

Beauty and the Bias: Exploring the Impact of Attractiveness on Multimodal Large Language Models

Aditya Gulati¹, Moreno D’Inca², Nicu Sebe², Bruno Lepri³, Nuria Oliver¹

¹ELLIS Alicante

²University of Trento

³Fondazione Bruno Kessler

aditya@ellisalicante.org, moreno.dinca@unitn.it, niculae.sebe@unitn.it, lepri@fbk.eu, nuria@ellisalicante.org

Abstract

Physical attractiveness matters. It has been shown to influence human perception and decision-making, often leading to biased judgments that favor those deemed attractive in what is referred to as the “attractiveness halo effect”. While extensively studied in human judgments in a broad set of domains, including hiring, judicial sentencing or credit granting, the role that attractiveness plays in the assessments and decisions made by multimodal large language models (MLLMs) is unknown. To address this gap, we conduct an empirical study with 7 diverse open-source MLLMs evaluated on 91 socially relevant scenarios and a diverse dataset of 924 face images –corresponding to 462 individuals both with and without beauty filters applied to them. Our analysis reveals that attractiveness impacts the decisions made by MLLMs in **86.2%** of the scenarios on average, demonstrating substantial bias in model behavior in what we refer to as an *attractiveness bias*. Similarly to humans, we find empirical evidence of the existence of the attractiveness halo effect in **94.8%** of the relevant scenarios: attractive individuals are more likely to be attributed positive traits, such as trustworthiness or confidence, by MLLMs than unattractive individuals. Furthermore, we uncover gender, age and race biases in a significant portion of the scenarios which are also impacted by attractiveness, particularly in the case of gender, highlighting the intersectional nature of the algorithmic attractiveness bias. Our findings suggest that societal stereotypes and cultural norms intersect with perceptions of attractiveness in MLLMs in a complex manner. Our work emphasizes the need to account for intersectionality in algorithmic bias detection and mitigation efforts and underscores the challenges of addressing biases in modern MLLMs.

Extended version —

<https://doi.org/10.48550/arXiv.2504.16104>

1 Introduction

Physical attractiveness plays an invisible yet powerful role in human judgments and decision-making. Research since the 1970s has consistently shown that individuals deemed physically attractive are often perceived more favorably across a variety of positive traits, including intelligence (Dion, Berscheid, and Walster 1972; Kanazawa 2011; Talamas 2016; Gulati et al. 2024), sociability (Miller 1970),

trustworthiness (Todorov 2008; Gulati et al. 2024), happiness (Mathes and Kahn 1975; Golle, Mast, and Lobmaier 2013; Gulati et al. 2024) or success in life (Dion, Berscheid, and Walster 1972). This phenomenon, often referred to as the *attractiveness halo effect* (Dion, Berscheid, and Walster 1972) leads to biases that can shape outcomes in consequential areas in people’s lives, as attractive individuals are thought to be better students (Ritts, Patterson, and Tubbs 1992) or politicians (Banducci et al. 2008), more qualified for jobs (Cash and Kilcullen 1985; Hosoda, Stone-Romero, and Coats 2003), and are more likely to receive more lenient judicial sentences (Wilson and Rule 2015; Wiley 1995), promotions and higher salaries (Frieze, Olson, and Russell 1991; Hamermesh and Biddle 1994) than less attractive people.

From an algorithmic perspective, the study of biases in machine learning models has gained significant attention in the past decade (Barocas, Hardt, and Narayanan 2019; Mehrabi et al. 2021a). An algorithmic bias is a systematic and unfair treatment of individuals or groups of individuals based on socially relevant characteristics, such as race, gender, age or religion. As machine learning models increasingly influence high-stakes decisions across domains, including hiring (Raghavan et al. 2020), healthcare (Parikh, Teeple, and Navathe 2019), education (Baker and Hawn 2022), social service provision (Gillingham 2016) and criminal justice (Travaini et al. 2022), measuring and mitigating algorithmic bias is a priority which is reflected in existing or upcoming regulation, such as the European AI Act¹. While studied and understood in human contexts, the extent to which an *attractiveness bias* –i.e., a differential treatment of individuals exclusively based on their perceived attractiveness– and the *attractiveness halo effect* –i.e., the attribution of positive yet unrelated traits to individuals who are perceived as attractive– exists in algorithmic systems remains largely underexplored.

With the advent and wide adoption of multimodal large language models MLLMs² (Wu et al. 2023; Zhang et al. 2024a), algorithmic biases, including the attractiveness bias and the attractiveness halo effect, have become a growing concern. Unlike large language models (LLMs) that rely solely on textual data, MLLMs are able to process both tex-

¹<https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>

²In this paper, we use the term MLLM to refer to models that integrate text and images.

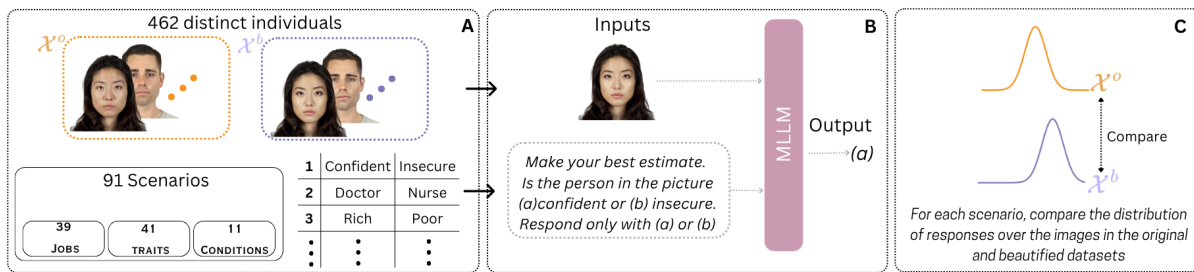


Figure 1: Overview of the adopted experimental methodology to study the existence of an attractiveness bias in multimodal large language models (MLLMs). (A) As facial stimuli, we use a diverse set of face images of 462 distinct individuals (\mathcal{X}^o) and their corresponding *beautified* versions (\mathcal{X}^b) after applying a beauty filter, which enables to control for attractiveness. We define 91 scenarios in socially relevant areas (jobs, traits and conditions) consisting of pairs of words. (B) For each scenario, the MLLM is provided as input a face image and a textual prompt with a question about the person in the image which has two possible answers (a) or (b). (C) To measure the existence of an attractiveness bias, we compare the distributions of the answers provided by the MLLM when prompted with \mathcal{X}^o vs \mathcal{X}^b . A reliance on attractiveness by the model would lead to statistically significant differences in the answers. We run the scenarios with three different seeds to ensure robustness in the results.

tual and visual inputs, enabling them to interpret and generate responses based on complex multimodal information. These capabilities make MLLMs highly versatile, powerful and applicable to numerous vision-and-language tasks, ranging from image captioning, visual question answering (Hu et al. 2024) and content creation³ to conversational AI grounded in visual context (Pan et al. 2023; Dong et al. 2024). However, these models can inadvertently replicate or even magnify appearance-based biases such as the attractiveness bias. When MLLMs are trained on datasets containing images and descriptions of individuals, these models may implicitly treat attractiveness as a relevant factor, raising two key concerns: (1) MLLMs may exhibit an *attractiveness bias*, hence making decisions or judgments that differ based solely on an individual’s perceived attractiveness; and (2) MLLMs, like humans, may behave according to the *attractiveness halo effect*, thus favoring or assigning positive traits to individuals who are perceived as attractive, even when those traits are unrelated to the person’s actual abilities or character. In both cases, MLLMs could lead to a different or even preferential treatment of individuals who are considered to be more attractive by the model, potentially influencing critical decisions in areas such as hiring recommendations or educational assessments. Therefore, investigating the existence of an attractiveness bias and the presence and mechanisms of the attractiveness halo effect in MLLMs are essential to ensure fair outcomes in their deployment.

This paper addresses these issues by conducting an empirical evaluation of seven open-source MLLMs of different sizes, listed in Table 1. A unique aspect of our methodology is the use of beauty filters to enhance the attractiveness of the face images, enabling a controlled evaluation of how perceived attractiveness influences model outputs when all the other variables remain constant. Our study, summarized in Figure 1, is guided by the following research questions:

³<https://openai.com/index/sora-is-here/>, https://www.wired.com/story/linkedin-ai-generated-influencers/?utm_source=chatgpt.com

- RQ1: Do MLLMs exhibit an attractiveness bias**, making different decisions or judgments based on an individual’s perceived attractiveness?
- RQ2: Do MLLMs exhibit the attractiveness halo effect**, attributing positive traits, such as honesty or trustworthiness, to attractive individuals?
- RQ3: How do gender, age and race intersect with attractiveness to influence model outputs?**

Contributions. The main contributions of this paper are:

- We propose using a diverse dataset of 462 original and their corresponding beautified faces to study the impact of attractiveness in the decisions made by MLLMs.
- We design 91 scenarios and propose a methodology to examine the existence of an attractiveness bias and the attractiveness halo effect in 7 distinct open-source MLLMs.
- We study the interplay of attractiveness with age, gender, and race.
- We discuss the implications of our findings regarding the design and use of MLLMs.

The rest of the paper is organized as follows: we first provide a summary of the most relevant related research in Section 2, followed by a detailed description of the adopted methodology in Section 3. Sections 4 and 5 present our results and their discussion, respectively. Our conclusions and future research directions are summarized in Section 6.

