Scaling Model Size and Robustness to Distribution Shifts in Tabular Data

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the impact of scaling model size on robustness to distribution shifts in tabular data, as measured by accuracy on TableShift benchmarks. 14 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: Benchmarking Distribution Shift in Tabular Data with TableShift. Research question: What is the impact of scaling model size on robustness to distribution shifts in tabular data, as measured by accuracy on TableShift benchmarks?.

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 3.8/10.

3 Results

15 papers retrieved. 14 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
Food insecurity affects more than 10% of households (13.5 million) across the United States in 2021.	×	0.01
The American Community Survey (ACS) is used as the data source for the Food Stamps task.	×	0.02
The Food Stamps task filters data for low-income adults aged 18-62 in households with at least one child.	×	0.02
An income threshold of $30000 is used based on the U.S. poverty threshold for a family with one child.	×	0.02
The ACS includes 10 regions: Puerto Rico; New England (Northeast region); Middle Atlantic (Northeast region); East North	×	0.00
East South Central (South region) is used as the holdout domain for the Food Stamps task.	×	0.02
The benchmark table shows a Δy of -34.49% for the ASSISTments Next Answer Correct task with a p-value of 0.011.	×	0.01
The benchmark table shows a Δy of -11.16% for the College Scorecard Low Degree Completion Rate task with a p-value of 0.	×	0.01
The benchmark table shows a Δy of -6.30% for the ICU Hospital Mortality task with a p-value of 0.008.	×	0.01
The benchmark table shows a Δy of -5.94% for the Hospital Readmission task with a p-value of 0.002.	×	0.01
The benchmark table shows a Δy of -4.48% for the Diabetes task with a p-value of 0.001.	×	0.01
The benchmark table shows a Δy of -3.39% for the ICU Length of Stay task with a p-value of 0.015.	×	0.01
The benchmark table shows a Δy of -2.58% for the Voting task with a p-value of 0.016.	×	0.01
The benchmark table shows a Δy of -4.48% for the Food Stamps task with a p-value of 0.001.	×	0.01

References

- http://arxiv.org/abs/2605.29330v1
- http://arxiv.org/abs/2206.02435v2
- http://arxiv.org/abs/2312.07577v3