Zhaoming Liu*

Cultural Bias in Large Language Models: A Comprehensive Analysis and Mitigation Strategies

https://doi.org/10.1515/jtc-2023-0019
Received December 26, 2023; accepted March 18, 2024; published online September 16, 2024

Abstract: This paper delves into the intricate relationship between Large Language Models (LLMs) and cultural bias. It underscores the significant impact LLMs can have on shaping a more equitable and culturally sensitive digital landscape, while also addressing the challenges that arise when integrating these powerful AI tools. The paper emphasizes the immense significance of LLMs in contemporary AI research and applications, underpinning many systems and algorithms. However, their potential role in perpetuating or mitigating cultural bias remains a pressing issue warranting extensive analysis. Cultural bias stems from various intertwined factors; the following analysis categorizes cultural bias shaping LLMs into three dimensions: data quality, algorithm design, and user interaction dynamics. Furthermore, the impacts of LLMs on cultural identity and linguistic diversity are scrutinized, highlighting the interplay between technology and culture. The paper advocates responsible AI development, outlining mitigation strategies such as ethical guidelines, diverse training data, user feedback mechanisms, and transparency measures. In conclusion, the paper emphasizes that cultural bias in LLMs is not solely a problem but also presents an opportunity. It can enhance our awareness and critical understanding of our own cultural biases while fostering curiosity and respect for diverse cultural perspectives.

Keywords: LLMs; AI; cultural bias; impact and mitigation strategies

1 Introduction

In an era characterized by the rapid advancement of artificial intelligence (AI) technologies, Large Language Models (LLMs) have emerged as transformative tools

***Corresponding author: Zhaoming Liu**, Shanghai University, Shanghai, China,
E-mail: liuzhaoming@shu.edu.cn. https://orcid.org/0009-0006-9104-3859

with far-reaching implications, underpinning many contemporary AI systems and applications (Bommasani et al., 2021). These powerful models, characterized by massive neural architectures and impressive language generation capabilities, hold significant importance in contemporary AI research. Their ubiquitous presence in everyday digital interactions, from virtual assistants to online content, underscores their potential influence on shaping cultural norms, perceptions, and identities.

LLMs, equipped with profound linguistic proficiency and generative capabilities, have revolutionized numerous sectors, including natural language understanding and content generation (Thoppilan et al., 2022). However, they also pose the risk of perpetuating societal biases and assumptions ingrained in their training data (Bender et al., 2021). Of particular concern is their potential to reinforce cultural bias, necessitating thorough examination. This paper aims to comprehensively explore the intricate relationship between LLMs and cultural bias, focusing on the impact, manifestations, and mitigation strategies. In an increasingly interconnected world, where diversity and cultural sensitivity are paramount, understanding the nuances of this relationship is pivotal for both AI researchers and society at large.

Based on LLMs, combining image understanding and generation techniques, AI image generators made it possible to create links between text and images but also pose a more pronounced risk of cultural bias. Notably, a report by The Washington Post introduced three prominent AI image generation models: DALL-E 2, Stable Diffusion, and Midjourney, and their reflection and amplification of biases and stereotypes regarding race, gender, and profession (Tiku et al., 2023). A reader named Carl commented:

> *I'll use my own ethnicity as an example. If I prompt with, "Show me someone who's Jewish," the image generator is likely going to include elements that make someone unambiguously Jewish, like a kipah (skullcap) or a tallit (prayer shawl). At that point, we want to say that the AI is stereotyping because most Jews don't wear a kipah at all times, and prayer shawls are worn only while praying. But what would training data look like for the AI to be able to come back with correct images for prompts like, "Show me an ordinary Jewish person on a city street," or "Show me someone who might be an ordinary Jewish person on the street?" The training data and the AI relying on it are cutoff from elements that are distinctively Jewish, and I'm going to get an ordinary person. To-date, AI has sidestepped this problem so that it can appear very successful, albeit problematically so, by always returning something unambiguous, at the expense of often making offensive choices in charged contexts.* The AI *achieves low ambiguity by "taking sides" on charged issues.*

The reader's comment is a critique of the current state of LLMs, especially image generation, and how it may produce and reinforce cultural bias. The reader argues that the underlying problem of generative AI is the handling of ambiguity, which is the uncertainty or vagueness of meaning or interpretation. The reader uses his own ethnicity, Jewish, as an example to illustrate how generative AI may generate images

that are either too stereotypical or too generic, depending on the prompt. The reader claims that generative AI models have sidestepped this problem by always returning something unambiguous but at the cost of making offensive choices in charged contexts.

We analyze the generative AI image case just as a sub-example here, exploring how it reflects and exacerbates biases. Additionally, it's obvious that LLMs also face similar challenges, although they cover a wider domain of language generation.

The topics of LLMs and cultural bias are both important and intricate, demanding careful and critical examination. One possible way to think about them is to consider the following questions: What are the sources and influences of cultural bias in LLMs? Is it mainly from the data, the algorithm, or the output, or a combination of them? What are the effects and impacts of cultural bias in generative AI? Is it harmful or beneficial, or both, to the users, the creators, and society? What are the responsibilities and roles of the users, the creators, and society in preventing or mitigating cultural bias in LLMs? How can they collaborate and communicate effectively and ethically?

2 Literature Review

LLMs have demonstrated remarkable proficiency in language understanding and generation (Chowdhery et al. 2023; Touvron et al. 2023), increasingly serving users from diverse cultural backgrounds.

