Adversarial Pretraining of Tabular Foundation Models on Synthetic Data Robustness Benchmarks

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How do tabular foundation models pretrained on adversarial synthetic datasets perform in robustness benchmarks like Tabular Adversarial Robustness Benchmark (TARB), as measured by accuracy. 16 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: Robust Tabular Foundation Models. Research question: How do tabular foundation models pretrained on adversarial synthetic datasets perform in robustness benchmarks like Tabular Adversarial Robustness Benchmark (TARB), as measured by accuracy degradation under perturbations compared to models pretrained on traditional generative priors?.

2 Methodology

Systematic literature search across multiple databases yielded 8 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

8 papers retrieved. 16 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
Tabular foundation models (TFMs) have emerged as a promising direction for classification and regression tasks with stru	×	0.13
TFMs rely on in-context learning (ICL).	×	0.03
TFMs can provide high-quality predictions on new datasets in milliseconds when GPU-accelerated.	×	0.08
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 still lag behind tree-based methods on some benchmarks.	×	0.06
The authors leverage the significant control provided by the data generation process to frame TFM training from an adver	×	0.08
The authors propose an efficient, model-agnostic two-stage adversarial training algorithm for TFMs, called ROBUST TABULA	✓	0.21
The authors apply RTFM to TabPFN V2, showing significant improvement in the ranking of TabPFN on several real-world tabu	×	0.11
Training TFMs relies on generating a large amount of diverse synthetic datasets.	×	0.09
The generation process relies on constructing structural causal models (SCMs) from which datasets can be sampled.	×	0.03
The structure of these SCMs is implicitly parameterized, giving significant control over the data generation process.	×	0.02
The authors formalize adversarial training over the SCM parameter space, allowing the model to adapt to challenging regi	×	0.09
The authors introduce an optimality gap concept and use it to target regions where the TFM underperforms relative to the	×	0.09
The authors use a black-box optimization algorithm to efficiently search the space for parameters with large optimality	×	0.04
For nds = 20 and e = 7, the estimated optimality gap b$\delta$$\theta$i could be computed in a matter of seconds when parallelized, gi	×	0.04

References

- http://arxiv.org/abs/2504.20900v1
- http://arxiv.org/abs/2512.03307v1
- http://arxiv.org/abs/2207.03208v2