What is the performance gap between TFMs pretrained on mixed synthetic-real data and those pretrained exclusiv

Assignee Research

June 10, 2026

Abstract

The development of tabular foundation models (TFMs) has accelerated in recent years, showing strong potential to outperform traditional ML methods for structured data. A key finding is that TFMs can be pretrained entirely on synthetic datasets, opening opportunities to design data generators that encourage desirable model properties. Prior work has mainly focused on crafting high-quality priors over generators to improve overall pretraining performance. Our insight is that parameterizing the generator distribution enables an adversarial robustness perspective: during training, we can adapt the

1 Introduction

This paper examines: Robust Tabular Foundation Models. Research question: What is the performance gap between TFMs pretrained on mixed synthetic-real data and those pretrained exclusively on real data, when evaluated on the TabMNAR benchmark with 30% structural missing data, and how does this gap vary with model capacity (e.g., 100M vs. 1B parameters)?.

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. 13 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 structur	×	0.12
TFMs can produce high-quality predictions on new datasets in milliseconds when GPU-accelerated.	×	0.08
Training TFMs relies on generating diverse synthetic datasets constructed from structural causal models (SCMs).	×	0.07
All current publicly available, competitive TFMs have been pretrained on datasets generated from a fixed prior distribut	×	0.06
Fixed priors in TFM training underrepresent certain regions of the parameter space, potentially degrading performance on	×	0.05
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) and is a model-agnostic two-stage adversarial tr	✓	0.22
Applying RTFM to TabPFN V2 with only 90k additional training datasets significantly improves its ranking on several real	×	0.10
The maximization stage of the proposed method uses a black-box optimization algorithm to search the parameter space for	×	0.05
In the described implementation, estimating the optimality gap with nds=20 and e=7 takes a matter of seconds when parall	×	0.03
The benchmark table includes synthetic dataset configurations with activation functions such as tanh, identity, elu, and	×	0.02
The benchmark table includes synthetic dataset configurations with noise distributions including uniform, exponential, a	×	0.02

References

- http://arxiv.org/abs/2403.13430v2
- http://arxiv.org/abs/2512.03307v1
- http://arxiv.org/abs/2601.21725v2