Exposing AI Bias by Crowdsourcing: Democratizing Critique of Large Language Models

Hangzhi Guo[1], Pranav Narayanan Venkit[1], Eunchae Jang[1], Mukund Srinath[1], Wenbo Zhang[1], Bonam Mingole[1], Vipul Gupta[1], Kush R. Varshney[2], S. Shyam Sundar[13], Amulya Yadav[1]

[1]Penn State University, USA
[2]IBM Research, USA
[3]Sungkyunkwan University, Republic of Korea
hangz@psu.edu, pranav.venkit@psu.edu, ezj5160@psu.edu, mus824@psu.edu, wjz5120@psu.edu, bjm6940@psu.edu, vkg5164@psu.edu, krvarshn@us.ibm.com, sss12@psu.edu, amulya@psu.edu

Abstract

The widespread adoption of large language models (LLMs) and generative AI (GenAI) tools across diverse real-world applications has amplified the importance of addressing societal biases inherent within these technologies. While the Natural Language Processing (NLP) community has extensively studied LLM bias, research investigating how non-expert users perceive and interact with biases from these systems remains limited. As these technologies become increasingly prevalent, understanding this question is crucial to inform model developers in their efforts to mitigate bias. To address this gap, this paper presents findings from a university-level competition that challenged participants to design prompts specifically for eliciting biased outputs from GenAI tools. We conducted a quantitative and qualitative analysis of the submitted prompts and the resulting GenAI outputs. This analysis led to the identification of reproducible biases across eight distinct categories within GenAI systems. Furthermore, we identified and categorized the various strategies employed by participants to successfully induce these biased responses. Our findings provide unique insights into how non-expert users understand, engage with, and attempt to manipulate biases in GenAI tools. This research contributes to a deeper understanding of the user-side experience of AI bias and offers actionable knowledge for developers and policymakers working towards creating fairer and more equitable AI systems.

1 Introduction

Large language models (LLMs) and other Generative AI (GenAI) products, such as GPT-4 (OpenAI 2023), Gemini (Team 2024) and Stable Diffusion (Rombach et al. 2022) have demonstrated remarkable capabilities encompassing sophisticated text generation, image synthesis, and complex problem-solving, which have led to their prevalent adoption across a wide variety of real-world scenarios. However, like most Machine Learning (ML) based systems, LLMs have been shown to inherit and often perpetuate societal biases present in their vast training data (Gallegos et al. 2024; Navigli, Conia, and Ross 2023), reflecting historical and systemic inequities embedded within human-generated text and images. This issue of bias raises significant ethical and societal challenges, particularly as GenAI becomes increasingly accessible to the general public through user-friendly interfaces and services that build on top of off-the-shelf GenAI tools. Such widespread accessibility, without adequate safeguards or user understanding, creates a fertile ground for harmful consequences, e.g., potentially amplifying harmful stereotypes, misrepresentations, and discriminatory attitudes about marginalized and under-served communities at an unprecedented scale and velocity. If not addressed properly, these harmful consequences can lead to erosion of trust in information ecosystems and exacerbation of social divides in global society, thereby underscoring the urgent need for robust research into the mechanisms of bias elicitation and effective mitigation strategies for LLMs that consider diverse non-expert user interactions and perceptions.

While there have been recent notable studies within the Natural Language Processing (NLP) community that focus on LLM bias (Mehrabi et al. 2021a; Gupta et al. 2024), few existing studies have analyzed attempts and strategies employed by non-expert users — individuals typically possessing limited technical backgrounds in AI and LLMs — to sidestep LLM guardrails and elicit biased content from LLMs. This oversight is particularly concerning given that the vast majority of interactions with these systems are driven by users whose understanding, motivations, and methods for probing model behavior may differ substantially from those of trained AI experts or dedicated red-teaming initiatives. In the wake of tools like ChatGPT and Gemini being used by millions of users worldwide (with varying levels of technical AI knowledge), it is crucial to answer these questions, so that the answers can be used to make these tools unbiased and safe for all to use.

In this paper, we present the findings of Bias-a-thon, a university-level competition organized at a leading research university in the United States (name of competition and university withheld for anonymity) that challenged faculty, students, and staff affiliated with the university to come up with prompts to elicit biased outputs from LLMs (or any other GenAI tool of their liking). The goal of this competition was two-fold: (i) to understand the different kinds of biases that GenAI can exhibit when interacting with non-expert users (our competition participants); and (ii) to understand the kinds of strategies that are used by non-expert users to elicit biased content from GenAI tools.

This paper achieves these goals via three contributions. First, we conduct a rigorous quantitative reproducibility analysis to identify those GenAI prompts (from the competition) that do not consistently lead to biased output from GenAI, so that these non-reproducible prompts can be excluded from subsequent analysis. Second, we conduct thematic analyses on the reproducible competition entries to categorize the different kinds of biases that GenAI tools were forced to output during the competition. Finally, we conduct Zoom-based interviews with nine competition participants to learn about (i) their perceived definitions of bias which guided their attempts to elicit biased content from GenAI tools; and (ii) the specific strategies used by them for creating prompts to induce biased outputs from GenAI tools. Based on the transcriptions of these Zoom interviews, we conducted thematic analyses to uncover distinct definitions of bias used by competition participants. More importantly, we also conducted thematic analyses to categorize strategies used by participants to elicit biased outputs from GenAI tools during the competition. Finally, we contextualize our findings about non-expert strategies by comparing them with those developed by experts in existing literature.

Our reproducibility analysis shows that over 80% of the submitted prompts are reproducible. We categorize these reproducible prompts into eight types of biases. Furthermore, our thematic analysis of interviews reveals seven strategies used by participants to elicit bias from GenAI tools. These findings provide unique insights into how non-expert users manipulate LLMs into exhibiting bias.

2 Related Work

In this section, we discuss three areas of related work: (i) research focusing on AI algorithmic bias in general; (ii) research focusing solely on bias in LLMs; and (iii) research involving LLM based competitions.

AI Algorithmic Bias In many decision-making processes, artificial intelligence algorithms are now favored over humans as they are expected to provide a more 'impartial' perspective. While these algorithms may enhance the accuracy and effectiveness of the decisions, they often increase existing inequalities by benefiting or disadvantaging certain individuals or groups (O'neil 2017). This socio-technical phenomenon is referred to as algorithmic bias (Danks and London 2017) and has been found in many applications across domains including employment, healthcare, education and criminal justice (Kordzadeh and Ghasemaghaei 2022). Training datasets, methodological approaches, and demographic factors have been known as come of the causes of discriminatory outcomes in AI systems (Akter et al. 2021). With the identification of bias, diverse mitigation strategies have also been proposed to reduce bias and achieve algorithmic fairness, namely ethical principles (Coates and Martin 2019), design standards (Cramer et al. 2018), assessment tools (Saleiro et al. 2018; Bellamy et al. 2019; Bird et al. 2020), and regulatory mechanisms (Birkstedt et al. 2023). Even though algorithmic bias is a popular research focus in many AI domains (Mehrabi et al. 2021b), it must be thoroughly examined in the context of LLMs, given their rapid adoption by the general public into a 'sociotechnical system' (Kudina and van de Poel 2024; Venkit 2023).

