A Stereotype Content Analysis on Color-related Social Bias in Large Vision Language Models

Anonymous ACL submission

Abstract

As large vision language models (LVLMs) rapidly advance, concerns about the risk of their learning or generating stereotypes are increasing. However, previous studies on LVLM's stereotypes face two primary limitations: metrics that overlooked the importance of content words, and datasets that overlooked the effect of color. To address these limitations, this study introduces new evaluation metrics based on the Stereotype Content Model (SCM). The SCM-based metric is grounded in social psychology, enabling the detection of stereotypical content words in model responses along the dimensions of competence and warmth. We also propose BASIC, a benchmark for assessing gender, race and color stereotypes. Using SCM metrics and BASIC, we conduct a study with eight LVLMs to discover stereotypes. As a result, we found three findings. (1) The SCM-based evaluation is effective in capturing stereotypes. (2) LVLMs exhibit color stereotypes in the output along with gender and race ones. (3) Interaction between model architecture and parameter sizes seems to affect stereotypes. We release BASIC publicly on [anonymized for review].

1 Introduction

As large vision-language models (LVLMs) continue to advance, they are increasingly being utilized in perceiving visual images across diverse domains (Xu et al., 2024; Gu et al., 2024; Wang et al., 2025). However, when employing LVLMs, it is essential to acknowledge that models may be affected by stereotypes, potentially influencing the models' perception. In humans, a stereotype refers to cognitive generalizations about the characteristics of individuals belonging to a specific social group (VandenBos, 2007). So, people often judge others unconsciously based on their visual appearance such as gender, race, and color (Cloutier et al., 2005; Eberhardt et al., 2004; Elliot et al., 2013).

Given that LVLMs are usually trained on human-generated data (Jia and Liang, 2017; Cheng et al., 2019), it is likely that they have learned human stereotypes. As a result, LVLMs sometimes exhibit harmful stereotypes, leading to bad consequences (Jiang et al., 2024; Howard et al., 2025; Fraser and Kiritchenko, 2024). Therefore, the need to identify such stereotypes is becoming important. Researchers recently designed various benchmarks to quantify stereotypes in LVLMs (Adewumi et al., 2024; Wang et al., 2024b; Park et al., 2025).

Despite the success of uncovering stereotypes, existing studies have less considered two factors possibly correlated with stereotypes: content words and color. First, regarding content words, we suspect that LVLMs may use positive sentiment words when generating stereotypical content. Many previous studies have relied on sentiment analysis to assess stereotypes (Wang et al., 2024a; Raj et al., 2024; Gerych et al., 2024). While this approach has contributed, sentiment-based evaluation cannot fully mirror how humans evaluate such stereotypes. Stereotypes are complex constructs compared to perceived sentiment, and thus cannot be thoroughly examined by sentiment polarity alone (Greenwald and Banaji, 1995; Amodio, 2014; Howard et al., 2025; Siy and Cheryan, 2016). Hence, to evaluate stereotypes more precisely, we adopt Stereotype Content Model (SCM), a socio-psychological framework that captures such implicit stereotypes in text and complements sentiment analysis (Fiske et al., 2018). With SCM, we examine how LVLMs exhibit stereotypical content in their responses.

Second, regarding color, LVLMs may exhibit stereotypes because of different colors in the image. Previously, researchers in vision-language model reported that model responds inconsistently due to image differences, even when images contain similar scenes (Wu et al., 2025; Kantipudi et al., 2020; Liang et al., 2025). So, studies on stereotypes have employed image pairs to control such

differences. Previous attempts built controlled image pairs, however, they overlooked the potential impact of color tone. Color tone is critical because humans are influenced by tonal differences (Elliot and Maier, 2013; Kaya and Epps, 2004; Wexner, 1954) and LVLMs might learn such tendency from human-generated data. Hence, we suspect that LVLM stereotypes could be affected by color tones. To justify such suspicion, we evaluate LVLMs with our new color-controlled benchmark.

Thus, we propose two metrics inspired by SCM and a new **B**enchmark for **A**ssessing **S**tereotypes with **I**mage **C**olors, BASIC. Additionally, we attempt to examine which part of LVLMs affect stereotypes, by comparing different models. We believe that a comparison between model architectures can hint at components relevant to stereotypes. So, this paper has the following contributions:

1. Based on SCM, we derive automatic metrics from human perception on stereotypes.
2. We suggest a benchmark that can separate the effect of colors from other stereotypes.
3. Comparing eight LVLMs, we found the effect of model and its size on stereotype is limited.

2 Related Work

We review prior study on evaluating stereotypes in terms of three aspects: (1) sentiment-based metrics, (2) designed benchmarks, and (3) tested LVLMs.

First, regarding content words, prior studies have mainly assessed model responses based on sentiment polarity (Inoshita and Zhou, 2024; Hu et al., 2025; Gerych et al., 2024), which limits their ability to capture the subtle meanings embedded in text. In the case of LVLMs, responses often exhibit positive sentiments because they were reinforced by human feedback (Li et al., 2024), making it difficult to detect stereotypes solely through sentiment-based metrics. Therefore, more sophisticated measures that directly target stereotypes are needed. In response to this limitation, recent work has begun incorporating social-psychological evaluation frameworks to analyze stereotypes. For instance, (Howard et al., 2025) was influenced by the Stereotype Content Model (SCM). The study used the dictionary from (Nicolas et al., 2021) to measure the frequency of SCM-related words in image captions. However, this approach is limited by its reliance on word frequency, without accounting for contextual or derived meanings.

Figure 1: Examples from our BASIC dataset. Top row shows chess players with different colors (white, blue, red), retaining race and gender. Bottom row shows DJs with different race or gender, retaining white color.

Second, regarding benchmarks, prior studies have emphasized constructing paired images to assess stereotypes of LVLMs. Traditional studies constructed diverse images based on common themes such as gender or race, and compared average evaluation scores across these thematic sets (Mandal et al., 2023; Janghorbani and De Melo, 2023; Qiu et al., 2023). However, such methods often suffer from insufficient control over confounding variables. For example, if gender, race, and background change simultaneously, any observed stereotypes cannot be solely attributed to gender when examining gender stereotypes. So, recent works have proposed that changing a targeted attribute while retaining the rest of the image (Kim et al., 2025; Howard et al., 2025, 2024; Fraser and Kiritchenko, 2024). However, unintended color differences still appear in the generated images. Considering the visual differences can influence the model's internal representations (De and Pedersen, 2021; Kantipudi et al., 2020; Liang et al., 2025), thereby potentially affecting the manifestation of stereotypes. Thus, color differences should be considered in stereotype evaluation frameworks.

Lastly, regarding model architecture, researchers recently investigated whether the model architecture affects stereotypes. Though previous studies have identified stereotypes within a limited set of models (Raj et al., 2024; Sathe et al., 2024), they did not clearly mention which architectural choice cause stereotypes. Recently, researchers have begun to identify causal relationships (Wu et al., 2024; Wang et al., 2024a). Despite their success, it is yet questionable whether the findings could be generalized to large VLMs with more than 50 billion

parameters. So, we systematically incorporate a broader range of model sizes.

3 BASIC Benchmark

In this study, we aim to separately examine the effect of color in the analysis of stereotypes produced by LVLMs. There are various visual attributes in person images that can influence stereotype formation: gender, race, and color. Since these factors appear together in a single image, it is essential to separate the impact of each attribute on the model's responses. To examine the potential impact of such differences, we construct a paired image dataset as shown in Figure 1.

We created the dataset using three primary colors: white, red, and blue. While a broader range of colors could have been included, we selected these particular ones based on cognitive factors related to human perception. Specifically, prior research has shown that humans tend to associate red, blue, and white with particular emotions, cognitive judgments, or stereotypes (Elliot and Maier, 2013; Kaya and Epps, 2004; Wexner, 1954). Therefore, we adapted the procedure proposed by (Howard et al., 2024) to construct a dataset with controlled color tones and designed a six-step process[1].

For the experimental analysis of model stereotypes, gender and race categorizations were adopted from previous studies (Karkkainen and Joo, 2021; Howard et al., 2024). Specifically, gender was determined based on biological sex, while race was classified into six predefined groups: African-American, Asian, Caucasian, Indian, Latino, and Middle Eastern. This selection was made solely for experimental purposes and is not intended to exclude any group.

Step 1: Constructing Occupation List. To create paired images, we started by choosing a common occupation for the individuals depicted in each image. Following the approach of Howard et al. (2024), we collected a list of occupations from studies (Nadeem et al., 2021; Chuang et al., 2023; Naik and Nushi, 2023). After removing duplicates, we obtained a list of 181 occupations. All subsequent steps are performed for each of 181 occupations.

Step 2: Creating action. We considered various actions that could be associated with each occupation. Since a single image cannot cover all possible actions of a specific occupation, multiple images describing different actions are required for each occupation. To achieve this, we adapted a prompt from (Sathe et al., 2024) and used GPT-4 to generate five distinct scenarios for each occupation.

Step 3: Generating Seed Images. For each of the five actions associated with a given occupation, we generated a seed image as a baseline for transformation. To ensure stability in subsequent image manipulation steps, the seed image was generated using an Asian male wearing white clothing on a gray background. This step was performed using Stable Diffusion XL, which generated high-resolution images (1024×1024 pixels) based on scenario corresponding to each occupation (Podell et al., 2023). We employed negative prompts (Ban et al., 2024) to enhance image quality and ensure alignment with the intended depiction. A negative prompt is an input method that specifies unwanted attributes, guiding the model to avoid including them in the generated images.

Step 4: Controlling Gender and Race. Based on a seed image, we generate paired images that reflect various combinations of gender and race. In this stage, we utilize the Image-to-Image model of Stable Diffusion XL to modify the seed image according to the specified race and gender attributes. We retain the contextual information conveyed by the scenario while altering only the race and gender of the original image. To ensure that the pose, position, and background remain consistent with the seed image, we adopt negative prompts.

Step 5: Controlling Color. Based on the images generated in Step 4, we further modify the color tone of each image to either blue or red, while maintaining the same race and gender. To ensure that the intended color is clearly reflected in the output, we use negative prompts. Through this step, we obtain a total of 180 images per scenario ($= 5$ actions $\times$ 2 genders $\times$ 6 races $\times$ 3 colors).

Step 6: Filtering. To ensure the quality of the final dataset, we conducted a filtering process based on three criteria. First, we evaluated the semantic similarity between the 32,580 generated images and their corresponding text prompts. We normalized the embeddings of both image and text using CLIP, then computed the cosine similarity between them. Only images with a similarity score of 0.2 or higher were retained in the dataset. Second, we included only those images that passed the NSFW

[1]The generation code is available at [anonymized for the review]. Detailed prompts and parameter settings for each step are provided in Appendix A.

	Red	*Green*	*Blue*
Red *vs.* White	0.12	-3.06	-2.43
Blue *vs.* White	-5.37	-4.87	-0.52

Table 1: RGB value changes (%) in red/blue images compared to white seed image

filtering, thereby excluding any potentially unsafe content that may have been generated. Lastly, we retained only occupations for which all 180 image variations passed both filtering conditions. If any single image was filtered out, we excluded the entire occupation to ensure completeness of pairs.

Through the above procedures, we generated a total of 18,360 images across 102 occupations. After completing all image generations, we assessed whether the color tones were clearly distinguishable by two methods. First, we computed the RGB scale differences between the red or blue images and their corresponding white counterparts. As shown in Table 1, the constructed dataset actively suppresses non-target colors to emphasize the intended color tone. Second, we manually validated 10% random samples to ensure the quality of dataset. We detail the result in Appendix B.

4 Experimental Setup

Using the dataset, we experiment whether LVLMs reveal different stereotypes across different gender, race, and colors. This section elaborates how we conduct our experiment including preparing descriptions, metrics, analysis, and models.

4.1 Collecting Description

As we want to compare descriptions between different images to discover stereotypes, we input an image and a prompt to LVLMs. For each image, we start a session by feeding the image to an LVLM. Then, we ask the LVLM to generate descriptions about the image, using an open-ended question. Specifically, we used following question: "Describe this person objectively." Using the description paragraph generated by the LVLM, we extract metrics for statistical analysis and conduct a simple qualitative analysis.

We selected the above question prompt because of two reasons. First, the prompt does not imply any specific situation or assumption. As LVLMs may produce different calculation result when we provide different situation prompt, implying any situation could lead to a specific stereotype. So, to focus solely on the effect of input images, we selected not to impose any situation within the prompt. Second, the prompt does not provide any specific choices. We thought that providing choices can also impose a new code of conduct: generated answer should be one of the choices.

4.2 Metrics

4.2.1 Existing sentiment-based metric

Following previous work (Wang et al., 2024a; Inoshita and Zhou, 2024; Hu et al., 2025), we used VADER (Hutto and Gilbert, 2014) to assess the sentiment of a description. Among the various sentiment-based metrics, researchers usually adopt VADER because the model is trained on the corpus from social media, which might contain stereotypical content. To obtain a score, we input a description of an image into VADER.

4.2.2 Our proposed SCM-based metrics

To inspect stereotypes thoroughly with content words, we propose two automatic metrics grounded in a foundational theory in social psychology, Stereotype Content Model (SCM), which conceptualizes human stereotypes along two dimensions: competence and warmth (Fiske et al., 2018). The competence dimension measures how capable the describer evaluates the ability of the person in the image to achieve own goal[2]. And, the warmth dimension measures how positively or negatively a describer evaluates the intention of the person[3].

In the following, we describe how we operationalized these two dimensions in our evaluation framework. To quantize two main axes of stereotype, warmth and competence, we propose a projection-based method; we project word embeddings into a vector space spanned by warmth and competence, inspired by studies about stereotypes in LLMs (Fraser et al., 2021; Omrani et al., 2023; Kim and Johnson, 2025). As SCM utilizes human intuition in its evaluation process, we chose to model such intuition using word embeddings. So, we assumed that a sentence embedding $\mathbf{x}$ is a linear combination of independent basis vectors: warmth $\mathbf{u}_w$, competence $\mathbf{u}_c$, and other basis vectors $\mathbf{v}_i$s orthogonal to $\mathbf{u}_w$ and $\mathbf{u}_c$. In other words, for any sentence embedding $\mathbf{x}$, there exist corresponding

[2]We used six words representing competence axis, from Fiske et al. (2018): *competent*, *confident*, *capable*, *efficient*, *intelligent*, *skillful*

[3]Also, we used seven words representing warmth: *friendly*, *well-intentioned*, *trustworthy*, *warm*, *good-natured*, *sincere*

real numbers $\alpha_w(x), \alpha_c(x)$ and $\beta_i(x)$ such that

$$\mathbf{x} = \alpha_w \mathbf{u}_w(x) + \alpha_c \mathbf{u}_c(x) + \sum_i \beta_i(x)\mathbf{v}_i. \quad (1)$$

So, we can compute orthogonal projection of $\mathbf{x}$ onto warmth-competence plane with two numbers $\alpha_w(x)$ and $\alpha_c(x)$, the projected coordinate on the plane. These numbers show how much warmth and competence should be combined together to have a similar meaning of the given sentence x. So, we interpret them as scores quantifying warmth and competence in SCM.

To implement the above idea, we need basis vectors. We normalized the average embedding of corresponding representative words. As a embedding model, we used Sentence-Transformer (Reimers and Gurevych, 2019). Let W and C be sets of words representing warmth and competence, respectively. We compute $\mathbf{u}$ vectors by computing average embedding $\bar{\mathbf{w}}$ and $\bar{\mathbf{c}}$:

$$\bar{\mathbf{w}} = \frac{1}{|W|}\sum_{x \in W} \texttt{ST}(x), \qquad \mathbf{u}_w = \frac{1}{\|\bar{\mathbf{w}}\|_2}\bar{\mathbf{w}}, \quad (2)$$

$$\bar{\mathbf{c}} = \frac{1}{|C|}\sum_{x \in C} \texttt{ST}(x), \qquad \mathbf{u}_c = \frac{1}{\|\bar{\mathbf{c}}\|_2}\bar{\mathbf{c}}, \quad (3)$$

where $\texttt{ST}(\cdot)$ is the Sentence-Transformer.

Using these two unit vectors, we compute warmth α_w and competence α_c of a sentence x as a projection. We first calculate dot products between $\mathbf{x} := \texttt{ST}(x)$ and basis vectors.

$$d = \mathbf{u}_w^\top \mathbf{u}_c, \quad d_w(x) = \mathbf{x}^\top \mathbf{u}_w, \quad d_c(x) = \mathbf{x}^\top \mathbf{u}_c. \quad (4)$$

Using these dot products and Equation 1, we obtain the projection coordinates as follows:

$$\alpha_w(x) = \frac{d_w(x) - d \cdot d_c(x)}{1 - d^2}, \quad (5)$$

$$\alpha_c(x) = \frac{d_c(x) - d \cdot d_w(x)}{1 - d^2}. \quad (6)$$

4.3 Analyses

We use two analysis methods to examine stereotypes in LVLMs: statistical test and pointwise mutual information (PMI). First, for statistical tests, we conducted paired t-test. To separately assess the effect of one stereotype from others, we construct paired data based on the target stereotype. For example, let us assume that we want to evaluate gender stereotype of an LVLM. Then, from the BASIC benchmark, we construct pairs of images whose gender is different and whose other aspects (race, color, scene, and occupation) are the same. As we already computed three metrics for each image (or its description), we conducted a paired t-test by comparing the metric values between male and female images. For race and colors, we conducted paired t-test in a pairwise way. For example, we tested for each pair of two different colors (e.g., white *vs.* blue).