Recent research has revealed that LLMs can encode biases that may have harmful consequences for downstream tasks (Bender et al., 2021; Dev et al., 2021; Kumar et al., 2022). These biases can emerge at various stages of LLM development, including data annotation, training data selection, model architecture design, and research methodology (Hovy and Prabhumoye, 2021; Sap et al., 2022).

Ideally, LLMs should be capable of understanding and respecting the cultural norms and beliefs of different communities, producing culturally-relevant content when generating text (Naous et al., 2023). However, AI systems often reflect the cultural values of Western, Educated, Industrialized, Economically Affluent, and Democratic societies (Prabhakaran et al., 2022). Modern LLMs exhibit significant cultural bias, defaulting to Western culture and norms even when operating in non-Western languages. For instance, ChatGPT, a prominent dialogue agent, strongly aligns with American culture when prompted with American context, but adapts less effectively to other cultural contexts (Cao et al., 2023). If an LLM disproportionately represents certain opinions, it risks imposing potentially undesirable effects, such as promoting hegemonic worldviews and homogenizing

people's perspectives and beliefs (Bender et al., 2021; Blodgett et al., 2020). The same research also indicates that English prompts reduce the variance in model responses, flattening cultural differences and biasing them towards American culture (Cao et al., 2023). This is likely because ChatGPT was trained on a vast multilingual corpus that inherently embeds biases and cultural nuances (Alshater 2022; McGee 2023).

LLMs can inherit harmful and stereotypical language from their training data, leading to downstream representational harms (Blodgett et al., 2020). Additionally, they internalize, spread, and amplify toxic language (e.g., offensiveness, hate speech, and insults) and social biases (e.g., stereotypes towards people with a particular demographic identity) existing in the training corpora (Gehman et al., 2020; Sheng et al., 2021).

One major challenge in mitigating bias in LLMs is constructing fairness datasets that comprehensively represent the diverse cultures and languages across the world (Ramesh et al., 2023). Researchers are exploring a variety of methods to reduce cultural bias in LLMs, such as training on more diverse datasets and developing new evaluation and mitigation approaches (Abid et al., 2021a, 2021b; Ahn and Oh, 2021; Cao et al., 2022; Nadeem et al., 2020; Nozza et al., 2021; Sheng et al., 2019).

The literature review offers valuable insights into existing research on the influence of LLMs on cultural bias. While prior research has established a solid foundation, certain limitations underscore the necessity for further investigation in this domain.

One limitation of current research is the prevalence of case-specific analyses, focusing on individual instances of cultural bias in LLM-generated content. While these case studies yield crucial insights, they often lack a comprehensive examination of broader patterns and systemic issues associated with cultural bias. Future research should aim to synthesize individual cases into a more holistic understanding.

3 Factors Contributing to Cultural Bias in LLMs

Cultural bias in LLMs can originate from various factors across domains such as data quality, algorithmic structure, and societal context. These factors are interconnected and mutually reinforcing. In the subsequent discussion, we will explore the critical elements contributing to this bias: the quality of data utilized, the design of algorithms, and user interaction dynamics.

3.1 Data

Data serves as the cornerstone of LLMs, shaping their learning and generative capacities. However, data often exhibits incompleteness, imbalance, lack of representation, or inaccuracy, mirroring human biases in collection, processing, and labeling. For instance, if a generative text model is trained on texts predominantly authored by male, white, and Western writers, it risks perpetuating their perspectives and stereotypes, neglecting diversity. Data bias undermines model quality, reliability, fairness, user satisfaction, creator confidence, and societal trust. Cultural bias in training data arises from two factors: Skewed Data and Historical Inherited Data.

Skewed Data refers to one-sided data lacking cultural diversity. Training data may inadequately represent the target population or domain, resulting in AI reflecting biases inherent in the dominant source. Limited exposure to diverse cultural perspectives renders the AI culturally insensitive and inaccurate in portraying different cultures. Conversely, if training data primarily originates from one culture, it may amplify embedded biases, potentially propagating hate speech and false assertions (Nadis, 2022).

Historical Inherited Data encompasses culturally outdated data and past prejudices. Biases entrenched in historical data can infiltrate AI models, perpetuating obsolete stereotypes. For instance, an AI trained on historical texts might associate certain occupations or roles with specific genders or ethnicities. A compelling real-world illustration of this phenomenon is the gender bias prevalent in job advertisements. This bias emanates from historical societal assumptions and stereotypes about gender roles, often reflected in the language of historical job postings. Evidence suggests that gendered wording in job advertisements sustains gender inequality. Studies examining the use of masculine and feminine language in job ads reveal that advertisements for male-dominated fields tend to employ more masculine language, dissuading women from applying and indicating lower belongingness (Gaucher et al., 2011). Training AI on such data could result in biased job descriptions, unwittingly reinforcing gender stereotypes.

3.2 Algorithm Design

Algorithms are central to LLMs, shaping their design, implementation, and functionality. However, they may also contain flaws or biases, reflecting human assumptions and constraints in their development, testing, and evaluation processes. Cultural bias in algorithm design can arise for two primary reasons: Choice of Metrics and Lack of Transparency.

Choice of metrics involves selecting appropriate and meaningful measures to assess the performance and fairness of AI algorithms. Different metrics may capture various aspects of an algorithm's behavior and impact, each with its trade-offs and implications. For instance, the COMPAS algorithm, utilized to predict defendants' likelihood of reoffending within the US court system, exhibited twice as many false positives for recidivism among black offenders (45 %) compared to white offenders (23 %) due to data, model selection, and overall algorithm creation process (Datatron, 2019), indicating potential racial bias in metric choice. Real-life examples of biased metrics extend to healthcare, where data representation of women or minority groups can influence predictive AI algorithm metrics. For example, computer-aided diagnosis (CAD) systems have lower accuracy rates for black patients compared to white patients (IBM Data and AI Team, 2023). Similarly, if training data is predominantly male, metrics may prioritize accuracy and precision for males, neglecting recall and sensitivity females, potentially resulting in underestimation or misdiagnosis of female patients' risks or conditions. Thus, metric selection should consider diverse representation of groups in data and utilize metrics capturing algorithm performance and fairness across all groups.