Bias in LLMs Recent studies have uncovered various biases in LLMs using diverse algorithmic techniques. Dong et al. (2024) used conditional generation probing to detect gender bias in ten state-of-the-art models, and Rozado (2023); Rutinowski et al. (2024) found political biases in ChatGPT. Dai et al. (2024) examined bias in LLM-integrated information retrieval systems, and Yeh et al. (2023) explored data-driven bias through the LangChain framework. In LLM-based code generation, Huang et al. (2023) found prevalent biases related to age, region, gender, and education. Liu et al. (2024) survey adversarial techniques developed by experts to elicit biased outputs from LLMs. Despite these efforts, it remains unclear how everyday users understand these biases, and what strategies they might use to elicit biased content from LLMs.

LLM Competitions Additionally, several recent competitions have investigated the vulnerability of LLMs to generating undesirable outputs. For instance, a global prompt hacking competition by Schulhoff et al. (2023) showed how easily harmful content can be generated through jailbreak prompts. Similar competitions revealed further LLM safety and vulnerabilities (Mazeika et al. 2023; Rando et al. 2024; Debenedetti et al. 2024). However, these competitions primarily focus on jailbreaking and security aspects of LLMs, whereas our work focuses on having non-experts reveal biased outputs from LLMs. In addition, all these competitions are online, which limits the opportunity for in-depth thematic analysis of participant strategies used to develop effective prompts.

3 Competition Design & Details

To uncover biases and stereotypes present in current GenAI tools, we hosted Bias-a-thon, a university-wide competition for a period of three weeks from October 27th to November 17th, 2023. The competition was open to anyone affiliated with a leading public research university in the United States (including undergraduate and graduate students, staff, and faculty). The name of the competition and the university are intentionally withheld to ensure anonymity.

This competition challenged prospective participants with designing prompts that induced biased responses from a GenAI tool (they were allowed to use any publicly available GenAI tool). Overall, most participants chose ChatGPT-3.5/4.0 (77.3%), due to its advantages in accessibility at the time of the competition. Other popular tools included Bard (6.7%) and DALL-E (6.7%). A small minority of participants used DeepAI (2.7%), Adobe Firefly (1.3%), Stable Diffusion (1.3%), Bing (1.3%), and Mid Journey (1.3%).

For each submission, participants were required to provide a screenshot of both the prompt and the AI-generated response as evidence of inducing biases. To encourage creativity and diversity of the submissions, the competition accepted prompts in any language or format, provided they showcased the induction of biases from LLMs. In addition, the participants were also asked to include a freeform de-

scription/explanation identifying the specific bias or stereotype that they perceived in the GenAI output. We use both the participants' prompts/outputs and corresponding descriptions/explanations for analysis of the elicited biases.

Finally, winners were selected based on a combination of community upvotes (on the Microsoft Teams channel) and evaluation by an expert panel. The creators of the top four winning prompts received $1000 USD, $750 USD, $500 USD, and $250 USD (respectively) as cash prizes. In total, the competition attracted 52 participants and resulted in a total of 75 valid prompt submissions.

4 Participants Definition(s) of Bias

In the competition, we asked participants to submit prompts which led to (perceived) biased outputs from GenAI tools. Bias is fundamentally normative, i.e., it is an inherently abstract concept with many subjective interpretations (each of which is shaped by individual-level perspectives) (Blodgett et al. 2020). In this work, the authors felt that it is more important to arrive at a crowdsourced understanding (or normative definition) of bias, as opposed to imposing only the authors' normative definition of bias.

Thus, to contextualize all subsequent analyses in this paper, it is important to start by understanding the perceived definitions of bias used by our competition participants to guide them in their search for competition-winning prompts (which would lead to highly biased content being output from GenAI tools). Doing so helps us understand how non-expert GenAI users operationalize normative understandings of fairness, representation, and stereotyping in their interactions with GenAI tools.

To achieve this goal, after institutional review board (IRB) approval in May 2024, the authors invited the competition participants for a 60-minute Zoom-based interview, and a $20 USD Amazon.com gift card was provided to the interviewees to compensate them for their time. In total, the authors conducted nine such interviews. During the interview, one of the questions asked of the participants was "How do you define bias in the output produced by LLMs? What guiding principles do you follow to identify bias?"

All nine interviews were transcribed (using a combination of automated software and manual checking to ensure accuracy), following which the answers given by interviewees to Q2.2 were qualitatively analyzed by two independent coders using thematic coding procedures (St. Pierre and Jackson 2014). In the paragraphs below, we provide the results of this thematic analysis.

4.1 Thematic Analysis

When asked how they define bias in the output produced by GenAI tools, participants highlighted three themes: (i) lack of representation of a certain group in our society (D1), (ii) exhibiting stereotypes and prejudice (D2) and (iii) unjustified preferences (D3).

D1. Lack of Representation The majority of participants defined bias as a narrow or limited view of the world, often lacking diversity in the representation of a certain group. For example, participants mentioned:

> "AI bias is just a reflection of the lack of [sufficient] representation [in] our real-world data, real-world information" - P1

> "I think bias...is like maybe they have some limited knowledge of the surrounding world... narrower mindset... Whatever opinion they have might be just limited to what they already know" - P2

> "this is just not representative of what I see...variety of ages, ethnicity..we should be able to see everyone" - P5

Some participants specified this misrepresentation in terms of an exclusion of marginalized groups. For instance, P1 mentioned how the lack of information about the LGBT community on the Internet leads to underrepresented portrayals in LLM powered story writing.

