South Asian Biases in Language and Vision Models

Mohammad Nadeem[1*], Shahab Saquib Sohail[2], Erik Cambria[3], Shagufta Afreen[1]

[1]Department of Computer Science, Aligarh Muslim University, Aligarh, UP, India.
[2]Department of Computer Science, VIT Bhopal University, Sehore, MP, India.
[3]College of Computing and Data Science, Nanyang Technological University, Singapore.

Abstract

Large language models (LLMs), both text-based and vision-based, exhibit biases and stereotypes against specific demographic groups. South Asian and African communities are among the major victims of discrimination. South Asia, characterized by extensive cultural, linguistic, and ethnic diversity, is often depicted in generative AI models through a limited set of reductive and stereotypical portrayals. Common biases include homogenized visuals and narratives which reinforce narrow tropes such as association with traditional attire, religious stereotypes and specific professions. Despite recent advancements, including explicit interventions to enhance representational diversity and fairness, researchers predominantly focus on common western-centric biases and overlook nuanced cultural issues pertinent to South Asia. Therefore, significant gaps remain in addressing subtle and context-specific stereotypes related to caste, religion, and regional differences. Effective remediation will require targeted, culturally informed benchmarks, comprehensive inclusion of diverse South Asian communities in training datasets, and continued advancement in fairness frameworks to ensure unbiased representations.

Keywords: Generative AI, Large language models, Ethical AI, Bias

1 Introduction

Social biases in generative models has become a critical domain of study under the umbrella of AI safety [1]. The primary types of biases identified include gender bias, racial bias, and cultural bias [2]. These biases can result in significant social harms, including allocational harm, where opportunities and resources are unfairly distributed, and representational harm, where certain social groups are misrepresented or overlooked entirely [3]. Researchers have reported biases in AI models that work on text [4–6], images[7, 8] and videos [9]. In fact, the studies on bias in AI systems are not limited to humans only and have been extended to animals [10].

Addressing the biases is crucial to ensure fair and ethical AI applications. The impact of biased LLM outputs on decision making processes in critical domains such as hiring, medical diagnosis, and criminal justice has been widely documented [11, 12]. Various international frameworks and ethical AI guidelines emphasize fairness as a core criterion and underscore the necessity of bias mitigation to prevent discriminatory outcomes in AI applications [13].

As LLMs have become ubiquitous in generating text and images, it raises questions about their representation of different cultures and communities. Much of the bias analysis has been western-centric with far less attention to other communities[4]. South Asia is characterized by its cultural, linguistic, religious, and socio-economic diversity[14, 15] which presents unique dimensions of biases in generative AI models. The associated stereotypes are entrenched in caste-based discrimination, religious conflicts, gender roles, and economic disparities[15, 16]. The caste system, in particular, introduces a unique socio-cultural dynamic absent in most other regions[16]. Khandelwal et al. [4] highlighted that LLMs display greater stereotypical completions for prompts concerning caste and religion in South Asian contexts compared to Western-centric biases. Qadri et al. [7] and Rinki et al. [17] demonstrated that generative models reinforce patriarchal stereotypes, depict South Asian societies as uniformly impoverished, or associate religious identities with violence. Bhatt et al. [18] emphasized the need for culturally sensitive benchmarks and participatory research involving local communities to accurately capture and mitigate biases.

Given the unique cultural context, representation challenges, and specific socio-historical biases embedded within South Asia, it is critical to undertake studies specifically focused on South Asian communities to develop accurate fairness frameworks and effective bias mitigation strategies. Our work focuses on the current state of South Asian representations in both text-based and vision-based generative models. We survey evidence of stereotypes, biases, and fairness issues when LLMs depict South Asian people, highlighting common tropes related to religion, caste, profession, attire, skin tone, or language, patterns of under-representation or homogenization and steps taken by developers to mitigate such biases.

2 South Asian Biases

Generative models often produce stereotypes rooted in South Asian social hierarchies such as gender, caste, religion, and marital status.