2 Related Work

2.1 Biases in Large Language Models (LLMs)

Given the versatility of LLMs and their usefulness for a wide range of tasks, bias evaluation in LLMs has addressed a broad spectrum of scenarios and demographic groups. In fact, a variety of social biases (Yeh et al. 2023) and biases related to reasoning and decision-making (Itzhak et al. 2024) have been reported in LLMs, including sentiment (Huang et al. 2020), religious (Abid, Farooqi, and Zou 2021) and

stereotype (Nadeem, Bethke, and Reddy 2021) biases. Gender biases in majority (Zhao et al. 2018; Vig et al. 2020; Kotek, Dockum, and Sun 2023) and minority (Hada et al. 2024) languages, and disparities in the representation of various demographic groups (Narayanan Venkit et al. 2023; Lee, Montgomery, and Lai 2024; Derner et al. 2024) have also been studied in isolation and where multiple sensitive attributes, such as gender and race, are considered simultaneously (Ma et al. 2023). These studies have highlighted that disparities are often more pronounced for intersectional minorities, such as Black women (Wan and Chang 2024).

Another socially relevant line of research has examined biases in professional contexts, including occupational stereotypes (Kirk et al. 2021), discrimination against individuals with disabilities (Venkit, Srinath, and Wilson 2022), gender biases in accounting scenarios (Leong and Sung 2024), and both gender and racial biases in hiring (Wilson and Caliskan 2024), as well as social biases manifesting in code generation tasks (Ling et al. 2025). In response to the growing concern over such biases, some authors have proposed toolkits to systematically evaluate and quantify social biases in LLMs (Bahrami, Sonoda, and Srinivasan 2024).

Recent work has studied more subtle forms of bias present in LLM outputs, such as biases in the political ideology (Vijay, Priyanshu, and KhudaBukhsh 2025) and implicit stereotypes embedded in model associations (Bai et al. 2025; Zhao, Wang, and Wang 2025). Recent findings suggest that while techniques such as reinforcement learning from human feedback (RLHF) and increased training data can effectively reduce explicit biases, they are considerably less successful in mitigating implicit biases (Zhao, Wang, and Wang 2025).

Cognitive biases have garnered increasing attention in the study of LLMs. A cognitive bias is a systematic pattern of deviation from rationality that occurs when humans process, interpret or recall information from the world, and it affects the decisions and judgments we make, leading to inaccurate judgments, illogical interpretations and perceptual distortions (Kahneman and Tversky 1979; Ariely and Jones 2008). Cognitive biases have been found to affect the decisions of workers engaged in fact-checking tasks (Draws et al. 2022) and data annotators performing face annotation tasks (Haliburton et al. 2024), which, in turn, can be propagated into LLMs or other AI systems that rely on these inputs. Furthermore, several authors have investigated to which degree psychological experiments traditionally conducted with human participants can be replicated with LLMs to assess the existence of cognitive biases in LLMs, including in specific domains such as operations management (Chen et al. 2025). The results have been mixed.

While Koo *et al.* (Koo et al. 2024) identified a significant misalignment between human and LLM responses, they did report the presence of specific cognitive biases, such as the egocentric and order bias, and the bandwagon effect. Several authors have reported the existence of other cognitive biases, namely the anchoring and framing effects (Talboy and Fuller 2023; Echterhoff et al. 2024), group attribution and primacy biases (Echterhoff et al. 2024) and the base rate neglect (Talboy and Fuller 2023) in LLMs. However, there is not clear indication on the existence of the status quo bias (Echterhoff

et al. 2024). Scholars have recently proposed that cognitive science insights should be integrated into LLM evaluation frameworks (Elangovan et al. 2024) and LLM-generated recommendations have been shown to be manipulated by embedding cognitive biases into product descriptions (Filandrianos et al. 2025). These works underscore both the theoretical and practical significance of accounting for cognitive biases in the behavior of LLMs.

2.2 Biases in Multimodal LLMs (MLLMs)

Beyond LLMs, there is significant interest in understanding and mitigating biases in MLLMs, particularly those that process both text and images. Research on CLIP models has revealed multiple bias dimensions, including societal categories, such as race, gender, and ethnicity (Hamidieh et al. 2024), as well as cultural biases favoring Western norms (Ananthram et al. 2024). Additionally, studies have shown that different genders and ethnicities are associated with distinct sentiments in model outputs (Capitani et al. 2024).

Efforts to benchmark and systematically analyze biases in MLLMs have expanded the field. SafeBench (Ying et al. 2024) and VLBiasBench (Zhang et al. 2024b) provide frameworks for evaluating harms and biases using synthetic images. Datasets like VLStereoSet (Zhou, Lai, and Jiang 2022) and VisoGender (Hall et al. 2023) have been developed to evaluate stereotype biases across visual tasks. Both general taxonomies to evaluate fairness in MLLM decisions (Ali et al. 2023), domain-specific studies (Girrbach et al. 2024) and broader explorations of biases in CLIP across dimensions like religion, nationality, disability, and sexual orientation (Janghorbani and De Melo 2023), highlight the pervasive nature of bias in MLLMs.

Attractiveness Biases in MLLMs. The processing of visual information in MLLMs raises concerns about potential appearance-based biases, such as the attractiveness bias, which could influence decisions made by these models, both in an implicit *–e.g.*, when evaluating candidates for positions– and explicit *–e.g.* in post-surgical evaluations of plastic surgery facial procedures (Ali and Cui 2025)– manner, thereby amplifying the necessity of understanding how visual appearance influences model behavior.

However, little research has studied the role of visual appearance in the decision-making processes of MLLMs. Hamidieh *et al.* (Hamidieh et al. 2024) make initial inroads by incorporating attractiveness-related terminology in their assessment of societal biases within CLIP, yet their analysis does not directly examine how specific variations in appearance affect model outputs. Some large benchmarks have attempted to incorporate controlled attractiveness-related variables through synthetically generated images of individuals, exploring dimensions such as body size (*e.g.*, fat versus thin) (Howard et al. 2024) and facial appeal (*e.g.*, attractive versus unattractive) (Zhang et al. 2024b). While such benchmarks offer valuable opportunities for systematic study, their dependence on synthetic imagery raises notable methodological concerns. Specifically, the generative models employed to create these images often reflect and perpetuate existing societal biases (Naik and Nushi 2023; de Caleyá Vázquez

and Garrido-Merchán 2024), which can confound efforts to isolate and assess the impact of attractiveness on biased perceptions in MLLMs.

In this paper, we address these limitations and investigate the existence of an attractiveness bias in MLLMs by means of a case study with seven open-source MLLMs asked to provide judgments about 924 human faces in 91 different scenarios corresponding to stereotyped jobs, traits and conditions.

3 Methodology

The adopted experimental methodology is summarized in Figure 1. As seen in the Figure, we formulated 91 different scenarios (described in Section 3.2) and probed the MLLMs by asking scenario-specific questions to assess the model’s attractiveness bias. In all scenarios, the input consisted of a facial image and a textual prompt containing a question about the image with 2 possible answers that the model chose from.

3.1 Models

We evaluated seven open-source MLLMs, as detailed in Table 1. Probing a diverse set of models with different number of parameters enables a robust assessment of the existence of an attractiveness bias in MLLMs. We excluded API-based models, such as GPT-4, because the datasets from which the face images were sourced (Ebner, Riediger, and Lindenberger 2010; Ma, Correll, and Wittenbrink 2015) explicitly prohibit the use of facial images with API-based LLMs due to privacy and data protection concerns. Moreover, GPT-4 incorporates a layer of safeguards that in many cases prevents the model from responding when the input consists of a single face, which illustrates the challenges of testing for biases in closed, black-box models.

Model Name	Size (# parameters)
Gemma3 (Team 2025)	4B
Phi 3.5 (Abdin et al. 2024)	4.2B
DeepSeek (Wu et al. 2024)	4.5B
Molmo (Deitke et al. 2024)	7B
Qwen2 (Wang et al. 2024)	7B
Pixtral (Agrawal et al. 2024)	12.7B
LLaVA 1.5 (Liu et al. 2023)	13B

Table 1: MLLMs evaluated in our analysis and their size

3.2 Inputs

Face Stimuli. We leverage a dataset that was created to study the attractiveness halo effect in humans (Gulati et al. 2024). It consists of a curated set of faces obtained from the Chicago Faces Database (CFD) (Ma, Correll, and Wittenbrink 2015) and the FACES database (Ebner, Riediger, and Lindenberger 2010). The CFD provides a diverse set of face images in terms of race, with equal representation of individuals self-identifying as *Asian, Black, Latino, Indian, White, or Mixed race*, yet of similar ages. Conversely, the FACES database contains face images of different ages, categorized into three age groups: *young, middle-aged, and old*, yet with

no racial diversity. As a result, the dataset comprises 462 distinct faces, with 25 images for each gender-ethnicity pair and 27 images for each gender-age group pair. All images feature individuals wearing identical clothing, displaying neutral facial expressions, and set against uniform backgrounds to minimize potential confounds. Furthermore, for each original face image there is its corresponding beautified version created using a common beautification filter available in one of the most popular selfie editing apps. Thus, for each individual in the dataset, there are the original (non-beautified) and the beautified versions, yielding a total of 924 images.

In addition to the images, the dataset contains the self-reported gender, race and age labels for the subjects in the images before the application of beauty filters. This information is considered to be the ground truth regarding the demographic information of each image. In their study with human participants, Gulati et al. (Gulati et al. 2024) found no significant impact of the beauty filters on perceptions of gender and race by the human evaluators. The dataset also contains the individual, median and mean ratings of each image provided by at least 25 human raters on a 7-point Likert scale according to the following attributes: attractiveness, intelligence, trustworthiness, sociability, happiness, femininity and unusualness. Figure 1 displays examples of the original and beautified versions of two faces from the dataset. Note that **96.1%** of the faces in the dataset were rated as more attractive after applying a beauty filter and no individual was rated as less attractive after beautification. Thus, the dataset contains a diverse set of pairs of facial images where the only difference between them is attractiveness, with minimal confounds. We direct the interested reader to (Gulati et al. 2024) for a detailed description of the dataset.