Lack of Transparency refers to situations where AI model internals or logic remain unknown or inexplicable to users or humans. This opacity often characterizes complex machine learning systems, termed "black box" systems (Kuang, 2017). Lack of Transparency presents various challenges, including difficulties in identifying and rectifying potential biases in AI models, establishing user trust in the model, and adhering to legal or ethical standards requiring AI model explainability (Hilliard, 2023). Without transparency, AI models may exert unfair or discriminatory impacts on individuals or groups. Notably, Apple's new credit card business faced allegations of sexist lending models (Vigdor 2019), while Amazon discontinued an AI hiring tool after discovering gender discrimination issues (Dastin, 2018). Thus, enhancing transparency in AI algorithms is crucial for fostering trust, fairness, and accountability in their deployment and usage.

3.3 User Interaction

When discussing "User Interaction" as a source of cultural bias in AI, the text explains how users directly or indirectly influence AI biases through interactions, delineating two main categories: feedback loops and misuse.

Feedback loops deepen biases as AI models interact with the world, gathering additional data reflecting existing biases. Users contribute to this loop by reinforcing biases through interactions or providing positive feedback to outputs reflecting their

own cultural biases. Over time, this creates a feedback loop, further amplifying biases as the AI interprets this feedback as a preference for biased content.

For instance, consider news recommender system (NRS) models, as discussed in research, can experience various feedback loops, such as sampling, individual, feature, model, and outcome feedback loops, affecting different biases and susceptibility to biased outcomes (Pagan et al., 2020, p. 3).

Furthermore, studies indicate that certain AI hiring algorithms based on resume analysis can perpetuate gender and racial biases. For instance, algorithms may favor resumes with masculine-sounding names or keywords associated with traditionally male-dominated fields, creating a feedback loop where qualified candidates from underrepresented groups are less likely to be selected, reinforcing existing biases within the algorithm (Dastin, 2018).

Misuse of AI involves intentionally exploiting AI systems for malicious purposes, such as spreading harmful stereotypes, propaganda, or misinformation. This can occur through feeding biased data to the AI, manipulating its training process, or exploiting vulnerabilities, reinforcing biases and undermining the trustworthiness and fairness of the AI. For example, cybercriminals misuse AI and machine learning for malicious activities like generating fake ads, attacking cybersecurity systems, creating automated malware, or forging identities and content. These activities pose threats to the security and privacy of individuals, businesses, and society, as well as the trustworthiness and fairness of AI (Trend Micro, UNICRI, & Europol, 2023).

Another example involves exploiting AI systems to create and spread deepfakes, audio and visual content manipulated using AI techniques to appear authentic. Deepfakes can be used for fraud, defamation, extortion, harassment, manipulation, or sabotage, posing risks to the reputation and interests of individuals and organizations, as well as societal stability. In 2018, actor and director Jordan Peele created a fake public service announcement featuring a deepfake video of former U.S. President Barack Obama warning about the ease of creating and potential harm from deepfake videos. Since then, the use of AI-generated videos and images for political disinformation has increased globally, appearing in Ukraine, Turkey, and other states (Frase & Daniels, 2023).

4 Impact on Cultural Identity and Diversity

In the preceding part, we explored how LLMs generate different sources of cultural bias. This part extends this discussion to examine their impact on cultural identity and diversity. LLMs wield significant influence on the intricate tapestry of cultural identity and diversity within our interconnected global society. These formidable AI systems serve as central figures in shaping discourse, molding the narratives that

shape perceptions, preservation and sharing of cultures. In this section, we embark on a comprehensive exploration of LLMs' profound impact on cultural identity and linguistic diversity. As we delve into the intricate interplay between technology, culture, and society, we will uncover how LLMs not only reflect the existing perceptions of cultural identity but also have the capacity to reshape them. Moreover, their influence on linguistic diversity, both in terms of preservation and potential alteration, revealing the delicate balance these models navigate in the modern, globalized digital landscape. This section is dedicated to unraveling the nuanced relationship between LLMs and the complex realm of cultural identity and diversity, a fundamental aspect of our contemporary, interconnected world.

4.1 Cultural Identity

Cultural identity, an inherently complex facet of human existence, intricately intertwines language, beliefs, customs, and traditions. Within the sphere of LLMs, the portrayal of cultural identity emerges as a nuanced and profoundly significant challenge.

At its essence, the crux of the matter concerning LLMs revolves around their ability to faithfully represent the diverse tapestry of cultural identities. These identities, far from monolithic, evolve through intricate interplays of historical legacies, geographical contexts, and complex sociopolitical dynamics. In the realm of content generation, LLMs must demonstrate a remarkable capacity to authentically reflect this intricate diversity of identities.

However, the terrain is not without pitfalls. Embedded biases in their design, data, or output may occasionally lead LLMs astray. In their endeavor to generate content, these models may inadvertently perpetuate stereotypes, oversimplify, or distort the nuanced aspects of cultural identities, resulting in incomplete or distorted portrayals that reinforce preconceived notions.

