup and outgroup variability: A meta-analytic integration. *Basic and Applied Social Psychology*, 10(3):233–252, 1989. ISSN 1532-4834. doi: 10.1207/s15324834basps1003_3.
- P. W. Linville, G. W. Fischer, and P. Salovey. 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): 165–188, August 1989. ISSN 0022-3514. doi: 10.1037//0022-3514.57.2.165.
- Ana Guinote, Charles M. Judd, and Markus Brauer. Effects of power on perceived and objective group variability: Evidence that more powerful groups are more variable. *Journal of Personality and Social Psychology*, 82(5):708–721, 2002. ISSN 1939-1315. doi: 10.1037/0022-3514.82.5.708.
- Fabio Lorenzi-Cioldi, Kay Deaux, and Anne-Claude Dafflon. Group homogeneity as a function of relative social status. *Swiss Journal of Psychology / Schweizerische Zeitschrift für Psychologie / Revue Suisse de Psychologie*, 57(4):255–273, 1998. ISSN 1662-0879.

- Susan T. Fiske. Controlling other people: The impact of power on stereotyping. *American Psychologist*, 48(6):621–628, 1993. ISSN 1935-990X. doi: 10.1037/0003-066X.48.6.621.
- Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.84.
- Messi H. J. Lee and Soyeon Jeon. Vision-Language Models Represent Darker-Skinned Black Individuals as More Homogeneous than Lighter-Skinned Black Individuals, December 2024.
- Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale, June 2023.
- Ranjita Naik and Besmira Nushi. Social Biases through the Text-to-Image Generation Lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 786–808, Montr' {e}al QC Canada, August 2023. ACM. ISBN 9798400702310. doi: 10.1145/3600211.3604711.
- Luhang Sun, Mian Wei, Yibing Sun, Yoo Ji Suh, Liwei Shen, and Sijia Yang. Smiling women pitching down: Auditing representational and presentational gender biases in image-generative AI. *Journal of Computer-Mediated Communication*, 29(1):zmad045, November 2023. ISSN 1083-6101. doi: 10.1093/jcmc/zmad045.
- Mansour Sami, Ashkan Sami, and Pete Barclay. 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)*, pages 391–396, October 2023. doi: 10.1109/ICSME58846.2023.00051.
- Kankan Zhou, Eason Lai, and Jing Jiang. VLStereoSet: A Study of Stereotypical Bias in Pre-trained Vision-Language Models. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, *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)*, pages 527–538, Online only, November 2022. Association for Computational Linguistics.
- Gabriele Ruggeri and Debora Nozza. A Multi-dimensional study on Bias in Vision-Language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6445–6455, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.403.
- Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink. The effects of category and physical features on stereotyping and evaluation. *Journal of Experimental Social Psychology*, 79:42–50, November 2018. ISSN 0022-1031. doi: 10.1016/j.jesp.2018.06.008.
- Robert W. Livingston and Marilyn B. Brewer. What are we really priming? Cue-based versus category-based processing of facial stimuli. *Journal of Personality and Social Psychology*, 82(1):5–18, 2002. ISSN 1939-1315, 0022-3514. doi: 10.1037/0022-3514.82.1.5.
- Irene V. Blair, Charles M. Judd, Melody S. Sadler, and Christopher Jenkins. The role of Afrocentric features in person perception: Judging by features and categories. *Journal of Personality and Social Psychology*, 83(1):5–25, July 2002. ISSN 0022-3514.
- Keith B. Maddox and Stephanie A. Gray. Cognitive Representations of Black Americans: Reexploring the Role of Skin Tone. *Personality and Social Psychology Bulletin*, 28(2):250–259, February 2002. ISSN 0146-1672. doi: 10.1177/0146167202282010.
- Claud Anderson and Rue L. Cromwell. "Black is Beautiful" and the Color Preferences of Afro-American Youth. *The Journal of Negro Education*, 46(1):76–88, 1977. ISSN 0022-2984. doi: 10.2307/2966874.
- Michael A. Strom, Leslie A. Zebrowitz, Shunan Zhang, P. Matthew Bronstad, and Hoon Koo Lee. Skin and bones: The contribution of skin tone and facial structure to racial prototypicality ratings. *PloS One*, 7(7):e41193, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0041193.
- Clara L. Wilkins, Cheryl R. Kaiser, and Heather Rieck. Detecting racial identification: The role of phenotypic prototypicality. *Journal of Experimental Social Psychology*, 46(6):1029–1034, November 2010. ISSN 0022-1031. doi: 10.1016/j.jesp.2010.05.017.
- Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, November 2022.

- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z.
- Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J. Kellman. Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, 14(12):e1006613, December 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006613.
- Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. Stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, 91:1–40, October 2019. ISSN 1548-7660. doi: 10.18637/jss.v091.i02.
- May I. Conley, Danielle V. Dellarco, Estee Rubien-Thomas, Alexandra O. Cohen, Alessandra Cervera, Nim Tottenham, and BJ Casey. The racially diverse affective expression (RADIATE) face stimulus set. *Psychiatry Research*, 270: 1059–1067, December 2018. ISSN 0165-1781. doi: 10.1016/j.psychres.2018.04.066.
- Nim Tottenham, James W. Tanaka, Andrew C. Leon, Thomas McCarry, Marcella Nurse, Todd A. Hare, David J. Marcus, Alissa Westerlund, BJ Casey, and Charles Nelson. The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, 168(3):242–249, August 2009. ISSN 0165-1781. doi: 10.1016/j.psychres.2008.05.006.
- Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S. Ryoo, and et al. xGen-MM (BLIP-3): A Family of Open Large Multimodal Models, August 2024.
- Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural Embedding Alignment for Multimodal Large Language Model, June 2024.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and et al. The Llama 3 Herd of Models, November 2024.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <https://arxiv.org/abs/1810.04805v2>, October 2018.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019.
- Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, August 2019.
- José C. Pinheiro and Douglas M. Bates. Linear Mixed-Effects Models: Basic Concepts and Examples. In *Mixed-Effects Models in S and S-PLUS*, pages 3–56. Springer, New York, NY, 2000. ISBN 978-0-387-22747-4. doi: 10.1007/0-387-22747-4_1.
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting Linear Mixed-Effects Models using lme4, June 2014.
- Sara J. Weston, Ian Shryock, Ryan Light, and Phillip A. Fisher. Selecting the Number and Labels of Topics in Topic Modeling: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 6(2):25152459231160105, April 2023. ISSN 2515-2459. doi: 10.1177/25152459231160105.
- Ning Hsu, Katie L. Badura, Daniel A. Newman, and Mary Eve P. Speech. Gender, “masculinity,” and “femininity”: A meta-analytic review of gender differences in agency and communion. *Psychological Bulletin*, 147(10):987–1011, 2021. ISSN 1939-1455. doi: 10.1037/bul0000343.
- Alice Eagly, Christa Nater, David Miller, Michèle Kaufmann, and Sabine Sczesny. Gender Stereotypes Have Changed: A Cross-Temporal Meta-Analysis of U.S. Public Opinion Polls From 1946 to 2018. *American Psychologist*, 75, July 2019. doi: 10.1037/amp0000494.
- Nayeon Lee, Yejin Bang, Holy Lovenia, Samuel Cahyawijaya, Wenliang Dai, and Pascale Fung. Survey of Social Bias in Vision-Language Models, September 2023.
- Richard E Lapchick, Andy Smith, Allison Straela, and Abraham Wade. The 2023 Racial and Gender Report Card™ National Basketball Association. 2023.
- Alexander M. Czopp and Margo J. Monteith. Thinking Well of African Americans: Measuring Complimentary Stereotypes and Negative Prejudice. *Basic and Applied Social Psychology*, 28(3):233–250, September 2006. ISSN 0197-3533. doi: 10.1207/s15324834basp2803_3.
- Myra Cheng, Tiziano Piccardi, and Diyi Yang. CoMPoS: Characterizing and Evaluating Caricature in LLM Simulations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.669.

- Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, January 2018.
- Kimmo Kärkkäinen and Jungseock Joo. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age, August 2019.
- Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. Social Bias Evaluation for Large Language Models Requires Prompt Variations, July 2024.
- Peter Green and Catriona J. MacLeod. SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4):493–498, 2016. ISSN 2041-210X. doi: 10.1111/2041-210X.12504.

S1 Power Analysis

We selected 10 images per demographic group. To determine adequate sample size, we conducted power analyses using the *simr* R package [Green and MacLeod, 2016], which estimates statistical power for mixed-effects models through Monte Carlo simulations. Based on effect sizes from previous research ($d = 0.30$; Lee et al. [2024]), we determined that 34 unique mean cosine similarity values per group would achieve 90% power to detect racial effects at $\alpha = .05$. Our study design, with 10 images per group yielding 45 unique pairwise combinations of images—each with distinct prototypicality ratings—exceeded this threshold, ensuring sufficient statistical power.

S2 Image Stimuli

Table S1: The randomly sampled RADIATE images used for data collection.

Race	Gender	Image IDs
Black	Men	BM01_NC, BM02_NC, BM03_NC, BM04_NC, BM05_NC, BM08_NC, BM09_NC, BM11_NC, BM12_NC, BM16_NC
Black	Women	BF09_NC, BF10_NC, BF12_NC, BF13_NC, BF14_NC, BF16_NC, BF17_NC, BF19_NC, BF21_NC, BF22_NC
White	Men	WM01_NC, WM03_NC, WM05_NC, WM06_NC, WM07_NC, WM08_NC, WM09_NC, WM10_NC, WM11_NC, WM12_NC
White	Women	WF02_NC, WF03_NC, WF07_NC, WF08_NC, WF09_NC, WF10_NC, WF11_NC, WF12_NC, WF14_NC, WF15_NC

S3 Demographic Breakdown of Raters

415 raters were recruited via Prolific research platform. Of the 415 participants, 263 (63.4%) identified as female, 144 (34.7%) as male, 5 (1.2%) as non-binary, and 3 (0.7%) did not disclose their gender. Regarding racial/ethnic identification, the sample comprised 277 (66.7%) White or Caucasian, 65 (15.7%) Black or Black American, 31 (7.5%) Asian, 25 (6.0%) Multiracial, 9 (2.2%) Other, 1 (0.2%) American Indian/Native American or Alaska Native, 1 (0.2%) Native Hawaiian or Other Pacific Islander participants, and 6 (1.4%) preferred not to disclose their race/ethnicity.

S4 Writing Prompt

Inside our writing prompts, we explicitly stated that the individual in the figure was American to prevent the model from associating the individual with other nationalities and emphasized that the individual in the figure was fictional to minimize non-compliance.

We initially tested broader writing prompts, such as, "Write a 50-word story about this American individual," but Llama-3.2 refused to generate stories in response to this prompt. We compiled variations of the writing prompt, such as, "In 50 words, describe a fictional day in this person's life that reveals their personality and values" and "Craft a 50-word fictional story imagining this person's daily life and a meaningful event they experienced," but they all resulted in non-compliances across models.

S5 Encoder Models

Among the many pre-trained models provided by the *sentence-transformers* package, we used the three models with the highest sentence embedding performance (as of Feb 7, 2025): *all-mpnet-base-v2*, *all-distilroberta-v1*, and *all-MiniLM-L12-v2*. The performance of pre-trained models were evaluated by assessing the similarity of text pairs across 14 different domains (e.g., Twitter, scientific articles, news). The models and their average performance scores can be found here: https://www.sbert.net/docs/pretrained_models.html.

S6 Results (Gender)

Table S2: Summary output of Gender and Prototypicality models (**GPT-4 Turbo** and **Ovis1.6**). A significantly positive Gender term indicates that cosine similarity of women was notably greater than men, and a significantly positive Prototypicality term indicates a positive relationship between mean gender prototypicality and cosine similarity.

	GPT-4 Turbo		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-2.82*** (0.30)	-3.06*** (0.27)	-1.89*** (0.29)
Gender (Women)	0.53*** (0.042)	0.46*** (0.038)	0.19*** (0.041)
Prototypicality	0.41*** (0.046)	0.46*** (0.041)	0.29*** (0.045)
Random Effects (σ^2)			
Pair ID Intercept	0.076	0.062	0.073
Residual	0.899	0.917	0.924
Observations	994,011	994,011	994,011
Log likelihood	-1,358,535	-1,368,671	-1,372,036
	Ovis1.6		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-1.62*** (0.38)	-2.12*** (0.32)	-2.07*** (0.36)
Gender (Women)	0.50*** (0.053)	0.60*** (0.045)	0.37*** (0.050)
Prototypicality	0.22*** (0.058)	0.29*** (0.049)	0.30*** (0.054)
Random Effects (σ^2)			
Pair ID Intercept	0.12	0.09	0.11
Residual	0.84	0.87	0.88
Observations	999,000	999,000	999,000
Log likelihood	-1,334,519	-1,347,937	-1,355,616

Table S3: Summary output of Gender and Prototypicality models (**BLIP-3** and **Llama-3.2**). A significantly positive Gender term indicates that cosine similarity of women was notably greater than men, and a significantly positive Prototypicality term indicates a positive relationship between mean gender prototypicality and cosine similarity.

	BLIP-3		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-1.43*** (0.22)	-1.12*** (0.23)	-1.21*** (0.27)
Gender (Women)	0.49*** (0.031)	0.19*** (0.032)	0.23*** (0.037)
Prototypicality	0.19*** (0.033)	0.17*** (0.035)	0.18*** (0.040)
Random Effects (σ^2)			
Pair ID Intercept	0.040	0.044	0.059
Residual	0.928	0.954	0.939
Observations	955,510	955,510	955,510
Log likelihood	-1,320,954	-1,334,469	-1,326,543
	Llama-3.2		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-0.82 (0.46)	-1.99*** (0.50)	-1.46** (0.49)
Gender (Women)	0.72*** (0.064)	0.82*** (0.070)	0.70*** (0.068)
Prototypicality	0.075 (0.070)	0.26*** (0.076)	0.18* (0.074)
Random Effects (σ^2)			
Pair ID Intercept	0.18	0.21	0.20
Residual	0.72	0.68	0.72
Observations	933,168	933,168	933,168
Log likelihood	-1,169,502	-1,147,999	-1,174,777

Table S4: Summary output of the Gender Interaction models (**GPT-4 Turbo** and **Ovis1.6**). A significantly positive Interaction term indicates that the relationship between mean gender prototypicality and cosine similarity of women is significantly greater than that of men.

