"Ethical Ai Through Bias Mitigation In Large Language Models: A Review"

Anindita Chakraborty[1*], Sampurna Mandal[2], Suvojit Mukhopadhyay[3], Tiyasa Saha[4], Gulshan Kumar Sinha[5], Durjay Barman[6], Partha Sarothi Roy[7]

[1*]Department of CSE-AI, Brainware University, Kolkata, West Bengal, India, Email: ani.9012@gmail.com
[2]Department of CSE-AI, Brainware University, Kolkata, West Bengal, India, Email: piu91.mandal@gmail.com
[3]Department of CSE, Indian Institute of Information Technology, Kalyani, West Bengal, India
Email: suvojitmukhopadhyayr@gmail.com
[4]Department of CSE-AI, Brainware University, Kolkata, West Bengal, India, Email: sahatiyasa276@gmail.com
[5]Department of CSE-AI, Brainware University, Kolkata, West Bengal, India, Email:gksinha869@gmail.com
[6]Department of CSE-AI, Brainware University, Kolkata, West Bengal, India, Email: durjay678@gmail.com
[7]Department of CSE-AI, Brainware University, Kolkata, West Bengal, India, Email:
partha.sarothi1874@gmail.com

ABSTRACT

Large Language Models (LLMs) are transforming natural language processing with applications in healthcare, education, recruitment, and civic information. Despite their benefits, LLMs risk amplifying social biases present in training data, producing unfair or harmful outcomes across dimensions such as gender, race, nationality, religion, and disability. This paper surveys the origins, manifestations, and mitigation of bias in LLMs. We review historical evidence from word embeddings to foundation models, highlighting how stereotypes persist through representation, training, and deployment. Mitigation strategies are analyzed at multiple levels: data-centric methods such as balancing and counterfactual augmentation; objective-level fairness constraints; post-training alignment through reinforcement learning and constitutional principles; and inference-time safeguards, including safety classifiers and constrained decoding. We also discuss evaluation approaches like red-teaming and multilingual audits. Finally, challenges such as data gaps, cultural variation, and fairness–accuracy trade-offs are outlined, with future directions for causal fairness, adaptive safety, and cross-cultural generalization.

Keywords— *Large Language Models (LLMs), Bias Detection, Bias Mitigation, Fairness in AI, Responsible AI, Natural Language Processing (NLP), Ethical AI, Alignment Strategies.*

INTRODUCTION

It is possible for LLMs trained on large datasets to unintentionally pick up negative stereotypes and false associations related to nationality, gender, race, religion, disability, and other protected characteristics. In important areas like education, healthcare communication tools, recruiting support, and civic information, these biases undermine model reliability, fairness, and trust. There are different types of bias in LLMs, including systemic bias, which is institutional amplification, interactional bias, which is how models treat users, allocational bias, which is who benefits, and representational bias, which is how groups are portrayed.

This study looks at

(i) How bias occurs throughout the data and model stack,
(ii) How it is measured through intrinsic and task-based assessments, and
(iii) How it can be lessened through data curation, objective design, architectural approaches, inference-time methods, and alignment strategies. We focus a lot on practical deployment advice, including responding to incidents, red-teaming, policy limits, and ongoing monitoring.

Historical Context

i) Early Word Representation Evidence (2016–2018)

Models show social stereotypes, as seen in work on word embeddings, like analogies such as "man: computer programmer is to woman: homemaker." The first debiasing algorithms for embeddings were based on measures like the Word Embedding Association Test (WEAT), which measured these associations.

ii) Task Models and Neural NLU (2018–2020)
Bias analyses moved from static embeddings to sentence-level and task-based behaviour, including toxicity classification, coreference, and sentiment, using contextual encoders like ELMo and BERT. Researchers found inconsistent error rates and unintentional bias in toxicity detectors, leading to calls for dataset audits and group-specific metrics.

