Table of Contents

[Chapter 3: Research Methodology 2](#_Toc206616476)

[3.1 Introduction 2](#_Toc206616477)

[3.2 Hypothesis and Research Pipeline 3](#_Toc206616478)

[3.2.2 Research Pipeline 3](#_Toc206616479)

[3.3 Dataset Design and Preparation 4](#_Toc206616480)

[3.3.1 Dataset Generation 5](#_Toc206616481)

[3.3.2 Protocol Coverage and Scope 5](#_Toc206616482)

[3.3.3 Error Design 5](#_Toc206616483)

[3.3.4 Validation Against CIS Benchmarks 6](#_Toc206616484)

[3.3.5 Dataset Realism 6](#_Toc206616485)

[3.3.6 Limitations and Contribution 7](#_Toc206616486)

[3.4 Test Case Design 7](#_Toc206616487)

[3.4.1 Broad Assessment 7](#_Toc206616488)

[3.4.2 Mid-Level Assessment 8](#_Toc206616489)

[3.4.3 Specific, CIS-Guided Assessment 8](#_Toc206616490)

[3.5 Prompt Construction 8](#_Toc206616491)

[3.6 Evaluation Procedure 8](#_Toc206616492)

[3.7 Analysis Methods 8](#_Toc206616493)

# Chapter 3: Research Methodology

## 3.1 Introduction

This chapter discusses the methodology behind three proposed test cases designed to evaluate the compliance of Large Language Models (LLMs) specifically GPT-4o and GPT-5. These tests were conducted to assess how effectively LLMs can identify security misconfigurations in Cisco configurations and determine adherence to industry standards. The approach taken in this research was primarily quantitative, measuring the accuracy of the models against a controlled dataset of configurations. In addition, qualitative observations were made regarding the model’s explanations and recurring error patterns, providing further insight into their interpretative strengths and weaknesses.

The test cases were selected to evaluate the level of specificity required for GPT models to provide accurate responses. Controlled datasets were used to introduce variety while maintaining realism, ensuring that the configurations reflected plausible network scenarios rather than artificial templates. This design allowed the evaluations to replicate a more realistic use case in which an analyst might assess a router configuration under varying conditions. The tests were also highly reproducible, as each could be repeated by simply providing the prompt, configuration and depending on the case, excerpts of the CIS Benchmarks. The inclusion of CIS Benchmarks was critical as they represent an internationally recognised standard for secure configuration practices. By grounding the evaluation in these guidelines, the study ensured that the assessment of GPT models was aligned with authoritative best practices in network security.

The following sections detail the dataset design, prompt construction and evaluation process before outlining the methods used to record and analyse results

## 3.2 Hypothesis and Research Pipeline

This study is based on the hypothesis that Large language models (LLMs) can improve the accuracy and efficiency of network configuration verification by detecting vulnerabilities and misconfigurations more effectively than traditional rule-based methods.

From this hypothesis, research questions were made to guide the methodology:

1. What techniques can be used to design a dataset of network configuration with security flaws that can be used for testing purposes?
2. How reliable are LLMs in identifying and flagging security vulnerabilities and misconfigurations for compliance with security standards?

### 3.2.2 Research Pipeline

This research followed a structured pipeline designed to ensure reproducibility and alignment with the research objectives. The pipeline consisted of five key stages:

1. Phase 01: Development of Test Cases

This phase involved obtaining the CIS Benchmarks, analysing their contents and reviewing related research to inform the design of suitable test cases. From this, the evaluation criteria and performance metrics were defined to provide a consistent basis for assessing the models

1. Phase 02: Formulation of Prompts

A series of prompts were developed for generating configurations and for evaluating them under different conditions. Prompts were designed for each protocol domain (AAA, EIGRP, OSPF, RIP), for merging multi-protocol configurations, for error introduction and for verification.

1. Phase 03: Dataset creation

Individual configurations were first generated per protocol before being merged into multi-protocol configurations that more accurately reflected enterprise environments. Errors were then systematically introduced, aligned with CIS requirements for misconfigurations and supplemented with typographical errors to simulate real-world scenarios.