	GPT-4 Turbo		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-0.30 (0.82)	-0.26 (0.74)	0.34 (0.80)
Gender (Women)	-2.33** (0.87)	-2.72*** (0.78)	-2.34** (0.85)
Prototypicality	0.03 (0.12)	0.03 (0.11)	-0.050 (0.12)
Interaction	0.44** (0.13)	0.49*** (0.12)	0.39** (0.13)
Random Effects (σ^2)			
Pair ID Intercept	0.074	0.060	0.071
Residual	0.899	0.917	0.924
Observations	994,011	994,011	994,011
Log likelihood	-1,358,531	-1,368,664	-1,372,033
	Ovis1.6		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-4.46*** (1.04)	-3.80*** (0.88)	-4.38*** (0.97)
Gender (Women)	3.74*** (1.10)	2.51** (0.94)	2.99** (1.03)
Prototypicality	0.65*** (0.16)	0.55*** (0.13)	0.66*** (0.15)
Interaction	-0.50** (0.17)	-0.29* (0.14)	-0.40* (0.16)
Random Effects (σ^2)			
Pair ID Intercept	0.12	0.09	0.10
Residual	0.84	0.87	0.88
Observations	999,000	999,000	999,000
Log likelihood	-1,334,515	-1,347,936	-1,355,614

Table S5: Summary output of the Gender Interaction models (**BLIP-3** and **Llama-3.2**). A significantly positive Interaction term indicates that the relationship between mean gender prototypicality and cosine similarity of women is significantly greater than that of men.

BLIP-3			
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-0.33 (0.60)	-0.15 (0.63)	-0.10 (0.73)
Gender	-0.76 (0.64)	-0.91 (0.67)	-1.03 (0.77)
Prototypicality	0.026 (0.092)	0.020 (0.096)	0.0089 (0.111)
Interaction	0.19 (0.098)	0.17 (0.103)	0.19 (0.119)
Random Effects (σ^2)			
Pair ID Intercept	0.040	0.044	0.059
Residual	0.928	0.954	0.939
Observations	955,510	955,510	955,510
Log likelihood	-1,320,954	-1,334,469	-1,326,542
Llama-3.2			
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	3.62** (1.24)	2.41 (1.37)	3.00* (1.32)
Gender	-4.33** (1.31)	-4.18** (1.45)	-4.37** (1.40)
Prototypicality	-0.60** (0.19)	-0.41* (0.21)	-0.50* (0.20)
Interaction	0.78*** (0.20)	0.77*** (0.22)	0.78*** (0.22)
Random Effects (σ^2)			
Pair ID Intercept	0.17	0.21	0.19
Residual	0.72	0.68	0.72
Observations	933,168	933,168	933,168
Log likelihood	-1,169,495	-1,147,994	-1,174,771

Table S6: Likelihood-ratio test results (Gender). Significant χ^2 statistic indicates that the effect of interest improved model fit.

VLM	Encoder Model	Mixed-Effects Model	Effect	χ^2	df	<i>p</i>
GPT-4 Turbo	all-mpnetbase-v2	Gender and Prototypicality model	Gender	133.82	1	< .001
		Gender and Prototypicality model	Prototypicality	74.86	1	< .001
		Interaction model	Interaction	138.03	2	< .001
	all-distilroberta-v1	Gender and Prototypicality model	Gender	128.12	1	< .001
		Gender and Prototypicality model	Prototypicality	107.12	1	< .001
		Interaction model	Interaction	137.44	2	< .001
	all-MiniLM-L12-v2	Gender and Prototypicality model	Gender	21.84	1	< .001
		Gender and Prototypicality model	Prototypicality	40.13	1	< .001
		Interaction model	Interaction	41.58	2	< .001
VLM	Encoder Model	Mixed-Effects Model	Effect	χ^2	df	<i>p</i>
Ovis1.6	all-mpnetbase-v2	Gender and Prototypicality model	Gender	81.02	1	< .001
		Gender and Prototypicality model	Prototypicality	14.46	1	< .001
		Interaction model	Interaction	100.94	2	< .001
	all-distilroberta-v1	Gender and Prototypicality model	Gender	147.48	1	< .001
		Gender and Prototypicality model	Prototypicality	34.76	1	< .001
		Interaction model	Interaction	173.32	2	< .001
	all-MiniLM-L12-v2	Gender and Prototypicality model	Gender	52.37	1	< .001
		Gender and Prototypicality model	Prototypicality	30.89	1	< .001
		Interaction model	Interaction	50.44	2	< .001
VLM	Encoder Model	Mixed-Effects Model	Effect	χ^2	df	<i>p</i>
BLIP-3	all-mpnetbase-v2	Gender and Prototypicality model	Gender	196.76	1	< .001
		Gender and Prototypicality model	Prototypicality	32.00	1	< .001
		Interaction model	Interaction	257.05	1	< .001
	all-distilroberta-v1	Gender and Prototypicality model	Gender	32.38	1	< .001
		Gender and Prototypicality model	Prototypicality	22.07	1	< .001
		Interaction model	Interaction	33.44	2	< .001
	all-MiniLM-L12-v2	Gender and Prototypicality model	Gender	36.99	1	< .001
		Gender and Prototypicality model	Prototypicality	19.06	1	< .001
		Interaction model	Interaction	38.16	2	< .001
VLM	Encoder Model	Mixed-Effects Model	Effect	χ^2	df	<i>p</i>
Llama-3.2	all-mpnetbase-v2	Gender and Prototypicality model	Gender	112.31	1	< .001
		Gender and Prototypicality model	Prototypicality	1.18	1	.28
		Interaction model	Interaction	212.22	2	< .001
	all-distilroberta-v1	Gender and Prototypicality model	Gender	119.44	1	< .001
		Gender and Prototypicality model	Prototypicality	11.14	1	< .001
		Interaction model	Interaction	177.82	2	< .001
	all-MiniLM-L12-v2	Gender and Prototypicality model	Gender	94.09	1	< .001
		Gender and Prototypicality model	Prototypicality	6.01	1	< .001
		Interaction model	Interaction	150.88	2	< .001

