Scaling Synthetic Tabular Data with Controlled Noise for Zero-Shot Generalization in Foundation Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: To what extent does scaling the size of synthetic tabular datasets with controlled noise levels improve the zero-shot generalization of tabular foundation models on out-of-distribution tasks, as. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: DeepSeq: High-Throughput Single-Cell RNA Sequencing Data Labeling via Web Search-Augmented Agentic Generative AI Foundation Models. Research question: To what extent does scaling the size of synthetic tabular datasets with controlled noise levels improve the zero-shot generalization of tabular foundation models on out-of-distribution tasks, as measured by accuracy on datasets like Yelp or CIFAR-10 tabularized versions?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

13 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
DeepSeq evaluates biological plausibility by matching extracted marker genes to canonical gene sets for known cell types	×	0.03
DeepSeq assesses the accuracy of cell-type predictions relative to ground truth annotations.	×	0.03
The DeepSeq pipeline outputs interpretable logs enabling traceability of decisions from filtering to LLM prompting.	×	0.01
DeepSeq integrates single-cell RNA-seq preprocessing with foundation model-driven cell-type annotation using large lang	×	0.13
The DeepSeq workflow includes filtering, clustering, marker gene extraction, prompting, and structured evaluation.	×	0.02
DeepSeq identifies top marker genes by ranking genes within each cluster.	×	0.00
When using GPT-4o, DeepSeq performs a web search via an OpenAI Agent to augment the prompt.	×	0.07
DeepSeq performs filtering using standard thresholding (e.g., $\geq$200 genes per cell), automated knee-point detection using	×	0.03
Raw single-cell data in DeepSeq is converted into the AnnData format.	×	0.08
Dimensionality reduction in DeepSeq is performed using PCA.	×	0.02
Cells in DeepSeq are clustered using the Leiden algorithm based on neighborhood graphs.	×	0.03
DeepSeq uses UMAP to embed cells in 2D.	×	0.02

References

- http://arxiv.org/abs/2506.13817v1
- http://arxiv.org/abs/2512.03307v1
- http://arxiv.org/abs/2402.01204v4