Investigating Biases in the SDXL-DPO Text2Image Model

S M Ahasanul Karim, Student nr. VTD561
University of Copenhagen

Abstract

This paper explores collective gender, race and ethnicity-specific bias in the Diffusion Model Alignment Using Direct Preference Optimization Text2Image model. The paper adopts the methodologies of some old research papers while also addressing their limitations and trying to mitigate them. The results depict deep historical bias underlying while also pointing out some racial stereotypes in job roles. The model also shows specific bias towards non-binary persons while generating images and assigning job roles. This research suggests improving the model training further to mitigate these problems.

Keywords: Generative AI, Stable Diffusion, Text-To-Image, Bias in AI, Equality, Diversity, and Inclusion (EDI).

1 Introduction

Generative AI has taken the world by storm. It has introduced wonders since its inception and is a strong potential advocate for the future of AI. It is dominating a wide number of sectors like content creation, sales, marketing, graphics design, education, genomics research etc. According to predictions by Bloomberg Intelligence (Bloomberg 2023) The Generative AI market can grow at a CAGR of 42% and over the next 10 years, demand for generative AI products could add about $280 billion of new software revenue. According to J.P. Morgan (2023), generative AI has the potential to surpass human production capacity by 2030 because of its broad range of output forms generation capabilities, which include text, code, graphics, and videos.

While Generative AI has already been advent, diffusion-based text-to-image systems are even newer machine learning approaches to generate images with text prompts. Every day models such as DALL· E, Imagen, Make-a-Scene, and Stable Diffusion are becoming more and more popular producing realistic and diverse pictures in response to user inputs. (Rombach et al., 2022, Ramesh et al., 2022, Gafni et al., 2022, Saharia et al., 2022). Several of these models have made their way through to stock image generation and graphic design(Lomas 2022, Moreno 2022).

But greater inventions also come with greater potential risks. These risks include issues like intellectual property rights, accuracy of output, explainability of results, and potential propagation of harmful biases. Generative AI models are employed to generate new material based on patterns from the training data, in contrast to standard AI models that are frequently used for categorization or prediction. Because there is no one "correct" output, it is challenging to quantify bias in these models. Rather, a variety of created information would need to be examined for bias-reflected patterns. Furthermore, newly created information—such as visual content—produced by these models has the power to directly influence users' views, reinforce negative stereotypes, and even warp their beliefs, particularly when the content is widely shared. Also, these models are trained through a collection of images from the internet which can not be controlled. To address any bias, verifying and updating the training data becomes extremely difficult when there is no control over the sources. The training data may contain a wide range of viewpoints, cultural norms, and beliefs, making it more difficult to identify and even fix the many biases that could unintentionally enter the model (Zhou et al., 2023).

For instance, stable diffusion is a technique that generates images from random Gaussian noise. To do this, a picture from the training dataset is blended with noise progressively at discrete time steps, until finally the noise overpowers the image. The model learns to reverse noise diffusion one step at a time during training. During training, diffusion models use both text and visuals. The text directs the denoising process by enhancing stages with token embeddings from pre-trained models such as CLIP. CLIP's combined training of image and text encoders provides latent space similarity, which helps the diffusion model generate results that are comparable to the input information. Both can introduce and magnify social biases at various points of the model training and deployment pipeline. Their interactions are complex and poorly understood (Luccioni and Akiki, et al. 2023).

Hence, externally probing for biases in various factors through a stable diffusion model is necessary. This paper approaches such techniques to find biases in the Diffusion Model Alignment Using Direct Preference Optimization Text2Image model. It has been described as a method to align diffusion models to text human preferences by directly optimizing on human comparison data. (Wallace et al, 2023). Using 1600 photos generated by this model of different races and genders, an investigation was conducted to determine the underrepresentation of genders and coloured persons, as well as stereotypes about work responsibilities based on gender and ethnicity.

