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

Saniya Bhaladhare

A thesis submitted in partial fulfillment of the

requirements for the degree of

Master of Science in Cybersecurity Engineering

University of Washington

2026

Committee:

Dr. Min Chen

Dr. Marc Dupuis

Dr. Yang Peng

Program Authorized to Offer Degree:

Computing and Software Systems

University of Washington

ABSTRACT

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

Saniya Bhaladhare

Chair of the Supervisory Committee: Dr. Min Chen

The rapid integration of Large Language Models into organizational workflows has outpaced the governance frameworks designed to oversee them. Standards such as NIST AI RMF 1.0 and ISO/IEC 42001 provide high-level principles for responsible AI but lack the operationalized, testable controls required for consistent compliance auditing of deployed LLM systems. This thesis designs, implements, and empirically evaluates an AI-powered compliance assessment framework that operationalizes these standards into a structured, interactive audit system.

The framework is grounded in a Design Science Research methodology. Its core artifact is an AI audit agent backed by a 30-control library spanning six risk domains, a 0-to-5 maturity scoring model anchored to the NIST SP 800-53A evidence hierarchy, and a rule-based adaptive selection layer that tailors the control set to the deployment context without sacrificing reproducibility. A key research finding is that NIST AI RMF 1.0 contains no dedicated subcategories for three OWASP LLM Top 10 (2025) risk categories: System Prompt Leakage (LLM07), Vector and Embedding Weaknesses (LLM08), and Unbounded Consumption (LLM10). This thesis constructs three composite controls to close this gap, a finding independently corroborated by NIST AI 600-1 (July 2024).

The agent is evaluated across three synthetic deployment scenarios representing low, moderate, and high organizational maturity, producing compliance scores of 23 percent, 33 percent, and 71 percent respectively. Two independent human assessors validated all 90 control assessments, producing an inter-assessor agreement rate of 72.2 percent and an LLM-to-human consensus agreement rate of 81.1 percent, confirming calibration within the expected range for AI-assisted assessment instruments.

This thesis contributes to the field in four respects. First, it provides the first operationalized, evidence-requesting 30-control compliance library that achieves full coverage of the OWASP LLM Top 10 (2025) with explicit standard citations to NIST AI RMF, ISO/IEC 42001, and ISO/IEC 27001. Second, it identifies and formally documents a structural gap in NIST AI RMF 1.0 through three

composite controls whose necessity is corroborated by NIST AI 600-1. Third, it demonstrates a replicable adaptive selection architecture that separates deterministic compliance logic from LLM-assisted language generation. Fourth, it establishes a validated empirical baseline characterizing AI compliance tool calibration behavior across the full maturity range, with two specific calibration gaps confirmed by independent human assessors for future scoring model refinement.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the guidance and support of many people. First and foremost, sincere gratitude is owed to Dr. Min Chen, committee chair, for his consistent direction, constructive feedback, and patience throughout this research. The contributions of committee members Dr. Yang Peng and Dr. Marc Dupuis, whose expertise shaped the rigor and depth of this work, are deeply appreciated.

This research is one part of a collaborative thesis project with Harsh Jannawar, whose work on the adversarial testing and vulnerability dimensions of the same overarching framework has been a valuable intellectual counterpart to the compliance and governance work documented here. The breadth of the combined framework reflects the strength of that collaboration.

Gratitude is also owed to the University of Washington Bothell community for providing a professional and academic environment that consistently challenged and encouraged this work.

TABLE OF CONTENTS

Contents

Chapter 1. Introduction

1.1 Background and Motivation

Large Language Models are no longer experimental technology. They are embedded in customer service platforms, legal document review tools, medical triage assistants, HR policy systems, financial advisory tools, and internal enterprise knowledge bases across virtually every industry sector. This adoption has been rapid enough that governance practices have not kept pace. Organizations deploying LLMs today face a practical challenge: they are expected to demonstrate responsible, risk-aware use of these systems, yet the frameworks that govern AI use are largely written at a level of abstraction that is difficult to translate into concrete compliance actions [2].

The NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0), published in January 2023, and ISO/IEC 42001, published in December 2023, both provide valuable frameworks for thinking about AI risk and governance [2, 3]. However, neither offers the specific, testable controls needed to audit a particular LLM deployment against a defined compliance baseline. Manual auditing is an option, but it is resource-intensive, inconsistent across assessors, and does not scale to the rate at which organizations are now deploying AI systems [4, 6].

The OWASP LLM Top 10 (2025) provides the most comprehensive publicly available taxonomy of LLM-specific vulnerabilities, covering ten distinct risk categories from Prompt Injection (LLM01) through Unbounded Consumption (LLM10) [11]. However, a taxonomy is not a testing or compliance tool. It identifies what can go wrong but does not tell an organization how to assess whether its deployment addresses those risks. The result is a growing class of AI systems deployed without clear risk visibility or assurance of regulatory alignment.