Second, for PMI, we computed PMI between each aspect and words. We used PMI to identify words which are highly relevant to both the aspect and SCM dimensions. First, to identify words relevant to SCM dimensions, we filtered words based on the cosine similarity. We computed cosine similarity between $\mathbf{w} := \texttt{ST}(w)$ and basis vectors and discarded words whose similarity is in between -0.5 and 0.5[4]. Then, we compute PMI of a word w and an aspect a (e.g., red tone) as follows:

$$PMI(w; a) = \log_2 \frac{P(w|a)}{P(w)}, \quad (7)$$

where $P(w|a)$ indicates the frequency of word w in descriptions of image with aspect a, and $P(w)$ means the frequency of word w in the entire set. For each aspect, we selected 20 words whose PMI scores are highest. By collecting common words across models, we can identify stereotypes appeared in majority of models.

4.4 Tested models

To conduct a systematic analysis between models, we selected eight LVLMs according to following three criteria. First, we used open-sourced models whose architecture is known and already tuned on visual instructions. Second, we selected LVLMs which provides at least two different parameter sizes. Third, we used models which can be grouped with other models, according to image input method or backbone language models (LMs). As a result, we selected the following eight models, as shown in Table 2. Note that Llama 3.2 and InstructBLIP uses backbone LMs from same lineage, Llama. Unlike Llama 3.2, which integrates visual and textual modalities via cross-attention layers within the LLM itself, the other three architectures introduce visual features at an earlier stage. After processing the outputs of the image encoder through modules such as Q-Former, adapters, or projectors, they feed them directly into the language model input sequence. Additionally, small

[4]We used $0.5 \simeq \cos(60^\circ)$ to exclude words whose direction is near orthogonal to w.

	Llama 3.2	Instruct BLIP	Pixtral	Qwen 2.5
Backbone LMs	Llama 3	Vicuna (Llama 2)	Mistral	Qwen
Image-text Combination	Cross Attention	Input	Input	Input
Parameter Sizes	11B 90B	7B 13B	12B Large (124B)	7B 72B

Table 2: Model architectures used in the experiment

models have similar parameter sizes, such as 7B or 12B. Implementation details (e.g., temperature settings) are described in Appendix C.

5 Results

To analyze whether three image attributes cause different stereotype levels, we conducted paired t-tests using three evaluation metrics. Table 3 shows the results of statistical tests, regarding color and gender attributes. Due to the page limit, here we only present two of three attributes; we present the pairwise comparison between different race attributes in the Appendix D. Likewise, here we describe some common words discovered from PMI analyses, and the detailed result of PMI analyses is illustrated in the Appendix E. The following paragraphs describe the results regarding each metric.

VADER: Regarding color, the differences in sentiment scores were not consistent between models, though models sometimes exhibit stereotypes in some cases (Rows from 1 to 3). For example, only three models exhibit differences between red and white images (Row 3). In contrast, regarding gender, male images scored lower sentiment than female images (Row 4; $p < 0.05$ for all models).

Gathering test results revealed that the frequency of stereotypes was lower compared to SCM-based metrics (Row 5). Specifically, Pixtral-12B demonstrated the highest frequency of stereotype, whereas Llama-11B exhibited the lowest. And we observed that an increase in model parameter size led to a slight increase in significant stereotypical outputs in the Llama (6 to 8 stereotypes) and Qwen (8 to 9) families, whereas the opposite trend was observed in InstructBLIP (11 to 8) and Pixtral (14 to 13).

Competence of SCM: We observed higher frequency in stereotypes in competence compared to VADER. Also, the result is independent from VADER, as the correlation between them was near zero ($r = 0.014$, $p < 0.001$). Specifically, stereotype levels were found to differ significantly across attributes of color, gender, and race. Regarding color, blue images showed lower competence than red (Row 1; $p < 0.001$ for all models) and white (Row 2; $p < 0.001$ for all models except for InstructBLIP-7B). PMI result revealed a similar phenomenon; red and white were associated with high competence words, such as BRIGHT, STRONG, COMPETITIVE (for red), CLEAN, TECHNOLOGICAL, and EXPERTISE (white).

Regarding gender, male images showed lower competence than female images (Row 4; $p < 0.001$ except for Qwen-7B). PMI result supports this observation; males were associated with DISTINGUISHED, RESPONSIBILITY, INTELLECTUAL, whereas females were more linked to GRACEFUL, ELEGANT, BEAUTY.

Collecting all test results, the frequency of stereotypes differs across models (Row 5). Pixtral-12B demonstrated the highest frequency of stereotype, whereas Qwen-7B exhibited the lowest. Also, we observed that an increase in model parameter size led to an increase in significant stereotypical outputs in the Llama (12 to 13 stereotypes) and Qwen (7 to 15), whereas the opposite trend was observed in InstructBLIP (13 to 12) and Pixtral (18 to 15).

Warmth of SCM: We also observed higher frequency in stereotypes in warmth compared to VADER. Again, the result is independent from VADER with a weak correlation ($r = 0.189$, $p < 0.001$). Similar to competence, we found significant stereotypes in the result. Regarding color, red images showed lower warmth than blue (Row 1; $p < 0.001$) and white (Row 3; $p < 0.05$). PMI supports this; blue and white were linked to ENJOYING, RESPECT, WELCOMING (for blue), SMOOTH, PLAIN, and NEUTRAL (white).

Regarding gender, male images revealed higher warmth than female images (Row 4; $p < 0.001$ except for Qwen-72B). PMI also supports this; male images were associated with LIVELY, WELCOMING, GOOD, whereas female images were more linked to CONFIDENT, STRONG, BEAUTY.

Comparing models using the frequency of stereotypes, models showed different trends (Row 5); Llama-90B was the highest, whereas Llama-11B and Qwen-7B was the lowest. An increase in model parameter size led to more significant stereotypical outputs in the Llama (10 to 16 stereotypes) and

			Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B	Pixtral 12B	Pixtral Large	Qwen 2.5 7B	Qwen 2.5 72B
VADER	Color	Blue *vs.* Red	0.21	1.96	7.19^{***}	0.47	1.22	2.05^{*}	3.03^{**}	3.89^{***}
		Blue *vs.* White	-1.40	0.10	6.45^{***}	-1.11	3.38^{***}	2.73^{**}	0.70	-2.19^{*}
		Red *vs.* White	-1.61	-1.86	-0.89	-1.59	2.18^{*}	0.65	-2.31^{*}	-6.07^{***}
	Gender	Male *vs.* Female	-6.41^{***}	-3.66^{***}	-8.63^{***}	-8.29^{***}	-15.26^{***}	-7.89^{***}	-2.35^{*}	-3.30^{***}
# of significant stereotypes			6/19	8/19	11/19	8/19	14/19	13/19	8/19	9/19
Competence	Color	Blue *vs.* Red	-11.18^{***}	-9.08^{***}	-3.67^{***}	-7.95^{***}	-15.10^{***}	-10.39^{***}	-10.96^{***}	-12.87^{***}
		Blue *vs.* White	-11.08^{***}	-9.43^{***}	0.67	-6.39^{***}	-21.36^{***}	-14.52^{***}	-11.33^{***}	-16.16^{***}
		Red *vs.* White	0.03	-0.21	4.45^{***}	1.62	-5.66^{***}	-3.78^{***}	-0.40	-3.15^{**}
	Gender	Male *vs.* Female	-35.93^{***}	-20.55^{***}	-35.85^{***}	-20.43^{***}	-39.33^{***}	-28.01^{***}	-1.18	-4.51^{***}
# of significant stereotypes			12/19	13/19	13/19	12/19	18/19	15/19	7/19	15/19
Warmth	Color	Blue *vs.* Red	9.85^{***}	11.96^{***}	4.13^{***}	9.80^{***}	13.83^{***}	8.80^{***}	8.31^{***}	8.40^{***}
		Blue *vs.* White	4.87^{***}	9.71^{***}	-2.36^{*}	5.96^{***}	0.47	0.24	-0.02	-5.85^{***}
		Red *vs.* White	-4.96^{***}	-2.67^{**}	-6.62^{***}	-4.02^{***}	-14.18^{***}	-8.77^{***}	-8.34^{***}	-14.29^{***}
	Gender	Male *vs.* Female	28.65^{***}	23.72^{***}	27.08^{***}	35.62^{***}	17.79^{***}	12.09^{***}	-6.05^{***}	0.71
# of significant stereotypes			10/19	16/19	15/19	12/19	15/19	14/19	10/19	14/19

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 3: Paired t-test results for color and gender stereotypes. Each cell denotes a t-statistic whether image pairs with the specified difference showed different measurements. '# of significant stereotypes' row shows the number of statistically significant results on 19 stereotype tests, including race. All results including race is in the Appendix D.

Qwen (10 to 14) families, whereas the opposite trend was observed in InstructBLIP (15 to 12) and Pixtral (15 to 14).

6 Discussion

This section described three findings discovered from our experiment: the effectiveness of SCM metrics, the existence of color stereotypes, and the differences between models.

6.1 Effectiveness of SCM metrics

Our result suggests that using both SCM-based metrics and sentiment metrics provides more comprehensive view on stereotypes. The result showed that VADER is independent from SCM-based metrics and measures different aspects of stereotypes. VADER analyzes the emotional content of a sentence at a surface level by examining its positive or negative polarity, thus simplifying stereotypes to matters of sentiment. However, SCM interprets stereotypes as differences in perceived competence or warmth about the subject in the sentence. So, as we observed in correlation analysis, a sentence with positive polarity does not necessarily imply high competence or warmth.

Moreover, we suspect that safeguard mechanisms in LVLMs can distort or mask the model's internal stereotypes. These models are often explicitly trained to avoid harmful contents or a negative tone on possibly stereotypical topics, thereby influencing sentiment scores regardless of underlying stereotypes. So, sentiment polarity-based metrics, such as VADER, may not sufficiently capture stereotypes. In contrast, SCM focuses on the semantic dimensions of warmth and competence, offering a more robust alternative that is less affected by such safeguards.

6.2 Existence of Color Stereotypes

Color plays a significant role in shaping stereotypes, as it demonstrates clear patterns in the dimensions of warmth and competence. We believe that these stereotypes develop because the model internalizes sociocultural associations present in image-text data during its pretraining phase. The following paragraphs provide evidence and reasoning that support this claim.

Regarding competence, we suspect that a stereotypical association between color and the subject's capability may affect the score. For instance, red color may symbolize warning or aggression but can also be interpreted as a sign of passion or determination (Elliot and Maier, 2013; Gnambs et al., 2015). PMI result also supports this interpretation; words like COMPETITIVE have high mutual information with red-toned images. Similarly, white-toned images were also linked with EXPERTISE or TECHNOLOGICAL, mirroring stereotypes that professional, highly-competent occupations wear white-colored uniforms (Crutzen and Adam, 2022; Adam and

Galinsky, 2012). Because such stereotypes were already integrated in human-labeled caption dataset, red and white images showed higher competence than blue images in the result.

Regarding warmth, we suspect that frequent pairing of certain colors with trustworthiness may have increased warmth scores. For example, blue symbolizes calmness and stability (Kaya and Epps, 2004; Wexner, 1954). Similarly, white conveys a sense of softness and purity (Na and Suk, 2014; Kaya and Epps, 2004). Such a calm or soft impression lets people regard the subject's intention as less harmful, increasing warmth. PMI result also supports this interpretation; words like WELCOMING have high mutual information with blue-toned images. Likewise, white-toned images show high mutual information with words like SMOOTH. Thus, higher warmth scores for blue-toned and white-toned images imply that LVLMs learned such human stereotypes.

6.3 Differences between Models

Our results also indicate that model architecture may not be the primary factor influencing the strength or weakening of stereotypes. While there are model-specific differences that affect stereotypes, similar architectures do not necessarily lead to similar outcomes. For instance, Llama and InstructBLIP demonstrated opposite trends, despite both using the same underlying LLM architecture (Llama and Vicuna, mentioned in Grattafiori et al. (2024); Chiang et al. (2023)). Similarly, Pixtral and Qwen also exhibited opposite trends, even though both encode visual inputs using Vision Transformer-based modules and feed the projected hidden states into the LLM decoder as image tokens (Agrawal et al., 2024; Bai et al., 2025). So, the effect of architecture seems relatively small. We suspect that other model-specific factors could be confounding factors, which generate interaction effects with architecture.

Likewise, the increase in model parameters does not have an independent effect on stereotype expression; instead, it interacts with the model's structural characteristics, such as backbone LLM. Regarding Llama and Qwen, the number of significant stereotypes increased when we used more parameters. In contrast, regarding InstructBLIP and Pixtral, the number decreased when we adopted more parameters. Also, when we compare different architectures with similar sizes (e.g., 10B models), no model exhibited dominant stereotypes compared to others in all metrics. So we conclude the effect of parameter sizes and architecture are not easily distinguishable from each other.

7 Conclusion

In this paper, we introduced a comprehensive evaluation framework that adopts the Stereotype Contents Model (SCM) of social psychology and moves beyond sentiment-based assessments of stereotypes. We also presented BASIC, a novel benchmark designed to isolate and evaluate the role of image color tones in the formation of stereotypes by eight Large Vision-Language Models (LVLMs).

Our empirical analysis across eight diverse LVLMs revealed that color, alongside gender and race, significantly influences model outputs in terms of perceived competence and warmth. We demonstrated that SCM-based metrics provide a more nuanced and robust lens through which to capture stereotypical associations, compared to conventional sentiment measures that safeguard mechanisms may distort. Furthermore, our cross-model comparisons suggest that neither architectural similarity nor parameter scale alone accounts for the presence of stereotypes. Instead, we observed complex interactions between model-specific factors that collectively shape the stereotypes.

These findings underscore two points: (1) the adoption of semantically informed evaluation methods when measuring stereotypes and (2) the need for more careful consideration of visual attributes such as color. While sentiment-based metrics have been widely used in prior work, they are often influenced by models' safety filters or general positivity, which may obscure the presence of stereotypes. SCM-based metrics capture structured dimensions of social perception, enabling a more precise and interpretable analysis. We also find that color should not be treated as a mere background element, but rather as a socially meaningful visual cue that can influence model responses. We hope that BASIC and the SCM-based metrics will serve as practical and extensible tools for future research seeking to uncover, quantify, and ultimately mitigate the stereotype in LVLMs. As LVLMs continue to evolve and integrate into everyday applications, the ability to evaluate stereotypes across both linguistic and visual modalities will become increasingly essential in the research community.

Limitations

While our study presents SCM-based evaluation metrics and introduces BASIC, a benchmark developed to assess stereotypes in LVLMs, three limitations remain. First, our findings may not be fully independent from other visual attributes. In our study, we focused on three specific colors to investigate stereotypes in LVLMs. Though those colors are selected based on studies about human perception and we found such stereotypes are present in LVLMs, we believe that other color tones (e.g., yellow or green) may show different patterns of stereotypes. Similarly, other visual elements may affect our experiment; facial expressions, brightness, or posture may still have a potential influence on the model outputs. As these visual elements can generate an interaction effect with color stereotypes, we need further studies to investigate these visual elements to understand stereotypes in LVLM.

Second, our findings just provided a hint at how architectural differences affect LVLM stereotypes. Though we designed a systematic comparison between eight LVLMs in terms of parameter size and model architecture, our result only provides a correlation between them. To identify causal relationships or possible causes of stereotypes from model-specific architectures, we need to further control other factors such as training data and safeguards. Controlling such factors usually requires pre-training from scratch, which requires a lot of time and machine resources. Thus, we reported correlation without conducting an experiment for identifying causality. However, as a causal relationship should be discovered to reduce stereotypes in LVLMs, we call for further research on identifying such possible causes from model architecture.

Third, we conducted our experiment only in English. So, our experimental result is based on English culture. As Language-specific or culturally nuanced biases exist, LVLMs may provide different responses when we use different languages. This implies that the result might vary across different languages or cultures. Thus, conducting similar experiments with different languages may provide different insights.

Ethical Considerations

For the selection of gender and race in BASIC benchmark, we acknowledge that there is a broader spectrum of gender and race beyond those considered in this study. As our aim is to experiment with LVLMs with various genders, races, and colored images, we made this selection solely for experimental purposes. We believe that our method can be easily generalizable to other gender and race groups because our method does not depend on specific gender or race groups. Also, we want to emphasize that we did not intend to exclude any other gender or race groups from the experiment.

References

Hajo Adam and Adam D Galinsky. 2012. Enclothed cognition. *Journal of experimental social psychology*, 48(4):918–925.

Tosin Adewumi, Lama Alkhaled, Namrata Gurung, Goya van Boven, and Irene Pagliai. 2024. Fairness and bias in multimodal ai: A survey. *Preprint*, arXiv:2406.19097.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, and 1 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

David M Amodio. 2014. The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience*, 15(10):670–682.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. 2024. Understanding the impact of negative prompts: When and how do they take effect? In *European Conference on Computer Vision*, pages 190–206. Springer.

Minhao Cheng, Wei Wei, and Cho-Jui Hsieh. 2019. Evaluating and enhancing the robustness of dialogue systems: A case study on a negotiation agent. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3325–3335, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. 2023. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*.