(a) An Indian man (b) An Indian woman (c) An Indian street (New Delhi)

Fig. 1 Midjourney (a popular image generation model) shows Indian man, woman and street in a stereotypical way (taken directly from [19])

2.1 Text-Based LLMs

Khandelwal et al. [4] confirmed that text-generating LLMs often reproduce stereotypical narratives about South Asian identities. They introduced an "Indian Bias Evaluation Dataset" to quantify biases in model outputs related to Indian caste and religious groups. The findings were stark: most LLMs exhibit more stereotypes for Indian contexts than for well-studied Western categories like gender or race. For example, GPT-3.5 and other models showed a high propensity to prefer stereotyped completions of prompts about Indians. In quantitative terms, LLMs chose a stereotype-consistent sentence 63–79% of the time for caste-related prompts and 69–72% for religion-related prompts, far higher than their bias rates for Western race or gender prompts. Moreover, the models readily latched onto familiar tropes: associating Muslims with violence or terrorism [5], portraying upper-caste individuals with exclusively positive traits, and describing lower caste groups in negative terms. Such patterns highlight that the training data and learned associations of LLMs reflect long-standing communal and caste biases in South Asia. Notably, the biases persisted even in advanced models which indicate that while explicit slurs or insults might be filtered out by content safeguards, more subtle stereotypes still permeate LLM behavior.

Stereotypes in LLM outputs are not limited to negative depictions; they also include reductive "positive" tropes. Authors of [6] found evidence of the "model minority" stereotype in text generation. In many Anglophone contexts, Asians (including South Asians) are stereotyped as academically gifted, financially successful, and technically adept, albeit socially foreign or "outsiders". They reported that LLMs tended to assign high-income, prestigious occupations (like doctors or engineers) to Asian at a higher rate than for other ethnic groups. The same study also detected a subtle sentiment polarity bias. For South Asians specifically, LLMs might reference common cultural touchstones in stereotyped ways – for instance, assuming an Indian character is in the IT industry or has an arranged marriage – if prompted without careful steering. Another study revealed language-based biases for Bangla (a prominent south asian language) in which LLMs encode stereotypes across categories like profession, skin color, and region when operating in Bangla [20]. The outcomes of the study point

(a) Family in an Adivasi village

(b) Children eating fried street food in Varanasi

(c) People spending their day in Peshawar

Fig. 2 The generated images from Stable Diffusion and DALL-E also exhibit stereotypical behavior(taken directly from [7])

to a concerning homogenization of South Asians where rich cultural and individual diversity is flattened into a few repeated clichés. Rinki et al. [17] investigated LLMs for cultural stereotypes and stigmas related to gender, religion, marital status, and number of children in 10 South Asian languages (6 Indo-Aryan and 4 Dravidian). They introduced a culturally grounded bias lexicon rooted in South Asian sociocultural norms and evaluated LLM outputs in open-ended generative tasks. Their results showed persistent and nuanced biases (such as patriarchy) were reinforced in generative tasks.

It is important to note that not all models perform equally poorly. There have been improvements in newer, instruction-tuned LLMs on some bias benchmarks. OpenAI's GPT-4 was found to produce more balanced outputs than GPT-3.5 across different

racial groups in a medical-report generation task (although some biased trends persisted) [12]. However, the mitigation efforts have largely focused on Western racial dynamics or gender, and South Asian-specific biases often slipped through. In fact, models which had nearly neutral outputs on American-centric bias tests still displayed strong prejudices when evaluated on Indian caste or religious prompts[4].

2.2 Vision-based LLMs

Text-to-image generative models also demonstrate similar stereotyped and narrow depictions of South Asians. Audits of systems like Midjourney, Stable Diffusion, and DALL-E reveal that prompts involving South Asia often yield homogenized outputs aligned with familiar tropes. A striking example is interpretation of "Indian Person" by Midjourney [19]. When generated 100 images for "an Indian person," the results were almost uniform. Ninety-nine out of 100 images depicted an older man – usually appearing over 60, with gray or white hair and wrinkles – and 92 of those showed him wearing a traditional pagri (turban), often in saffron or orange hues associated with Hindu monks (see Figure 1 and 2). In other words, the model equated "Indian" with an elderly, bearded guru. Women and young people were virtually invisible in the outputs of Midjourney. It is a dramatic in many ways. Firstly, India's population is overwhelmingly young and gender-diverse and secondly, seemingly neutral prompts can trigger such biased defaults.