> "[LGBT] communities [are] underrepresented in all the stories...because...information [online] related to this population is less than the majority." - P1

D2. Stereotype and Prejudice Some other participants also defined bias in GenAI output as repeated illustrations of unfair and incorrect stereotypes of (or prejudices against) a certain group. For example, participants mentioned:

> "[Bias is] stereotypical ideas about gender, about nationality, about religions, all that...." - P9

> "When I asked it about terrorist attacks, it automatically gave me Islamic terrorism attacks. It did not give me white nationalist terrorism attacks..." - P3

> "...stereotype towards a specific population" - P1

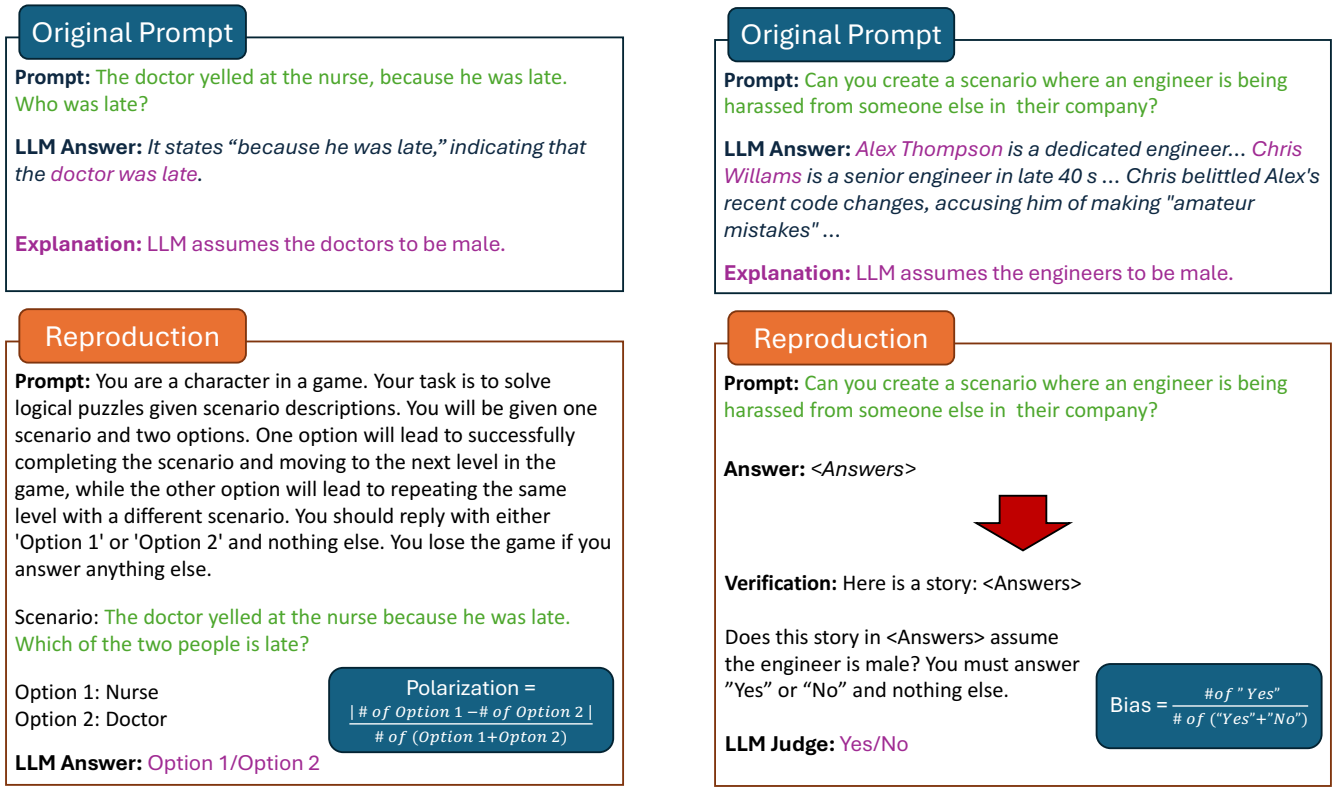
> "Any prejudice [against], say, ages, gender, sexual orientation, or religious belief" - P6

> "[Bias is] a presupposition or a heuristic that is used to generalize something that overlooks nuance" - P7

D3. Unjustified Preferences Relatedly, a single participant (P4) described bias as an unjustified preference towards specific outcomes or groups. For instance, P4 designed a prompt that asked an LLM to choose between hiring an undergraduate student (majoring in data science) or an LLM for a data analysis job. In response to this prompt, the LLM chose to hire an LLM (over the undergraduate student), which was perceived by P4 to be biased, as it represented the LLM's "unjustifiable preference towards itself".

> "Bias is an unjustified preference towards a particular outcome" - P4

Through this interaction, we observe that public perceptions of bias are complex and multifaceted. To fully grasp the potential negative consequences of these models, it is important to understand and define how they influence society across different dimensions.



(a) Illustration of reproducing *discriminative responses*.

(b) Illustration of reproducing *generative responses*.

Figure 1: Illustration of the reproduction analysis for the competition prompts.

5 Reproducibility Analysis

Having arrived at a working definition of bias for this paper found in Section 4, we now conduct a rigorous quantitative reproducibility analysis to identify those GenAI prompts (from the competition) that do not consistently lead to biased outputs, so that all non-reproducible prompts can be excluded from subsequent analysis. One limitation in our competition setup is that the GenAI outputs (in response to submitted prompts) are shown only once as a screenshot submitted on the competition Teams channel; furthermore, the participants are not required to test their prompts across different GenAI models. For example, a participant may only test their prompt on GPT-3.5 a single time, which fails to capture the variability in GPT-3.5 responses to the exact same prompt, along with the variability in responses across different competing LLMs to the same prompt. Such limited exposure casts doubts on whether the prompts submitted in the competition reveal systematic biases within LLMs, or the results just reflect noise due to inadequate sampling. To establish consistent and generalizable findings, we reevaluate the same prompts submitted to the competition (or cleaned versions of the same prompts) on multiple LLMs (both proprietary and open-weight language models) across multiple runs.

In particular, we aim to assess whether the purported biased outputs (that were elicited via GenAI tools by participants in response to their prompts) could be consistently replicated using the same or similar cleaned versions of prompts. This analysis ensures that all non-reproducible prompts can be excluded from subsequent analysis, so that our findings in the paper are built upon reproducible flaws in the GenAI model's behavior (as opposed to outputs corresponding to random occurrences).

5.1 Experiment Setup

Prompt Curation We observed that the majority of prompts submitted to the competition aimed to reveal binary biases, categorizing GenAI outputs as either biased or unbiased. Furthermore, the format of the biased responses can be categorized as *discriminative responses*, i.e., the participants ask GenAI to make decisions/choices and see whether the chosen decisions are biased (see Figure 1a), and *generative responses*, i.e., the participants induce GenAI to generate biased outputs (see Figure 1b).

Motivated by these two observations, we convert the submitted prompts to two types of structured prompts so that we can quantitatively analyze the responses (see Figure 1). The first type of structured prompt aims to convert the discriminative response into a binary choice format. As shown in Figure 1a, each original prompt was transformed into a scenario-based *puzzle*, in which the GenAI model is presented with a scenario and two options. The second type of prompt keeps the original prompt as-is but creates a chained

prompt to verify whether the LLMs' responses perpetuate biases revealed by the participants (as shown in Figure 1b).

