Proceedings of the CONF-MLA 2026 Symposium: Explainable Computing, Modeling & Data Science in Complex Systems
DOI: 10.54254/2755-2721/2026.GF37475

Bias in Large Language Models Through the Lens of Data Feminism

Lihao Wang

Qingdao No.58 High School, Qingdao, China
lihaoqd@outlook.com

Abstract. This paper systematically examines bias in Large Language Models (LLMs) through the lens of Data Feminism. The research demonstrates that bias in LLMs is not an issue of technical neutrality, but rather a systematic reproduction of societal power structures embedded in data collection, algorithmic design, and output generation. This analysis reviews relevant empirical studies and reveals both progress and limitations in current research on bias detection and measurement. The paper argues that future research should transcend mere technical adjustments by developing intersectional bias evaluation frameworks, exploring the feasibility of constructing "Feminist LLMs," and translating the critical insights of Data Feminism into actionable ethical standards and policy guidelines. These steps are critical to advancing a more equitable and responsible AI ecosystem.

Keywords: Data Feminism, Large Language Models, Algorithmic Bias, AI Ethics

1. Introduction

Large language models (LLMs) have demonstrated powerful capabilities across various tasks and are increasingly integrated into critical social domains. However, LLMs are not value-neutral technical tools; during training they may absorb and amplify social biases, especially along dimensions such as gender, race, class, and region [1]. For instance, models often associate certain professions with specific genders, reinforcing stereotypes [2]. Such biases extend not only to the model's semantic outputs, but also to upstream stages such as data collection, annotation, and optimization decisions [3].

"Data Feminism" offers a critical framework rooted in power, justice, and situated knowledge to interrogate and rectify algorithmic bias (a term encompassing bias in all algorithmic systems, including LLMs) [3]. It emphasizes that data and algorithms are embedded within social structures and power relations; feminist and intersectional perspectives can help reveal and resist hidden biases within technical systems.

This paper reviews the application of data feminism principles in bias research on LLMs, assesses existing findings, and outlines future research directions. Theoretically, it extends data feminism from traditional data science research to cutting-edge discussions in AI ethics, enriching the theoretical toolbox for algorithmic critique. Practically, it provides developers, policymakers, and the public concrete pathways for addressing bias in LLMs, contributing to fairer and more responsible artificial intelligence systems.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GF37475

2. Theoretical foundation: integrating algorithmic bias and data feminism

2.1. Algorithmic bias and fairness research trajectory

Algorithmic bias refers to situations where algorithmic systems systematically disadvantage certain groups in inputs, processing, or outputs [4]. In computer science and social science research, studies typically focus on fairness metrics, bias detection, mitigation strategies, and ethical frameworks. In more recent interdisciplinary work, scholars emphasize that algorithmic bias is not merely a technical issue but a reflection of socio-technical systems andthe reproduction of power [5].

Kasirzadeh argues that mainstream algorithmic fairness research tends to adopt distributive justice models and neglect structural injustice [6]. He advocates bringing feminist political philosophy into algorithmic fairness to create "responsible algorithmic fairness" that addresses deeper structural inequities [4]. Moreover, the myth of technical neutrality is contested: technological systems are not value-neutral; designers' choices, data selection, objective formulation, and task framing all embed normative judgments [7].

This theoretical trajectory in algorithmic bias research provides the conceptual foundation for understanding bias in LLMs, where similar socio-technical dynamics operate on a larger scale and with greater complexity.

2.2. Theoretical foundation: data feminism

Data Feminism, proposed by D'Ignazio and Klein, argues that data and algorithms are never neutral —they are products of social power relations [3]. It sets out seven core principles: examining and challenging power, rejecting binaries, valuing context, recognizing embodiment, embracing multiple methods, and making labor visible—which were later extended to include environmental impact and consent.

Applied to LLMs, these principles help move beyond purely technical fixes toward a structural understanding of bias. Because LLMs are trained on large internet corpora, they inevitably inherit social stereotypes. Bolukbasi et al. demonstrated this through the famous analogy *"man:programmer : woman:homemaker,"* where occupations like *nurse* were closer to *she* than *he* [2]. Such findings reveal how models reproduce social inequalities embedded in data.

However, identifying stereotypical associations in LLMs is only the first step. If bias is seen merely as an incorrect association between words, then the solution may seem limited to technical fixes or statistical adjustments. Data Feminism pushes the analysis further by asking why these associations exist, whose experiences are represented in the training data, whose perspectives are marginalized, and who has the authority to define fairness. Therefore, the presence of stereotypes in LLMs should not only be treated as a problem of model output, but also as evidence of deeper power relations embedded in data production, model design, and evaluation practices.

From a Data Feminism perspective, research on bias in LLMs should emphasize five dimensions [3]: (1) Power-identify who controls data and development; promote transparency and accountability; (2) Context-consider how bias differs across languages and cultures; (3) Methodological diversity-combine quantitative tests (e.g., embedding association, prompt audits) with qualitative and participatory studies; (4) Empowerment-involve affected communities in data collection and evaluation; (5) Sustainability and ethics-address labor and environmental costs, and ensure informed consent in data use.