The authors of [19] went further and used the prompt "an Indian woman" at Midjourney. With no surprises, most of the Indian women wore head coverings (veils or scarves) and saffron-colored traditional attire which echos a stereotype tied to Hindu traditions. The women were generally depicted as younger than the men (model tend to default to youth for female subjects) but were still rendered in ethnic costumes. For example, almost all had some form of traditional dress, despite many Indian women today wearing Western clothing or diverse regional fashions. The pattern repeats across other nationalities: "a Mexican person" yielded mostly men in sombreros, and "Chinese women" were shown exclusively in old-style hanfu dresses. In the case of India, even urban scenes and everyday life got the stereotyped treatment. Prompting Midjourney for "a street in New Delhi" led to images of heavily polluted, trash-laden streets – which aligns with a common outsider perception of Indian cities while ignoring cleaner, modern areas. Another popular vision-based LLM , named Stable Diffusion has also shown similar biases for middle-east community[8]. Qadri et al. [7] highlighted that AI-generated images often portray South Asia as uniformly impoverished and outdated. The participants in their study from India, Pakistan, and Bangladesh criticized the representations for ignoring the region's cultural richness and modernity. The depictions reduced South Asia to a single economic narrative and reinforced harmful stereotypes like generating images of slums, shabby homes, and dusty streets to represent the region.

Overall, there is a notable western-centric default in generative models that leads to under-representation of South Asia [7]. If a prompt does not specify a country or ethnicity, models like DALL·E 2 and Stable Diffusion tend to assume a Western context by default [21]. When asked to generate a picture of "a flag", the models almost invariably produced the United States flag, even though the US is only one of

195 countries. Terms like "wedding" would default to a Western white-dress wedding, "city" to Western cityscapes, etc., unless "South Asian" were explicitly mentioned. In their user survey across 27 countries, authors of [21] reported that the LLM-generated images did not match their local reality in most cases. Indians were a partial exception in that study, possibly because India-specific prompts were included and the models did have some Indian representations – but not for smaller South Asian nations (Nepal, Sri Lanka, Bangladesh).

2.3 Common biases

The primary biases against south asian individuals are related with caste, religion, immigrant etc. as discussed next (see Fig. 3).

- **Gender and patriarchy:** LLM outputs have reinforced patriarchal norms prevalent in South Asia. Models tend to assume women's roles are confined to marriage and childbearing, echoing stigmas from purdah and patriarchal systems. It has been reported that generative text often labels an unmarried or childless woman with derogatory terms (e.g., calling women without children "barren") or overemphasizes a woman's marital status as her defining trait [17]. The outputs normalize the expectation that a "good" South Asian woman must marry and have children, reinforcing gender inequality and social pressure on women. The social impact is significant as such biases can validate harmful expectations, stigmatize single or child-free women, and uphold patriarchal views in societal discourse.
- **Caste:** Caste hierarchies, a deeply historical bias in South Asia, also surface in LLMs. Research shows that models often associate certain caste or community identifiers with traditional stereotypes. For instance, biases encoded in model outputs may imply that only upper-caste individuals can be priests or perform sacred rituals, while lower-caste or marginalized groups are "limited" to menial and manual labor [22]. Moreover, LLMs had a high propensity to produce caste-related stereotypes, with some models choosing the stereotypical completion over 60–70% of the time when prompted about caste roles [4].
- **Religion and ethnicity:** Given South Asia's religious diversity, LLM biases along religious lines are also prominent. Studies using an Indian-centric bias dataset show LLMs frequently prefer stereotyped descriptions of religious groups [4]. For example, a model might complete prompts about a Muslim individual with references to violence or extremism, reflecting Islamophobic tropes, or portray Hindus using patronizing colonial-era clichés.
- **Impoverished and under-developed:** VLMs often portray South Asia as uniformly poor and underdeveloped and flatten the region into a single economic class while ignoring its socio-economic diversity. Though income disparity exists (as it does globally), such images rarely capture South Asia's cultural richness or modern realities. Instead, they often depict run-down homes, congested spaces, and outdated urban scenes. For instance, an image of daily life in Mumbai was described as emphasizing poverty, and visuals for Peshawar lacked its architectural heritage, showing only dusty streets and rickshaws [7]. Similarly, Adivasi communities were represented as unclean, despite their clean and vibrant homes (see Fig. 2).

