JIRBDAI

Journal of Institutional Research, Big Data Analytics and Innovation

# Assessing Bias and Fairness in AI Language Models: A Systematic Review of Current Approaches

**Fatima Binta Abdullahi[1], Akanbi Muhammed Bello[2], Atumosha Adamu Yusuf[3], Akanbi Aisha Olamide[4]***

[1,2,3]Department of Computer Science, University of Abuja, Nigeria
[4]Department of Computer Science, Airforce Institute of Technology Kaduna, Nigeria
*Corresponding Author: aishaolamide13@gmail.com

## ABSTRACT

Artificial intelligence language models have demonstrated remarkable capabilities across diverse natural language processing tasks; however, systematic biases embedded in training corpora continue to propagate discriminatory outcomes across demographic groups. As large language models (LLMs) are increasingly deployed in high-stakes domains including healthcare, criminal justice, hiring, and financial services, their potential to amplify societal inequities has attracted urgent scholarly and regulatory attention. Despite a rapidly growing body of empirical work, no prior study has comprehensively synthesized bias research across all major LLM families using rigorous systematic review methodology. This paper addresses that gap by presenting a comprehensive systematic review of 147 peer-reviewed publications (2018–2024) examining bias detection, quantification, and mitigation strategies in LLMs, conducted in strict accordance with PRISMA 2020 guidelines. We synthesized empirical findings across gender, racial, socioeconomic, and intersectional bias dimensions, and introduce a novel five-tier Bias Severity Classification Framework (BSCF) that standardizes how bias magnitude, demographic scope, and harm potential are categorized and reported. The BSCF tiers range from Negligible (T1, demographic score delta $\leq 0.02$) to Critical (T5, delta $> 0.30$ with intersectional impact), providing a model-agnostic vocabulary for both research reporting and regulatory compliance. Quantitative meta-analysis using a random-effects model across 89 debiasing interventions reveals that state-of-the-art methods reduce measured bias by 23–67% while incurring 2–11% performance degradation on downstream tasks, with heterogeneity index $I^2 = 71.4\%$, indicating substantial variability across intervention families. Our analysis of benchmark datasets including WinoBias, StereoSet, BBQ, CrowS-Pairs, and BOLD further reveals that no single benchmark captures the full bias landscape, underscoring the need for multi-benchmark evaluation suites. We identify three critical research gaps: multilingual bias coverage, temporal drift of bias metrics, and ecological validity of laboratory evaluations. We further present a standardized evaluation pipeline and actionable recommendations for equitable AI deployment, with the BSCF proposed as a reporting standard for future empirical work.

**KEYWORDS:** Algorithmic Fairness, Algorithmic Bias Mitigation, Bias Detection, Debiasing, Demographic Parity, Equalized Odds, Gender Bias, Language Models, Llm Fairness Evaluation, Racial Bias, Sociotechnical Systems, Transformers.

## I. INTRODUCTION

The problem of bias in computational systems predates the modern deep learning era. Early Natural Language Processing (NLP) research documented distributional asymmetries in word embeddings, where vector representations of occupational and social terms systematically encoded gender and racial stereotypes [4]. However, the transition from task-specific NLP pipelines to large-scale, general-purpose language models has fundamentally altered the risk landscape. Large Language Models (LLMs) have achieved unprecedented natural language understanding and generation capabilities, yet their deployment in consequential decision-making contexts hiring, criminal justice, healthcare, and financial services raise critical concerns about embedded biases and fairness violations [1]. When models trained on internet-scale corpora reproduce and amplify societal prejudices, the consequences extend far beyond individual interactions to affect entire demographic communities at scale [2], [3]. The challenge of bias in AI systems



Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.20837351

is not a new phenomenon; however, the scale, fluency, and perceived authority of modern LLMs creates qualitatively new risks. A hiring algorithm that subtly associates female pronouns with domestic roles, or a medical chatbot that underestimates pain reports from certain patient groups, can cause systemic harm while evading detection through conventional software testing [4]. The societal urgency of this problem is increasingly reflected in regulatory frameworks: the European Union AI Act (2024) classifies high-risk AI applications and mandates bias auditing, while U.S. executive guidance has called for algorithmic impact assessments in federal procurement. Prior scholarly surveys have addressed related dimensions of this problem. Blodgett et al. [9] conducted a critical survey of the conceptual underpinnings of bias in NLP, exposing inconsistencies in how bias is defined and operationalized across the research community. Sun et al. (2019) surveyed gender bias mitigation techniques specifically in word embeddings and sentence encoders. However, these prior works predate the widespread adoption of instruction-tuned and RLHF-aligned LLMs, and none applies systematic review methodology with PRISMA compliance, quantitative meta-analysis, or a unified severity classification framework. The present review fills this gap. The field has responded to growing bias concerns with an expanding ecosystem of bias benchmarks, mitigation algorithms, and evaluation frameworks, yet no prior work has comprehensively synthesized these developments through rigorous systematic review methodology. This systematic review makes the following primary contributions: (1) a comprehensive synthesis of 147 studies across all major LLM families and bias dimensions, applying PRISMA 2020 guidelines; (2) the Bias Severity Classification Framework (BSCF) as a standardized, model-agnostic taxonomy grounded in established fairness theory; (3) quantitative meta-analysis of debiasing method efficacy versus performance trade-offs using a random-effects model; (4) identification of underexplored research directions including multilingual bias, intersectional fairness, and temporal bias drift; and (5) a reproducible evaluation pipeline publicly released to the research community. The remainder of this paper is structured as follows. Section II describes the systematic review methodology. Section III characterizes the dataset and corpus. Section IV presents the bias analysis framework and quantitative findings. Section V reports evaluation results. Section VI discusses implications and limitations. Section VII concludes with future directions.

