Exploring Gender and Religious Bias in Large Language Models

Introduction

Large Language Models (LLMs) have transformed how we communicate, process information, and interact with technology. LLMs, like OpenAI's GPT, process vast amounts of text data from diverse sources to predict and construct coherent responses in applications ranging from virtual assistants to content creation tools. While they showcase remarkable capabilities, their growing influence raises important questions about fairness and inclusivity. Embedded in training data, these models often perpetuate systemic biases that reflect and reinforce societal inequalities. This paper explores gender and religious biases in LLMs, analyzing how they encode and reproduce cultural narratives. By integrating perspectives from computer science, women's studies, and theological ethics, we propose strategies for developing more equitable and inclusive language technologies.

Understanding Gender and Religious Bias in Language Models

Gender bias in language models is deeply rooted in historical and cultural contexts. These models, trained on vast datasets from literature, media, and public discourse, absorb patterns that reflect traditional roles and stereotypes. For example, they might associate leadership, assertiveness, or technical expertise with men while linking caregiving, emotional labor, or teaching to women. These subtle but pervasive patterns influence how language models respond to and generate content, affecting societal perceptions and the content generated by LLMs.

A striking instance of gender bias emerged in Amazon’s experimental recruitment algorithm in 2014 [1]. The tool was designed to streamline the hiring process by scoring resumes. However, it quickly became apparent that resumes with words like *executed* or *captained*, terms culturally coded as masculine, received higher scores than others. As a result, the algorithm systematically favored male candidates, reflecting biases from the historical data on which it was trained. While not an LLM, this example underscores the risks of bias in automated systems, a concern that extends to LLMs trained on similarly biased datasets. Amazon ultimately discontinued the tool in 2018, underscoring the consequences of allowing historical inequities to shape modern decision-making technologies.

Large Language Models are not just tools; they are also influencers. They reinforce gender expectations subtly but persistently. In professional contexts, for instance, they may suggest male pronouns when describing CEOs or inventors but default to female pronouns for roles like nurses or teachers. These biases extend beyond text completion; they shape how individuals view themselves and others, limiting aspirations and reinforcing societal divides.

Religious bias in language models is equally troubling, though it manifests in unique ways. These biases often stem from the dominance of Western perspectives in training data and the scarcity of nuanced representations of minority faiths. When language models interact with religious themes, they frequently reflect societal stereotypes or oversimplifications.

For instance, while Christianity might be represented with depth and nuance, religions such as Islam and Judaism are more likely to face misrepresentation or stigmatization. A request for religious advice might generate thoughtful responses grounded in Christian values but produce dismissive or harmful outputs or even toxic response when referencing minority religions. Such patterns not only reinforce existing prejudices but also risk alienating individuals from faith traditions that are already marginalized in mainstream discourse.

Additionally, religious bias can intersect with geopolitical dynamics. One such instances was during the early stages of the COVID-19 pandemic, when social media platforms faced widespread criticism for failing to contain conspiracy theories that targeted specific religious groups. For example, in the United States, far-right extremists falsely accused Jewish communities of orchestrating the virus's spread, invoking anti-Semitic tropes with deep historical roots [2]. Similarly, in other regions, religious minorities were scapegoated for outbreaks, exacerbating social divisions and fostering discrimination. Automated moderation systems on these platforms struggled to manage the surge of misinformation, inadvertently amplifying harmful narratives. These events illustrate how digital spaces can deepen pre-existing societal tensions, creating environments that privilege dominant perspectives while marginalizing others. Language models trained on such biased data risk perpetuating and spreading these narratives in new contexts, further entrenching digital inequities and undermining efforts to foster inclusivity and understanding across diverse communities.

How Bias Becomes Embedded in Language Models

Bias in language models is an inevitable consequence of their training processes. These models are developed using vast datasets that reflect the realities of human language, both its creativity and its prejudices. Gender bias, for example, often originates from occupational stereotypes embedded in job descriptions or historical inequalities in professional fields. Similarly, religious bias emerges from data that disproportionately reflects the perspectives of dominant cultural groups, leaving minority voices underrepresented.

Consider the datasets that fuel language models. Large corpora often prioritize Western-centric content, including news, literature, and social media. While this ensures broad coverage, it excludes underrepresented communities, whose perspectives might differ significantly from mainstream narratives. For example, idioms, rituals, or even names from non-Western religions might be misinterpreted or overlooked entirely. The absence of diverse voices perpetuates a cycle of erasure and misrepresentation, directly influencing the outputs of LLMs.

Without intervention, these biases become self-perpetuating. Language models learn to predict patterns, and if those patterns include stereotypes or prejudices, they are reinforced with each use. Addressing this issue requires rethinking how we curate data and train models to prioritize equity alongside accuracy.

Ethical Frameworks for Bias-Free Language Models

Ethical frameworks offer critical insights into addressing bias in language models. Feminist ethics, which emphasize dismantling oppressive systems, and liberation theology [3], which advocates for justice and inclusivity, both provide valuable principles for creating fairer technologies.

One guiding concept is philosopher Martha Nussbaum's "capabilities approach," which argues that technology should enhance human potential and uphold dignity. Applying this approach to language models means designing systems that actively counteract bias, ensuring they contribute to equity rather than perpetuating inequality.

From a theological perspective, the concept of *imago Dei* – the belief that all humans are created in the image of God – provides a moral imperative to design technologies that honor human dignity and diversity. By embracing these ethical principles, developers can create systems that respect the inherent worth of every individual, regardless of their gender or religion.

Challenges and Opportunities in Addressing Bias

Recognizing bias is the first step; addressing it is far more complex. Gender bias often manifests in overt ways, such as reinforcing stereotypes about professions or leadership. Religious bias, by contrast, is more subtle, shaped by deeply ingrained cultural narratives and historical injustices.