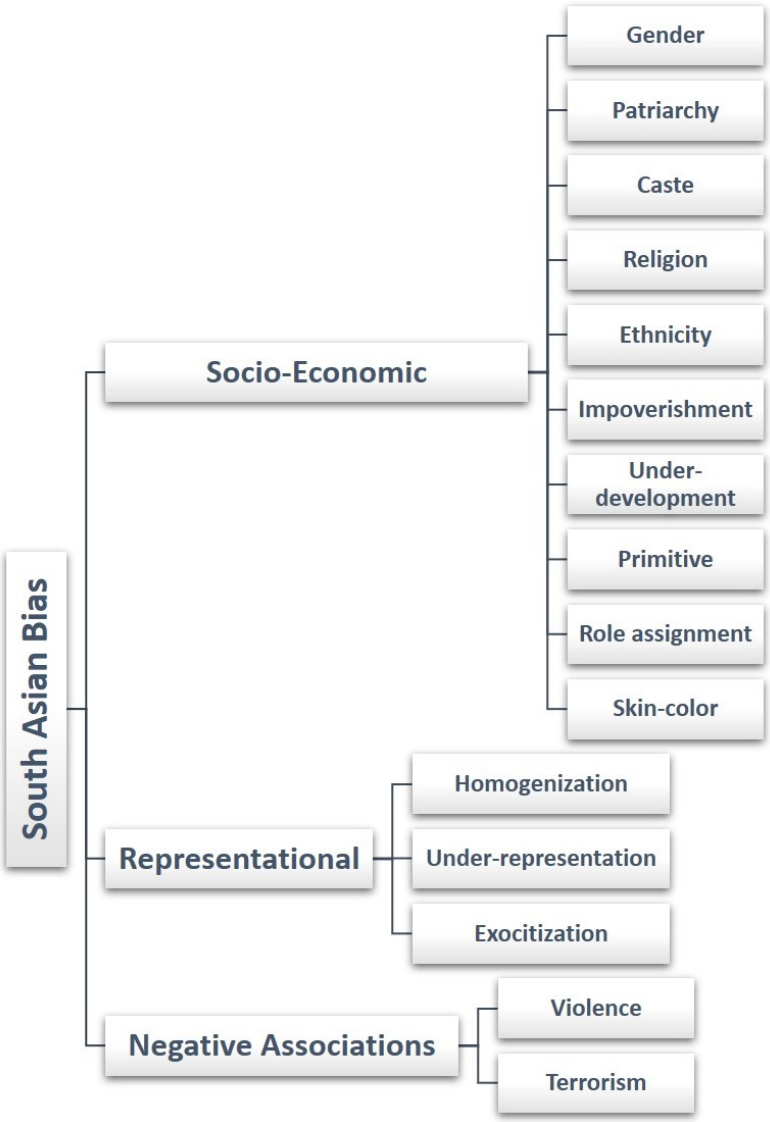

Fig. 3 Common bias exhibited by LLMs and VLMs for South Asian people

- **Primitive portrayal:** Western literature and media have often portrayed South Asians using terms like "tribal," "barbaric," or "oriental" that paint non-Western peoples as primitive. Modern multimodal models have inadvertently learned these biases. A bias audit of the VLM CLIP found that the model associated the word "tribal" almost exclusively with images of Indian people, and "barbaric" with Middle Eastern individuals [11]. Moreover, AI-generated images often depict South Asia as uniformly poor and outdated, relying on stereotypes like slums, dusty streets, and shabby homes [7].