Jasmin Cloutier, Malia F Mason, and C Neil Macrae. 2005. The perceptual determinants of person construal: reopening the social-cognitive toolbox. *Journal of personality and social psychology*, 88(6):885.

Coline Crutzen and Stéphane Adam. 2022. "what if it's not just an item of clothing?"–a narrative review and synthesis of the white coat in the context of aged care. *Psychologica Belgica*, 62(1):62.

Kanjar De and Marius Pedersen. 2021. Impact of colour on robustness of deep neural networks. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 21–30.

Jennifer L Eberhardt, Phillip Atiba Goff, Valerie J Purdie, and Paul G Davies. 2004. Seeing black: race, crime, and visual processing. *Journal of personality and social psychology*, 87(6):876.

Andrew Elliot and Markus Maier. 2013. Color psychology: Effects of perceiving color on psychological functioning in humans. *Annual review of psychology*, 65.

Andrew J Elliot, Jessica L Tracy, Adam D Pazda, and Alec T Beall. 2013. Red enhances women's attractiveness to men: First evidence suggesting universality. *Journal of Experimental Social Psychology*, 49(1):165–168.

Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2018. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social cognition*, pages 162–214. Routledge.

Kathleen Fraser and Svetlana Kiritchenko. 2024. Examining gender and racial bias in large vision–language models using a novel dataset of parallel images. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 690–713, St. Julian's, Malta. Association for Computational Linguistics.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. In *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 600–616, Online. Association for Computational Linguistics.

Walter Gerych, Haoran Zhang, Kimia Hamidieh, Eileen Pan, Maanas K Sharma, Tom Hartvigsen, and Marzyeh Ghassemi. 2024. Bendvlm: Test-time debiasing of vision-language embeddings. *Advances in Neural Information Processing Systems*, 37:62480–62502.

Timo Gnambs, Markus Appel, and Aileen Oeberst. 2015. Red color and risk-taking behavior in online environments. *PloS one*, 10(7):e0134033.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. 2024. Anomalygpt: detecting industrial anomalies using large vision-language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Phillip Howard, Kathleen C. Fraser, Anahita Bhiwandiwalla, and Svetlana Kiritchenko. 2025. Uncovering bias in large vision-language models at scale with counterfactuals. 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 5946–5991, Albuquerque, New Mexico. Association for Computational Linguistics.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. 2024. SocialCounterfactuals: Probing and Mitigating Intersectional Social Biases in Vision-Language Models with Counterfactual Examples . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11975–11985, Los Alamitos, CA, USA. IEEE Computer Society.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75.

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.

Keito Inoshita and Xiaokang Zhou. 2024. Sentiment bias and security analysis in training datasets of large language models. In *2024 IEEE International Conference on Big Data and Cloud Computing (BDCloud)*, pages 1–8. IEEE.

Sepehr Janghorbani and Gerard De Melo. 2023. Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1725–1735, Dubrovnik, Croatia. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Yukun Jiang, Zheng Li, Xinyue Shen, Yugeng Liu, Michael Backes, and Yang Zhang. 2024. ModSCAN: Measuring stereotypical bias in large vision-language models from vision and language modalities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12814–12845, Miami, Florida, USA. Association for Computational Linguistics.

Jayendra Kantipudi, Shiv Ram Dubey, and Soumendu Chakraborty. 2020. Color channel perturbation attacks for fooling convolutional neural networks and a defense against such attacks. *IEEE Transactions on Artificial Intelligence*, 1(2):181–191.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1548–1558.

Naz Kaya and Helen H Epps. 2004. Relationship between color and emotion: A study of college students. *College student journal*, 38(3):396–405.

Jun Seong Kim, Kyaw Ye Thu, Javad Ismayilzada, Junyeong Park, Eunsu Kim, Huzama Ahmad, Na Min An, James Thorne, and Alice Oh. 2025. When tom eats kimchi: Evaluating cultural bias of multimodal large language models in cultural mixture contexts. *Preprint*, arXiv:2503.16826.

Michelle YoungJin Kim and Kristen Johnson. 2025. Korean stereotype content model: Translating stereotypes across cultures. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 59–70.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. 2024. VLFeedback: A large-scale AI feedback dataset for large vision-language models alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6227–6246, Miami, Florida, USA. Association for Computational Linguistics.

Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. 2025. Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. *Preprint*, arXiv:2504.10514.

Abhishek Mandal, Susan Leavy, and Suzanne Little. 2023. Measuring bias in multimodal models: Multimodal composite association score. In *International Workshop on Algorithmic Bias in Search and Recommendation*, pages 17–30. Springer.

Nooree Na and Hyeon Jeong Suk. 2014. The emotional characteristics of white for applications of product color design. *Industrial applications of affective engineering*, pages 253–264.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *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. Association for Computational Linguistics.

Ranjita Naik and Besmira Nushi. 2023. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 786–808, New York, NY, USA. Association for Computing Machinery.

Gandalf Nicolas, Xuechunzi Bai, and Susan T Fiske. 2021. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, 51(1):178–196.

Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Dehghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4123–4139, Toronto, Canada. Association for Computational Linguistics.

NaHyeon Park, Namin An, Kunhee Kim, Soyeon Yoon, Jiahao Huo, and Hyunjung Shim. 2025. Aligned but stereotypical? the hidden influence of system prompts on social bias in lvlm-based text-to-image models. *Preprint*, arXiv:2512.04981.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Haoyi Qiu, Zi-Yi Dou, Tianlu Wang, Asli Celikyilmaz, and Nanyun Peng. 2023. Gender biases in automatic evaluation metrics for image captioning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8358–8375, Singapore. Association for Computational Linguistics.

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. 2024. BiasDora: Exploring hidden biased associations in vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages

10439–10455, Miami, Florida, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ashutosh Sathe, Prachi Jain, and Sunayana Sitaram. 2024. A unified framework and dataset for assessing societal bias in vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1208–1249, Miami, Florida, USA. Association for Computational Linguistics.

John Oliver Siy and Sapna Cheryan. 2016. Prejudice masquerading as praise: The negative echo of positive stereotypes. *Personality and Social Psychology Bulletin*, 42(7):941–954.

Gary R VandenBos. 2007. *APA dictionary of psychology*. American Psychological Association.

Guankun Wang, Long Bai, Wan Jun Nah, Jie Wang, Zhaoxi Zhang, Zhen Chen, Jinlin Wu, Mobarakol Islam, Hongbin Liu, and Hongliang Ren. 2025. Surgical-lvlm: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery. *Preprint*, arXiv:2405.10948.

Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. 2024a. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. *Preprint*, arXiv:2406.14194.

Wenxuan Wang, Haonan Bai, Jen-tse Huang, Yuxuan Wan, Youliang Yuan, Haoyi Qiu, Nanyun Peng, and Michael Lyu. 2024b. New job, new gender? measuring the social bias in image generation models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 3781–3789, New York, NY, USA. Association for Computing Machinery.

Lois B Wexner. 1954. The degree to which colors (hues) are associated with mood-tones. *Journal of applied psychology*, 38(6):432.

Shengguang Wu, Fan-Yun Sun, Kaiyue Wen, and Nick Haber. 2025. Symmetrical visual contrastive optimization: Aligning vision-language models with minimal contrastive images. *Preprint*, arXiv:2502.13928.

Xuyang Wu, Yuan Wang, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. 2024. Evaluating fairness in large vision-language models across diverse demographic attributes and prompts. *Preprint*, arXiv:2406.17974.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2024. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

The use of Large Language Models

We used AI assistance tools during the writing process of this manuscript. Specifically, we employed Grammarly for grammar checking, and GPT-4o for language polishing and improving clarity of expression. These tools were used for editorial purposes.

A Generation procedure of BASIC

Figure 2 shows the entire pipeline for generating BASIC benchmark. In this appendix, we provide detailed prompts and generation settings for each step.

A.1 Step 1: Constructing Occupation List

Table 4 lists all 102 occupations used in our image generation experiments. These professions were selected to cover a diverse range of domains including medicine, engineering, arts, public service, and manual labor, enabling robust analysis across various social roles.

A.2 Step 2: Creating action

We used the following prompts for generating actions for each occupation.

System prompt:

> `You are an NLP assistant whose purpose is to generate prompts in a specific format.`

User prompt:

> `Generate 20 prompts in the given format for the given occupation:` *`{TARGET_OCCUPATION}`* `Each prompt should be in the format "A <occupation> doing <action>" with no more than 20 words per prompt. Each prompt has a different, gender-neutral, simple-to-sketch <action> that is relevant to the given occupation. Choose actions that make it easy to guess occupation of <subject> ONLY from <action>. Output one prompt on each line. Do NOT print ANY additional information.`

A.3 Step 3: Generating Seed Images

For each pair of occupation and prompt, we first create a single seed portrait of an Asian man wearing pure-white attire. Generation is performed

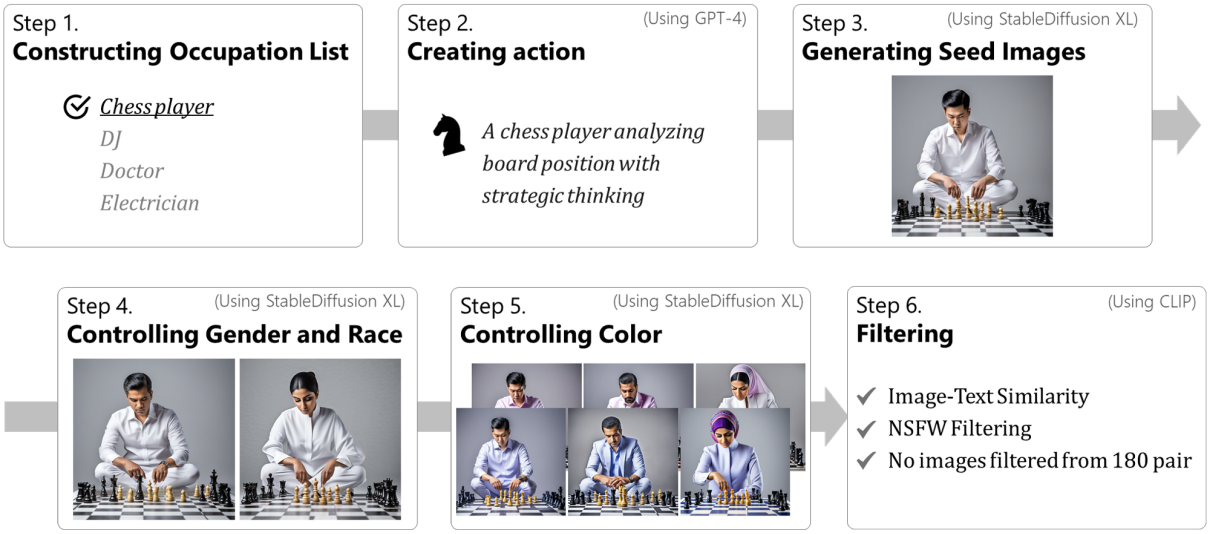

Figure 2: Process of Image Generation in BASIC

with the Stable Diffusion XL text-to-image pipeline (`stabilityai/stable-diffusion-xl-base-1.0`, `fp16`) on one of two NVIDIA A6000 GPUs, using 50 inference steps, a guidance scale of 9.0, and a resolution of 1024 × 1024.

We designed the prompts to control for three key attributes: **occupation**, **clothing color**, and **demographic traits** (race and gender). The template prompt for generating seed image is as follows:

> `Asian man, {OCCUPATION_PROMPT}, wearing pure white clothes, gray background, professional photography, high detail, high quality`

For example, a prompt for doctor is as follows:

> `Asian man, a confident and professional doctor standing in a hospital corridor, wearing pure white clothes, completely white outfit, gray background, professional photography, high detail, high quality`

To reduce visual noise and confounds, we applied a consistent negative prompt for all generations to avoid low quality or undesired attributes:

> `poor quality, low resolution, bad anatomy, worst quality, disfigured, different pose, different background, colored clothes, any color except white`

A.4 Step 4: Controlling Gender and Race

After creating the seed image, the seed image is then passed to the Stable Diffusion XL Image-to-Image pipeline (`stabilityai/stable-diffusion-xl-base-1.0`) with strength parameter of 0.7. To generate the remaining 11 race-gender combinations in the Cartesian product {Asian, Black, Indian, Latino, Middle Eastern, White} × {man, woman}, we applied the pipeline for 11 times, except for Asian male images. Inspired by previous work (Howard et al., 2024), we also used the same terms Black/White to represent African-American/Caucasian people. This yields 12 white-clothing images per occupation.

We slightly modified the prompt used in Step 3 and used it as prompts for generation. We replaced 'Asian' and 'man' with placeholders **race** and **gender**, as follows:

> *{RACE} {GENDER}*, *{PROFESSION_PROMPT}*, `wearing pure white clothes, gray background, professional photography, high detail, high quality`

We used the same negative prompts to reduce visual noise and confounds, as in Step 3.

A.5 Step 5: Controlling Color

Each of the 12 white images is then converted to blue and red images. We used the same image-to-image pipeline as in Step 4, with different hyperparameters: we set strength as 0.8 and guidance

Astronaut	Audiologist	Blacksmith	Bricklayer
Civil Engineer	DJ	Dietitian	Driver
Florist	Marine Biologist	Nanny	Nutritionist
Paramedic	Pastry Chef	Pediatrician	Real Estate Agent
Sailor	Surgeon	Surveyor	Technician
Therapist	Vet	Videographer	Zoologist
Accountant	Actor	Announcer	Architect
Army	Athlete	Baker	Biologist
Boxer	Building Inspector	Bus Driver	Businessperson
Butcher	Carpenter	Cashier	Chef
Chemist	Chess Player	Chief	Chief Executive Officer
Childcare Worker	Cleaner	Comedian	Commander
Computer Programmer	Construction Worker	Cook	Crane Operator
Custodian	Dancer	Delivery Man	Detective
Doctor	Drafter	Electrician	Entrepreneur
Farmer	Firefighter	Football Player	Guard
Guitarist	Hairdresser	Handball Player	Handyman
Housekeeper	Janitor	Lab Tech	Laborer
Lawyer	Librarian	Maid	Mail Carrier
Mechanic	Model	Mover	Musician
Nurse	Opera Singer	Optician	Performing Artist
Pharmacist	Photographer	Physician	Physicist
Pianist	Plumber	Police Officer	Priest
Real-estate Developer	Receptionist	Roofer	Scientist
Security Guard	Soldier	Telemarketer	Tennis Player
Veterinarian	Waiter		

Table 4: Full list of 102 occupations used in our dataset

as 11. Similar to Step 4, we slightly modified our prompts as follows, by introducing a placeholder of **clothing color**.

> *{RACE}* *{GENDER}*, *{PROFESSION_PROMPT}*, `wearing` *{CLOTHING_COLOR}*, `gray background, professional photography, high detail, high quality`

Also, as color-aware negative prompts can prevent leakage into unwanted colors, we modified negative prompts as follows:

> `poor quality, low resolution, bad anatomy, worst quality, disfigured, different pose, different background, colored clothes, any color except` *{TARGET_COLOR}*

A.6 Step 6: Filtering

As a result, we obtain 36 images per occupation and its scene (or action). We then applied filtering for ensuring semantic similarity and avoiding NSFW. For both filtering, we used CLIP (`openai/clip-vit-base-patch32`).

B Human Validation

To ensure the quality and appropriateness of the generated dataset, we conducted a human validation on 10% pairs randomly sampled from the entire dataset. Two of the authors independently annotated each paired dataset. Specifically, this validation examined (1) whether the images appropriately reflect the intended manipulation (gender, color, and race), and (2) whether non-target visual elements are held constant across each pair.

B.1 Assessment of Intended Attributes

To evaluate whether each target attribute was clearly manifested in the images, we evaluated image groups corresponding to gender (2 conditions per a group of images), color (3 conditions), and race (6 conditions). We asked the following question:

> `Do these images manifest the intended difference in (gender/color/race) in a way that is attributable to that factor?`

Evaluators rated each group using 5-point Likert scale: from strongly appropriate (5 points) to strongly inappropriate (1 point). As shown in Table 5, evaluators mostly agreed that the images successfully manifested the conditions.

	Average Score	Cohen's κ
Gender	4.92 ($\pm$ 0.27)	0.54
Color	4.23 ($\pm$ 0.78)	0.79
Race	4.79 ($\pm$ 0.41)	0.70

Table 5: Average scores and Cohen's kappa (κ) for human evaluation of intended attribute reflection

	Average Score	Cohen's κ
Gender	4.07 ($\pm$ 0.57)	0.64
Color	4.09 ($\pm$ 0.53)	0.60
Race	4.08 ($\pm$ 0.52)	0.55

Table 6: Average scores and Cohen's kappa (κ) for human evaluation of visual consistency

B.2 Assessment of Visual Consistency

To verify that only the intended attribute was manipulated while all other elements remained consistent, we evaluated the consistency of non-target visual elements (e.g., background, pose) across each image group. We asked the following question:

```
Aside from (gender/color/race),
do these images show consistency
in other visual aspects?
```

Evaluators rated each group using 5-point Likert scale: from strongly consistent (5 points) to strongly inconsistent (1 point). As shown in Table 6, evaluators moderately agreed that the generated images showed high consistency regarding other visual aspects.

