# Causal Data Augmentation and Adversarial Training for Robust Language Model Reasoning

Assignee Research

June 9, 2026

**Abstract**

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does causal data augmentation compare to adversarial training in improving the robustness of large language models on reasoning benchmarks. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Improving Commonsense Causal Reasoning by Adversarial Training and Data Augmentation. Research question: How does causal data augmentation compare to adversarial training in improving the robustness of large language models on reasoning benchmarks?.

## 2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

## 3 Results

12 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.7/10.

## 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 |
|---|---|---|
| Training with augmented data in addition to the original COPA data boosts model performance over training on just the Ba | × | 0.08 |
| All augmentation methods on their own seem to be able to provide at least a decent learning signal, with GPT-2 and all c | × | 0.03 |
| When training with all augmentation data, models on average reach as high as 77% on the dev and test sets, and around 73 | × | 0.04 |
| When adding the original COPA to the augmented data sets, we consistently achieve higher model performance. | × | 0.06 |
| A single one of our models trained on COPA alone achieves 91% on the test set, however, this comes at the price of a sev | × | 0.02 |
| The single best model according to the test set is based on GPT-2-generated data in addition to the COPA training data, | × | 0.08 |
| When training on all data combined, the mean results on the COPA test set are only slightly lower. | × | 0.05 |
| The first approach is an application of an adversarial example generation through perturbations of the original COPA dat | × | 0.05 |
| The second approach consists of gathering causally linked clauses from the web with the help of a Penn Discourse TreeBan | × | 0.10 |
| Both approaches lead to varying improvements on the performance of the RoBERTa model exhibited by higher average accurac | × | 0.03 |

## References


- http://arxiv.org/abs/2504.07887v2
- http://arxiv.org/abs/2101.04966v1
- http://arxiv.org/abs/2407.15549v3