Scaling Behavior of Causal TabPFN vs Standard TabPFN on Reasoning Benchmarks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the scaling behavior of TabPFN with causal structure (e.g., number of features, training time) compare to standard TabPFN when evaluated on a suite of downstream reasoning benchmarks like. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Causal Pre-training Under the Fairness Lens: An Empirical Study of TabPFN. Research question: How does the scaling behavior of TabPFN with causal structure (e.g., number of features, training time) compare to standard TabPFN when evaluated on a suite of downstream reasoning benchmarks like GSM8K or HellaSwag?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

12 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
The Heart dataset contains 303 samples and 13 features, with sensitive attributes being age and sex.	×	0.01
The Bank dataset contains 5,000 samples and 14 features, with sensitive attributes being education and family.	×	0.01
The Law dataset contains 21,000 samples and 8 features, with sensitive attributes being race and sex.	×	0.01
The Adult dataset contains 48,000 samples and 14 features, with sensitive attributes being race and sex.	×	0.02
Logistic Regression (LR) is evaluated with L2 regularization (C=0.1).	×	0.00
Random Forest (RF) is evaluated with 50 trees of maximum depth 5 and minimum samples per split of 10.	×	0.02
Multi-Layer Perceptron (MLP) is evaluated with one hidden layer of 50 units, L2 regularization (α=0.01), and up to 300 t	×	0.01
TabPFN is evaluated in zero-shot mode with pretraining limits disabled, n_estimators=2, and inference using 5000 subsamp	×	0.03
FT-TabPFN is fine-tuned for 10 epochs via Adam (lr=1$\times$10-5), meta-batch size 1, inner batch size 5000, and cross-entropy	×	0.05
TabPFN is designed for small- to medium-sized tabular datasets (up to roughly 10k samples).	×	0.04
For the Adult and Law datasets, TabPFN is evaluated on subsets of 500, 1k, and 10k samples to analyze how fairness varie	×	0.05
FT-TabPFN and TabPFN exhibit superior flip consistency, reaching 0.82–0.97 across datasets and leading the pack on Heart	×	0.02
FT-TabPFN and TabPFN have the highest accuracy on most datasets (Bank 0.97, Heart/Adult 0.75-0.88).	×	0.04

References

- http://arxiv.org/abs/2511.07236v1
- http://arxiv.org/abs/2603.10254v1
- http://arxiv.org/abs/2601.17912v2