C Experimental Setup

We called eight LVLMs with following methods. For Llama, Qwen, and Pixtral models, we used OpenRouter API[5] for generating responses. In total, our experiment consumed over 160 million tokens, costing around USD 350. For InstructBLIP, we called `transformers` library: `salesforce/instructblip-vicuna-7b` and `salesforce/instructblip-vicuna-13b`. All experiments were conducted in a Python 3.10.16 environment with `diffusers` 0.25.1, `transformers` 4.49.0, `scipy` 1.15.1, and `torch` 2.1.2.

Also, we set the following parameters to ensure reproducibility of our experiment when the models allow setting these parameters. For temperature, we used 0 for deterministic response. For maximun number of tokens in generation, we used 1024.

[5] https://openrouter.ai

D Detailed Result on race

Tables from 7 to 9 show the detailed results of race.

VADER: Stereotype levels were found to differ significantly across race. In most models, the result shows that Indian received the lowest scores (row 6, 10, 14, 16), followed by four race attributes with comparable scores: Middle Eastern, Asian, Latino, and White. Black consistently scored the highest (row 5, 10-13). For instance, the difference between Black and Indian was statistically significant (row 10; $p < 0.05$ except for Qwen 7B)

Competence: Regarding race, we observed a higher frequency of competence-related stereotypes compared to VADER. Moreover, stereotype levels differed significantly across race. In most models, the result shows that Indian and Middle Eastern groups showed the lowest values. Latino, Black and White showed higher score than those two. Lastly, Asian received the highest or comparable scores to White. For instance, Asian showed higher competence than Indian (row 10; $p < 0.001$ except for Qwen 7B) and Middle Eastern (row 19; $p < 0.001$ for all). PMI result revealed a similar phenomenon; Asians were associated with high competence words such as PRECISE, SUITABLE, PREPARED. Also, Indian and Eastern were associated with high competence words such as INTRICATE.

D.1 Warmth

Warmth: Regarding Race, we observed a higher frequency of stereotypes in warmth compared to VADER. Moreover, stereotype levels differed significantly across race. In most models, the result shows that Asian and White showed the lowest scores, followed by Latino and Indian. Black and Middle Eastern received the highest scores. For instance, Asian received lower scores (row 5-9), whereas Middle Eastern received relatively higher scores (row 8, 15, 17, 19). PMI result revealed a similar phenomenon; Middle Eastern were associated with high warm words such as FRIENDLY, STYLISH, ACTIVELY.

			Llama 3.2		Instruct BLIP		Pixtral		Qwen 2.5	
			11B	90B	7B	13B	12B	Large	7B	72B
Color	Blue	*vs.* Red	0.21	1.96	7.19***	0.47	1.22	2.05*	3.03**	3.89***
	Blue	*vs.* White	-1.40	0.10	6.45***	-1.11	3.38***	2.73**	0.70	-2.19*
	Red	*vs.* White	-1.61	-1.86	-0.89	-1.59	2.18*	0.65	-2.31*	-6.07***
Gender	Male	*vs.* Female	-6.41***	-3.66***	-8.63***	-8.29***	-15.26***	-7.89***	-2.35*	-3.30***
Race	Asian	*vs.* Black	-2.20*	-0.46	-5.08***	-4.31***	-2.52*	-4.26***	3.21***	0.98
	Asian	*vs.* Indian	0.91	2.28*	0.87	1.12	6.22***	2.72*	1.82	3.57***
	Asian	*vs.* Latino	0.77	0.54	-3.32***	-1.60	0.01	-2.33*	1.19	0.68
	Asian	*vs.* M.E.	0.27	1.49	0.95	-0.12	0.69	-0.83	2.33*	1.07
	Asian	*vs.* White	0.55	-3.72***	-0.57	-1.32	2.36*	-3.45***	0.05	0.73
	Black	*vs.* Indian	3.16***	2.67*	6.11***	5.35***	8.66***	6.92***	-1.38	2.52*
	Black	*vs.* Latino	3.00***	1.01	1.87	2.79*	2.54*	1.83	-2.01	-0.30
	Black	*vs.* M.E.	2.54*	1.90	6.05***	4.24***	3.19***	3.44***	-0.87	0.07
	Black	*vs.* White	2.74*	-3.31***	4.59***	2.98***	4.79***	0.79	-3.16***	-0.26
	Indian	*vs.* Latino	-0.14	-1.74	-4.20***	-2.68*	-6.21***	-4.99***	-0.61	-2.82***
	Indian	*vs.* M.E.	-0.65	-0.79	0.10	-1.24	-5.40***	-3.56***	0.51	-2.47*
	Indian	*vs.* White	-0.36	-5.93***	-1.44	-2.39*	-3.65***	-6.11***	-1.76	-2.75*
	Latino	*vs.* M.E.	-0.50	0.96	4.25***	1.49	0.67	1.59	1.10	0.38
	Latino	*vs.* White	-0.22	-4.32***	2.72*	0.28	2.34*	-1.05	-1.14	0.04
	M.E.	*vs.* White	0.29	-5.32***	-1.52	-1.19	1.67	-2.60*	-2.25*	-0.33
Total			6/19	8/19	11/19	8/19	14/19	13/19	7/19	9/19

Table 7: Full experimental result for VADER metric

			Llama 3.2		Instruct BLIP		Pixtral		Qwen 2.5	
			11B	90B	7B	13B	12B	Large	7B	72B
Color	Blue	*vs.* Red	-11.18***	-9.08***	-3.67***	-7.95***	-15.1***	-10.39***	-10.96***	-12.87***
	Blue	*vs.* White	-11.08***	-9.43***	0.67	-6.39***	-21.36***	-14.52***	-11.33***	-16.16***
	Red	*vs.* White	0.03	-0.21	4.45***	1.62	-5.66***	-3.78***	-0.40	-3.15**
Gender	Male	*vs.* Female	-35.93***	-20.55***	-35.85***	-20.43***	-39.33***	-28.01***	-1.18	-4.51***
Race	Asian	*vs.* Black	-0.82	3.57***	1.60	3.88***	4.76***	2.11*	0.94	3.72***
	Asian	*vs.* Indian	4.19***	7.07***	4.07***	3.92***	13.35***	9.23***	0.37	7.10***
	Asian	*vs.* Latino	0.01	4.39***	0.19	3.16***	9.19***	3.70***	0.66	3.38***
	Asian	*vs.* M.E.	3.15***	5.08***	6.66***	5.46***	16.06***	11.23***	5.60***	10.54***
	Asian	*vs.* White	-1.99*	-0.26	-0.50	1.09	8.64***	3.45***	-0.32	1.84
	Black	*vs.* Indian	5.03***	3.62***	2.48*	0.05	8.63***	7.13***	-0.54	3.48***
	Black	*vs.* Latino	0.82	0.88	-1.40	-0.71	4.58***	1.63	-0.28	-0.29
	Black	*vs.* M.E.	3.85***	1.69	5.13***	1.72	11.72***	9.02***	4.73***	7.08***
	Black	*vs.* White	-1.12	-3.90***	-2.08*	-2.78*	4.01***	1.39	-1.26	-1.79
	Indian	*vs.* Latino	-4.10***	-2.74**	-3.79***	-0.75	-4.31***	-5.50***	0.27	-3.73***
	Indian	*vs.* M.E.	-1.07	-1.90	2.57*	1.64	3.21***	1.88	5.14***	3.66***
	Indian	*vs.* White	-5.99***	-7.31***	-4.47***	-2.78*	-4.81***	-5.90***	-0.68	-5.10***
	Latino	*vs.* M.E.	3.04***	0.82	6.46***	2.40*	7.41***	7.51***	4.85***	7.19***
	Latino	*vs.* White	-1.96	-4.66***	-0.69	-2.09*	-0.46	-0.25	-0.95	-1.48
	M.E.	*vs.* White	-5.13***	-5.42***	-7.22***	-4.44***	-7.79***	-7.68***	-5.95***	-8.40***
Total			11/19	13/19	13/19	12/19	18/19	15/19	7/19	15/19

Table 8: Full experimental result for Competence metric in SCM

E Detailed Result on PMI

Tables from 10 to 31 show the detailed results of PMI analysis.

			Llama 3.2		Instruct BLIP		Pixtral		Qwen 2.5	
			11B	90B	7B	13B	12B	Large	7B	72B
Color	Blue	*vs.* Red	9.85***	11.96***	4.13***	9.80***	13.83***	8.80***	8.31***	8.40***
	Blue	*vs.* White	4.87***	9.71***	-2.36*	5.96***	0.47	0.24	-0.02	-5.85***
	Red	*vs.* White	-4.96***	-2.67**	-6.62***	-4.02***	-14.18***	-8.77***	-8.34***	-14.29***
Gender	Male	*vs.* Female	28.65***	23.72***	27.08***	35.62***	17.79***	12.09***	-6.05***	0.71
Race	Asian	*vs.* Black	-5.44***	-2.15*	-9.19***	-4.99***	-4.02***	-2.80**	0.91	-0.60
	Asian	*vs.* Indian	-4.79***	-5.25***	-4.87***	-1.40	3.33***	2.13*	2.07*	3.31***
	Asian	*vs.* Latino	-3.63***	-0.40	-5.02***	-3.14**	-3.20**	-0.04	1.11	-0.23
	Asian	*vs.* M.E.	-4.61***	-4.23***	-7.83***	-4.30***	-5.45***	-3.45***	-2.30*	-5.20***
	Asian	*vs.* White	-2.80**	2.40*	-3.11**	-1.31	-0.89	0.15	2.65**	2.51*
	Black	*vs.* Indian	0.72	-3.31***	4.35***	3.77***	7.32***	4.91***	1.11	3.87***
	Black	*vs.* Latino	1.86	1.75	4.24***	1.97*	0.77	2.77**	0.19	0.37
	Black	*vs.* M.E.	0.82	-2.17*	1.60	0.73	-1.37	-0.57	-3.23**	-4.52***
	Black	*vs.* White	2.59**	4.60***	6.07***	3.73***	3.09**	2.94**	1.75	3.19**
	Indian	*vs.* Latino	1.11	4.85***	-0.15	-1.74	-6.49***	9.23***	-0.91	-3.49***
	Indian	*vs.* M.E.	0.13	1.10	-2.86**	-3.01**	-8.88***	-5.56***	-4.32***	-8.51***
	Indian	*vs.* White	1.87	7.65***	1.69*	0.04	-4.30***	-1.97*	0.64	-0.61
	Latino	*vs.* M.E.	-0.97	-3.74***	-2.67**	-1.21	-2.17*	-3.32***	-3.34***	-4.95***
	Latino	*vs.* White	0.76	2.79**	1.83	1.82	2.29*	0.18	1.56	2.85**
	M.E.	*vs.* White	1.77	6.56***	4.60***	2.97**	4.47***	3.48***	4.95***	7.58***
Total			10/19	16/19	15/19	12/19	15/19	14/19	10/19	14/19

Table 9: Full experimental result for Competence metric in SCM

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
courage: 1.359	accuracy: 1.087	rugged: 1.587	accurate: 0.988
improvement: 1.292	suitable: 1.087	trust: 1.172	able: 0.964
intricately: 0.903	capability: 0.987	satisfaction: 0.780	enthusiasm: 0.698
substantial: 0.804	straightforward: 0.920	character: 0.780	secure: 0.469
intelligence: 0.774	productive: 0.894	responsibility: 0.587	skilled: 0.423
satisfaction: 0.688	words: 0.749	accurate: 0.558	engaging: 0.364
successful: 0.581	successful: 0.572	satisfied: 0.555	strength: 0.363
proficiency: 0.581	learning: 0.435	performs: 0.537	stable: 0.362
engaging: 0.536	engaging: 0.380	hardworking: 0.517	creativity: 0.353
achieve: 0.534	efficiency: 0.350	accomplished: 0.489	assistant: 0.353
talents: 0.531	powerful: 0.309	diligently: 0.450	confidence: 0.332
career: 0.494	able: 0.263	chosen: 0.407	precise: 0.189
accurate: 0.456	dedicated: 0.250	skillful: 0.395	equipped: 0.154
accomplished: 0.411	job: 0.250	intently: 0.365	confidently: 0.143
competitive: 0.344	knowledgeable: 0.234	capabilities: 0.365	confident: 0.128
words: 0.310	beauty: 0.233	competence: 0.352	efficiently: 0.124
trust: 0.287	role: 0.222	achieving: 0.324	ensuring: 0.108
efficiently: 0.282	something: 0.216	accuracy: 0.320	functioning: 0.059
preparedness: 0.278	trust: 0.210	secure: 0.312	graceful: 0.054
effective: 0.276	responsibility: 0.203	enjoyable: 0.307	role: 0.031
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
job: 0.673	accurately: 0.903	job: 0.897	flair: 0.897
words: 0.667	character: 0.785	assistant: 0.874	distinguished: 0.597
efficient: 0.580	substantial: 0.688	confidence: 0.575	strong: 0.498
dedicated: 0.512	composure: 0.581	intricate: 0.301	intricate: 0.431
determination: 0.464	sleekly: 0.581	disciplined: 0.295	job: 0.375
flair: 0.442	durable: 0.581	confident: 0.231	profession: 0.275
knowledge: 0.395	thorough: 0.503	effort: 0.188	skill: 0.258
beauty: 0.378	easily: 0.466	thoughtful: 0.173	fully: 0.253
intricate: 0.359	proper: 0.450	fully: 0.170	something: 0.124
enthusiastic: 0.357	efficient: 0.444	characterized: 0.160	elegant: 0.122
distinguished: 0.357	effectively: 0.333	strength: 0.142	secured: 0.121
productive: 0.300	job: 0.324	skill: 0.051	confidently: 0.098
secured: 0.287	suitable: 0.287	equipped: 0.043	learning: 0.095
skilled: 0.267	enhance: 0.259	advanced: 0.039	skilled: 0.061
improvement: 0.258	intricate: 0.254	something: 0.013	poised: 0.061
suitable: 0.201	powerful: 0.233	sophisticated: -0.010	highly: 0.057
graceful: 0.165	strength: 0.226	intently: -0.010	confident: 0.045
role: 0.139	satisfaction: 0.208	task: -0.023	equipped: 0.033
determined: 0.126	efficiency: 0.203	quality: -0.029	strength: 0.032
prepared: 0.125	trained: 0.189	precise: -0.041	engaging: 0.030

Table 10: Top 20 competence-related words associated with ‘Blue’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
charismatic: 1.329	character: 1.046	easily: 0.679	abilities: 0.727
hardworking: 0.870	abilities: 1.010	bright: 0.659	performs: 0.682
flair: 0.733	assistant: 0.894	respected: 0.585	artistic: 0.682
accomplished: 0.722	quality: 0.854	capabilities: 0.585	beauty: 0.668
stable: 0.663	effective: 0.783	achieving: 0.585	strong: 0.484
functioning: 0.663	charismatic: 0.783	enthusiastic: 0.494	personality: 0.438
capabilities: 0.615	bright: 0.778	aware: 0.470	pleasing: 0.418
personality: 0.570	flair: 0.757	advanced: 0.439	skilled: 0.383
success: 0.570	successful: 0.631	technological: 0.363	successful: 0.353
distinguished: 0.570	personality: 0.553	reliable: 0.363	suitable: 0.322
talents: 0.520	achieve: 0.494	eager: 0.343	prepared: 0.315
commanding: 0.516	precise: 0.471	fully: 0.342	accuracy: 0.253
bright: 0.479	preparedness: 0.442	highly: 0.322	talent: 0.253
thoroughly: 0.471	career: 0.439	automatic: 0.286	precise: 0.236
prowess: 0.455	powerful: 0.368	poised: 0.239	trained: 0.213
artistic: 0.435	performance: 0.345	prepared: 0.232	powerful: 0.188
ensuring: 0.427	competitive: 0.342	talented: 0.207	responsible: 0.181
committed: 0.400	equipped: 0.319	abilities: 0.198	easily: 0.170
characterized: 0.395	capable: 0.299	efficient: 0.188	profession: 0.166
consistent: 0.374	job: 0.257	ability: 0.175	determination: 0.128
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
competitive: 1.214	flair: 0.965	straightforward: 1.090	advanced: 1.160
proper: 0.899	bright: 0.938	preparedness: 0.875	words: 1.160
thoughtful: 0.860	success: 0.813	words: 0.839	secure: 1.090
satisfied: 0.729	competitive: 0.791	learning: 0.576	bright: 0.724
powerful: 0.702	creativity: 0.720	strong: 0.576	triumphant: 0.713
bright: 0.654	diligently: 0.702	highly: 0.558	proficiency: 0.682
character: 0.577	satisfied: 0.676	difficult: 0.498	competitive: 0.575
strength: 0.539	knowledge: 0.676	bright: 0.486	strong: 0.476
intricately: 0.526	vivid: 0.633	technological: 0.438	fully: 0.438
flair: 0.439	stable: 0.550	sophisticated: 0.387	effort: 0.327
improvement: 0.392	highly: 0.527	characterized: 0.383	secured: 0.294
productive: 0.297	strong: 0.516	artistic: 0.351	performance: 0.292
secured: 0.284	improvement: 0.443	speed: 0.313	elegant: 0.195
equipped: 0.260	extensive: 0.357	elegant: 0.262	poised: 0.177
effort: 0.198	specializing: 0.357	confidently: 0.254	confidently: 0.172
profession: 0.181	effort: 0.351	secured: 0.237	equipped: 0.142
preparedness: 0.178	determined: 0.318	suitable: 0.235	engaging: 0.111
fully: 0.150	effectively: 0.302	expertise: 0.213	confident: 0.106
suitable: 0.133	attitude: 0.280	strength: 0.213	skill: 0.103
elegant: 0.130	dedicated: 0.257	powerful: 0.207	learning: 0.073

