Synthetic Data Generation Techniques and Their Impact on Tabular Foundation Model Alignment

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How do different synthetic data generation techniques influence the alignment of tabular foundation models with real-world distributions, as measured by cross-domain accuracy on TabBench tasks. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Tabular Data Augmentation for Machine Learning: Progress and Prospects of Embracing Generative AI. Research question: How do different synthetic data generation techniques influence the alignment of tabular foundation models with real-world distributions, as measured by cross-domain accuracy on TabBench tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

11 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 4.2/10.

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
Tabular data is heterogeneous, typically containing both dense numerical and sparse categorical attributes.	×	0.05
Tabular data has complicated structure, such as row and column permutation invariance and hierarchical organization, whe	×	0.05
Many TDA tasks involve large-scale table pools, sometimes encompassing millions of tables.	×	0.05
These tables often have inconsistent attribute naming and value formatting, and table pools themselves are dynamic, chan	×	0.02
TDA methods can be broadly categorized into retrieval-based approaches, which involve retrieving data from table pools,	×	0.11
The initial ML model yields sub-optimal results due to insufficient data and numerous missing or incorrect values.	×	0.04
Augmentation can be achieved through retrieval-based methods or generation-based methods that synthesize new data.	×	0.14
The augmented table () combines the original and new data.	×	0.04
After augmentation, evaluation steps evaluate the effectiveness of TDA process.	×	0.08
The result-TDA table enables the scientist to train a more accurate price prediction model.	×	0.04
The original training set () has limited features and records.	×	0.04
The original table has missing and incorrect values.	×	0.05
The augmented table includes additional attributes (columns), records (rows), and corrected values (cells).	×	0.03
The table pools can contain various table sources, including databases and Web tables.	×	0.04
Generative methods include statistical approaches such as MICE, deep generative models like diffusion models, and langua	×	0.09
The augmented analytics market is growing, as indicated by market research reports.	×	0.02

References

- http://arxiv.org/abs/2407.21523v1
- http://arxiv.org/abs/2507.07829v1
- http://arxiv.org/abs/2512.03307v1