At the same time, the conversational and reasoning capabilities of LLMs themselves present an opportunity: they can serve as the engine of a compliance assessment tool that is interactive, context-aware, and dynamically tailored to an organization's deployment context [9, 10]. This thesis investigates that opportunity by designing, building, and evaluating an AI-powered compliance framework that operationalizes existing standards into a structured, auditable assessment system for LLM deployments.

1.2 Problem Statement

Despite the rapid integration of LLM-powered systems into enterprise workflows, organizations lack a practical, evidence-based method for assessing their compliance posture against the governance standards that apply to those systems. Three compounding gaps define this problem.

The first is an operationalization gap. The distance between a high-level framework principle, such as NIST AI RMF's instruction to map AI risks and impacts, and a concrete audit checkpoint that can be tested against a specific deployment, must be bridged by expert human judgment. This judgment is expensive, inconsistent across assessors, and does not produce replicable results [4, 6].

The second is an evidence validation gap. Compliance assessment requires not only asking the right questions but evaluating whether submitted documentation actually satisfies a control's intent. Without a structured evaluation mechanism, organizations rely on self-reported compliance claims that may not reflect actual practice [17].

The third is a scalability gap. The demand for AI compliance assessment is growing faster than the supply of qualified human auditors. A framework that requires expert human time for every control assessment cannot scale to the pace of AI adoption [1, 3].

The intended primary users of this framework are internal compliance officers, AI risk managers, and GRC practitioners within organizations that have deployed or are preparing to deploy LLM systems. Secondary users include external auditors conducting third-party AI governance assessments, and AI product teams seeking a structured self-assessment mechanism prior to a formal audit. The tool is not designed for end-users of the LLM system being assessed; it is designed for the organizational roles responsible for governance oversight of that system. This target user profile informs the tool's design: it assumes familiarity with compliance concepts, relies on professional judgment for the human override mechanism, and produces outputs structured for stakeholder reporting rather than end-user communication.

This research addresses all three gaps through a single integrated contribution: a 30-control compliance library operationalized from NIST AI RMF 1.0, ISO/IEC 42001:2023, and OWASP LLM Top 10 (2025), implemented as an interactive AI audit agent that automates evidence analysis, score generation, and report production.

1.3 Research Questions

This thesis is organized around four research questions derived from the problem statement.

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 a rule-based adaptive selection layer accurately filter and prioritize compliance controls based on LLM deployment type, and how does deployment-aware control selection affect the relevance and completeness of generated assessments?

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 compared to manual audit processes?

These four questions map respectively to the four core research contributions of this thesis: the 30-control library and its composite controls (RQ1), the adaptive control selection architecture (RQ2), the empirical evaluation of report completeness and calibration (RQ3), and the human assessor comparison and override analysis (RQ4).

1.4 Contributions

The primary contribution of this thesis is a set of research findings that advance the understanding of automated, standards-aligned compliance assessment for LLM deployments. An implemented audit agent was developed and evaluated as the vehicle through which those findings were produced and validated. Within this broader contribution, four specific research contributions distinguish this work from prior efforts in AI governance and compliance automation.

The first contribution is a 30-control compliance library that operationalizes requirements from NIST AI RMF 1.0, ISO/IEC 42001:2023, ISO/IEC 27001:2022, and OWASP LLM Top 10 (2025) into structured, evidence-requesting audit questions organized across six risk domains. Every OWASP LLM Top 10 (2025) risk category is addressed by at least one control in the library, providing comprehensive coverage of the current LLM threat landscape.

The second contribution is the identification and construction of three composite controls for OWASP risk categories that have no dedicated native subcategory in NIST AI RMF 1.0 or ISO/IEC 42001:2023: System Prompt Leakage (LLM07), Vector and Embedding Weaknesses (LLM08), and Unbounded Consumption (LLM10). Each composite control combines two existing standard subcategories around a new LLM-specific control objective. The same structural gap in these source standards is also acknowledged by NIST in the July 2024 release of NIST AI 600-1 [21], a companion

document released to extend NIST AI RMF 1.0 with supplementary guidance for generative AI systems that the original 2023 standard did not adequately address.

The third contribution is a rule-based adaptive control selection architecture that filters and prioritizes the 30 controls based on deployment type, enabling context-aware assessments that reflect the risk profile of the specific system being audited without introducing LLM non-determinism into control selection logic.

The fourth contribution is an empirical evaluation dataset of 90 control assessments across three synthetic deployment scenarios of graduated maturity, with calibration findings documented in sufficient detail to support both replication and extension by future researchers.

1.5 Scope and Collaborative Division of Work

This thesis focuses on the compliance and governance dimensions of LLM deployment: the design, implementation, and evaluation of a structured assessment framework that operationalizes existing AI governance standards into a practical audit tool. The framework is validated using synthetic organizational evidence across three representative deployment scenarios.