Generating content presents a significant challenge for LLMs in ensuring accurate and respectful cultural representations. To illustrate this challenge, we examine real-world case studies demonstrating how LLMs can inadvertently perpetuate or misconstrue cultural stereotypes and misconceptions, potentially leading to harmful consequences. These case studies underscore the urgency to directly address this issue and develop methods and principles to mitigate the risks of bias, insensitivity, and appropriation in LLMs.

To begin, we delve into a paper by Kotek, Dockum, and Sun (2023), which investigates gender bias and stereotypes in LLMs. They found that LLMs are prone to selecting stereotypical occupations for men and women, amplifying bias beyond what is reflected in people's perceptions or official statistics. Furthermore, they

demonstrated that LLMs can generate sexist and harmful sentences when prompted with gendered words or scenarios. They propose potential strategies to reduce gender bias in LLMs, including using de-biased data, applying gender-swapping techniques, and incorporating human feedback.

Next, we explore a paper by Lee, Montgomery, and Lai (2024), which investigates the phenomenon of homogeneity bias in LLMs. They define homogeneity bias as the tendency of LLMs to portray socially subordinate groups as more uniform than the dominant group. For instance, they observed that LLMs tend to depict African, Asian, and Hispanic Americans as more homogeneous than White Americans, potentially affecting the quality and diversity of generated content. They propose metrics to measure homogeneity bias in LLMs and suggest strategies to mitigate it, such as employing more diverse and representative data and implementing diversity-aware sampling methods.

Following this, we review a paper by Dev et al. (2023), which proposes a method for building stereotype repositories with LLMs and community engagement. They argue that LLMs can generate content that reflects and perpetuates existing stereotypes about various social groups, including race, ethnicity, gender, sexuality, religion, and disability. They propose that creating and maintaining stereotype repositories can help identify and monitor the stereotyping harms of LLMs, providing a resource for developing interventions and evaluations. They emphasize the importance of covering diverse and intersectional identities in stereotype repositories and involving affected communities in the creation and maintenance process.

The portrayal of cultural identity within the realm of LLMs presents a multifaceted and intricate challenge, replete with nuances. Addressing this challenge extends beyond rectifying biases within training data; it necessitates LLMs to authentically and respectfully depict the complex diversity of cultures and identities.

4.2 Linguistic Diversity

As previously discussed, accurately portraying cultural identity poses a multifaceted challenge for LLMs, as they may inadvertently perpetuate stereotypes or misrepresentations without careful oversight. Similarly, ensuring linguistic diversity is crucial for preserving cultural richness and requires LLMs to undergo rigorous design and training.

Language and linguistic diversity, like cultural identity, are deeply intertwined facets of human existence and society. Within the realm of LLMs, accurately representing linguistic diversity also emerges as a nuanced and significant challenge.

At its essence, faithfully depicting linguistic diversity entails LLMs' ability to authentically represent the diversity of languages and dialects. Linguistic forms evolve through intricate interplays of histories, regions, and sociopolitical environments. In content generation, LLMs must demonstrate the capability to authentically reflect this diversity of languages.

However, potential pitfalls abound. Inherent biases in training data may mislead LLMs. In their quest for content generation, models may inadvertently prioritize dominant forms, neglect marginalized dialects, or oversimplify linguistic variations, potentially resulting in skewed or incomplete portrayals that fail to capture the nuances of linguistic diversity. Several cases illustrate such challenges.

For instance, a study by Guo et al. (2023) examined the consequences of training LLMs on synthetic data generated by their predecessors. They found that this practice leads to a noticeable decrease in the diversity of the models' outputs over successive iterations, undermining the preservation of linguistic richness.

As the influence of LLMs extends globally, addressing their safety challenges in multilingual contexts becomes a pressing focus for alignment research. A study by Shen et al. (2024) explores the phenomenon of language barriers, where LLMs trained on high-resource languages struggle to generalize to low-resource languages, and the diverse safety challenges faced by LLMs across various languages. Surprisingly, while training with high-resource languages enhances model alignment, training in lower-resource languages yields marginal improvement. This suggests that the bottleneck of cross-lingual alignment originates from the pretraining phase.

An article by Lappin (2023) evaluated the extent to which LLMs shed light on human cognitive abilities in language learning and linguistic representation. While they surpass human performance in numerous linguistically significant Natural Language Processing (NLP) tasks, their reliance on vast amounts of data for language acquisition exceeds human capabilities. It remains unclear whether they acquire and encode linguistic knowledge akin to human cognitive processes. This raises concerns about the potential implications of LLMs on linguistic diversity, as their learning mechanisms may not fully capture the intricacies and nuances of diverse languages and dialects.

In conclusion, LLMs face a significant challenge in ensuring accurate and respectful cultural and linguistic representations. Language and linguistic diversity, like cultural identity, are complex aspects of human life and society that require LLMs to undergo rigorous design and training to faithfully represent them. However, LLMs may encounter various pitfalls, including inherent biases in training data, language barriers, and cognitive limitations, potentially resulting in skewed or incomplete portrayals of linguistic diversity.

5 Mitigating Cultural Bias in LLMs

Amidst the ongoing evolution and utilization of LLMs, addressing cultural bias has emerged as a paramount issue. Recognizing the significant impact of these models on content generation and cross-cultural interactions, there is an urgent need to explore robust strategies for mitigating such biases. This section aims to scrutinize various approaches intended to alleviate and rectify cultural biases embedded within LLMs.

