Synthetic-to-Real Data Ratio Effects on Tabular Foundation Model F1 Score Variance in CausalMixFT

Assignee Research

June 12, 2026

Abstract

Large language models (LLMs), exemplified by ChatGPT, have gained considerable attention for their excellent natural language processing capabilities. Nonetheless, these LLMs present many challenges, particularly in the realm of trustworthiness. Therefore, ensuring the trustworthiness of LLMs emerges as an important topic. This paper introduces TrustLLM, a comprehensive study of trustworthiness in LLMs, including principles for different dimensions of trustworthiness, established benchmark, evaluation, and analysis of trustworthiness for mainstream LLMs, and discussion of open challenges and f

1 Introduction

This paper examines: TrustLLM: Trustworthiness in Large Language Models. Research question: How does the ratio of synthetic-to-real data in CausalMixFT affect the F1 score variance of tabular foundation models on TabFact across multiple random seeds?.

2 Methodology

Systematic literature search across multiple databases yielded 6 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.8/10.

3 Results

6 papers retrieved. 7 claims extracted; 6 independently verified. Quality review score: 7.8/10.

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
TrustLLM proposes a set of principles for trustworthy LLMs spanning eight different dimensions.	✓	0.19
TrustLLM establishes a benchmark covering six dimensions: truthfulness, safety, fairness, robustness, privacy, and machi	✓	0.19
The TrustLLM study evaluates 16 mainstream LLMs.	✓	0.15
The TrustLLM benchmark consists of over 30 datasets.	×	0.10
In general, trustworthiness and utility (functional effectiveness) in LLMs are positively related.	✓	0.21
Proprietary LLMs generally outperform most open-source counterparts in terms of trustworthiness.	✓	0.30
A few open-source LLMs perform close to proprietary ones in terms of trustworthiness.	✓	0.23

References

- https://doi.org/10.48550/arxiv.2401.05561
- https://doi.org/10.1038/s41398-023-02592-2
- https://doi.org/10.18653/v1/2020.repl4nlp-1