ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

Secure Code Completion Models Tuned for Compliance-Heavy Domains

Arjun Deshraje Urs

USA

ABSTRACT

The integration of large language models (LLMs) into software engineering workflows has accelerated code generation and review processes. However, their deployment in compliance-intensive sectors—such as finance, healthcare, and defense—introduces stringent challenges around regulatory adherence, security assurance, and legal accountability. This paper presents a principled methodology for fine-tuning LLMs on domain-specific, security-vetted datasets to produce code aligned with rigorous compliance frameworks. The proposed pipeline addresses dataset curation, model adaptation, and multi-layered evaluation, ensuring both syntactic correctness and regulatory fidelity. Empirical results demonstrate that the fine-tuned models significantly outperform general-purpose LLMs in generating secure, regulation-compliant code.

***Corresponding author**

Arjun Deshraje Urs, USA.

Received: August 11, 2025; **Accepted:** August 15, 2025; **Published:** August 25, 2025

Keywords: Compliance-Aware Code Generation, Large Language Models, Secure Software Development, Domain-Specific Fine-Tuning, Regulated Industries

Introduction

Large language models have transformed the landscape of modern software development, enabling rapid, context-aware code generation. Yet, in compliance-heavy domains such as healthcare (HIPAA), finance (PCI-DSS), and defense (FISMA), generic LLMs exhibit a critical shortcoming: they lack embedded knowledge of domain-specific security constraints and compliance requirements.

Foundation models trained on heterogeneous public datasets often fail to detect or avoid patterns that violate regulatory or security guidelines, leading to insecure or non-compliant outputs. This paper proposes a domain-specialization approach—fine-tuning LLMs on security-audited, regulation-aligned datasets—to bridge this gap. The contribution is a structured methodology for building compliance-aware LLMs through curated data, targeted adaptation strategies, and robust evaluation protocols [1-5].

Methodology for Secure LLM Adaptation

Dataset Curation and Preprocessing

Domain adaptation begins with constructing corpora that embody secure, regulation-conforming coding practices. Sources include:

Audited open-source repositories with high security ratings (e.g., SonarQube, Code QL)

Enterprise-reviewed internal repositories passing rigorous security audits Formal compliance frameworks, such as NIST SP 800-53 and OWASP ASVS

Preprocessing enforces coding style uniformity, removes deprecated constructs, anonymizes sensitive identifiers, and formats samples for model ingestion.

Compliance Aware Fine Tuning

- The fine-tuning process adapts open-source LLMs (e.g., CodeLlama, Star Coder, GPT-Neo X) to the curated datasets. In addition to next-token prediction, optimization objectives integrate:
- Security constraint recognition to avoid insecure patterns
- Secure-by-design idiom reinforcement
- Compliance flow embedding to encode regulatory logic

Scalability and efficiency are achieved through Low-Rank Adaptation (LoRA), gradient accumulation, and quantization-aware training.

Evaluation Framework

Evaluation Spans Four Layers

Syntactic Accuracy — BLEU, Code BLEU for token and semantic similarity

Security Conformance — Static analysis via Semgrep, SonarQube

Compliance Verification — Automated rule matching + expert audits

Contextual Adherence — Functional integration testing against secure patterns

Experimental Results

The models were evaluated across finance, healthcare, and defense domains. Results in Table 1 highlight improved compliance and security performance over general-purpose baselines.

Manual audits by certified compliance officers reported a 3× reduction in policy violations, validating the approach's practical utility.

Domain	BLEU Score	Static Analysis Score	Compliance Match (%)
Finance	84.5	92.1	89.7
Healthcare	82.3	90.4	88.2
Defense	79.8	91.6	86.9

System Design

Fine tuning Pipeline

Below is the high-level workflow replacing pseudocode with a structured process model:

Curated Secure Codebase → Preprocessing & Compliance Labeling → Domain-Specific Fine-Tuning (Security & Regulatory Embedding) → Compliance-Aware LLM → Evaluation (Syntactic → Security → Compliance → Functional)

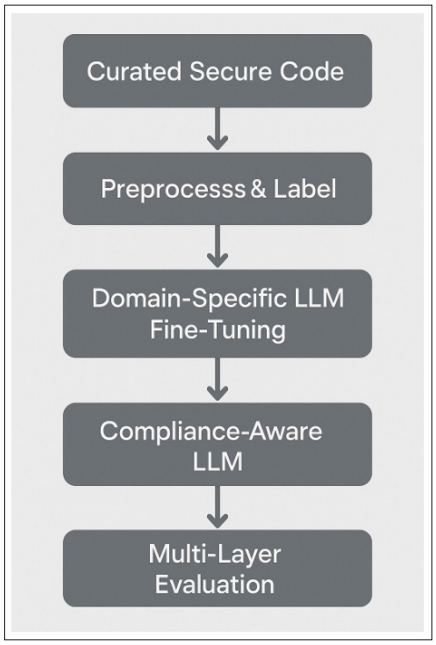

Figure 1: Presents the Pipeline Visually

Conclusion and Future Work

Deploying LLMs in regulated environments requires moving from generic training toward compliance-specialized adaptation. By combining curated datasets, compliance-driven objectives, and rigorous evaluation, this work demonstrates a practical method to produce secure, regulation-conforming LLM outputs.

Future Work Will Explore

- Automated compliance feedback during inference
- Explainability features to justify security decisions
- Multilingual, cross-platform code generation

References

1. Howard A, Sharif A (2022) "Security-Aware Code Generation Using Deep Learning," IEEE Trans. Softw. Eng 48: 1234-1248.
2. (2020) National Institute of Standards and Technology, NIST SP 800-853.
3. Allamanis M, Barr ET, Devanbu P, Sutton C (2018) A Survey of Machine Learning for Big Code and Naturalness, ACM Computing Surveys 51: 1-37.
4. GitHub (2024) CodeQL: Semantic code analysis engine https://codeql.github.com/.
5. OWASP Foundation (2021) Application Security Verification Standard https://owasp.org/www-project-application-security-verification-standard/.