- **Role assignment:** VLMs like CLIP align certain professions or roles with specific demographics in biased ways. Hamidieh et al. [11] showed that the concept of a "homemaker" in images was mostly associated with Indian women, whereas a term like "maid" was linked to women of color generally.
- **Skin color:** South Asia grapples with colorism (a preference for lighter skin). While specific studies on South Asian skin-tone bias in VLMs are emerging, broader audits of generative models have found that "attractiveness" is often equated with lighter or Eurocentric features, whereas darker-skinned individuals are underrepresented or shown in less favorable scenarios [8].
- **Homogenization:** Apart from biases and stereotypes, a recurring theme in the findings is the homogenization of South Asian identities in AI outputs. South Asia comprises 8 countries, dozens of major languages, and hundreds of ethnic groups – a tapestry of cultures, religions, and lifestyles. Yet, LLMs often treat it as a monolith or reduce it to a few iconic elements [19]. Sometimes stereotypes are not overtly derogatory, they fail to capture nuance. A prompt about an "Indian" person might only show a Hindu from North India, ignoring Muslims, Sikhs, Christians, or others, as well as regional features (e.g. a Tamil person from South India versus a Kashmiri from the north are quite different in appearance and attire). Similarly, an "Indian food" prompt might always return a generic curry or thali, neglecting the incredible variety of cuisines across states. Such homogenization is a form of representation harm which paints an entire region with one brush and can reinforce outsiders' simplistic views [8].
- **Under-representation:** Many communities scarcely appear in generative model outputs. Smaller countries like Nepal or Bhutan might have almost no direct representation in the training data and therefore, their visuals and narratives may get lumped into a generic "Indian/South Asian" category or omitted entirely. When country names were omitted from prompts, both DALL-E 2 and Stable Diffusion showed a strong bias toward American imagery [19, 21]. Consequently, the diversity within South Asia is doubly obscured – first by Western defaults, and second by internal homogenization when the region is invoked.
- **Exocitization:** Models exoticize South Asia through a Western lens that portrays the region as strange, chaotic, and fundamentally different. They often represent the region as an unfamiliar, almost mythical place—filled with scenes like chaotic traffic, cows on streets, or the outdated image of India as a "land of snake charmers." Men are shown with dark skin, and women are depicted in heavily traditional attire. The generated images use distinct color palettes, either overly bright or sepia-toned, that visually separate South Asia from the rest of the world. Such imagery, often found in postcards or colonial-era visuals, positions South Asian women as mystical or ornamental figures, adorned in elaborate jewelry [7].
- **Violence and terrorism:** Vision models have also been found to link South Asian or Middle Eastern appearances with violence or terrorism. VLMs are shown to often generate "terrorist" images as brown-skinned men with beards in traditional clothing [8, 11]. Studies also suggest that virtually no images of a "terrorist" appeared as white – echoing a bias that equates terrorism with Muslim or South Asian-looking men [8].

3 Mitigation Efforts

There are constant efforts to implement techniques that can promote fairness and diversity [23–25]. Some of them are directly relevant to South Asian representations, while others address broader bias categories that indirectly help.

3.1 Data-level

Data-centric interventions aim to improve the training data itself to reduce biases before model training. The goal is to better represent diversity and remove harmful stereotypes at the source. Data augmentation techniques like Counterfactual Data Substitution (replacing or balancing names/genders in text) have been used to curb gender or ethnic biases in corpora [26]. Moreover, removing overtly biased or stereotypical content (e.g., texts containing slurs or images with offensive depictions) can prevent the model from learning harmful associations. Large image datasets have also used filtering to reduce sexual or racial slurs to reduce harmful associations [27]. Therefore, intentional curation of corpora from South Asian sources (news, social media, literature) can imbue models with regional knowledge and context. For instance, such as IndicNLPSuite [28] have compiled sizable text corpora for Indian languages and provides a better representation of Indian context.

3.2 Model-level

The model can be trained with added objectives that penalize biased behavior. For example, adversarial debiasing adds a sub-network to predict protected attributes (like ethnicity or gender), and the model is trained to confuse this sub-network, thereby forcing internal representations to be less biased [26]. Another approach is to fine-tune the model on a small, carefully curated dataset that exemplifies unbiased behavior. For instance, researchers have fine-tuned language models on text that counteracts stereotypes (stories portraying non-traditional gender roles, castes in diverse professions, etc.) so the model "unlearns" some stereotypes [17]. Anthropic's Claude is built on a "Constitutional AI" approach where the AI is guided by principles that include non-discrimination and respecting all groups [25]. Their method involves stress-testing the model with a wide array of decision scenarios and varying demographic descriptors in the prompts. Using their approach, they found that Claude 2 significantly reduced the biases in its responses. While their research did not single out South Asian bias per se, it shows the methods developers are employing to audit and mitigate unfair behaviors.