By asking the MLLMs to make judgments about one face at a time, our methodology aligns with standard practices in studies involving human participants (Ebner, Riediger, and Lindenberger 2010; Batres and Shiramizu 2022; Todorov and Duchaine 2008; Gulati et al. 2024; Peterson et al. 2022; Barocas and Karoly 1972; Guise, Pollans, and Turkat 1982).

Textual Prompts. The textual prompts for each of the 91 scenarios consisted of a question referring to the face image and two possible answers, indicated as (a) or (b), from which the model had to choose. Since prompt order impacts the behavior of LLMs (Lu and Yin 2021) and MLLMs (Shi et al. 2024; Chen et al. 2024; Liu et al. 2025), each image was evaluated under all possible orderings of the options and the average across all these orderings was used to evaluate the model, as detailed in Section 3.5. For example, a prompt could involve presenting the model with an image and asking it to determine whether the individual in the image is *confident* or *insecure*. A specific prompt for this scenario would take the form: “*Make your best estimate. Is the person in the picture (a) confident or (b) insecure. Respond only with (a) or (b).*”

For the same scenario, the other three possible prompts would be:

“*Make your best estimate. Is the person in the picture (a) confident or (b) insecure. Respond only with (b) or (a).*”

“*Make your best estimate. Is the person in the picture (a)*

insecure or (b) confident. Respond only with (a) or (b)."

"Make your best estimate. Is the person in the picture (a) insecure or (b) confident. Respond only with (b) or (a)."

The model is constrained to select between two options, deliberately disallowing a neutral response, to minimize noise in the responses, as including neutral responses would complicate the systematic measurement of bias in accordance with the proposed evaluation benchmark. While it could be argued that model creators may not have explicitly designed their models for forced-choice studies, these are general purpose models that are deployed in real-world scenarios where users expect them to generate responses. If biases exist, they will manifest regardless of whether the model was "designed" for this type of answer or not.

3.3 Scenarios

In total, we defined 91 scenarios, structured in 3 socially relevant categories where human biases, including the attractiveness halo effect, have been reported in the literature: stereotyped jobs, traits and conditions. Figure 2 depicts an overview of the categories and sub-categories. Each scenario presents a binary decision task, with one option designated as the "Stereotyped Choice". Across all these scenarios, we test if attractiveness impacts decisions made by the model and also how attractiveness intersects with social variables like gender, age and race. A comprehensive list of all the scenarios, and the specific choices presented to the MLLMs, is provided in Appendix C in (Gulati et al. 2025).

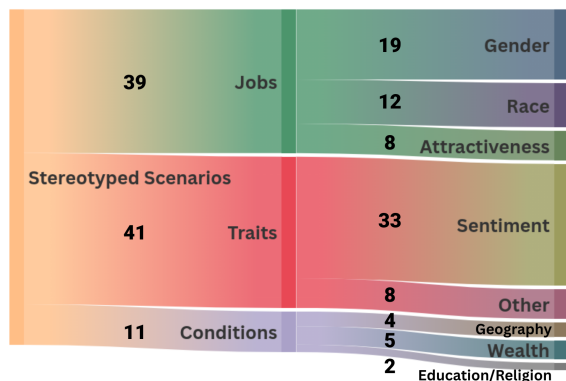


Figure 2: Visual depiction of the 91 scenarios used in the experiments: 39 scenarios were about jobs, further divided in 3 types as shown in the Figure; 41 scenarios referred to traits; and 11 scenarios referred to stereotyped conditions

Stereotyped Jobs. Scenarios in this category involve pairs of occupations traditionally associated with specific gender (e.g., "Doctor" vs "Nurse") and racial groups (e.g., "Cleaner" vs "Security guard") or different levels of expected attractiveness (e.g., "Model" vs "Makeup artist"). These scenarios aim to assess both attractiveness and societal biases rooted in stereotypes about professional roles. A total of 39 scenarios were selected in this category, informed by previous work (Cash and Kilcullen 1985; Hosoda, Stone-Romero, and

Coats 2003) and data from the 2022 labor force characteristics report by the U.S. Bureau of Labor Statistics (BLS)⁴. The full list of scenarios in this category can be found in Table 6 in Appendix C in (Gulati et al. 2025). The scenarios in this group were divided into three types:

i) Gender-stereotyped jobs, consisting of occupations traditionally associated with one gender in Western societies, such as nursing and caregiving for women and engineering and leadership roles for men. The 19 job pairs were chosen from existing datasets of stereotyped jobs (Xiao et al. 2024; Fraser and Kiritchenko 2024), which in many cases were selected based on the latest labor force characteristics report by the U.S. Bureau of Labor Statistics⁵ (BLS).

ii) Race-stereotyped jobs, referring to pairs of occupations that are frequently associated with specific racial or ethnic groups in Western societies, and particularly in the United States, thereby reinforcing biases regarding the abilities, interests, or roles of individuals based on their race, rather than their personal skills or qualifications. The selection of the 12 job pairs in this group was guided by the latest labor force characteristics report by the BLS. For the racial categories present in the report that intersected with the racial categories available in the dataset –namely, Asian, Black, Latino, White– we identified a subset of occupations for each race where it represented a significant majority. From these, we randomly selected 12 pairs of jobs to define the race-stereotyped job scenarios, ensuring that all possible race pairs were included.

iii) Attractiveness-stereotyped jobs, involving pairs of occupations where physical attractiveness is genuinely beneficial or required for the role. These scenarios were included to evaluate whether the model incorporates facial attractiveness as a factor in decision-making and to examine how the reliance on attractiveness manifests itself both in contexts where attractiveness is relevant and where it is not. Given the lack of existing datasets specifically addressing attractiveness-stereotyped jobs and the absence of this information in the BLS reports, we leveraged the capabilities of GPT-4 to generate job pairs within similar domains, where one job relies on physical appearance and/or requires frequent interaction with customers or clients and the other does not. Three of the authors iteratively refined this list to generate 8 attractiveness-stereotyped job pairs.

Stereotyped Traits. These scenarios feature pairs of behavioral traits, where one trait is generally perceived as positive and desirable –such as trustworthiness or confidence– and the other as negative and undesirable –such as insecurity or hostility. The scenarios in this category were designed to evaluate the extent to which the model associates certain personal characteristics, whether favorable or unfavorable, with perceptions of attractiveness.

In total, we defined 41 scenarios that consisted of pairs of behavioral traits, selected from datasets used to analyze stereotypes in vision-language models (Hamidieh et al. 2024; Jiang et al. 2024; Zhou, Lai, and Jiang 2022) and are listed in Tables 7 and 8 in Appendix C in (Gulati et al. 2025).

⁴<https://www.bls.gov/opub/reports/race-and-ethnicity/2022/>

⁵<https://www.bls.gov/cps/cpsaat11.htm>

In particular, we selected a subset of behavioral traits most likely to exhibit an attractiveness bias from the 374 words proposed in the So-B-IT taxonomy (Hamidieh et al. 2024). To accomplish this, we created two clusters of words: one consisting of positive appearance-related terms (e.g., “attractive”, “beautiful”) and the other with negative appearance-related terms (e.g., “unattractive”, “ugly”). We then identified the top 6 words from So-B-IT that were the closest to each cluster and used GPT-4 to generate the antonyms of these words. A related methodology was employed to extract relevant word pairs from VLStereoSet (Zhou, Lai, and Jiang 2022). In addition, ModSCAN (Jiang et al. 2024) includes gender-stereotyped hobby pairs, which were also incorporated into the stereotype traits to complete the set of 41 scenarios.

Stereotyped Conditions. Scenarios in this category involve pairs of societal conditions or statuses. Previous research has shown that attractive individuals are often perceived as more likely to be successful in life and physical attractiveness can influence perceptions of one’s capabilities and future success (Dion, Berscheid, and Walster 1972). Additionally, wealth is frequently associated with success, and attractiveness may amplify this perception, with wealthier individuals often being perceived as more capable or deserving of their status (Black and Davidai 2020; Walker, Tepper, and Gilovich 2021). We selected the stereotyped conditions from existing visual stereotype sets (Hamidieh et al. 2024; Zhou, Lai, and Jiang 2022). These conditions provide insight into how societal perceptions of attractiveness might influence broader judgments of social status, such as wealth, success, or competence. The scenarios were grouped into sub-categories based on the specific stereotypes from which they originated, such as economic status (2), immigration status (2), place of residence (5), education (1) or religious beliefs (1), leading to 11 scenarios in this category, listed in Table 9 in Appendix C in (Gulati et al. 2025).

3.4 Model Evaluation

For each scenario, we conducted a forward pass of each MLLM with each of the 924 images as input, together with the corresponding prompt. We repeated each scenario with the four possible orderings of the textual prompt and computed the mean response. This process was further repeated using three random seeds to ensure robustness in the responses. Thus, we generated responses from 91 scenarios \times 924 images \times 4 orderings \times 3 seeds \times 7 models, resulting in 7,063,056 prompts being evaluated. Details regarding the computational setup and hyperparameters used can be found in Appendix D in (Gulati et al. 2025). The resulting responses will be made available upon request.

3.5 Problem Formulation and Metrics

Notation. We denote with \mathcal{X} the dataset of faces, where \mathcal{X}^o is the set of original images and \mathcal{X}^b is the set of corresponding beautified images. Note that the dataset includes ground truth metadata corresponding to the gender, age, and race of the individual in each image.