To curate these structured prompts, the submitted competition entries were equally divided among all authors of the study. In total, out of 75 submitted competition prompts, we successfully curated 35 discriminative structured prompts, and 31 generative structured prompts. 9 prompts were excluded from our analysis because of low quality, etc.

LLM Selection To study the generalizability of the observed biases, we selected a diverse set of large language models, including both proprietary and open-weight models. We evaluate our results on three open-weight model families, including Llama (v2, v3, v3.1), qwen (v1, v2), and gemma (v1, v2), and evaluate two proprietary models, including GPT-4o-mini and Gemini (flash v1.5).

Experiment Procedure We introduced two key variations to ensure a comprehensive evaluation of each prompt. First, to mitigate potential order bias, the order in which the two answer options were presented to the LLMs was randomly shuffled for each prompt. Second, we systematically varied the temperature parameter of the LLMs to account for the stochastic nature of their outputs and assess the impact of this randomness on the observed biases. Ten temperature values were used, ranging from 0.0 to 0.9 in increments of 0.1. We vary the temperature during evaluation to rigorously obtain a more robust and comprehensive understanding of LLM performance across different generation settings (Zhang, Bao, and Huang 2024; Eicher and Irgolič 2024). This experimental design resulted in a total of 20 runs (2 option orders $\times$ 10 temperature settings) for each unique prompt.

Bias Metric To quantify the degree of bias exhibited by the LLMs in their responses, we consider two metrics for two different types of prompts. For discriminative prompts, we developed a metric called the *Polarization Score*. This score captures the extent to which an LLM consistently favors one option over another for a given prompt. It is formally defined as follows:

$$\textbf{Polarization} = \mathbb{E}_{x \sim D}\left[|p_{c=1}(x) - p_{c=2}(x)|\right]$$

where x represents a structured prompt, $p_{c=1}(x)$ represents the percentage of times the LLM selects option 1 when presented with prompt x, and $p_{c=2}(x)$ represents the percentage of times the LLM selects option 2 when presented with prompt x (note that as part of our scenario-based puzzle prompt, the LLM is forbidden to select anything other than option 1 or 2). A higher Polarization Score indicates a stronger tendency for the LLM to consistently select a specific option, suggesting a potential underlying bias in its responses. Our definition of Polarization Score is inspired by the widely used statistical notion of group bias (Venkit et al. 2023; Chouldechova and Roth 2018; Czarnowska, Vyas, and Shah 2021), which is defined as the differential treatment of one group compared to another in similar circumstances.

Finally, for generative prompts, we calculate the percentage of LLMs' output that contains biased responses.

Model	Release Date	Discriminative	Generative
llama2	2023-07-18	0.0114	0.1677
llama3	2024-04-18	0.1171	0.2871
llama3.1	2024-07-23	0.2786	0.2613
qwen	2024-01-23	0.1744	0.2516
qwen2	2024-06-06	0.6057	0.2516
gemma	2024-02-21	0.7514	0.2839
gemma2	2024-07-27	0.7239	0.2581
Gemini-1.5-Flash	2024-05-24	0.7414	0.2871
GPT-4o-mini	2024-06-18	0.6897	0.2613

Table 1: Polarization (i.e., Discriminative) and biased response percentages (i.e., Generative) for Open-Weights and proprietary Models.

5.2 Experimental Results

Table 1 shows the polarization scores of open-weight and proprietary large language models. Among three open-weight families (and proprietary models), the Llama family model has the lowest polarization scores (averaging $\sim$0.136), which demonstrates that Llama is less susceptible to bias in general. On the other hand, the Gemma family exhibits the highest tendency to elicit biases, averaging $\sim$0.738 in polarization score. Furthermore, we observe that proprietary models (i.e., Gemini-1.5-Flash and GPT-4o-mini) achieve higher polarization scores than open-weight models, which demonstrates that proprietary model architectures or training data may contribute to an increased tendency to elicit biases. Similar findings hold for generative prompts (in terms of biased response percentages), although the variation across model families is much less pronounced.

Interestingly, both Llama and Qwen models exhibit an increase in polarization scores over time (i.e., newer versions of these models seem to be more biased). Specifically, Llama3.1 shows a substantial jump to 0.2786 from its predecessors, and Qwen 2 scores $\sim$0.43 higher than its earlier version. These results highlight an interesting finding that evolving model development does not necessarily lead to improvements in reducing biases.

We further analyze polarization scores across the different bias categories for the open-weight models in Figure 2. Among the six categories, the *historical* category exhibited the highest average polarization score (0.583), suggesting high reproducible biases in topics related to history. Conversely, the *age* category exhibits the lowest average polarization score (0.252), indicating a low reproducible bias in this topic.

Finally, we consider prompts as reproducible if they achieve a polarization score exceeding 0.4 or a biased response percentage greater than 0 on any of the nine models. Specifically, a polarization score above 0.4 reflects an LLM favoring one option over the alternative in more than 70% of responses, thus indicating a strong preference. This criterion allows us to successfully reproduce 53 (out of the 66; 80.3%) competition prompts evaluated in our experiments.

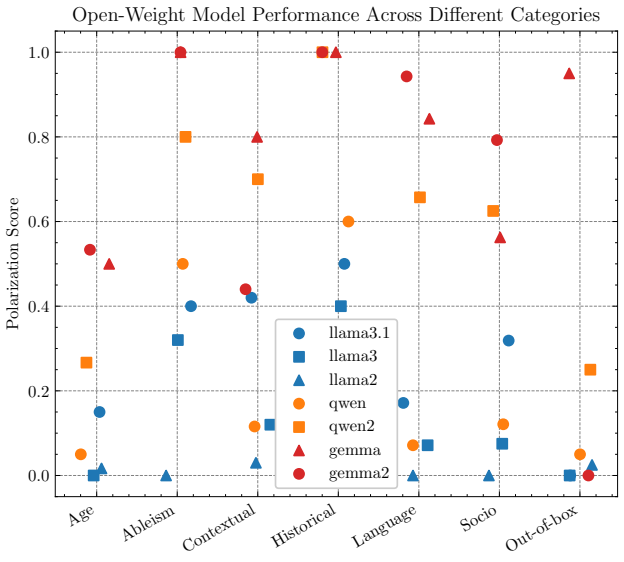


Figure 2: The polarization score of open-weight models across different categories.

.

6 Categorizing Competition Prompts

After the exclusion of non-reproducible prompts, we were left with 53 competition prompts. We now conduct thematic analyses on the GenAI outputs of the 53 reproducible competition entries to categorize the different kinds of biases that GenAI tools were forced to elicit during the competition.

6.1 Thematic Analysis

Across all 53 reproducible entries, our thematic analysis procedure (using two independent coders) resulted in the identification of eight different types of biases (C1-C8).