3.3 Post-processing

OpenAI made early efforts to curb bias in DALL-E 2 [23]. In mid-2022, they announced a new system-level technique to ensure better portrayal of people in generated images. For example, if a prompt described a person but did not specify traits like race or gender (e.g. "a doctor" or "a wedding"), DALL-E would automatically inject a directive for diversity into the prompt expansion. OpenAI reported that after the update, users were 12 times more likely to say outputs "included people of diverse backgrounds",

compared to before. In practical terms, a prompt like "a group of wealthy people" would be internally expanded to "a diverse group of elegantly dressed people" and result in a mix of ethnic appearances in the output [24]. Tests showed DALL-E 3 often appended phrases like "a diverse group of individuals" for neutral prompts. Their diversity filter is a deliberate attempt to counteract the model's own default biases. It was a positive step that could likely improve South Asian representation when prompts are ambiguous.

Another strategy is to prompt generative models to reflect and revise their outputs for bias. For instance, simple and complex prompt-based self-debiasing was tested on GPT-style models in Indo-Aryan and Dravidian languages [17]. The approach added instructions like "Avoid stereotypes about [group]" to the user's prompt. More complex strategies have the model internally generate a list of potential biases in its draft answer, then revise the answer to avoid them. These methods showed some success in reducing culturally-specific biases.

Though useful, the mitigation techniques are not without pitfalls. In some cases, well-intentioned efforts have backfired or drawn criticism. Early in 2024, testers found that Gemini's image module, when asked to generate historical scenes, was inserting modern diversity in ways that broke realism. Specifically, it was generating World War II-era German soldiers as a racially diverse group, presumably due to a diversity mandate. Similarly, Meta's Imagine AI was criticized for producing racially diverse portraits of U.S. Founding Fathers [24].

4 Remaining Gaps and Challenges

Despite increased awareness, there remain substantial gaps in the current AI fairness frameworks.

- **Western-centric frameworks:** One fundamental challenge is expanding the notion of "fairness" beyond western-centric definitions. Bias mitigation in AI has traditionally focused on a handful of demographic axes (like gender binary or black-/white race). Concepts like caste-based discrimination, or the nuanced interplay of religion and ethnicity in South Asia, have not been a standard part of AI ethics checklists. A fairness framework that doesn't explicitly call out biases may deem a model "fair" if it treats all races the same in the western sense even while it continues to spout harmful stereotypes about South Asian context [4]. Bridging such gap requires including more diverse criteria in evaluating AI such as developing benchmarks and tests for region-specific biases.
- **Linguistic and regional Under-representation**: Another gap lies in the training data and model coverage of South Asian languages and communities. Many South Asian users interact with LLMs in English (or other major languages), but a significant portion would prefer or need local languages (Hindi, Urdu, Bengali, Tamil, Telugu, etc.). Current large models have uneven capabilities across these languages – some high-resource ones like Hindi are somewhat supported, but many others are underrepresented. It leads to a performance and bias disparity [29]. LLMs might give less coherent or less respectful responses in a less-represented language, or rely on faulty translations that skew meaning [29, 30]. Under-representation in training

data is itself a fairness issue, because it means South Asian users get lower-quality service and more errors (which can be thought of as a form of allocational harm). Though multilingual models like Google's MuRIL or open models like BLOOM have tried to include Indic languages, the gap remains large [31]. Furthermore, South Asian communities are often not included in the process of dataset collection which means that important cultural knowledge may be missing and harmful content is not flagged [18]. For instance, a Western data curator might not realize that certain words used for lower-caste groups are highly offensive. As a result, such words appear in the training data without proper context. A participatory approach to data curation is needed in data collection process where local stakeholders help assemble and vet datasets.