5.1 The Role of Ethical Guidelines in Mitigating Cultural Bias

Primarily, they provide a structured framework for identifying and rectifying biases within LLMs. These guidelines emphasize the paramount importance of continuous monitoring and auditing of model behavior, enabling the detection and mitigation of bias. By offering a systematic approach, they empower developers to proactively address issues related to cultural bias, such as the choice of metrics and the lack of transparency in algorithm design, as discussed in Chapter 3.2. For example, ethical guidelines may require developers to disclose data sources, training methodologies, and potential biases associated with their models, as well as to use appropriate and fair evaluation metrics that reflect the diversity of users and contexts.

Furthermore, ethical guidelines underscore the importance of user involvement and education. Developers should proactively solicit and value feedback from people and communities who may be affected by cultural bias and educate users about the strengths and weaknesses of LLMs. This way, developers can ensure that LLMs are aligned with diverse cultural values and expectations, and that users are informed of the possible benefits and harms of using LLMs, as explained in Section 3.3. For instance, ethical guidelines may advise developers to provide transparent and understandable information about how LLMs work, their reliability, and responsible usage, as well as to design user interfaces that prevent abuse and feedback loops.

User engagement and education are crucial for mitigating cultural bias in LLMs.

Conducting user research and testing to understand the expectations, preferences, and concerns of different user groups, particularly those marginalized or underrepresented in the data used to train LLMs (Turner Lee et al., 2019). This can help developers identify and address potential sources of cultural bias and design inclusive LLMs respectful of diverse values and norms.

Establishing feedback and reporting mechanisms to enable users to share opinions, suggestions, and complaints about LLMs, as well as report any issues or problems they encounter (Ghosh, 2023). This can help developers monitor and evaluate the performance and impact of LLMs and improve them based on user feedback.

Educating users about the ethical and social implications of using LLMs, such as potential misuse, abuse, or harm, as well as user rights and responsibilities. This can help users make informed and responsible decisions about when and how to use LLMs and respect the privacy and dignity of others.

Furthermore, ethical guidelines mandate developers to be transparent about their data sources, training methodologies, and potential biases in their models. Liao & Vaughan (2023) argues that transparency is crucial for the responsible development and deployment of LLMs, despite the challenges posed by their complex capabilities, massive architectures, proprietary nature, diverse stakeholders, and evolving public perception.

Transparency and accountability are essential for mitigating cultural bias in LLMs. Transparency entails developers disclosing their data sources, training methodologies, and potential biases in their models, while accountability means developers are responsible for their models' performance and impact. By being transparent and accountable, developers can foster user trust and enable informed decisions when interacting with LLM-generated content. Responsible development practices involve a holistic approach to model creation, including:

Using diverse and representative training data covering various cultural contexts, languages, and demographics to expose LLMs to a wide range of cultural nuances and reduce the risk of bias due to inadequate data representation.

Engaging with affected communities to solicit input and feedback, especially those disproportionately impacted by cultural bias, ensuring LLMs align with diverse cultural needs and sensitivities.

Comprehensive testing and validation of LLMs across various cultural contexts to identify and rectify bias in content generation, ensuring equitable performance across diverse user interactions (Tao et al., 2023).

In summary, ethical guidelines and responsible development practices lay the foundation for more equitable and culturally sensitive AI systems that cater to the diverse needs of our global society. They act as ethical compasses, guiding developers toward bias rectification while fostering transparency.

5.2 A Deep Dive into Diverse Training and Mitigation Techniques for LLMs

Ensuring diversity in training data is pivotal for mitigating cultural bias within LLMs, constituting the cornerstone of responsible LLM development. By exposing

these models to a wide range of cultural contexts, languages, and demographics during training, substantial progress can be achieved in reducing cultural bias.

The inclusion of diverse training data is imperative because LLMs learn from the data they encounter. When training data lacks diversity, biases can inadvertently persist. Conversely, diverse training data reflects the rich tapestry of human cultures, equipping LLMs with the understanding needed to navigate global diversity effectively.

Exposing LLMs to a diverse array of cultural sources heightens the likelihood of generating content that is inclusive and culturally sensitive. This approach not only helps mitigate biases arising from underrepresented cultures but also fosters a deeper appreciation and respect for the cultural diversity present in our world.

While advocating for the positive impact of LLMs trained on diverse data and advocating for inclusive data practices, it's essential to acknowledge and address the potential negative consequences of LLMs trained on biased data. Typically trained on vast amounts of uncurated Internet-based data, LLMs inherit stereotypes, misrepresentations, derogatory language, and other harmful behaviors that disproportionately affect already-vulnerable and marginalized communities (Bender et al. 2021; Dodge et al. 2021). While LLMs often reflect existing biases, they can also amplify them, further solidifying systems of inequity (Benjamin, 2020).

To address the negative impacts of LLMs trained on uncurated Internet-based data, three types of data processing techniques are employed: data augmentation, filtering, and reweighting.

Data augmentation tackles bias in natural language models by diversifying training data, particularly for underrepresented or misrepresented groups. A notable technique, counterfactual data augmentation (CDA), alters words reflecting protected attributes, such as gendered pronouns or job titles, to achieve balance. Lu et al. (2020) pioneered CDA to combat gender bias in occupation descriptions by generating paired sentences subtly differing in gendered terms. Ghanbarzadeh et al. (2023) introduced a novel CDA approach involving masking and predicting gendered words using a language model and fine-tuning with the original label. Alternatively, Dixon et al. (2018) proposed balancing toxicity across groups by incorporating non-toxic examples where toxicity overrepresentation occurs.

Data filtering and reweighting techniques involve modifying, selecting, or reweighting specific examples within a dataset based on properties such as bias levels or demographic information. These methods address the limitations of data augmentation, which may introduce grammatical errors or rely on incomplete word lists.