In sum, Data Feminism provides a framework to move from mere bias detection toward redesigning the social and institutional conditions that shape AI systems.

DOI: 10.54254/2755-2721/2026.GF37475

3. Origins and evaluation methods of bias in LLMs

3.1. Origins of bias in LLMs

To understand bias in large language models, it is necessary to examine not only biased outputs, but also the broader processes through which such outputs are produced. From a Data Feminism perspective, bias is not simply an accidental technical error. Rather, it emerges from the interaction among data, social power relations, model design, and deployment contexts. Therefore, the origins of LLM bias can be traced across multiple stages of the AI development pipeline.

First, bias often originates from training data. LLMs are trained on large-scale text corpora collected from books, websites, news articles, online forums, and other digital sources. These sources reflect the societies that produce them, including their stereotypes, inequalities, and dominant cultural assumptions. If certain professions are repeatedly described as masculine, or if certain religious, racial, or regional groups are associated with negative language, models may learn these associations as statistical patterns. In this sense, LLMs do not create bias from nothing; they often reproduce and amplify patterns that already exist in society.

Second, bias may arise from unequal representation in data collection. Not all communities are equally visible in digital datasets. English-language content, Western perspectives, urban experiences, and voices from socially dominant groups are often overrepresented, while non-English languages, marginalized communities, rural populations, and lower-income groups may be underrepresented. This imbalance affects what the model learns as "normal" or "general" knowledge. For Data Feminism, this is not only a technical issue of dataset coverage, but also a question of power: whose knowledge is included, whose experience is excluded, and who has the authority to define what constitutes valid data.

Third, bias can be introduced or reinforced during annotation and human feedback processes. Many LLMs are refined using human evaluation, content moderation, and reinforcement learning from human feedback. While these processes aim to make models safer and more useful, they depend on human judgments about what is appropriate, harmful, neutral, or helpful. These judgments may vary across cultures, languages, and social groups. If annotators or evaluators share similar backgrounds or follow narrow guidelines, the resulting model may reflect limited assumptions about social norms and acceptable expression.

Fourth, bias can also be influenced by model optimization and deployment. Developers often optimize models for fluency, helpfulness, safety, and user satisfaction. However, these goals do not automatically guarantee fairness. A model may provide fluent and confident answers while still reproducing stereotypes. It may also avoid sensitive topics in ways that erase the experiences of marginalized groups, or present dominant viewpoints as neutral knowledge. Once deployed in education, employment, healthcare, or public communication, these biases can have real social consequences. Therefore, LLM bias should be understood as a socio-technical issue that extends beyond the model itself.

3.2. Evaluation methods for bias in LLMs

Because bias in LLMs appears in different forms, researchers have developed various methods to detect and evaluate it. Early studies often relied on embedding-based methods. For example, word embedding association tests measure whether certain words are statistically closer to specific social categories in the model's representation space. If "programmer" is closer to male-associated words and "homemaker" is closer to female-associated words, this suggests that the model has learned

DOI: 10.54254/2755-2721/2026.GF37475

gendered occupational stereotypes. Such methods are useful because they provide measurable evidence of bias inside language representations.

Recent studies have increasingly adopted prompt-based evaluation. In this method, researchers design prompts involving different identities, professions, or social groups, and then compare the model's responses. For example, a model may be asked to complete sentences about doctors, nurses, immigrants, religious groups, or people from different regions. Researchers then examine whether the outputs contain stereotypes, toxic associations, unequal sentiment, or different levels of respect and competence. Prompt-based evaluation is especially important for LLMs because these models generate open-ended text rather than just fixed categories.

Another common approach relies on benchmark datasets and quantitative metrics. Bias benchmarks allow researchers to test models systematically across categories such as gender, race, religion, nationality, and occupation. These benchmarks allow model comparisons and help track whether mitigation strategies reduce biased outputs. However, benchmark-based evaluation also has limitations. Fixed datasets may fail to capture subtle, contextual, or culturally specific forms of bias. They may also treat social categories as isolated variables, while real-world discrimination often occurs through intersecting identities such as gender, class, race, and region.

For this reason, Data Feminism advocates contextual, qualitative, and participatory approaches to bias evaluation. Technical metrics can reveal patterns, but they cannot fully determine whether an output is harmful or fair in a specific social context. A feminist approach would involve affected communities inevaluation and ask not only whether bias exists, but also who is harmed by it, how that harm is experienced, and what forms of accountability are needed. This approach expands bias evaluation from a narrow technical exercise into a broader ethical and political practice.

Overall, examining the origins and evaluationof LLM bias helps bridge theoretical critique and empirical research. It shows that bias is embedded throughout the AI lifecycle, from data collection to deployment. It also demonstrates why Data Feminism is useful for LLM bias research: it does not only ask how bias can be measured, but also how power, representation, context, and accountability shape the production of AI knowledge.

