Closing the Loop: Embedding Observability in the GenAI Product Lifecycle for Systematic Bias Mitigation

FREYAM MEHTA, International Institute of Information Technology, India
NIMMI RANGASWAMY, International Institute of Information Technology, India

In exploring the realm of Generative Artificial Intelligence (GenAI), this study introduces a significant step towards addressing the ethical challenge of inherent biases - specifically gender bias - within its applications. By presenting an observability and governance framework, highlighted through the BiasAware initiative, we offer a consolidated approach to identifying and mitigating biases during the early stages of the GenAI product lifecycle. This framework not only equips developers with the necessary methodologies and tools for bias detection and mitigation but also emphasizes the importance of continuous improvement through an Observability and Governance Layer. This layer ensures GenAI systems are continuously aligned with ethical standards, thereby promoting responsible AI development. While our primary focus is gender bias, the initiative is designed with the flexibility to extend its methodology to a broader spectrum of social biases, aligning with the emerging capabilities of GenAI and its role in shaping future human-AI interactions. Through this work, we aim to contribute to the development of GenAI technologies that are both innovative and inclusive, reflecting diverse societal values and norms.

CCS Concepts: • **Applied computing → IT governance**; • **Computing methodologies → Natural language processing**; • **Human-centered computing → HCI design and evaluation methods**.

Additional Key Words and Phrases: Observability and Governance Framework, Bias Mitigation, Generative Artificial Intelligence (GenAI), AI Training Dataset Bias

1 INTRODUCTION

Generative AI, at its core, embodies the potential for a paradigm shift across various sectors, promising innovations that span the spectrum from creative endeavors in art to critical applications in healthcare. Yet, this promise is shadowed by the specter of bias. As demonstrated in instances where AI in healthcare showed lower diagnostic accuracies for black patients compared to white patients due to underrepresented data, and in recruitment tools that disproportionately favored resumes featuring traditionally male-associated action words, the repercussions of unaddressed biases in AI are both immediate and far-reaching. These examples underscore not only the pervasive nature of bias in AI, but also the imperative to address it from the ground up, beginning with the training data itself.

In the enterprise AI ecosystem, Large Language Models (LLMs) powered applications are predominantly deployed using two architectures: Retrieval-Augmented Generation (RAG) and fine-tuning processes. Both approaches rely heavily on the quality and composition of their underlying datasets, but they do so in uniquely impactful ways. RAGs dynamically incorporate external data during the inference time, augmenting pre-trained models with relevant information from a broader dataset. Fine-tuning, on the other hand, adjusts the parameters of pre-trained models using specific datasets to tailor outputs to more narrowly defined needs. The pivotal role of data in these architectures brings into sharp focus the ethical imperatives and challenges inherent in designing responsible GenAI systems.

At the heart of the BiasAware initiative is a commitment to scrutinizing the very lifeblood of GenAI systems—their training data. Training Data Bias occurs when the data on which AI systems are trained contain skewed representations or outright exclusions of certain demographics, thereby leading to outputs that are biased. This paper emphasizes a targeted approach to mitigate this form of bias by implementing robust observability and governance mechanisms right

from the earliest stages of GenAI product development. We firmly believe that by addressing Training Data Bias, we can significantly reduce the incidence of AI systems that inadvertently perpetuate societal biases. This brings us to the heart of our inquiry: How can we leverage feedback loops within an observability and governance framework to systematically uncover and mitigate social biases at the early stages of GenAI product development?

2 RELATED WORK

Awareness of social biases in Generative Artificial Intelligence, specifically in Large Language Models (LLMs), has not only led to significant ethical concerns but also underscored the critical need for embedding bias mitigation mechanisms across the AI technology stack. The transformative potential of these technologies is marred by the risk of reinforcing societal biases present in their training data, resulting in potential dissemination of harmful stereotypes and a loss of public trust. This challenge highlights the imperative of scrutinizing and improving the AI framework—from data collection to application interaction—to ensure its ethical application in decision-making processes.