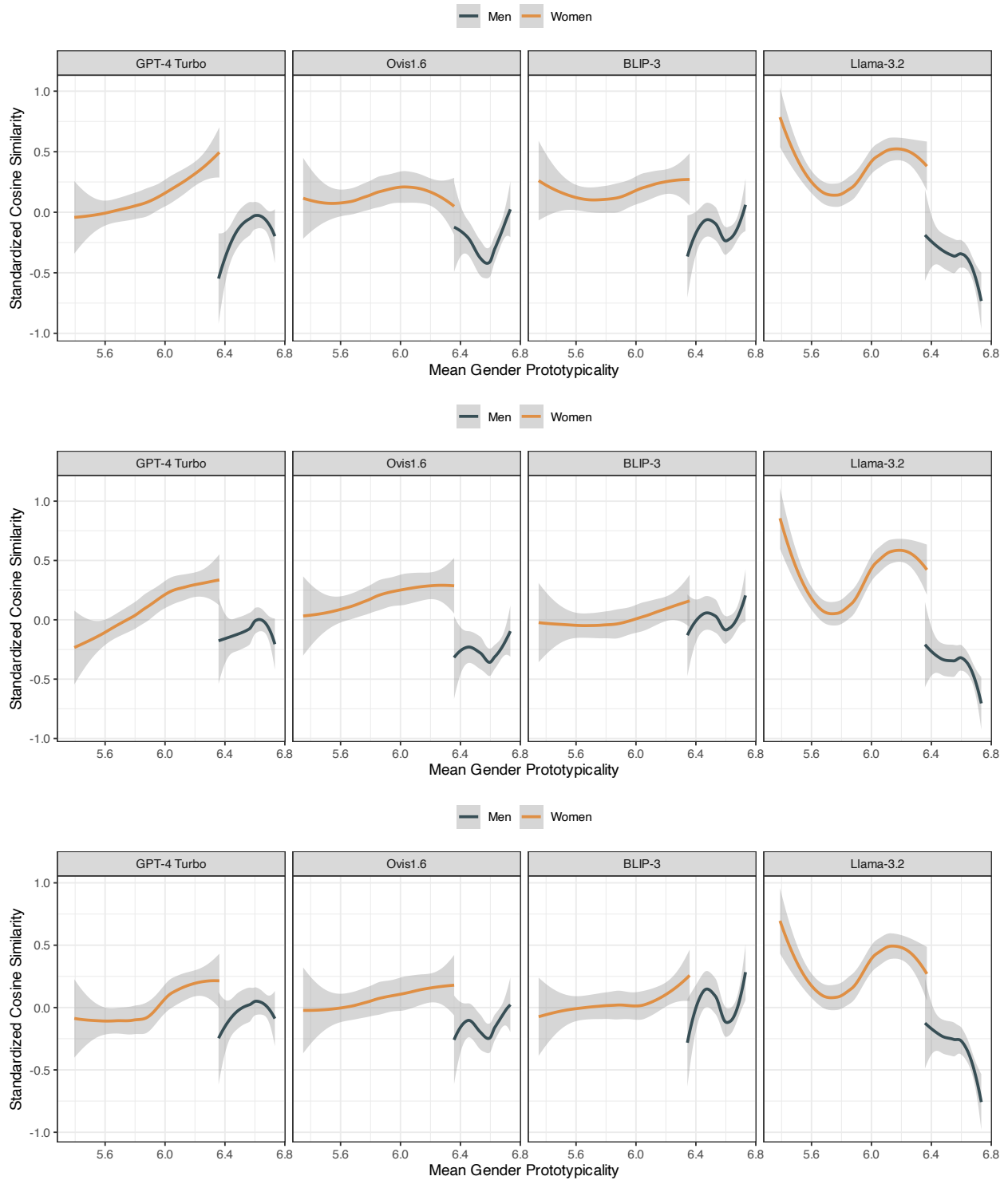


Figure S2: Standardized cosine similarity (1,000 random samples for each gender group) by prototypicality, calculated using all three encoder models. The top and bottom 10% of prototypicality values were excluded to minimize tail effects.

Table S7: STM output table containing the expected proportion of each topic (%), the top three FREX words, words that are both frequent and exclusive to each topic identified by the STM, and example texts by topic.

Topic	Proportions	FREX Words	Topic Label	Example
1	9.29%	local, communiti, share, school, basketbal	Basketball	On a typical day, our protagonist rises early, driven by a passion for basketball. After a quick breakfast, he heads to school, where he excels academically and plays on the varsity team. In the afternoons, he practices at the local court, honing his skills. As the sun sets, he enjoys time with friends, discussing future dreams and aspirations. The day concludes with quiet reflection, fueling his commitment to excellence and dreams of a brighter future.
2	14.40%	help, gentl, reflect, began, vibrant	–	In the crisp morning light, Sarah awoke in her cozy Brooklyn apartment, a blend of soft city sounds greeting her day. She dressed in her favorite light gray scarf, preparing for a day filled with the creative challenges of her graphic design job. Her commute was a gentle reminder of the city’s vibrant energy. At the studio, she immersed herself in crafting digital art, surrounded by the hum of creativity and collaboration. As the day wound down, she cherished the quiet moments, reflecting on the joy of turning ideas into reality.
3	15.36%	dinner, project, good, gym, break	–	As a freelance writer, Emma starts her day with a cup of coffee and a quick review of her schedule. She spends the morning working on articles, responding to client emails, and brainstorming new ideas. In the afternoon, she heads to the gym for a workout, followed by a relaxing yoga session. Dinner and a movie with her friends mark the end of her day, before she finally calls it a night.
4	19.89%	jazz, novel, paint, sketch, everi	Art	Amidst bustling New York streets, he brews morning coffee, savoring its aromatic warmth. He jogs through Central Park, embracing the rhythm of city life. At work, creativity flourishes within digital confines. Evenings find him strumming guitar chords, blending melodies with dreams before nightfall embraces another bustling tomorrow.
5	8.33%	alarm, stretch, rub, yawn, kitchen	–	The young man woke up to the sound of his alarm blaring in his ear. He rubbed the sleep from his eyes and swung his legs over the side of the bed. After a quick shower, he grabbed a granola bar and headed out the door to catch the bus to work.
6	8.95%	patient, lab, care, hospital, nurs	Healthcare	As a nurse, Jenna starts her day early, waking up at 5am to prepare for a 12-hour shift at the hospital. She puts on her scrubs and begins her rounds, checking on patients and administering medication. Throughout the day, she remains focused and compassionate, providing comfort and care to those in need. After her shift, Jenna returns home to her husband and two children, grateful for the opportunity to make a difference in people’s lives.
7	17.42%	thought, lost, gaze, rose, sip	–	As she sipped her coffee, Sarah gazed out the window, lost in thought. The sun rose over the city, casting a warm glow over the bustling streets. She took a deep breath, feeling the stress melt away. Today was a new day, full of possibilities and promise. She smiled, ready to take on whatever came her way.
8	6.37%	hair, shirt, cur, beard, woman	Appearance	A young woman with short dark hair and a shaved headline. She’s wearing a white shirt and has a serious expression on her face. She’s standing in a white background, staring off into the distance. She’s not smiling, but her expression is calm and composed.

