Scaling Laws of Vulnerability Classification Accuracy in Llama3, Codestral, and DeepSeek R1

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the vulnerability classification accuracy of Llama3, Codestral, and Deepseek R1 scale with increasing model size when evaluated on technical domain benchmarks like Big-Vul compared to. The rapid development of open-source large language models (LLMs) has been truly remarkable. However, the scaling law described in previous literature presents varying conclusions, which casts a dark cloud over scaling LLMs. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. Research question: How does the vulnerability classification accuracy of Llama3, Codestral, and Deepseek R1 scale with increasing model size when evaluated on technical domain benchmarks like Big-Vul compared to general conversational datasets?.

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 3.3/10.

3 Results

14 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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 LLM was evaluated on a series of public benchmarks both in English and Chinese.	×	0.10
The evaluation included multi-subject multiple-choice datasets such as MMLU, C-Eval, and CMMLU.	×	0.03
Language understanding and reasoning datasets evaluated included HellaSwag, PIQA, ARC, OpenBookQA, and BigBench Hard (BB	×	0.05
Closed-book question answering datasets included TriviaQA and NaturalQuestions.	×	0.00
Reading comprehension datasets included RACE and DROP, C3.	×	0.01
Reference disambiguation datasets included WinoGrande and CLUEWSC.	×	0.00
Language modeling datasets included Pile.	×	0.03
Chinese understanding and culture datasets included CHID and CCPM.	×	0.02
Math datasets included GSM8K, MATH, and CMath.	×	0.00
Code datasets included HumanEval and MBPP.	×	0.02
Standardized exams included AGIEval.	×	0.00
Perplexity-based evaluation was applied to datasets that require answers to be chosen from several options, including He	×	0.03
The model scale representations 6N1 and 6N2 either overestimate or underestimate the computational cost in models of dif	×	0.02
The discrepancy in model scale representations is particularly pronounced in small-scale models, with differences reachi	×	0.05
After adopting M to represent the model scale, the objective is to find the optimal model scale Mopt and data scale Dopt	×	0.03

References

- http://arxiv.org/abs/2503.10486v2
- http://arxiv.org/abs/2401.02954v1
- http://arxiv.org/abs/2505.02390v2