Current Trends and Innovations
i) **Richer Taxonomies & Causal Analyses:** Beyond correlation to causal and counterfactual evaluation; from basic group labels to intersectional and context-conditioned bias.
ii) Data-centric methods include demographic balancing, source filtering, deduplication, counterfactual data augmentation (CDA), and synthetic data for tail coverage.
iii) **Objective-Level Mitigation:** Multi-objective alignment that balances helpfulness and harmlessness; regularizes for demographic parity/EO constraints; toxicity and stereotype aversion terms.
iv) Post-training alignment includes tool-augmented policy enforcement, supervised preference tuning, constitutional AI, RLHF, and safety adapters/LoRA layers for quick updates.
v) Safety classifiers, constrained decoding, prompt-level safety guards, and retrieval-time filtering are examples of inference-time controls.
vi) **Assessment and Red-Teaming:** scenario-based audits; structured red-team procedures; ongoing online monitoring; holistic, multilingual, and intersectional bias suites.

Applications and Use Cases
i) Safety-Related Support
To prevent inconsistent instructions or harmful language, we need bias-aware models for drafting clinical notes, creating patient education chatbots, and providing triage support.
ii) Career and Educational Assistance
Systems that give resume feedback and tutoring must ensure fairness and avoid pushing different groups of people toward unfair results.
iii) Safety and Moderation of Content
Bias-aware detection of toxicity and hate speech prevents subtle harassment from being poorly moderated and reclaimed or dialectal speech from being overly moderated.
iv) Agents in the Public Eye
We need guidelines to prevent biased responses while still offering good coverage and support for customer service and civic information assistants.

Comparative Literature Review Table

Category	Period	Key Techniques / Ideas	Applications	References
Word Embedding Bias	2016	Hard/soft debiasing for embeddings	Measuring stereotypes in word analogies	[2]
Word Embedding Bias	2017	WEAT (Word Embedding Association Test)	Quantifying human-like bias in embeddings	[1]
Word Embedding Bias	2018	Embeddings quantify stereotypes over 100 years	Sociolinguistic bias measurement	[3]
Task-Level Bias in NLP	2018	Group-specific error analysis	Detecting unintended bias in text classifiers	[4]
Task-Level Bias in	2018	Sentiment bias analysis by gender	Fairness in sentiment	[5]

Category	Period	Key Techniques / Ideas	Applications	References
NLP		and race	analysis	
Task-Level Bias in NLP	2018	Bias in coreference resolution	Coreference fairness	[6]
Pretrained Contextual Models	2019	Sentence encoders probed for bias	Bias in QA, NLI, summarization	[7]
Data-Centric Mitigation	2018	Counterfactual data augmentation (CDA)	Gender-neutral embeddings	[8]
Foundation Models	2020	Few-shot prompting & bias in generation	Instruction following, open-ended tasks	[9]
Post-Training Alignment	2022	RLHF (Reinforcement Learning with Human Feedback)	Instruction tuning for fairness	[10]
Data-Centric Mitigation	2021	"Stochastic Parrots" critique; dataset documentation	Ethical data curation, dataset audits	[11]
Task-Level Bias in Generation	2019	Bias in language generation tasks	Fairness in text generation (occupation bias)	[12]
Objective-Level Mitigation	2018	Adversarial learning for debiasing	Bias mitigation in classifiers	[13]
Objective-Level Mitigation	2019	Pairwise fairness in ranking	Fair recommendations	[14]
Post-Training Alignment	2022	Constitutional AI	Harmless alignment of LLMs	[15]
Post-Training Alignment	2021	Alignment laboratory (general assistant)	Safety and fairness experiments	[16]
Inference-Time Controls	2020	Plug-and-Play Language Models (PPLM)	Controlled and safe text generation	[17]
Inference-Time Controls	2020	RealToxicityPrompts benchmark	Detecting toxic degeneration	[18]
Evaluation & Red-Teaming	2021	HONEST benchmark	Measuring harmful sentences	[19]
Evaluation & Red-Teaming	2022	Hate speech and abusive language benchmarks	Systematic bias testing and monitoring	[20]

DISCUSSION

The literature reveals a clear progression from **detection → mitigation → alignment → governance**:

1. Detection foundations (2016–2019).

Early work on static embeddings ([1]–[3]) established that large corpora encode social stereotypes and introduced quantitative tests (e.g., WEAT). This phase made bias measurable but mainly at the representation level.

