Comparative Analysis of Retrieval-Augmented Few-Shot Prompting and Fine-Tuned CodeBERT on Big-Vul Robustness

Assignee Research

June 11, 2026

Abstract

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 domains. In this work, we examine retrieval-augmented prompting as a strategy to improve few-shot performance in code vulnerability detection, where the goal is to identify one or more security-relevant weaknesses present in a given code snippet from a predefined set of vulnerability categories. We perform a systematic

1 Introduction

This paper examines: Retrieval-Augmented Few-Shot Prompting Versus Fine-Tuning for Code Vulnerability Detection. Research question: How does retrieval-augmented few-shot prompting with Llama-3.1-8B compare to fine-tuned CodeBERT in terms of false positive rates and robustness against adversarial code perturbations on the Big-Vul dataset?.

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 7.8/10.

3 Results

13 papers retrieved. 10 claims extracted; 9 independently verified. Quality review score: 7.8/10.

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, may require access to model weights, and entails non-trivial tr	✓	0.25
Few-shot prompting suffers from high variance depending on the quality and relevance of in-context examples.	✓	0.24
Retrieval-augmented prompting achieves an F1 score of 74.05% and a partial match accuracy of 83.90% with 20 shots.	✓	0.35
Fine-tuned 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.	✓	0.20
Fine-tuning smaller open-source models like DistilBERT and DistilGPT2 achieves lower performance compared to retrieval-a	✓	0.15
Semantic retrieval of in-context examples significantly enhances few-shot prompting effectiveness.	✓	0.20
Retrieval-augmented prompting achieves substantial gains over other prompting strategies and fine-tuned LLMs like Gemini	✓	0.22
Retrieval-augmented prompting requires no model training.	×	0.10
Retrieval-augmented prompting is a practical alternative to fine-tuning in many real-world settings where resources may	✓	0.26

References

- http://arxiv.org/abs/2512.04106v1
- http://arxiv.org/abs/2504.16584v1
- http://arxiv.org/abs/2306.11066v2