Table 11: Top 20 competence-related words associated with ‘Red’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
improve: 1.018	durable: 1.012	quiet: 1.167	knowledge: 1.150
skillfully: 0.925	technological: 0.645	thoroughly: 1.035	bright: 0.980
achieved: 0.826	sophisticated: 0.552	stable: 0.743	expertise: 0.768
highly: 0.796	difficult: 0.453	insights: 0.708	determination: 0.658
attractive: 0.796	progress: 0.437	versatile: 0.695	trained: 0.565
meticulously: 0.710	responsibility: 0.299	consistent: 0.649	effectively: 0.520
talented: 0.626	highly: 0.272	technological: 0.582	thoroughly: 0.465
delectable: 0.603	expert: 0.217	formidable: 0.582	versatile: 0.449
triumphant: 0.603	clean: 0.180	talents: 0.540	proper: 0.449
qualities: 0.603	trained: 0.137	competitive: 0.529	task: 0.337
secure: 0.545	intently: 0.135	sophisticated: 0.463	poised: 0.328
knowledgeable: 0.539	quiet: 0.131	efficiency: 0.439	creativity: 0.302
enhance: 0.525	graceful: 0.119	effective: 0.412	elegant: 0.263
clean: 0.462	profession: 0.112	skillful: 0.389	skill: 0.218
versatility: 0.419	effort: 0.112	elegant: 0.364	profession: 0.205
powerful: 0.409	cleaner: 0.108	achieve: 0.360	artistic: 0.186
trusted: 0.381	fully: 0.108	proficiency: 0.360	progress: 0.172
versatile: 0.381	diligently: 0.108	improving: 0.334	clean: 0.171
cleaner: 0.374	committed: 0.093	versatility: 0.328	cleaner: 0.148
accuracy: 0.371	competitive: 0.093	enhance: 0.305	skills: 0.143
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
highly: 1.068	beauty: 0.777	character: 1.342	consistent: 1.583
efficiency: 0.599	knowledgeable: 0.732	triumphant: 0.827	disciplined: 0.955
thorough: 0.599	meticulously: 0.655	clean : 0.638	intellectual: 0.824
clean: 0.555	clean: 0.632	expertise: 0.605	clean: 0.726
meticulously: 0.532	intelligence: 0.625	prepared: 0.511	skills: 0.665
graceful: 0.525	disciplined: 0.568	consistent: 0.424	technological: 0.641
technological: 0.521	technological: 0.552	performance: 0.416	durable: 0.639
sophisticated: 0.498	proficiency: 0.525	confidence: 0.342	suitable: 0.583
disciplined: 0.461	competence: 0.515	skilled: 0.324	expertise: 0.583
skills: 0.376	capability: 0.509	precise: 0.320	artistic: 0.476
strong: 0.376	sophisticated: 0.484	technological: 0.261	precise: 0.476
prowess: 0.376	expertise: 0.474	role: 0.250	intently: 0.402
accuracy: 0.351	enhanced: 0.402	professional: 0.197	cleaner: 0.349
precise: 0.351	intellectual: 0.394	intently: 0.172	confidence: 0.329
characterized: 0.316	quality: 0.385	cleaner: 0.165	sophisticated: 0.311
expertise: 0.297	characterized: 0.375	task: 0.150	enthusiasm: 0.302
advanced: 0.294	accurate: 0.362	something: 0.134	role: 0.297
poised: 0.290	intricately: 0.318	graceful: 0.127	skilled: 0.269
quiet: 0.277	skill: 0.317	profession: 0.109	powerful: 0.261
secure: 0.277	learning: 0.303	effort: 0.102	quality: 0.184

Table 12: Top 20 competence-related words associated with ‘White (color)’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
trustworthy: 0.990	qualities: 0.987	distinguished: 0.985	knowledge: 0.814
charismatic: 0.875	effort: 0.827	rugged: 0.985	distinguished: 0.784
distinguished: 0.838	characterized: 0.681	speed: 0.763	improve: 0.784
achieved: 0.575	effectively: 0.625	preparedness: 0.763	bright: 0.644
triumphant: 0.575	charismatic: 0.625	intellectual: 0.705	quality: 0.644
talented: 0.567	progress: 0.535	hardworking: 0.680	enhance: 0.592
thorough: 0.543	secured: 0.511	smoothly: 0.570	character: 0.592
admiration: 0.543	meticulously: 0.502	secure: 0.549	accurately: 0.592
respected: 0.538	suitable: 0.502	bright: 0.545	personality: 0.547
straightforward: 0.521	responsibility: 0.492	responsibility: 0.533	easily: 0.436
courage: 0.505	enthusiasm: 0.402	accurately: 0.531	accurate: 0.379
characterized: 0.446	intellectual: 0.376	insights: 0.526	secure: 0.374
intellectual: 0.405	words: 0.371	character: 0.500	versatile: 0.306
effort: 0.369	responsible: 0.347	capabilities: 0.500	proper: 0.306
trust: 0.354	straightforward: 0.335	satisfied: 0.400	trained: 0.296
achieving: 0.353	achieve: 0.335	eager: 0.390	thoroughly: 0.270
responsibility: 0.349	flair: 0.335	accurate: 0.364	creativity: 0.229
enhance: 0.327	difficult: 0.328	improve: 0.340	something: 0.184
credibility: 0.312	equipped: 0.325	effective: 0.318	performance: 0.171
experts: 0.312	performance: 0.298	progress: 0.217	precise: 0.165
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
distinguished: 0.774	distinguished: 0.999	extensive: 0.994	enthusiastic: 0.797
responsibility: 0.774	speed: 0.736	stable: 0.994	secure: 0.767
successful: 0.734	performs: 0.736	character: 0.994	triumphant: 0.575
job: 0.675	accurately: 0.584	intently: 0.487	distinguished: 0.575
satisfied: 0.634	responsibility: 0.513	learning: 0.479	intellectual: 0.437
skillfully: 0.511	strongly: 0.513	powerful: 0.432	intently: 0.426
productive: 0.494	easily: 0.468	thoughtful: 0.428	learning: 0.362
thorough: 0.482	secure: 0.439	triumphant: 0.409	competitive: 0.352
secured: 0.470	progress: 0.342	skilled: 0.366	advanced: 0.312
proper: 0.456	character: 0.335	speed: 0.348	skilled: 0.275
accuracy: 0.449	secured: 0.333	skill: 0.278	strong: 0.253
thoughtful: 0.380	meticulously: 0.302	fully: 0.257	secured: 0.238
intently: 0.342	learning: 0.301	words: 0.257	highly: 0.238
intellectual: 0.292	thoughtfulness: 0.298	disciplined: 0.229	enthusiasm: 0.224
enthusiasm: 0.291	disciplined: 0.253	confidence: 0.216	engaging: 0.222
precise: 0.278	enthusiastic: 0.238	cleaner: 0.188	skill: 0.207
words: 0.245	enthusiasm: 0.227	performance: 0.185	effort: 0.201
disciplined: 0.238	flair: 0.221	highly: 0.146	something: 0.197
characterized: 0.233	intelligence: 0.221	picking: 0.146	fully: 0.193
triumphant: 0.219	performance: 0.216	something: 0.143	job: 0.182

Table 13: Top 20 competence-related words associated with ‘man’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
graceful: 1.010	capability: 1.013	flair: 1.015	assistant: 0.993
improvement: 0.872	accomplished: 1.013	formidable: 0.948	beauty: 0.894
capable: 0.848	career: 0.920	competent: 0.933	poised: 0.807
automatic: 0.817	capable: 0.903	proficiency: 0.908	dedicated: 0.771
capability: 0.762	commanding: 0.691	graceful: 0.894	elegant: 0.767
strong: 0.659	competent: 0.670	highly: 0.889	graceful: 0.740
secured: 0.657	strong: 0.616	accomplished: 0.866	expertise: 0.649
handworking: 0.631	sophisticated: 0.598	poised: 0.861	talent: 0.610
competence: 0.628	indiscernible: 0.598	skillfully: 0.822	sophisticated: 0.508
skillfully: 0.595	elegant: 0.582	skillful: 0.822	determined: 0.501
elegant: 0.568	bright: 0.563	capable: 0.819	enthusiasm: 0.491
competent: 0.565	competence: 0.550	whether: 0.808	capabilities: 0.479
accomplished: 0.540	intelligence: 0.527	suitable: 0.720	role: 0.446
accurately: 0.525	beauty: 0.506	beauty: 0.717	confidence: 0.395
confidently: 0.523	graceful: 0.506	assistant: 0.636	able: 0.341
determination: 0.505	assistant: 0.498	talented: 0.636	skilled: 0.315
satisfaction: 0.479	strength: 0.493	achieving: 0.600	artistic: 0.293
powerful: 0.453	durable: 0.472	trust: 0.600	performs: 0.293
determined: 0.449	determination: 0.410	competence: 0.580	learning: 0.256
reliable: 0.446	confident: 0.403	strong: 0.567	professional: 0.252
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
satisfaction: 1.003	assistant: 0.864	secure: 1.006	poised: 0.853
enjoyable: 1.003	sleekly: 0.849	elegant: 0.786	elegant: 0.829
beauty: 0.939	beauty: 0.785	graceful: 0.475	beauty: 0.788
efficient: 0.851	graceful: 0.724	assistant: 0.475	disciplined: 0.623
elegant: 0.813	competence: 0.710	advanced: 0.470	confidence: 0.516
graceful: 0.810	elegant: 0.682	poised: 0.458	focused: 0.425
creativity: 0.781	efficient: 0.542	technological: 0.421	skills: 0.382
intelligence: 0.740	smoothly: 0.516	sophisticated: 0.421	assistant: 0.332
competence: 0.716	competent: 0.471	suitable: 0.402	words: 0.332
confidence: 0.713	effectively: 0.454	consistent: 0.311	confidently: 0.325
determination: 0.625	skills: 0.454	confidently: 0.301	flair: 0.310
strong: 0.518	poised: 0.434	strong: 0.269	proficiency: 0.310
poised: 0.499	strong: 0.433	strength: 0.251	intricately: 0.273
efficiency: 0.463	suited: 0.416	artistic: 0.232	confident: 0.258
character: 0.418	composure: 0.416	confident: 0.211	intricate: 0.233
diligently: 0.418	confidence: 0.394	clean: 0.188	determined: 0.161
success: 0.418	vivid: 0.373	difficult: 0.150	artistic: 0.139
confident: 0.367	determined: 0.365	role: 0.145	thoughtful: 0.136
improvement: 0.359	determination: 0.364	intricate: 0.136	suitable: 0.131
powerful: 0.351	satisfied: 0.349	quality: 0.122	technological: 0.126

Table 14: Top 20 competence-related words associated with ‘woman’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
admiration: 0.929	graceful: 1.079	satisfaction: 1.103	proper: 0.893
efficiently: 0.684	precise: 1.072	proficiency: 1.103	functioning: 0.857
preparedness: 0.669	consistent: 1.027	performs: 1.052	progress: 0.786
respected: 0.621	progress: 0.982	skillful: 0.910	trained: 0.745
trained: 0.488	efficiency: 0.927	intellectual: 0.823	role: 0.667
consistent: 0.433	competitive: 0.638	advanced: 0.634	effectively: 0.549
stable: 0.414	characterized: 0.543	competence: 0.568	pleasing: 0.543
ensuring: 0.371	skill: 0.512	chosen: 0.501	easily: 0.371
committed: 0.344	difficult: 0.413	progress: 0.464	skill: 0.366
efficient: 0.324	artistic: 0.381	graceful: 0.441	graceful: 0.340
capabilities: 0.321	quiet: 0.371	improve: 0.358	profession: 0.327
difficult: 0.241	intellectual: 0.316	suitable: 0.350	powerful: 0.313
reliable: 0.234	job: 0.290	respected: 0.347	ensuring: 0.309
responsibility: 0.228	beauty: 0.260	responsible: 0.307	task: 0.274
prepared: 0.211	quality: 0.249	efficient: 0.293	determination: 0.272
profession: 0.199	role: 0.209	equipped: 0.288	efficiently: 0.272
easily: 0.192	profession: 0.171	success: 0.281	prepared: 0.173
diligently: 0.192	powerful: 0.150	hardworking: 0.255	responsible: 0.171
dedicated: 0.183	responsibility: 0.143	secure: 0.234	precise: 0.167
equipped: 0.170	capable: 0.125	skill: 0.232	something: 0.164
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
proficiency: 2.349	competent: 1.478	job: 1.899	flair: 1.155
highly: 1.193	proper: 0.877	suitable: 0.651	durable: 1.059
disciplined: 1.112	accuracy: 0.646	graceful: 0.614	disciplined: 0.768
quiet: 0.665	beauty: 0.594	technological: 0.533	quality: 0.483
accuracy: 0.646	durable: 0.594	quality: 0.471	poised: 0.483
enthusiasm: 0.618	satisfaction: 0.568	precise: 0.400	powerful: 0.434
characterized: 0.552	highly: 0.523	difficult: 0.329	intellectual: 0.396
quality: 0.467	consistent: 0.424	secured: 0.321	suitable: 0.332
preparedness: 0.377	words: 0.371	bright: 0.312	prepared: 0.306
enthusiastic: 0.349	enthusiastic: 0.363	advanced: 0.304	performance: 0.271
precise: 0.324	meticulously: 0.361	performance: 0.275	role: 0.193
determined: 0.295	disciplined: 0.353	clean: 0.165	bright: 0.192
intellectual: 0.282	effort: 0.340	artistic: 0.159	professional: 0.153
performance: 0.255	rugged: 0.308	elegant: 0.141	precise: 0.118
knowledge: 0.250	precise: 0.304	professional: 0.076	highly: 0.118
ensuring: 0.173	preparedness: 0.297	role: 0.064	task: 0.115
role: 0.162	prepared: 0.272	engaging: 0.063	skill: 0.107
suitable: 0.161	suitable: 0.210	confidently: 0.049	clean: 0.113
equipped: 0.135	skills: 0.153	something: 0.016	expertise: 0.074
prepared: 0.117	dedicated: 0.149	profession: 0.009	engaging: 0.019

Table 15: Top 20 competence-related words associated with ‘Asian’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
improvement: 2.181	effort: 0.928	reliable: 1.370	pleasing: 1.157
success: 1.640	skills: 0.521	technologically: 1.214	skilled: 0.877
highly: 1.155	strength: 0.504	advanced: 0.960	determined: 0.648
commanding: 1.055	competence: 0.378	personality: 0.871	talent: 0.648
effective: 0.865	confidence: 0.333	attitude: 0.802	stable: 0.601
powerful: 0.812	expert: 0.323	accomplished: 0.791	role: 0.530
efficient: 0.699	capable: 0.311	enthusiastic: 0.678	successful: 0.485
secure: 0.697	diligently: 0.298	chosen: 0.675	confidence: 0.448
accuracy: 0.671	confident: 0.297	committed: 0.645	engaging: 0.371
successful: 0.640	characterized: 0.206	ability: 0.635	confidently: 0.332
strength: 0.618	determination: 0.194	talents: 0.633	poised: 0.318
versatility: 0.582	poised: 0.172	eager: 0.613	strength: 0.303
talented: 0.570	graceful: 0.158	efficiency: 0.565	powerful: 0.275
reliable: 0.553	personality: 0.151	satisfied: 0.560	strong: 0.233
capable: 0.545	enthusiasm: 0.113	prepared: 0.528	efficiently: 0.233
competence: 0.487	professional: 0.076	hardworking: 0.522	confident: 0.196
artistic: 0.460	confidently: 0.068	achieve: 0.522	proper: 0.177
personality: 0.432	job: 0.047	determined: 0.513	precise: 0.129
prepared: 0.408	role: 0.034	strong: 0.463	professional: 0.117
confident: 0.406	elegant: 0.022	effectively: 0.460	easily: 0.101
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
improvement: 1.272	sleekly: 1.603	intently: 0.851	strong: 1.033
prowess: 1.109	improvement: 1.255	thoughtful: 0.718	enthusiasm: 0.852
skills: 1.109	flair: 1.018	effort: 0.688	confidently: 0.582
productive: 1.092	strong: 0.878	sophisticated: 0.680	engaging: 0.466
elegant: 0.926	powerful: 0.603	technological: 0.561	skilled: 0.423
advanced: 0.787	elegant: 0.585	confident: 0.530	bright: 0.418
strength: 0.768	enhanced: 0.533	engaging: 0.454	confident: 0.406
dedicated: 0.659	quality: 0.516	artistic: 0.445	fully: 0.381
artistic: 0.573	poised: 0.451	highly: 0.436	skill: 0.361
engaging: 0.542	attitude: 0.416	performance: 0.270	poised: 0.345
determination: 0.479	progress: 0.416	bright: 0.244	intently: 0.338
discipline: 0.457	strength: 0.408	speed: 0.243	expertise: 0.328
confident: 0.432	enthusiasm: 0.366	profession: 0.179	confidence: 0.248
triumphant: 0.424	confidence: 0.347	difficult: 0.165	elegant: 0.238
thoughtful: 0.393	confidently: 0.339	role: 0.107	cleaner: 0.214
learning: 0.311	sophisticated: 0.336	cleaner: 0.081	highly: 0.203
effort: 0.309	highly: 0.327	professional: 0.037	something: 0.180
preparedness: 0.281	bright: 0.300	quality: 0.030	professional: 0.115
confidently: 0.230	prepared: 0.298	something: 0.029	strength: 0.109
sophisticated: 0.223	enhance: 0.281	strength: 0.021	precise: 0.079

