Synthetic Data Diversity and Robustness in Tabular Foundation Models for Retail

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the impact of synthetic data diversity on the robustness of tabular foundation models when evaluated against distribution shifts in retail datasets. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Enhancing Robustness of Foundation Model Representations under Provenance-related Distribution Shifts. Research question: What is the impact of synthetic data diversity on the robustness of tabular foundation models when evaluated against distribution shifts in retail datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.5/10.

3 Results

4 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 2.5/10.

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 focuses on confounding shift where label distributions among subpopulations differ in the training and testing	×	0.04
The problem formulation excludes the distribution of predictor variables (X), which are derived from language.	×	0.03
The evaluation framework is designed for binary classification with two subpopulations where Y $\in${0, 1} and Z $\in${0, 1}.	×	0.01
The parameter Ptrain(z = 1) is set equal to Ptest(z = 1) (denoted as Cz) to eliminate confounding factors related to dif	×	0.04
The parameter Ptrain(y = 1) is set equal to Ptest(y = 1) (denoted as Cy) to negate effects of different background posit	×	0.02
The study introduces auxiliary variables αtrain and αtest to measure differences in site-specific class distribution, de	×	0.04
Robustness to distribution shift is measured by the difference between αtrain and αtest.	×	0.11
Model stability is quantified by fitting a linear regression line between the log-transformed αtest and the AUPRC (Area	×	0.01
A lower absolute value of the fitted regression coefficient indicates higher model robustness to confounding shift.	×	0.06
A fitted coefficient value of zero indicates equivalent model performance irrespective of the confounding shift.	×	0.03
Backdoor adjustment calculates P(y|x) by summing P(y|x, z)P(z) over all values of z.	×	0.07
Landeiro and Culotta developed an approach using logistic regression to estimate P(y|X, z = c) in the presence of confou	×	0.03

References

- http://arxiv.org/abs/2312.05435v1

- http://arxiv.org/abs/2512.03307v1
- http://arxiv.org/abs/2307.05284v6