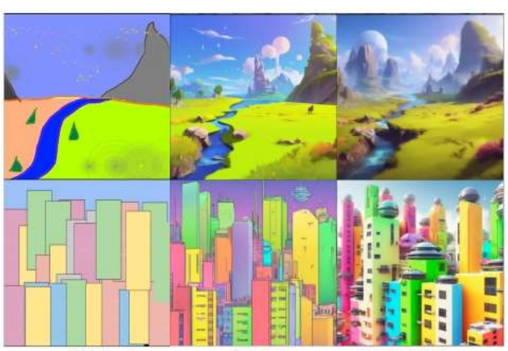


Fig 1: Diffusion-DPO generates more visually appealing images in the downstream image-to-image translation task. Comparisons of using SDEdit from color layouts. (Wallace et al, 2023).

2 Motivation & Literature Review

It is a matter of consolation that the bias in Generative AI has eventually become a popular discussion. There has been much research on the fairness and detection of bias in models, datasets etc. Fabbrizzi et al., have meticulously examined existing research on methods for uncovering and quantifying biases within datasets. It scrutinizes initiatives undertaken to create datasets that are attuned to biases, emphasizing the challenges associated with bias detection and measurement in the visual domain. Importantly, the research concludes that achieving a completely bias-free dataset is an elusive goal. Instead, researchers and practitioners should cultivate awareness regarding biases inherent in their datasets and transparently acknowledge them(Fabbrizzi et al., 2021).

While existing bias mitigation strategies often focus on gender parity, they have limitations. To address this, Simone et al. presented DiffusionWorldViewer, a tool enabling analysis and manipulation of these models' attitudes, values, and narratives that influence image generation. This categorizes the demographics of the generated images and offers interactive methods to align them with user worldviews (De Simone et al., 2023). Ghosh et al.'s study on 136 prompts with Stable Diffusion, using CLIP's cosine similarity, found that it tends to portray individuals as European/North American men, with a preference for European/North American men. This leads to the overemphasis of Australian/New Zealander identity over Papua New Guinean, erasing Indigenous Oceanic peoples. The study also revealed an unexpected pattern of over-sexualization of women from Latin America, Mexico, India, and Egypt, raising concerns about perpetuating Western fetishization and stereotypical representation(Ghosh et al., 2023). Fraser et al. used social psychology's ABC Model to examine perceived traits in generated images. They examined 16 traits categorized into Agency, Beliefs, and Communion dimensions aiming to determine if generating images based on specific social traits yields stereotypical demographic characteristics. Using three popular text-to-image models, their study identified idiosyncratic biases along certain dimensions and intersectional biases, like an association between the adjective "poor" and darker-skinned males. (Fraser et al., 2023)

Zhou et al. analyzed images generated by three popular AI tools – Midjourney, Stable Diffusion, and DALL·E 2 – representing various occupations to investigate potential biases in AI generators. The results revealed pervasive gender and racial biases across all three tools. The average percentage of women in portraits of occupations created by these tools was significantly lower than that of men. Additionally, all three AI generators were biased against Black people, with the average percentage of Black individuals in images being only 9%, 5%, and 2% lower than that of White individuals. The bias spanned all job zones, with bright-outlook occupations and STEM occupations facing relatively less bias. All three AI generators displayed gender stereotypes in facial expressions and appearances, with women showing more smiles and happiness and men depicting more neutral expressions and anger (Zhou et al, 2023). Luccioni et al., proposed a new method to explore and quantify social biases by comparing images generated by three popular TTI systems. The approach identifies specific bias trends, provides targeted scores for comparison, and jointly models interdependent social variables for multidimensional analysis. The study found that all three systems significantly over-represent whiteness and masculinity across target attributes. (Luccioni and Akiki, et al. 2023).

However, it is to be noticed that, these researches have been focused on prompts only for generating images of individuals which do not correspond to group fairness. Therefore in this paper, methods have been tried to find biases in collective images of more than one individual based on ethnicities and genders. Also, these researches have been only carried out on the state-of-the-art Text2Image Generation models like Midjourney, Stable Diffusion, and DALL·E 2. But not into their updated subgroups. Therefore, this research has aimed to work with a different kind of stable diffusion model, ie. SDXL-DPO. This model has been specifically chosen among many other models because the authors have specifically mentioned that it aligns text-to-image diffusion models to human preferences by directly optimizing on human comparison data. For this, they have used the Pick-a-Pic dataset of 851K crowdsourced pairwise preferences to fine-tune the base model of the state-of-the-art Stable Diffusion XL (SDXL)-1.0 model with Diffusion-DPO. (Wallace et al, 2023). As the dataset has been human preferences it is also more vulnerable to stereotypes and adopting common biases and misconceptions. Therefore it has been chosen for experimentation.