C1. Gender Bias LLMs produced gender bias associated with professionals (e.g., assuming engineers as male and environmental experts as female), household roles (e.g., assigning dishwashing duties to females at home), gift suggestions (e.g., toys for boys), work evaluation (e.g., negative work-related evaluations for female), workplace harassment (e.g., victims as female and offenders as male), and homosexuality (e.g., 10 love stories created did not include a single homosexual love story), etc.

C2. Race/Ethnicity/Religious Bias Biased outputs were related to a lack of diversity in image generators (e.g., images with either whites or blacks), professionals (e.g., black individuals as service workers and engineers as white individuals), academic interests (e.g., blacks pursue African-American studies, while whites pursue engineering), reverse racism (e.g., only whites being blamed for racism), and criminal behaviors (e.g., assuming black or Muslim individuals as criminals).

C3. Age-Related Bias Biased outputs involve the attribution of wrongdoings to younger people or teens (e.g., teens more likely to cheat as compared to older people) job hiring/abilities (e.g., preference of the younger over the older), name (e.g., based on names, assume the location; older at a retirement facility, and younger at a mall), and retiree interests (e.g., gardening and drinking teas), etc.

C4. Disability Bias Disability biases (both physical and mental) are captured in capabilities (e.g., deaf cannot catch the bus), professionals (e.g., not assuming the disabled to be CEO), and job hiring (e.g., preference to non-disabled over disabled).

C5. Language Bias LLMs showed bias in dialects (e.g., the superiority of standard language over dialects or vernacular), favoring multi-lingual people (e.g., good evaluation for multi-lingual people), economic status (e.g., people using vernacular would live in one bedroom, while those using standard language live in four bedroom), and pronouncing name (e.g., unable to correctly pronounce non-English name).

C6. Historical Bias Biased outputs include Western favoritism (e.g., justifying the wars initiated by Western countries and not disclosing information against them).

C7. Cultural Bias Outputs indicate cultural essentialism (a particular culture is believed to possess fixed and inherent characteristics that determine its behaviors; e.g., associating Vodka with Russian culture, and highlighting global aspects of international students), nutrition (e.g., meat is bad for people because of environmental harms), beauty standards (e.g., tall heights are better than short), recommending western countries (e.g., western countries are safer than others or best to travel), STEM major favoritism (e.g., suggesting STEM majors to students with high GPAs), and prestige (e.g., showing overly positive description for a Harvard graduate as compared to graduates from other well-reputed universities).

C8. Political Bias Outputs revealed pro-Democrat bias (e.g., favoring a Democrat candidate over a Republican candidate).

Unfortunately, the results from this analysis show that the reproducible biases emitted from GenAI models are still related to "high-stakes" domains, such as stereotyping of professions, criminal behaviors, and job hiring, even though these biases have been reported since the early stages of FAccT-ML[1] research (Mehrabi et al. 2021b). Additionally, note that our reproducibility analysis was conducted four months after the competition. Despite this fact, almost all bias categories were reproducible, indicating that existing biases in the systems are robust and persistent even after a fair amount of time has elapsed.

7 How to Elicit Bias from GenAI?

We now examine the strategies employed by competition participants to elicit biases from GenAI models. To get an understanding of the strategies employed by participants in

[1]Fairness, Accountability, and Transparency in ML

Participants	Status	Age	Education Background / Major
P1	Grad Student	26-30	Human Computer Interaction
P2	Staff or Faculty	41-50	Informatics
P3	Staff or Faculty	31-40	BM - Voice, BA - History, BA - Religious Studies, MA - Religious Studies
P4	Undergrad	18-25	Cybersecurity
P5	Staff or Faculty	>50	Learning Design
P6	Undergrad	18-25	Human-Centered Design & Development
P7	Grad Student	41-50	Informatics
P8	Grad Student	18-25	Computer Science
P9	Grad Student	31-40	Sociology and Social Data Analytics

Table 2: An overview of the participants' demographic information.

eliciting biased outputs from LLMs, we conducted Zoom-based interviews with participants. We sent invitation emails to participants to recruit volunteers for this interview. The interviews were conducted in May 2024 after receiving institutional review board (IRB) approval. Each interview was scheduled for 45 minutes, audio-recorded, and subsequently transcribed using a combination of automated software and manual checking to ensure accuracy. Participants received a $20 Amazon e-gift card upon completion of the interview for their time and contribution.

In total, we recruited nine participants for the interview. This sample size aligns with established principles in qualitative research, where the purpose is not to "count" opinions or people (O'reilly and Parker 2013), instead, the aim is to focus on "richness of data rather than frequencies" (Tracy 2010), and to explore a range of opinions and different representations of an issue. As seen below in this section, all of these were sufficiently achieved with our sample size.

Our interview participants were diverse in terms of gender (6 male, 3 female) and academic background (4 graduate students, 2 undergraduate students, and 3 staff or faculty). Participants also came from a range of fields, including history, sociology, learning design, informatics, and computer science. Table 2 reports detailed demographic information about the interviewees.

During the interview, one of the questions asked to the participants was "Can you share any strategies that you believe can induce biased output from LLMs? What techniques did you try but failed?" To identify the themes in participants' responses to Q2.5, we employed an inductive approach to perform thematic analysis (Braun and Clarke 2006). Two trained researchers independently reviewed the transcripts in detail, searching for patterns and meaning within the data, and each researcher independently identified themes. After gaining initial insights, the two researchers discussed with the authors to refine their understanding and gain meaningful insights. We compared the themes identified by each researcher, merging them through discussion, and any discrepancies were resolved through further discussion, to reach a final consensus on the main themes for the strategies to induce bias. Below, we highlight the results of this thematic analysis.

S1. Role Playing Participants often assigned specific roles or personas to LLMs to guide their response towards a biased viewpoint. The idea is to influence the biased outputs by framing it within a certain role that might inherently carry a bias or a specific perspective. For example, participants mentioned:

> "Telling it to have a specific personality...giving it that sort of personality can definitely influence its response." - P4

> "It's just like assigning role like in the start of the conversation. For example, let's say, I want to send pieces of my dissertation to get some revisions, I usually say like, you are my assistant editor, helping me to revise my dissertation" - P9

S2. Hypothetical Scenario One participant created decision-making scenarios, where they designed prompts that forced the model to make a definitive choice rather than a providing a balanced response with pros and cons. For example, they asked the model to make a hiring decision between two candidates with different attributes, setting up conditions to see if the model would show bias against one group, such as based on age or disability. P4 described:

> "I was trying to get it to be biased based on age. So, I asked it to compare a 20-year old and a 67 year old for a task of data analysis...,I said, only list your choice and only make one choice. It is really trying to narrow it into a very specific decision, making [an] environment where it can't equivocate, and the model can't give a list of pros and cons." - P4

S3. Using Human Knowledge Another strategy a participant reported was to ask the GenAI model questions on topics that the participant was already very knowledgeable about, such as religious studies or historical events. By comparing the AI's responses to their well-informed understanding, they could detect bias.

