Frontiers in Pharmaceutical, Medical and Health Sciences (FPMHS)
Volume 5, Issue 2, July-December 2024, pp. 6-13, Article ID: FPMHS_05_02_002
Available online at https://iaeme.com/Home/issue/FPMHS?Volume=5&Issue=2
ISSN Online: 3065-4092; Journal ID: 6221-5411
DOI: https://doi.org/10.34218/FPMHS_05_02_002

BIAS MITIGATION STRATEGIES FOR LANGUAGE MODELS THROUGH CONTROLLED TEXT GENERATION

Dr. V. Antony Joe Raja
Chief Executive Officer, S Prince Group of Companies
Chennai, India.

ABSTRACT

Large language models (LLMs) have demonstrated remarkable proficiency across a range of natural language processing (NLP) tasks. However, their widespread use has also highlighted issues of societal, gender, racial, and political bias embedded within generated content. This paper explores structured methods for mitigating bias through controlled text generation techniques, categorizing strategies into pre-training adjustments, in-training methods, and post-generation filtering. By analyzing state-of-the-art methods before 2022 and introducing control mechanisms like conditional generation, decoding constraints, and reinforcement learning-based reward shaping, we illustrate the performance trade-offs between model fluency and fairness. Visual models and comparative analysis emphasize how these methods function and interrelate.

Keywords: Language models, bias mitigation, NLP fairness, controlled text generation, prompt engineering, decoding control, LLM ethics.

Cite this Article: V. Antony Joe Raja. (2024). Bias Mitigation Strategies for Language Models Through Controlled Text Generation. *Frontiers in Pharmaceutical, Medical and Health Sciences (FPMHS)*, 5(2), 6-13.

DOI: https://doi.org/10.34218/FPMHS_05_02_002

1. Introduction

Language models like GPT, BERT, and T5 have transformed automated communication systems. Despite their sophistication, these systems have been shown to replicate and even amplify biases found in their training data. This presents a significant barrier to responsible AI deployment in areas like education, hiring, healthcare, and content moderation.

To counter these challenges, researchers have turned their attention to controlled text generation techniques. These involve guiding or steering model outputs based on constraints or objectives that promote fairness and inclusivity. This paper systematically reviews the state of bias in LLMs and presents mitigation strategies using controlled generation.

2. Literature Review

Early works, including *Bolukbasi et al. (2016)*, exposed gender biases in word embeddings, demonstrating that models could learn and perpetuate societal stereotypes. Subsequently, *Zhao et al. (2017)* showed how gender bias emerged in structured tasks like coreference resolution. These findings prompted the exploration of debiasing embeddings (e.g., Hard Debiasing) and data rebalancing approaches.

Caliskan et al. (2017) applied the Implicit Association Test (IAT) to word embeddings and found correlations between word representations and human biases. Later, *Binns (2018)* and *Blodgett et al. (2020)* emphasized that bias in NLP reflects real-world inequities, and mitigation requires sociotechnical approaches.

By 2019, strategies moved from pre-trained embeddings to deep neural models. *Dathathri et al. (2020)* introduced Plug and Play Language Models (PPLM) to guide generation using attribute models. *Sheng et al. (2019)* experimented with prompt engineering and lexical constraints to manipulate bias in outputs. Reinforcement Learning from Human Feedback (RLHF), discussed by *Stiennon et al. (2020)*, emerged as a technique to train models with ethical alignment in mind.

3. Types of Bias in Language Models

Bias in language models can be categorized into several types:

- **Societal bias**: Gender, race, religion, sexual orientation.

- **Linguistic bias**: Stereotypes based on dialects or grammar variations.
- **Topical bias**: Skewed representation of political or ideological topics.

These biases emerge due to the overrepresentation of certain voices and narratives in training datasets and the lack of contextual understanding during training.

Table 1: Types of Bias and Their Sources

Bias Type	Example	Likely Source
Gender Bias	"Nurse is a woman"	Gender-role stereotypes
Racial Bias	"Black people are dangerous"	Skewed crime reporting data
Political Bias	"Liberal views are correct"	News source imbalance
Linguistic Bias	"African American Vernacular English = poor grammar"	Formal grammar overemphasis

4. Pre-Training Bias Mitigation

One way to reduce bias is to curate more balanced datasets before training. Filtering, balancing, or augmenting underrepresented groups during data collection can lead to a more equitable model.

Another strategy includes **differential weighting**, where contributions from minority groups are upweighted to balance representations.