This work is part of a joint research project with Harsh Jannawar, whose contribution addresses the adversarial security testing dimension of the same overarching problem. Harsh's contribution covers the operationalization of OWASP LLM Top 10 controls into automated adversarial test cases, the implementation of an evolutionary attack engine, and the empirical validation of that testing pipeline against target LLM deployments. The two bodies of work address non-overlapping problem spaces: this thesis operationalizes governance standards into structured compliance assessment, while Harsh’s thesis operationalizes attack taxonomies into automated adversarial testing. Together they form a complete framework covering both the compliance posture and the technical vulnerability surface of an LLM deployment, though each is independently evaluable and independently documented.

The contribution documented in this thesis covers the derivation of precise audit criteria from NIST AI RMF and ISO/IEC 42001, the construction of the 30-control library including the three composite controls, the design and implementation of the adaptive selection layer, the evidence analysis and scoring mechanism, and the empirical evaluation across all three deployment scenarios.

1.6 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 surveys the existing literature on AI governance frameworks, compliance automation, LLM security risks, and process maturity models,

positioning this framework relative to prior work and identifying the specific gaps it addresses. Chapter 3 presents the research methodology, covering the design science research approach, scenario design, evidence pack construction, and the human assessor validation protocol. Chapter 4 describes the framework design in full, covering the six-domain control library, the three composite controls, the maturity scoring model, and the audit dialogue flow. Chapter 5 documents the system design and implementation of the AI audit agent, including the tool architecture, LLM backend selection, evidence file handling pipeline, scoring mechanism, and adaptive selection layer. Chapter 6 presents the evaluation methodology, covering the three deployment scenarios and their evidence packs. Chapter 7 presents the empirical results across all three scenarios, including domain scores, calibration findings, and the AI-to-human assessor comparison. Chapter 8 discusses the findings in relation to the four research questions and addresses limitations. Chapter 9 concludes with a summary of contributions and directions for future work.

Chapter 2. Literature Review and Related Work

2.1 AI Governance Frameworks

The NIST AI RMF [2], published in January 2023, organizes AI risk governance around four core functions: Govern, Map, Measure, and Manage. The Govern function establishes organizational risk culture and accountability structures applicable across all stages of the AI lifecycle. The Map, Measure, and Manage functions address contextualizing, assessing, and responding to AI risks respectively. A companion Playbook [5] provides suggested actions and informative references aligned to each subcategory within these four functions. As Shetty (2024) observes, applying these high-level principles to specific operational deployments remains a significant challenge, and the framework stops short of providing concrete, testable criteria for systematic auditing [6].

ISO/IEC 42001 [3], published in December 2023, is the first international certification standard specifically addressing AI management systems. It adopts the same clause-based structure as ISO/IEC 27001, making it familiar to organizations already operating information security management systems. Its Annex A provides AI-specific controls covering areas such as data governance, AI system lifecycle, and supplier relationships. However, its coverage of LLM-specific risks such as hallucination, prompt injection, and model drift remains limited, and the standard does not offer operationalized, testable audit criteria for individual LLM deployment configurations.

The OECD AI Principles [7], first adopted in May 2019 and updated in May 2024, provide normative background on trustworthy AI organized around five values: inclusive growth and well-being, human rights and democratic values, transparency and explainability, robustness and safety, and accountability. While influential as a global intergovernmental standard, the Principles operate at a higher level of abstraction than either NIST AI RMF or ISO/IEC 42001 and do not translate directly into audit criteria for specific AI deployments.

A critical observation about all three frameworks, and one that is central to the design of this thesis, is that each was written before the widespread enterprise deployment of LLMs made certain risk categories mainstream. NIST AI RMF 1.0 was published in January 2023, at a point when RAG architectures and system prompt extraction attacks were not yet well-documented LLM deployment risks. ISO/IEC 42001:2023, while published later in December 2023, is a general AI management system standard not designed to address LLM-specific attack surfaces. NIST itself acknowledged this gap in July 2024 with the release of NIST AI 600-1 [21], the Generative AI Profile, specifically designed to extend the original framework with supplementary guidance for generative AI systems. The existence of this

companion document confirms the operationalization gap that motivates this thesis. The three frameworks, NIST AI RMF, ISO/IEC 42001, and ISO/IEC 27001 for access control, serve as the primary source standards from which the 30-control library in this research is derived, as documented in Chapter 4.