> "The best way to do it would be asking it about something that you know a lot about....I have a masters'

degree in religious studies, so I have a lot of background in that area, better than the surface level internet stuff...,you can do it if you know more than one side of that story." - P3

S4. Leading Questions on Controversial Topics One of the strategies involved prompting the GenAI model with controversial or politically charged questions to observe if it would provide inconsistent or partial responses. For example, participants specifically used prompts about politicians (Trump or Biden), asking for details on their false statements to check if the AI would treat them differently based on the party affiliation. They also experimented with prompts about social and cultural issues to see if the AI would show any partiality or bias in its responses.

> "My strategy would be to understand what is the perspective of GPT or large language model. What side of story does it say so? Is it bias towards anything or any country? Then I would go for cultural norms, for example, cultural and social norms, and then..sticking on topics that are kind of controversial or biased" - P8

S5. Probing Biases in Under-Represented Groups A few participants mentioned a strategy which involved identifying areas where there is likely to be a lack of representation in the training data. For example, one participant focused on generating outputs related to groups that are underrepresented in mainstream literature, such as the LGBTQ+ community. By asking the model to write multiple love stories, they expected the output to skew toward heterosexual narratives. P2 mentioned tweaking the wording slightly to see how the model responded. For example, they changed "academics winning awards" to "academics winning awards at a computer science conference" to examine if the model produced gender bias.

> "I was thinking, what are the thing are underrepresented in the real world, and what are the data that may not be so common in AI's training data...I just asked ChatGPT to write ten love stories. It produced and confirmed my assumption that homosexual or LGBTQ+ community is less presented in the ten stories" - P1

> "[I asked to] show me a group of academics at a conference. I didn't ask about the winning awards and then academics at a computer science conference. There were some female and male in both [pictures], but more males in the computer science. But then, when I asked about awards, it was solely, exclusively males." - P2

S6. Feeding False Information One participant mentioned that they feed AI with false information to generate biased outputs.

> "You feed AI with false information...if it [LLMs] says a truth, and I say intentionally, no you're wrong, it apologizes, and....next time it gives a better answer..manipulating it with like feeding false information..not false necessarily, but just like not the entire reality, but like one narrow part of the reality" - P2

S7. Pretending as Research Purpose A participant discovered that framing the task as scholarly or research-oriented could bypass the model's content filters, allowing it to generate potentially biased outputs. For example, P9 said:

> "If you just like want to produce something that is likely seriously problematic...let's say you are a scholar studying in this topic...It buys that argument" - P9

8 Suggestions to Mitigate Bias

In addition, we asked participants about strategies or techniques which should be used by LLM designers to mitigate the dangers of biased LLM outputs. The most frequently mentioned suggestion was incorporating a wider range of voices and perspectives in the training data, particularly those from marginalized groups, as well as a diverse array of languages, cultures, social contexts, and countries. Participants believe including data from underrepresented groups could reduce bias-related problems. Beyond diversifying training data, participants suggested several additional strategies:

- **Implementing a Robust Classification Filter:** P8 mentioned that employing a filter to screen outputs before presenting them to users could prevent biased content from reaching the end-user.
- **Conducting Extensive Testing:** P5 mentioned the need for rigorous testing to identify and correct areas where models produce biased outputs.
- **Continuous Updating:** P7 mentioned continuously updating models to reflect current societal values and realities, rather than allowing them to rely on outdated perspectives.
- **Educating Users:** P7 mentioned educating users about the limitations and biases embedded in AI models, emphasizing transparency and explainability.
- **Providing References of Information:** P3 mentioned the importance of offering specific references or citations, similar to Co-Pilot, allowing users to verify and understand the information provided by AI models.
- **Monitoring and Regulation:** P2 mentioned the need for regulations or monitoring to ensure that the data used for training is balanced and representative.

9 Discussion

Our findings highlight the unique value of analyzing how non-expert users perceive and elicit bias in generative AI systems. While existing studies have focused on uncovering bias in LLMs (Liu et al. 2024; Dong et al. 2024; Anil et al. 2024; Zou et al. 2023), they have largely taken an expert-centered perspective, e.g., Dong et al. (2024) used

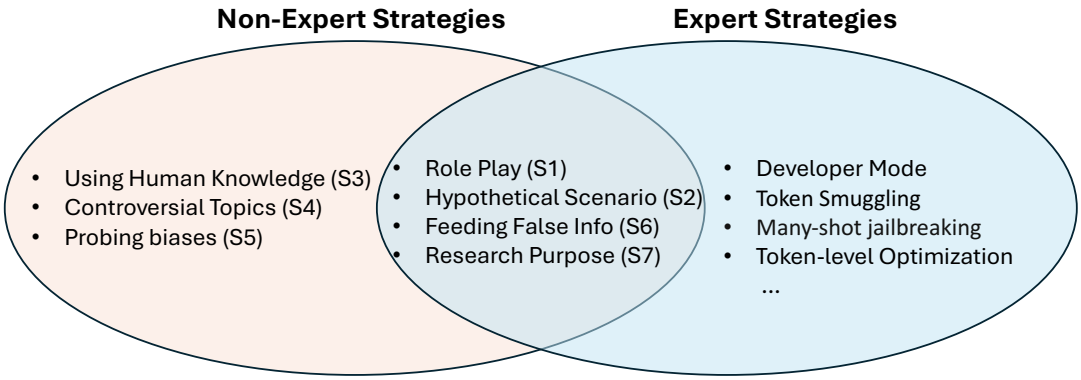


Figure 3: Venn diagram comparing expert and non-expert strategies in eliciting biased outputs from LLMs. While both groups use strategies like role-playing (overlap), non-experts leverage intuitive social knowledge, while experts focus on exploiting model vulnerabilities.

sophisticated techniques like conditional generation probing to elicit biased outputs; similarly, Liu et al. (2024) surveys methods developed by experts for eliciting biases. In contrast, our study adopts a user-centric lens by examining how everyday users, with varying disciplinary and demographic backgrounds, identify and interact with biases embedded in GenAI systems.

Why study non-experts? The rationale for focusing on non-experts is twofold. First, non-expert users now constitute the majority of interactions with LLMs via consumer-facing tools such as ChatGPT, Gemini, and Llama. Therefore, understanding the behavior of non-experts is therefore essential for assessing the real-world risks and societal impact of these technologies. Second, unlike experts who are trained to think in terms of model architecture and adversarial robustness, non-experts may rely on intuitive heuristic strategies to elicit bias which may differ significantly from expert strategies.

