

Synthetic-Pretrained Tabular Foundation Models vs. Tree Ensembles on Zero-Shot TabBench Performance

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How do inference latency and throughput metrics differ between large-scale synthetic-pretrained tabular foundation models and optimized fine-tuned tree ensembles during zero-shot evaluation on. 6 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: Cost-Aware Model Selection for Text Classification: Multi-Objective Trade-offs Between Fine-Tuned Encoders and LLM Prompting in Production. Research question: How do inference latency and throughput metrics differ between large-scale synthetic-pretrained tabular foundation models and optimized fine-tuned tree ensembles during zero-shot evaluation on TabBench?.

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 4.5/10.

3 Results

12 papers retrieved. 6 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 |
|--|----------|------------|
| The study analyzes performance metrics jointly with inference latency and monetary cost through Pareto frontier projecti | ✓ | 0.18 |
| The released artifacts are intended to function as living reference points that facilitate sustainable system evolution. | × | 0.01 |
| The methodology is grounded in operational assumptions that commonly constrain production NLP systems. | × | 0.04 |
| Model selection in production-grade NLP systems is a knowledge-based decision process that considers system-level constr | × | 0.10 |
| Benchmark results serve as reusable decision evidence for principled model selection under realistic deployment regimes. | × | 0.08 |
| The study jointly quantifies predictive quality, inference latency, and economic cost across representative datasets. | × | 0.08 |

References

- <http://arxiv.org/abs/2512.03307v1>
- <http://arxiv.org/abs/2511.02802v3>
- <http://arxiv.org/abs/2602.06370v1>