Advancing Fairness in Generative AI Through Intrinsic and Extrinsic Bias Evaluation and Mitigation

Mina Arzaghi[1,2]

[1] HEC Montréal, 3000 Chem. de la Côte-Sainte-Catherine, Montréal, QC H3T 2A7, Canada
[2] Mila – Quebec Artificial Intelligence Institute, 6666 Rue Saint-Urbain, Montréal, QC H2S 3H1, Canada
mina.arzaghi@mila.quebec

Abstract

This research focuses on understanding and mitigating bias in generative AI models, specifically by examining both intrinsic and extrinsic biases. The project aims to develop a unified evaluation framework and bias mitigation strategies to promote fairness across real-world applications, such as finance and healthcare. The goal is to ensure that generative AI systems do not propagate harmful societal biases, and the research explores bias detection and mitigation across various deployment stages.

Introduction

Foundation models are powerful due to their training on vast amounts of data, which enable them to perform well across a wide range of tasks. However, several studies have shown that these models often encode social biases present in the training data, which are then transferred to downstream tasks (Rombach et al. 2022; Nadeem et al. 2020). For instance, Stable Diffusion has been found to depict "poverty" with stereotypical images of Africa (Bianchi et al. 2023), while Llama2 and Falcon have been observed associating Black people with poverty (Arzaghi et al. 2024) when integrated into real-world applications, such as educational content creation, advertising, or chatbots. These biased models can lead to significant harms, including the amplification of stereotypes and allocational harms that affect marginalized groups such as unfair loan denials for minority groups, or biased healthcare recommendations that compromise patient care. Addressing these biases is essential to ensure that generative AI systems operate fairly and do not perpetuate harmful societal biases in high-stakes environments. Although bias evaluation and mitigation in these models have been widely studied (Gallegos et al. 2024; Wan, et al. 2024), there are still contradictions in the findings. Bias in foundation models is often categorized into two types (Gallegos et al. 2024): *Intrinsic bias*, which refers to biases embedded within the model itself, and *Extrinsic bias*, which measures bias in specific applications using metrics like equality of opportunity (Hardt et al. 2016). Some studies suggest that mitigating former bias can improve the latter (Jin et al. 2020), while others have shown no correlation or even negative correlation between the two (Goldfarb-Tarrant et al. 2020). To address fairness in foundation models, understanding the relationship between intrinsic and extrinsic bias is an important direction to ensure that bias mitigation strategies are effective, and ultimately promoting fairness in real-world applications.

Objective

The primary objective of this research is to develop a unified framework for evaluating and mitigating biases in foundation models. This involves:

- Understanding the relationship between intrinsic and extrinsic biases.
- Developing metrics that effectively capture both intrinsic and extrinsic biases across different downstream tasks.
- Proposing mitigation strategies that ensure fairness in generative models' real-world applications.

The outcome will be a bias evaluation and mitigation framework that can be applied to models such as GPT and Llama, addressing biases in applications like healthcare recommendations and credit approval systems.

Methodology

The project is divided into three work packages (WPs):

WP1: Survey of Intrinsic and Extrinsic Bias Relationship

This work package aims to understand the relationship between intrinsic and extrinsic biases across different adaptation techniques such as full fine-tuning (FF) (Weng 2024), Low-Rank Adaptation (LoRA) (Hu et al. 2021), In-Context Learning (RAG) (Lewis et al. 2020), and reward modeling (RLHF) (Bai et al. 2022). The survey will examine the impact of intrinsic bias mitigation on extrinsic outcomes by comparing scenarios where the parameter space changes (e.g., FF, LoRA) to those where it remains constant (e.g., RAG). Prior studies have yielded contradictory findings on whether mitigating intrinsic bias improves extrinsic outcomes (Goldfarb-Tarrant et al. 2020; Jin et al. 2020). Even though these studies provide a valuable foundation, they are limited by the use of older models, restricted intrinsic metrics, and evaluations confined to limited applications due to a lack of available datasets. With

recent advancements in foundation models and evaluation techniques, a comprehensive survey is needed. Using open-source datasets like Diabetes, German Credit, and Adult converted into language-model-friendly formats, this survey will evaluate biases across real-world applications (e.g., credit approval, healthcare, salary predictions) using modern foundation models such as Llama, and GPT.

