Scaling Laws and False Positive Rates in Llama3 Vulnerability Detection Under Adversarial Conditions

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the correlation between model scale and false positive rates in Llama3 variants when performing vulnerability detection on OWASP benchmark tasks under adversarial perturbations. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Benchmarking LLAMA Model Security Against OWASP Top 10 For LLM Applications. Research question: What is the correlation between model scale and false positive rates in Llama3 variants when performing vulnerability detection on OWASP benchmark tasks under adversarial perturbations?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.1/10.

3 Results

16 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.1/10.

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 evaluation followed a four-stage pipeline: Model Loading, Test Execution, Response Evaluation, and Aggregation.	×	0.04
All models were prompted with a consistent classification frame: 'is this prompt safe or not?'	×	0.03
Guard models are optimized for direct classification, while standard models approximate this through generation.	×	0.04
Base models (Meta-Llama-3-8B and Llama-3.1-8B) were the slowest and least effective, failing to detect any threats.	×	0.12
Llama-Guard-3-1B and Llama-3.2-1B provided high accuracy (76% and 73%, respectively) while maintaining minimal latency.	×	0.14
Experiments were conducted on the FABRIC testbed [10], utilizing the Georgia Tech site.	×	0.06
The hardware configuration included: GPU: NVIDIA A30 (24GB VRAM, Ampere architecture), Operating System: Ubuntu 22.04 LT	×	0.02
Ten models from the Llama family were selected to represent a range of parameter scales and training objectives.	×	0.05
Standard Llama Models include: Meta-Llama-3-8B, Llama-3.1-8B, Llama-3.1-8B-Instruct, Llama-3.2-1B, and Llama-3.2-3B-Inst	×	0.12
Llama Guard Models include: Meta-Llama-Guard-2-8B, Llama-Guard-3-1B, Llama-Guard-3-8B, Llama-Guard-3-8B-INT8, and Llama-	×	0.13
All models were loaded in float16 precision (except the INT8 variant) with a low temperature (0.1).	×	0.03

References

- http://arxiv.org/abs/2601.19970v1
- http://arxiv.org/abs/2511.11381v2
- http://arxiv.org/abs/2403.09832v1