Causal vs. Correlational Synthetic Data Augmentation for Few-Shot Logical Deduction in LLMs

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does causal synthetic data augmentation compare to correlational methods in improving the reasoning accuracy of large language models on few-shot logical deduction benchmarks. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.9/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: FlexKBQA: A Flexible LLM-Powered Framework for Few-Shot Knowledge Base Question Answering. Research question: How does causal synthetic data augmentation compare to correlational methods in improving the reasoning accuracy of large language models on few-shot logical deduction benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.9/10.

3 Results

14 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 3.9/10.

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
FlexKBQA achieved an Exact Match (EM) score of 62.8 and an F1 score of 69.4 on the GrailQA test set with only 25 annotat	×	0.03
FlexKBQA outperforms the previous state-of-the-art model that adopted the entity linking results from ELQ (Li et al. 202	×	0.09
The performance of the model trained by synthetic data is significantly worse than that trained by real data.	×	0.10
The model performance in the 'real program, synthetic question' setting is just slightly weaker than the model trained b	×	0.07
FlexKBQA achieved an EM score of 60.6 and an F1 score of 67.2 on the GrailQA dataset with 100 shots.	×	0.03
FlexKBQA achieved an EM score of 70.29 and an F1 score of 67.20 on the GrailQA dataset with 60 shots.	×	0.02
FlexKBQA achieved an EM score of 46.83 on the EmbedKGQA dataset with 100 shots.	×	0.02
FlexKBQA achieved an EM score of 70.04 and an F1 score of 66.61 on the GrailQA dataset with 10 shots.	×	0.02

References

- http://arxiv.org/abs/2510.21391v1
- http://arxiv.org/abs/2308.12060v3
- http://arxiv.org/abs/2407.08029v1