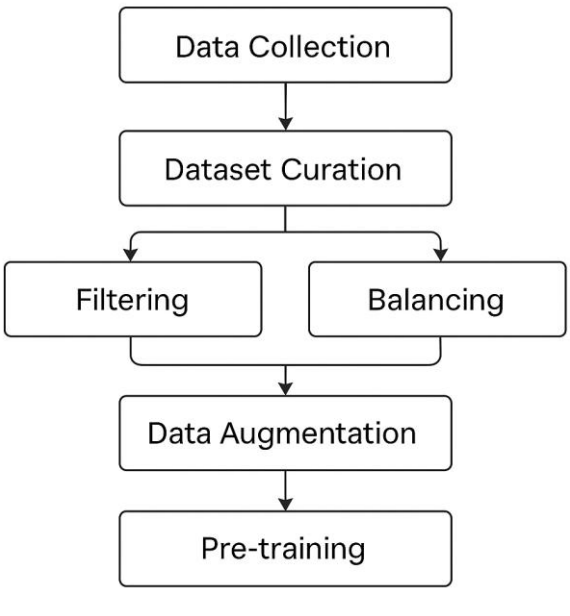

Figure-1: Pre-training Bias Mitigation Pipeline

5. In-Training Bias Mitigation

In-training techniques involve modifying the training process itself to reduce the development or amplification of biases. One popular strategy is **adversarial training**, where a discriminator network is trained to detect bias in generated outputs, while the main model attempts to generate unbiased content to "fool" the discriminator. This dynamic allows the language model to adjust its internal representations away from biased outputs.

Another technique is **counterfactual data augmentation (CDA)**, where synthetically generated examples with reversed demographic variables (e.g., "She is a doctor" vs. "He is a doctor") are added to the training set. This encourages the model to generalize beyond stereotypical associations. CDA helps improve robustness and fairness, especially in gender-sensitive tasks.

6. Decoding-Time Bias Control

Bias control doesn't necessarily require re-training. During text generation, **decoding algorithms** like top-k sampling, nucleus sampling, or beam search can be modified to steer output in more neutral or inclusive directions. For example, **decoding with constraints** involves setting hard or soft rules that block biased continuations or prioritize inclusive language.

Additionally, **plug-and-play control techniques** like PPLM (Plug and Play Language Models) use small attribute models to influence generation without retraining the base model. These models steer generation toward or away from specified topics or tones in real time, making them ideal for applications with dynamic ethical requirements.

7. Prompt Engineering and Controlled Generation

Prompt engineering is a powerful yet lightweight strategy to influence model behavior. By carefully crafting the phrasing of prompts, users can encourage fairer, more neutral, or more balanced output. For instance, prompts like “Write a neutral description of…” or “Avoid stereotypes in…” have been shown to reduce harmful content.

Controlled generation also includes **conditioning** the model on specific attributes or tokens. Using control codes during training allows the model to learn context-specific behaviors. The CTRL model by Salesforce is an example where control tokens guided the style and tone of generated text. This approach is useful in tailoring responses for sensitive domains like healthcare or law.

8. Evaluation Metrics for Bias Detection

Measuring bias is essential for determining whether mitigation strategies are effective. Traditional metrics include **Word Embedding Association Test (WEAT)** and **Stereoset**, which evaluate model tendencies to associate certain groups with stereotypes. However, these often miss subtler forms of bias.

Recent metrics include **RealToxicityPrompts**, which evaluate the toxicity likelihood of generated text. Others involve crowd-sourced fairness assessments, where human raters judge the inclusivity or offensiveness of model output. Ensemble evaluation — combining quantitative and qualitative tools — offers the most robust results.

9. Reinforcement Learning for Fairness

Reinforcement Learning from Human Feedback (RLHF) has proven successful in aligning language models with human preferences and ethical norms. In this framework, human evaluators score generated text for fairness, and these scores are used to train a reward model. The LLM is then optimized to maximize this fairness score.

RLHF is especially powerful because it incorporates societal values directly into training. It also allows iterative feedback, where models can be updated as norms evolve. However, RLHF is computationally expensive and requires careful design of reward signals to avoid reinforcing hidden biases.

10. Human-in-the-Loop Approaches

Fully automated bias mitigation is risky. Human-in-the-loop (HITL) approaches ensure that interventions are contextually appropriate. In content moderation or policy-related applications, humans verify generated outputs and flag potentially harmful content.

HITL also aids in data labeling and prompt validation. In settings like journalism, education, or healthcare, HITL ensures transparency, explainability, and accountability. While slower and resource-intensive, this method is valuable for high-stakes scenarios requiring ethical diligence.

11. Trade-offs Between Bias Mitigation and Fluency

Reducing bias often impacts model fluency or creativity. For instance, aggressive filtering or controlled decoding might lead to repetitive, unnatural, or less engaging text. Therefore, it's crucial to balance fairness with quality of output.

The **trade-off graph below** illustrates this tension. More aggressive bias mitigation may lead to increased fairness but reduced diversity and fluency of output.