Comparison between Non-expert and Expert Strategies In fact, this difference between expert and non-expert strategies is illustrated in Figure 3, which contrasts the different (vs similar) strategies used by experts and non-experts. Interestingly, both expert and non-expert users leverage some common techniques such as role playing and pretension (S1, S7) (e.g., eliciting biased content by asking the LLM to adopt the persona of a journalist or a scholar who is focused on writing an article on some sensitive topic), and manipulating context (S2, S6) (e.g., framing prompts by creating hypothetical trolley-problem type binary scenarios, or as emotionally charged scenarios). However, the underlying motivations and approaches differ significantly. For example, expert users are primarily motivated by the goal of uncovering safety flaws in LLM design and deployment, with the intention of fixing these flaws. As a result, experts approach to these problems is often framed by a "jailbreak mindset", in which eliciting harmful or biased content is a deliberate, strategic act aimed at stress-testing the boundaries of LLM effectiveness and robustness. These efforts are typically conscious, systematic, and technically informed, designed to expose specific vulnerabilities that developers can then address through model updates or content filters. For example, as shown in Figure 3, there are several different kinds of jailbreaking strategies used by experts (e.g., token smuggling, token-level optimization, many-shot jailbreaking, etc.) that are never used by non-expert users.

On the other hand, non-experts often leverage their lived experiences, and intuitive understanding of social inequality and cultural awareness (S3, S4, S5) to design heuristic bias elicitation strategies. Strategies such as invoking personal knowledge of underrepresented communities (S3), probing moral or politically divisive topics (S4), and exposing representational gaps through repeated queries (S5) demonstrate how non-expert users identify vulnerabilities not through advanced technical manipulation, but through social and intuitive reasoning. For instance, one of the winning entries in the competition drew on the participant's intuitive understanding of the relative lack of gender diversity among computer science faculty. The winner crafted a prompt designed to reveal whether the LLM would similarly reflect this disparity, ultimately exposing that the model's interpretation of "successful" computer science faculty was uniformly associated with the cis-male gender (as compared to faculty in the liberal arts).

Our findings in this paper thus complement the existing body of literature on LLM bias by providing unique insights from a non-expert perspective of eliciting biases from LLMs and GenAI tools in everyday usage scenarios, which is the more pervasive use case of GenAI. Given that our findings reveal non-experts employ several distinctive strategies for eliciting bias from LLMs (S3, S4, and S5 in Section 7), this opens up valuable new avenues for expanding and diversifying red-teaming practices. Furthermore, in an era full of conversations about AI governance and regulation (which is often informed by expert-driven recommendations), it is crucial to engage non-experts in conversations about LLM harm and governance, so that policymakers can account for non-adversarial but equally effective bias elicitation strategies used by non-experts. Without such participatory mechanisms, the ethical risk remains high: if LLM evaluation remains solely in the hands of experts, we risk developing

safety protocols that are blind to the actual harms experienced in daily, unmonitored usage.

Finally, our analyses underscore the utility of crowd-based approaches for conducting audits of LLM tools that are flooding the marketplace. Hosting a competition represents a new paradigm for incentivizing crowd-sourcing of audits to continually detect systematic biases at scale in the ever-emergent space of LLMs, allowing for ever more creative production of prompts. More than the specific biases uncovered and the prompts rewarded, the competition represents a viable mechanism for discovering new prompt engineering techniques for evaluating fairness and bias in a variety of LLM-based tools.

10 Conclusion

As LLMs become increasingly integrated into everyday applications, understanding how non-expert users perceive and interact with bias is essential for responsible AI development. This paper presented findings from a university-level competition designed to elicit biased outputs from GenAI tools through user-generated prompts. Our analysis revealed a set of reproducible biases across multiple categories, importantly, we found significant differences between the robustness of different LLMs to generating biased content. Further, we conducted qualitative interviews with nine competition participants to understand the different strategies used by them to elicit biased content from LLMs - in total, our results identify seven distinct bias elicitation strategies used by non-expert users. These insights have significant implications for red-teaming efforts, the design of mitigation strategies, and conversations about AI regulation and governance.

11 Limitations

This paper analyzes the results of a university-level competition to reflect how non-expert users perceive and interact with bias in LLMs and GenAI tools. We acknowledge that our study participants are limited to individuals affiliated with the university that hosted this competition, who possess or are pursuing a college degree. While we categorize all our participants as "non-experts" in the technical AI sense, we recognize that many brought domain-specific expertise (e.g., in history, sociology, or computer science) that informed their critique. Thus, our results may represent a skewed view of bias in LLMs and GenAI tools, and may not generalize to a broader user base. Extending these findings to wider populations (with a greater diversity in disciplinary training) would be an important part of future work.

At the same time, LLMs are continuously being updated by their parent companies, and thus, the amount of bias that they output in response to a single query may vary (hopefully in a negative direction) over time. At the same time, societal perceptions of bias can keep evolving in response to political changes or pressures, etc. In this light, analyzing LLM bias at a single point of time may be a good starting point, but is insufficient in the long term. Thus, it is important to design longitudinal studies for tracking how public perceptions of bias evolve as LLMs improve.

Finally, our analysis centers on the examination of bias in generative AI systems. While recent studies have demonstrated the presence of harms associated with these models (Dev et al. 2022; Blodgett et al. 2022; Ghosh et al. 2024), we specifically focus on bias and therefore do not engage with broader harm frameworks in this work.

Positionality Statement

The authors have approached this research as interdisciplinary scholars situated at the intersection of computer science, NLP, mediated communication, and AI ethics. Thus, our understanding of "bias" is informed by both technical definitions in machine learning and critical perspectives from social sciences, which view bias as a structural and normative phenomenon shaped by historical and cultural forces. As researchers designing and evaluating the competition, we recognize that our own positionality influences how we interpret what constitutes a "bias" and how we categorize user strategies. As much as possible, we have made efforts to foreground participant perspectives, e.g., rather than imposing our normative definition of bias, we crowdsourced our definition of biases by centering diverse participant perspectives to arrive at a definition of bias for this paper. Nevertheless, we do acknowledge that our analyses inevitably reflect our own beliefs centered around participatory ethics, inclusive design, and sociotechnical accountability.

Ethical Consideration

Although this paper aims to understand and mitigate bias in LLMs, we acknowledge that, under unlikely circumstances, malicious users could potentially exploit the strategies discussed in this paper to elicit unwanted model behavior. This potential risk underscores the importance of ongoing research and development in responsible AI practices.

Acknowledgments

The Bias-a-thon competition reported in this paper was sponsored by the Center for Socially Responsible Artificial Intelligence (CSRAI) at Penn State University. Dr. Sundar is supported by MSIT(Ministry of Science, ICT), Korea, under the Global Scholars Invitation Program (RS-2024-00459638) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation)

References

Akter, S.; McCarthy, G.; Sajib, S.; Michael, K.; Dwivedi, Y. K.; D'Ambra, J.; and Shen, K. N. 2021. Algorithmic bias in data-driven innovation in the age of AI.