The nuanced behavior of LLMs, displaying complex social dynamics yet marred by inherent biases, draws attention to the essential scrutiny required at every level of the AI technology stack. As illustrated in Figure 2, starting from the Data Layer, it is crucial to enhance data quality and address inherent biases to ensure that the foundation upon which AI models are built is both ethical and of high quality. At the Model Layer, integrating inputs in a way that promotes unbiased and balanced outcomes is key, emphasizing the importance of model transparency and adaptability to prevent bias.

Efforts to detect and understand social biases in LLMs, such as model compression analysis [3] and benchmarks like StereoSet [6] and CrowS-Pairs [7], have been foundational in revealing how biases infiltrate AI systems. They emphasize the necessity for comprehensive approaches not only in model training but throughout all stages of AI applications. This drives home the importance of efficient coordination across the AI stack through advanced orchestration mechanisms, crucial for delivering relevant and unbiased AI outputs as emphasized at the Orchestration Layer. At the user-facing Application Layer, the direct impact of biases is most apparent, necessitating strict monitoring and corrective measures to ensure equitable and unbiased user experiences.

The evolution of proprietary and domain-specific data applications in the advanced AI context necessitates a streamlined integration strategy that systematically addresses social biases [9]. This nuanced approach prioritizes enhancing the entire lifecycle—from data acquisition through to application deployment—to preemptively identify and neutralize biases. Highlighting the overemphasis on model layer innovations [2, 5, 8, 10] at the expense of foundational data integrity, this integrated review aims to shift AI development towards a more inclusive and proactive stance on bias identification and neutralization, promoting an ethical, equitable, and responsible application of AI technologies. Through addressing the diversity of social biases and the critical examination of the AI technology stack, our approach seeks to advance the field in a socially responsible and technologically effective manner.

3 CLOSING THE LOOP WITH BIASAWARE

The BiasAware initiative plays a pivotal role in the contemporary landscape of ethical AI by providing a nuanced approach to identifying and mitigating gender biases present within AI training datasets. Developed in a collaborative effort with the AI Vulnerability Database (AVID), BiasAware serves as an exemplary model of leveraging open platforms for responsible AI development. This initiative is underpinned by a suite of specific methodologies, crafted with the intent to systematically address and reduce gender biases from the foundational data used in AI applications. Operationalized through an interactive portal hosted by Hugging Face (available at Hugging Face Spaces:

https://huggingface.co/spaces/avid-ml/biasaware), BiasAware offers a comprehensive tool set for the scrutiny and enhancement of both locally stored and remotely hosted datasets. This integration paves the way for a more inclusive and fair handling of data, well-aligned with the broader principles of fairness and inclusiveness that are increasingly recognized as crucial in the field of AI.

At the core of BiasAware's functionality are methodologies designed to quantitatively assess and address gender bias (as illustrated in Figure 1):

- Gender Distribution (Term Identity Diversity): Through this approach, BiasAware quantifies the representation of gender within text datasets by measuring the occurrence rates of gender-specific terms. Utilizing a lexicon of gender-related terms [4], this method accounts for the proportional representation of different gender identities, classifying texts into categories based on the predominant gender terms' presence. This methodological approach sheds light on potential gender biases in topics associated with different gender identities.
- Gender Profession Bias (Lexical Evaluation): This method actively seeks out patterns linking gender pronouns to professional roles within textual data. By employing predefined lexicons alongside sophisticated regular expressions, BiasAware meticulously catalogs connections between gender pronouns and professions. The insights derived from this analysis aim to uncover and mitigate biases in professional representation, crucial for fostering equitable AI applications that eschew gender-based professional stereotypes.
- GenBiT (Microsoft Gender Bias Tool): Further enhancing BiasAware's toolkit is the integration of GenBiT, a part of Microsoft's Responsible AI Toolbox [1], which provides a versatile framework for evaluating gender bias in language datasets through statistical analyses of word co-occurrences. GenBiT's broader applicability to various bias dimensions amplifies its utility in the endeavor to create bias-aware AI systems.