2.2 Compliance Automation Research

Gajbhiye et al. (2024) examine the application of AI compliance tools to regulatory requirements in application security, identifying value in automated policy enforcement and audit trail management, but acknowledge that their analysis does not address the unique challenges of LLM deployments [1]. Ali et al. (2023) demonstrate that structured compliance control mappings can be implemented programmatically for critical infrastructure security, using a machine learning framework to recommend cybersecurity standards, audits, and compliance measures across nine international and national standards [4]. Their work validates the core premise of this research, namely that compliance logic can be implemented as automated, structured tooling, while stopping short of interactive evidence collection or LLM-specific control selection.

Kothandapani (2025) applies LLMs to financial regulatory text interpretation and compliance code generation, demonstrating that LLMs can convert regulatory text into executable compliance rules and reduce the manual effort of translating regulatory requirements into actionable procedures [8]. This work is the closest near-neighbor to the current research in its use of LLMs as a compliance automation engine, but it does not address the LLM-as-subject-of-audit dimension: in that work, LLMs are the tool for interpreting regulations about financial systems, not the system being assessed for compliance with AI governance standards.

Mohammed (2023) and Dambe (2023) survey AI applications in cybersecurity auditing more broadly [9, 10], documenting how AI capabilities in pattern recognition, anomaly detection, and automated reporting are being applied to compliance workflows. Both works provide supporting evidence for the general viability of AI-assisted compliance assessment while acknowledging that LLM-specific security controls remained outside the scope of existing automated audit tools at the time of writing.

Doshi (2023) argues that embedding compliance checking into development workflows produces more consistent regulatory alignment than periodic manual audits [14], supporting the interactive, tool-mediated approach taken in this research. The argument for continuous, integrated compliance tooling over episodic manual review is a recurring theme across this body of work and directly motivates the design of the audit agent in this thesis.

Taken together, the compliance automation literature confirms two things. First, automated compliance tooling is technically feasible and organizationally valuable across a range of security and regulatory domains. Second, no prior work addresses the specific combination this thesis targets: operationalizing AI governance standards from NIST AI RMF and ISO/IEC 42001 into a structured, evidence-based compliance assessment tool for LLM deployments specifically. The LLM-as-subject-of-audit problem, where the system being assessed for compliance is itself an LLM deployment, remains unaddressed by existing compliance automation research.

2.3 LLM Security Risks

The OWASP LLM Top 10 (2025) [11] provides the most comprehensive publicly available taxonomy of security vulnerabilities specific to LLM applications, organized into ten categories: Prompt Injection (LLM01), Sensitive Information Disclosure (LLM02), Supply Chain Vulnerabilities (LLM03), Data and Model Poisoning (LLM04), Improper Output Handling (LLM05), Excessive Agency (LLM06), System Prompt Leakage (LLM07), Vector and Embedding Weaknesses (LLM08), Misinformation (LLM09), and Unbounded Consumption (LLM10). The OWASP taxonomy was selected as the primary completeness criterion for the control library in this research because it is operationally grounded, maintained by a practitioner community, and directly maps to real-world deployment risks rather than theoretical harm categories. The requirement that every OWASP category be addressed by at least one control ensures that the framework covers the full documented attack surface of LLM deployments as understood by the security community.

A complementary academic taxonomy by Weidinger et al. [20], published at the 2022 ACM Conference on Fairness, Accountability, and Transparency, organizes LLM risks around six harm areas: discrimination and hate speech, information hazards, misinformation harms, malicious uses, human-computer interaction harms, and environmental and socioeconomic harms. This taxonomy provides theoretical grounding for the broader risk landscape within which the OWASP operational categories sit. The two taxonomies are complementary rather than competing: OWASP addresses the specific attack surfaces and failure modes of deployed LLM systems, while Weidinger et al. address the broader societal and ethical harm categories that such failures can produce.

A key finding of the framework design process is that NIST AI RMF 1.0 does not contain dedicated subcategories for three of the ten OWASP categories: System Prompt Leakage (LLM07), Vector and Embedding Weaknesses (LLM08), and Unbounded Consumption (LLM10). This structural gap, documented in detail in Section 4.5, required the construction of three composite controls, each combining two existing standard subcategories around a new LLM-specific control objective.

The phenomenon of hallucination and misinformation in LLM outputs has been extensively documented by Ji et al. (2023) in a comprehensive survey covering causes, measurement approaches, and mitigation strategies across multiple NLG tasks [12]. This framework addresses the hallucination risk through a dedicated validity and reliability control requiring evaluation against verified benchmark datasets with results tracked over time. Bias in model outputs, a related concern with significant implications for fairness and legal liability in organizational deployments, is recognized by NIST SP 1270 (2022) [18] as a distinct risk category requiring systematic identification across three categories of bias: statistical and computational bias, human cognitive bias, and systemic bias embedded in institutional processes. This framework addresses bias through the data quality and data provenance controls in the Data Integrity domain, which require documented evidence of training data sourcing, quality assessment, and lineage tracking.

