Causal Synthetic Training Data Effects on Multimodal Model Alignment Stability

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the impact of causally-generated synthetic training data on the alignment stability of multimodal models when evaluated against standard safety and capability metrics. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.9/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Estimating Test Performance for AI Medical Devices under Distribution Shift with Conformal Prediction. Research question: What is the impact of causally-generated synthetic training data on the alignment stability of multimodal models when evaluated against standard safety and capability metrics?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.9/10.

3 Results

16 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.9/10.

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 source dataset for CIFAR-10 is the original CIFAR-10 (25,000 train / 25,000 validation / 10,000 test), while the tar	×	0.03
The source dataset for Fitzpatrick17K is the DermaAmin atlas (7295 train / 811 validation / 3474 test), while the target	×	0.05
Distribution shift for DMIST is evaluated between four scanner types (1: 12421, 2: 47896, 3: 41311, 4: 6562) and from DM	×	0.06
Distribution shift for WILDS-Camelyon17 is evaluated under source hospitals (302,436 train / 34,904 validation / 33,560	×	0.07
The study evaluates a variety of network architectures and performs five training runs for each model.	×	0.03
The study compares results with and without temperature scaling (TS) to study the effects of softmax calibration on test	×	0.03
The study reports state-of-the-art test accuracy prediction using CPC on several medical imaging datasets without TS and	×	0.10
The performance gains with CPC are especially noticeable on the Fitzpatrick17K dataset, which contains a much larger num	×	0.03
Temperature scaling (TS) moderates the effect of overestimation of predicted test accuracy across methods and architectu	×	0.05
The study hypothesizes that further analysis of the effect of model architecture on different medical datasets and tasks	×	0.10
The study suggests that follow-up work might extend test estimation methods to other clinically relevant tasks such as s	×	0.09
The majority of work incorporating AI for clinical classification tasks has traditionally focused on model accuracy and	×	0.07
AI models tend to break down when inferring on out-of-distribution (OOD) data, which could be images acquired from diffe	×	0.03

References

- http://arxiv.org/abs/2207.05796v1
- http://arxiv.org/abs/2410.20971v2
- http://arxiv.org/abs/1905.11374v5