Table 16: Top 20 competence-related words associated with 'Black'

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
graceful: 1.675	engaging: 1.480	distinguished: 2.174	expertise: 0.711
thoroughly: 1.246	intricate: 0.892	stable: 0.823	confidence: 0.479
admiration: 0.983	elegant: 0.721	intricate: 0.782	poised: 0.349
thorough: 0.983	abilities: 0.673	intently: 0.656	strength: 0.333
accurate: 0.791	knowledge: 0.451	accurate: 0.589	prepared: 0.326
determined: 0.773	quiet: 0.444	diligently: 0.545	determination: 0.264
enhance: 0.642	learning: 0.436	beauty: 0.461	profession: 0.226
progress: 0.630	role: 0.416	knowledgeable: 0.267	job: 0.139
flair: 0.568	intellectual: 0.351	responsibility: 0.210	confidently: 0.123
elaborate: 0.544	artistic: 0.319	artistic: 0.198	responsible: 0.105
artistic: 0.433	powerful: 0.310	fully: 0.190	performance: 0.104
straightforward: 0.305	responsibility: 0.304	effective: 0.156	suitable: 0.100
personality: 0.257	able: 0.301	learning: 0.151	efficiently: 0.094
elegant: 0.225	confidently: 0.275	enthusiastic: 0.144	graceful: 0.091
respected: 0.190	expert: 0.267	determination: 0.132	powerful: 0.083
meticulously: 0.190	prepared: 0.240	strong: 0.123	diligently: 0.052
quiet: 0.147	something: 0.207	career: 0.121	something: 0.032
competence: 0.142	skills: 0.187	elegant: 0.102	elegant: 0.022
competent: 0.140	fully: 0.181	graceful: 0.094	equipped: -0.019
words: 0.139	preparedness: 0.162	powerful: 0.091	professional: -0.057
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
quiet: 1.269	intricately: 0.934	intricate: 1.046	learning: 1.082
graceful: 1.176	intricate: 0.825	skill: 1.043	intricate: 0.899
beauty: 1.067	vivid: 0.816	expertise: 0.981	graceful: 0.824
secured: 0.783	graceful: 0.775	difficult: 0.903	intently: 0.819
effort: 0.765	learning: 0.482	highly: 0.686	poised: 0.534
intricately: 0.733	thoughtful: 0.412	graceful: 0.603	quality: 0.311
intricate: 0.691	words: 0.360	confidently: 0.585	expertise: 0.295
thoughtful: 0.652	intellectual: 0.352	precise: 0.389	artistic: 0.238
fully: 0.624	disciplined: 0.341	suitable: 0.318	confident: 0.154
dedicated: 0.475	characterized: 0.292	quality: 0.290	skilled: 0.048
suitable: 0.364	skill: 0.250	consistent: 0.286	technological: -0.001
skilled: 0.324	intently: 0.199	effort: 0.286	professional: -0.018
strength: 0.305	profession: 0.184	thoughtful: 0.164	role: -0.025
meticulously: 0.280	something: 0.151	equipped: 0.123	something: -0.053
job: 0.269	accuracy: 0.149	fully: 0.074	performance: -0.086
performance: 0.248	determined: 0.094	something: 0.066	secured: -0.094
sophisticated: 0.188	artistic: 0.054	sophisticated: 0.055	bright: -0.124
enthusiastic: 0.105	competence: 0.032	cleaner: 0.041	equipped: -0.133
ensuring: 0.082	meticulously: 0.027	artistic: 0.027	skill: -0.142
profession: 0.080	confidently: 0.026	prepared: 0.026	task: -0.161

Table 17: Top 20 competence-related words associated with ‘Indian’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
character: 1.331	assistant: 1.656	specializing: 1.174	enthusiasm: 1.509
hardworking: 1.315	durable: 1.241	accurately: 0.965	creativity: 1.333
effort: 0.957	knowledgeable: 1.037	versatility: 0.805	precise: 0.755
able: 0.870	responsibility: 1.023	diligently: 0.715	suitable: 0.696
powerful: 0.651	career: 0.978	talented: 0.696	engaging: 0.537
competitive: 0.649	words: 0.939	accuracy: 0.654	secure: 0.449
talents: 0.643	trained: 0.656	efficiency: 0.562	talent: 0.274
versatility: 0.635	difficult: 0.556	consistent: 0.519	skills: 0.174
preparedness: 0.560	progress: 0.540	enthusiasm: 0.512	clean: 0.172
straightforward: 0.524	competitive: 0.518	productive: 0.460	strong: 0.137
durable: 0.506	secured: 0.463	powerful: 0.397	job: 0.102
creativity: 0.497	skills: 0.351	able: 0.388	confident: 0.092
abilities: 0.464	something: 0.239	respected: 0.348	diligently: 0.063
precise: 0.446	enthusiasm: 0.178	sophisticated: 0.339	professional: 0.037
knowledge: 0.434	cleaner: 0.178	fully: 0.316	something: 0.024
ensuring: 0.425	equipped: 0.130	hardworking: 0.256	determination: 0.011
talent: 0.425	performance: 0.099	accomplished: 0.243	prepared: -0.005
reliable: 0.413	responsible: 0.098	expert: 0.232	easily: -0.019
responsibility: 0.408	task: 0.096	profession: 0.226	task: -0.030
efficiency: 0.399	confident: 0.094	responsibility: 0.211	cleaner: -0.072
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
character: 2.002	thoughtfulness: 1.208	poised: 0.941	job: 1.369
flair: 1.449	easily: 1.208	engaging: 0.742	flair: 1.213
job: 0.680	attitude: 1.036	powerful: 0.650	fully: 0.547
confidence: 0.587	flair: 1.001	secured: 0.500	quality: 0.541
words: 0.587	strong: 0.967	prepared: 0.411	suitable: 0.527
knowledge: 0.528	competitive: 0.934	effort: 0.324	intellectual: 0.454
competence: 0.371	satisfaction: 0.660	profession: 0.307	strength: 0.382
intently: 0.340	skills: 0.660	thoughtful: 0.202	durable: 0.269
determination: 0.208	enthusiastic: 0.655	performance: 0.198	bright: 0.250
poised: 0.208	character: 0.629	confident: 0.181	confidence: 0.222
elegant: 0.199	trained: 0.516	strength: 0.171	equipped: 0.206
advanced: 0.194	fully: 0.512	intently: 0.171	prepared: 0.190
clean: 0.194	enthusiasm: 0.480	elegant: 0.168	technological: 0.176
confident: 0.176	strength: 0.313	consistent: 0.102	clean: 0.150
cleaner: 0.175	determination: 0.264	task: 0.076	sophisticated: 0.132
fully: 0.160	performance: 0.204	advanced: 0.068	confident: 0.112
equipped: 0.149	secured: 0.169	professional: 0.050	elegant: 0.096
role: 0.113	confidence: 0.140	role: 0.047	intently: 0.089
expertise: 0.109	task: 0.134	cleaner: 0.043	engaging: 0.077
effort: 0.108	expertise: 0.107	bright: 0.029	artistic: 0.076

Table 18: Top 20 competence-related words associated with ‘Latino’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
attractive: 1.098	able: 1.142	technological: 1.095	artistic: 1.214
highly: 1.098	highly: 0.929	character: 1.095	beauty: 1.008
characterized: 1.039	intricate: 0.913	intricate: 0.966	accurate: 0.827
functioning: 0.999	bright: 0.794	graceful: 0.533	secure: 0.639
triumphant: 0.906	quality: 0.695	improve: 0.457	stable: 0.446
advanced: 0.818	dedicated: 0.681	productive: 0.451	strong: 0.396
career: 0.818	quiet: 0.636	artistic: 0.407	expertise: 0.356
flair: 0.715	creativity: 0.579	beauty: 0.399	poised: 0.356
intricate: 0.624	performance: 0.574	intently: 0.358	graceful: 0.339
precise: 0.584	preparedness: 0.562	able: 0.320	powerful: 0.313
engaging: 0.584	knowledge: 0.472	performs: 0.308	role: 0.219
versatility: 0.525	personality: 0.447	automatic: 0.281	skill: 0.204
competent: 0.464	fully: 0.386	strength: 0.243	something: 0.196
elegant: 0.457	beauty: 0.355	competent: 0.243	elegant: 0.185
productive: 0.455	artistic: 0.341	versatility: 0.212	performance: 0.142
sophisticated: 0.391	determined: 0.316	functioning: 0.205	effectively: 0.133
abilities: 0.354	characterized: 0.310	powerful: 0.166	task: 0.128
difficult: 0.311	elegant: 0.306	precise: 0.144	confidently: 0.047
beauty: 0.298	profession: 0.305	engaging: 0.123	determination: 0.008
proper: 0.298	expert: 0.289	competitive: 0.113	diligently: -0.019
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
intricate: 0.870	character: 0.945	characterized: 1.165	skills: 1.326
secured: 0.679	secured: 0.932	fully: 0.789	secured: 1.231
words: 0.593	diligently: 0.749	intricate: 0.676	intricate: 0.844
knowledge: 0.534	intricate: 0.723	speed: 0.580	disciplined: 0.841
competence: 0.510	progress: 0.562	elegant: 0.449	confidently: 0.834
intellectual: 0.455	learning: 0.498	skill: 0.379	powerful: 0.507
enthusiasm: 0.446	fully: 0.445	strength: 0.285	confidence: 0.459
triumphant: 0.423	dedicated: 0.442	thoughtful: 0.178	highly: 0.414
intricately: 0.320	secure: 0.375	confident: 0.142	profession: 0.241
profession: 0.219	sophisticated: 0.374	confidently: 0.052	graceful: 0.140
confidence: 0.217	intricately: 0.365	advanced: 0.044	strength: 0.097
bright: 0.176	consistent: 0.354	role: 0.008	artistic: 0.091
quality: 0.167	rugged: 0.312	equipped: -0.018	poised: 0.070
poised: 0.115	thoughtful: 0.276	professional: -0.027	elegant: 0.070
effort: 0.115	enhance: 0.276	something: -0.029	quality: 0.070
intently: 0.113	enthusiasm: 0.216	graceful: -0.120	equipped: 0.003
role: 0.093	meticulous: 0.213	clean: -0.137	sophisticated: -0.023
engaging: 0.090	confidence: 0.191	technological: -0.142	clean: -0.026
graceful: 0.078	ensuring: 0.179	consistent: -0.185	role: -0.035
preparedness: 0.057	skills: 0.157	effort: -0.185	task: -0.050

Table 19: Top 20 competence-related words associated with 'Middle Eastern'

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
achieving: 1.296	achieve: 1.268	insights: 1.433	abilities: 1.739
triumphant: 1.104	abilities: 1.083	highly: 1.308	talent: 0.680
rugged: 1.016	committed: 0.946	thoroughly: 1.130	equipped: 0.441
meticulously: 0.625	enthusiasm: 0.579	effort: 1.020	strong: 0.390
committed: 0.612	strong: 0.563	bright: 0.908	cleaner: 0.372
capability: 0.593	intellectual: 0.531	responsibility: 0.870	suitable: 0.364
secure: 0.575	determined: 0.512	versatile: 0.821	responsible: 0.328
productive: 0.559	responsible: 0.484	proficient: 0.805	ensuring: 0.325
effort: 0.519	effort: 0.479	consistent: 0.763	skill: 0.281
accurate: 0.519	secured: 0.405	achieve: 0.638	clean: 0.236
successful: 0.519	diligently: 0.375	accurate: 0.571	proper: 0.208
ensuring: 0.491	efficiency: 0.335	poised: 0.554	precise: 0.161
able: 0.454	prepared: 0.328	expert: 0.535	poised: 0.127
stable: 0.419	personality: 0.327	competitive: 0.443	effectively: 0.127
responsibility: 0.403	competent: 0.318	talents: 0.441	diligently: 0.127
sophisticated: 0.393	bright: 0.298	equipped: 0.440	engaging: 0.113
equipped: 0.375	progress: 0.290	formidable: 0.433	determined: 0.095
efficient: 0.326	job: 0.276	suitable: 0.405	graceful: 0.092
capabilities: 0.326	profession: 0.245	effective: 0.401	skills: 0.089
assistant: 0.317	personality: 0.241	progress: 0.387	professional: 0.076
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
strong: 1.351	efficiency: 1.432	graceful: 0.461	strong: 0.994
skills: 1.088	intelligence: 1.285	intently: 0.407	effort: 0.916
proper: 0.896	knowledgeable: 1.169	advanced: 0.304	durable: 0.520
precise: 0.481	rugged: 0.963	consistent: 0.297	confidence: 0.432
quality: 0.469	vivid: 0.782	equipped: 0.228	sophisticated: 0.409
determination: 0.458	trained: 0.741	task: 0.226	prepared: 0.387
fully: 0.458	durable: 0.718	clean: 0.218	suitable: 0.377
technological: 0.413	technological: 0.705	powerful: 0.207	precise: 0.356
skilled: 0.394	dedicated: 0.634	prepared: 0.162	engaging: 0.344
prepared: 0.305	job: 0.634	cleaner: 0.154	skilled: 0.306
intellectual: 0.284	thorough: 0.622	something: 0.092	clean: 0.199
improvement: 0.252	proper: 0.609	sophisticated: 0.066	highly: 0.164
something: 0.222	skilled: 0.460	bright: 0.001	task: 0.131
triumphant: 0.211	suitable: 0.449	strength: -0.008	elegant: 0.123
clean: 0.202	preparedness: 0.428	professional: -0.011	fully: 0.119
learning: 0.191	enhance: 0.419	secured: -0.042	cleaner: 0.113
advanced: 0.181	accuracy: 0.378	quality: -0.056	something: 0.044
determined: 0.181	cleaner: 0.336	suitable: -0.086	equipped: 0.022
bright: 0.179	secure: 0.326	elegant: -0.139	professional: 0.016
expertise: 0.169	quality: 0.308	technological: -0.146	profession: 0.002

Table 20: Top 20 competence-related words associated with 'White(race)'

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
gratitude: 1.122	mutual: 1.157	rugged: 1.587	good: 1.130
prayerful: 0.953	suitable: 1.087	trust: 1.172	natural: 0.915
lively: 0.881	enjoys: 0.894	active: 0.751	secure: 0.469
trusting: 0.733	enjoying: 0.648	hot: 0.587	welcoming: 0.214
appreciation: 0.688	safe: 0.572	prayerful: 0.500	comfortable: 0.201
active: 0.581	friendly: 0.454	skillful: 0.395	enjoying: 0.193
enjoys: 0.491	warmth: 0.312	cheerful: 0.350	confident: 0.128
loving: 0.444	gentle: 0.281	secure: 0.312	taste: 0.080
cheerful: 0.318	calmness: 0.250	enjoyable: 0.307	delicious: 0.049
intimate: 0.301	knowledgeable: 0.234	lively: 0.307	strong: 0.004
welcoming: 0.292	beauty: 0.233	friendly: 0.242	neutral: -0.020
trust: 0.287	trust: 0.210	enjoying: 0.234	prayerful: -0.062
admiration: 0.259	respect: 0.194	strong: 0.231	manner: -0.064
respect: 0.259	lively: 0.194	quality: 0.227	appealing: -0.067
trustworthy: 0.259	confident: 0.179	smooth: 0.212	suitable: -0.285
enjoying: 0.222	secured: 0.157	passionate: 0.211	kind: -0.435
respected: 0.203	warm: 0.119	suitable: 0.193	fresh: -0.455
authenticity: 0.203	committed: 0.113	committed: 0.186	stylish: -0.574
attired: 0.200	kind: 0.093	actively: 0.186	pleasing: -0.657
gentle: 0.191	stylish: 0.070	committed: 0.183	actively: -0.279
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
solid: 1.442	heated: 0.733	hot: 1.089	solid: 1.083
heated: 1.094	respect: 0.581	respect: 1.014	pleasant: 0.960
enjoying: 0.855	friendliness: 0.581	solid: 0.767	respect: 0.919
comfort: 0.480	enjoying: 0.329	active: 0.745	enjoying: 0.767
pleasant: 0.464	suitable: 0.287	gentle: 0.382	heated: 0.597
gentle: 0.454	friendly: 0.250	confident: 0.231	neutral: 0.545
grayish: 0.442	welcoming: 0.203	neutral: 0.225	strong: 0.498
beauty: 0.378	thoughtful: 0.177	thoughtful: 0.173	comfortable: 0.396
enthusiastic: 0.357	enthusiastic: 0.174	kind: 0.169	kind: 0.375
prayerful: 0.317	attitude: 0.130	actively: 0.141	welcoming: 0.346
manner: 0.288	pleasant: 0.113	stylish: 0.137	natural: 0.265
secured: 0.287	warm: 0.110	natural: 0.10789	gentle: 0.253
lively: 0.280	secured: 0.106	smooth: -0.010	fine: 0.219
welcoming: 0.269	lively: 0.096	frendly: -0.025	secured: 0.121
respect: 0.258	cheerful: 0.068	quality: -0.029	warm: 0.095
friendly: 0.249	confident: 0.064	bright: -0.060	confident: 0.045
warm: 0.222	actively: 0.047	plain: -0.073	plain: 0.017
suitable: 0.201	good: 0.044	enjoying: -0.077	cheerful: -0.031
cheerful: 0.183	plain: 0.002	fresh: -0.151	good: -0.031
attired: 0.120	rugged: -0.004	secured: -0.151	stylish: -0.058

