DeepSeek R1, Llama3, and Codestral Trade-offs in Vulnerability Detection Efficiency and Accuracy

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the trade-off between inference efficiency (latency, throughput) and F1-score performance for Llama3, Codestral, and Deepseek R1 when deployed for vulnerability detection across multiple. This study investigates the performance of the DeepSeek R1 language model on 30 challenging mathematical problems derived from the MATH dataset, problems that previously proved unsolvable by other models under time constraints. Unlike prior work, this research removes time. 10 claims were extracted from source literature; 2 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: Token-Hungry, Yet Precise: DeepSeek R1 Highlights the Need for Multi-Step Reasoning Over Speed in MATH. Research question: What is the trade-off between inference efficiency (latency, throughput) and F1-score performance for Llama3, Codestral, and Deepseek R1 when deployed for vulnerability detection across multiple programming languages in Big-Vul under constrained hardware conditions?.

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. 10 claims extracted; 2 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
DeepSeek R1's architecture relies on token-based reasoning, facilitating accurate solutions through a more deliberate, m	✓	0.21
DeepSeek R1 achieves high accuracy in solving complex mathematical problems at the cost of significantly higher token us	✓	0.16
DeepSeek R1's average token count (4717.5) is an order of magnitude higher than that of the other models tested.	×	0.07
Llama 3.1 only achieved correct results at a temperature of 0.4, highlighting the sensitivity of certain models to tempe	×	0.07
The MATH dataset contains problems that require multi-step reasoning and symbolic manipulation, posing significant chall	×	0.10
Prior research demonstrated that specific problems from the MATH dataset remained unsolved by several language models un	×	0.12
Transformer-based models have significantly improved the ability of LLMs to process and generate mathematical text.	×	0.04
Even state-of-the-art models struggle to achieve high accuracy on the MATH dataset, particularly under resource constrai	×	0.07
The DeepSeek R1 model is of particular interest due to its documented reliance on token-based reasoning steps, suggestin	×	0.14
The influence of temperature settings on model outputs affects the balance between creativity and coherence in mathemati	×	0.08

References

- http://arxiv.org/abs/2501.18576v1
- http://arxiv.org/abs/2106.16020v1
- http://arxiv.org/abs/2505.02390v2