Table S8: The effect of gender, gender prototypicality, and their interactions in individual VLMs across the five topics. A significant positive gender term indicates that the topic is significantly more prevalent for women than men, a significant positive prototypicality term indicates that a unit increase in prototypicality is associated with a significant increase in topic prevalence, and a significant interaction term indicates that the effect of prototypicality on topic prevalence is significantly greater for women than men.

	Gender			
	Basketball	Art	Healthcare	Appearance
GPT-4 Turbo	0.014* [0.00050, 0.027]	-0.029* [-0.046, -0.012]	0.0041 [-0.0073, 0.016]	0.00054 [-0.0094, 0.010]
Ovis1.6	0.0060 [-0.021, 0.033]	-0.0037 [-0.029, 0.022]	0.0056 [-0.015, 0.026]	0.021* [0.0034, 0.038]
BLIP-3	-0.0044 [-0.027, 0.018]	0.00032 [-0.025, 0.026]	0.041* [0.017, 0.066]	0.000052 [-0.018, 0.019]
Llama-3.2	-0.0094 [-0.033, 0.014]	0.0094 [-0.016, 0.035]	0.014 [-0.0058, 0.034]	-0.022* [-0.039, -0.0042]
	Interaction			
	Basketball	Art	Healthcare	Appearance
GPT-4 Turbo	-0.0088 [-0.053, 0.036]	0.0042 [-0.043, 0.051]	0.0020 [-0.039, 0.043]	0.013 [-0.023, 0.049]
Ovis1.6	0.096 [0.015, 0.18]	-0.012 [-0.086, 0.062]	0.026 [-0.048, 0.10]	0.026 [-0.037, 0.089]
BLIP-3	0.029 [-0.048, 0.11]	0.0015 [-0.071, 0.074]	-0.027 [-0.12, 0.064]	0.043 [-0.033, 0.12]
Llama-3.2	-0.0059 [-0.081, 0.069]	0.0067 [-0.066, 0.080]	0.0078 [-0.064, 0.080]	-0.0048 [-0.068, 0.059]

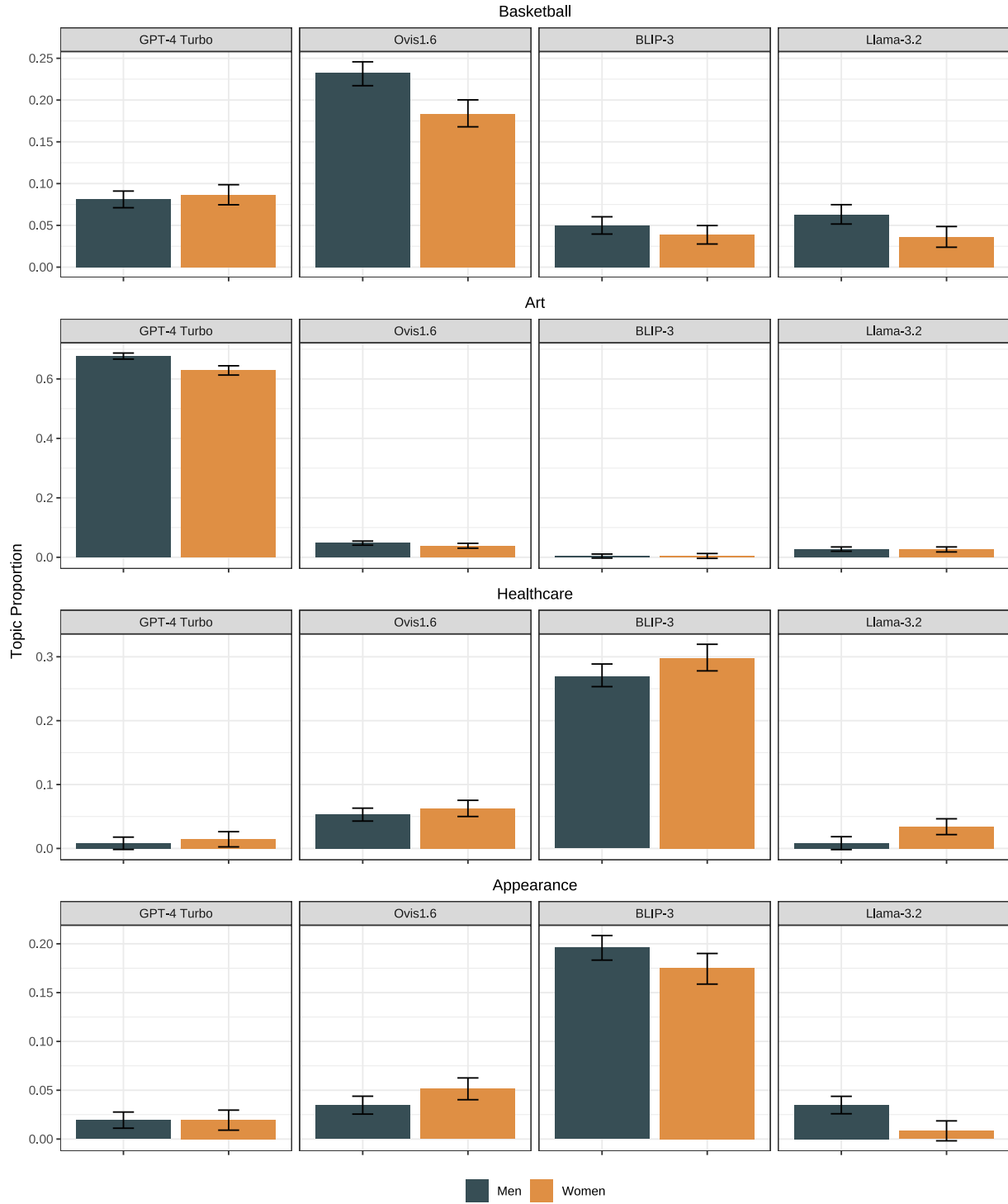


Figure S3: Prevalence of all four topics across four VLMs. Error bars indicate 95% confidence intervals.

S7 Results (Race)

Table S9: Summary output of the Race and Prototypicality models (**GPT-4 Turbo** and **Ovis1.6**). A significantly positive Race term indicates that cosine similarity of Black Americans was notably greater than White Americans. A significantly positive Prototypicality term indicates a positive relationship between mean prototypicality and cosine similarity.

	GPT-4 Turbo		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-0.26 (0.22)	0.040 (0.25)	0.18 (0.19)
Race	-0.46*** (0.038)	-0.36*** (0.043)	-0.44*** (0.032)
Prototypicality	0.084* (0.036)	0.027 (0.041)	0.011 (0.030)
Random Effects (σ^2)			
Pair ID Intercept	0.15	0.19	0.10
Residual	0.80	0.78	0.85
Observations	995,010	995,010	995,010
Log likelihood	-1302151.00	-1292041.00	-1334234.00
	Ovis1.6		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	0.54 (0.28)	0.086 (0.35)	0.76** (0.29)
Race	0.0021 (0.047)	0.030 (0.059)	-0.081 (0.048)
Prototypicality	-0.085 (0.045)	-0.011 (0.056)	-0.11* (0.046)
Random Effects (σ^2)			
Pair ID Intercept	0.23	0.36	0.24
Residual	0.78	0.65	0.76
Observations	999,000	999,000	999,000
Log likelihood	-1,292,578	-1,202,381	-1,282,517

Table S10: Summary output of the Race and Prototypicality models (**BLIP-3** and **Llama-3.2**). A significantly positive Race term indicates that cosine similarity of Black Americans was notably greater than White Americans. A significantly positive Prototypicality term indicates a positive relationship between mean prototypicality and cosine similarity.