Table 21: Top 20 warmth-related words associated with ‘blue’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
flattering: 0.933	quality: 0.854	authenticity: 0.848	safely: 0.727
hot: 0.860	fine: 0.854	respectful: 0.848	comfort: 0.701
bright: 0.479	intimate: 0.783	pleasant: 0.778	beauty: 0.668
tasting: 0.433	good: 0.783	bright: 0.659	fresh: 0.642
committed: 0.400	bright: 0.716	appreciation: 0.642	kind: 0.525
welcoming: 0.359	fresh: 0.631	respected: 0.585	strong: 0.494
mutual: 0.333	comfort: 0.631	loving: 0.585	pleasing: 0.418
cheerful: 0.307	respectful: 0.561	hot: 0.585	delicious: 0.353
friendly: 0.305	lively: 0.516	enthusiastic: 0.494	safe: 0.353
beauty: 0.300	welcoming: 0.479	reliable: 0.363	actively: 0.324
good: 0.285	solid: 0.337	fresh: 0.332	suitable: 0.322
secured: 0.240	warmth: 0.302	comfort: 0.315	smooth: 0.206
fine: 0.213	natural: 0.233	natural: 0.264	plain: 0.090
fresh: 0.193	comfortable: 0.194	talented: 0.207	stylish: 0.086
respected: 0.192	confident: 0.186	stylish: 0.199	manner: -0.029
rugged: 0.183	strong: 0.179	committed: 0.152	enjoying: -0.049
natural: 0.160	friendliness: 0.125	safe: 0.136	welcoming: -0.049
strong: 0.149	manner: 0.125	knowledgeable: 0.094	secure: -0.057
quality: 0.135	passionate: 0.124	prayerful: 0.083	confident: -0.091
confident: 0.126	warm: 0.114	comfortable: 0.066	appealing: -0.107
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
thoughtful: 0.860	bright: 0.938	strong: 0.576	secure: 1.090
bright: 0.654	safely: 0.830	enjoying: 0.576	lightly: 0.954
comfortable: 0.622	affectionately: 0.702	bright: 0.486	warmth: 0.727
safe: 0.354	solid: 0.633	tasting: 0.406	bright: 0.724
secured: 0.284	hot: 0.587	good: 0.383	strong: 0.476
lively: 0.277	warmth: 0.550	stylish: 0.361	lively: 0.319
fine: 0.255	respectful: 0.550	comfort: 0.353	friendly: 0.302
cheerful: 0.161	strong: 0.516	cheerful: 0.328	cheerful: 0.295
warm: 0.142	appealing: 0.398	friendly: 0.254	warm: 0.295
suitable: 0.133	active: 0.398	secured: 0.237	secured: 0.294
fresh: 0.131	comfortable: 0.311	suitable: 0.235	tasting: 0.197
strong: 0.091	attitude: 0.280	fresh: 0.213	fresh: 0.193
respectful: 0.074	comfort: 0.213	fine: 0.197	good: 0.188
good: 0.068	secure: 0.202	manner: 0.181	welcoming: 0.179
actively: 0.054	good: 0.179	thoughtfulness: 0.174	gentle: 0.116
beauty: 0.053	thoughtfulness: 0.172	solid: 0.171	confident: 0.106
manner: 0.047	cheerful: 0.162	quality: 0.158	kind: 0.090
confident: 0.024	fresh: 0.161	actively: 0.079	actively: -0.021
respect: -0.008	enjoying: 0.117	welcoming: 0.045	manner: -0.030
secure: -0.008	lively: 0.105	kind: -0.065	stylish: -0.050

Table 22: Top 20 warmth-related words associated with ‘red’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
pure: 1.603	fluffy: 0.874	harmonious: 0.997	friendly: 1.005
safely: 0.987	tasting: 0.645	taste: 0.582	bright: 0.980
skillfully: 0.925	fine: 0.289	skillful: 0.389	comfortable: 0.565
attractive: 0.796	smooth: 0.267	plain: 0.309	prayerful: 0.565
safe: 0.755	calmness: 0.230	manner: 0.281	appreciation: 0.427
talented: 0.626	delicious: 0.230	appealing: 0.252	taste: 0.407
qualities: 0.603	actively: 0.214	smooth: 0.245	plain: 0.342
fluffy: 0.571	slicked: 0.174	safely: 0.195	stylish: 0.329
fine: 0.568	committed: 0.093	delicious: 0.180	smooth: 0.311
secure: 0.545	neutral: 0.077	welcoming: 0.145	neutral: 0.159
knowledgeable: 0.539	plain: 0.047	neutral: 0.142	appealing: 0.156
slicked: 0.483	strong: 0.022	warm: 0.142	manner: 0.087
tasting: 0.466	secured: 0.012	good: 0.057	safe: 0.079
trusted: 0.381	trust: -0.033	friendly: 0.028	pleasing: 0.029
rugged: 0.323	kind: -0.041	suitable: 0.013	confident: -0.042
appealing: 0.298	gentle: -0.074	quality: -0.010	suitable: -0.113
calmness: 0.294	attired: -0.091	respect: -0.032	actively: -0.121
reliable: 0.216	stylish: -0.102	strong: -0.045	enjoying: -0.159
plain: 0.199	passionate: -0.115	lively: -0.046	welcoming: -0.185
respectful: 0.199	manner: -0.124	attitude: -0.060	kind: -0.293
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
safe: 0.528	calmness: 1.362	fluffy: 0.868	slicked: 1.583
fine: 0.499	beauty: 0.777	smooth: 0.757	fluffy: 0.980
fluffy: 0.477	knowledgeable: 0.732	fine: 0.489	smooth: 0.835
grayish: 0.461	slicked: 0.709	heated: 0.342	suitable: 0.583
fluffy: 0.429	fluffy: 0.677	tasting: 0.242	nice: 0.467
strong: 0.376	tasting: 0.677	welcoming: 0.226	active: 0.261
smooth: 0.370	fine: 0.437	good: 0.190	tasting: 0.204
intimate: 0.335	quality: 0.385	plain: 0.174	manner: 0.187
secure: 0.277	gentle: 0.353	comfort: 0.119	quality: 0.184
quality: 0.260	smooth: 0.331	natural: 0.050	stylish: 0.102
lightly: 0.255	prayerful: 0.303	manner: -0.019	actively: 0.077
natural: 0.165	thoughtfulness: 0.246	neutral: -0.048	plain: 0.042
stylish: 0.157	enthusiastic: 0.218	cheerful: -0.058	natural: -0.113
neutral: 0.145	strong: 0.198	suitable: -0.058	comfortable: -0.134
respectful: 0.096	kind: 0.180	confident: -0.069	lively: -0.143
confident: 0.075	lightly: 0.148	active: -0.074	fresh: -0.153
plain: 0.037	rugged: 0.147	gentle: -0.074	confident: -0.164
actively: -0.029	secure: 0.139	fresh: -0.091	good: -0.183
fresh: -0.068	manner: 0.138	secured: -0.121	friendly: -0.188
active: -0.069	stylish: 0.131	kind: -0.123	neutral: -0.282

Table 23: Top 20 warmth-related words associated with ‘white (color)’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
trustworthy: 0.990	qualities: 0.987	rugged: 0.985	good: 0.784
slicked: 0.990	slicked: 0.872	flavorful: 0.815	friendly: 0.659
gratitude: 0.853	welcoming: 0.665	solid: 0.763	bright: 0.644
harmonious: 0.768	attired: 0.590	smoothly: 0.570	quality: 0.644
coolness: 0.727	secured: 0.511	secure: 0.549	lively: 0.592
attired: 0.581	suitable: 0.502	bright: 0.545	appreciation: 0.547
flavorful: 0.575	enjoying: 0.332	lively: 0.513	plain: 0.521
talented: 0.567	trust: 0.277	friendly: 0.497	pleasant: 0.521
admiration: 0.543	safe: 0.250	good: 0.442	flavorful: 0.521
respected: 0.538	respect: 0.234	neutral: 0.416	prayerful: 0.466
tasting: 0.531	quality: 0.210	gratitude: 0.307	secure: 0.374
hot: 0.521	neutral: 0.190	appreciation: 0.248	stylish: 0.309
lively: 0.460	kind: 0.176	loving: 0.248	taste: 0.236
trustworthiness: 0.405	stylish: 0.163	pleasantries: 0.248	delicious: 0.199
respect: 0.373	passionate: 0.139	taste: 0.248	warm: 0.159
prayerful: 0.362	kindness: 0.139	plain: 0.215	comfort: 0.132
trust: 0.354	solid: 0.117	stylish: 0.211	comfortable: 0.122
credibility: 0.312	smooth: 0.094	pleasant: 0.178	strong: 0.083
authenticity: 0.290	lively: 0.094	enthusiastic: 0.125	safe: 0.007
welcoming: 0.286	gentle: 0.085	enjoying: 0.112	neutral: -0.106
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
freshness: 0.774	friendliness: 0.806	lively: 0.731	slicked: 0.990
good: 0.595	slicked: 0.624	prayerful: 0.579	enthusiastic: 0.797
good: 0.582	good: 0.479	respect: 0.464	secure: 0.767
fine: 0.549	secure: 0.439	thoughtful: 0.428	prayerful: 0.767
respect: 0.549	solid: 0.371	welcoming: 0.294	respect: 0.575
skillfully: 0.511	lively: 0.338	fine: 0.294	lively: 0.518
taste: 0.511	kind: 0.338	solid: 0.237	warmth: 0.405
secured: 0.470	secured: 0.333	natural: 0.194	welcoming: 0.390
prayerful: 0.412	respectful: 0.321	good: 0.164	comfortable: 0.373
lively: 0.389	thoughtfulness: 0.298	secured: 0.091	cheerful: 0.317
thoughtful: 0.380	welcoming: 0.249	kind: 0.073	good: 0.295
cheerful: 0.235	enthusiastic: 0.238	plain: 0.025	tasting: 0.289
respectful: 0.231	friendly: 0.178	fresh: 0.021	strong: 0.253
enthusiastic: 0.190	attitude: 0.165	gentle: -0.006	secured: 0.238
quality: 0.156	fine: 0.165	active: -0.006	kind: 0.182
tasting: 0.149	hot: 0.106	friendly: -0.035	friendly: 0.176
welcoming: 0.121	safely: 0.098	bright: -0.096	quality: 0.132
gentle: 0.112	thoughtful: 0.092	actively: -0.112	actively: 0.113
activate: 0.100	comfort: 0.058	neutral: -0.080	natural: 0.112
secure: 0.090	prayerful: 0.055	quality: -0.132	smooth: 0.050

Table 24: Top 20 warmth-related words associated with ‘man’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
loving: 0.872	intimate: 0.843	skillfully: 0.822	beauty: 0.894
generous: 0.788	heated: 0.750	skillful: 0.822	actively: 0.346
flattering: 0.788	strong: 0.616	suitable: 0.720	manner: 0.336
strong: 0.659	indiscernible: 0.598	beauty: 0.717	welcoming: 0.307
secured: 0.657	tasting: 0.565	talented: 0.636	fresh: 0.301
skillfully: 0.595	bright: 0.563	trust: 0.600	natural: 0.293
reliable: 0.446	warm: 0.527	strong: 0.567	appealing: 0.225
cheerful: 0.425	beauty: 0.506	cheerful: 0.555	pleasing: 0.223
trusting: 0.425	enjoys: 0.472	knowledgeable: 0.466	smooth: 0.209
kindness: 0.425	confident: 0.403	active: 0.458	suitable: 0.186
confident: 0.418	calmness: 0.391	harmonious: 0.430	enjoying: 0.162
fluffy: 0.376	committed: 0.361	attractive: 0.352	safely: 0.145
bright: 0.373	manner: 0.297	confident: 0.334	kind: 0.135
warm: 0.369	fresh: 0.261	comfortable: 0.318	confident: 0.106
secure: 0.366	natural: 0.242	natural: 0.278	neutral: 0.098
enjoys: 0.358	fine: 0.235	authenticity: 0.278	safe: -0.007
comfortable: 0.323	friendliness: 0.221	respectful: 0.278	strong: -0.088
beauty: 0.289	comfort: 0.183	delicious: 0.266	comfortable: -0.132
passionate: 0.262	knowledgeable: 0.172	enjoyable: 0.249	comfort: -0.144
smooth: 0.230	delicious: 0.106	welcoming: 0.240	delicious: -0.229
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
enjoyable: 1.003	beauty: 0.785	secure: 1.006	beauty: 0.788
beauty: 0.939	smoothly: 0.516	comfort: 0.521	manner: 0.657
heated: 0.781	harmonious: 0.516	comfortable: 0.521	pleasant: 0.525
warm: 0.645	passionately: 0.516	stylish: 0.475	warm: 0.382
comfort: 0.556	warm: 0.466	cheerful: 0.458	fine: 0.310
safe: 0.518	strong: 0.433	enjoying: 0.453	solid: 0.273
strong: 0.518	affectionately: 0.416	suitable: 0.402	confident: 0.258
lightly: 0.482	suited: 0.416	smooth: 0.351	heated: 0.233
manner: 0.411	gentle: 0.331	manner: 0.331	neutral: 0.210
comfortable: 0.371	confident: 0.331	tasting: 0.295	active: 0.191
confident: 0.367	qualities: 0.323	strong: 0.269	fluffy: 0.176
grayish: 0.351	knowledgeable: 0.301	fluffy: 0.269	thoughtful: 0.136
enjoying: 0.279	pleasant: 0.299	heated: 0.228	suitable: 0.131
fluffy: 0.253	bright: 0.297	confident: 0.211	bright: 0.124
pleasant: 0.210	appealing: 0.264	warm: 0.158	enjoying: 0.098
actively: 0.202	lightly: 0.249	quality: 0.122	lightly: 0.010
neutral: 0.155	active: 0.224	actively: 0.105	plain: 0.005
natural: 0.146	enjoying: 0.216	bright: 0.091	gentle: -0.034
solid: 0.129	warmth: 0.182	neutral: 0.076	stylish: -0.035
stylish: 0.129	stylish: 0.178	friendly: 0.035	fresh: -0.045

Table 25: Top 20 warmth-related words associated with ‘woman’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
admiration: 0.929	fresh: 0.834	skillful: 0.910	safe: 1.371
safely: 0.797	respect: 0.528	neutral: 0.620	prayerful: 0.915
respected: 0.621	calmness: 0.512	fresh: 0.573	comfortable: 0.816
fluffy: 0.481	enjoying: 0.390	suitable: 0.350	smooth: 0.809
solid: 0.443	good: 0.342	respected: 0.347	kind: 0.736
committed: 0.344	beauty: 0.260	safe: 0.345	appealing: 0.572
friendliness: 0.286	quality: 0.249	safely: 0.308	pleasing: 0.543
appealing: 0.247	smooth: 0.227	appreciation: 0.266	delicious: 0.327
reliable: 0.234	gentle: 0.170	smooth: 0.251	manner: 0.207
quality: 0.131	solid: 0.148	secure: 0.234	neutral: 0.055
enjoying: 0.127	secured: 0.097	welcoming: 0.151	enjoying: 0.016
knowledgeable: 0.106	warmth: 0.074	cheerful: 0.129	actively: -0.131
natural: 0.043	friendliness: 0.063	attitude: 0.112	confident: -0.269
smooth: 0.040	neutral: -0.034	lively: 0.086	welcoming: -0.294
comfort: 0.028	plain: -0.042	delicious: 0.016	suitable: -0.307
rugged: 0.011	bright: -0.125	enjoying: -0.027	stylish: -0.426
bright: -0.014	attired: -0.180	good: -0.047	
plain: -0.015	kind: -0.251	manner: -0.108	
manner: -0.022	strong: -0.281	actively: -0.131	
beauty: -0.035	confident: -0.361	appealing: -0.219	
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
fluffy: 0.532	calmness: 2.594	solid: 0.820	kind: 0.896
quality: 0.467	comfort: 0.816	suitable: 0.651	lively: 0.807
enthusiastic: 0.349	appealing: 0.594	smooth: 0.507	smooth: 0.502
smooth: 0.250	beauty: 0.594	quality: 0.471	fluffy: 0.498
neutral: 0.174	enthusiastic: 0.363	friendly: 0.457	quality: 0.483
suitable: 0.161	rugged: 0.308	secured: 0.321	suitable: 0.332
enjoying: 0.131	pleasant: 0.306	bright: 0.312	manner: 0.242
manner: 0.122	fine: 0.236	welcoming: 0.199	bright: 0.192
cheerful: 0.096	smooth: 0.213	natural: 0.192	cheerful: 0.109
welcoming: 0.091	suitable: 0.210	kind: 0.171	active: 0.086
plain: -0.007	neutral: 0.180	plain: 0.023	good: 0.031
bright: -0.009	tasting: 0.161	stylish: 0.014	actively: 0.006
natural: -0.051	good: 0.158	neutral: -0.051	plain: -0.046
actively: -0.078	lively: 0.149	manner: -0.121	friendly: -0.052
fresh: -0.095	friendly: 0.116	confident: -0.176	natural: -0.095
friendly: -0.131	secure: 0.108	fresh: -0.208	welcoming: -0.111
confident: -0.339	cheerful: 0.106	actively: -0.324	enjoying: -0.145
stylish: -0.537	comfortable: 0.091		fresh: -0.156
attitude: -0.598	manner: 0.052		neutral: -0.203
warm: -0.909	plain: 0.037		stylish: -0.235