One challenge is that language models are designed to process patterns rather than understand context. For example, a model might associate "Muslim" with "terrorism" because of biased news coverage in the training data, even though this association is baseless and harmful. Correcting such biases requires rethinking both the data sources used and the algorithms that process them.

Another challenge lies in the global nature of bias. While much attention has been given to gender and racial biases in Western contexts, less focus has been placed on how these issues intersect with class, religion, or ethnicity in non-Western settings. Tackling these global dimensions requires diverse teams and a commitment to inclusivity in both research and implementation.

Gender Bias Challenges

Gender bias often appears in overt ways, such as reinforcing traditional stereotypes about roles, characteristics, or abilities associated with men and women. For example, language models might complete the phrase *"A brilliant scientist is..."* with a male pronoun, reflecting ingrained societal associations of brilliance with men. This perpetuates historical inequalities and discourages women from aspiring to certain roles or professions.

Additionally, the intersectionality of gender bias is not accounted for. Gender is not experienced uniformly; a woman's experience of bias can differ based on her race, ethnicity, socioeconomic status, or other intersecting identities. However, most language models fail to capture these nuances, treating gender as a binary construct rather than as a spectrum of identities. This limitation erases the experiences of transgender, nonbinary, and gender-nonconforming individuals, who are often underrepresented in training data.

Addressing gender bias is also complicated by its subtlety. While overtly sexist outputs are easy to identify and correct, implicit biases – such as consistently describing women in nurturing or subordinate roles – are harder to detect but equally harmful. These implicit patterns shape societal norms in ways that may not immediately appear discriminatory but have profound cumulative effects.

Religious Bias Challenges

Religious bias might often manifest in more subtle but equally damaging ways. These biases are shaped by societal narratives and geopolitical histories that marginalize certain faiths. For instance, Islamophobia and antisemitism, pervasive in some regions' media and public discourse, become encoded into language models, which may then generate harmful stereotypes or biased associations.

The scarcity of diverse religious perspectives in training data is another major obstacle. Faith traditions outside of dominant cultural narratives are often underrepresented, leading to mischaracterizations or omissions. For example, while references to Christianity may draw from rich and varied sources, smaller or less dominant religious traditions may only appear in contexts of conflict or controversy, distorting their true essence.

Moreover, religious bias is often context-sensitive, making it difficult to address universally. The same phrase or sentiment may be interpreted differently depending on cultural or theological backgrounds. This variability complicates the development of standardized tools for detecting and mitigating bias, requiring developers to engage deeply with the cultural contexts of the communities they aim to serve.

Opportunities for Addressing Both Biases

Despite these challenges, addressing gender and religious biases offers significant opportunities to advance fairness and inclusivity in language models.

1. Intersectional Analysis: Recognizing the interplay between gender, religion, and other identities can lead to more nuanced approaches to bias detection and correction. For instance, focusing on how gendered language varies across different religious contexts could uncover patterns invisible through isolated analysis.
2. Improved Training Data Curation: By deliberately sourcing data that represents diverse genders and religions, developers can reduce the dominance of Western or male-centric narratives. This includes collecting texts that reflect women’s contributions across fields, as well as literature from marginalized faith communities.
3. Collaborative Development: Addressing these biases requires interdisciplinary expertise. Sociologists, theologians, and gender studies scholars can collaborate with computer scientists to identify and address biases that may otherwise go unnoticed.
4. Continuous Testing: Regularly evaluating language models for bias in real-world applications allows developers to address emerging issues. Developers can employ fine-tuning with curated datasets or implement reinforcement learning from human feedback (RLHF) to mitigate such biases.

5. Developing Bias Detection Tools: Algorithms that identify and correct bias during training can prevent harmful patterns from becoming entrenched.

6. Transparent Metrics: Setting clear benchmarks for fairness and inclusivity can guide both developers and users toward more ethical outcomes.

7. Cultural Contextualization: Incorporating cultural awareness into language models can help them navigate sensitive topics more effectively. For example, instead of responding generically to religious or gender-related queries, models could tailor responses to reflect the diversity of perspectives within different traditions and identities.

Ultimately, the challenges posed by gender and religious biases in language models highlight the need for ethical vigilance and interdisciplinary solutions. These biases are not merely technical flaws but reflections of societal inequalities that require a concerted effort to dismantle. Addressing them is not just a matter of improving technology – it is a step toward fostering equity and understanding in a diverse, interconnected world.

Conclusion

LLMs, such as GPT, have become central to how we interact with technology, from daily tasks to shaping global conversations. As these models gain prominence in sectors ranging from healthcare to education, addressing the gender and religious biases they inherit becomes even more crucial. These biases reflect broader societal inequalities, but they are not an inevitable part of technological progress. By integrating insights from computer science, women's studies, and theology, we have an opportunity to shape these tools into instruments that foster empowerment, inclusivity, and fairness. Tackling bias in language models goes beyond improving technology – it is about cultivating a more just and equitable world, where technology supports and amplifies marginalized voices, helping to create a digital landscape rooted in our shared values of fairness and human dignity. As the role of LLMs continues to expand, the urgency of addressing these biases becomes clearer, underscoring the responsibility we hold in ensuring these technologies serve all of humanity equitably.

Works Cited

1. Amazon's sexist recruiting algorithm reflects a larger gender bias | Mashable
2. Preventing violent extremism during and after the COVID-19 pandemic
3. Religious Ethics in the Age of Artificial Intelligence and Robotics: Exploring Moral Considerations and Ethical Perspectives - AI and Faith