1. Phase 04: Analysis of Collected Data

Model responses were collected for each test case and assessed against the defined criteria and metrics.

1. Phase 05: Evaluation of Results

Finally, the outcomes were evaluated at two levels:

Criteria-based evaluation, determining whether CIS—aligned misconfigurations were correctly identified

Metric-based evaluation, quantifying accuracy across test cases and protocols

To achieve the objective, the steps outlined in Figure 1 were followed.

A diagram of a company

AI-generated content may be incorrect.

Figure 1: Research Pipeline

## 3.3 Dataset Design and Preparation

A custom dataset of Cisco IOS configurations was created for the purpose of evaluation the ability of large language models to detect security misconfigurations against the CIS Benchmarks. The dataset forms the foundation of this research, providing a structured yet realistic set of test inputs that enabled repeatable and controlled evaluation of the models. Since no openly available configuration datasets exist for Cisco platforms such as the C7200 router, it was necessary to generate the dataset specifically for this dissertation. While originally designed for this study, the dataset has potential for reuse and extension in future research.

### 3.3.1 Dataset Generation

The dataset was generated using OpenAI’s API with carefully engineered prompts. These prompts were designed to mimic the work of a network architect constructing realistic Cisco IOS configurations. For example, the instructions required the generation of router configurations with randomised OSPF area IDs, unique hostnames and unique subnets drawn from the 192.168.\*.\*/24 address space. Interfaces were configured consistently with the advertised networks, and authentication commands were randomly generated to reflect operational diversity. Each generation produced 10 distinct router configurations with the resulting dataset consisting of 80 unique configurations, each approximately 150 lines in length

The use of prompts ensured that configurations were not repetitive or based on rigid templates but instead reflected the complexity of production-grade enterprise environments. Features such as VLANs, multiple routing protocols and varied authentication methods were incorporated, ensuring that the dataset simulated the types of configurations typically encountered in corporate networks. The configurations were subsequently tested in GNS3 on C7200 router images to confirm that they were syntactically correct and operationally valid

### 3.3.2 Protocol Coverage and Scope

The dataset focused on four protocol domains explicitly covered in the CIS Benchmarks and widely deployed in enterprise environments: AAA, EIGRP, OSPF and RIP. Each domain contained 20 configurations, giving a balanced distribution across protocol types. The decision to focus on these four areas was pragmatic, they represent high-impact configuration domains where errors or omissions can have serious security consequences, while also being sufficiently well documented within the CIS Benchmarks to support structured evaluation. Extending the dataset to cover additional protocols such as BGP, SNMP or IPSec would have increased its variety, but this was beyond the scope and time constraints of this dissertation.

### 3.3.3 Error Design

Within each protocol domain, configurations were subdivided into four categories:

* Five configurations with a single misconfiguration
* Five with two misconfigurations
* Five with three misconfigurations, and
* Five containing typographical or syntactic errors.

The misconfigurations were created by systematically removing benchmark-mandated commands, ensuring that each violation directly corresponded to a CIS requirement. For example:

AAA errors: included the absence of commands such as **aaa new-model**, missing login authentication lists, lack of accounting for EXEC sessions or omission of enable secret

EIGRP errors: included configurations missing **authentication mode md5,** incomplete key chains or omission of passive-interface commands.

OSPF errors: included the absence of **area X authentication message-digest**, or missing per-interface MD5 authentication keys

RIP errors: included reverting to RIP version 1, omission of **no auto-summary**, or failure to enable MD5 authentication

By contrast, the typographical errors were not aligned with CIS Benchmarks but were deliberately introduced to test the model’s robustness in identifying realistic human mistakes. These included misspelled commands (e.g. **interfce** instead of **interface**). This distinction ensured that the dataset tested both benchmark-related compliance failures and more practical day-to-day issues faced by network engineers.

