Alignment Stability of LLMs on TabPFN Synthetic Data Under Distribution Shifts

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: Does the alignment stability of LLMs trained on TabPFN-generated synthetic data with causal structure degrade under distribution shifts (e.g., domain adaptation tasks) compared to LLMs trained on. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Improving TabPFN's Synthetic Data Generation by Integrating Causal Structure. Research question: Does the alignment stability of LLMs trained on TabPFN-generated synthetic data with causal structure degrade under distribution shifts (e.g., domain adaptation tasks) compared to LLMs trained on standard TabPFN data, measured by response fidelity scores on SuperBench or MMLU?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

4 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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
Synthetic data quality is evaluated using Correlation Matrix Difference (CMD), k-Marginal Total Variation Distance (kMTV	×	0.05
CMD quantifies structural dependency preservation by computing the Frobenius norm of the difference between real and syn	×	0.03
The study replaces Pearson correlation with Spearman's rank correlation for numerical–numerical pairs to capture monoton	×	0.01
Mixed correlation matrices combine Cramr's V for categorical–categorical pairs, the correlation ratio η for categorical	×	0.00
kMTVD measures pairwise distributional fidelity by discretizing continuous variables into 20 quantile-based bins.	×	0.03
kMTVD is calculated as the mean Total Variation Distance across all variable pairs.	×	0.01
NNAA assesses privacy preservation by quantifying the distinguishability between synthetic and real data based on neares	×	0.06
In the NNAA metric, values near 0.5 indicate that synthetic and real data are hard to distinguish.	×	0.04
Statistical significance of differences between conditioning strategies is assessed using the Wilcoxon signed-rank test	×	0.02
Effect sizes are quantified using the Hodges–Lehmann estimator.	×	0.01
Experiments are conducted on three dataset classes: fully controlled hand-crafted settings, public benchmark datasets, a	×	0.03
A custom four-variable Structural Causal Model (SCM) containing a collider was designed to evaluate TabPFN's sensitivity	×	0.13
TabPFN is pre-trained on millions of synthetic datasets derived from Structural Causal Models (SCMs).	×	0.07
Generation methods that ignore causal dependencies may create spurious correlations that differ from the true data-gener	×	0.09
Inaccurate estimation of treatment effects from flawed synthetic data could lead to costly trials on ineffective drugs o	×	0.06

References

- http://arxiv.org/abs/1907.02664v2
- http://arxiv.org/abs/2603.10254v1
- http://arxiv.org/abs/2406.15126v1