Causal Data Augmentation Effects on Harmlessness and Reasoning in Small Language Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the impact of causal data augmentation on the trade-off between harmlessness rates and reasoning accuracy in small language models evaluated on BigBench Hard. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: SLM-Bench: A Comprehensive Benchmark of Small Language Models on Environmental Impacts–Extended Version. Research question: What is the impact of causal data augmentation on the trade-off between harmlessness rates and reasoning accuracy in small language models evaluated on BigBench Hard?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

11 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 7.5/10.

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.

References

- http://arxiv.org/abs/2508.15478v2
- http://arxiv.org/abs/2601.08844v1
- http://arxiv.org/abs/2510.00071v2