	BLIP-3		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	0.38 (0.21)	0.43 (0.24)	0.51* (0.23)
Race	-0.17*** (0.036)	-0.11** (0.041)	-0.35*** (0.040)
Prototypicality	-0.045 (0.034)	-0.057 (0.039)	-0.050 (0.038)
Random Effects (σ^2)			
Pair ID Intercept	0.13	0.18	0.16
Residual	0.86	0.82	0.81
Observations	955,627	955,627	955,627
Log likelihood	-1,287,785	-1,265,156	-1,257,435
	Llama-3.2		
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	0.70* (0.31)	1.08*** (0.31)	0.78* (0.36)
Race	-0.15** (0.052)	-0.27*** (0.053)	-0.097 (0.061)
Prototypicality	-0.10* (0.049)	-0.15** (0.051)	-0.11* (0.058)
Random Effects (σ^2)			
Pair ID Intercept	0.28	0.30	0.39
Residual	0.72	0.69	0.62
Observations	933,396	933,396	933,396
Log likelihood	-1,172,468	-1,155,491	-1,100,526

Table S11: Summary output of the Race Interaction models (**GPT-4 Turbo** and **Ovis1.6**). A significantly positive Interaction term indicates that the relationship between mean prototypicality and cosine similarity of Black Americans is significantly greater than that of White Americans.

GPT-4 Turbo			
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	-0.20 (0.56)	-0.40 (0.63)	-0.98 (0.46)
Race (Black)	-0.53 (0.61)	0.17 (0.69)	0.93 (0.50)
Prototypicality	0.074 (0.090)	0.098 (0.10)	0.20** (0.074)
Interaction	0.011 (0.098)	-0.085 (0.11)	-0.22** (0.081)
Random Effects (σ^2)			
Pair ID Intercept	0.15	0.19	0.10
Residual	0.80	0.78	0.85
Observations	995,010	995,010	995,010
Log likelihood	-1302152.00	-1292042.00	-1334232.00
Ovis1.6			
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	2.73*** (0.68)	1.39 (0.86)	1.56* (0.71)
Race (Black)	-2.59*** (0.74)	-1.52 (0.94)	-1.04 (0.77)
Prototypicality	-0.44*** (0.11)	-0.22 (0.14)	-0.24* (0.11)
Interaction	0.42*** (0.12)	0.25 (0.15)	0.15 (0.13)
Random Effects (σ^2)			
Pair ID Intercept	0.22	0.36	0.24
Residual	0.78	0.65	0.76
Observations	999,000	999,000	999,000
Log likelihood	-1,292,573	-1,202,381	-1,282,517

Table S12: Summary output of the Race Interaction models (**BLIP-3** and **Llama-3.2**). A significantly positive Interaction term indicates that the relationship between mean prototypicality and cosine similarity of Black Americans is significantly greater than that of White Americans.

BLIP-3			
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	0.38 (0.21)	0.43 (0.24)	0.51* (0.23)
Race (Black)	-0.17*** (0.036)	-0.11** (0.041)	-0.35*** (0.040)
Prototypicality	-0.045 (0.034)	-0.057 (0.039)	-0.050 (0.038)
Interaction	0.19* (0.092)	0.13 (0.11)	0.088 (0.10)
Random Effects (σ^2)			
Pair ID Intercept	0.13	0.18	0.16
Residual	0.86	0.82	0.81
Observations	955,627	955,627	955,627
Log likelihood	-1,287,785	-1,265,157	-1,257,436
Llama-3.2			
	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L12-v2
Fixed Effects			
Intercept	0.70* (0.31)	0.78* (0.36)	1.08*** (0.31)
Race (Black)	-0.15** (0.052)	-0.097 (0.061)	-0.27*** (0.053)
Prototypicality	-0.10* (0.049)	-0.11* (0.058)	-0.15** (0.051)
Interaction	0.13 (0.13)	0.13 (0.16)	0.07 (0.14)
Random Effects (σ^2)			
Pair ID Intercept	0.28	0.39	0.30
Residual	0.72	0.62	0.69
Observations	933,396	933,396	933,396
Log likelihood	-1,172,469	-1,100,526	-1,155,492

Table S13: Likelihood-ratio test results (Race). Significant χ^2 statistic indicates that the effect of interest improved model fit.

VLM	Encoder Model	Mixed-Effects Model	Effect	χ^2	df	<i>p</i>
GPT-4 Turbo	all-mpnetbase-v2	Race and Prototypicality model	Race	98.16	1	< .001
		Race and Prototypicality model	Prototypicality	1.17	1	.28
		Interaction model	Interaction	97.67	2	< .001
	all-distilroberta-v1	Race and Prototypicality model	Race	13.94	1	< .001
		Race and Prototypicality model	Prototypicality	0.04	1	.84
		Interaction model	Interaction	13.89	2	< .001
	all-MiniLM-L12-v2	Race and Prototypicality model	Race	145.15	1	< .001
		Race and Prototypicality model	Prototypicality	3.81	1	.051
		Interaction model	Interaction	146.94	2	< .001
VLM	Encoder Model	Mixed-Effects Model	Effect	χ^2	df	<i>p</i>
Ovis1.6	all-mpnetbase-v2	Race and Prototypicality model	Race	0.00	1	.97
		Race and Prototypicality model	Prototypicality	3.63	1	.057
		Interaction model	Interaction	3.57	2	.15
	all-distilroberta-v1	Race and Prototypicality model	Race	0.27	1	.61
		Race and Prototypicality model	Prototypicality	0.04	1	.84
		Interaction model	Interaction	0.44	2	.80
	all-MiniLM-L12-v2	Race and Prototypicality model	Race	2.84	1	.092
		Race and Prototypicality model	Prototypicality	6.03	1	.014
		Interaction model	Interaction	7.89	2	0.19
VLM	Encoder Model	Mixed-Effects Model	Effect	χ^2	df	<i>p</i>
BLIP-3	all-mpnetbase-v2	Race and Prototypicality model	Race	22.17	1	< .001
		Race and Prototypicality model	Prototypicality	1.79	1	1.81
		Interaction model	Interaction	21.65	2	< .001
	all-distilroberta-v1	Race and Prototypicality model	Race	7.66	1	.006
		Race and Prototypicality model	Prototypicality	2.12	1	.15
		Interaction model	Interaction	8.64	2	.013
	all-MiniLM-L12-v2	Race and Prototypicality model	Race	70.16	1	< .001
		Race and Prototypicality model	Prototypicality	1.76	1	.19
		Interaction model	Interaction	69.17	2	< .001
VLM	Encoder Model	Mixed-Effects Model	Effect	χ^2	df	<i>p</i>
Llama-3.2	all-mpnetbase-v2	Race and Prototypicality model	Race	7.98	1	.005
		Race and Prototypicality model	Prototypicality	4.07	1	.044
		Interaction model	Interaction	10.67	2	.005
	all-distilroberta-v1	Race and Prototypicality model	Race	2.50	1	.11
		Race and Prototypicality model	Prototypicality	3.91	1	.048
		Interaction model	Interaction	5.70	2	.058
	all-MiniLM-L12-v2	Race and Prototypicality model	Race	24.96	1	< .001
		Race and Prototypicality model	Prototypicality	8.89	1	.003
		Interaction model	Interaction	30.60	2	< .001

