Adversarial Perturbations in Visual Inputs Degrade LLaVa-1.8-7B Reasoning Accuracy

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Can adversarial perturbations in visual inputs reduce the reasoning accuracy of LLaVa-1.8-7B, and how does its robustness compare to text-only models on equivalent mathematical reasoning tasks. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Robust Image Classification: Defensive Strategies against FGSM and PGD Adversarial Attacks. Research question: Can adversarial perturbations in visual inputs reduce the reasoning accuracy of LLaVa-1.8-7B, and how does its robustness compare to text-only models on equivalent mathematical reasoning tasks?.

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 4.2/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.2/10.

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.

References

- http://arxiv.org/abs/2408.13274v1
- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2104.09369v1