

Scaling Synthetic Pretraining Data for Few-Shot Tabular Foundation Models

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: Does scaling synthetic pretraining data volume improve few-shot learning performance of tabular foundation models compared to traditional machine learning baselines on heterogeneous datasets. 11 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Robust Tabular Foundation Models. Research question: Does scaling synthetic pretraining data volume improve few-shot learning performance of tabular foundation models compared to traditional machine learning baselines on heterogeneous datasets?.

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 4.4/10.

3 Results

11 papers retrieved. 11 claims extracted; 1 independently verified. Quality review score: 4.4/10.

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 foundation models (TFMs) rely on in-context learning (ICL) for classification and regression tasks with structured	×	0.11
TFMs can produce high-quality predictions on new datasets in milliseconds when GPU-accelerated.	×	0.07
Training TFMs relies on generating large amounts of diverse synthetic datasets constructed from structural causal models	×	0.07
All current publicly available, competitive TFMs have been pretrained on datasets generated from a fixed prior distribution	×	0.06
Fixed priors underrepresent certain regions of the parameter space, potentially degrading performance on real-world data	×	0.06
State-of-the-art TFMs lag behind tree-based methods on some benchmarks.	×	0.06
The proposed method formalizes adversarial training over the SCM parameter space to allow models to adapt to challenging	×	0.06
The proposed algorithm is named ROBUST TABULAR FOUNDATION MODELS (RTFM).	✓	0.18
Applying RTFM to TabPFN V2 with only 90k additional training datasets significantly improves its ranking on several real	×	0.12
The optimality gap is estimated by sampling a fixed number of generators and datasets and computing the difference between	×	0.09
In the described implementation, the estimated optimality gap could be computed in a matter of seconds when parallelized	×	0.04

References

- <http://arxiv.org/abs/1706.01513v2>
- <http://arxiv.org/abs/2311.10051v1>
- <http://arxiv.org/abs/2512.03307v1>