These techniques fall into two main categories. The first involves selecting a subset of examples to emphasize during fine-tuning. Garimella et al. (2022) curated and filtered text from historically disadvantaged gender, racial, and geographical

groups to diversify the model's understanding during fine-tuning. Similarly, Borchers et al. (2022) proposed techniques for data selection prioritizing underrepresented or low-bias examples; they developed a low-bias dataset of job advertisements by selecting the 10 % least biased examples based on a gendered word list frequency.

The second category involves reweighting instances to de-emphasize them during training. Han et al. (2021) used instance reweighting to equalize the importance of each class during training, assigning weights inversely proportional to both the instance's label and an associated protected attribute. Orgad & Belinkov (2022) suggested methods focused on downweighting examples containing social group information, even without explicit labels. They employed auxiliary classifiers to identify potentially biased examples, either through a shallow model trained on a small subset of the data (Utama et al., 2020) or by assessing the predicted success of a pre-trained model (Orgad & Belinkov, 2022). Their argument was that reducing the influence of stereotypical shortcuts could help mitigate bias during fine-tuning.

6 Challenges in Cultural Bias Mitigation

This exploration encompasses a range of critical aspects and offers insights into the multifaceted challenges and promising future prospects associated with LLMs in the context of mitigating cultural bias. We began by examining data challenges, which focus on the availability, quality, and potential biases present in training data. It is evident that the training data profoundly shapes the behavior of LLMs and their effectiveness in reducing bias.

A subsequent focus was on technological challenges, which are related to the limitations of current LLM algorithms in bias reduction. We delved into the complexities of addressing bias while preserving creative content generation and explored the ethical considerations that arise in this context.

Another area of consideration revolves around regulatory and ethical challenges. This includes the role of government regulations in shaping LLM bias and the ethical dilemmas associated with LLM deployment and representation. As LLMs become integrated into various aspects of society, these challenges gain increasing significance.

6.1 Data Challenges in Cultural Bias Mitigation

The availability and quality of data serve as fundamental determinants in shaping the behavior and efficacy of LLMs regarding bias mitigation.

A critical starting point involves identifying biases within the training data, as these biases often underpin cultural biases in LLMs. Thorough data analysis is essential to recognize and understand these biases (Balayn et al., 2021). This analysis extends beyond overt biases to include subtler, less conspicuous ones that can permeate the data.

One major challenge is the presence of historical biases within existing datasets. Many training corpora have been amassed over decades and may reflect biases that were prevalent at the time of collection (Roselli et al., 2019). Consequently, LLMs can inadvertently perpetuate stereotypes and misconceptions.

Another dimension of data bias relates to underrepresentation (Kuhlman et al., 2020). Certain cultures, languages, and demographic groups may be inadequately represented in training data. This underrepresentation can create an imbalance in how LLMs respond to different cultural contexts, unintentionally favoring majority groups while marginalizing minority voices.

The acquisition of diverse and unbiased data poses its own set of challenges, as we delve into the complexities involved in this process. This includes the difficulties of data collection when aiming to represent a wide spectrum of cultural contexts and languages.

One challenge is the sheer volume of data required. LLMs rely on vast amounts of text data, and collecting a sufficient volume to comprehensively represent global cultural diversity can be a daunting task. Furthermore, data collection efforts must navigate linguistic, ethical, and privacy considerations to ensure the responsible acquisition of diverse data.

6.2 Technological Challenges in Addressing Cultural Bias

This subsection navigates through the intricate technological hurdles encountered when addressing cultural bias within LLMs. While these models have the potential to reshape AI applications, they are not without their limitations and complexities.

A primary technological challenges lies in the limitations of current LLM algorithms in reducing bias. LLMs, like other machine learning models, are trained on vast datasets, and their behavior is shaped by the patterns present in the data. While efforts are made to remove bias during training, complete elimination is often elusive.

The challenge here is twofold. Firstly, biases in training data can be deeply ingrained, making them challenging to identify and rectify. Secondly, the fine line between bias reduction and censorship poses ethical dilemmas. Striking the right balance between reducing bias and preserving the model's ability to generate creative and contextually relevant content is a nuanced endeavor (Park et al., 2023).

In this context, the ethical considerations surrounding bias reduction and creativity in LLMs warrant exploration. Navigating the delicate balance between reducing bias and fostering innovation necessitates meticulous ethical deliberation. Additionally, apprehensions arise regarding who defines and controls the boundaries of bias reduction. Decisions about what constitutes bias and how it should be reduced can have profound implications for freedom of expression and the representation of diverse cultural perspectives.

6.3 Regulatory and Ethical Challenges in Cultural Bias Mitigation

This subsection delves into the regulatory and ethical challenges surrounding the mitigation of cultural bias within LLMs. As these models become increasingly integrated into various facets of society, the need for regulatory frameworks and ethical considerations becomes paramount.

A significant regulatory challenge is navigating the role of government regulations in shaping LLM bias. Governments worldwide are beginning to recognize the significance of addressing bias in AI systems and are formulating regulations and guidelines aimed at ensuring equitable and unbiased AI technologies. However, the challenge here lies in striking a balance between regulatory oversight and stifling innovation. While regulations are essential for curbing harmful biases, they should not hinder the development of creative and valuable AI applications (Chang, 2024). Finding the right equilibrium is a complex task that requires collaboration between policymakers, technologists, and ethicists.

Exploring ethical considerations surrounding LLM deployment and representation reveals dilemmas related to how LLMs should represent and engage with cultural content and expressions. One significant challenge is the risk of cultural appropriation, where LLMs generate content inspired by diverse cultural backgrounds without appropriate context or respect for cultural sensitivities. Addressing this issue through ethical guidelines is crucial to ensure respectful representation.