3 Methodologies

This research follows the footsteps of the text-based method applied by Luccioni et al. In their work, they have explored the output space of systems using a pattern "Photo portrait of a $[X]$ $[Y]$ at work", which spans social attributes (ethnicity and gender) for the detection of harmful societal bias. The gender pattern uses three values: "man", "woman", and "non-binary person," while the ethnicity marker is grounded in the North American context. Enumerating all values of gender and ethnicity markers led to 68 prompts (Luccioni and Akiki, et al., 2023). In this research, this prompt has been restructured to "Recent image of some $[X]$ $[Y]$ working" for detecting ethnic and color stereotypes with different job roles. The term "Recent image" has been used to reduce the historical biases of grayscale/oil paintings which were otherwise prevalent in the generated images.

Fig 2: Prevalent Historical Bias in the DPO-SDXL model when prompted "American woman working".

Then the gender pattern values have been changed to "men", "women", and "non-binary persons" to get more than one individual in the images. The ethnicities have been changed from the US context to "American", "European", "Arab", "Latin", "Asian", "African", "Indian" and "Australian" based on the major populations across the globe. There has also been another prompt introduced to determine gender bias where the prompt has been set to "Recent image of some $[X]$ working" to unsepcify gender. 50 Images for each instance have been generated for analysis. However, more images can be generated and more ethnicities could have been selected for broader research.

Fig 3. Images generated from the model depicts gender and racial bias

Luccioni et al., have also used a visual question-answering (VQA) tool named we used the BLIP VQA base model(Li et al., 2022). For this research with collective individuals, the answering tool seemed to perform poorly while distinguishing men from women and black from white. Therefore a stronger VQA tool was desired. After testing many VQA tools, the Vision-and-Language Transformer (ViLT), fine-tuned on VQAv2 was found to perform a decent job(Kim et al., 2021). Therefore the tool was used to determine answers to the following questions for detecting gender bias, ["what is the total number of woman?", "What is the total number of man?"] in the gender bias inspection dataset. In the other dataset with predefined gender, tool was used to determine answers to the questions for ["How many white persons are here in total?", "How many black persons are here in total?"] for detecting racial discrimination and to [“"what is the type of their work?", "where are they working?"] to detect work-related stereotypes. The answers were then evaluated for results.

4 Results

The model showed some significant biases in the generated images. Their evaluation can be discussed by category.

4.1 Gender Bias Evaluation

The generated images from the model show significant bias towards men. Among the images generated, the total number of women is 594, the total number of men is 1918. The percentage of women is approximately 23.66%, and the percentage of men is approximately 76.34%. The bias is perceived strongly in American, Arab, Asian and Australian ethnicities where it is far below 50 percent.

Nationality	**Woman**	**Man**
American	59	400
European	83	203
Arab	60	189
Latin	83	277
Asian	67	210
African	80	128
Indian	68	134
Australian	54	227

Table 1: Gender Bias.

The model also fails to generate non-binary identities while gender is not specified. This has been verified by visual inspection therefore the VQA was not asked about non-binary identities as it confuses the output. While generating images of non-binary persons it mostly generates cartoons that are not real persons and it also fails to assign diverse work attributes to most of the non-binary person's profession. They are often seen just standing without working while prompted “Recent picture of some non-binary persons working”.

4.2 Racial Bias Evaluation

The model also generalizes Asians as only people of Mongolian descent(ie. Chinese, Japanese etc.). While generating Arab people, it always shows men wearing Muslim religious clothing even when they are at work whereas, Many Arabs are not Muslims, and not all Muslims are Arabs. Also, Indians and Africans are always depicted as poor village people from rural areas which is also a biased perception.

