

# Synthetic-Only vs. Mixed-Data Pretraining for Tabular Foundation Models on Small-Sample Tasks

Assignee Research

June 9, 2026

## Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: Does pretraining tabular foundation models exclusively on synthetic data degrade performance on small-sample downstream tasks compared to mixed-data pretraining strategies. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Causal Data Augmentation for Robust Fine-Tuning of Tabular Foundation Models. Research question: Does pretraining tabular foundation models exclusively on synthetic data degrade performance on small-sample downstream tasks compared to mixed-data pretraining strategies?.

## 2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

## 3 Results

15 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.2/10.

## 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.

## References

- <http://arxiv.org/abs/2601.04110v2>
- <http://arxiv.org/abs/2203.07057v2>
- <http://arxiv.org/abs/2207.03208v2>