AI Security Compliance and Risk Assessment Framework for Large Language Model Systems

Saniya Bhaladhare

A Master's Thesis Proposal submitted

in partial fulfillment of the

requirements of the degree of

Master of Science in Cybersecurity Engineering

University of Washington

7th October 2025

Project Committee:

- Dr. Min Chen, Committee Chair
- Dr. Yang Peng, Committee Member
- Dr. Marc Dupuis, Committee Member

I. Introduction

The rapid adoption of Artificial Intelligence, particularly Large Language Models, has transformed how organizations operate, automate, and deliver value. From internal knowledge management to customer service automation, LLMs are now embedded in critical business workflows across industries. However, this widespread integration has outpaced the maturity of compliance and security practices. Organizations are increasingly confronted with the limitations of broad regulatory frameworks such as the NIST AI Risk Management Framework (AI RMF) and ISO/IEC 42001, which offer essential principles but lack actionable, AI-specific guidance for real-world implementation. As a result, compliance efforts often rely on manual audits that demand specialized expertise, significant time investment, and still yield inconsistent results across contexts.

This thesis investigates whether interactive, LLM-driven dialogue systems can serve as a viable method for conducting AI compliance and risk assessments. The research aims to model how organizational context can be captured through structured LLM interactions, how applicable regulatory controls can be dynamically identified, and how tailored audit paths can be generated in alignment with recognized standards. It further evaluates whether such a framework can produce traceable, control-level findings and strategic recommendations that are both actionable and scalable. By grounding the work in regulatory theory and empirical validation, this research contributes toward a more adaptive and automated compliance methodology for modern AI deployments.

II. Project Goals / Vision

A. Goals of Completed Work

The overarching goal of this thesis is to design and evaluate an AI-powered framework that can automate and contextualize security compliance assessments for Large Language Model systems. The research investigates whether interactive LLM-based dialogue systems can be used not only to guide organizational stakeholders through complex regulatory frameworks but also to dynamically generate tailored audit paths, collect and validate evidence, and output both control-level and executive-level risk evaluations.

This work is part of a joint thesis collaboration. Harsh will focus on the adversarial and technical vulnerability dimension of LLM security. Specifically, his contributions will include formalizing specifications from the OWASP LLM Top 10, implementing both static and dynamic analysis tools for detecting prompt injection and manipulation vulnerabilities, and validating those tools against benchmark LLM systems in real-world conditions.

My role in the thesis will focus on the compliance alignment and risk assessment side. I will develop and refine a research-backed, LLM-interactive assessment framework that maps compliance standards to real-world LLM deployment contexts. This includes refining a baseline control checklist, engineering dialogue flows to extract organizational attributes, and integrating automated control selection logic. I will also conduct literature-informed analysis on the limitations of current compliance guidance for AI systems and propose evaluation metrics for assessing audit coverage, control relevance, and reporting clarity.

B. Identification of Problem or Opportunity

While LLMs are increasingly embedded in high-impact business operations, most organizations lack reliable methods to assess their compliance posture against evolving AI governance standards. Frameworks such as the NIST AI Risk Management Framework and ISO/IEC 42001 provide high-level guidance but fall short on offering concrete, testable criteria for LLM-specific deployments [1], [2]. Manual audits, meanwhile, are resource-intensive, inconsistent, and not easily scalable [3], [4]. This creates a significant operational blind spot, where LLM-driven systems are deployed without clear risk visibility or assurance of regulatory alignment. However, the adaptive reasoning and conversational capabilities of LLMs themselves present a promising opportunity: enabling compliance assessments that are interactive, context-aware, and dynamically tailored to an organization's regulatory footprint [5], [6]. This research aims to formalize and test such a framework, addressing a growing need for scalable, standards-aligned compliance tooling in enterprise AI environments.

C. Stakeholders and Beneficiaries of Research

1. **Academic researchers in AI governance and compliance** will gain a replicable framework for operationalizing AI regulations, and empirical insights into AI risk evaluation methodologies [2], [6].
2. **Enterprise compliance and risk management teams**, who will benefit from automated, context-aware audit tooling that aligns with evolving frameworks like NIST AI RMF and ISO/IEC 42001 [1], [3].
3. **Regulatory bodies and third-party auditors**, who can leverage standardized, traceable reporting artifacts that enhance oversight and reduce ambiguity in AI system evaluations [4], [5].
4. **LLM service providers and solution architects**, seeking to embed transparent, testable compliance checks into deployment workflows, strengthening trust with enterprise customers and regulators [7].

