Pretraining Objective Impact on CodeT5 Adversarial Robustness in CWE-200 Benchmarks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the choice of pretraining objective in CodeT5 affect its performance on the CWE-200 benchmark when evaluated using adversarial robustness metrics such as success rate of adversarial attacks. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Attack-SAM: Towards Attacking Segment Anything Model With Adversarial Examples. Research question: How does the choice of pretraining objective in CodeT5 affect its performance on the CWE-200 benchmark when evaluated using adversarial robustness metrics such as success rate of adversarial attacks or perturbation tolerance?.

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 4.7/10.

3 Results

14 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 4.7/10.

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
In the vanilla version of SAM, a single point is chosen in the image and a single mask near the given point is predicted	×	0.03
The attack succeeds in the mask removal task if Maskadv is empty or has a much smaller area than Maskclean.	×	0.10
Clean images for visualization were randomly selected from SAM project demo images or the SA-1B dataset.	×	0.03
FGSM and PGD attacks generate adversarial images with imperceptible perturbations.	×	0.08
Both FGSM and PGD attacks are able to remove the area of Maskclean.	×	0.03
The PGD attack performs better than the FGSM attack in the mask removal task.	×	0.10
The study adopts the IoU metric for quantitative evaluation of adversarial attacks on SAM.	×	0.07
mIoU is calculated as the average IoU between predicted masks of clean images (Maskclean) and predicted masks of adversa	×	0.03
The maximum value of mIoU equals 1 when the perturbation vector is zero (no attack).	×	0.04
With the proposed ClipMSE, the mIoU drops from 1 to close to zero after a PGD attack.	×	0.04
The FGSM attack achieves an mIoU much smaller than 1.	×	0.02
SAM generates masks as outputs by taking both images and prompts as inputs.	×	0.03
Masks generated by SAM do not have semantic labels for each mask.	×	0.04
In SAM, a pixel is marked within the mask area if the predicted confidence value is positive (larger than zero).	×	0.05
The final predicted masks (Maskpred) in SAM are binary matrices with dimensions H*W.	×	0.04

References

- `http://arxiv.org/abs/2305.00866v2`

- http://arxiv.org/abs/2103.15670v3
- http://arxiv.org/abs/2008.07651v1