Prompt injection, one of the earliest and most extensively studied LLM-specific attack vectors, was systematically characterized by Perez and Ribeiro (2022), who demonstrated that adversarial user inputs could reliably override operator-defined system prompt instructions in production LLM deployments [13]. Their work established prompt injection as a reproducible vulnerability class rather than an edge case and directly informed the design of the adversarial testing and secure development controls in the Technical Reliability and Security domain of this framework.

2.4 Process Maturity Models

The maturity scoring model in this research draws on two established process maturity frameworks: CMMI [15] and ISO/IEC 33001 [16]. Together, these provide the conceptual foundation for a structured, evidence-anchored 0-to-5 scoring scale that maps directly to the quality of compliance evidence an organization can produce.

CMMI defines five maturity levels: Initial, Managed, Defined, Quantitatively Managed, and Optimizing. At Level 1 (Initial), processes are ad hoc and unpredictable. At Level 2 (Managed), work is planned and tracked but not yet standardized. At Level 3 (Defined), processes are formally documented and consistently applied across the organization. At Level 4 (Quantitatively Managed), quantitative metrics are used to monitor and control process performance over time. At Level 5 (Optimizing), the organization engages in continuous, data-driven process improvement across successive cycles. This five-level hierarchy translates naturally into a compliance scoring framework, where the progression from ad hoc verbal claims through formal documentation to quantitative tracking and continuous improvement mirrors exactly how organizational compliance maturity develops in practice.

ISO/IEC 33001:2015 provides the standardized terminology and concepts for process assessment, including guidance on how assessment results are applied to the conduct of process management. Its vocabulary for describing process capability and assessment rigor provides the definitional foundation that makes the scoring scale in this framework interpretable and consistent across different assessors and deployment contexts.

NIST SP 800-53A [17] provides the third layer of the scoring model's foundation: the evidence hierarchy. NIST SP 800-53A defines three assessment methods: examine, interview, and test. In the context of this framework, these three methods map to the lower, middle, and upper ranges of the scoring scale respectively. Verbal claims without documentation correspond to the interview method at its lowest depth, producing scores of 0 to 1. Examination of submitted policy documents and architecture artifacts corresponds to the examine method, producing scores of 2 to 3 depending on the formality and consistency of the documentation. Technical evidence including test reports, audit logs, and quantitative metrics corresponds to the test method at depth, producing scores of 4 to 5 when supplemented by longitudinal trend data demonstrating continuous improvement.

The 0-to-5 scale used in this framework adapts all three standards to the AI compliance context, where the evidence artifacts of interest include policy documents, architecture diagrams, adversarial test reports, monitoring dashboards, and quantitative performance metrics. The alignment with NIST SP 800-53A ensures that the scoring logic is grounded in an established federal evidence standard, strengthening the framework's defensibility in organizational compliance settings and ensuring that scores reflect the depth of demonstrable compliance rather than self-reported assertions.

2.5 Positioning of This Framework Relative to Prior Work

The literature surveyed in this chapter establishes three things. First, the AI governance framework landscape is well developed at the principles level but operationally incomplete for LLM-specific deployments: NIST AI RMF, ISO/IEC 42001, and the OECD AI Principles collectively provide governance structure but not the testable, evidence-anchored audit criteria that organizations need to demonstrate compliance in practice. Second, individual components of an automated compliance assessment pipeline have been demonstrated in isolation: structured control mappings have been implemented programmatically, LLMs have been applied to regulatory text interpretation, and AI-assisted audit workflows have been shown to be viable. Third, no prior work has integrated these components into a single, unified, operationally deployable compliance assessment tool grounded in empirically validated LLM-specific control logic.

The urgency of this gap is significant. As LLM deployments proliferate across regulated industries including financial services, healthcare, and legal services, the absence of practical compliance tooling creates a situation where organizations are expected to demonstrate responsible AI governance without reliable, standardized methods for doing so. Commercial AI governance platforms such as Microsoft Azure AI Content Safety and IBM OpenPages partially address this space but from a different angle: they monitor running model behavior through telemetry and API-level filtering rather than assessing an organization's governance posture against external standards. They are complementary to structured compliance assessment rather than substitutes for it, their dashboards and logs are precisely the kind of evidence artifacts an assessor would submit to a framework like this one. The frameworks exist; what is missing is the operationalized, tooling-supported assessment infrastructure that makes them practically applicable.

This thesis addresses the gap by combining three elements that prior work has not brought together: a control library that operationalizes all ten OWASP LLM Top 10 (2025) risk categories into structured, evidence-requesting audit questions with explicit standard citations; an interactive AI audit agent that automates evidence analysis and score generation against those controls; and an adaptive selection architecture that tailors the control set to the specific deployment type without introducing LLM non-determinism into compliance logic. The three composite controls for LLM07, LLM08, and LLM10 constitute a research finding in their own right, demonstrating that NIST AI RMF 1.0 alone is insufficient to cover the current LLM threat landscape and that operationalization work of the kind performed in this thesis is a necessary step between publishing a governance standard and applying it to real LLM deployments.

