Abstract

Introduction

Motivation

Foundational challenges in assuming safety of large language models

Related Work

Gender Bias in Artificial Intelligence:

Previous studies explores biases in LLM, OpenAI, GPT2, ChatGPT, LLaama (UNESCO).

1. Bias in pre-trained language models
   The community has developed a gamut of datasets and methods to measure and mitigate biases in text-only LLMs (Bordia and Bowman, 2019; Liang et al., 2020; Ravfogel et al., 2020; Webster et al., 2020; Lauscher et al., 2021; Smith et al., 2022; Kumar et al., 2023; Nadeem et al., 2021; Nangia et al., 2020).
   Bias in pre-trained vision models
   The use of vision models on various tasks has been hindered by bias in vision, as demonstrated by multiple studies Buolamwini and Gebru (2018); DeVries et al. (2019); Wilson et al. (2019); Rhue (2018); Shankar et al. (2017); Steed and Caliskan (2021). Numerous studies have been conducted to measure the extent of biases present in vision models Steed and Caliskan (2021); Shankar et al. (2017); DeVries et al. (2019); Buolamwini and Gebru (2018).
   Bias in Vision and Language models
   Image-to-text : Hall et al. (2023) introduced a novel portrait based dataset for benchmarking social biases in VLMs for both pronoun resolution and retrieval settings. Srinivasan and Bisk (2021) measure the associations between small set of entities and gender in visual-linguistic models using template based masked language modeling.Zhou et al. (2022); Janghorbani and de Melo (2023) study stereotypes in VLMs.
   Text-to-image: Cho et al. (2023) highlights a bias towards generating male figures for job-related prompts and limited skin tone diversity, while probing miniDALL-E Kim et al. (2021) and stable diffusion Rombach et al. (2022b). The prompts used to generate images explicitly specify the profession. Fraser et al. (2023); Ghosh and Caliskan

(2023) further highlights stereotypical depictions of people within text-to-image models

References:

https://twitter.com/DavidSKrueger/status/1779900511627452467

Foundational challenges in safety of Large Language models https://arxiv.org/pdf/2404.09932

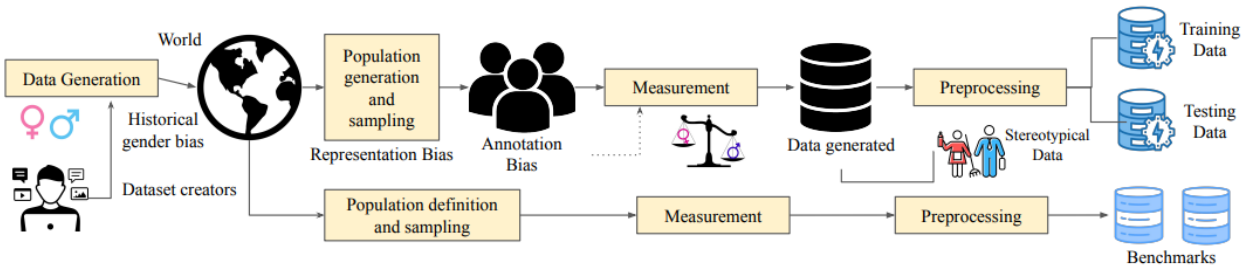

Fig. 4: Gender Bias induced from Data Generation