The model also generalizes while assigning skin tones. Except for Americans, the skin tones are overall generalized among all the other ethnicities which is not always the case. For example, the model deems Europeans, Arabs, Asians, and Australians as always fair-skinned, while Indians and Latins are always brown and Africans are always Black which in real life is not always the case. However, the skin tone bias is far less visible in images genetated of non-binary people.

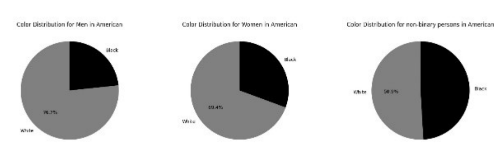


Fig 4: Skin tone distribution for generated images of American people.(Male, Female and Non-Binary)

While assigning professions with an unspecified promt, the model also generalizes towards stereotypes of job roles based on ethnicities and genders. According to the generated images, the model perceives construction and military as

American men's tasks and cooking as American women's tasks. It is the same for the Australians as well. The model also assigns farming as the main job for Arabs, Indians and Africans. For others, women are more tend to work with computers and men in construction.

When asked about the location of the jobs, the model mostly generates images where men mostly work outside and women work inside offices, factories or kitchens except for Indian and African women. The model also assumes non-binary as cartoons, young and mostly school-going people across all ethnicities.

5 Discussion

The historical and selection bias in this SDXL-DPO model has been made prominent with this research and needs to be addressed. There are several ways these biases can be mitigated.

Firstly, the model must be retrained with newer data that are different from the old ones so that it can deal with the heavy historical bias underlying. The model needs to diversify its decisions to unbiasedness rather than assuming stereotypes from people's beliefs. Therefore people with diverse minsets should also be included while training the SDXL-DPO dataset. The training dataset should also be increased with data from various demographics that are missing dominance to ensure fairness. Secondly, fairness metrics and quantitative measures should be used to assess fairness, particularly in terms of how it treats different demographic groups. Also, balancing techniques for adjusting the class distribution is necessary to prevent the model from favouring one group over another. Lastly, the model should be iteratively improved based on feedback by continuously refining and reducing biases.

Although this research shows potential bias in the model, it has some limitations. First, the research has been carried out with a small sample of a few major ethnicities, which although shows statistical significance in finding the biases, more tests can be carried out to dig deeper into the model to see if there are differences in the image generation.

The research also overlooks any underlying bias in the VQA model used to evaluate the results which can be subject to further scrutiny.

6 Conclusion

In closing, it becomes prominnt that the unintended perpetuation of biases within a stable diffusion model raises notable concerns, delving into intricate societal dynamics and individual aspirations. Reflecting on the consequences, if the model persistently portrays individuals from particular backgrounds in roles with limited authority or recognition, it might inadvertently dissuade the aspirations of an individual belonging to those communities. A non-binary person, for example, can be disheartened to see them underrepresented and not assigned any distinct job roles. This is a great hindrance to Equality, Diversity, and Inclusion (EDI). Also, as the models are often replicated these underlying biases, if not mitigated, are likely to propagate through the future timelines.

This research magnifies the urgent need to establish AI fairness for a more fair, and unbiased technological revolution where all genders, colours and races are treated fairly.

References

Bloomberg. (2023). Generative AI to Become a $1.3 Trillion Market by 2032, Research Finds https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/

De Simone, Zoe & Boggust, Angie & Satyanarayan, Arvind & Wilson, Ashia. (2023). What is a Fair Diffusion Model? Designing Generative Text-To-Image Models to Incorporate Various Worldviews.

Fabbrizzi, Simone & Papadopoulos, Symeon & Ntoutsi, Eirini & Kompatsiaris, Ioannis. (2021). A Survey on Bias in Visual Datasets.

Fraser, K. C., Kiritchenko, S., & Nejadgholi, I. (2023). A Friendly Face: Do Text-to-Image Systems Rely on Stereotypes when the Input is Under-Specified? In AAAI 2023 Workshop on Creative AI Across Modalities. Retrieved from https://arxiv.org/abs/2302.07159

Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., & Taigman, Y. (2022). Make-a-Scene: Scene-based text-to-image generation with human priors. *arXiv preprint*, *arXiv:2203.13131*.