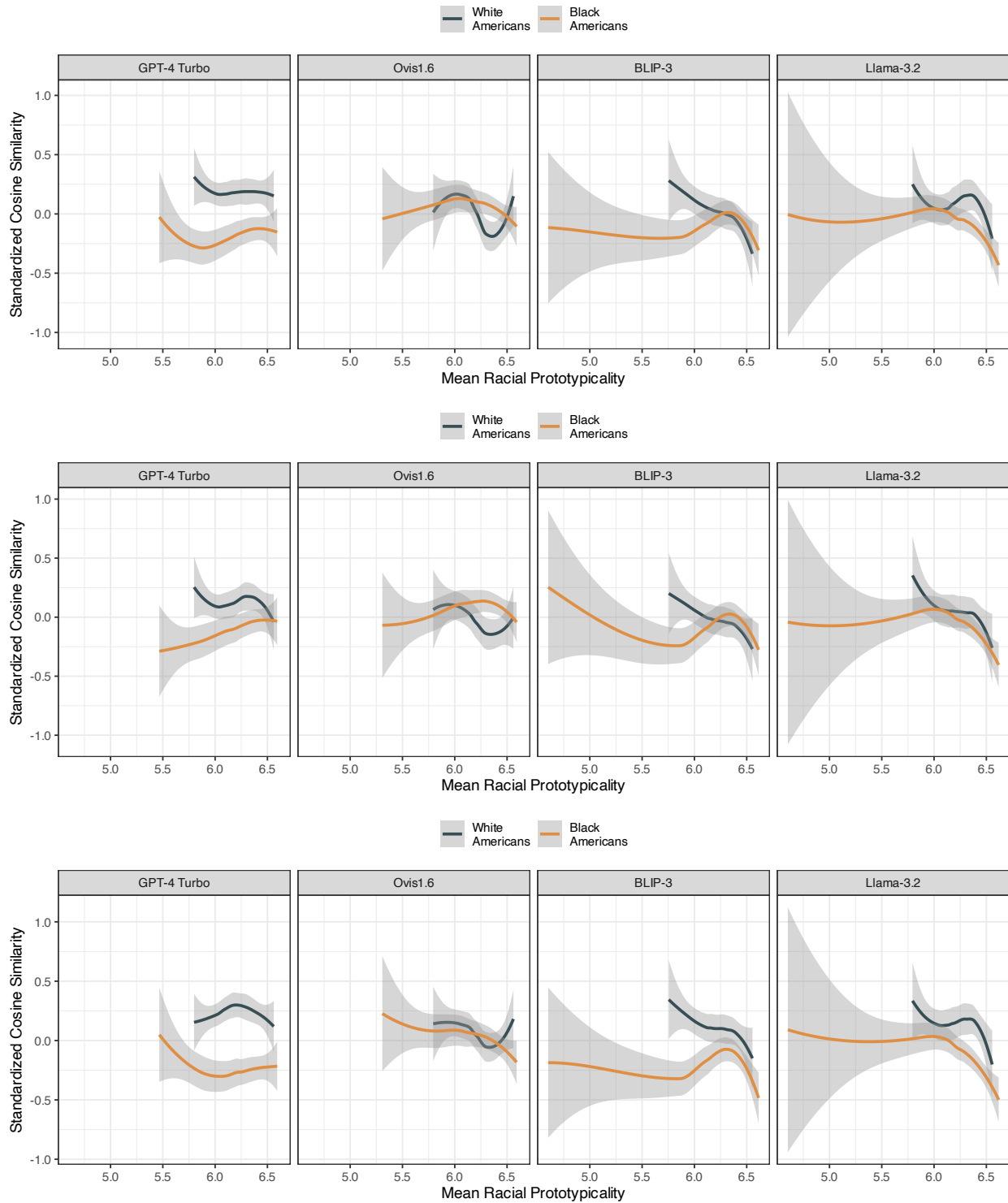


Figure S4: Standardized cosine similarity (10,000 random samples for each racial group) by prototypicality, calculated using all three encoder models. The top and bottom 10% of prototypicality values were excluded to minimize tail effects.

Table S14: Topics identified from the Race STM.

Topic	Proportions	FREX Words	Topic Label	Example
1	8.46%	communiti, school, local, basketball, share	School	On a typical day, our protagonist rises early, driven by a passion for basketball. After a quick breakfast, he heads to school, where he excels academically and plays on the varsity team. In the afternoons, he practices at the local court, honing his skills. As the sun sets, he enjoys time with friends, discussing future dreams and aspirations. The day concludes with quiet reflection, fueling his commitment to excellence and dreams of a brighter future.
2	14.77%	help, gentl, reflect, began, simpl	–	In the heart of a bustling American city, the day began with a soft morning light. Alex, with his casual yet stylish attire, embraced the day with a quiet resolve. A quick breakfast fueled him for a productive day at his tech startup. Amidst the city's vibrant pulse, he found moments of tranquility, reflecting on the day's challenges and successes.
3	15.65%	dinner, good, project, gym, break	–	As a busy software engineer, Michael often starts his day with a cup of coffee and a quick review of his schedule. He spends his mornings coding and collaborating with his team, solving complex problems and creating innovative solutions. After lunch, he might take a short walk around the office building to clear his head and recharge. In the late afternoon, he enjoys a quick chat with his colleagues before diving back into work. As the day winds down, he wraps up any remaining tasks and prepares for the next day.
4	19.87%	jazz, novel, everi, sketch, paint	Art	Each morning, she brews coffee, savoring its aroma. She bikes to her art studio, savoring San Francisco's breeze. Hours pass amidst vivid canvases, painting visions of freedom. At noon, she strolls the golden city's bustling streets, seeking inspiration. Evening arrives with soothing guitar melodies under a starry sky, dreams alight."
5	8.86%	alarm, check, yawn, rub, clock	–	The young man woke up to the sound of his alarm blaring in his ear. He rubbed the sleep from his eyes and swung his legs over the side of the bed. After a quick shower, he grabbed a granola bar and headed out the door to catch the bus to work.
6	9.17%	patient, care, lab, hospit, nurs	Healthcare	Kelly wakes up early to get ready for her day at work. She brushes her teeth, takes a shower, and gets dressed in her work clothes. Kelly works as a nurse in a busy hospital, so she knows that her day will be busy and challenging. She makes sure to stay focused and take care of her patients to the best of her ability. After her shift, Kelly goes home to rest and prepare for her night shift. She loves her job and the people she works with, and she is always grateful for the opportunity to help others.
7	16.74%	thought, lost, gaze, rose, sip	–	As he sipped his morning coffee, he gazed out the window, lost in thought. The sun rose higher, casting a warm glow over the city. He took a deep breath, feeling the weight of the day ahead. With a quiet determination, he rose from his chair, ready to face whatever challenges lay in store. Today would be a good day.
8	6.48%	hair, shirt, cur, woman, beard	Appearance	A young woman with short dark hair and a shaved headline. She's wearing a white shirt and has a serious expression on her face. She's standing in a white background, staring off into the distance. She's not smiling, but her expression is calm and composed.