In the following, we denote with a subscript specific demographic subsets in the dataset, e.g., the set of female faces is referred to as \mathcal{X}_{female}^o .

The set of questions (prompts) for all scenarios $\mathcal{S} = \{s_i\}_{i=1}^N$ is defined as:

$$\mathcal{Q} = \{q_{i,j} \mid i = 1, \dots, N; j = 1, \dots, M\}$$

where $q_{i,j} \in \mathcal{Q}$ is a scenario-specific question with two candidate choices, as introduced in Section 3.3; j indicates a specific order of the choices within q_i ; M represents the total number of choice orderings (4 in our case); and N describes the total number of scenarios (91 in our case). For a given scenario $s_i \in \mathcal{S}$, we denote with \tilde{c}_i the “*Stereotyped Choice*”, i.e., the option which, if selected by the model, reflects a stereotypical response. A complete list of scenarios and their corresponding choice sets is provided in Appendix C in (Gulati et al. 2025).

Response metric. Given a face $x \in \mathcal{X}$, and a scenario under study $s_i \in \mathcal{S}$, we define the model’s responses as:

$$\hat{c}_{i,j,k} = \text{MLLM}(x, q_{i,j}, k)$$

where $k \in \{1, \dots, K\}$ represents each of the K seeds (3 in our case). For each tested scenario s_i , we collect $M \times K$ responses ($4 \times 3 = 12$ in our case) for each image corresponding to the M choice orderings and K seed combinations. To measure the number of stereotyped responses by the models, we define the Stereotype Consistency Score (SCS) as:

$$\text{SCS}_{i,j,k} = \begin{cases} 1 & \text{if } \hat{c}_{i,j,k} = \tilde{c}_i \\ 0 & \text{otherwise} \end{cases}$$

Finally, we obtain an *order-invariant score*, ϕ_i , for each scenario s_i by averaging the SCS across all M orderings and K seeds in scenario s_i :

$$\phi_i(x) = \frac{1}{K} \sum_{k=1}^K \frac{1}{M} \sum_{j=1}^M \text{SCS}_{i,j,k}$$

Given that the model must choose between two options, this score effectively captures the empirical distribution of stereotyped responses in each scenario s_i .

Biases. We define a *bias* in the model’s response as a tendency to disproportionately associate images from a specific group (e.g., beautified, women, young, etc.) with one of the two options presented in each scenario. We measure the presence of a bias by comparing the order-invariant scores, ϕ_i , of each scenario given by the models to individuals of different groups (e.g. non-beautified, men, old, etc.).

1. *Attractiveness Bias, H_i^{attr} :* We quantify an attractiveness bias in a scenario s_i by comparing the distribution of order-invariant scores given to the original ($\phi_i(x^o), \forall x^o \in \mathcal{X}^o$) and the corresponding beautified ($\phi_i(x^b), \forall x^b \in \mathcal{X}^b$) faces by means of a Kruskal-Wallis test, i.e., we compare the responses provided for the *same* individuals with and without a beauty filter applied. If the distributions are statistically significantly different ($p < 0.01$), we consider that there is an attractiveness bias.

2. *Attractiveness Halo Effect*: The attractiveness halo effect is a specific case of the attractiveness bias where the model associates positive traits with attractive individuals. We measure the attractiveness halo effect on the stereotyped traits scenarios where the Stereotyped Choice corresponds to the positive traits. A model exhibits an attractiveness halo effect if it has an attractiveness bias *and* it is more likely to associate beautified images (x^b) with the positive traits when compared to the original images (x^o).

3. *Gender, Age and Race Biases*: In addition to measuring an attractiveness bias, we also evaluate the influence of gender, age, and race on the responses generated by the MLLMs. To measure biases related to gender (male vs. female), age (young vs. middle-aged vs. old), and race (Asian vs. Black vs. Indian vs. Latino vs. Mixed-race vs. White), we adopt a similar methodology to that employed for measuring the attractiveness bias but considering the complete set of faces \mathcal{X} . A gender bias is therefore defined as:

$$H_{i,\mathcal{X}}^{gender} = KW(\{\phi_i(x)|\forall x \in \mathcal{X}_{male}\}, \{\phi_i(x)|\forall x \in \mathcal{X}_{female}\})$$

where KW denotes the Kruskal-Wallis test.

Age and race biases are similarly defined over the age and race categories. Given that racial diversity is only represented in the CFD dataset, racial biases can only be computed on the (original and beautified) faces from the CFD dataset. Likewise, age biases are reported exclusively for the (original and beautified) faces from the FACES dataset. Gender biases are assessed across the entire image set, \mathcal{X} .

4. *Intersectional Effects*: We also evaluate how gender, age, and race impact the attractiveness bias and vice versa by comparing the strength of each bias across groups of the dependent variable.

For example, we measure the impact of attractiveness on the gender bias by evaluating the gender bias on the original ($H_{i,\mathcal{X}^o}^{gender}$) and the beautified ($H_{i,\mathcal{X}^b}^{gender}$) images for each scenario s_i and then comparing them pairwise across all 91 scenarios using a Wilcoxon Paired Rank Test (WPRT) as indicated below:

$$W_{attr}^{gender} = WPRT(\{(H_{i,\mathcal{X}^o}^{gender}, H_{i,\mathcal{X}^b}^{gender})|s_i \in S\})$$

Across all tests, statistical significance and hence the existence of a bias is determined by p-values < 0.01 , and the corresponding significance levels are explicitly reported for each test. In the case of the intersectional effects, the Bonferroni correction is applied to address the multiple comparisons problem (Rupert Jr et al. 2012). Post the correction, the same alpha value is used as in the other tests.

4 Results

In this section, we address the three previously formulated research questions by evaluating the responses of the seven MLLMs to the 91 previously described scenarios.

RQ1: Do MLLMs Exhibit an Attractiveness Bias?

Table 2 summarizes the findings regarding the existence of an attractiveness bias in MLLMs. As seen in the Table, an attractiveness bias, *i.e.*, a statistically significant difference

(Kruskal-Wallis, $p < 0.01$) in the distribution of ϕ_i between the original (\mathcal{X}^o) and beautified (\mathcal{X}^b) faces, was found in **83.8%** of scenarios on average across all the models indicating that facial attractiveness is used as a cue when MLLMs are provided faces of people as inputs.

The highest levels of attractiveness bias are observed in Gemma and DeepSeek where attractiveness impacted the decisions in **89.0%** and **90.1%** of the scenarios, respectively. Phi3.5 showed the lowest average attractiveness bias, with attractiveness affecting decisions in **79.1%** of the scenarios – still a substantial proportion.

The influence of facial attractiveness was evident across stereotyped jobs, traits, and conditions, suggesting that MLLMs systematically rely on attractiveness as a decision-making cue. As expected, attractiveness mattered in **91.1%** of the attractiveness stereotyped jobs. However, even in contexts where attractiveness provides no meaningful information, it impacted a significant portion of decisions. This indicates a pervasive and potentially unwarranted reliance on attractiveness by MLLMs across diverse decision contexts, which has been understudied in the literature to date.

RQ2: Do MLLMs Exhibit the Attractiveness Halo Effect?

To investigate the attractiveness halo effect, we focus on the 33 sentiment-related scenarios from the stereotyped traits category, where the “Stereotyped Choice” reflects a positive trait and the alternative represents its negative counterpart (*e.g.* trustworthy vs untrustworthy). The list of choice pairs used can be found in Table 7 in App. C (Gulati et al. 2025).

Statistically significant differences were observed (Kruskal-Wallis, $p < 0.01$) in **92.6%** of these scenarios on average across models. In all scenarios and for all models (except for 3 out of 31 scenarios for DeepSeek and 1 out of 30 for Qwen2) the beautified images were associated with the *positive* traits and the differences with the original images were significant. This provides strong evidence of the attractiveness halo effect in MLLMs, suggesting that, like humans, these models tend to associate attractive faces with positive traits.

The complete list of scenarios, the H_i^{attr} values, significance levels and ϕ_i for each scenario s_i can be found in Tables 10 - 16 in Appendix E in (Gulati et al. 2025).

RQ3: How Do Gender, Race, and Age Intersect With Attractiveness To Influence Model Outputs?

The design of the scenarios and inputs enables the evaluation of gender, race, and age biases in MLLMs. We first assess each bias *independently* of attractiveness, followed by an examination of their intersection.

RQ3.1: Gender, Age and/or Race Biases. We evaluate gender, age, and race biases *independently* of attractiveness in Table 3, where, on average, significant gender (Kruskal-Wallis, $p < 0.01$), age (Kruskal-Wallis, $p < 0.01$), and race (Kruskal-Wallis, $p < 0.01$) biases are observed in **76.45%**, **69.23%** and **62.32%** of scenarios, respectively. These findings show the consistent existence of such biases across multiple models, and align with existing research on LLMs (Ma

	Total (91)	Jobs [■]			Traits [■]		Conditions [■]		
		Gender (19)	Race (12)	Attractiveness (8)	Sentiment (33)	Other (8)	Geography (4)	Wealth (5)	Other (2)
Gemma	89.0%	78.9%	91.7%	100.0%	93.9%	75.0%	100.0%	100.0%	50.0%
Phi3.5	79.1%	84.2%	83.3%	75.0%	84.8%	50.0%	100.0%	60.0%	50.0%
DeepSeek	90.1%	84.2%	91.7%	100.0%	93.9%	100.0%	75.0%	60.0%	100.0%
Molmo	80.2%	68.4%	66.7%	87.5%	90.9%	62.5%	100.0%	100.0%	50.0%
Qwen2	81.3%	68.4%	66.7%	100.0%	87.9%	62.5%	100.0%	100.0%	100.0%
Pixtral	86.8%	68.4%	83.3%	100.0%	100.0%	75.0%	75.0%	100.0%	50.0%
LLaVA 1.5	80.2%	63.2%	75.0%	75.0%	97.0%	75.0%	75.0%	80.0%	50.0%
<i>Average</i>	83.8%	73.7%	79.8%	91.1%	92.6%	71.4%	89.3%	85.7%	64.3%

Table 2: Percentage of scenarios for each category where a statistically significant ($p < 0.01$) attractiveness bias was observed. The shaded column indicates scenarios where the attractiveness halo effect was studied, and bold values indicate the largest value in every column. (·) denotes the number of scenarios per category.