### 3.3.4 Validation Against CIS Benchmarks

To ensure that the misconfigurations represented genuine violations, the removed or altered commands were cross-checked against the official CIS Cisco IOS Benchmark for IOS 15. This validation process ensured that the dataset-maintained fidelity to recognised industry standards and that any error introduced would be considered non-compliant under CIS rules. In this way, the dataset served not only as a testbed for GPT models, but also as a controlled approximation of the compliance-checking process used in professional network security audits

### 3.3.5 Dataset Realism

To replicate real-world conditions, the configurations were not isolated per protocol but instead merged into composite files containing multiple features, for instance, an OSPF-focused test configuration also included AAA and EIGRP sections, reflecting the reality that enterprise routers rarely operate with a single protocol enabled. This approach ensured that the dataset challenged the models to parse through complex, multi-protocol configurations in order to identify the relevant issues.

### 3.3.6 Limitations and Contribution

The dataset is subject to several limitations. First, it covers only four protocol domains, leaving other areas of the CIS Benchmark unexplored. Second, the configurations are synthetic and AI-Generated rather than drawn from real production environments. While they were tested in GNS3 for validity, they do not capture the full variability of operational networks. Finally, the dataset is limited in size, consisting of 80 configurations, which, while sufficient for proof-of-concept evaluation, does not constitute an exhaustive benchmark.

Despite these limitations, the dataset represents a meaningful contribution to the field. No comparable open-source Cisco IOS configuration dataset exists, particularly for C7200 devices. By combining benchmark-aligned misconfigurations with realistic, multi-protocol enterprise configurations, this dataset offers a foundation for future research in automated compliance checking, network misconfiguration detection, and the application of large language models in cybersecurity contexts.

## 3.4 Test Case Design

To evaluate the performance of GPT-4o and GPT-5 in identifying misconfigurations and assessing compliance with industry standards, three test cases were designed. These test cases varied in the degree of guidance provided to the model and t he presence or absence of CIS Benchmark references, enabling a comparison of general reasoning, implicit benchmark knowledge and explicit benchmark application.

### 3.4.1 Broad Assessment

The first test case adopted a broad, open-ended approach. The model was prompted as a network security analyst reviewing a router configuration for deployment in a mid-sized corporate network. The task was to determine whether the configuration was secure, and if not, to identify any security issues, misconfigurations, typographical errors or best practice violations. No reference was made to CIS Benchmarks in this case. The purpose of this test was to measure the model’s general interpretative ability to detect errors in network configurations without external guidance, simulating a scenario where ana analyst requires a high- level review

### 3.4.2 Mid-Level Assessment

The second test case introduced CIS Benchmarks into the prompt, instructing the model to assess the configuration according to these standards. However, no excerpts of the CIS documents were provided, requiring the model to rely on its internal knowledge and training data to recall relevant compliance requirements. This case tested whether GPT could correctly apply benchmark-driven reasoning when prompted, despite not being given the official text. It also assessed the extend to which the models had internalised knowledge of the CIS Benchmarks and could distinguish compliance-related issues from general misconfigurations.

### 3.4.3 Specific, CIS-Guided Assessment

The third test case was the most constrained, providing the model with both a narrow focus and explicit benchmark excerpts, in this case, the model was instructed to review only one protocol within the configuration (e.g. RIP, OSPF, AAA or EIGRP) and to assess compliance against the attached section of the CIS Benchmark specific to that protocol. The prompt emphasized that the assessment should be limited strictly to the given protocol. This test case examined the model’s ability to apply prescriptive compliance rules when explicitly provided, and to restrict their analysis to a defined scope. By doing so, it measured how effectively GPT could operate under conditions closely aligned with professional compliance audits.

Together, these three test cases provided a layered evaluation of Gpt-4o and GPT-5. The Broad case tested open-ended reasoning, The Mid case assessed the models internalised benchmark knowledge, and the Specific case evaluated benchmark-driven compliance checking within a constrained scope. This progression allowed the study to capture not only overall detection accuracy, but also the effects of prompt specificity, context and external guidance on the model’s performance.

### 3.5 Prompt Construction

### 3.6 Evaluation Procedure

### 3.7 Analysis Methods