Scaling of False Positive Rates in Llama-3.1-8B Security Flaw Detection on Big-Vul Versus CodeLlama and LLaVA

Assignee Research

June 11, 2026

Abstract

The recent success of large language models (LLMs) has sparked a growing interest in training large-scale models. As the model size continues to scale, concerns are growing about the depletion of high-quality, well-curated training data. This has led practitioners to explore training approaches like Federated Learning (FL), which can leverage the abundant data on edge devices while maintaining privacy. However, the decentralization of training datasets in FL introduces challenges to scaling large models, a topic that remains under-explored. This paper fills this gap and provides qualitative in

1 Introduction

This paper examines: Scaling Law Analysis in Federated Learning: How to Select the Optimal Model Size?. Research question: How does the false positive rate of Llama-3.1-8B in identifying security flaws scale with model size when evaluated on the Big-Vul benchmark compared to other models like CodeLlama or LLaVA?.

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 7.2/10.

3 Results

14 papers retrieved. 11 claims extracted; 8 independently verified. Quality review score: 7.2/10.

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
Estimating the optimal model size in federated scenarios should depend on the average training compute across clients.	✓	0.29
The study empirically validates its results with extensive training runs on different models, network settings, and data	✓	0.19
Clients generally have fewer computational resources than the server in federated learning scenarios.	×	0.13
Most existing works on applying FL to large-scale models follow the intuition to reduce model size by tailoring architec	✓	0.19
Few studies have explored the modified scaling behavior of language models in federated scenarios prior to this work.	×	0.11
Previous studies on federated scaling behavior primarily offered observational insights based on empirical evidence.	✓	0.18
The authors model federated training as an SGD optimization problem over distributed data.	×	0.12
The authors derive an analytic solution for the optimal model size to quantify the impact on scaling.	✓	0.15
Stochastic gradient descent (SGD) is a widely used optimization method in machine learning.	✓	0.18
The generalization performance of stochastic algorithms can be quantified using a PAC-Bayes upper bound.	✓	0.21
The PAC-Bayes upper bound has been applied to explore aspects including algorithm convergence.	✓	0.17

References

- http://arxiv.org/abs/2604.14171v1
- http://arxiv.org/abs/2511.12188v1
- http://arxiv.org/abs/2402.04177v3