4. Empirical landscape of bias in LLMs

Empirical studies, building on the above discussion of bias origins and evaluation, consistently show that LLMs inherit and amplify societal biases, confirming Data Feminism's concerns.Bolukbasi et al. employed word embedding association tests to reveal that the Word2Vec model systematically associated "programmer" with men and "homemaker" with women, reflecting clear occupational gender bias. By identifying a "gender direction" within the embedding space, they quantified these associations and introduced neutralization and equalization techniques to mitigate bias. This seminal study demonstrated that bias is deeply embedded in training data and model representations, providing a methodological foundation for subsequent research on bias detection and mitigation in LLMs [2].

Caliskan et al. demonstrated that word embeddings trained on large text corpora capture human-like implicit biases. Using the Word Embedding Association Test (WEAT), they found that the GloVe model associated male names with career-related words and female names with family-related words, mirroring patterns observed in human implicit association tests. The study also revealed racial and age-related stereotypes and found that such linguistic associations strongly correlate with real-world labor statistics [8].

Sheng et al. examined gender bias in generative language models for dialogue tasks and found that models tend to generate stereotyped replies [9]. Bender et al. criticized the environmental

DOI: 10.54254/2755-2721/2026.GF37475

unsustainability of LLMs and their neglect of marginalized languages and cultures, calling for the development of more ethically aware language technologies [10]. Abid et al. studied bias in GPT-3 when describing Muslim groups, finding that it was more likely to associate Muslims with violence, further reinforcing religious stereotypes [11].

Collectively, these studies progress from identifying specific biases (e.g., gender in embeddings) to demonstrating their prevalence in model outputs, and finally to critiquing the structural foundations that allow such biases to persist. This evolution mirrors the Data Feminism call to move from detection to structural redesign.

5. Critical discussion and future research directions

5.1. Limitations of existing research

Building on the empirical review, current research exhibits these major limitations:

Current research predominantly focuses on revealing and measuring bias, lacking effective governance mechanisms. Moreover, most studies are still based on English corpora and Western contexts, and there is little research on bias in non-Western and non-English settings.

Methodologically, although recent studies have increasingly examined bias in large generative models through prompt-based audits and benchmark evaluations, some earlier foundational studies relied on static word embeddings or relatively limited corpora. Thus, existing methods still struggle to capture dynamic, contextualized manifestations of bias.

Finally, the literature largely examines biases along single axes (e.g., gender or race), failing to adequately address how these biases intersect and compound, a core concern of Data Feminism.

5.2. Future research directions

Strengthen Policy and Ethical Research. Explore how Data Feminism insights can be incorporated into legal and ethical norms for LLM development and deployment.

Deepen Localized Research. In the Chinese context, examine how LLM bias manifests across dimensions such as gender, region, and urban-rural divide, and explore governance approaches suited to local socio-cultural contexts.

Explore the Feasibility of "Feminist LLMs". Is it possible to build a new, ethically conscious language model from the ground up guided by data feminism principles?

Develop Intersectional Bias Evaluation Frameworks. Construct benchmark datasets and evaluation metrics capable of simultaneously assessing multidimensional biases such as gender, race, and class.

6. Conclusion

This paper systematically elaborates the theoretical framework, origins, evaluation methods, empirical progress, and future directions of bias in LLMs from a Data Feminism perspective. Data Feminism not only provides theoretical tools for critiquing bias in LLMs but also suggests actionable governance pathways. Addressing bias in LLMs is not only a technical challenge but also a political struggle concerning knowledge, power, and representation. Future research should integrate feminist epistemologies with technical innovation to design governance frameworks that are both globally ethical and locally contextualized.

DOI: 10.54254/2755-2721/2026.GF37475

References

[1] Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F.,…Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics, 50*(3), 1097-1179. https://doi.org/10.1162/coli_a_00524

[2] Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Barcelona, Spain.

[3] Klein, L., & D'Ignazio, C. (2024). Data feminism for AI. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, Rio de Janeiro, Brazil. https://doi.org/10.1145/3630106.3658543

[4] Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems, 31*(3), 388-409. https://doi.org/10.1080/0960085X.2021.1927212

[5] Ukanwa, K. (2024). Algorithmic bias: Social science research integration through the 3-D dependable AI framework. *Current Opinion in Psychology, 58*, 101836. https://doi.org/https://doi.org/10.1016/j.copsyc.2024.101836

[6] Kasirzadeh, A. (2022). Algorithmic fairness and structural injustice: Insights from feminist political philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (p. 348–356). https://doi.org/10.1145/3514094.3534188

[7] Castaneda, J., Jover, A., Calvet, L., Yanes, S., Juan, A. A., & Sainz, M. (2022). Dealing with gender bias issues in data-algorithmic processes: A social-statistical perspective. *Algorithms, 15*(9), 303. https://www.mdpi.com/1999-4893/15/9/303

[8] 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

[9] Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019, November). The woman worked as a babysitter: On biases in language generation. In K. Inui, J. Jiang, V. Ng, & X. Wan, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* Hong Kong, China.

[10] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event, Canada. https://doi.org/10.1145/3442188.3445922

[11] Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-Muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, Virtual Event, USA. https://doi.org/10.1145/3461702.3462624