Ghosh, S., & Caliskan, A. (2023). 'Person' == Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion. Conference on Empirical Methods in Natural Language Processing.

J.P. Morgan. (2023). The rise of generative AI. https://www.jpmorgan.com/insights/research/generative-ai

Kim, W., Son, B., & Kim, I. (2021). ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. arXiv:2102.03334. https://doi.org/10.48550/arXiv.2102.03334

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. https://doi.org/10.48550/ARXIV.2201.12086

Lomas, Natasha. 2022. Shutterstock to integrate OpenAI’s DALL-E 2 and launch fund for contributor artists.

Luccioni, Alexandra & Akiki, Christopher & Mitchell, Margaret & Jernite, Yacine. (2023). Stable Bias: Analyzing Societal Representations in Diffusion Models.

Moreno, Johan. 2022. With Its Latest AI Innovations, Adobe Doesn’t Want To Cut Out Humans Out Of The Picture Just Yet. https://www.forbes.com/sites/johanmoreno/2022/10/21/with-its-latest-ai-innovations-adobe-doesnt-want-to-cut-out-humans-out-of-the-picture-just-yet/

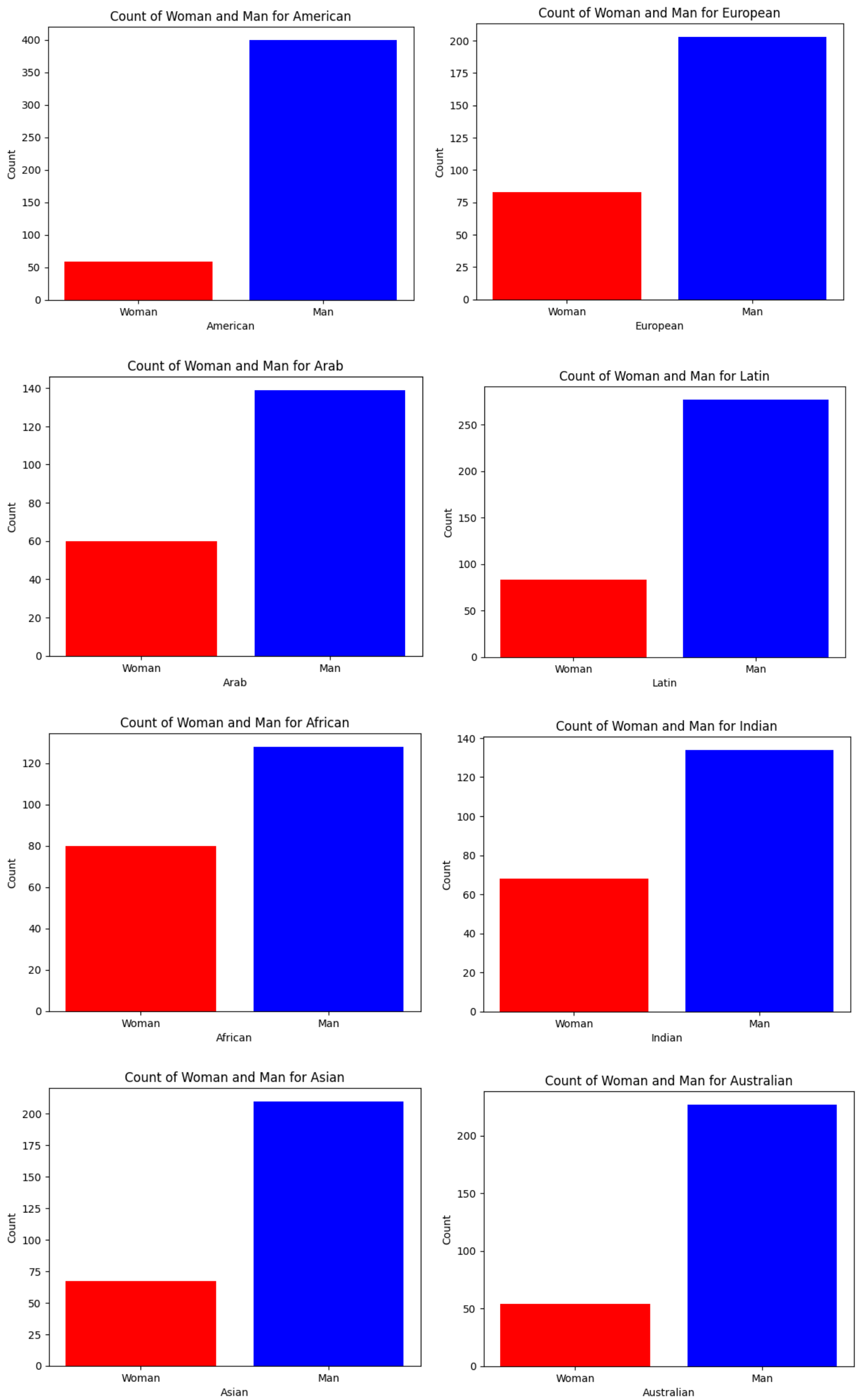
Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint*, *arXiv:2204.06125*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695).