The introduction of an "Observability and Governance Layer" post-application phase, as illustrated in Figure 3, brings into focus a dynamic framework geared towards monitoring, evaluating, and refining the AI systems with an overarching goal of bias mitigation and ethical alignment. This layer acts as a critical feedback mechanism, channeling insights back to the foundational data layer, thereby setting the stage for a holistic BiasAware GenAI ecosystem. Here's an articulated insight into how this layer integrates and amplifies the BiasAware initiative:

(1) Purpose and Functionality: Positioned beyond the application layer, this layer serves a multifaceted purpose: it scrutinizes AI operations for adherence to security protocols, ethical standards, and provenance tracking. Moreover, it assesses the AI systems against a suite of bias metrics identified by the BiasAware framework, ensuring a comprehensive audit of social biases across all phases of AI development and deployment.
(2) Acting as a Feedback Loop: A distinguishing attribute of the Observability and Governance layer is its role in informing and refining the data layer. Utilizing metrics and insights gleaned from BiasAware's rigorous bias detection methodologies, this layer propels a system-wide introspection and recalibration process, aiming to iteratively enhance the quality of data and model outputs. It ensures a dynamic, responsive approach to bias mitigation, adapting to evolving societal norms and values.
(3) BiasAware Integration: At the heart of this layer's BiasAware integration are methodologies and tools designed for deep analytical assessment and bias rectification. By leveraging the Gender Distribution (Term Identity Diversity), Gender Profession Bias (Lexical Evaluation), and tools like GenBiT from Microsoft's Responsible AI Toolbox, the layer enriches the AI system with nuanced, actionable insights into potential bias manifestations.
(4) A Pathway for Continuous Improvement: By instituting a Governance regime that emphasizes ethical AI development and deployment, this layer fosters an environment of continuous improvement and accountability

within the AI ecosystem. It champions the principles of responsible AI, encouraging developers, stakeholders, and regulators to prioritize fairness, transparency, and inclusiveness.

4 DISCUSSION

As we advance the BiasAware initiative, we pivot towards three critical domains: broadening the spectrum of evaluating social biases embedding comprehensive feedback loops, and enhancing industry adoption. Initially focused on gender bias, it's critical to widen our scope to include racial, age, disability, and socioeconomic biases among others, tailoring our methodology for a more inclusive analysis. Addressing this diversity of biases ensures BiasAware remains relevant and effective across varying societal and ethical landscapes.

A cornerstone of our future strategy is the integration of a holistic feedback loop mechanism that spans all layers of the AI development lifecycle. This entails not just the data and model phases but extends to orchestration and application, enabling a governance model that iterates and enhances continuously. Such an approach ensures bias mitigation is an evolving commitment, leveraging real-time insights from AI application impacts to inform early-stage development processes.

Lastly, realizing the potential of BiasAware hinges on its industry-wide adoption and integration. Efforts will concentrate on forging partnerships with industry stakeholders, demonstrating the practical benefits of embracing BiasAware to foster fair and equitable AI applications. This includes advocating for BiasAware's methodologies to be recognized as foundational in the development of industry standards for ethical AI, promoting a broader cultural and operational shift towards responsible AI development.

5 CONCLUSION

This study pioneers a critical advancement in the ethical AI domain by instituting the BiasAware initiative, embedding observability into the generative AI lifecycle for proactive bias mitigation. Our approach, emphasizing early detection and continuous governance, paves the way for creating more inclusive and fair AI systems. By addressing gender bias with the potential for broader application, we champion a shift towards a more ethically aligned AI development process. This framework signifies a step toward realizing AI's promise without compromising societal values, guiding the industry towards a future where technology serves all segments of society equitably.