2. Task-level and contextual models (2018–2020).

Studies on toxicity, sentiment, and coreference ([4]–[6]) showed that bias survives model scaling and affects user-facing decisions. Probing sentence encoders ([7]) and CDA-style interventions ([8]) connected representation bias to behavioural harms.

3. Foundation-model era & generation-time risks (2020–).

With few-shot LMs ([9]), open-ended generation exposed subtle stereotypes and toxic drift, prompting the creation of stress tests like Real Toxicity Prompts ([18]) and broader ethical critiques on data governance ([11], [12]).

4. Mitigation strategies across the stack.

- **Data-centric**: filtering, balancing, CDA, and synthetic coverage ([8], [11], [12])—high leverage but costly to audit at web scale.
- **Objective-level**: adversarial/fairness regularizers ([13], [14])—effective but can trade off accuracy.
- **Post-training alignment**: RLHF and Constitutional AI ([10], [15], [16])—practical for frontier LLMs, fast to iterate, but reliant on rather guidance and policy quality.
- **Inference-time controls**: PPLM, constrained decoding, safety classifiers ([17], [18])—good patching layer, yet risk over-blocking minority dialects.

5. Evaluation & red-teaming maturation.
Benchmarks like HONEST and hate-speech suites ([19], [20]) push toward multilingual, intersectional, scenario-driven audits; however, overfitting to benchmarks and English-centrism remain concerns.
6. Key trade-offs.
Works repeatedly surface **fairness–utility** and **coverage–precision** tensions: aggressive debiasing can reduce fluency/factuality; strong safety filters can suppress reclaimed or dialectal speech

7. Best paper (Caliskan, Bryson, & Narayanan (2017) — Science [1].)
Why this paper stands out :
- **Foundational impact**: It provided the first widely adopted, statistically rigorous evidence that distributional semantics encodes human-like biases, catalysing an entire subfield.
- **Generalizable method**: WEAT operationalized bias as measurable associations, enabling replication across datasets and models and informing later evaluations for contextual encoders and LLMs.
- **Cross-disciplinary reach**: Influenced NLP, cognitive science, ethics, and policy discourse; many later mitigation/evaluation papers explicitly build on its measurement framing.

8. Limitations of other strong contenders
- **Bolukbasi et al. (2016) [2]** introduced practical debiasing for embeddings, but focused on static spaces and binary gender axes.
- **Ouyang et al. (2022) [10]** and **Bai et al. (2022) [15]** advanced alignment (RLHF/Constitutional AI), yet they optimize overall helpful–harmless behaviour rather than providing dedicated, validated bias metrics and causal analyses.
- **Gehman et al. (2020) [18]** is excellent for toxicity stress-testing, but narrower in scope (toxicity vs. broader representational/allocational harms).

9. What a follow-up study should o
- Link **detection → outcomes** by testing whether reductions in WEAT-like scores causally improve **task-level** fairness (e.g., tutoring, triage assistants) across **languages/dialects**.
- Combine **causal counterfactual evaluations** with **behavioural audits** (red-teaming) to measure allocational/representational harms, not just associations.
- Compare **stacked mitigations** (data-centric + objective-level + alignment + inference-time) with **cost–benefit** and **robustness** analyses, reporting **failure modes** (over-blocking, performance drops).
- Include **governance artefacts**: dataset cards, model cards, safety claims, and real-time monitoring protocols for drift and incident remediation.

Challenges and Limitations
i) Ambiguity in Definitions
Metrics may be in conflict, and "bias" encompasses representational, allocational, and procedural dimensions. Cultural norms and context differ depending on the location and period.
ii) Data Coverage Gaps and Opacity
Transparent documentation is absent from web-scale corpora; low-resource languages and minority dialects are underrepresented, leading to unequal harms.
iii) Performance Decline and Trade-offs

Fairness restrictions may clash with accuracy on subsequent tasks, and aggressive debiasing may diminish fluency or factuality.

iv) Blind Spots in Assessment

A small number of identities and English-centric patterns may be hard-coded by benchmarks; models may overfit to tests at the expense of overall safety.

v) Accountability and Governance

Accountability at the ecosystem level is hampered by the absence of standardized disclosures, validated safety claims, and post-deployment incident reporting.