Furthermore, concerns arise regarding the reinforcement of stereotypes and misconceptions by LLMs. These models have the potential to perpetuate existing stereotypes or inadvertently create new ones. Ethical considerations necessitate active efforts by developers and users to counteract harmful narratives, ensuring that LLM-generated content respects the dignity and authenticity of all cultures.

7 Conclusions

This paper set out to conduct a thorough analysis of the relationship between LLMs and cultural bias. Through an examination of existing literature and a systematic categorization of sources, impacts, and challenges, we explored this issue from multiple perspectives.

Our literature review revealed that while past studies have offered useful insights, a holistic and cross-cutting analysis was still needed. By investigating factors like data, algorithms, and user interactions, we identified how LLMs can introduce and perpetuate biases at different stages.

Cultural bias in LLMs was found to originate from various intertwined influences, including skewed and historically inherited training data, choices in model design and metrics, as well as feedback loops and misuse during deployment. These biases can then manifest in the generation of stereotypical content and the inaccurate portrayal of cultural identities and linguistic diversity.

Additionally, we observed how LLMs present both opportunities and risks regarding representation – while they have potential to spread cultural knowledge, there is a constant threat of appropriation, homogenization and skewed narratives if not properly guided.

To address these issues, we proposed a range of mitigation strategies focusing on transparency, responsible practices, and diversifying training methodology through tools like ethical guidelines, community engagement and unbiased data augmentation.

Despite the progress made, significant challenges still lie ahead in areas like data collection, algorithm design, and regulatory oversight. Achieving fully representative and unbiased LLMs requires extensive, cooperative efforts across technical, social and policy domains.

In conclusion, through this comprehensive analysis, we have provided a deeper understanding of cultural bias in LLMs and its nuanced relationship with technology and society. Most importantly, we emphasized that bias mitigation is not just a problem but an opportunity – one that can enhance self-awareness while fostering appreciation for diverse perspectives in both AI systems and their users. Continued research on this topic is vital for building more equitable and culturally sensitive technologies.

References

Abid, A., Farooqi, M., & Zou, J. (2021a). Large language models associate Muslims with violence. *Nature Machine Intelligence*, *3*(6), 461–463.

Abid, A., Farooqi, M., & Zou, J. (2021b). Persistent anti-Muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)* (pp. 298–306). New York, NY, USA: Association for Computing Machinery.

Ahn, J., & Oh, A. (2021). Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 533–549). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Alshater, M. (2022). Exploring the role of artificial intelligence in enhancing academic performance: A case study of ChatGPT (December 26, 2022). Available at SSRN: https://ssrn.com/abstract=4312358 or http://dx.doi.org/10.2139/ssrn.4312358

Balayn, A., Lofi, C., & Houben, G. J. (2021). Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, *30*(5), 739–768.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (FAccT '21)* (pp. 610–623). New York, NY, USA: Association for Computing Machinery.

Benjamin, R. (2020). *Race after technology: Abolitionist tools for the new Jim code*. Polity.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. (pp. 5454–5476). Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.485.

Cao, Y. T., Sotnikova, A., Daumé III, H., Rudinger, R., & Zou, L. (2022). Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*. (pp. 1276–1295). Seattle, United States: Association for Computational Linguistics.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Wang, W., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Borchers, C., Gala, D. S., Gilburt, B., Oravkin, E., Bounsi, W., Asano, Y. M., & Kirk, H. R. (2022). Looking for a handsome carpenter! debiasing GPT-3 job advertisements. *arXiv preprint arXiv:2205.11374*.

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

Chang, E. Y. (2024). SocraSynth: Multi-LLM Reasoning with Conditional Statistics. *arXiv preprint arXiv: 2402.06634*.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., … Fiedel, N. (2023). PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, *24*(240), 1–113.

Dastin, J. (2018).Rpt-insight-amazon scraps secret ai recruit- ing tool that showed bias against women. Reuters, 2018. https://www.reuters.com/article/amazoncom-jobs-automation/rpt-insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSL2N1WP1RO.

Datatron (2019). Real-life examples of discriminating artificial intelligence. https://datatron.com/real-life-examples-of-discriminating-artificial-intelligence/

Dev, S., Goyal, J., Tewari, D., Dave, S., & Prabhakaran, V. (2023). Building socio-culturally inclusive stereotype resources with community engagement. *arXiv:2307.10514*.

Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J. M., & Chang, K. W. (2021). Harms of gender exclusivity and challenges in non-binary representation in language technologies. *arXiv preprint arXiv:2108.12084*.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018, December). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) (AIES '18). Association for Computing Machinery, New York, NY, USA, 67–73. https://doi.org/10.1145/3278721.3278729

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., … Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Frase, H., & Daniels, O. (2023). *Understanding AI harms: An overview*. Center for Security and Emerging Technology. https://cset.georgetown.edu/article/understanding-ai-harms-an-overview/

Garimella, A., Mihalcea, R., & Amarnath, A. (2022, November). Demographic-aware language model fine-tuning as a bias mitigation technique. *In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pp. 311–319, 2022.

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, *101*(1), 109–128.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). Real Toxicity Prompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 3356–3369). Association for Computational Linguistics. https://aclanthology.org/2020.findings-emnlp.301.

Ghanbarzadeh, S., Huang, Y., Palangi, H., Moreno, R. C., & Khanpour, H. (2023). Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. *arXiv preprint arXiv:2307.10522*.

Ghosh, B. (2023). Ways to Monitor LLM Behavior. Medium. Retrieved from https://medium.com/@bijit211987/ways-to-monitor-llm-behavior-c068fba53932.

