DeepSeek R1 Token Reasoning and Latency Trade-offs in Big-Vul Classification

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the trade-off between model size and inference latency for Llama3, Codestral, and Deepseek R1 when classifying software vulnerabilities in the Big-Vul dataset. 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. 11 claims were extracted from source literature; 1 was 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 model size and inference latency for Llama3, Codestral, and Deepseek R1 when classifying software vulnerabilities in the Big-Vul dataset?.

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 4.7/10.

3 Results

4 papers retrieved. 11 claims extracted; 1 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
Prior research demonstrated that specific MATH dataset problems remained unsolved by several language models when operat	×	0.14
DeepSeek R1 has a documented reliance on token-based reasoning steps.	✓	0.16
A previous study imposed strict time limits on response generation which significantly hindered the performance of the D	×	0.11
The average token count for deepseek-r1:8b on the tested MATH dataset subset is 4717.5.	×	0.10
The average token count for gemini-1.5-flash-8b on the tested MATH dataset subset is 359.283333.	×	0.10
The average token count for gpt-4o-mini-2024-07-18 on the tested MATH dataset subset is 462.398268.	×	0.15
The average token count for llama3.1:8b on the tested MATH dataset subset is 390.02.	×	0.08
The average token count for mistral-8b-latest on the tested MATH dataset subset is 191.75.	×	0.10
The average token count for DeepSeek R1 is an order of magnitude higher than that of the other models tested in this stu	×	0.09
Llama 3.1 only achieved correct results at a temperature of 0.4 in this experiment.	×	0.04
DeepSeek R1 solved complex mathematical problems that eluded other models in a previous constrained experiment.	×	0.11

References

- http://arxiv.org/abs/2602.06370v1
- http://arxiv.org/abs/2505.00025v2
- http://arxiv.org/abs/2501.18576v1