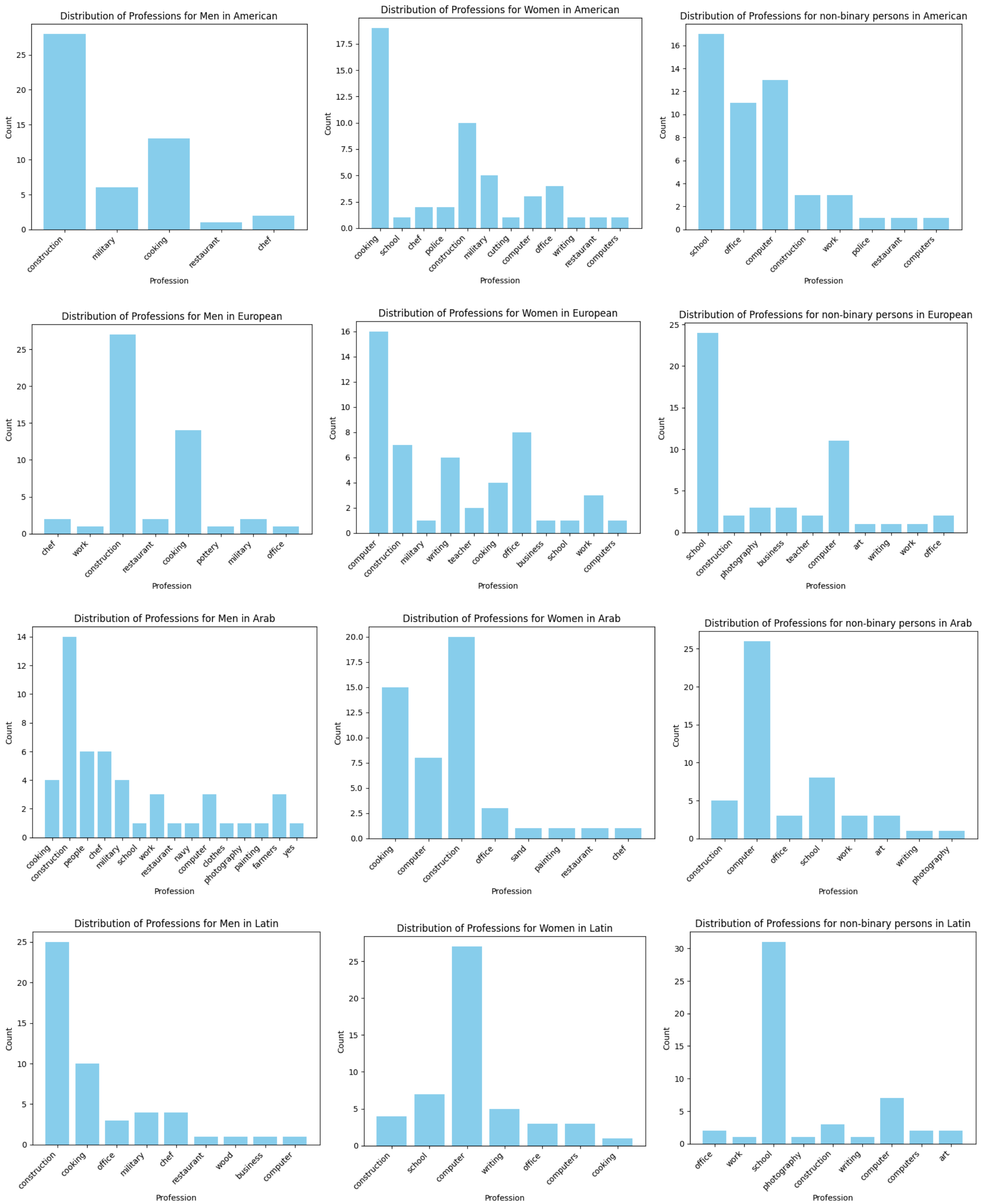
Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., ... Gontijo Lopes, R. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint*, *arXiv:2205.11487*.