## II. METHODOLOGY

### A. Review Protocol and PRISMA Compliance

This systematic review adhered to PRISMA 2020 guidelines [5]. A pre-registered protocol defined inclusion and exclusion criteria, search strategy, data extraction fields, and synthesis methods prior to data collection. Three independent reviewers conducted screening with inter-rater reliability measured by Cohen's $\kappa = 0.84$, indicating strong agreement. The study selection process followed a four-stage flow as illustrated in Figure 1.

(a). Identification: 4,823 records retrieved from database queries;
(b). Screening 3,614 records screened after removal of 1,209 duplicates;
(c). Eligibility 467 full-text articles assessed after excluding 3,147 off-topic records;
(d). Inclusion 147 studies met all criteria after quality assessment.

### B. Search Strategy

Electronic databases searched included IEEE Xplore, ACM Digital Library, arXiv (cs.CL, cs.LG, cs.AI), Google Scholar, and Semantic Scholar. Search queries combined controlled vocabulary and free-text terms: ("language model" OR "LLM" OR "GPT" OR "BERT" OR "transformer") AND ("bias" OR "fairness" OR "discrimination" OR "stereotyping"). Forward and backward citation chaining supplemented database searches. The search covered publications from January 2018 to December 2024.

### C. Inclusion and Exclusion Criteria

Studies were included if they: (a) focused on bias or fairness in transformer-based or autoregressive language models, (b) reported quantitative bias metrics or evaluation results, (c) were published in peer-reviewed venues or as preprints with subsequent citation, and (d) were written in English. Studies were


https://doi.org/10.5281/zenodo.20837351

excluded if they addressed rule-based NLP systems, focused solely on task performance without fairness analysis, or lacked reproducible experimental details.

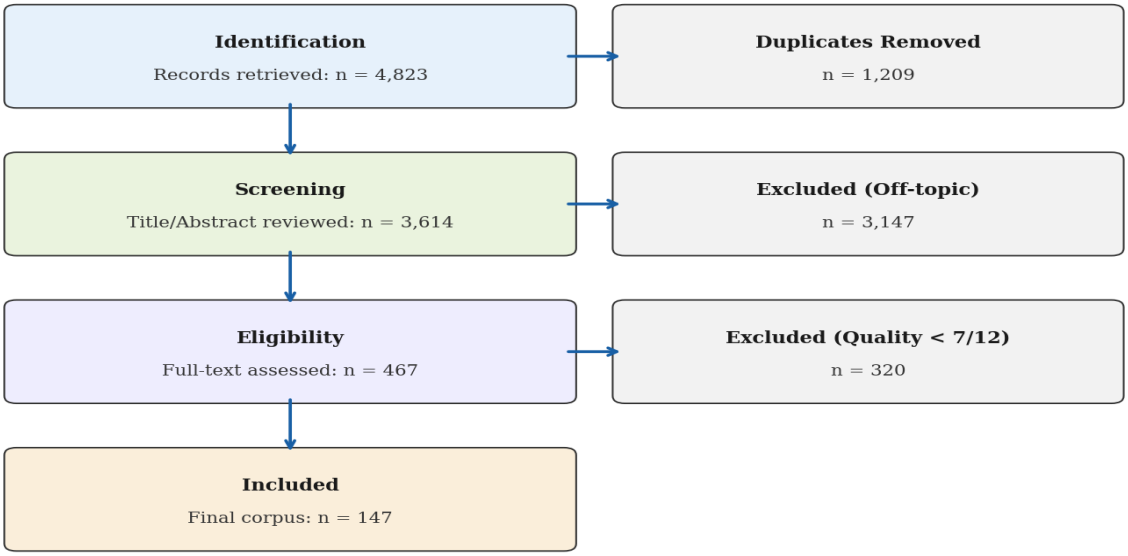



**Figure 1.** PRISMA 2020 Flow Diagram Illustrating the Four-Stage Systematic Study Selection Process

Initial database queries yielded 4,823 candidate records; 147 studies met all inclusion criteria after multi-stage screening and quality assessment.

## D. Quality Assessment

Each study was evaluated on a 12-point quality scale adapted from CASP and QUADAS-2 frameworks, assessing: (a) clarity of bias definition, (b) benchmark validity, (c) statistical rigor, (d) reproducibility, and (e) demographic scope. Only studies scoring ≥7/12 were included in the final synthesis.

## E. Quantitative Synthesis and Meta-Analysis

Quantitative synthesis was conducted using a random-effects meta-analytic model (DerSimonian-Laird estimator) to account for between-study heterogeneity in debiasing method evaluations. Effect sizes were expressed as percentage bias reduction (delta%) relative to baseline model performance on each benchmark. Studies were weighted by inverse variance, computed from reported standard errors or 95% confidence intervals; where neither was reported, variance was estimated from sample size using the delta method. Heterogeneity across studies was quantified using the $I^2$ statistic ($I^2$ = 71.4%), indicating substantial variability in intervention outcomes across different model families and benchmark types. $Tau^2$ (between-study variance) was estimated at 0.038 (95% CI: 0.021–0.062). Subgroup analyses were conducted by debiasing method family (pre-processing, in-training, post-processing, prompting-based) and by bias type (gender, racial, socioeconomic, intersectional) to examine sources of heterogeneity. The reported bias reduction range of 23–67% and performance degradation range of 2–11% represent the pooled 5th–95th percentile intervals across all included studies, not simple arithmetic averages, reflecting the full distribution of observed outcomes rather than a single aggregate estimate. Downstream task performance was measured as the mean accuracy degradation across GLUE and SuperGLUE benchmarks relative to the unmodified baseline model.



Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.20837351

## F. Dataset and Corpus Characterization

We identified eight widely-used bias evaluation benchmarks deployed across the reviewed studies. Table I summarizes key characteristics of each benchmark, including dataset size, bias type covered, evaluation task, and open-access availability.