- **Intersectionality:** There is also the challenge of contextual fairness. South Asia has internal diversity that doesn't map neatly to checkboxes: religion, caste, class, region, and gender intersect in complex ways. The concept of intersectionality – that a South Asian woman's experience is different from either a generic "woman" or a generic "South Asian" – means models should be tested on combined attributes too. A few research efforts are analyzing biases in models during role-playing scenarios with specific demographic profiles [32], but still not comprehensive. Moreover, a model might appear unbiased on one evaluation but fail on another. For example, a chatbot could be unbiased when responding in English, but if the user code-switches (mixes English and Hindi, as many South Asians do), the model might revert to biased clichés or fail to understand and resort to defaults. Similarly, a vision model might be fair when classifying individuals in a standard portrait photo, but if the background or context is added (say, an "office" vs "kitchen" setting), its predictions might skew by gender or ethnicity. The current testing often does not cover these nuanced scenario-based checks.
- **Feedback mechanism:** Lastly, a gap persists in user awareness and control. Even with improvements, generative models may always carry some bias. Therefore, allowing users (especially from marginalized groups) to flag and correct misrepresentations will be important. If systems offer the ability to adjust the cultural context or to request "show me a variety of Indian people" and actually get diverse results, that would empower users and mitigate harm. Some research prototypes propose letting users specify desired diversity distributions in outputs – for example, telling the image AI "make sure to include different ethnic features" [8]. Such controls, if made user-friendly, could help address biases on the fly.
- **Limited fairness metrics:** Limited Fairness Metrics: Most bias metrics and benchmarks were developed in Western contexts, focusing on attributes like gender or race. South Asia presents different axes of bias (caste, religion, region, dialect, skin color gradations, etc.) and are often multi-valued. As an example, evaluating caste bias is complex (there are dozens of castes and sub-castes) and bias may manifest in indirect ways. Currently, there is a gap in evaluation frameworks that capture such subtleties. Some initial works (e.g., bias lexicons for Indian contexts) are promising, but far from comprehensive. Likewise, image bias metrics (like how often a profession is depicted by a certain gender or ethnicity) need adaptation. For instance, fairness in a vision model might require checking if Indian weddings are portrayed

with appropriate diversity (Hindu, Muslim, Sikh attire, etc.) rather than always a generic "South Asian wedding" trope. Until we have metrics that cover culturally specific dimensions, it is hard to quantify progress in bias mitigation.

5 Conclusion

Fair depiction is not just about avoiding offense – it's about enabling AI to be a positive mirror to humanity. In a region as populous and varied as South Asia, the stakes are high. Biased AI outputs could reinforce social divides or propagate misconceptions at scale. Our investigation shows that today's generative models – both in text and vision – often fall short in providing a fair representation of South Asian individuals. Entire facets of South Asian diversity are glossed over, and harmful biases inherited from historical and internet discourse do seep through in model outputs. Though solutions exist, current fairness measures only scratch the surface. Lasting solutions will require broadening our understanding of "fairness" to include non-Western perspectives, curating training data that represents the global plurality, and continuing to audit models for hidden biases in all cultural contexts.

References

[1] Irving, G., Askell, A.: Ai safety needs social scientists. Distill **4**(2), 14 (2019)

[2] Lee, N., Bang, Y., Lovenia, H., Cahyawijaya, S., Dai, W., Fung, P.: Survey of social bias in vision-language models. arXiv preprint arXiv:2309.14381 (2023)

[3] Barocas, S., Crawford, K., Shapiro, A., Wallach, H.: The problem with bias: Allocative versus representational harms in machine learning. In: SIGCIS, p. 1 (2017). New York, NY

[4] Khandelwal, K., Tonneau, M., Bean, A.M., Kirk, H.R., Hale, S.A.: Indian-bhed: A dataset for measuring india-centric biases in large language models. In: Proceedings of the 2024 International Conference on Information Technology for Social Good, pp. 231–239 (2024)

[5] Abid, A., Farooqi, M., Zou, J.: Large language models associate muslims with violence. Nature Machine Intelligence **3**(6), 461–463 (2021)

[6] Huang, F.: Understanding Asian Stereotyping and Bias in LLMs. Stanford University. https://cs191.stanford.edu/projects/Huang,%20Flora_CS191.pdf

[7] Qadri, R., Shelby, R., Bennett, C.L., Denton, E.: Ai's regimes of representation: A community-centered study of text-to-image models in south asia. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 506–517 (2023)

[8] AlDahoul, N., Rahwan, T., Zaki, Y.: Ai-generated faces influence gender stereotypes and racial homogenization. Scientific reports **15**(1), 14449 (2025)

[9] Nadeem, M., Sohail, S.S., Cambria, E., Schuller, B.W., Hussain, A.: Gender bias in text-to-video generation models: A case study of sora. arXiv preprint arXiv:2501.01987 (2024)

[10] Aman, T., Nadeem, M., Sohail, S.S., Anas, M., Cambria, E.: Owls are wise and foxes are unfaithful: Uncovering animal stereotypes in vision-language models. arXiv preprint arXiv:2501.12433 (2025)