Guo, Y., Shang, G., Vazirgiannis, M., & Clavel, C. (2023). The curious decline of linguistic diversity: Training language models on synthetic text. *arXiv preprint arXiv:2311.09807*.

Han, X., Baldwin, T., & Cohn, T. (2021). Balancing out bias: Achieving fairness through balanced training. *arXiv preprint arXiv:2109.08253*.

Hilliard, A. (2023). *What is AI transparency?* Holistic AI. https://www.holisticai.com/blog/ai-transparency

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8), e12432.

IBM Data and AI Team. (2023). Shedding light on AI bias with real world examples. https://www.ibm.com/blog/shedding-light-on-ai-bias-with-real-world-examples/

Kotek, H., Dockum, R., & Sun, D. Q. (2023). Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*. Retrieved from https://api.semanticscholar.org/CorpusID:261276445

Kuang, C. (2017). Can A.I. be taught to explain itself? *The New York Times Magazine*. https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html?_r=0

Kuhlman, C., Jackson, L., & Chunara, R. (2020). No computation without representation: Avoiding data and algorithm biases through diversity. *arXiv preprint arXiv:2002.11836*.

Kumar, S., Balachandran, V., Njoo, L., Anastasopoulos, A., & Tsvetkov, Y. (2022). Language generation models can cause harm: So what can we do about it? An actionable survey. *arXiv preprint arXiv:2210.07700*.

Lappin, S. (2023). Assessing the strengths and weaknesses of Large Language Models. *Journal of Logic, Language and Information*, *15*, 1–12.

Lee, M. H., Montgomery, J. M., & Lai, C. K. (2024). The effect of group status on the variability of group representations in LLM-generated text. *arXiv preprint arXiv:2401.08495*.

Liao, Q. V., & Vaughan, J. W. (2023). Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941*.

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2020). Gender bias in neural natural language processing. *Logic, language, and security: Essays dedicated to Andre Scedrov on the occasion of his 65th birthday*, 189–202.

McGee, R. W. (2023). Is chat gpt biased against conservatives? An empirical study (February 15, 2023). Available at SSRN: https://ssrn.com/abstract=4359405 or http://dx.doi.org/10.2139/ssrn.4359405

Nadeem, M., Bethke, A., & Reddy, S. (2020). StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*. https://doi.org/10.48550/arXiv.2004.09456.

Nadis, S. (2022). *Subtle biases in*. AI can influence emergency decisions. MIT News Office. https://news.mit.edu/2022/when-subtle-biases-ai-influence-emergency decision1216#:~:text=But%20the%20harm%20from%20a,an%20MIT%20team%20has%20show.

Naous, T., Ryan, M. J., & Xu, W. (2023). Having beer after prayer? Measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Nozza, D., Bianchi, F., & Hovy, D. (2021). HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*. (pp. 2398–2406). Association for Computational Linguistics.

Orgad, H., & Belinkov, Y. (2022). BLIND: Bias removal with no demographics. *arXiv preprint arXiv: 2212.10563*.

Pagan, N., Baumann, J., Elokda, E., De Pasquale, G., Bolognani, S., & Hannák, A. (2020). A classification of feedback loops and their relation to biases in automated decision-making systems. *ACM Transactions on Internet Technology*, *20*(4). Article 49.

Park, S., Choi, K., Yu, H., & Ko, Y. (2023). Never too late to learn: Regularizing gender bias in coreference resolution. *In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM'*23) *(pp. 15–23).* New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3539597.3570473

Ramesh, K., Sitaram, S., & Choudhury, M. (2023). Fairness in Language Models beyond English. In *Findings of the association for computational linguistics: EACL 2023*. Association for Computational Linguistics.

Roselli, D., Matthews, J., & Talagala, N. (2019, May). Managing bias in AI. In *Companion proceedings of the 2019 world wide web conference* (pp. 539-544). https://doi.org/10.1145/3308560.3317590

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., & Smith, N. A. (2022). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*. (pp. 5884–5906). Seattle, United States: Association for Computational Linguistics.

Shen, L., Tan, W., Chen, S., Chen, Y., Zhang, J., Xu, H., Zheng, B., Koehn, P., & Khashabi, D. (2024). The language barrier: Dissecting safety challenges of llms in multilingual contexts. *arXiv preprint arXiv: 2401.13136*.

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. pp. (3407–3412). Hong Kong, China: Association for Computational Linguistics.

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2021). Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*. (pp. 4275–4293). Association for Computational Linguistics. https://aclanthology.org/2021.acl-long.330.

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2023). Auditing and mitigating cultural bias in llms. *arXiv preprint arXiv:2311.14096*.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J. (2022). Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Tiku, N., Schaul, K., & Chen, S. Y. (2023). *These fake images reveal how AI amplifies our worst stereotypes. The Washington Post*. https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes/

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-tuned Chat Models*. arXiv preprint arXiv:2307.09288.

Trend Micro, UNICRI, & Europol (2023). *Malicious uses and abuses of artificial intelligence.* TrendMicro. https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/malicious-uses-and-abuses-of-artificial-intelligence.

Turner Lee, N., Resnick, P., & Barton, G. (2019). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Brookings. https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Utama, P. A., N. S. Moosavi, and I. Gurevych. 2020. “Towards debiasing NLU models from unknown biases.” In *Proceedings of the 2020 conference on empirical methods in natural language processing*, 7597–7610. Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp -main.613.

Vigdor, N. (2019). Apple card investigated after gender discrimination complaints. *The New York Times*, November 10. https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html.