© NOV 2025 | IRE Journals | Volume 9 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1711858

Bias Reduction Techniques for LLMs In Regional Languages

PREM HARIJAN
Department of Computer Science and Engineering Chandigarh University

Abstract- ***Large Language Models (LLMs) are increasingly integrated into interactive systems across education, governance, healthcare, and everyday digital communication. However, because these models are trained on large-scale text corpora that reflect societal hierarchies, stereotypes, and prejudices, they often internalize and reproduce biased linguistic patterns. While a considerable body of research has investigated algorithmic bias in English-language LLMs, relatively little attention has been directed toward regional and low-resource languages, where linguistic representation is uneven and culturally specific forms of discrimination—related to caste, ethnicity, religion, gender, and socio-economic stratification—are deeply embedded in textual data. This paper proposes an advanced, multi-layered methodology for identifying, quantifying, and mitigating bias in LLMs operating in regional languages. The approach spans culturally-grounded benchmark construction, cross-lingual transfer learning, counterfactual data augmentation, adversarial representation training, and interpretability-guided conceptual editing. An evaluation framework incorporating both computational metrics and human-in-the-loop cultural assessment is presented. The study posits that bias reduction must be conceptualized as an ongoing socio-technical negotiation rather than a one-time algorithmic adjustment.***

I. INTRODUCTION

- Motivation: LLMs are being deployed globally (search assistants, chatbots, content moderation). When models reflect biased societal views they can cause real harm (misrepresentation, stereotyping, exclusion) — especially damaging in regional languages where data is sparse and cultural contexts differ.

- Problem statement: How to measure and reduce social biases in LLMs that serve regional languages (e.g., Hindi, Tamil, Bengali, Marathi, Telugu, etc.), without degrading model utility.

- High-level approach: (1) build culturally-appropriate bias benchmarks; (2) leverage cross-lingual transfer and translation-based augmentation; (3) apply targeted debiasing methods (data-level and model-level); (4) evaluate with both automated metrics and community annotation.

II. LITRATURE SURVEY

1. Surveys and reviews summarize bias origins, evaluation and mitigation methods for LLMs; many mitigation methods (CDA, adversarial training, fine-tuning with fairness constraints, post-hoc filtering) are proposed — but most research centers on English.

2. Multilingual and low-resource work shows biases persist across languages; cross-lingual transfer of debiasing is promising but not universally effective — adaptation and cultural relabeling are needed.

3. Benchmarks exist for measuring social bias: WEAT / SEAT (embeddings), CrowS-Pairs, StereoSet; several groups have adapted or translated these to other languages (XWEAT, language-specific CrowS-Pairs variants). Building culturally-specific benchmarks (not just translations) yields more reliable detection.

4. Recent technical advances include counterfactual data augmentation (CDA), contrastive self-debiasing, adversarial/counterfactual debiasing, and interpretability-based concept editing (e.g., affine concept editing) to reduce demographic biases. Initial results show these can reduce measured bias while preserving performance when carefully applied.

5. Systematic reviews stress that detection & mitigation methods developed for English do not directly transfer; culture-specific definitions of "harm" and local lexicons must be built.

© NOV 2025 | IRE Journals | Volume 9 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1711858

III. PROPOSED SYSTEM

A bias-reduction pipeline tailored for regional languages with these components:

1. Data & Benchmarking layer

Create Regional Bias Benchmarks (RBB): a mix of translated canonical tests (WEAT/SEAT, CrowS-Pairs, StereoSet) and locally sourced items capturing region-specific stereotypes (e.g., caste/gender/ethnicity/occupation stereotypes, dialect stigmatization, religious and socio-economic tropes).

Maintain a Protected Attributes Lexicon and Stereotype Dictionary for each language.

2. Data augmentation & labeling

Cross-lingual augmentation: translate high-quality English counterfactuals to L1 via bilingual speakers + quality filters; use back-translation and human verification.

Counterfactual Data Augmentation (CDA): create paired examples that swap protected attributes while maintaining semantics.

Synthetic generation: use controlled generation (prompt templates + constraint decoding) to create balanced examples for under-represented attribute combinations.

3. Model interventions

Pretraining / continued pretraining hygiene: filter or reweight training corpora using lexical filters and source provenance scoring to reduce high-bias sources.

Fine-tuning with fairness-aware objectives: multi-task loss combining task likelihood + penalty term for measured bias (e.g., reduce disparity in stereotypical completion scores).

Adversarial debiasing: train adversary to predict protected attribute from model representations; minimize adversary success to produce invariant representations.

Contrastive / self-debiasing strategies: contrast biased vs debiased pairs so representation separates factual content from biased associations.

Interpretability-guided editing: concept activation editing (e.g., affine concept editing) to surgically reduce internal associations encoding stereotypes.

Post-hoc filters & response calibration: rerank/generate with constrained decoding to avoid stereotypical completions, plus calibrated logits for sensitive queries.