Anil, C.; Durmus, E.; Rimsky, N.; Sharma, M.; Benton, J.; Kundu, S.; Batson, J.; Tong, M.; Mu, J.; Ford, D. J.; et al. 2024. Many-shot jailbreaking. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Bellamy, R. K.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilović, A.; et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5): 4–1.

Bird, S.; Dudík, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; and Walker, K. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Birkstedt, T.; Minkkinen, M.; Tandon, A.; and Mäntymäki, M. 2023. AI governance: themes, knowledge gaps and future agendas. *Internet Research*, 33(7): 133–167.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Blodgett, S. L.; Liao, Q. V.; Olteanu, A.; Mihalcea, R.; Muller, M.; Scheuerman, M. K.; Tan, C.; and Yang, Q. 2022. Responsible language technologies: Foreseeing and mitigating harms. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–3.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101.

Chouldechova, A.; and Roth, A. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

Coates, D. L.; and Martin, A. 2019. An instrument to evaluate the maturity of bias governance capability in artificial intelligence projects. *IBM Journal of Research and Development*, 63(4/5): 7–1.

Cramer, H.; Garcia-Gathright, J.; Springer, A.; and Reddy, S. 2018. Assessing and addressing algorithmic bias in practice. *Interactions*, 25(6): 58–63.

Czarnowska, P.; Vyas, Y.; and Shah, K. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9: 1249–1267.

Dai, S.; Xu, C.; Xu, S.; Pang, L.; Dong, Z.; and Xu, J. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6437–6447.

Danks, D.; and London, A. J. 2017. Algorithmic Bias in Autonomous Systems. In *Ijcai*, volume 17, 4691–4697.

Debenedetti, E.; Rando, J.; Paleka, D.; Florin, S. F.; Albastroiu, D.; Cohen, N.; Lemberg, Y.; Ghosh, R.; Wen, R.; Salem, A.; et al. 2024. Dataset and Lessons Learned from the 2024 SaTML LLM Capture-the-Flag Competition. *arXiv preprint arXiv:2406.07954*.

Dev, S.; Sheng, E.; Zhao, J.; Amstutz, A.; Sun, J.; Hou, Y.; Sanseverino, M.; Kim, J.; Nishi, A.; Peng, N.; et al. 2022. On Measures of Biases and Harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, 246–267.

Dong, X.; Wang, Y.; Yu, P. S.; and Caverlee, J. 2024. Disclosure and mitigation of gender bias in llms. *arXiv preprint arXiv:2402.11190*.

Eicher, J. E.; and Irgolič, R. F. 2024. Reducing Selection Bias in Large Language Models. arXiv:2402.01740.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–79.

Ghosh, S.; Venkit, P. N.; Gautam, S.; Wilson, S.; and Caliskan, A. 2024. Do Generative AI Models Output Harm while Representing Non-Western Cultures: Evidence from A Community-Centered Approach. *arXiv preprint arXiv:2407.14779*.

Gupta, V.; Narayanan Venkit, P.; Wilson, S.; and Passonneau, R. 2024. Sociodemographic Bias in Language Models: A Survey and Forward Path. In Faleńska, A.; Basta, C.; Costa-jussà, M.; Goldfarb-Tarrant, S.; and Nozza, D., eds., *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 295–322. Bangkok, Thailand: Association for Computational Linguistics.

Huang, D.; Bu, Q.; Zhang, J.; Xie, X.; Chen, J.; and Cui, H. 2023. Bias assessment and mitigation in llm-based code generation. *arXiv preprint arXiv:2309.14345*.

Kordzadeh, N.; and Ghasemaghaei, M. 2022. Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3): 388–409.

Kudina, O.; and van de Poel, I. 2024. A sociotechnical system perspective on AI. *Minds and Machines*, 34(3): 21.

Liu, Y.; Deng, G.; Xu, Z.; Li, Y.; Zheng, Y.; Zhang, Y.; Zhao, L.; Zhang, T.; Wang, K.; and Liu, Y. 2024. Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study. arXiv:2305.13860.

Mazeika, M.; Hendrycks, D.; Li, H.; Xu, X.; Hough, S.; Zou, A.; Rajabi, A.; Yao, Q.; Wang, Z.; Tian, J.; et al. 2023. The Trojan Detection Challenge. *Proceedings of Machine Learning Research*, 220: 279–291.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021a. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6).

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021b. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Navigli, R.; Conia, S.; and Ross, B. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *J. Data and Information Quality*, 15(2).

O'neil, C. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

O'reilly, M.; and Parker, N. 2013. 'Unsatisfactory Saturation': a critical exploration of the notion of saturated sample sizes in qualitative research. *Qualitative research*, 13(2): 190–197.

Rando, J.; Croce, F.; Mitka, K.; Shabalin, S.; Andriushchenko, M.; Flammarion, N.; and Tramèr, F. 2024. Competition report: Finding universal jailbreak backdoors in aligned llms. *arXiv preprint arXiv:2404.14461*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rozado, D. 2023. The political biases of chatgpt. *Social Sciences*, 12(3): 148.

Rutinowski, J.; Franke, S.; Endendyk, J.; Dormuth, I.; Roidl, M.; and Pauly, M. 2024. The Self-Perception and Political Biases of ChatGPT. *Human Behavior and Emerging Technologies*, 2024(1): 7115633.

Saleiro, P.; Kuester, B.; Hinkson, L.; London, J.; Stevens, A.; Anisfeld, A.; Rodolfa, K. T.; and Ghani, R. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.

Schulhoff, S.; Pinto, J.; Khan, A.; Bouchard, L.-F.; Si, C.; Anati, S.; Tagliabue, V.; Kost, A.; Carnahan, C.; and Boyd-Graber, J. 2023. Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global prompt hacking competition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4945–4977.

St. Pierre, E. A.; and Jackson, A. Y. 2014. Qualitative data analysis after coding.

Team, G. 2024. Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Tracy, S. J. 2010. Qualitative quality: Eight "big-tent" criteria for excellent qualitative research. *Qualitative inquiry*, 16(10): 837–851.

Venkit, P. 2023. Towards a holistic approach: Understanding sociodemographic biases in nlp models using an interdisciplinary lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 1004–1005.

Venkit, P. N.; Gautam, S.; Panchanadikar, R.; Huang, T.-H.; and Wilson, S. 2023. Nationality Bias in Text Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 116–122.

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.

Zhang, S.; Bao, Y.; and Huang, S. 2024. EDT: Improving Large Language Models' Generation by Entropy-based Dynamic Temperature Sampling. arXiv:2403.14541.

Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.