SOVEREIGN: Does cross-domain fine-tuning (e.g., pre-training on general code vs. security-focused code) combined with tar

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Large Language Models (LLMs) have significantly impacted numerous domains, including Software Engineering (SE). Many recent publications have explored LLMs applied to various SE tasks. Nevertheless, a comprehensive understanding of the application, effects, and possible limitations of LLMs on SE is still in its early stages. To bridge this gap, we conducted a systematic literature review (SLR) on LLM4SE, with a particular focus on understanding how LLMs can be exploited to optimize processes and outcomes. We select and analyze 395 research papers from January 2017 to January 2024 to answer fou

1 Introduction

Analysis of: Large Language Models for Software Engineering: A Systematic Literature Review. Research goal: Does cross-domain fine-tuning (e.g., pre-training on general code vs. security-focused code) combined with targeted preprocessing improve the generalization of Llama3, Codestral, and Deepseek R1 for vulnerability classification, as measured by F1-scores across unseen programming languages?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

4 papers retrieved. 9 claims extracted, 6 verified. Tribunal: 7.4/10 $\rightarrow$ REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
The study is a systematic literature review (SLR) on Large Language Models for Software Engineering (LLM4SE).	✓	0.27
The review analyzes 395 research papers.	×	0.10
The selected research papers were published between January 2017 and January 2024.	×	0.14
The study addresses four key research questions (RQs).	×	0.14
RQ1 categorizes different LLMs employed in SE tasks and characterizes their distinctive features and uses.	✓	0.20
RQ2 analyzes methods used in data collection, preprocessing, and application for LLMs in SE.	✓	0.21
RQ3 investigates strategies employed to optimize and evaluate the performance of LLMs in SE.	✓	0.27
RQ4 examines specific SE tasks where LLMs have shown success to date.	✓	0.26
The artifacts for this study are publicly available at https://github.com/xinyi-hou/LLM4SE_SLR.	✓	0.21

References

- https://doi.org/10.48550/arxiv.2312.02003
- https://doi.org/10.1145/3597503.3639222
- https://doi.org/10.48550/arxiv.2308.10620