Chapter 3. Research Methodology

3.1 Research Design and Approach

This thesis adopts a Design Science Research (DSR) methodology, which is the appropriate paradigm for research that produces a purposefully designed artifact as its primary contribution and evaluates that artifact empirically against a defined set of research objectives. The foundational DSRM process model was established by Peffers et al. (2007), who defined design science research as a six-step process encompassing problem identification and motivation, definition of objectives for a solution, design and development, demonstration, evaluation, and communication of results [22]. DSR is the appropriate choice for this research rather than alternatives such as action research or case study research because the central contribution is not an observation about an existing phenomenon but a novel artifact, the 30-control compliance framework and the AI audit agent, that did not exist before this research and whose construction constitutes a research act in itself. Design science research differs from purely empirical research in that the act of designing and constructing the artifact is itself a research contribution, not merely a means to an end. The artifact embodies the researcher's theoretical claims about how the problem should be solved, and its empirical evaluation provides evidence for or against those claims.

In this research, the artifact is the AI audit agent and its underlying 30-control compliance framework. The theoretical claims embedded in its design are threefold. First, that compliance requirements from NIST AI RMF and ISO/IEC 42001 can be operationalized into structured, evidence-requesting audit logic that produces consistent, defensible maturity scores. Second, that a rule-based adaptive selection layer can meaningfully filter and prioritize controls based on deployment type without introducing LLM non-determinism into compliance logic. Third, that an AI-powered evidence analysis mechanism can produce audit findings that are sufficiently calibrated to support human review and override at an operationally useful rate. The empirical evaluation in Chapter 7 provides the evidence against which these three claims are assessed.

The research followed the DSR process model across five phases. The problem identification phase produced the literature review in Chapter 2 and the gap analysis establishing that no existing tool achieves the combination of OWASP coverage, framework operationalization, and adaptive deployment context that the identified gap requires. The objectives definition phase produced the four research questions and the success criteria for each. The design and development phase produced the framework and tool documented in Chapters 4 and 5. The demonstration phase applied the tool to three synthetic deployment scenarios as documented in Chapter 6. The evaluation phase produced the scored results,

calibration findings, and human assessor comparison reported in Chapter 7. The communication phase is fulfilled by this thesis document and any subsequent publications derived from it.

3.2 Research Questions Operationalization

Each of the four research questions maps to a specific research activity within the DSR process, and each is answered through a distinct methodological strand with defined success criteria.

RQ1 is answered through the framework design documented in Chapter 4. The operationalization process treats each governance standard requirement as a specification to be translated into a concrete, evidence-requesting audit question with an explicit standard citation and an LLM-specific control objective. The success criterion for RQ1 is whether the resulting controls produce differentiated, evidence-grounded scores across three materially different deployment scenarios, and whether the tool assigns zero hallucinated standard references across all 90 assessments. Both are measurable outcomes reported in Chapter 7.

RQ2 is answered through the adaptive selection layer documented in Section 5.7. The rule-based PRIORITY_MAP assigns each control a priority tier of High, Medium, Low, or N/A per deployment type, filtering and reordering the 30-control set based on the session's deployment context. The success criteria for RQ2 are the correctness of N/A exclusions for each deployment type, the appropriateness of priority assignments relative to the documented risk profile of each deployment category, and the correct behavior of the RAG architecture override when a vector store is declared at session start.

RQ3 is answered through the calibration findings documented in Section 7.6. The proportion of controls for which the agent produces coherent, evidence-grounded rationales without assessor correction, and the specific patterns of scoring behavior observed across all three scenarios, constitute the primary evidence for RQ3. The success criterion is whether the tool generates complete, evidence-referenced rationales for all active controls without requiring assessor intervention to correct the scoring logic itself.

RQ4 is answered through the human assessor comparison documented in Section 7.7. The divergence between AI-assigned scores and human assessor scores, analyzed by domain, evidence type, and maturity level, provides the empirical evidence for RQ4. The success criterion is whether the AI-assigned scores agree with independent human assessors at a rate that would make the tool operationally useful as a first-pass compliance assessment instrument, with divergences concentrated at identifiable maturity boundary cases rather than distributed randomly across all controls.

3.3 Framework Design Methodology

The 30-control library was constructed through a systematic process grounded in three source standards: NIST AI RMF 1.0, ISO/IEC 42001:2023, and OWASP LLM Top 10 (2025). The construction process proceeded in four sequential steps, each with a defined decision rule to ensure reproducibility.