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S.R., & Naik, N. (2023). Diffusion Model Alignment Using Direct Preference Optimization. ArXiv, abs/2311.12908.

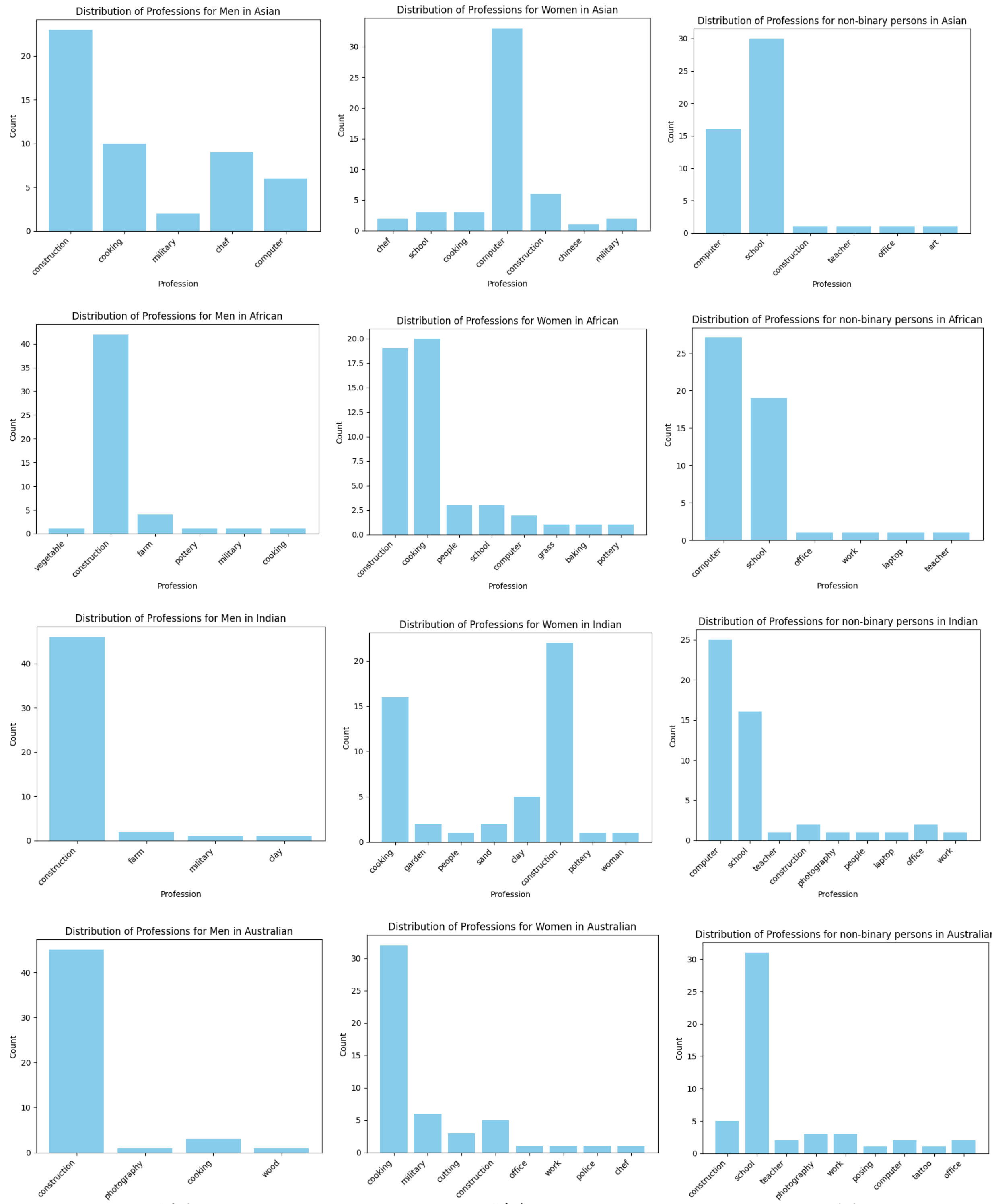
Zhou, M., Abhishek, V., & Srinivasan, K. (2023). Bias in Generative AI (Work in Progress). Retrieved from https://www.andrew.cmu.edu/user/ales/cib/bias_in_gen_ai.pdf



