Parameter Scaling Effects on RLHF-Aligned Models in Legal Knowledge Benchmarking

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does increasing parameter scale from 7B to 70B affect the accuracy of RLHF-aligned models on the Legal Knowledge level of the LawBench benchmark. 13 claims were extracted from source literature; 6 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Can Open Large Language Models Catch Vulnerabilities?. Research question: How does increasing parameter scale from 7B to 70B affect the accuracy of RLHF-aligned models on the Legal Knowledge level of the LawBench benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.1/10.

3 Results

8 papers retrieved. 13 claims extracted; 6 independently verified. Quality review score: 6.1/10.

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 study evaluates three LLMs: Llama3, Codestral, and Deepseek R1.	×	0.15
The evaluation uses a carefully filtered subset of the Big-Vul dataset.	✓	0.17
The dataset subset is annotated with eight representative Common Weakness Enumeration (CWE) categories.	✓	0.18
The study adopts a closed-world classification setup.	×	0.13
The models were assessed on their ability to identify the presence of vulnerabilities.	×	0.07
The models were assessed on their ability to map vulnerabilities to the correct CWE label.	×	0.10
The evaluated models demonstrated high detection rates for vulnerabilities.	×	0.11
The evaluated models demonstrated markedly poor classification accuracy for CWE labels.	×	0.12
The models exhibited frequent overgeneralization and misclassification of vulnerabilities.	×	0.12
The study analyzes model-specific biases and common failure modes.	✓	0.17
Current LLMs have limitations in performing fine-grained security reasoning.	✓	0.21
LLMs are being adopted as learning aids in educational contexts.	✓	0.18
Key challenges must be addressed before LLMs can be reliably deployed in security-sensitive environments.	✓	0.26

References

- https://doi.org/10.4230/oasics.icpec.2025.4

- https://doi.org/10.48550/arxiv.2309.16289
- https://doi.org/10.48550/arxiv.2308.12950