Multimodal Context Enhances Codestral Vulnerability Detection on MBXD Benchmarks

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the inclusion of multimodal context (e.g., code structure visualizations or natural language vulnerability descriptions) in prompting affect Codestral's performance in vulnerability. 17 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluating Open-Source Vision-Language Models for Multimodal Sarcasm Detection. Research question: How does the inclusion of multimodal context (e.g., code structure visualizations or natural language vulnerability descriptions) in prompting affect Codestral's performance in vulnerability detection compared to text-only instruction mixing, evaluated using accuracy and latency on MBXD benchmarks?.

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 4.8/10.

3 Results

4 papers retrieved. 17 claims extracted; 3 independently verified. Quality review score: 4.8/10.

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
Current models achieve moderate success in binary sarcasm detection.	✓	0.27
Models are still not able to generate high-quality explanations without task-specific fine-tuning.	✓	0.24
Sarcasm is a nuanced form of communication where the intended meaning diverges from the literal expression conveying iro	×	0.01
Automatic sarcasm detection has been widely studied in texts.	×	0.06
Sarcasm is often achieved by pairing text with images on social media, and the ironic effect arises from the mismatch be	×	0.04
Multimodal sarcasm detection (MSD) has attracted increasing attention due to the proliferation of multimedia content on	×	0.06
Vision–language models (VLMs) pre-trained on large image–text corpora exhibit powerful zero- and few-shot abilities acro	×	0.14
Models like Flamingo and VILA have demonstrated capabilities in zero-shot and few-shot learning, adapting to new tasks w	×	0.06
Studies like Lin et al. and Yang et al. include MSD within 14 broader multimodal benchmark datasets used to evaluate VLM	×	0.06
Only minimal sarcasm data is present in these studies.	×	0.09
Recent work on Multimodal Sarcasm Explanation (MuSE) requires models to generate human-style explanations for sarcastic	✓	0.19
The study assesses whether off-the-shelf open-source VLMs can detect and explain multimodal sarcasm purely through in-co	×	0.13
Models evaluated include Flamingo, VILA, GPT-4, LLaMA, OpenFlamingo, BLIP2, and InstructBLIP.	×	0.06
Models like Blip2 2.7B, OpenFlamingo 3B, LLaVA 7B, PaliGemma 3B, Qwen-VL 7B, and Gemma3 27B are used in the methodology.	×	0.09
In-Context Prompting is applied with a unified global prompt instruction template across all-shots settings.	×	0.01
Zero-shot prompting includes no examples or reference explanations.	×	0.09
One-shot / Few-shot prompting prepends 1-3 exemplar triplets (caption, reference explanation, 'Yes.' or 'No.') drawn fro	×	0.06

References

- http://arxiv.org/abs/2008.13369v1
- http://arxiv.org/abs/2510.11852v1
- http://arxiv.org/abs/2511.10212v1