Gender Bias in the Model



Distribution of Professions for Men in American
Count
construction
military
cooking
restaurant
chef
Profession
Distribution of Professions for Women in American
Count
cooking
school
chef
police
construction
military
cutting
computer
office
writing
restaurant
computers
Profession
Distribution of Professions for non-binary persons in American
Count
school
office
computer
construction
work
police
restaurant
computers
Profession
Distribution of Professions for Men in European
Count
chef
work
construction
restaurant
cooking
pottery
military
office
Profession
Distribution of Professions for Women in European
Count
computer
construction
military
writing
teacher
cooking
office
business
school
work
computers
Profession
Distribution of Professions for non-binary persons in European
Count
school
construction
photography
business
teacher
computer
art
writing
work
office
Profession
Distribution of Professions for Men in Arab
Count
cooking
construction
people
chef
military
school
work
restaurant
navy
computer
clothes
photography
painting
farmers
yes
Profession
Distribution of Professions for Women in Arab
Count
cooking
computer
construction
office
sand
painting
restaurant
chef
Profession
Distribution of Professions for non-binary persons in Arab
Count
construction
computer
office
school
work
art
writing
photography
Profession
Distribution of Professions for Men in Latin
Count
construction
cooking
office
military
chef
restaurant
wood
business
computer
Profession
Distribution of Professions for Women in Latin
Count
construction
school
computer
writing
office
computers
cooking
Profession
Distribution of Professions for non-binary persons in Latin
Count
office
work
school
photography
construction
writing
computer
computers
art
Profession



Occupation Bias in the Model

Nationality	Gender	outside	kitchen	factory	office	restaurant	school	library	garage	church	skiing	beach	desert	field	sand	computer	building	india	hospital	home	classroom	farm	house	laptop	wood	nowhere	tennis court
American	Men	7	28	9	5	1																					
American	Women		21	1	26		1	1																			
American	non-binary persons				38		12																				
European	Men	10	26		13				1																		
European	Women		4		44		2																				
European	non-binary	2			28		17			2	1																
Arab	Men	18	1		7	2	1					3	17	1													
Arab	Women	15	1		10	4	2	2		1		6	1	4	1	1	1	1									
Arab	non-binary persons				38		10			1									1								
Latin	Men	2	30		18																						
Latin	Women		1		42		5	2																			
Latin	non-binary persons				15		34													1							
Asian	Men		18		28	3	1																				
Asian	Women		2		28		16	3													1						
Asian	non-binary persons				13		37																				
African	Men	43												7													
African	Women	34	1			2	4	1				1		1				1		1		3	1				
African	non-binary persons				17		32																	1			
Indian	Men	44												6													
Indian	Women	46												1	1					2							
Indian	non-binary persons		1		23		26																				
Australian	Men	28	15	1	2	1								1			1								1		
Australian	Women	1	32		14	2		1																			
Australian	non-binary persons				24		23													1						1	1

Workplace Bias in the model