Pre-Training Data Contamination and Adversarial Robustness in Open-Weight Mathematical Models

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the correlation between pre-training data contamination and robustness to adversarial perturbations in university-level mathematics subsets for open-weight models like Mistral and Gemma. 7 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Targeted Nonlinear Adversarial Perturbations in Images and Videos. Research question: What is the correlation between pre-training data contamination and robustness to adversarial perturbations in university-level mathematics subsets for open-weight models like Mistral and Gemma?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.8/10.

3 Results

4 papers retrieved. 7 claims extracted; 0 independently verified. Quality review score: 2.8/10.

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
A suitable choice for the perturbation generator P is any model architecture able to produce meaningful perturbations on	×	0.02
In the present case, P was taken as a convnet for generating perturbations on both images and videos.	×	0.12
The perturbation method is applied to image and video classification models.	×	0.09
The perturbations are generated by deep convolutional neural networks (convnets) able to learn complex (yet minimalistic	×	0.08
The studied models do consider some features that are indeed relevant to humans.	×	0.09
The models are easily confused by the introduction of seemingly subtle perturbations that do not change the image or vid	×	0.08
Data augmentation procedures can be improved through the addition of perturbations known to drastically affect the perfo	×	0.06

References

- http://arxiv.org/abs/2103.15670v3
- http://arxiv.org/abs/1907.02664v2
- http://arxiv.org/abs/1809.00958v1