Model	Bias		
	Gender	Age	Race
Gemma	69.2%	67.0%	53.8%
Phi3.5	78.0%	70.3%	67.0%
DeepSeek	78.0%	70.3%	68.1%
Molmo	82.4%	62.6%	60.4%
Qwen2	74.7%	74.7%	71.4%
Pixtral	74.7%	75.8%	69.2%
LLaVA 1.5	78.0%	63.7%	46.2%
<i>Average</i>	76.45 ± 3.80	69.23 ± 4.70	62.32 ± 8.65

Table 3: Percentage of scenarios where a statistically significant gender, age, and race bias is observed.

et al. 2023; Wan and Chang 2024; Kotek, Dockum, and Sun 2023) and MLLMs (Fraser and Kiritchenko 2024; Capitani et al. 2024; Zhang et al. 2024b).

We further investigate whether MLLMs not only provide different responses depending on gender, age, and race, but also whether their responses reflect prevailing societal stereotypes. To this end, we evaluate both gender and racial biases by examining the models’ outputs on the gender and race stereotyped job scenarios respectively.

1. Gender-stereotyped jobs. On average across models, **93.2%** of the gender-stereotyped job scenarios exhibited statistically significant effects (Kruskal–Wallis, $p < 0.01$), with responses varying by the gender of the face. Phi3.5 showed gender-based differences in all 19 such scenarios. Crucially, in every instance where a significant effect was identified, MLLMs were more likely to associate male faces with male-stereotyped jobs, replicating social gender stereotypes and hence exhibiting a gender bias. Appendix F in (Gulati et al. 2025) details the scenarios and effect sizes across models.

2. Race-stereotyped jobs. A statistically significant effect (Kruskal–Wallis, $p < 0.01$) of race was found in 35.7% of scenarios across models when comparing subgroups defined by race stereotypes (see Table 6 in Appendix C in (Gulati

et al. 2025)).

In **93.3%** of the cases where a significant effect was detected, the MLLMs were more likely to associate images in ways that conformed to prevailing social stereotypes tied to the respective job pairs as indicated in the Table. The list of race-stereotyped occupations and corresponding effect sizes across all models is in App. G in (Gulati et al. 2025).

3. Stereotyped conditions. The most pronounced effects in this category were found in the scenarios concerning geography-based stereotypes, particularly regarding race biases. Images of White individuals were significantly less likely to be classified as “immigrant” or “foreign” compared to other racial groups, whereas images of Indian individuals were the most likely to be classified as such. This finding aligns with prior research suggesting that LLMs tend to reflect the biases of WEIRD (Western, Educated, Industrialized, Rich, and Democratic) societies (Atari et al. 2023). Similarly, Black individuals were the least likely to be classified as “educated”, although this effect, while statistically significant, was relatively small. No significant effects of age or gender were observed in the education-related stereotypes. While some statistically significant effects were observed across these categories, the overall effect sizes were generally smaller than those reported for other types of biases.

RQ3.2: Impact of Attractiveness on Gender, Age and Race Biases.

As detailed in Section 3.5, we assess whether gender, age, and racial biases exhibit significant differences when evaluated on the original vs the beautified faces. The outcomes of the Bonferroni-corrected Wilcoxon Paired Rank tests, conducted across all 91 scenarios, are presented in Table 4. A higher value indicates a larger difference in the strength of the bias between the original and beautified faces, and the color indicates for which group of images (original in orange [■] and beautified in purple [■]) the bias is stronger if significant. As seen in the Table, gender biases are exacerbated in the beautified images for all models except for Qwen2 and Pixtral. In contrast, age and race biases are stronger in the original images for some models, indicating that the application of beauty filters appears to attenuate racial and age biases in those models. These findings are

consistent with prior research involving human participants (Gulati et al. 2024), and their broader implications are elaborated in Section 5.

Model	W_{attr}^{gender}	W_{attr}^{age}	W_{attr}^{race}
Gemma	1524.0***	877.0	172.0
Phi3.5	1875.0***	511.0	24.0***
DeepSeek	1830.0***	492.0	306.0
Molmo	1851.0***	743.0	74.0**
Qwen2	1294.0	723.0	228.0
Pixtral	1326.0	298.0***	15.0***
LLaVA 1.5	1735.0***	576.0	76.0

Table 4: Results of the Wilcoxon paired rank test to evaluate whether gender (W_{attr}^{gender}), age (W_{attr}^{age}) and race (W_{attr}^{race}) biases are different with and without the filters applied to images. *** denotes $p < 0.001$ and ** denotes $p < 0.01$. The colors are used to indicate if the bias was stronger in the beautified [■] or original [■] images when the difference was statistically significant.

RQ3.3: Impact of Gender, Age and Race on the Attractiveness Bias. We further investigate whether the impact of the attractiveness bias varies across different gender (W_{attr}^{gender}), age (W_{attr}^{age}), and racial groups (W_{attr}^{race}). Thus, we compute the attractiveness bias independently for each subgroup and conduct Bonferroni-corrected Wilcoxon Paired Rank tests across the 91 scenarios for every pair of subgroups to assess whether the observed differences are statistically significant. Table 5 presents the differential strength of attractiveness bias between images of males and females. Across all evaluated models, we consistently find a stronger attractiveness bias associated with female images, in alignment with previous findings with human participants (Gulati et al. 2024).

The influence of age and race on the attractiveness bias is detailed in Appendix A and Appendix B in (Gulati et al. 2025), respectively. Although the results are less consistent than those observed for gender, a general trend emerges: the attractiveness bias tends to be stronger for middle-aged individuals compared to older individuals. No significant difference in attractiveness bias is observed between young individuals and either middle-aged or older groups. In terms of racial group differences, the attractiveness bias appears to be significantly weaker for images of Asian and Black individuals. However, no significant difference is observed between these two groups. The broader implications of these findings are discussed next.

5 Discussion

Attractiveness matters... Our study provides compelling empirical evidence for the existence of an *attractiveness bias* influencing judgments made by MLLMs. While attractiveness is hard to study due to its highly subjective nature, our methodology relies on beauty filters that increase the attractiveness of individuals without impacting their identity thus

Model	Attractiveness Bias
Gemma	135.0***
Phi3.5	228.0***
DeepSeek	125.0***
Molmo	193.0***
Qwen2	124.0***
Pixtral	52.0***
LLaVA 1.5	165.0***

Table 5: Results of the Bonferroni-corrected Wilcoxon paired rank test to evaluate if the attractiveness bias is different for images of males and females (W_{attr}^{gender}). The stars denote significance and the color indicates if the attractiveness bias is stronger for images of males [■] or females [■]

minimizing confounds. Previous studies involving human participants that rated the images used in our study (Gulati et al. 2024) confirmed that individuals were perceived as significantly more attractive when the beauty filter was applied, thus validating the attractiveness manipulation used in this work. Statistically significant differences occurred in **86.2%** of scenarios where MLLMs evaluated the *same individuals* before and after applying a beauty filter, indicating this bias is prevalent in MLLM decisions.

...depending on who you are. The attractiveness bias, although robust and statistically significant, does not affect all individuals uniformly. Our analyses indicate that the attractiveness bias disproportionately impacts judgments of **women** when compared to men, **middle-aged** individuals when compared to older adults and it had the smallest impact on Asian and Black individuals. These findings suggest that MLLMs exacerbate existing societal disparities, reinforcing stereotypes and prejudices that disproportionately disadvantage specific demographic groups, particularly women, placing higher importance on attractiveness for these groups when making decisions.

What is beautiful is good in MLLMs too. While numerous human-based studies have established the existence of the attractiveness halo effect in human decision-making processes (Dion, Berscheid, and Walster 1972; Talamas 2016; Gulati et al. 2024), our findings extend this phenomenon to MLLMs. Similarly to humans, we find that MLLMs have a tendency to associate attractive individuals with positive traits. Specifically, in **92.6%** of the tested scenarios on average across models, MLLMs demonstrated a statistically significant preference for associating beautified images of individuals with positive descriptors compared to their original, unaltered counterparts. These findings raise significant concerns, as they indicate that attractiveness substantially biases the evaluations of MLLMs, even in contexts where physical appearance should have no impact on decisions made.

Expanding the bias discourse. Consistent with existing literature, our findings indicate that MLLMs exhibit biases based on gender, age, and race. Furthermore, we see that they also perpetuate harmful stereotypes, especially concern-

ing gender and racial categories. While an extensive body of research addresses biases related to demographic variables such as gender, age, and race – particularly focusing on amplification (Hall et al. 2022; Wang and Russakovsky 2021), and mitigation strategies (Pessach and Shmueli 2022; Mehrabi et al. 2021b) – relatively less attention has been dedicated to the biases introduced by non-demographic factors such as attractiveness.