REFERENCES

[1] Shikha Bordia and Samuel R. Bowman. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. arXiv:1904.03035 [cs.CL]

[2] Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. 2023. Queer People are People First: Deconstructing Sexual Identity Stereotypes in Large Language Models. arXiv:2307.00101 [cs.CL]

[3] Gustavo Gonçalves and Emma Strubell. 2023. Understanding the Effect of Model Compression on Social Bias in Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2663–2675. https://doi.org/10.18653/v1/2023.emnlp-main.161

[4] Sophie Jentzsch and Cigdem Turan. 2022. Gender Bias in BERT - Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen (Eds.). Association for Computational Linguistics, Seattle, Washington, 184–199. https://doi.org/10.18653/v1/2022.gebnlp-1.20

[5] Yan Leng and Yuan Yuan. 2024. Do LLM Agents Exhibit Social Behavior? arXiv:2312.15198 [cs.AI]

[6] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416

[7] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154

[8] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic Biases in LLM Simulations of Debates. arXiv:2402.04049 [cs.CL]

[9] Kassym-Jomart Tokayev. 2023. Ethical Implications of Large Language Models A Multidimensional Exploration of Societal, Economic, and Technical Concerns. *International Journal of Social Analytics* 8, 9 (Sep. 2023), 17–33. https://norislab.com/index.php/ijsa/article/view/42

[10] Kai-Ching Yeh, Jou-An Chi, Da-Chen Lian, and Shu-Kai Hsieh. 2023. Evaluating Interfaced LLM Bias. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, Jheng-Long Wu and Ming-Hsiang Su (Eds.). The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taipei City, Taiwan, 292–299. https://aclanthology.org/2023.rocling-1.37

A FIGURES

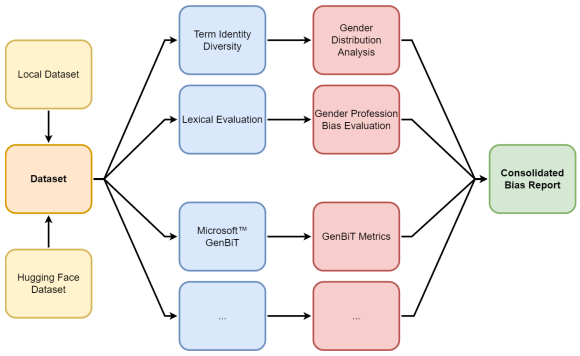

Fig. 1. Bias Mitigation in AI Training Datasets (Hugging Face Space: https://huggingface.co/spaces/avid-ml/biasaware).

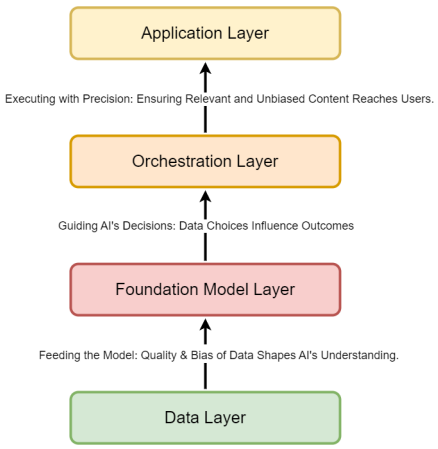

Fig. 2. The Current GenAI Product Lifecycle

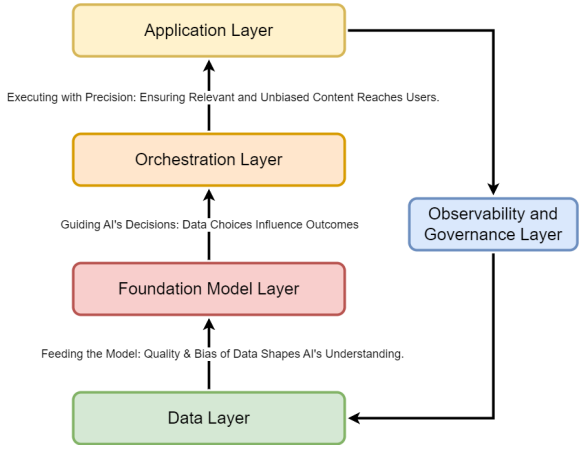

Fig. 3. The expanded 'BiasAware' GenAI Product Lifecycle