Debiasing Static Embeddings' Impact on Fairness-Accuracy Trade-off in Contextualized Models

Assignee Research

June 11, 2026

Abstract

Rapid advancements of large language models (LLMs) have enabled the processing, understanding, and generation of human-like text, with increasing integration into systems that touch our social sphere. Despite this success, these models can learn, perpetuate, and amplify harmful social biases. In this paper, we present a comprehensive survey of bias evaluation and mitigation techniques for LLMs. We first consolidate, formalize, and expand notions of social bias and fairness in natural language processing, defining distinct facets of harm and introducing several desiderata to operationalize fair

1 Introduction

This paper examines: Bias and Fairness in Large Language Models: A Survey. Research question: What is the impact of debiasing methods derived from static embeddings on the fairness-accuracy trade-off in contextualized models evaluated on the BiasBios dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.2/10.

3 Results

9 papers retrieved. 9 claims extracted; 8 independently verified. Quality review score: 8.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Large language models (LLMs) enable the processing, understanding, and generation of human-like text.	✓	0.26
LLMs are increasingly integrated into systems that touch the social sphere.	×	0.14
LLMs can learn, perpetuate, and amplify harmful social biases.	✓	0.21
The paper proposes three taxonomies: two for bias evaluation (metrics and datasets) and one for mitigation.	✓	0.21
The taxonomy of metrics for bias evaluation organizes metrics by the levels at which they operate: embeddings, probabili	✓	0.28
The taxonomy of datasets for bias evaluation categorizes datasets by their structure as counterfactual inputs or prompts	✓	0.28
The taxonomy of datasets for bias evaluation identifies targeted harms and social groups.	✓	0.26
The authors release a consolidation of publicly-available datasets for bias evaluation.	✓	0.21
The taxonomy of bias mitigation techniques classifies methods by intervention stage: preprocessing, in-training, intra-	✓	0.27

References

- https://doi.org/10.48550/arxiv.2402.06196
- https://doi.org/10.48550/arxiv.2309.00770
- https://doi.org/10.1145/3534678.3539396