

# Scaling Laws of Tabular Foundation Models on Out-of-Distribution Generalization

Assignee Research

June 7, 2026

## Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Do larger tabular foundation models exhibit better generalization performance on out-of-distribution benchmarks like SuperTab compared to smaller models, and what is the scaling relationship between. 11 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: Do larger tabular foundation models exhibit better generalization performance on out-of-distribution benchmarks like SuperTab compared to smaller models, and what is the scaling relationship between model size and accuracy improvement?.

## 2 Methodology

Systematic literature search across multiple databases yielded 14 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

14 papers retrieved. 11 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 a data source for the Food Stamps benchmark task.	×	0.03
The Food Stamps benchmark 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	×	0.01
East South Central (South region) is used as the holdout domain for the Food Stamps task.	×	0.02
The correlation between other forms of shift showed roughly $\rho \approx -0.2$ (see Figure 7).	×	0.05
The datasets in TableShift do not contain personally identifiable information, offensive content, or proprietary informa	×	0.02
For data collected from human subjects, the conditions of collection and the ethics approval under which the data were c	×	0.03
The benchmark tasks in TableShift are based on publicly available data sources.	×	0.12
Some datasets in TableShift require training or authorization but are available to the public.	×	0.04

## References

- <http://arxiv.org/abs/2506.16791v4>
- <http://arxiv.org/abs/2512.03307v1>
- <http://arxiv.org/abs/2312.07577v3>