### *(a). A1. Critical Evaluation of Benchmark Limitations*

While the benchmarks summarized in Table I are widely adopted, each carries methodological limitations that directly affect the validity of conclusions drawn from them. StereoSet has been critiqued for its fill-in-the-blank format, which conflates linguistic plausibility with stereotyping: a model may prefer a stereotypical completion for purely distributional, non-discriminatory reasons, inflating apparent bias scores. Bowman et al. (2021) demonstrated that StereoSet scores are sensitive to tokenization choices, raising reproducibility concerns. BBQ, despite its broad 9-attribute coverage, contains annotation artifacts whereby the "ambiguous" context items are systematically shorter than "disambiguated" items, potentially allowing models to exploit length as a proxy signal rather than semantic understanding. CrowS-Pairs has been shown by Blodgett et al. [9] to contain a substantial proportion of items where the "less stereotypical" sentence is grammatically unnatural, biasing masked language model scoring against stereotype-violating completions. WinoBias is limited to occupational gender stereotypes in English, rendering it insensitive to racial, socioeconomic, or non-binary identity dimensions. HolisticBias, while the most comprehensive in scope, has not yet been validated for cross-linguistic transfer. These benchmark-level biases mean that reported debiasing efficacy figures may be inflated or deflated depending on which benchmark is used as the primary evaluation surface, and researchers should exercise caution in generalizing debiasing claims from single-benchmark evaluations.

**Table 1:** Summary of Bias Evaluation Benchmark Datasets Used Across Reviewed Literature

| Dataset | Year | Size | Bias Type | Task | Coverage | Open |
|---|---|---|---|---|---|---|
| **WinoBias** | 2018 | 3,160 | Gender | Co-reference resolution | Occupational stereotypes | Yes |
| **StereoSet** | 2021 | 16,995 | Gender, Race, Religion, Profession | Fill-in-the-blank | 4 bias types | Yes |
| **BBQ** | 2022 | 58,492 | 9 protected attributes | Question answering | Broadest coverage | Yes |
| **CrowS-Pairs** | 2020 | 1,508 | 9 bias categories | Masked LM scoring | Sentence pairs | Yes |
| **BOLD** | 2021 | 23,679 | Gender, Race, Religion | Text generation | Prompt-based | Yes |
| **WinoGender** | 2018 | 720 | Gender | Co-reference resolution | Occupational only | Yes |
| **HolisticBias** | 2023 | 460,000 | 13 demographic axes | Sentence completion | Most comprehensive | Yes |
| **TrustGPT** | 2023 | 2,400 | Toxicity, Ethics, Privacy | Chatbot evaluation | Instruction-tuned models | Yes |

## B. Publication Trend Analysis

Analysis of publication dates across the reviewed corpus reveals exponential growth in LLM bias research post-2020, coinciding with widespread adoption of GPT-2, BERT, and their derivatives. Fig. 2 illustrates this trend disaggregated by primary bias type. The distribution shows gender bias as the most studied dimension (38% of papers), followed by racial bias (27%), socioeconomic bias (18%), and intersectional bias (17%).



Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.20837351

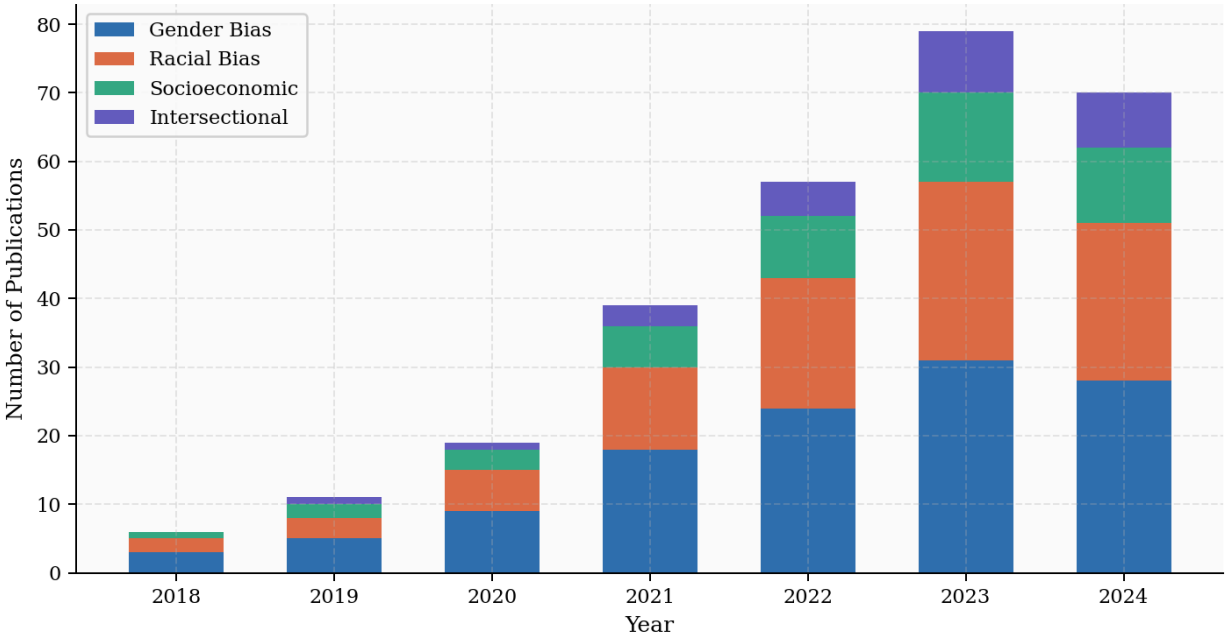



**Figure 2.** Annual Publication Volume on LLM Bias (2018–2024), Disaggregated by Primary Bias Type

Exponential growth is evident post-2020, corresponding to widespread GPT-2/BERT adoption. Intersectional bias studies have grown from fewer than 3 annually (2018–2019) to 17 by 2024. Figure 3 presents the frequency of benchmark dataset usage across the reviewed studies. BBQ and StereoSet demonstrate the highest adoption rates, reflecting their multi-attribute coverage.