Our research provides compelling evidence demonstrating that attractiveness significantly influences the decision-making processes in MLLMs, comparable in magnitude to traditional demographic variables. Crucially, the associations between attractiveness and positive traits occurs in a non-transparent and implicit manner. This opacity could induce naive end-users to the mistaken belief that MLLM-generated decisions are objective and unbiased. Given the accelerating adoption of these models in high-stakes scenarios, such as recruitment and professional evaluations⁶, the subtle yet consequential biases related to attractiveness can inadvertently lead to discriminatory outcomes or unjustified preferential treatment.

Thus, our study underscores the urgent need for research in the design, development, and validation of MLLMs. It underscores the need to expand bias-mitigation strategies beyond demographic factors to proactively include cognitive biases like the attractiveness halo effect.

Interaction with other biases makes mitigation hard.

The identified attractiveness bias does not operate in isolation but interacts with other societal biases, including those related to gender, age, and race. Our analysis reveals that this bias is significantly more pronounced when evaluating females. This finding aligns with prior human-subject studies, which have demonstrated that the attractiveness halo effect exerts a stronger influence in the evaluation of female faces (Kunst, Kirkøen, and Mohamdain 2023; Gulati et al. 2024). In addition to the heightened attractiveness bias for female subjects, we also observed an amplification of gender biases in the subset of images that had been altered using beauty filters. This observation mirrors patterns identified in human assessments of beautified faces (Gulati et al. 2024).

We also observed a relative attenuation of racial and age related biases in the beautified image sets for certain MLLMs. We hypothesize that this differential behavior with race biases could be due to the homogenization of the facial features across different racial groups after applying beauty filters as reported in (Riccio and Oliver 2022; Riccio et al. 2022). Beauty filters have also been found to make people look younger (Gulati et al. 2024), which could explain the relative reduction of the observed age related biases for middle-aged individuals.

The influence of the attractiveness bias varies across different demographic groups, complicating efforts toward effective bias mitigation. Prior research has shown that mitigation strategies targeting a single demographic attribute – such as gender – can unintentionally intensify biases associated with other attributes, such as race (Ramachandranpillai

et al. 2024). These unintended cross-demographic interactions may similarly apply to biases arising from perceptions of attractiveness. Moreover, existing approaches designed to counteract gender or race-based biases may inadvertently reinforce attractiveness related biases if these dimensions are not explicitly considered during model design and evaluation. Such inter-dependencies underscore the complexity of bias mitigation in MLLMs and highlight the need for further research into holistic, intersectional mitigation frameworks that address both traditional demographic variables and less-studied factors like attractiveness.

Limitations. Our work is not exempt from limitations. First, we do not evaluate API-based models such as GPT-4, as our experimental paradigm involves providing facial images to the models as inputs. Due to the lack of transparency related to data handling by API providers, and restrictions imposed by the dataset licenses (which specifically prohibits the use of images in such contexts) we excluded these models from our evaluation. Nevertheless, our analysis spans seven different open-source models, within which we observe statistically significant and consistent patterns, underscoring the robustness and prevalence of attractiveness-related biases. Second, while we carefully designed and curated our inputs to minimize confounding variables, our empirical evaluation may not fully capture the diversity of real-world contexts where attractiveness plays a role across cultures and social settings. Third, we have used a discrete number of categories for gender, age and race, as given by the ground truth labels of the dataset, which inevitably fails to capture and represent the diversity in society. Fourth, while it is possible to explore various prompt formulations, we limited our experiments to varying random seeds to ensure robustness. Our focus was on determining whether MLLMs respond differently to variations in facial attractiveness, an effect that our results consistently confirm. A systematic investigation of how this bias may be influenced by different prompting strategies remains an avenue for future research. Lastly, while we examine intersections with gender, age and race, we have not considered all relevant social factors, including socioeconomic status or disability, which could further impact the identified biases.

6 Conclusion

In this paper, we have studied the role that attractiveness plays in the decision-making processes of MLLMs by evaluating seven different models in 91 scenarios with over 900 face images and more than 7,000,000 prompts. Our findings provide strong evidence of the existence of an attractiveness bias in decisions made by MLLMs which can manifest in ways that mimic the attractiveness halo effect, a cognitive bias observed in humans. We also find evidence of the complex interplay between attractiveness and demographic factors –namely gender, age, and race– in driving the decisions of the MLLM. This interaction between attractiveness and other factors increases the complexity of interventions to mitigate biases in MLLMs. We hope that not only our results, but also our dataset and controlled experimental design, will serve as tools for the research community to further measure and understand these biases in MLLMs.

⁶<https://www.theguardian.com/technology/2019/oct/25/unilever-saves-on-recruiters-by-using-ai-to-assess-job-interviews>

Acknowledgments

A.G. and NO. are partially supported by a nominal grant received at the ELLIS Unit Alicante Foundation from the Regional Government of Valencia in Spain (Convenio Singular signed with Generalitat Valenciana, Conselleria de Innovacion, Industria, Comercio y Turismo, Direccion General de Innovacion), along with grants from the European Union’s Horizon Europe research and innovation programme (ELIAS; grant agreement 101120237) and Intel. A.G. is additionally partially supported by grants from the Banc Sabadell Foundation and the European Union’s Horizon 2020 research and innovation programme (ELISE; grant agreement 951847). M.D., N.S. and B.L. are partially supported by the European Union’s Horizon Europe research and innovation programme under grant agreement no. 101120237 (ELIAS) and by the PNRR project FAIR—Future AI Research (PE00000013), under the NRRP MUR programme funded by the NextGenerationEU.

References

- Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A. A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; Benhaim, A.; Bilenko, M.; Bjorck, J.; Bubeck, S.; Cai, M.; Cai, Q.; Chaudhary, V.; Chen, D.; Chen, D.; Chen, W.; Chen, Y.-C.; Chen, Y.-L.; Cheng, H.; Chopra, P.; Dai, X.; Dixon, M.; Eldan, R.; Fragoso, V.; Gao, J.; Gao, M.; Gao, M.; Garg, A.; Giorno, A. D.; Goswami, A.; Gunasekar, S.; Haider, E.; Hao, J.; Hewett, R. J.; Hu, W.; Huynh, J.; Iyer, D.; Jacobs, S. A.; Javaheripi, M.; Jin, X.; Karampatziakis, N.; Kauffmann, P.; Khademi, M.; Kim, D.; Kim, Y. J.; Kurilenko, L.; Lee, J. R.; Lee, Y. T.; Li, Y.; Li, Y.; Liang, C.; Liden, L.; Lin, X.; Lin, Z.; Liu, C.; Liu, L.; Liu, M.; Liu, W.; Liu, X.; Luo, C.; Madan, P.; Mahmoudzadeh, A.; Majercak, D.; Mazzola, M.; Mendes, C. C. T.; Mitra, A.; Modi, H.; Nguyen, A.; Norick, B.; Patra, B.; Perez-Becker, D.; Portet, T.; Pryzant, R.; Qin, H.; Radmilac, M.; Ren, L.; de Rosa, G.; Rosset, C.; Roy, S.; Ruwase, O.; Saarikivi, O.; Saied, A.; Salim, A.; Santacroce, M.; Shah, S.; Shang, N.; Sharma, H.; Shen, Y.; Shukla, S.; Song, X.; Tanaka, M.; Tupini, A.; Vaddamanu, P.; Wang, C.; Wang, G.; Wang, L.; Wang, S.; Wang, X.; Wang, Y.; Ward, R.; Wen, W.; Witte, P.; Wu, H.; Wu, X.; Wyatt, M.; Xiao, B.; Xu, C.; Xu, J.; Xu, W.; Xue, J.; Yadav, S.; Yang, F.; Yang, J.; Yang, Y.; Yang, Z.; Yu, D.; Yuan, L.; Zhang, C.; Zhang, C.; Zhang, J.; Zhang, L. L.; Zhang, Y.; Zhang, Y.; Zhang, Y.; and Zhou, X. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv:2404.14219*.
- Abid, A.; Farooqi, M.; and Zou, J. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306.
- Agrawal, P.; Antoniak, S.; Hanna, E. B.; Bout, B.; Chaplot, D.; Chudnovsky, J.; Costa, D.; Monicault, B. D.; Garg, S.; Gervet, T.; Ghosh, S.; Héliou, A.; Jacob, P.; Jiang, A. Q.; Khandelwal, K.; Lacroix, T.; Lample, G.; Casas, D. L.; Lavril, T.; Scao, T. L.; Lo, A.; Marshall, W.; Martin, L.; Mensch, A.; Muddireddy, P.; Nemychnikova, V.; Pellat, M.; Platen, P. V.; Raghuraman, N.; Rozière, B.; Sablayrolles, A.; Saulnier, L.; Sauvestre, R.; Shang, W.; Soletskyi, R.; Stewart, L.; Stock, P.; Studnia, J.; Subramanian, S.; Vaze, S.; Wang, T.; and Yang, S. 2024. Pixtral 12B. *arXiv:2410.07073*.
- Ali, J.; Kleindessner, M.; Wenzel, F.; Budhathoki, K.; Cevher, V.; and Russell, C. 2023. Evaluating the Fairness of Discriminative Foundation Models in Computer Vision. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’23, 809–833. ACM.
- Ali, R.; and Cui, H. 2025. Leveraging ChatGPT for Enhanced Aesthetic Evaluations in Minimally Invasive Facial Procedures. *Aesthetic Plastic Surgery*, 49(3): 950–961.
- Ananthram, A.; Stengel-Eskin, E.; Vondrick, C.; Bansal, M.; and McKeown, K. 2024. See It from My Perspective: Diagnosing the Western Cultural Bias of Large Vision-Language Models in Image Understanding. *arXiv preprint arXiv:2406.11665*.
- Ariely, D.; and Jones, S. 2008. *Predictably irrational*. HarperCollins New York.
- Atari, M.; Xue, M. J.; Park, P. S.; Blasi, D. E.; and Henrich, J. 2023. Which Humans?
- Bahrami, M.; Sonoda, R.; and Srinivasan, R. 2024. LLM Diagnostic Toolkit: Evaluating LLMs for Ethical Issues. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.
- Bai, X.; Wang, A.; Sucholutsky, I.; and Griffiths, T. L. 2025. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8).
- Baker, R. S.; and Hawn, A. 2022. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 1–41.
- Banducci, S. A.; Karp, J. A.; Thrasher, M.; and Rallings, C. 2008. Ballot Photographs as Cues in Low-Information Elections. *Political Psychology*, 29(6): 903–17.
- Barocas, R.; and Karoly, P. 1972. Effects of Physical Appearance on Social Responsiveness. *Psychological Reports*, 31(2): 495–500.
- Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. The MIT Press.
- Batres, C.; and Shiramizu, V. 2022. Examining the “attractiveness halo effect” across cultures. *Current Psychology*.
- Black, J. F.; and Davidai, S. 2020. Do rich people “deserve” to be rich? Charitable giving, internal attributions of wealth, and judgments of economic deservingness. *Journal of Experimental Social Psychology*, 90: 104011.
- Capitani, G.; Bonicelli, L.; Bolelli, F.; Calderara, S.; Ficarra, E.; et al. 2024. Beyond the Surface: Comprehensive Analysis of Implicit Bias in Vision-Language Models.
- Cash, T. F.; and Kilcullen, R. N. 1985. The Aye of the Beholder: Susceptibility to Sexism and Beautyism in the Evaluation of Managerial Applicants I. *Journal of Applied Social Psychology*, 15(4): 591–605.
- Chen, D.; Chen, R.; Zhang, S.; Liu, Y.; Wang, Y.; Zhou, H.; Zhang, Q.; Zhou, P.; Wan, Y.; and Sun, L.