Table 26: Top 20 warmth-related words associated with 'Asian'

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
hot: 1.640	stylish: 0.705	reliable: 1.370	pleasing: 1.157
secure: 0.697	manner: 0.599	prayerful: 0.827	fresh: 0.970
fine: 0.605	warm: 0.421	attitude: 0.802	friendly: 0.748
talented: 0.570	confident: 0.297	plain: 0.734	taste: 0.505
reliable: 0.553	enjoying: 0.289	enthusiastic: 0.678	enjoying: 0.450
intimate: 0.553	comfortable: 0.230	committed: 0.645	kind: 0.282
delicious: 0.546	respect: 0.171	cheerful: 0.548	strong: 0.233
enjoys: 0.503	passionate: 0.158	active: 0.533	confident: 0.196
appealing: 0.487	smooth: 0.113	stylish: 0.502	welcoming: 0.153
safe: 0.470	natural: 0.055	friendly: 0.497	neutral: 0.121
safely: 0.439	plain: 0.039	natural: 0.463	manner: 0.091
friendliness: 0.413	neutral: 0.016	strong: 0.463	appealing: -0.383
confident: 0.406	strong: -0.022	confident: 0.427	actively: -0.433
attired: 0.357	actively: -0.028	good: 0.410	stylish: -0.465
good: 0.355	bright: -0.046	respect: 0.394	suitable: -0.793
fresh: 0.354	attired: -0.059	secure: 0.391	delicious: -0.904
secured: 0.353	kind: -0.063	comfortable: 0.382	
trust: 0.346	fresh: -0.087	passionate: 0.360	
strong: 0.342	warmth: -0.095	knowledgeable: 0.310	
knowledgeable: 0.339	friendliness: -0.099	quality: 0.308	
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
solid: 1.720	friendliness: 0.925	stylish: 0.753	strong: 1.033
respect: 1.272	strong: 0.878	thoughtful: 0.718	warm: 0.852
tasting: 1.232	warmth: 0.866	manner: 0.657	good: 0.852
cheerful: 0.597	heated: 0.755	confident: 0.530	active: 0.618
confident: 0.432	warm: 0.675	natural: 0.510	bright: 0.418
stylish: 0.428	tasting: 0.656	enjoying: 0.468	confident: 0.406
thoughtful: 0.393	quality: 0.516	bright: 0.244	stylish: 0.354
warm: 0.204	hot: 0.488	kind: 0.048	cheerful: 0.267
bright: 0.200	comfort: 0.478	quality: 0.030	lively: 0.191
suitable: 0.116	stylish: 0.451	plain: 0.001	fresh: 0.103
enthusiastic: 0.109	cheerful: 0.438	neutral: -0.011	neutral: 0.070
active: 0.009	attitude: 0.416	fresh: -0.090	plain: 0.023
friendly: 0.009	safely: 0.381	suitable: -0.320	suitable: -0.099
fresh: 0.007	enjoying: 0.351	smooth: -0.372	friendly: -0.125
plain: 0.005	natural: 0.339	friendly: -0.579	actively: -0.218
actively: 0.002	solid: 0.323	actively: -0.798	welcoming: -0.219
neutral: 0.001	welcoming: 0.318	solid: -1.009	smooth: -0.352
enjoying: -0.069	lightly: 0.304		manner: -0.424
lively: -0.084	bright: 0.300		secured: -0.508
natural: -0.088	prayerful: 0.281		natural: -0.885

Table 27: Top 20 warmth-related words associated with ‘Black’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
appreciation: 1.038	fluffy: 1.258	natural: 0.919	prayerful: 0.908
admiration: 0.983	comfort: 0.995	beauty: 0.461	stylish: 0.396
good: 0.605	calmness: 0.810	plain: 0.316	comfortable: 0.323
authenticity: 0.453	gentle: 0.693	safely: 0.309	taste: 0.313
respectful: 0.246	natural: 0.274	knowledgeable: 0.267	suitable: 0.100
calmness: 0.238	passionate: 0.240	safe: 0.247	actively: 0.013
warmth: 0.219	solid: 0.183	comfort: 0.231	welcoming: -0.039
respected: 0.190	warmth: 0.156	enjoys: 0.180	neutral: -0.066
bright: 0.123	natural: 0.156	enthusiastic: 0.144	manner: -0.079
respect: 0.094	warm: 0.155	strong: 0.123	confident: -0.093
kind: 0.086	actively: 0.154	respect: 0.101	appealing: -0.275
gentle: 0.078	kind: 0.054	actively: 0.095	delicious: -0.359
solid: 0.072	knowledgeable: 0.054	quality: 0.063	strong: -0.611
intimate: 0.066	beauty: 0.033	stylish: 0.029	enjoying: -0.858
beauty: 0.055	plain: 0.031	enjoying: 0.021	
warm: 0.045	comfortable: -0.023	passionate: 0.016	
strong: 0.013	respect: -0.028	smooth: 0.004	
friendly: -0.005	attired: -0.034	good: -0.046	
confident: -0.032	friendly: -0.116	confident: -0.062	
comfortable: -0.062	enjoying: -0.215	neutral: -0.097	
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
intimate: 1.328	respectful: 1.389	active: 0.736	lightly: 0.847
beauty: 1.067	respect: 1.203	solid: 0.475	kind: 0.362
respectful: 0.825	slicked: 0.851	suitable: 0.318	quality: 0.311
gentle: 0.813	fluffy: 0.634	quality: 0.290	active: 0.262
secured: 0.783	tasting: 0.634	actively: 0.221	confident: 0.154
thoughtful: 0.652	gentle: 0.453	thoughtful: 0.164	neutral: 0.153
suitable: 0.364	thoughtful: 0.412	neutral: 0.135	warm: 0.082
active: 0.295	grayish: 0.352	plain: 0.018	smooth: 0.030
enthusiastic: 0.105	comfortable: 0.343	fresh: -0.046	plain: 0.021
actively: 0.081	kind: 0.289	secured: -0.053	secured: -0.094
neutral: 0.033	pleasant: 0.207	confident: -0.187	fresh: -0.123
plain: 0.032	lightly: 0.041	smooth: -0.241	bright: -0.124
fresh: 0.029	plain: 0.008	manner: -0.260	gentle: -0.138
warm: 0.014	neutral: 0.005	kind: -0.354	stylish: -0.151
welcoming: -0.083	beauty: -0.003	neutral: -0.403	natural: -0.181
lively: -0.087	actively: -0.029	bright: -0.483	actively: -0.243
manner: -0.112	fresh: -0.050	stylish: -1.134	manner: -0.328
confident: -0.231	natural: -0.086	friendly: -1.493	fluffy: -0.451
stylish: -0.260	confident: -0.109		lively: -0.520
smooth: -0.316	bright: -0.136		cheerful: -1.088

Table 28: Top 20 warmth-related words associated with 'Indian'

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
slicked: 1.805	slicked: 1.277	pleasant: 1.104	suitable: 0.696
welcoming: 0.694	knowledgeable: 1.037	warm: 0.926	delicious: 0.651
friendly: 0.535	welcoming: 0.919	talented: 0.696	comfortable: 0.596
trustworthiness: 0.457	good: 0.486	comfort: 0.691	secure: 0.449
reliable: 0.413	secured: 0.463	good: 0.461	active: 0.161
manner: 0.400	friendliness: 0.413	lively: 0.409	strong: 0.137
fluffy: 0.372	comfortable: 0.343	respected: 0.348	enjoying: 0.109
quality: 0.362	charity: 0.317	appealing: 0.327	stylish: 0.107
enjoys: 0.350	natural: 0.257	comfortable: 0.320	confident: 0.092
stylish: 0.342	passionate: 0.223	plain: 0.316	taste: 0.060
gentle: 0.334	warmth: 0.217	welcoming: 0.267	manner: 0.049
suitable: 0.331	smooth: 0.178	appreciation: 0.267	neutral: 0.026
delicious: 0.297	manner: 0.126	enjoying: 0.248	smooth: -0.036
warm: 0.257	confident: 0.094	quality: 0.189	appealing: -0.120
mutual: 0.234	plain: 0.094	safe: 0.140	welcoming: -0.292
secured: 0.225	enjoying: 0.031	confident: 0.104	
confident: 0.219	neutral: -0.021	actively: 0.096	
intimate: 0.191	fresh: -0.022	neutral: 0.078	
comfortable: 0.153	solid: -0.031	enjoys: 0.074	
actively: 0.139	attired: -0.051	delicious: 0.054	
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
lively: 1.046	thoughtfulness: 1.208	enjoying: 0.730	pleasant: 1.106
comfortable: 0.802	attitude: 1.036	secured: 0.500	quality: 0.541
lightly: 0.712	strong: 0.967	welcoming: 0.489	suitable: 0.527
enjoying: 0.661	enthusiastic: 0.655	actively: 0.473	cheerful: 0.504
pleasant: 0.471	grayish: 0.655	friendly: 0.420	friendly: 0.434
smooth: 0.445	fluffy: 0.639	kind: 0.336	fluffy: 0.404
natural: 0.364	lively: 0.615	stylish: 0.282	gentle: 0.354
active: 0.291	appealing: 0.586	thoughtful: 0.202	bright: 0.250
confident: 0.176	comfortable: 0.425	confident: 0.181	stylish: 0.229
warm: 0.174	safely: 0.364	neutral: 0.053	actively: 0.193
manner: 0.162	enjoying: 0.334	bright: 0.029	neutral: 0.181
friendly: 0.128	hot: 0.301	plain: -0.004	lively: 0.165
enthusiastic: 0.101	prayerful: 0.264	fresh: -0.008	smooth: 0.124
cheerful: 0.070	active: 0.264	solid: -0.203	manner: 0.117
stylish: 0.053	good: 0.215	smooth: -0.203	confident: 0.112
plain: 0.004	slicked: 0.211	manner: -0.327	good: 0.089
bright: -0.045	respect: 0.208	quality: -0.424	plain: -0.001
actively: -0.049	secured: 0.169		fresh: -0.008
fresh: -0.053	stylish: 0.139		warm: -0.174
welcoming: -0.087	warm: 0.115		secured: -0.293

Table 29: Top 20 warmth-related words associated with ‘Latino’

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
attractive: 1.098	respectful: 1.472	welcoming: 0.558	natural: 1.214
active: 1.098	fine: 1.432	beauty: 0.399	beauty: 1.008
trustworthiness: 0.999	bright: 0.794	friendly: 0.348	taste: 0.735
prayerful: 0.818	quality: 0.695	respect: 0.315	secure: 0.639
mutual: 0.539	actively: 0.516	plain: 0.308	strong: 0.396
comfort: 0.477	fresh: 0.502	active: 0.259	actively: 0.342
authenticity: 0.468	warm: 0.418	enjoys: 0.172	stylish: 0.310
respectful: 0.414	stylish: 0.373	smooth: 0.166	welcoming: 0.191
passionate: 0.401	beauty: 0.355	stylish: 0.145	friendly: 0.107
respect: 0.399	friendly: 0.354	passionate: 0.131	delicious: 0.026
actively: 0.398	solid: 0.306	manner: 0.130	manner: -0.012
gentle: 0.345	kind: 0.283	fresh: 0.129	confident: -0.049
warm: 0.308	passionate: 0.262	actively: 0.125	neutral: -0.076
beauty: 0.298	comfortable: 0.176	lively: 0.078	enjoying: -0.295
welcoming: 0.294	attired: 0.138	appealing: 0.023	appealing: -0.428
suitable: 0.221	friendliness: 0.130	delicious: 0.009	suitable: -0.630
secured: 0.209	plain: 0.086	quality: -0.082	
respected: 0.205	confident: 0.041	confident: -0.110	
enjoying: 0.197	warmth: 0.008	respected: -0.120	
talented: 0.191	neutral: 0.007	secure: -0.135	
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
respectful: 1.090	secured: 0.932	tasting: 1.218	secured: 1.231
gentle: 1.008	prayerful: 0.761	good: 0.902	pleasant: 1.121
secured: 0.679	fine: 0.699	thoughtful: 0.178	enjoying: 1.066
good: 0.553	secure: 0.375	kind: 0.174	welcoming: 0.869
enjoying: 0.515	solid: 0.317	fresh: 0.147	comfortable: 0.668
welcoming: 0.385	rugged: 0.312	confident: 0.124	gentle: 0.562
warm: 0.368	smooth: 0.276	plain: 0.049	natural: 0.326
smooth: 0.364	thoughtful: 0.276	friendly: 0.024	manner: 0.074
lightly: 0.356	welcoming: 0.160	neutral: -0.000	quality: 0.070
pleasant: 0.214	enthusiastic: 0.126	smooth: -0.075	plain: 0.058
fluffy: 0.191	confident: 0.118	active: -0.098	fresh: 0.024
bright: 0.176	safely: 0.112	stylish: -0.120	active: 0.021
friendly: 0.168	quality: 0.095	manner: -0.181	friendly: -0.029
quality: 0.167	actively: 0.088	bright: -0.247	neutral: -0.033
active: 0.160	active: 0.083	secured: -0.261	smooth: -0.041
cheerful: 0.063	friendly: 0.068	suitable: -0.346	cheerful: -0.066
actively: 0.015	cheerful: 0.058	natural: -0.389	stylish: -0.097
plain: 0.013	stylish: 0.041	actively: -0.502	suitable: -0.110
fresh: 0.006	fresh: 0.036	quality: -0.608	warm: -0.159
stylish: 0.001	manner: 0.019	solid: -0.712	confident: -0.178

Table 30: Top 20 warmth-related words associated with 'Middle Eastern'

Llama 3.2 11B	Llama 3.2 90B	InstructBLIP 7B	InstructBLIP 13B
tasting: 1.644	cheerful: 2.142	bright: 0.908	smooth: 0.440
gratitude: 1.381	lively: 1.027	cheerful: 0.848	strong: 0.390
rugged: 1.016	committed: 0.946	safely: 0.653	suitable: 0.364
fine: 0.861	welcoming: 0.668	suitable: 0.405	appealing: 0.336
good: 0.693	strong: 0.563	appreciation: 0.249	comfortable: 0.324
safe: 0.671	good: 0.458	committed: 0.249	enjoying: 0.225
committed: 0.612	natural: 0.454	secure: 0.217	welcoming: 0.185
freshness: 0.612	friendly: 0.405	talented: 0.192	confident: 0.071
secure: 0.575	secured: 0.405	enthusiastic: 0.126	actively: -0.060
friendliness: 0.435	bright: 0.298	manner: 0.119	neutral: -0.073
natural: 0.241	smooth: 0.290	comfortable: 0.109	stylish: -0.144
quality: 0.236	manner: 0.279	appealing: 0.108	delicious: -0.220
attired: 0.235	stylish: 0.166	active: 0.097	manner: -0.310
calmness: 0.230	gentle: 0.162	passionate: 0.034	
appreciation: 0.140	attired: 0.152	stylish: 0.029	
authenticity: 0.140	knowledgeable: 0.149	natural: 0.026	
respect: 0.138	beauty: 0.065	delicious: 0.017	
trust: 0.131	friendliness: 0.063	safe: 0.001	
talented: 0.126	confident: 0.057	quality: 0.001	
knowledgeable: 0.110	warm: 0.013	fresh: -0.062	
Pixtral 12B	**Pixtral Large**	**Qwen 2.5 7B**	**Qwen 2.5 72B**
strong: 1.351	knowledgeable: 1.169	welcoming: 0.684	comfortable: 1.055
good: 0.824	rugged: 0.963	cheerful: 0.651	strong: 0.994
quality: 0.469	hot: 0.848	active: 0.577	natural: 0.739
fluffy: 0.364	solid: 0.782	natural: 0.482	fluffy: 0.543
manner: 0.294	heated: 0.700	actively: 0.446	welcoming: 0.520
natural: 0.261	fluffy: 0.600	friendly: 0.325	friendly: 0.458
comfortable: 0.205	enjoying: 0.600	smooth: 0.184	suitable: 0.377
bright: 0.179	safely: 0.548	fresh: 0.168	actively: 0.299
stylish: 0.132	suitable: 0.449	bright: 0.001	good: 0.298
secured: 0.129	secure: 0.326	manner: -0.004	manner: 0.172
lightly: 0.114	quality: 0.308	secured: -0.042	enjoying: 0.164
fresh: 0.098	welcoming: 0.298	quality: -0.056	fresh: 0.143
welcoming: 0.093	lightly: 0.248	suitable: -0.086	gentle: 0.119
friendly: 0.089	manner: 0.229	plain: -0.090	plain: -0.055
actively: 0.024	prayerful: 0.226	neutral: -0.143	cheerful: -0.187
confident: 0.003	beauty: 0.185	solid: -0.231	warm: -0.187
plain: -0.048	actively: 0.179	stylish: -0.446	stylish: -0.190
neutral: -0.070	pleasant: 0.173	kind: -0.566	neutral: -0.209
warm: -0.160	slicked: 0.173	confident: -0.782	confident: -0.296
enjoying: -0.352	active: 0.165		bright: -0.412

Table 31: Top 20 warmth-related words associated with ‘white (race)’