Scaling Laws of Synthetic-to-Real Data Ratios in Tabular Pretraining

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: How does the computational cost of pretraining with different synthetic-to-real data ratios scale with dataset size, and what is the trade-off between pretraining throughput and downstream task. 8 claims were extracted from source literature; 6 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Accurate predictions on small data with a tabular foundation model. Research question: How does the computational cost of pretraining with different synthetic-to-real data ratios scale with dataset size, and what is the trade-off between pretraining throughput and downstream task accuracy on TabBench OOD sets?.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.4/10.

3 Results

7 papers retrieved. 8 claims extracted; 6 independently verified. Quality review score: 7.4/10.

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 data are ubiquitous across scientific fields, including biomedicine, particle physics, economics, and climate sc	✓	0.26
Gradient-boosted decision trees have dominated tabular data learning for the past 20 years.	✓	0.25
TabPFN outperforms all previous methods on datasets with up to 10,000 samples.	✓	0.22
TabPFN uses substantially less training time than previous methods.	×	0.12
In a classification setting, TabPFN achieves superior performance in 2.8 seconds compared to an ensemble of the stronges	×	0.11
TabPFN is a generative transformer-based foundation model.	✓	0.22
TabPFN supports fine-tuning, data generation, density estimation, and learning reusable embeddings.	✓	0.21
TabPFN is a learning algorithm that was learned across millions of synthetic datasets.	✓	0.21

References

- https://doi.org/10.1038/s41586-024-08328-6
- https://doi.org/10.3390/iot6010013
- https://doi.org/10.1016/j.csbj.2024.07.005