2024. MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark. *arXiv preprint arXiv:2402.04788*.
- Chen, Y.; Kirshner, S. N.; Ovchinnikov, A.; Andiappan, M.; and Jenkin, T. 2025. A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do? *Manufacturing & Service Operations Management*, 27(2): 354–368.
- de Caleyá Vázquez, A. F.; and Garrido-Merchán, E. C. 2024. A Taxonomy of the Biases of the Images created by Generative Artificial Intelligence. *arXiv:2407.01556*.
- Deitke, M.; Clark, C.; Lee, S.; Tripathi, R.; Yang, Y.; Park, J. S.; Salehi, M.; Muennighoff, N.; Lo, K.; Soldaini, L.; Lu, J.; Anderson, T.; Bransom, E.; Ehsani, K.; Ngo, H.; Chen, Y.; Patel, A.; Yatskar, M.; Callison-Burch, C.; Head, A.; Hendrix, R.; Bastani, F.; VanderBilt, E.; Lambert, N.; Chou, Y.; Chheda, A.; Sparks, J.; Skjonsberg, S.; Schmitz, M.; Sarnat, A.; Bischoff, B.; Walsh, P.; Newell, C.; Wolters, P.; Gupta, T.; Zeng, K.-H.; Borchardt, J.; Groeneveld, D.; Nam, C.; Lebrecht, S.; Wittlif, C.; Schoenick, C.; Michel, O.; Krishna, R.; Weihs, L.; Smith, N. A.; Hajishirzi, H.; Girshick, R.; Farhadi, A.; and Kembhavi, A. 2024. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. *arXiv:2409.17146*.
- Derner, E.; de la Fuente, S. S.; Gutiérrez, Y.; Moreda, P.; and Oliver, N. 2024. Leveraging Large Language Models to Measure Gender Bias in Gendered Languages. *arXiv preprint arXiv:2406.13677*.
- Dion, K.; Berscheid, E.; and Walster, E. 1972. What is beautiful is good. *Journal of Personality and Social Psychology*, 24(3): 285–290.
- Dong, X.; Zhang, P.; Zang, Y.; Cao, Y.; Wang, B.; Ouyang, L.; Wei, X.; Zhang, S.; Duan, H.; Cao, M.; et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.
- Draws, T.; La Barbera, D.; Soprano, M.; Roitero, K.; Cecolin, D.; Checco, A.; and Mizzaro, S. 2022. The Effects of Crowd Worker Biases in Fact-Checking Tasks. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 2114–2124. ACM.
- Ebner, N. C.; Riediger, M.; and Lindenberger, U. 2010. FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42(1): 351–362.
- Echterhoff, J. M.; Liu, Y.; Alessa, A.; McAuley, J.; and He, Z. 2024. Cognitive Bias in Decision-Making with LLMs. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 12640–12653. Miami, Florida, USA: Association for Computational Linguistics.
- Elangovan, A.; Liu, L.; Xu, L.; Bodapati, S. B.; and Roth, D. 2024. ConSiDERS-The-Human Evaluation Framework: Rethinking Human Evaluation for Generative Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1137–1160. Bangkok, Thailand: Association for Computational Linguistics.
- Filandrianos, G.; Dimitriou, A.; Lymperaïou, M.; Thomas, K.; and Stamou, G. 2025. Bias Beware: The Impact of Cognitive Biases on LLM-Driven Product Recommendations. *arXiv preprint arXiv:2502.01349*.
- Fraser, K.; and Kiritchenko, S. 2024. Examining Gender and Racial Bias in Large Vision–Language Models Using a Novel Dataset of Parallel Images. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 690–713. St. Julian’s, Malta: Association for Computational Linguistics.
- Frieze, I. H.; Olson, J. E.; and Russell, J. 1991. Attractiveness and Income for Men and Women in Management. *Journal of Applied Social Psychology*, 21(13): 1039–1057.
- Gillingham, P. 2016. Predictive risk modelling to prevent child maltreatment and other adverse outcomes for service users: Inside the ‘black box’ of machine learning. *The British Journal of Social Work*, 46(4): 1044–1058.
- Girrbach, L.; Huang, Y.; Alaniz, S.; Darrell, T.; and Akata, Z. 2024. Revealing and Reducing Gender Biases in Vision and Language Assistants (VLAs). *arXiv preprint arXiv:2410.19314*.
- Golle, J.; Mast, F. W.; and Lobmaier, J. S. 2013. Something to smile about: The interrelationship between attractiveness and emotional expression. *Cognition and Emotion*, 28(2): 298–310.
- Guise, B. J.; Pollans, C. H.; and Turkat, I. D. 1982. Effects of Physical Attractiveness on Perception of Social Skill. *Perceptual and Motor Skills*, 54(3_suppl): 1039–1042.
- Gulati, A.; D’Inca, M.; Sebe, N.; Lepri, B.; and Oliver, N. 2025. Beauty and the Bias: Exploring the Impact of Attractiveness on Multimodal Large Language Models. *arXiv:2504.16104*.
- Gulati, A.; Martínez-García, M.; Fernández, D.; Lozano, M. A.; Lepri, B.; and Oliver, N. 2024. What is beautiful is still good: the attractiveness halo effect in the era of beauty filters. *Royal Society Open Science*, 11(11).
- Hada, R.; Husain, S.; Gumma, V.; Diddee, H.; Yadavalli, A.; Seth, A.; Kulkarni, N.; Gadiraju, U.; Vashistha, A.; Seshadri, V.; and Bali, K. 2024. Akal Badi ya Bias: An Exploratory Study of Gender Bias in Hindi Language Technology. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1926–1939. ACM.
- Haliburton, L.; Ghebremedhin, S.; Welsch, R.; Schmidt, A.; and Mayer, S. 2024. *Investigating Labeler Bias in Face Annotation for Machine Learning*, 145–161. IOS Press. ISBN 9781643685229.
- Hall, M.; van der Maaten, L.; Gustafson, L.; Jones, M.; and Adcock, A. 2022. A Systematic Study of Bias Amplification. *arXiv:2201.11706*.
- Hall, S. M.; Abrantes, F. G.; Zhu, H.; Sodunke, G.; Shtedritski, A.; and Kirk, H. R. 2023. VisoGender: A dataset for benchmarking gender bias in image-text pronoun resolution. *arXiv:2306.12424*.