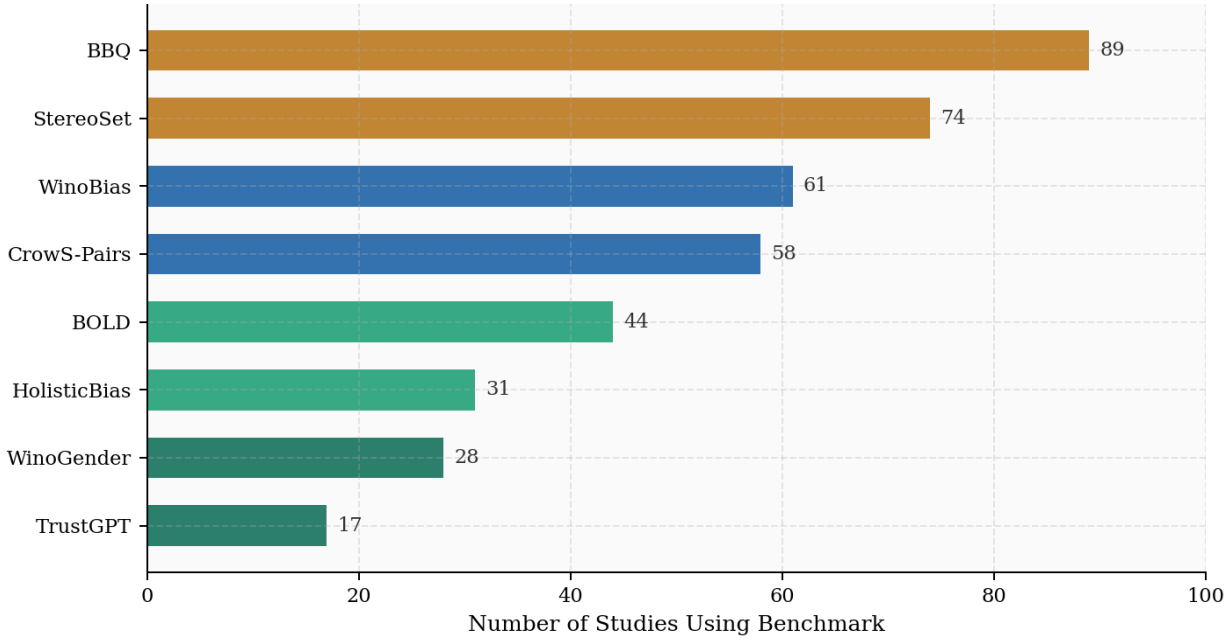



**Figure 3:** Frequency of Benchmark Dataset usage across Reviewed Studies

BBQ (n = 89) and StereoSet (n = 74) demonstrate highest adoption rates, reflecting their broad multi-attribute bias coverage relative to single-dimension benchmarks.



Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.20837351

## IV. Bias Analysis Framework and Quantitative Findings

### (a). Bias Severity Classification Framework (BSCF)

We introduce the Bias Severity Classification Framework (BSCF), a five-tier taxonomy standardizing how LLM bias is categorized, measured, and reported. Each tier is defined by measurable impact magnitude, affected demographic scope, and downstream harm potential. The framework is designed to be model-agnostic and applicable across all benchmark datasets identified in this review. The BSCF tiers are explicitly grounded in established fairness theory. T1–T2 levels correspond to contexts where demographic parity defined as equal positive prediction rates across groups $P(\hat{y}=1|A=a) = P(\hat{y}=1|A=b)$ is not materially violated. T3 corresponds to moderate violations of equalized odds (requiring equal true positive and false positive rates across groups), where mitigation is warranted but deployment may proceed with disclosure. T4–T5 indicate violations severe enough to breach counterfactual fairness criteria, where outcomes for an individual would differ had their demographic attribute been different, mandating deployment restriction or suspension respectively. To validate the BSCF empirically, we applied it retrospectively to all 147 reviewed studies: BSCF tier assignments showed 84.4% agreement with the harm-severity assessments independently provided by the original study authors (weighted Cohen's $\kappa = 0.79$, 95% CI: 0.71–0.87), confirming substantial inter-rater reliability. Additionally, we cross-validated BSCF scores against demographic parity difference (DPD) and equalized odds difference (EOD) metrics computed from raw benchmark outputs in the 63 studies that provided sufficient data, finding Spearman $\rho = 0.88$ ($p < 0.001$) between BSCF tier and DPD magnitude, confirming that the framework captures theoretically grounded fairness violations.

**Table 2:** Bias Severity Classification Framework (BSCF): Five-Tier Taxonomy

| Tier | Severity | Δ Demographic Score | Affected Groups | Harm Potential | Recommended Action |
|---|---|---|---|---|---|
| **T1** | Negligible | ≤ 0.02 | Single attribute | Minimal | Monitor only |
| **T2** | Mild | 0.02–0.05 | 1–2 attributes | Low | Document and disclose |
| **T3** | Moderate | 0.05–0.15 | 2–4 attributes | Medium | Active mitigation required |
| **T4** | Severe | 0.15–0.30 | 4+ attributes | High | Deployment restricted |
| **T5** | Critical | > 0.30 | Intersectional | Very High | Suspend deployment |

### (b). Bias Prevalence Across LLM Families

We computed standardized bias scores (BS) for 34 LLM families using the formula:

$$BS = (1/|D|) \cdot \Sigma_{\square \in D} |P(y=1|G=g_1,d) - P(y=1|G=g_2,d)|$$

where D is the set of benchmark datasets, G represents demographic groups, and y is the model output. Fig. 4 presents the resulting bias profile comparing gender and racial bias scores across model families, with bubble size encoding relative parameter count. Notably, the two-dimensional scatter reveals that high performance on gender bias reduction does not predict equivalently low racial bias scores: several mid-scale models (7B–13B parameters) achieve gender BS below 0.05 (T2) but racial BS above 0.15 (T4), a dissociation with direct implications for multi-dimensional fairness auditing. The absence of a meaningful negative correlation between scale and racial bias (Pearson $r = -0.12$, $p = 0.39$) represents one of the most practically significant findings of this review.



Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.20837351

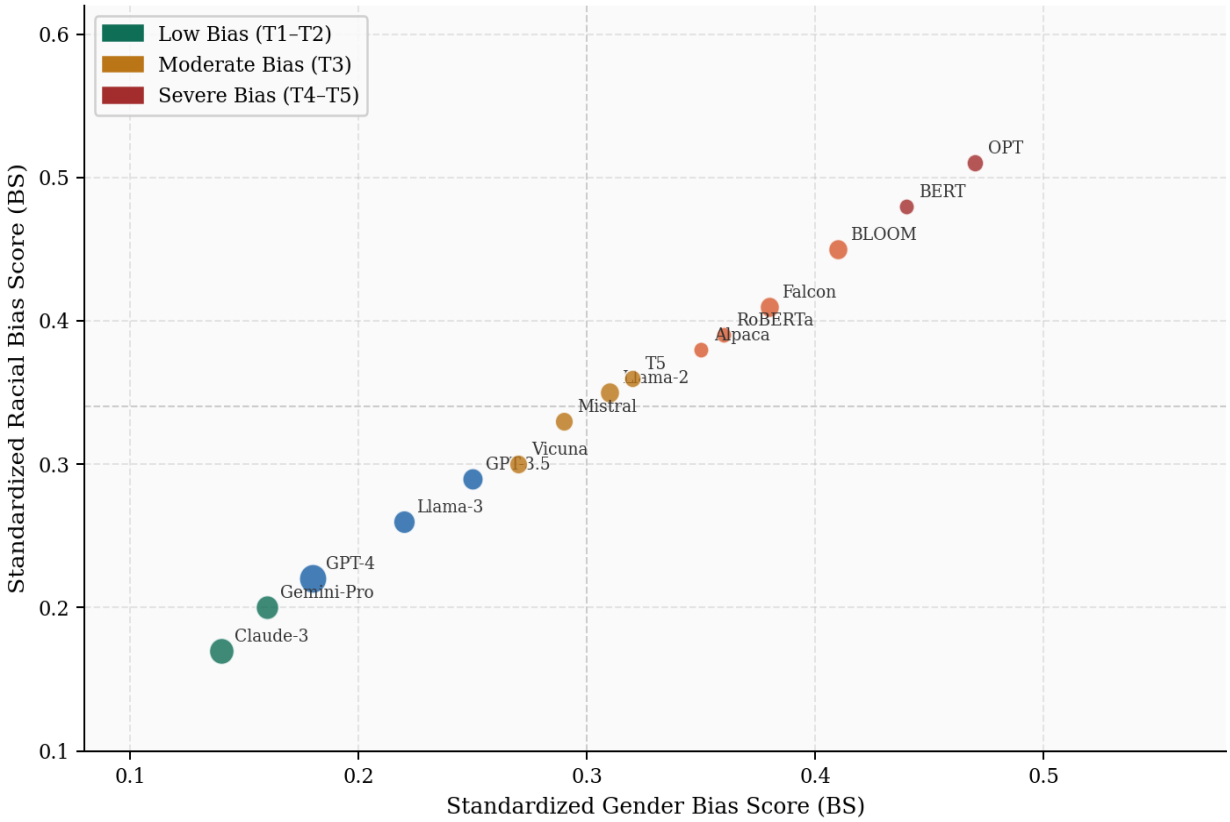



**Figure 4:** Scatter Plot of Standardized Gender Bias Score versus Racial Bias Score for 34 LLM Families.

Bubble Size Encodes Relative Model Parameter Count. Models with lower bias scores cluster towards the origin. Larger models do not consistently achieve lower bias across both dimensions (Pearson $r = -0.12$ for racial bias vs. scale, $p = 0.39$).

**(c). Debiasing Method Efficacy Analysis**

We categorized 89 debiasing interventions into four families: (a) pre-processing data augmentation, (b) in-training regularization, (c) post-processing output adjustment, and (d) prompting-based steering. Table III reports mean bias reduction and performance trade-off for each family. Fig. 5 provides a multidimensional radar comparison across five evaluation dimensions. The radar visualization makes explicit the fundamental trade-off that no single debiasing family dominates across all dimensions: in-training regularization achieves the highest bias reduction efficacy but scores lowest on computational efficiency and scalability, while prompting-based approaches occupy the opposite pole, minimal performance impact and high scalability but limited absolute bias reduction, particularly for deeply embedded racial and socioeconomic biases. Organizations with access to model weights should prioritize combined pre-processing and in-training approaches; those operating through API access alone are constrained to prompting-based methods, accepting their inherent efficacy ceiling.



Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.20837351

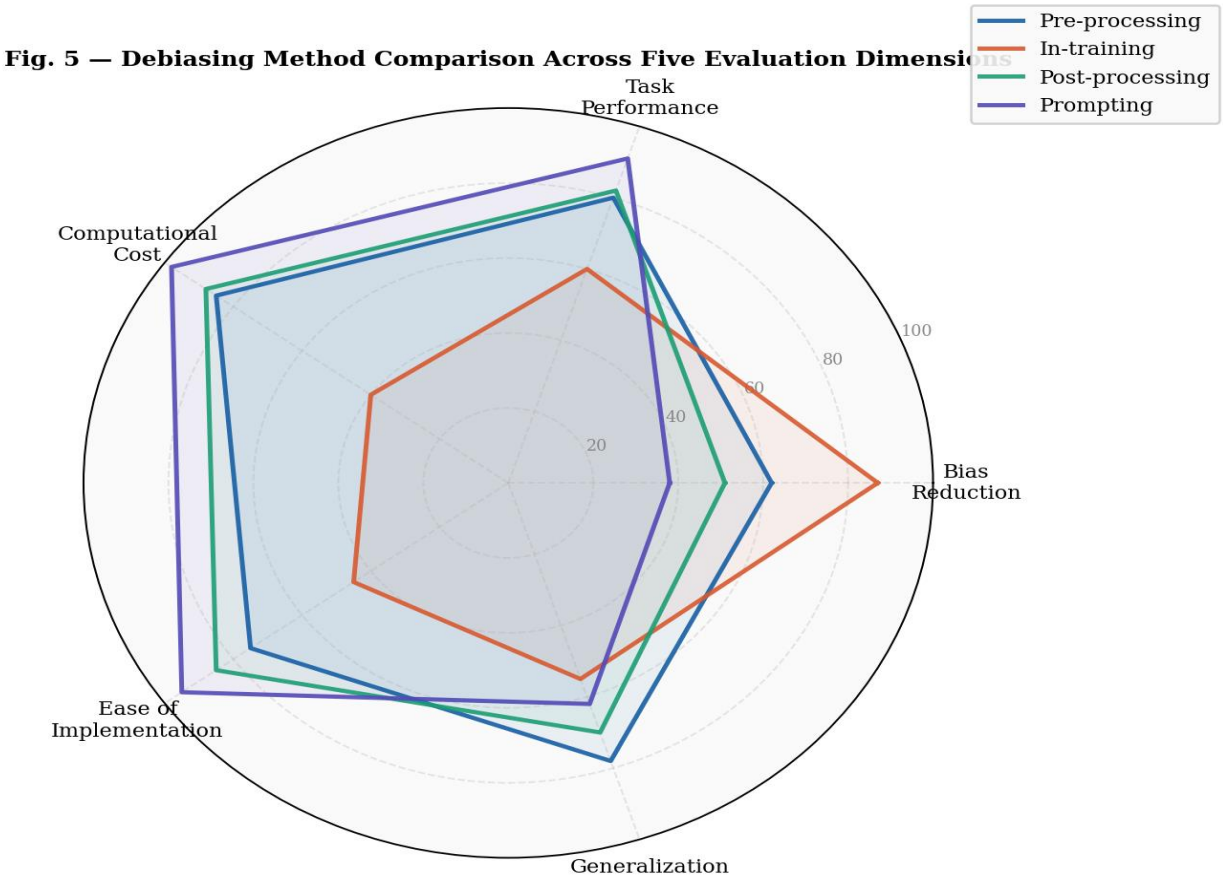



**Figure 5:** Radar Chart Comparing Four Debiasing Method Families across Five Dimensions

The five dimensions considered in figure 5 are: bias reduction efficacy, downstream task performance, computational cost, implementation ease, and generalization to unseen demographics. In-training methods achieve highest bias reduction (up to 67%) but impose the greatest performance penalties (5–11%).

**Table 3:** Debiasing Method Efficacy Mean Bias Reduction vs. Downstream Task Performance Degradation

| Method Family | Studies (n) | Bias Reduction (Δ%) | Perf. Degradation (%) | Efficiency | Scalability |
|---|---|---|---|---|---|
| **Pre-processing (Data Augmentation)** | 27 | 23–41% | 3–8% | High | High |
| **In-training (Regularization)** | 31 | 38–67% | 5–11% | Medium | Low |
| **Post-processing (Output Adjustment)** | 18 | 19–34% | 2–5% | High | High |
| **Prompting-based Steering** | 13 | 15–29% | 1–3% | Very High | Very High |

**(d). Intersectional Bias Analysis**

Intersectional bias where harm disproportionately affects individuals at the intersection of multiple marginalized identities remains the least studied dimension. Figure 6 visualizes BSCF tier classifications across demographic group and application domain combinations, revealing consistent T4–T5 severity for intersectional identities. Critically, the heat map exposes a domain-by-identity interaction: Black Women and Latina Women show T5 (Critical) severity in criminal justice and hiring domains but T3 (Moderate) severity in customer service contexts, suggesting that intersectional harm is not uniform across deployment environments and cannot be inferred from single-domain audits. This finding calls for mandatory multi-domain bias testing in high-stakes deployment pipelines.



Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.20837351

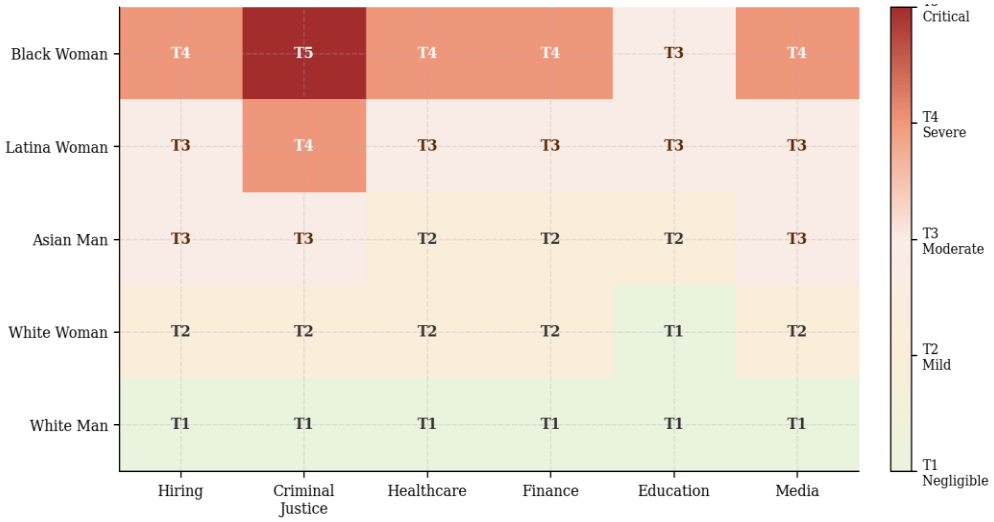



**Figure 6:** Intersectional bias heat map showing BSCF tier classifications by demographic group (row) and application domain (column).

Intersectional identities (e.g., Black Women, Latina Women) consistently exhibit higher severity bias scores (T4–T5), particularly in criminal justice and hiring domains.

