Error Analysis Effects on Fairness Metrics in Discrimination-Aware Language Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does error analysis impact the fairness metrics (e.g., demographic parity, equal opportunity) in discriminatory-aware machine learning models when evaluated on language model benchmarks. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Multi-Objective Evaluation Framework for Analyzing Utility-Fairness Trade-Offs in Machine Learning Systems. Research question: How does error analysis impact the fairness metrics (e.g., demographic parity, equal opportunity) in discriminatory-aware machine learning models when evaluated on language model benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

16 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
Fairness criteria in machine learning include demographic parity, equality of opportunity, equalized odds, and predictiv	×	0.12
Sources of bias in ML systems can stem from data under-representation, algorithms prioritizing accuracy over fairness, o	×	0.04
Perspectives of fairness include individual fairness, group fairness, and subgroup fairness.	×	0.08
Methodologies to enforce fairness involve preprocessing data, in-processing adjustments, and post-processing prediction	×	0.10
To the best of the authors' knowledge, there are no existing frameworks that enable a comprehensive comparison of ML sys	×	0.09
The presented work introduces an evaluation framework supported by Multi-Objective Optimization (MOO) principles for com	✓	0.25
The framework was applied to the analysis of ML systems for three medical imaging tasks.	×	0.10
In the benchmark results, System1 has a Utility-Diagnostic (UD) score of 0.54.	×	0.04
In the benchmark results, System2 has a Utility-Diagnostic (UD) score of 0.64.	×	0.04
In the benchmark results, System1 has an OS score of 0.45.	×	0.02
In the benchmark results, System2 has an OS score of 0.05.	×	0.02
In the benchmark results, System1 has an HV score of 0.55.	×	0.02
In the benchmark results, System2 has an HV score of 0.21.	×	0.02
In the benchmark results, System1 has an ONVG score of 8 and an ONVGR score of 0.80.	×	0.03
In the benchmark results, System2 has an ONVG score of 2 and an ONVGR score of 0.66.	×	0.03

References

- http://arxiv.org/abs/2601.19035v1

- http://arxiv.org/abs/2503.11120v2
- http://arxiv.org/abs/2406.17974v3