# Abstract

This dissertation investigated the use of Large Language Models (LLMs) for vulnerability discovery and verification in network environments, specifically Cisco IOS 15 router configurations. The central hypothesis was that LLMs could improve the accuracy and efficiency of network configuration verification by detecting vulnerabilities and misconfigurations more effectively than traditional rule-based methods. By applying structured evaluation methods, the study aimed to assess whether prompt engineering could guide models toward more consistent and benchmark-aligned compliance checks.

The study addressed a key research gap in the evaluation of LLMs within applied cybersecurity auditing. While prior work had demonstrated the potential of LLMs in general reasoning and code analysis, little attention ahs been given to their effectiveness in network configuration compliance. Furthermore, the absence of openly available datasets for Cisco IOS devices has hindered systematic testing. This research contributes by introducing a controlled dataset of enterprise-grade configurations with misconfigurations deliberately aligned to CIS Benchmark requirements, enabling reproducible evaluation and comparison across prompt designs.

A layered experimental approach was used. Configurations were generated using automated prompt engineering with embedded Cisco IOS 15 documentation, validated for syntactic and operational correctness in GNS3, and then modified with benchmark-aligned errors. Misconfigurations were introduced programmatically using protocol-specific regex rules, while additional Mistype errors simulated human mistakes. The evaluation framework applied three levels of prompt specificity, Broad, Mid and Specific, progressively increasing guidance to the LLM. Outputs were collected in isolated runs, recorded manually, and evaluated against ground-truth errors using the Perfect Prediction (PP) Score.

The results showed that prompt specificity had a decisive impact on performance. The Broad prompt achieved weak and inconsistent detection, identifying 28% of misconfigurations. The Mid prompt improved this slightly to 33%, while the Specific prompt, with attached CIS excerpts, reached 61%. Mistype detection was stronger, with overall accuracy of 85% under the Specific prompt. Despite these gains, the results highlight persistent inconsistency, with critical requirements often overlooked and frequent reliance on external references, such as Cisco, NIST and NSA publications, even when explicit CIS benchmarks were attached.

These findings underscore both the potential and limitations of LLMs in network compliance verification. While they demonstrate clear improvements when guided by authoritative benchmarks, their variability and reliance on external knowledge prevent them from being considered production-ready tools. Nevertheless, the study confirms that structured prompt engineering and benchmark-driven testing can substantially improve performance and provide meaningful insights into the reasoning behaviour of LLMs in security contexts.

Future research should expand the dataset to cover additional CIS domains, scale to real-world production configurations and explore automated scoring mechanisms to reduce evaluator subjectivity. Investigation of multi-run variance, fine-tuned models and hybrid workflows that combine LLM reasoning with rule-based may further enhance reliability. As LLMs evolve rapidly, continued refinement of datasets, evaluation frameworks and prompt strategies will be essential for advancing their role in automated compliance auditing

**Keywords:** Large Language Models, Network Security, Cisco IOS, CIS Benchmarks, Misconfiguration Detection, Prompt Engineering