Contents

[Chapter 5: Conclusions and Recommendations 2](#_Toc207907610)

[5.1 Summary of Findings 2](#_Toc207907611)

[5.2 Dissertation Limitations 3](#_Toc207907612)

[5.3 Recommendations for Future research 4](#_Toc207907613)

# Chapter 5: Conclusions and Recommendations

## 5.1 Summary of Findings

This dissertation hypothesised that large language models (LLMs) can improve the accuracy and efficiency of network configuration verification by detecting vulnerabilities and misconfiguration more effectively than traditional rule-based methods. The underlying assumption was that, with carefully engineered prompts and a structured evaluation framework, thesis models could deliver consistent compliance checks against recognised security benchmarks, such as the CIS standards, while also capturing more practical errors like typographical mistakes.

The findings of this research partially confirmed this hypothesis. While the models demonstrated an ability to detect security misconfigurations and compliance failures, their performance was highly dependent on the quality and specificity of the prompt. Broad and unguided instructions led to weak and inconsistent results, whereas prompt that explicitly referenced compliance frameworks or provided benchmark excerpts yielded stronger outputs. Even so, the results showed that although LLMs have potential in this domain, their inconsistencies prevent them from being considered production-ready tools.

The study systematically assessed benchmark aligned misconfigurations and Mistype errors across four major protocol domains: AAA, EIGRP, OSPF and RIP. The results showed a clear trend, the more structured and specific the prompt, the more effective the model was at detecting errors. The strongest outcomes were achieved with protocol-specific prompts supported by benchmark excerpts, which enabled the models to provide detailed and compliance-oriented reasoning. However, the models frequently overlooked certain critical requirements, and in some cases, they relied on external references not provided in the prompt, demonstrating the interpretive variability of LLM reasoning.

The approach taken in this research, combining AI- generated configurations which were validated in GNS3, systematic error injection and manual inspection of outputs, was central to uncovering these findings. By deliberately structuring the dataset and controlling the evaluation proves, it was possible to isolate prompt specificity as the primary variable influencing accuracy. This approach not only highlighted the practical limitations of current LLMs but also revealed qualitative patterns, such as their tendency to hallucinate issues or misinterpret valid commands under unguided conditions.

To support these evaluations, a layered framework was developed, progressing from broad to mid-level and then to specific prompt types. This framework was designed to measure reasoning performance across increasing levels of guidance, while using a transparent metric, the PP Score, to provide consistency in accuracy assessment. The framework proved effective in identifying both strengths and limitations, demonstrating that while LLMs can achieve meaningful detection rates under guided conditions, they remain prone to inconsistency when applied without structure, this layered design ultimately provided the evidence needed to critically evaluate the hypothesis and show where LLMs stand today in terms of readiness for network auditing and compliance verification.

Another important finding relates to the references employed by the models when producing their assessments. Despite being explicitly instructed to rely on the attached CIS Benchmark excerpts in the Specific test case, GPT frequently cited or implicitly drew from external sources such as Cisco documentation, NIST publications and broader security frameworks. While this behaviour added contextual depth to its reasoning, it also revealed a tendency to blend authoritative standards with prior training knowledge, raising concerns over consistency and traceability. This highlights both the interpretative strength and the limitation of LLMs in compliance contexts.

## 5.2 Dissertation Limitations

This dissertation was subject to several limitations which shape the scope and interpretation of its findings. The first limitation relates to the coverage of security domains. Only four protocols were examined meaning that many other areas of the CIS Benchmarks, such as SNMP, logging or access control lists were not included in the evaluation. As such, the study does not claim to provide a comprehensive assessment of LLM capability across the full spectrum of Cisco IOS security requirements.

A second limitation is the synthetic nature of the dataset. All router configuration were generated using LLM-based prompts rather than being collected from real-world production environments. While these configurations were validated for operational correctness in GNS3, they inevitably lack the unpredictability, legacy practices and context-specific design choices that characterise enterprise networks. This limitation was caused by the absence of openly available Cisco IOS 15 datasets, particularly for the C7200 router, which necessitated the creation of an artificial dataset tailored to this research.