III. Criteria

A. Level of Success

1. Minimum

The research will produce a baseline compliance assessment framework that maps key requirements from the NIST AI Risk Management Framework and ISO/IEC 42001 to structured, LLM-compatible audit questions. It will demonstrate the ability to programmatically capture organizational attributes through guided prompts, identify applicable controls, and generate an initial compliance checklist. The system will output basic control mappings and support rationale, enabling structured evaluation of governance coverage across different deployment contexts [1],[2].

2. Expected

At this level, the framework will include a functioning LLM-driven interface that dynamically adapts audit flows based on organizational inputs such as industry, model usage, and regulatory scope. It will implement rule-based logic to generate control applicability scores, automate evidence collection guidance, and produce formalized compliance reports. The output will be benchmarked

across at least three distinct LLM deployment scenarios, with evaluation metrics covering report completeness, control relevance, and usability for compliance analysts [3], [4].

3. Aspirational

To reach its full potential, the research will incorporate intelligent control reasoning using LLM-generated justifications for control applicability, integrate a scoring engine for risk severity and mitigation priority, and support continuous compliance tracking over time. It will propose a method for integrating these outputs with organizational governance dashboards or CI/CD tooling, enabling near real-time compliance visibility. The final contribution will be a publishable methodology demonstrating how LLMs can drive practical, scalable alignment with AI governance standards [5], [6].

B. Targets

1. Design and implement a structured compliance mapping system that aligns core functions from NIST AI RMF and clauses from ISO/IEC 42001 with context-aware audit questions for LLM deployments.
2. Develop an adaptive logic engine that selects and ranks applicable compliance controls based on organizational inputs across at least three industry scenarios.
3. Generate complete compliance reports, achieving maximum coverage of relevant controls per scenario, with stakeholder-rated clarity and usability scores above 70% in qualitative feedback.

IV. Proposed Thesis

Prior work has demonstrated the growing relevance of AI-powered compliance tooling in application and infrastructure security, but many efforts stop short of delivering dynamic, LLM-specific, and standards-aligned audit workflows. For example, Gajbhiye et al. explored the use of AI technologies for regulatory compliance automation in application security, highlighting capabilities such as policy enforcement, audit trail management, and real-time monitoring, but did not address how these tools might adapt to LLM-specific controls or deployment contexts [1]. Similarly, Kothandapani examined AI-driven compliance in the financial sector and introduced the use of LLMs for text-to-code translation and regulatory interpretation yet did not explore evidence validation or interactive audit sequencing for LLM-based systems [2].

Ali et al. proposed an automated compliance framework for critical infrastructure security using AI, but the design remained limited to static compliance mappings, with little integration of dynamic control selection or regulatory tailoring based on business attributes [3]. Meanwhile, Shetty's perspective on AI and security reinforced the need for organizations to align their AI systems with frameworks like NIST AI RMF yet emphasized the challenges of adapting these principles to operational deployments without offering automation strategies [4].

In contrast to these efforts, this thesis introduces a compliance framework that actively engages with an organization's operational profile to deliver dynamic, control-relevant audit flows for LLM deployments. It leverages the reasoning capabilities of LLMs to translate high-level governance mandates into contextualized audit paths, supported by structured evidence collection and explainable control selection. Specifically, this research:

1. Operationalizes ISO/IEC 42001 and NIST AI RMF using logic-based mapping to drive tailored control applicability.
2. Automates risk assessment interviews via interactive, LLM-based dialogue agents that adjust based on regulatory scope, sector, and data use.
3. Produces structured, standards-aligned reports with mapped controls, rationale, and evidence pointers to support internal audit and external review.

A. Research Questions:

This thesis investigates whether Large Language Models can serve as an effective interface for automating and contextualizing AI compliance assessments. The following research questions guide the study:

RQ1: How can compliance requirements from NIST AI RMF and ISO/IEC 42001 be operationalized into structured, LLM-driven audit logic?

RQ2: To what extent can an interactive LLM interface accurately identify applicable compliance controls based on organizational inputs (e.g., domain, deployment scope, regulatory obligations)?

RQ3: How effective is the proposed framework in generating complete, standards-aligned audit reports with minimal manual intervention?

RQ4: What metrics can be used to evaluate the usefulness, accuracy, and coverage of AI-generated compliance assessments in comparison to manual audit processes?