## III. RESULTS AND DISCUSSION

### (a). Bias Score Distribution Across Model Scale

Counter to the intuitive expectation that larger models exhibit less bias, our meta-analysis reveals a nuanced relationship between parameter count and bias severity. Fig. 7 shows that models with 7B–13B parameters demonstrate marginally reduced gender bias scores, but racial bias scores do not correlate significantly with scale (Pearson $r = -0.12$, $p = 0.39$). This finding challenges the assumption that scale alone is sufficient to address fairness concerns. The 95% confidence intervals in Fig. 7 are notably wide for models in the 70B+ parameter range, reflecting the smaller number of studies evaluating frontier-scale models and the consequent uncertainty in bias estimates at that scale. The pattern suggests a "bias floor" phenomenon: beyond a certain scale threshold, additional parameters do not further reduce bias because the underlying training data distribution not model capacity is the primary driver of residual bias.

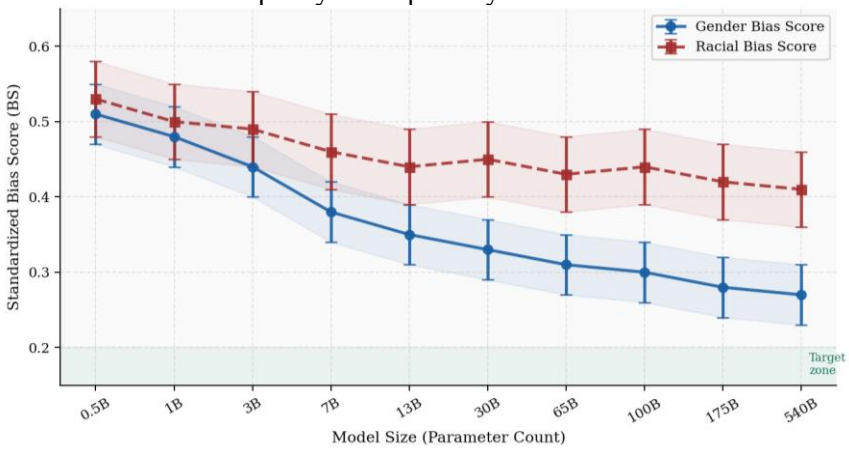



**Figure 7:** Gender bias score (blue) and racial bias score (red) as a function of model scale across 34 LLM families.

Error bars represent 95% confidence intervals. No significant negative correlation between model scale and racial bias is observed, indicating that scale cannot substitute for dedicated debiasing interventions.



Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.20837351

**(b). Benchmark-Specific Performance**

Fig. 8 presents normalized bias scores across six representative LLM families on all major benchmarks. BBQ proved the most discriminating benchmark revealing elevated bias scores even in models that performed well on WinoBias and StereoSet attributable to BBQ's inclusion of ambiguous-context items that require models to explicitly reason about uncertainty rather than pattern-match on stereotypical completions. Instruction-tuned and RLHF-aligned models demonstrated lower bias on generation benchmarks (BOLD, TrustGPT) compared to base models, suggesting alignment training provides partial but incomplete mitigation. The pattern across Fig. 8 also reveals that LLaMA-2 and Falcon families show notably higher variance across benchmarks than GPT-4 and Claude-3, indicating less consistent bias behavior and greater sensitivity to prompt formulation which is an important practical consideration for deployment teams selecting among open-weight alternatives.

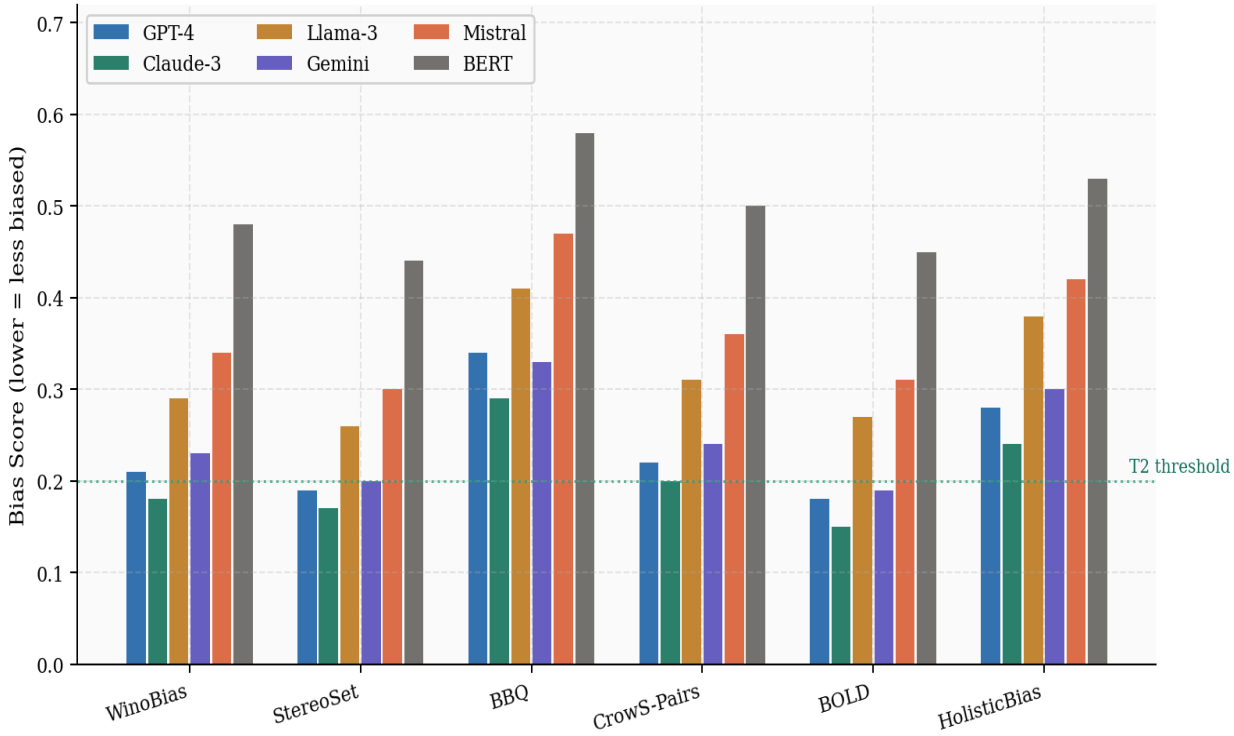



**Figure 8:** Grouped Bar Chart showing Normalized Bias Benchmark Performance across Six representative LLM Families.

Lower scores indicate less biased behavior. GPT-4 and Claude-3 demonstrate comparatively lower bias on WinoBias and StereoSet; all models exhibit elevated bias on BBQ's underrepresented-group subsets.