The third limitation concerns dataset size. The study employed 80 configurations in total, which is sufficient for controlled experimentation and proof-of-concept testing but remains small in scale compared to what would be required for a full benchmarking suite. This limited sample size means that results should be interpreted as indicative rather than exhaustive, capturing patterns in model behaviour without representing all possible network misconfigurations or error types.

In conclusion, these limitations reflected the boundaries within which the research was conducted. Importantly, LLM technology is evolving rapidly, with frequent changes in model behaviour and capability. Findings presented here are therefore time-bound and future work may yield different outcomes as models improve, datasets expand and broader CIS domains are incorporated into evaluation frameworks.

## 5.3 Recommendations for Future research

The findings of this dissertation highlight both the promise and the current limitations of using Large Language Models (LLMs) for network configuration auditing. While the evaluation demonstrated that structured prompt engineering can guide models toward improved compliance assessment, it also revealed inconsistencies that prevent immediate deployment in production environments. Based on these insights, several recommendations for future research are proposed to extend and refine this line of investigation.

First, future work should broaden the scope of evaluation to include additional CIS Benchmark domains beyond AAA, EIGRP, OSPF and RIP. This study focused on four core protocols area due to time and dataset constraints, yet the CIS Benchmarks cover a wide range of services and security requirements. Expanding coverage to domains such as SNMP, NTP and IPv6 would provide a more comprehensive assessment of model performance across the full breadth of enterprise-grade configurations. Doing so would not only test the scalability of the evaluation framework but also generate deeper insights into whether LLMs generalise effectively across diverse security standards.

Second, dataset expansion is critical. The present research relied on 80 synthetic configurations generated through carefully engineered prompts, which served as a valid proof-of-concept but did not capture the full variability of operational networks. Future studies should aim to incorporate a larger dataset and, where feasible, include anonymised real-world configurations. This would help validate whether the patterns observed here, such as the tendency to detect surface-level issues more reliably than protocol-specific misconfigurations, remain consistent when tested on authentic, complex environments.

Third, refinement of the evaluation methodology is recommended. Manual inspection was the most effective approach in this study, ensuring that outputs were carefully verified against CIS requirements. However, as models evolve, incorporating automated scoring pipelines could improve scalability and reduce evaluator subjectivity. In particular, future research should consider multi-run variance testing, since repeated outputs may differ in subtle but important ways. Tracking this variability would provide a clearer picture of model consistency and reliability across multiple attempts.

Fourth, prompt engineering strategies should continue to be refined and tested under different levels of specificity. This dissertation demonstrated that increasing guidance significantly improved detection rates, yet models still drew on external references even when instructed to rely solely on CIS Benchmarks. Future research should investigate mechanisms to enforce stricter alignment with provided reference materials, potentially through fine-tuning approaches. Additionally, exploring how interactive prompts or multi-step dialogues influence accuracy could offer valuable insights for real-world applications, where iterative auditing is common practice.

Finally, while this dissertation confirmed that LLMs show potential as auditing assistants, it also made clear that they are not yet reliable replacements for human auditors. The field of AI is advancing rapidly, and models are continually updated with larger training sets, longer memory windows and improved reasoning capabilities. Future research should revisit the evaluation framework presented here at regular intervals, testing new models against the same CIS-aligned datasets to track progress over time.

In conclusion, this dissertation has provided valuable insight into how prompt engineering and dataset design can be leveraged to assess LLM performance in security compliance auditing. The recommendations outlined above point toward a path for strengthening both the datasets and methodologies used in future studies. By expanding protocol coverage, incorporating real-world data, adoption scalable evaluation methods and refining prompt engineering strategies, subsequent research can build on this work to move closer to integrating LLMs into professional compliance workflows.