B. Initial Baseline Checklist

Below is a preliminary version of the baseline compliance checklist. Each entry maps a governance requirement to a control objective, provides an example of how it applies to LLM systems, and outlines expected evidence types.

Standard	Function/ Clause	LLM-Relevant Control Objective	Example Implementation	Evidence Type
NIST AI RMF	Govern (GOV.1-GOV.3)	Define and document AI system use, purpose, and risk posture	Maintain documented LLM use cases and approval workflows	Governance policy, use case registry

NIST AI RMF	Map (MAP.1)	Identify context, users, data, and impacts	Classify training inputs and intended user audiences	Data inventory, context mapping report
ISO/IEC 42001	Clause 8.2.1	Control access to AI models and interfaces	Implement RBAC for LLM APIs and fine-tuning workflows	IAM logs, access control matrix
ISO/IEC 42001	Clause 6.1.2	Risk assessment procedure for AI-specific deployment	Evaluate misuse scenarios like prompt injection or output abuse	Threat model, risk register
NIST AI RMF	Manage (MAN.4)	Incident response for AI-generated harms	Escalation plan for harmful completions or output anomalies	IR procedure, incident log samples

V. Project Plan

A. Detailed Milestones for Key Deliverables

Quarter	Weeks	Tasks and Milestones
Autumn 2025	Weeks 1–4	Literature review on NIST AI RMF, ISO/IEC 42001; study regulatory gaps in LLM systems
	Weeks 5–8	Develop initial baseline compliance checklist; map standards to LLM-specific controls
	Weeks 9–12	Design an interactive audit engine; define dialogue flow and control applicability logic

Winter 2026	Weeks 1–4	Build a dynamic audit generator and control selection engine using simulated deployment scenarios
	Weeks 5–8	Implement a structured evidence collection module; define validation and scoring rules
	Weeks 9–12	Generate formalized compliance reports and evaluate framework on 3 deployment cases; gather feedback and finalize write-up

B. Work to be Completed

The project will follow an iterative, research-driven development model. The process begins with a grounded literature review to inform architecture and compliance mappings. Design will follow a modular approach, enabling early testing of each component (e.g., audit generation, control logic) before full system integration. Evaluation will involve scenario-based walkthroughs using simulated LLM deployment cases with synthetic or placeholder evidence. Metrics will include control coverage, reporting accuracy, and stakeholder usability feedback. A mix of qualitative and quantitative methods will be used to assess outcomes.

VI. Constraints, Risks, and Resources

A. Key Constraints

1. **Scope Limitation:** This thesis will focus on a representative subset of ISO/IEC 42001 controls and selected subcategories from all four NIST AI RMF Functions (Govern, Map, Measure, Manage). The framework is designed to be extensible to the full standards, but validation is limited to a focused set of controls to ensure feasibility within the thesis timeline.
2. **Access to Representative Use Cases**: The effectiveness of the compliance framework depends on applying it to varied realistic LLM deployment scenarios. Lack of access to real-world enterprise configurations may limit generalizability. Synthetic or placeholder evidence will be used rather than real enterprise data.
3. **Time and Scope**: As the project involves both system design and research validation, time must be carefully managed to ensure completion without overextending into feature-heavy implementations.
4. **Evolving Standards**: Both NIST AI RMF and ISO/IEC 42001 are subject to revisions and evolving interpretations. The framework will need to work with current documentation, with clear assumptions about versioning.

5. **Domain Familiarity**: While having experience in cybersecurity and compliance, hands-on familiarity with LLM development and orchestration is limited. Additional time and study may be required to fully prototype complex LLM-based dialogue flows and logic systems.

B. Resources Needed for Success

1. **Technical Stack**: Open-source LLMs (e.g., Hugging Face Transformers), Python, LangChain or similar orchestration tools, and vector databases for storing audit sessions.

2. **Framework Documents**: Official documentation and mappings for NIST AI RMF and ISO/IEC 42001; supporting literature from industry and academic sources.

3. **Test Environments**: Simulated or anonymized enterprise LLM deployment configurations to test control applicability and audit output quality.

C. Anticipated Risks

1. **Framework Complexity**: The logic required to dynamically select and adapt compliance controls may grow complex, potentially leading to scalability issues or unexpected edge cases.
2. **LLM Output Variability**: As the dialogue engine relies on LLMs, non-deterministic outputs or hallucinations may introduce inconsistency in audit paths or responses unless carefully constrained and validated.