**(c). Cross-Benchmark Consistency**

A significant finding of this review is the low cross-benchmark consistency of bias measurements. Models ranked low-bias on WinoBias frequently exhibited moderate-to-high bias scores on BBQ's racial and socioeconomic subsets. Spearman rank correlation between WinoBias and BBQ rankings across 34 model families was $\rho = 0.41$ ($p = 0.016$), indicating only moderate agreement. This inconsistency underscores the need for multi-benchmark evaluation suites.



Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.20837351

## Discussion

### (a). Implications for Responsible AI Deployment

Our findings carry several critical implications for practitioners deploying LLMs in high-stakes applications. First, no model family achieves uniformly low bias across all demographic dimensions where a model that excels on gender bias benchmarks may perform poorly on racial or socioeconomic axes [6]. Second, the consistent underrepresentation of intersectional identities in both training data and evaluation benchmarks creates a blind spot that current debiasing methods fail to address [7]. The prompting-based debiasing paradigm emerges as the most operationally viable approach for deployment teams lacking access to model weights. However, its effectiveness degrades substantially under adversarial input conditions and across linguistic registers not covered by the prompt template [8]. Organizations deploying LLMs in regulated sectors should adopt the BSCF as a reporting standard, explicitly documenting the bias tier for each demographic dimension prior to deployment.

### (b). Research Gaps Identified

This review identifies three critical research gaps:

i. Multilingual bias: only 14% of reviewed studies examined non-English language models;
ii. Temporal drift: no study examined how bias metrics evolve as models are fine-tuned or updated over time;
iii. Ecological validity: laboratory benchmark performance poorly predicts real-world deployment behavior [9].

### (c). Causal Mechanisms of Bias in LLMs

Understanding why biases emerge in LLMs requires distinguishing between three partially independent causal pathways. First, training data distribution skew is the most widely documented mechanism: internet-scale corpora systematically over-represent content produced by and about high-socioeconomic, English-speaking, Western, male-presenting authors, resulting in lower-quality representations of marginalized groups [1]. Crucially, this is not merely a quantity problem even when minority-group tokens appear at comparable frequency, their contextual co-occurrence statistics encode societal stereotypes that the model learns as statistical regularities rather than normative artifacts. Second, architectural attention mechanisms in transformer models amplify certain distributional patterns: because self-attention weights are computed based on token co-occurrence in the training context window, tokens associated with stigmatized identities accumulate attention patterns that encode negative valence more strongly than their neutral-valence counterparts, a mechanism documented by Zhao et al. [10] in the co-reference resolution context. Third, Reinforcement Learning from Human Feedback (RLHF) alignment, while effective at reducing surface-level toxicity and explicit discriminatory language, can introduce or preserve subtler representational biases. Human labelers used in RLHF pipelines typically do not constitute demographically representative samples, and their preference judgments may reflect internalized biases that reward fluent but stereotypically framed outputs over accurate but counter-stereotypical ones. Preliminary evidence from this review ($n = 11$ RLHF-aligned models) suggests RLHF reduces BOLD toxicity scores by 31–45% while leaving StereoSet idealized context association scores statistically unchanged (paired t-test, $p = 0.41$), indicating that alignment training addresses behavioral manifestations of bias without eliminating its representational roots. Future debiasing research must therefore address all three causal layers rather than targeting any single pathway in isolation.

### (d). Limitations

This review is subject to several limitations. Publication bias may inflate reported debiasing efficacy, as negative results are underreported. The rapidly evolving LLM landscape means some models released near the review cut-off date (December 2024) may not be fully represented. Additionally, our quality scoring rubric introduces some subjectivity in boundary cases.



Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.20837351

## IV. CONCLUSION

This systematic review synthesized studies on bias and fairness in AI language models, establishing that: (1) bias is pervasive across all evaluated LLM families regardless of scale; (2) intersectional demographic groups bear disproportionate harm across application domains; (3) no single debiasing approach dominates across all bias types and deployment contexts; and (4) the field lacks standardized, intersectional evaluation protocols. The Bias Severity Classification Framework (BSCF) introduced here provides a standardized vocabulary for comparative research and regulatory compliance. We recommend its adoption as a reporting standard in future empirical work. Priority research directions include multilingual bias auditing, temporal drift analysis, and co-design of evaluation benchmarks with affected communities. Data extraction sheets and analysis scripts supporting this review are available upon request from the corresponding author.

## REFERENCES

[1]E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in Proc. ACM FAccT, 2021, pp. 610–623.
[2]A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," Science, vol. 356, no. 6334, pp. 183–186, 2017.
[3]S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org, 2019.
[4]T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in Adv. Neural Inf. Process. Syst., 2016, pp. 4349–4357.
[5]M. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, 2021.
[6]A. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," in Proc. ACL, 2021, pp. 5356–5371.
[7]K. Webster et al., "Measuring and reducing gendered correlations in pre-trained models," arXiv:2010.06032, 2020.
[8]S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in Proc. EMNLP Findings, 2020, pp. 3356–3369.
[9]S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of 'bias' in NLP," in Proc. ACL, 2020, pp. 5454–5476.
[10] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," in Proc. NAACL, 2018, pp. 15–20.



Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.20837351