Table S15: The effect of race and the interaction between race and racial prototypicality in individual VLMs across three topics. A significant positive race effect indicates that the topic is more prevalent for Black Americans than White Americans. A significant interaction term suggests that prototypicality has a stronger influence on topic prevalence for Black Americans compared to White Americans. Significant effects are marked with * when the 95% confidence intervals do not overlap with 0.

	Race			
	Basketball	Art	Healthcare	Appearance
GPT-4 Turbo	0.051* [0.040, 0.063]	-0.020* [-0.033, -0.0074]	0.0015 [-0.010, 0.013]	0.00031 [-0.0097, 0.010]
Ovis1.6	0.14* [0.12, 0.16]	-0.0073 [-0.028, 0.013]	-0.053* [-0.074, -0.032]	-0.0021 [-0.020, 0.016]
BLIP-3	0.027* [0.0066, 0.047]	-0.00087 [-0.021, 0.019]	0.0030 [-0.023, 0.029]	0.057* [0.037, 0.078]
Llama-3.2	0.052* [0.032, 0.072]	0.0023 [-0.018, 0.022]	0.012 [-0.0080, 0.033]	0.00028 [-0.017, 0.018]
	Interaction			
	Basketball	Art	Healthcare	Appearance
GPT-4 Turbo	0.023* [0.0029, 0.043]	-0.031 [-0.063, 0.000036]	0.0027 [-0.019, 0.024]	0.0064 [-0.012, 0.025]
Ovis1.6	0.047* [0.0082, 0.085]	0.0050 [-0.042, 0.052]	-0.030 [-0.068, 0.0085]	0.0071 [-0.026, 0.040]
BLIP-3	0.0061 [-0.028, 0.041]	0.0020 [-0.045, 0.049]	0.019 [-0.030, 0.067]	-0.0088 [-0.045, 0.028]
Llama-3.2	0.020 [-0.015, 0.055]	0.0058 [-0.041, 0.052]	0.0051 [-0.032, 0.043]	0.0012 [-0.032, 0.034]

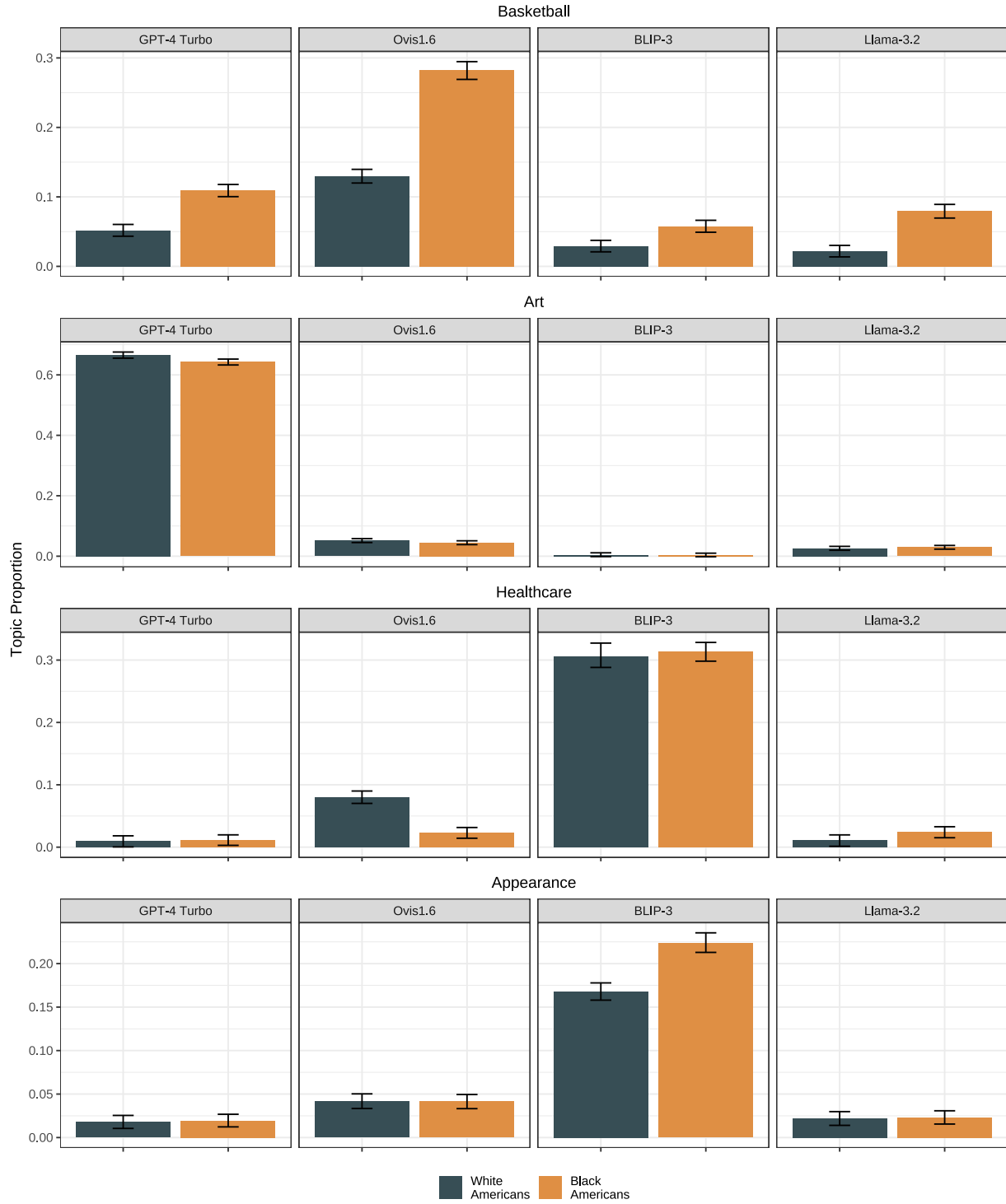


Figure S5: Prevalence of all four topics across four VLMs. Error bars indicate 95% confidence intervals.