Fairness Alignment Gains in Code Generation via Tabular Data Rebalancing Strategies

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: Can rebalancing strategies derived from tabular data theory improve the fairness alignment scores of code generation models when evaluated on imbalanced programming task datasets. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Understanding and Mitigating Bias Inheritance in LLM-based Data Augmentation on Downstream Tasks. Research question: Can rebalancing strategies derived from tabular data theory improve the fairness alignment scores of code generation models when evaluated on imbalanced programming task datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.2/10.

3 Results

15 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 2.2/10.

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
The proportion of negative adjectives decreases across all bias ratios for the Spanish culture after adding bias augment	×	0.06
For the Arabic culture, noticeable increases in the proportion of negative adjectives are observed at higher bias data r	×	0.03
Bias inheritance gets amplified over multiple rounds and eventually extends to majority groups.	×	0.08
For classification, performance declines across all demographic groups over multiple rounds.	×	0.02
For hiring recommendations, the proportion of Arabic candidates steadily decreases, while Spanish candidates become incr	×	0.01
For salary recommendations, predicted salaries for male candidates rise over time, while those for female candidates dec	×	0.01
The model's bias toward minority groups accumulates and spreads over time, leading to a decline in performance across al	×	0.04
In the salary recommendation task, the average predicted salaries for female candidates decreased, while the average pre	×	0.02
The average predicted salaries for female candidates decreased, while the average predicted salaries for male candidates	×	0.00
The proportion of Arabic candidates steadily decreases, while Spanish candidates become increasingly favored in hiring r	×	0.01

References

- http://arxiv.org/abs/2409.05215v1
- http://arxiv.org/abs/2308.08638v2
- http://arxiv.org/abs/2502.04419v3