The first step was completeness mapping. Every OWASP LLM Top 10 (2025) risk category was treated as a required coverage target, ensuring that the framework addresses the full documented LLM threat landscape. The decision to use the OWASP LLM Top 10 as the completeness criterion, rather than the source governance standards alone, was deliberate: the OWASP list represents the current practitioner-validated attack surface of deployed LLM systems, meaning any framework that fails to cover it would have demonstrable gaps against real-world deployment risks. This completeness requirement drove the decision to include 30 controls rather than a smaller subset.

The second step was standard mapping. For each OWASP risk category, the governance, data, technical, and monitoring controls in NIST AI RMF and ISO/IEC 42001 that most directly address that risk were identified. The selection criterion was LLM-specificity: controls that addressed only generic IT security concerns without a meaningful LLM-specific dimension were excluded. This criterion ensured that every control in the library has an audit question that would not be equally applicable to a non-LLM software system.

The third step was composite control construction. For three OWASP risk categories with no dedicated native subcategory in any source standard, composite controls were constructed by combining two existing standard subcategories around a new LLM-specific control objective. The rationale for and composition of each composite control is documented in full in Section 4.5.

The fourth step was domain organization. The 30 selected controls were organized into six domains of five controls each. The balanced allocation of five controls per domain serves two purposes: it ensures that no single risk area dominates the overall compliance score, and it enables direct domain-level comparison in the compliance report since all domain averages share the same denominator.

3.4 Evaluation Scenario Design

The empirical evaluation of the framework follows a controlled scenario-based testing design with three conditions corresponding to three target deployment scenarios. The experimental design addresses a fundamental challenge in compliance tool evaluation: the absence of a ground-truth compliance oracle. Unlike software testing tools that can be validated against programs with known bugs,

a compliance assessment tool cannot be validated against a definitive catalog of known compliance states because no such catalog exists for the specific organizations being assessed.

The evaluation design addresses this challenge through two methodological choices. First, the deployment scenarios were designed within this research, which allowed certainty about the intended compliance posture of each scenario. Each scenario was assigned an intended maturity profile, and evidence packs were constructed to represent that profile consistently across all 30 controls. This controlled design provides a form of construct validity: the expected scoring pattern is derivable from the scenario's known design, allowing the tool's actual scoring pattern to be compared against a pre-defined expected baseline rather than post-hoc judgment.

Second, the three scenarios represent materially different deployment types and maturity profiles, providing meaningful variation across which the framework's behavior can be observed. The public-facing chatbot scenario represents low and uneven maturity, with seven controls having no evidence at all. The internal knowledge assistant scenario represents uniform moderate maturity, where documentation exists for every control but each document contains exactly one identifiable gap. The financial advisory tool scenario represents uniformly high maturity, with Board-approved policies, external red team reports, and quantitative monitoring metrics throughout. This range from early-stage governance to mature, regulated deployment tests the tool's ability to differentiate at all three meaningful regions of the scoring scale.

3.5 Evidence Pack Construction Methodology

Each deployment scenario is accompanied by a synthetic evidence pack comprising the documents an organization would submit to demonstrate compliance with each control. Synthetic evidence was used in preference to real organizational data for three reasons: access to real enterprise AI deployment documentation is restricted by confidentiality agreements, real data introduces privacy constraints that would prevent public disclosure of the evaluation dataset, and synthetic evidence allows the intended compliance posture to be precisely controlled, enabling the expected score ranges to serve as a meaningful validation baseline.

The expected score ranges for each control in each scenario were determined by the researcher before the evaluation runs began, based on the content of the evidence files and the maturity scale definitions. For example, a control where the evidence pack contains a formally documented policy signed by named executives but with one documented gap was assigned an expected range of 2 to 3, reflecting formal documentation that is not fully complete. These pre-defined targets were recorded in a

scenario overview file included in each evidence pack and were not modified after the evaluation runs began.

Evidence packs were designed to exercise the full file handling pipeline of the audit agent. Files span all supported formats including PDF, DOCX, CSV, and TXT. Each pack includes policy documents, risk registers, architecture diagrams, red team reports, monitoring data, and the scenario overview file. The file format diversity was intentional: the evaluation tests not only the scoring logic but also the file processing pipeline, since format handling failures would be a material finding about the tool's reliability in practice.

3.6 Validation Methodology

The validation methodology addresses the question of whether the audit agent's outputs can be trusted as accurate compliance findings. Two independent human assessors validated the tool's scores to establish the reliability of its output as a compliance assessment instrument.

The first assessor is a graduate peer with prior experience conducting security audits in academic research contexts, providing a baseline representing the judgment of a trained but non-practitioner assessor. The second is an industry professional with hands-on compliance auditing experience in organizational settings, providing a baseline representing practitioner-level judgment. Using assessors with these two distinct backgrounds serves a specific analytical purpose: it allows the study to detect whether score divergence from the AI tool is systematic across all assessors or concentrated in assessor-specific judgment patterns that reflect individual training and experience rather than tool calibration errors.