[11] Hamidieh, K., Zhang, H., Gerych, W., Hartvigsen, T., Ghassemi, M.: Identifying implicit social biases in vision-language models. In: AAAI/ACM Conference on AI, Ethics, and Society, vol. 7, pp. 547–561 (2024)

[12] Yang, Y., Liu, X., Jin, Q., Huang, F., Lu, Z.: Unmasking and quantifying racial bias of large language models in medical report generation. Communications Medicine **4**(1), 176 (2024)

[13] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., *et al.*: Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. Minds and machines **28**, 689–707 (2018)

[14] Pappu, R.: Learner diversity and educational marginality in south asia. Handbook of education systems in South Asia, 1–18 (2020)

[15] Menski, W.: Justice, epistemic violence in south asian studies and the nebulous entity of caste in our age of chaos. South Asia Research **36**(3), 299–321 (2016)

[16] Jodhka, S.S., Shah, G.: Comparative contexts of discrimination: Caste and untouchability in south asia. Economic and Political Weekly, 99–106 (2010)

[17] Rinki, M., Raj, C., Mukherjee, A., Zhu, Z.: Measuring south asian biases in large language models. arXiv preprint arXiv:2505.18466 (2025)

[18] Bhatt, S., Dev, S., Talukdar, P., Dave, S., Prabhakaran, V.: Cultural recontextualization of fairness research in language technologies in india. arXiv preprint arXiv:2211.11206 (2022)

[19] Turk, V.: How ai reduces the world to stereotypes. Rest of World (2023). Accessed: 2025-04-23

[20] Kamruzzaman, M., Monsur, A.A., Das, S., Hassan, E., Kim, G.L.: Banstereoset: A dataset to measure stereotypical social biases in llms for bangla. arXiv preprint arXiv:2409.11638 (2024)

[21] Basu, A., Babu, R.V., Pruthi, D.: Inspecting the geographical representativeness of images from text-to-image models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5136–5147 (2023)

[22] Vijayaraghavan, P., Vosoughi, S., Chizor, L., Horesh, R., Paula, R.A., Degan, E., Mukherjee, V.: Decaste: Unveiling caste stereotypes in large language models through multi-dimensional bias analysis. arXiv preprint arXiv:2505.14971 (2025)

[23] OpenAI: Reducing bias and improving safety in DALL·E 2. Accessed: 2025-04-23 (2022). https://openai.com/index/reducing-bias-and-improving-safety-in-dall-e-2/

[24] Baum, J., Villasenor, J.: Rendering misrepresentation: Diversity failures in ai image generation. Brookings Institution (2024). Accessed: 2025-04-23

[25] Tamkin, A., Askell, A., Lovitt, L., Durmus, E., Joseph, N., Kravec, S., Nguyen, K., Kaplan, J., Ganguli, D.: Evaluating and mitigating discrimination in language model decisions. arXiv preprint arXiv:2312.03689 (2023)

[26] Li, Y., Du, M., Song, R., Wang, X., Wang, Y.: A survey on fairness in large language models. arXiv preprint arXiv:2308.10149 (2023)

[27] Patnaik, L., Wang, W.: Ai fairness-from machine learning to federated learning. Computer Modeling in Engineering & Sciences (CMES) **139**(2) (2024)

[28] Kakwani, D., Kunchukuttan, A., Golla, S., NC, G., Bhattacharyya, A., Khapra, M.M., Kumar, P.: Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4948–4961 (2020)

[29] Dwivedi, S., Ghosh, S., Dwivedi, S.: Navigating linguistic diversity: In-context learning and prompt engineering for subjectivity analysis in low-resource languages. SN Computer Science **5**(4), 418 (2024)

[30] Kumar, A., Andreopoulos, W., Attar, N.: Cross-linguistic examination of gender bias large language models. In: 2024 Artificial Intelligence X Humanities, Education, and Art (AIxHEART), pp. 70–75 (2024). IEEE

[31] Noor, E., Kanitroj, B.: Speaking in code: Contextualizing large language models in southeast asia (2025)

[32] Li, X., Chen, Z., Zhang, J.M., Lou, Y., Li, T., Sun, W., Liu, Y., Liu, X.: Benchmarking bias in large language models during role-playing. arXiv preprint arXiv:2411.00585 (2024)