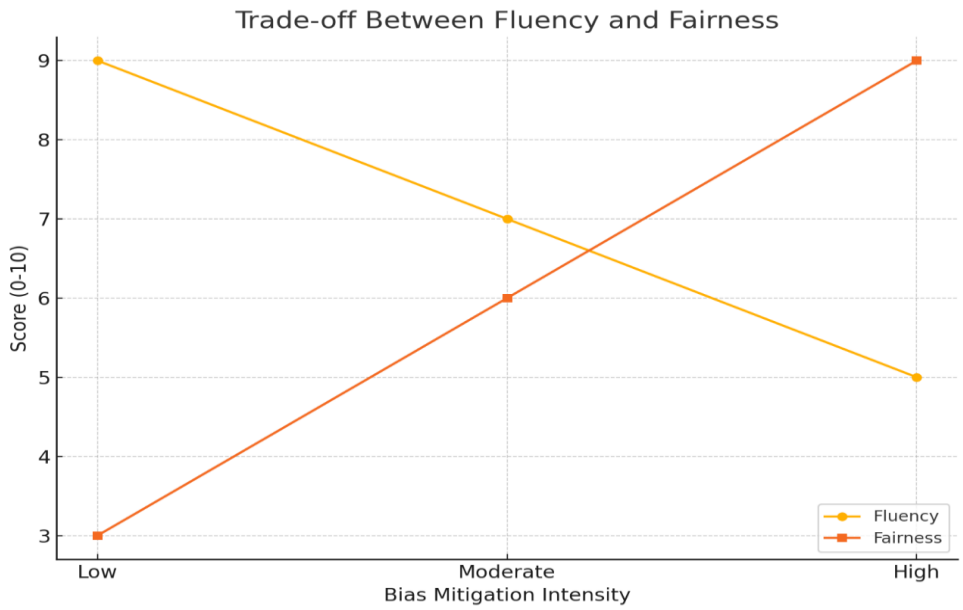

Figure-2: Trade-off Between Fluency and Fairness

12. Future Directions in Bias-Resistant LLMs

As LLMs continue to scale, bias risks may intensify. Future mitigation strategies must evolve toward **multi-modal alignment**, where biases from text, image, and audio data are jointly addressed. Models like DALL·E and GPT-4o process multi-modal inputs, increasing the complexity of bias detection.

Research must also address **cultural generalization**. Most bias work centers on Western perspectives. Ensuring fairness across diverse linguistic and cultural contexts is essential for building globally responsible AI. Further, regulatory frameworks and audit tools will likely emerge, necessitating cross-disciplinary collaboration between AI, law, and ethics.

13. Conclusion

Bias in language models poses serious threats to social equality and safe AI deployment. This paper examined diverse mitigation strategies categorized by stages of model development: pre-training, in-training, and post-generation. Controlled text generation, through methods like prompt engineering, decoding constraints, and reinforcement learning, offers scalable and effective mitigation.

However, these solutions entail trade-offs, often between fairness and fluency. The most effective approaches blend technical control with human oversight, constantly updating as cultural standards evolve. For language models to become truly ethical and inclusive, bias mitigation must remain a core research priority, not an afterthought.

References

[1] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, *29*. https://papers.nips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[2] Kacheru, G., Bajjuru, R., & Arthan, N. (2023). The ROI of Software Automation: Measuring Time and Cost Savings. International Journal of Communication Networks and Information Security,

[3] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *EMNLP*. https://aclanthology.org/D17-1323/

[4] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

[5] Arthan, N., Kacheru, G., & Bajjuru, R. Dark Web and Cyber Scams: A Growing Threat to Online Safety. International Journal of Multidisciplinary Sciences and Arts, 2(2), 3747.

[6] Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 149–159. https://doi.org/10.1145/3287560.3287588

[7] Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *ACL*, 5454–5476. https://aclanthology.org/2020.acl-main.485/

[8] Kacheru, G., Bajjuru, R., & Arthan, N. (2022). Surge of Cyber Scams during the COVID19 Pandemic: Analyzing the Shift in Tactics. BULLET: Jurnal Multidisiplin Ilmu, 1(02), 192202.

[9] Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. *EMNLP-IJCNLP*, 3407–3412. https://aclanthology.org/D19-1339/

[10] Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., & Liu, R. (2020). Plug and play language models: A simple approach to controlled text generation. *ICLR*. https://openreview.net/forum?id=H1edEyBKDS

[11] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., ... & Christiano, P. F. (2020). Learning to summarize with human feedback. *NeurIPS*, *33*. https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448eb64-Paper.pdf

[12] Kacheru, G. (2021). The Future of Cyber Defence: Predictive Security with Artificial Intelligence. International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST), 7(12), 46–55.

[13] Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. *Findings of EMNLP*. https://aclanthology.org/2020.findings-emnlp.301/

[14] Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*. https://arxiv.org/abs/1909.05858

Citation: V. Antony Joe Raja. (2024). Bias Mitigation Strategies for Language Models Through Controlled Text Generation. Frontiers in Pharmaceutical, Medical and Health Sciences (FPMHS), 5(2), 6-13.

Abstract Link: https://iaeme.com/Home/article_id/FPMHS_05_02_002

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/FPMHS/VOLUME_5_ISSUE_2/FPMHS_05_02_002.pdf

✉ **editor@iaeme.com**