Causal Structure Complexity and Synthetic Data Quality in Tabular Foundation Models

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the causal structure complexity in Structural Causal Models (SCMs) affect the trade-off between synthetic data quality and fine-tuning efficiency when scaling tabular foundation models. 12 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: TabularARGN: A Flexible and Efficient Auto-Regressive Framework for Generating High-Fidelity Synthetic Data. Research question: How does the causal structure complexity in Structural Causal Models (SCMs) affect the trade-off between synthetic data quality and fine-tuning efficiency when scaling tabular foundation models?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

13 papers retrieved. 12 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
NADE architecture is not adapted to handle multi-categorical datasets with variables of cardinality $\geq$2, nor mixed-type d	×	0.06
Bayesian Networks (BNs) represent dependencies between variables using a directed acyclic graph and can be used for gene	×	0.06
Variational Autoencoders (VAEs) explicitly model joint distributions by learning a latent representation of the data.	×	0.04
Generative Adversarial Networks (GANs) implicitly approximate joint distributions through an adversarial training proces	×	0.04
Both VAEs and GANs have been adapted to handle mixed-type tabular synthetic data.	×	0.14
Token-based transformers and auto-regressive transformer models have been applied to generate tabular synthetic data.	×	0.11
Auto-regressive transformer models designed for tabular data leverage the inherent structure of tabular data to enhance	✓	0.16
Hybrid approaches that combine transformers with diffusion have been explored to model discrete features.	×	0.04
Diffusion models have been adapted for generating tabular synthetic data and have shown promise in capturing intricate p	×	0.10
Training separate, independent models for each sub-column results in summed validation losses that exceed the converged	×	0.03
Using shared embeddings across columns confers an advantage by capturing and transferring relationships across different	×	0.02
The TabularARGN model for sequential tables is auto-regressive along both the column dimension and the time/sequence dim	×	0.09

References

- `http://arxiv.org/abs/2304.14109v1`

- http://arxiv.org/abs/2305.10308v2
- http://arxiv.org/abs/2501.12012v2