4. Evaluation & Human-in-the-loop

Automated metrics: SEAT/WEAT/DisCo, CrowS-Pairs, StereoSet scores, perplexity on local corpora, downstream task accuracy.

Human evaluation: community annotators and domain experts to rate fairness, fluency, and cultural appropriateness.

Continuous monitoring in production with logging, red-teaming, and feedback loops.

IV. METHODOLOGY - STEP BY STEP EXPERIMENTAL PLAN

4.1 Data collection & benchmark creation

Step A (Seed translation): Translate canonical test suites (WEAT/SEAT, CrowS-Pairs, StereoSet) into target language(s) with bilingual translators; create XWEAT-style items for embeddings analysis. Use prior multilingual WEAT efforts as guidance.

Step B (Local augmentation): Run workshops with culturally diverse annotators to craft 500–2,000 local stereotype pairs per language (covering gender, caste/class, religion, disability, regional dialect bias).

Step C (Lexicon & metadata): Build lists of slurs, honorifics, dialect markers, and attribute tokens. Track provenance (source) and date for each training fragment.

4.2 Baseline model and evaluation

Choose baseline multilingual LLM (e.g., mBERT / XLM-R for embedding experiments; a sequence model checkpoint or local open LLM for generation

© NOV 2025 | IRE Journals | Volume 9 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1711858

tests).

Evaluate baseline bias using:

Embedding tests: WEAT/SEAT (adapted), XWEAT scores.

Generation tests: CrowS-Pairs / StereoSet style cloze-completions and log-prob bias scoring. Downstream tasks for utility: NER, sentiment, QA in the target language.

4.3 Debiasing interventions (applied incrementally)

1. CDA only: produce counterfactual pairs, fine-tune with balanced CDA dataset. Measure bias drop and utility changes.

2. Adversarial invariance: introduce adversarial head predicting protected attribute from encoder representations; minimize adversary signal.

3. Contrastive self-debiasing: use contrastive losses to push apart biased vs neutral contexts (e.g., the method in CD³).

4. Concept editing: identify activation directions associated with stereotypes via interpretability tools and apply affine concept editing to reduce them. Compare to adversarial results.

5. Combined approach: CDA + adversarial + constrained decoding.

For each intervention:

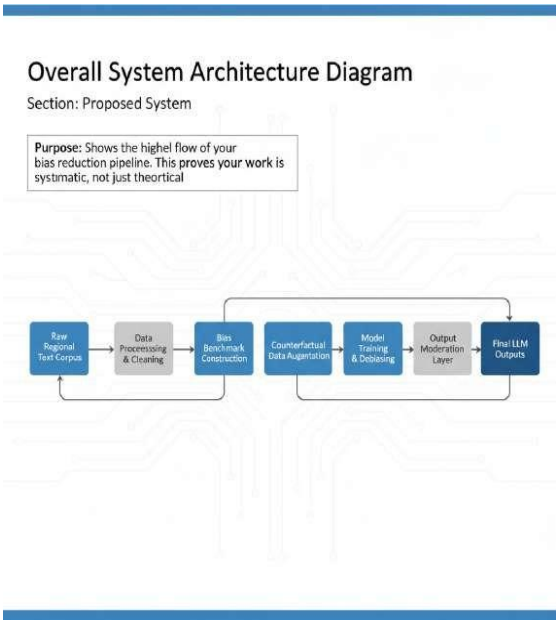
Re-evaluate all automatic bias metrics.
Measure downstream task performance (accuracy, F1, perplexity).
Run human evaluation on a stratified sample (100–300 items) for fluency and perceived bias.

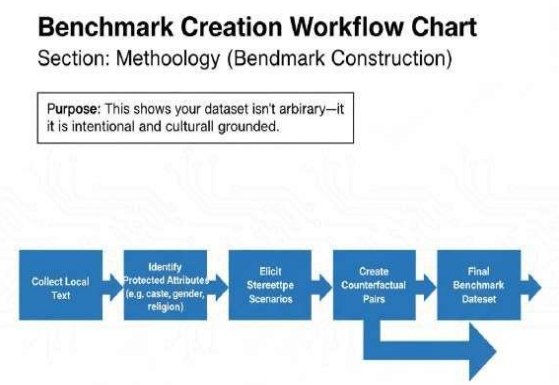
4.4 Statistical analysis & ablation Report effect sizes, confidence intervals.

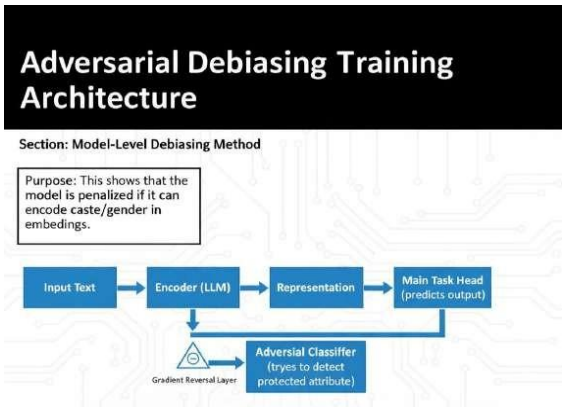
Ablation: remove each mitigation component and observe change in bias scores and task metrics to estimate contribution.