VII. Research Methods & Design

This thesis follows a design science research methodology, focusing on the development and evaluation of a novel compliance framework for LLM systems. The objective is not only to build a functional tool, but to investigate its effectiveness, reproducibility, and alignment with recognized AI governance frameworks.

A. Framework Design Approach

The framework will be developed as a modular system, consisting of:

- An LLM-driven dialogue interface to capture organizational context.
- A logic engine that maps inputs to applicable controls from NIST AI RMF and ISO/IEC 42001.
- A questionnaire generator that produces a custom audit path.
- A reporting module that outputs both control-level and executive-level summaries.

Design decisions will be informed by:

- Prior literature on AI compliance tooling [1][2].
- Structure and semantics of the chosen standards.
- Realistic organizational deployment scenarios collected via case study templates.

B. Evaluation Methodology

The framework will be evaluated using scenario-based testing and expert review. This includes:

- **Applying the system to three simulated LLM deployment cases:**
    1. **Financial Services Assistant**: focusing on accountability, access control, and risk assessment.
    2. **Customer Service Chatbot**: focusing on incident handling, transparency, and monitoring outputs.
    3. **University Knowledge Assistant**: focusing on data governance, fairness, and continuous monitoring.
- **Measuring output completeness**: whether the system surfaces maximum relevant controls for each scenario.
- **Assessing accuracy**: whether controls selected are appropriate for the scenario context (e.g., financial risk controls appear in the financial assistant use case).
- **Evaluating clarity**: reviewers rate report readability, structure, and actionability on a 5-point scale.
- **Usability review**: qualitative feedback from compliance professionals or graduate peers on how well the reports support decision-making across different scenarios.

C. Validation Criteria

To assess the framework's effectiveness, the following criteria will be used:

- Control Coverage: Percentage of relevant controls accurately surfaced based on input parameters.
- Report Quality: Expert-rated usefulness of compliance reports using a structured rubric (e.g., completeness, actionability).
- Audit Path Efficiency: Reduction in total controls shown compared to full checklist, without compromising relevance.
- Repeatability: Whether similar inputs produce consistent, traceable outputs across multiple sessions.

VIII. Appendix

Initial Baseline Compliance Checklist (Full Table)

- Expanded table version of the five selected controls from NIST AI RMF and ISO/IEC 42001
- Includes control ID, standard source, objective, LLM-specific application, and evidence type

Sample Audit Output

- Mockup of a system-generated audit report
- Includes: relevant controls selected, rationale, evidence prompts, and summary dashboard for stakeholders

Evaluation Rubric (Draft)

- Rubric used to score coverage, relevance, and usability of the compliance framework
- Criteria for completeness, control accuracy, and reporting clarity

IX. References

[1] B. Gajbhiye, S. Khan, and O. Goel, "Regulatory Compliance in Application Security Using AI Compliance Tools," *International Research Journal of Modernization in Engineering, Technology and Science*, vol. 6, no. 8, pp. 2583–2585, Aug. 2024. [Online]. Available: https://www.irjmets.com

[2] H. P. Kothandapani, "AI-Driven Regulatory Compliance: Transforming Financial Oversight through Large Language Models and Automation," *Emerging Science Research*, vol.3, Jan. 2025. [Online]. Available: https://www.researchgate.net/publication/388231248

[3] S. M. Ali et al., "An Automated Compliance Framework for Critical Infrastructure Security Through Artificial Intelligence," *IEEE Access*, vol. 11, pp. 74459–74472, 2023, doi: 10.1109/ACCESS.2023.3275907.

[4] P. Shetty, "AI and Security: From an Information Security and Risk Manager Standpoint," *IEEE Access*, vol. 12, pp. 77468–77480, Jun. 2024, doi: 10.1109/ACCESS.2024.3408144.

[5] S. Dambe, "The Role of Artificial Intelligence in Enhancing Cybersecurity and Internal Audit," in *2023 3rd Int. Conf. on Advancement in Electronics & Communication Engineering (AECE)*, IEEE, pp. 85–90, doi: 10.1109/AECE59614.2023.10428353.

[6] A. Mohammed, "AI in Cybersecurity: Enhancing Audits and Compliance Automation," *Innovative Computer Science Journal*, vol. 7, no. 1, pp. 1–4, 2023. [Online]. Available: https://innovatesci-publishers.com

[7] K. Doshi, "Revolutionizing Compliance with Automation and AI," *Int. Journal of Science, Engineering and Technology*, vol. 11, no. 5, pp. 120–124, Sep. 2023, doi: 10.61463/ijset.vol.11.issue5.567.