WP2: Development of a Bias Evaluation Framework

Evaluating bias in foundation models is complex, as biases can emerge at different stages, both within the model itself and in its downstream applications. Intrinsic bias metrics, such as embedding-based (Caliskan et al. 2017), probability-based (Nadeem et al. 2020), and generated text-based (Hardt et al. 2016) metrics, evaluate inherent biases within the model itself, which may increase or decrease as a result of applying the model to specific downstream tasks. However, extrinsic bias evaluations (Dhamala, et al. 2021) are often limited to specific tasks and datasets, making it difficult to generalize bias detection across applications. This work package aims to develop a unified bias evaluation framework that effectively captures both intrinsic biases and extrinsic performance, bridging the existing gap in bias evaluation methods. The framework will leverage the datasets generated in WP1 and our previous paper (Arzaghi et al. 2024), using contrastive clustering(Li et al. 2021) and entropy-based metrics to evaluate bias comprehensively. Contrastive clustering will help form demographic group clusters from language model embeddings using my previous work dataset (Arzaghi et al. 2024), enabling comparison of how models assign biased sentences, which provides insights into intrinsic bias. Entropy will be used to assess uncertainty across these clusters, identifying whether certain attributes are disproportionately assigned in both intrinsic tasks (e.g., fill-in-the-mask) and downstream tasks (e.g., loan approvals). By combining these techniques, we can systematically examine how biases evolve from model training to real-world applications. This evaluation framework will unify bias analysis across all stages of model deployment. The outcome of this work package will include open-source code and tools made available to the community.

WP3: Bias Mitigation Framework Development

Building on the insights from WP1 and WP2, this work package focuses on developing an integrated bias mitigation framework for generative models, targeting both intrinsic and extrinsic biases. WP1 provides a deeper understanding of the relationship between intrinsic biases and downstream tasks, helping to identify key areas for mitigation. WP2 offers a comprehensive approach to evaluating biases, creating a solid foundation for mitigation efforts. Leveraging these insights, we propose using Reinforcement Learning with Human Feedback (RLHF) (Bai et al. 2022) to effectively address both types of biases. Current techniques often fall short in consistently mitigating biases within internal representations and downstream outputs. To overcome these limitations, RLHF will be adapted based on the specific nature of the bias. For intrinsic biases, such as in text generation tasks, RLHF will fine-tune model parameters based on human preferences, promoting fairer internal associations. The adaptable dataset developed in our previous work (Arzaghi et al. 2024) will support this phase, as it can be customized for various demographic groups and types of biases, such as stereotyping and can be easily annotated based on the model's output. For extrinsic biases, in tasks like classification (e.g., predicting loan approval), the mitigation may involve updating only the classifier using human feedback (without modifying the foundation model's parameters), or using a combination of both approaches. By applying this framework to models like GPT, Llama, and Stable Diffusion, we can rigorously test our approach. The outcome of this work package will include open-source code and tools made available to the community.

Expected Contributions

This research will contribute significantly to the field of AI fairness by:

- Providing a clear understanding of the relationship between intrinsic and extrinsic biases.
- Offering a unified evaluation framework that can be used to assess fairness across various tasks.
- Developing open-source tools and methods for bias mitigation in generative models, enhancing their fairness in real-world applications.

Conclusion

As foundation models become increasingly integrated into business and everyday life, ensuring fairness is crucial to prevent the amplification of biases that could harm marginalized communities. This project aims to address these concerns by developing methods to evaluate and mitigate biases across generative AI models, contributing to the more equitable deployment of AI technologies.

Acknowledgements

Funding support for project activities has been partially provided by Canada CIFAR AI Chair, Google award, NSERC Discovery Grants program , and IVADO. I also express my gratitude to Compute Canada for their support in providing facilities for my evaluations.

References

Arzaghi, M.; et al. 2024. Understanding Intrinsic Socioeconomic Biases in Large Language Models. *arXiv preprint arXiv:2405.18662*.

Bai, Y.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bianchi, F.; et al. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493–1504.

Caliskan, A.; et al. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Dhamala, J.; ; et al. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 862–872.

Gallegos, I.; et al. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–79.

Goldfarb-Tarrant, S.; et al. 2020. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*.

Hardt, M.; et al. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hu, E. J.; et al. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jin, X.; et al. 2020. On transferability of bias mitigation effects in language model fine-tuning. *arXiv preprint arXiv:2010.12864*.

Lewis, P.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, Y.; et al. 2021. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 8547–8555.

Nadeem, M.; et al. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Rombach, R.; et al. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Wan, Y.; ; et al. 2024. Survey of Bias In Text-to-Image Generation: Definition, Evaluation, and Mitigation. *arXiv preprint arXiv:2404.01030*.

Weng, B. 2024. Navigating the Landscape of Large Language Models: A Comprehensive Review and Analysis of Paradigms and Fine-Tuning Strategies. *arXiv preprint arXiv:2404.09022*.