- Hamermesh, D. S.; and Biddle, J. 1994. Beauty and the labor market. *The American Economic Review*, 84(5): 1174–1194.
- Hamidieh, K.; Zhang, H.; Gerych, W.; Hartvigsen, T.; and Ghassemi, M. 2024. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 547–561.
- Hosoda, M.; Stone-Romero, E. F.; and Coats, G. 2003. The effects of physical attractiveness on job-related outcomes: A meta-analysis of experimental studies. *Personnel Psychology*, 56(2): 431–462.
- Howard, P.; Fraser, K. C.; Bhiwandiwala, A.; and Kiritchenko, S. 2024. Uncovering Bias in Large Vision-Language Models at Scale with Counterfactuals. arXiv:2405.20152.
- Hu, W.; Xu, Y.; Li, Y.; Li, W.; Chen, Z.; and Tu, Z. 2024. BLIVA: A Simple Multimodal LLM for Better Handling of Text-Rich Visual Questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3): 2256–2264.
- Huang, P.-S.; Zhang, H.; Jiang, R.; Stanforth, R.; Welbl, J.; Rae, J.; Maini, V.; Yogatama, D.; and Kohli, P. 2020. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 65–83.
- Itzhak, I.; Stanovsky, G.; Rosenfeld, N.; and Belinkov, Y. 2024. Instructed to Bias: Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias. *Transactions of the Association for Computational Linguistics*, 12: 771–785.
- Janghorbani, S.; and De Melo, G. 2023. Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision language models. *arXiv preprint arXiv:2303.12734*.
- Jiang, Y.; Li, Z.; Shen, X.; Liu, Y.; Backes, M.; and Zhang, Y. 2024. ModSCAN: Measuring Stereotypical Bias in Large Vision-Language Models from Vision and Language Modalities. *arXiv preprint arXiv:2410.06967*.
- Kahneman, D.; and Tversky, A. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2): 263.
- Kanazawa, S. 2011. Intelligence and physical attractiveness. *Intelligence*, 39(1): 7–14.
- Kirk, H. R.; Jun, Y.; Volpin, F.; Iqbal, H.; Benussi, E.; Dreyer, F.; Shtedritski, A.; and Asano, Y. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems*, 34: 2611–2624.
- Koo, R.; Lee, M.; Raheja, V.; Park, J. I.; Kim, Z. M.; and Kang, D. 2024. Benchmarking Cognitive Biases in Large Language Models as Evaluators. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 517–545. Bangkok, Thailand: Association for Computational Linguistics.
- Kotek, H.; Dockum, R.; and Sun, D. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, 12–24.
- Kunst, J. R.; Kirkøen, J.; and Mohamdain, O. 2023. Hacking attractiveness biases in hiring? The role of beautifying photo-filters. *Management Decision*, 61(4): 924–943.
- Lee, M. H.; Montgomery, J. M.; and Lai, C. K. 2024. Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1321–1340. ACM.
- Leong, K.; and Sung, A. 2024. Gender stereotypes in artificial intelligence within the accounting profession using large language models. *Humanities and Social Sciences Communications*, 11(1).
- Ling, L.; Rabbi, F.; Wang, S.; and Yang, J. 2025. Bias Unveiled: Investigating Social Bias in LLM-Generated Code. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26): 27491–27499.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning.
- Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, 216–233. Springer.
- Lu, Z.; and Yin, M. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.
- Ma, D. S.; Correll, J.; and Wittenbrink, B. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47(4): 1122–1135.
- Ma, W.; Chiang, B.; Wu, T.; Wang, L.; and Vosoughi, S. 2023. Intersectional Stereotypes in Large Language Models: Dataset and Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8589–8597. Association for Computational Linguistics.
- Mathes, E. W.; and Kahn, A. 1975. Physical Attractiveness, Happiness, Neuroticism, and Self-Esteem. *The Journal of Psychology*, 90(1): 27–30.
- Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021a. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.
- Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021b. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6).
- Miller, A. G. 1970. Role of physical attractiveness in impression formation. *Psychonomic Science*, 19(4): 241–243.
- Nadeem, M.; Bethke, A.; and Reddy, S. 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)*, 5356–5371.
- Naik, R.; and Nushi, B. 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, 786–808. ACM.
- Narayanan Venkit, P.; Gautam, S.; Panchanadikar, R.; Huang, T.-H.; and Wilson, S. 2023. Unmasking Nationality Bias: A Study of Human Perception of Nationalities in AI-Generated Articles. In *Proceedings of the 2023 AAAI/ACM*

- Conference on AI, Ethics, and Society*, AIES '23, 554–565. ACM.
- Pan, X.; Dong, L.; Huang, S.; Peng, Z.; Chen, W.; and Wei, F. 2023. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*.
- Parikh, R. B.; Teeple, S.; and Navathe, A. S. 2019. Addressing bias in artificial intelligence in health care. *Jama*, 322(24): 2377–2378.
- Pessach, D.; and Shmueli, E. 2022. A Review on Fairness in Machine Learning. *ACM Comput. Surv.*, 55(3).
- Peterson, J. C.; Uddenberg, S.; Griffiths, T. L.; Todorov, A.; and Suchow, J. W. 2022. Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, 119(17).
- Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 469–481.
- Ramachandranpillai, R.; Sampath, K.; Mohammad, A.; and Alikhani, M. 2024. Fairness at Every Intersection: Uncovering and Mitigating Intersectional Biases in Multimodal Clinical Predictions. *arXiv:2412.00606*.
- Riccio, P.; and Oliver, N. 2022. Racial Bias in the Beautyverse: Evaluation of Augmented-Reality Beauty Filters. In *European Conference on Computer Vision*, 714–721. Springer.
- Riccio, P.; Psomas, B.; Galati, F.; Escolano, F.; Hofmann, T.; and Oliver, N. 2022. OpenFilter: a framework to democratize research access to social media AR filters. *Advances in Neural Information Processing Systems*, 35: 12491–12503.
- Ritts, V.; Patterson, M. L.; and Tubbs, M. E. 1992. Expectations, Impressions, and Judgments of Physically Attractive Students: A Review. *Review of Educational Research*, 62(4): 413–426.
- Rupert Jr, G.; et al. 2012. Simultaneous statistical inference.
- Shi, L.; Ma, C.; Liang, W.; Ma, W.; and Vosoughi, S. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*.
- Talamas, S. N. 2016. *Perceptions of intelligence and the attractiveness halo*. Ph.D. thesis, University of St Andrews.
- Talbot, A. N.; and Fuller, E. 2023. Challenging the appearance of machine intelligence: Cognitive bias in LLMs and Best Practices for Adoption. *arXiv:2304.01358*.
- Team, G. 2025. Gemma 3.
- Todorov, A. 2008. Evaluating Faces on Trustworthiness: An Extension of Systems for Recognition of Emotions Signaling Approach/Avoidance Behaviors. *Annals of the New York Academy of Sciences*, 1124(1): 208–224.
- Todorov, A.; and Duchaine, B. 2008. Reading trustworthiness in faces without recognizing faces. *Cognitive Neuropsychology*, 25(3): 395–410.
- Travaini, G. V.; Pacchioni, F.; Bellumore, S.; Bosia, M.; and De Micco, F. 2022. Machine learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction. *International journal of environmental research and public health*, 19(17): 10594.
- Venkit, P. N.; Srinath, M.; and Wilson, S. 2022. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, 1324–1332.
- Fig, J.; Gehrmann, S.; Belinkov, Y.; Qian, S.; Nevo, D.; Sakennis, S.; Huang, J.; Singer, Y.; and Shieber, S. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.
- Vijay, S.; Priyanshu, A.; and KhudaBukhsh, A. R. 2025. When Neutral Summaries Are Not That Neutral: Quantifying Political Neutrality in LLM-Generated News Summaries (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(28): 29514–29516.
- Walker, J.; Tepper, S. J.; and Gilovich, T. 2021. People are more tolerant of inequality when it is expressed in terms of individuals rather than groups at the top. *Proceedings of the National Academy of Sciences*, 118(43).
- Wan, Y.; and Chang, K.-W. 2024. White Men Lead, Black Women Help: Uncovering Gender, Racial, and Intersectional Bias in Language Agency. *arXiv preprint arXiv:2404.10508*.
- Wang, A.; and Russakovsky, O. 2021. Directional Bias Amplification. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 10882–10893. PMLR.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024. Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.
- Wiley, D. L. 1995. Beauty and the beast: Physical appearance discrimination in American criminal trials. *St. Mary’s Law Journal*, 27.
- Wilson, J. P.; and Rule, N. O. 2015. Facial Trustworthiness Predicts Extreme Criminal-Sentencing Outcomes. *Psychological Science*, 26(8): 1325–1331.
- Wilson, K.; and Caliskan, A. 2024. Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7: 1578–1590.
- Wu, J.; Gan, W.; Chen, Z.; Wan, S.; and Philip, S. Y. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, 2247–2256. IEEE.
- Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; Xie, Z.; Wu, Y.; Hu, K.; Wang, J.; Sun, Y.; Li, Y.; Piao, Y.; Guan, K.; Liu, A.; Xie, X.; You, Y.; Dong, K.; Yu, X.; Zhang, H.; Zhao, L.; Wang, Y.; and Ruan, C. 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. *arXiv:2412.10302*.

- Xiao, Y.; Liu, A.; Cheng, Q.; Yin, Z.; Liang, S.; Li, J.; Shao, J.; Liu, X.; and Tao, D. 2024. GenderBias-VL: Benchmarking Gender Bias in Vision Language Models via Counterfactual Probing. *arXiv preprint arXiv:2407.00600*.
- Yeh, K.-C.; Chi, J.-A.; Lian, D.-C.; and Hsieh, S.-K. 2023. Evaluating interfaced llm bias. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, 292–299.
- Ying, Z.; Liu, A.; Liang, S.; Huang, L.; Guo, J.; Zhou, W.; Liu, X.; and Tao, D. 2024. SafeBench: A Safety Evaluation Framework for Multimodal Large Language Models. *arXiv preprint arXiv:2410.18927*.
- Zhang, D.; Yu, Y.; Dong, J.; Li, C.; Su, D.; Chu, C.; and Yu, D. 2024a. Mm-llms: Recent advances in multimodal large language models. *ACL*.
- Zhang, J.; Wang, S.; Cao, X.; Yuan, Z.; Shan, S.; Chen, X.; and Gao, W. 2024b. VLBiasBench: A Comprehensive Benchmark for Evaluating Bias in Large Vision-Language Model. *arXiv preprint arXiv:2406.14194*.
- Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics.
- Zhao, Y.; Wang, B.; and Wang, Y. 2025. Explicit vs. Implicit: Investigating Social Bias in Large Language Models through Self-Reflection. *arXiv:2501.02295*.
- Zhou, K.; Lai, E.; and Jiang, J. 2022. VLStereoSet: A Study of Stereotypical Bias in Pre-trained Vision-Language Models. In He, Y.; Ji, H.; Li, S.; Liu, Y.; and Chang, C.-H., eds., *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 527–538. Online only: Association for Computational Linguistics.