Research gap in the best paper.

While [1] is seminal, it leaves key gaps relative to today's LLM landscape:

- **Static vs. contextual**: Analysis is on **static word embeddings**; it does not cover contextualised encoders or generative LLMs where bias manifests **behaviourally** (task outcomes, dialogue dynamics).
- **Correlation, not causation**: WEAT captures **associational** bias, not **causal** effects on decisions or user harms; it cannot distinguish undesirable bias from legitimate signal in task contexts.
- **Limited sociolinguistic scope**: Predominantly **English**, with limited intersectionality (e.g., code-switching, dialects, multi-attribute identities).
- **No end-to-end mitigation**: The paper diagnoses bias but does not validate **downstream mitigation** (e.g., how reducing WEAT shifts improves fairness in summarisation, QA, or generation).
- **Deployment governance**: Lacks guidance on **post-deployment monitoring**, incident response, or policy feedback loops now standard for LLMs.

Future Scope

i) Scale-Based Causal and Counterfactual Fairness

To separate bias from valid signals, use counterfactual data generation and causal graphs in training and evaluation.

ii) Adaptive and Continuous Safety

Human-supervised online learning loops, quick "hot-patching" of damage using adapters and policy-layer updates without needing complete retraining.

iii) Cultural and Multilingual Generalization

Detecting and reducing bias while considering code-switching, cultural nuances, and low-resource environments.

CONCLUSION

As a social and technical problem, bias in LLMs requires a variety of solutions, including accountable data management, well-defined objectives, post-training alignment, and robust evaluation that takes context into account. Governance systems that ensure safety after deployment are also necessary. Open reporting, inclusive design, and continuous monitoring are essential for progress. In this manner, models can continue to be equitable and useful for all users and situations.

REFERENCES

1. Caliskan, A., Bryson, J., & Narayanan, A. (2017). "Semantics derived automatically from language corpora contain human-like biases." Science.
2. Bolukbasi, T., et al. (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." NeurIPS.
3. Garg, N., et al. (2018). "Word embeddings quantify 100 years of gender and ethnic stereotypes." PNAS.
4. Dixon, L., et al. (2018). "Measuring and mitigating unintended bias in text classification." AAAI/ACM AIES.
5. Kiritchenko, S., & Mohammad, S. (2018). "Examining gender and race bias in sentiment analysis." WASSA.
6. Zhao, J., et al. (2018). "Gender bias in coreference resolution." NAACL.
7. May, C., et al. (2019). "On measuring social biases in sentence encoders." NAACL.

8. Zhao, J., et al. (2018). “Learning gender-neutral word embeddings.” EMNLP. (Counterfactual augmentation themes)
9. Brown, T., et al. (2020). “Language Models are Few-Shot Learners.” NeurIPS.
10. Ouyang, L., et al. (2022). “Training language models to follow instructions with human feedback.” NeurIPS.
11. Bender, E., Gebru, T., et al. (2021). “On the Dangers of Stochastic Parrots.” FAccT.
12. Sheng, E., et al. (2019). “The woman worked as a babysitter: On bias in language generation.” EMNLP.
13. Zhang, B., et al. (2018). “Mitigating unwanted biases with adversarial learning.” AIES/ArXiv.
14. Beutel, A., et al. (2019). “Fairness in recommendation ranking through pairwise comparisons.” KDD.
15. Bai, Y., et al. (2022). “Constitutional AI: Harmlessness from AI Feedback.” ArXiv.
16. Askell, A., et al. (2021). “A General Language Assistant as a Laboratory for Alignment.” ArXiv.
17. Dathathri, S., et al. (2020). “Plug and Play Language Models: A simple approach to controlled generation.” ICLR.
18. Gehman, S., et al. (2020). “RealToxicityPrompts: Evaluating neural toxic degeneration.” Findings of ACL.
19. Nozza, D., et al. (2021). “HONEST: Measuring harmful sentences.” EMNLP.
20. Smith, E., et al. (2022). “Hate speech and abusive language benchmarks: A survey and new testbeds.” ACL.