Intrinsic Meets Extrinsic Fairness: Assessing the Downstream Impact of Bias Mitigation in Large Language Models

Large Language Models (LLMs) are increasingly deployed in sensitive domains such as finance, where intrinsic representational biases can propagate into extrinsic harms in downstream tasks. High-stakes applications such as credit scoring are especially vulnerable, as biased model behavior can reinforce existing inequities and result in harmful disparities across demographic groups [1]. While prior research has questioned whether intrinsic bias truly translates into extrinsic unfairness [4], this connection remains poorly understood. To address this gap, we propose a four-stage evaluation framework that systematically examines the relationship between intrinsic and extrinsic fairness. In Stage 1, we establish a baseline by training models such as logistic regression, LLM embeddings, and fine-tuned classifiers without any mitigation strategy, providing reference points for fairness and accuracy. In Stage 2, we evaluate task-level mitigation through Counterfactual Data Augmentation (CDA) [3], which balances gender representation by generating counterfactual training instances, allowing us to assess improvements in extrinsic fairness. In Stage 3, we adapt concept unlearning [2] as an intrinsic bias mitigation method, encouraging LLMs to forget socioeconomic stereotypes while preserving fluency and predictive utility, and we evaluate how this intervention impacts downstream fairness. Finally, in Stage 4, we combine CDA with unlearning to test whether dual mitigation further enhances fairness. We conduct experiments on three datasets (Adult Census Income, ACS Employment, and German Credit) using instruction-tuned LLMs (LLaMA-3.1, Phi-3, and Gemma-2) in both frozen embedding and fine-tuned classifier settings, evaluating performance with predictive accuracy and group fairness metrics, including Demographic Parity, Accuracy Parity, and Equality of Odds.

Our experiments demonstrate that intrinsic bias mitigation through unlearning is highly effective; in Phi-3, for instance, it reduces gender socioeconomic stereotype gaps by 94.9% while maintaining language fluency. In downstream tasks, unlearning consistently improves group fairness metrics while preserving predictive accuracy, whereas CDA primarily enhances demographic parity but can introduce accuracy tradeoffs. For instance, on the ACS Employment dataset, unlearned Gemma-2 improved Accuracy Parity from 0.199 to 0.104 (48% gain), and combining CDA with unlearning on Llama-3.1 reduced Demographic Parity from 0.080 to 0.014 (82% gain). On the Adult dataset, all three models maintained accuracy above 0.82 while showing reduced fairness gaps, and on German Credit, unlearning consistently outperformed CDA by improving group fairness metrics without sacrificing predictive performance. Overall, CDA and unlearning exhibit complementary effects, with their combination yielding the strongest fairness improvements across models and datasets.

This work contributes to bias mitigation and fairness in LLMs in two ways. First, we adapt concept unlearning to mitigate socioeconomic stereotyping, showing that intrinsic bias reduction improves both representational and downstream fairness. Second, we introduce a unified evaluation framework that links intrinsic and extrinsic fairness, enabling systematic comparison of mitigation strategies. The framework is flexible, applying to both fine-tuned and frozen LLMs, and offers actionable guidance for deploying fairer models in finance and other high-stakes domains.

References

[1] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

[2] Omkar Dige, Diljot Singh, Tsz Fung Yau, Qixuan Zhang, Borna Bolandraftar, Xiaodan Zhu, and Faiza Khan Khattak. Mitigating social biases in language models through unlearning. *arXiv preprint arXiv:2406.13551*, 2024.

[3] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

[4] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*, 2020.