DeepSeek R1, Llama3, and Codestral Efficiency Trade-offs in Big-Vul Vulnerability Classification

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the computational efficiency trade-off between Llama3, Codestral, and Deepseek R1 when performing vulnerability classification on the Big-Vul dataset, measured in tokens per second and. 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. 9 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. 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 computational efficiency trade-off between Llama3, Codestral, and Deepseek R1 when performing vulnerability classification on the Big-Vul dataset, measured in tokens per second and accuracy?.

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 5.2/10.

3 Results

14 papers retrieved. 9 claims extracted; 2 independently verified. Quality review score: 5.2/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
DeepSeek R1 achieves high accuracy on complex mathematical problems from the MATH dataset when given sufficient computat	✓	0.18
DeepSeek R1 uses significantly more tokens (4717.5 on average) compared to other models like gemini-1.5-flash-8b (359.28	✓	0.22
Llama 3.1 only achieved correct results at a temperature of 0.4, highlighting the sensitivity of certain models to tempe	×	0.08
Previous research demonstrated that specific problems from the MATH dataset remained unsolved by several language models	×	0.11
The MATH dataset contains problems that often require multi-step reasoning and symbolic manipulation, posing significant	×	0.11
Transformer-based models have significantly improved the ability of LLMs to process and generate mathematical text.	×	0.04
DeepSeek R1's architecture relies heavily on token-based reasoning steps, suggesting a potential mechanism for enhanced	×	0.13
The influence of temperature settings on model outputs affects the balance between creativity and coherence in mathemati	×	0.08
Strict time limits on response generation significantly hindered the performance of the DeepSeek R1 model in previous ex	×	0.09

References

- http://arxiv.org/abs/2503.10486v2
- http://arxiv.org/abs/2508.11281v3
- http://arxiv.org/abs/2501.18576v1