Adversarial Perturbation Magnitude and Robustness in Tabular Foundation Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the correlation between the magnitude of adversarial perturbations in synthetic pretraining data and the subsequent robustness of tabular foundation models against natural distribution shifts. 17 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: Enhancing Robustness of Foundation Model Representations under Provenance-related Distribution Shifts. Research question: What is the correlation between the magnitude of adversarial perturbations in synthetic pretraining data and the subsequent robustness of tabular foundation models against natural distribution shifts?.

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 4.2/10.

3 Results

12 papers retrieved. 17 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.

5 Extracted Claims

Claim	Verified	Confidence
The problem formulation does not include the distribution of predictor variables, X, which in our case is derived from l	×	0.04
The approach for synthetic injection of confounding shift is built upon to develop an evaluation framework for binary cl	×	0.03
The parameters to construct a testing scenario are set as Ptrain(y = 1|z = 0), Ptrain(y = 1|z = 1), Ptrain(z = 1) = Ptes	×	0.01
Equation (3) aims to eliminate a potential confounding factor where the proportion of training examples from each source	×	0.06
Equation (4) is implicitly enforced to negate effects of different background positive rates in the train and test sets.	×	0.01
The objective of these constraints is to focus on shifts related to provenance.	×	0.12
In contrast to the work of Landeiro and Culotta, two auxiliary variables for measuring differences in site-specific clas	×	0.04
During evaluation, desired ranges for variables (1), (2), (3), and αtest are specified.	×	0.01
All combinations of these parameters are applied to govern selection of corresponding samples to construct multiple trai	×	0.03
The goal is to examine a model's robustness to these different degrees of distribution shift, measured by the difference	×	0.12
To quantify robustness, αtest is first log-transformed and a linear regression line is fit against a target evaluation m	×	0.03
The coefficient measures the slope of a line that relates changes in the performance metric of interest to changes in αt	×	0.02
The lower the absolute value of the fitted coefficient, the more robust a model is to confounding shift, with a value of	×	0.04
Backdoor adjustment is a technique to make adjustments on predictions when confounding variable (z) exists, originally p	×	0.08
Backdoor adjustment is defined as P(y|x) = Σ_z P(y|x, z)P(z).	×	0.08
A similar approach was developed by Landeiro and Culotta for text classification in the presence of confounding bias.	×	0.04
A logistic regression model is fit to estimate P(y|X, z = c) as logit(yc) = β0 + β1X + β2zc + ϵ.	×	0.01

References

- http://arxiv.org/abs/2312.05435v1
- http://arxiv.org/abs/2512.03307v1
- http://arxiv.org/abs/2307.05284v6