Fine-Tuning vs Retrieval-Augmented Prompting for Code Security in Llama3-70B and Gemini 1.5 Pro

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does fine-tuning affect the zero-shot and few-shot performance of Llama3-70B and Gemini 1.5 Pro on the CodeXGLUE security subset compared to retrieval-augmented approaches. Few-shot prompting has emerged as a practical alternative to fine-tuning for leveraging the capabilities of large language models (LLMs) in specialized tasks. However, its effectiveness depends heavily on the selection and quality of in-context examples, particularly in complex. 11 claims were extracted from source literature; 5 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Retrieval-Augmented Few-Shot Prompting Versus Fine-Tuning for Code Vulnerability Detection. Research question: How does fine-tuning affect the zero-shot and few-shot performance of Llama3-70B and Gemini 1.5 Pro on the CodeXGLUE security subset compared to retrieval-augmented approaches?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.0/10.

3 Results

13 papers retrieved. 11 claims extracted; 5 independently verified. Quality review score: 6.0/10.

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
Fine-tuning large language models is resource-intensive and may require access to model weights, with non-trivial traini	×	0.08
Few-shot prompting avoids the need for model retraining by embedding labeled input-output examples directly into the pro	×	0.06
Few-shot prompting suffers from high variance depending on the quality and relevance of in-context examples.	×	0.10
Retrieval-augmented prompting with 20 shots achieves an F1 score of 74.05% and a partial match accuracy of 83.90%.	✓	0.37
Fine-tuning Gemini-1.5-Flash achieves an F1 score of 59.31% and a partial match accuracy of 53.10%.	✓	0.34
Retrieval-augmented prompting surpasses fine-tuned Gemini-1.5-Flash without any training overhead.	✓	0.18
Fine-tuning smaller open-source models like DistilBERT and DistilGPT2 achieves lower performance compared to retrieval-a	✓	0.18
Semantic retrieval of in-context examples significantly enhances few-shot prompting effectiveness.	×	0.11
Retrieval-augmented prompting achieves substantial gains over other prompting strategies and fine-tuned LLMs like Gemini	✓	0.17
Retrieval-augmented prompting requires no model training.	×	0.11
Retrieval-augmented prompting is positioned as a practical alternative to fine-tuning in real-world settings where resou	×	0.13

References

- http://arxiv.org/abs/1909.03564v2
- http://arxiv.org/abs/2512.04106v1
- http://arxiv.org/abs/2310.03059v8