Adversarial Training Batch Size Effects on Codestral Cross-Domain Code Generation

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Does adversarial training with different batch sizes improve the cross-domain generalization of Codestral as measured by accuracy on unseen code generation benchmarks like HumanEval. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: How do SGD hyperparameters in natural training affect adversarial robustness?. Research question: Does adversarial training with different batch sizes improve the cross-domain generalization of Codestral as measured by accuracy on unseen code generation benchmarks like HumanEval?.

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 2.4/10.

3 Results

14 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 2.4/10.

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 architectures M1 and StdCNN are used for experiments as given in Table 1.	×	0.02
For CIFAR-10 based experiments, the models C1 and ResNet18 architecture are used.	×	0.04
Input training data was augmented with random cropping and random horizontal flips by default.	×	0.03
Architectures M1 used for MNIST and Fashion MNIST experiments and C1 for CIFAR-10 experiments are as given in [18].	×	0.02
All the PGD based attack results in the Appendix for the corresponding FGSM attack based plots in the paper were plotted	×	0.04
For the benchmark alone, momentum is set to 0.9.	×	0.08
As the batch size increases, the test accuracy decreases.	×	0.12
The associated FGSM test accuracy also drops with increasing batch size.	×	0.07

References

- http://arxiv.org/abs/2410.21676v4
- http://arxiv.org/abs/2504.07887v2
- http://arxiv.org/abs/2006.11604v1