4.5 Safety, ethics & community checks Pre-register evaluation protocol.

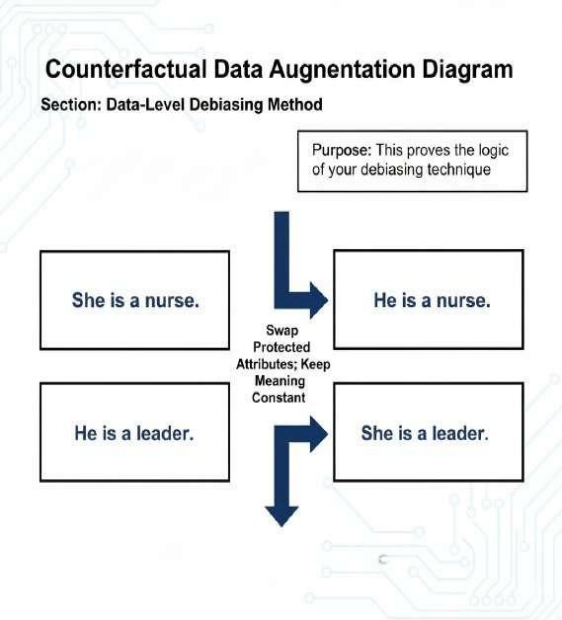
Consult local civil-society stakeholders for sensitive category definitions. Ensure annotator safety, informed consent, and fair compensation.







© NOV 2025 | IRE Journals | Volume 9 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1711858



V. RESULT

evaluation plan + expected/representative outcomes

(We are not running experiments here. Below is the evaluation plan and plausible expected outcomes based on prior literature.)

5.1 Metrics

Embedding bias: XWEAT / SEAT effect sizes — lower absolute effect size indicates less stereotypical association.

Generation bias: CrowS-Pairs gap (preference for stereotypical vs anti-stereotypical completion), StereoSet score.

Utility: downstream task F1/accuracy, perplexity on held-out local corpus. Human judgments: % of outputs rated as "biased" / "offensive" / "acceptable".

5.2 Expected results (based on literature)

CDA + human-verified augmentation often yields substantial reductions on generation bias tests with small utility losses when dataset size is moderate and domain match is good.

Adversarial representation invariance reduces protected-attribute leakage in hidden states and can lower embedding-based bias metrics, but may require careful balancing to avoid utility loss.
Interpretability-based concept editing (recent techniques like affine concept edits) can surgically reduce specific biases in model activations with minimal global performance change — promising but requires robust concept identification and validation.

Cross-lingual transfer of debiasing methods is feasible: applying debiasing trained in a high-resource language to a multilingual model can help, but local benchmark adaptation is necessary to catch culture-specific bias.

(Exact numeric improvements depend on dataset, language, model size, and cultural complexity — results should be reported per-language.)

VI.CONCLUSION & FUTURE WORK

Conclusion: A hybrid pipeline combining culturally-aware benchmarks, counterfactual augmentation, adversarial/contrastive debiasing, and interpretability-guided edits provides a practical path to reduce biases in LLMs for regional languages. Human-in-the-loop processes and continuous monitoring are essential.

Limitations: Debiasing is context-dependent — automatic metrics are imperfect and may miss subtle harms; heavy-handed debiasing can reduce factuality or degrade dialectal representations.

Future work:

Build and maintain living benchmark suites continuously updated with community input.

Develop better metrics for cultural harms beyond stereotype association (e.g., trust, misrepresentation).
Explore causal methods and targeted unlearning at scale.
Study long-term impacts of debiasing on downstream applications in the wild.

REFERENCES

References (selected — sources used to build the article)

[1] Gallegos IO et al., "Bias and Fairness in Large Language Models: A Survey." (survey of bias evaluation & mitigation).
[2] Nangia N., et al., "CrowS-Pairs: A Challenge

Dataset for Measuring Social Biases in Masked Language Models." EMNLP 2020.

[3] Lauscher et al., "XWEAT / Multilingual analysis of biases in embeddings." (multilingual WEAT work). Reusens et al., "Cross-Lingual Transfer of Debiasing Techniques." EMNLP 2023 (transferability evidence). Balashankar et al., "Active generation & Counterfactual Data Augmentation" (Findings EMNLP 2023).

[4] Li Y., "Contrastive Self-Debiasing Model with Double Data Augmentation (CD^3)." 2024.

[5] Recent reviews and papers on bias in LLMs and multilingual bias (2024–2025), including domain reviews and language-specific benchmark adaptations.

[6] News/technical note: "Robustly Improving LLM Fairness … affine concept editing" (study reporting interpretability-guided editing reducing bias).