Both assessors were provided with access to the evidence packs and the scenario overview files, including the expected score ranges, but without access to the AI agent's scores prior to completing their own independent evaluations. This independence protocol ensures that human judgments are formed without anchoring to the tool's output. Each assessor scored all 30 controls per scenario using the same 0-to-5 maturity scale, with access to the control definitions and scoring criteria documented in Chapter 4.

The divergence between AI scores and human assessor scores is analyzed by domain, evidence type, and maturity level. Controls where evidence submission failures occurred during the AI session are analyzed as a separate category, distinct from controls where evidence was correctly submitted. This separation is methodologically important because the tool's lower score at submission-failure controls reflects the correct behavior of an evidence-based audit instrument, not a calibration error, and conflating the two categories would systematically overstate the tool's divergence from human judgment.

3.7 Threats to Validity

Following the validity threat framework established by Wohlin et al. for empirical software engineering research [23], threats are identified across four validity dimensions: internal, external, construct, and conclusion validity.

Internal Validity: The primary internal validity threat is the self-designed evaluation environment. Because the deployment scenarios and evidence packs were designed within this research, there is a potential for unconscious bias in scenario construction that could inflate or deflate scoring targets. This threat is mitigated by documenting the intended maturity profiles and expected score ranges for each control before the evaluation runs began, and by analyzing the AI tool's scores against those pre-defined targets rather than against post-hoc expectations. The separation of evidence design from evaluation execution, with scenario overview files locked before any tool assessment was run, provides a documented audit trail against this threat.

External Validity: The primary external validity threat is the generalizability of findings beyond the synthetic evidence environment. Scoring behavior observed against purpose-built policy documents and architecture files may not accurately predict scoring behavior against real organizational evidence with its inherent complexity, incompleteness, and ambiguity. The evaluation results are therefore framed throughout this thesis as demonstrated capabilities within the designed test environment rather than population-level performance estimates. Extending the evaluation to real organizational evidence under authorized data-sharing agreements is identified as the primary direction for future work in Chapter 9.

Construct Validity: The primary construct validity threat is whether the 0-to-5 maturity scale accurately measures what it claims to measure, specifically whether a score reflects genuine compliance maturity rather than evidence presentation skill. The anchoring of each maturity level to the NIST SP 800-53A evidence hierarchy addresses this threat by grounding scores in specific, externally defined evidence standards rather than assessor judgment alone. A secondary construct validity concern is whether the three deployment scenarios adequately represent the range of real-world LLM deployment contexts. This is partially mitigated by selecting scenarios that span public-facing, internal, and regulated financial contexts, but the framework's applicability to deployment types not represented in the evaluation, such as healthcare or agentic systems, remains unvalidated.

Conclusion Validity: The primary conclusion validity threat is the limited number of scenarios and the use of synthetic rather than real organizational evidence. With three scenarios, it is not possible to estimate variance in scoring behavior across the full population of LLM deployment contexts. The

evaluation results are therefore framed as existence proof and calibration documentation rather than distributional performance estimates. The specific calibration findings documented in Chapter 7 are valid as qualitative conclusions about the tool's scoring behavior regardless of scenario count, but quantitative claims about population-level accuracy rates would require a substantially larger evaluation dataset.

Chapter 4. Framework Design

4.1 Framework Design Philosophy

The framework is designed around a single organizing principle: every assessment question must trace directly to a named requirement in a recognized governance standard, and every score must reflect the quality of evidence submitted rather than the assessor's subjective judgment about an organization's intent. This traceability requirement drives every design decision in the control library, the scoring model, and the dialogue flow.

This design philosophy has two practical consequences. First, it excludes controls that address general organizational security hygiene without a meaningful LLM-specific dimension. Controls that could apply equally to any software system without modification are not included in the library. Every control in the library has an LLM-specific control objective that describes a risk or governance requirement that arises specifically from the deployment of an LLM system, not from software deployment generally. Second, it requires that every score from 0 to 5 correspond to a specific category of evidence rather than a general level of organizational maturity. This anchoring ensures that scores are reproducible across assessors who apply the same criteria to the same evidence, which is a prerequisite for using the tool as a compliance instrument rather than a subjective assessment aid.

The tool is designed as an advisory accessor: it gathers and evaluates submitted evidence but does not make binding compliance determinations. Human judgment remains the authoritative decision point at every control. This distinction is operationalized through the human override mechanism described in Section 5.6, where every LLM-assigned score is presented to the assessor for review and correction before it is recorded as final. The tool's role is to surface evidence-grounded findings consistently and efficiently; the assessor's role is to apply professional judgment, organizational context, and regulatory experience that no automated system can fully replicate. This positioning as an advisory instrument rather than a decision-making authority is a deliberate design choice, not a limitation to be engineered away in future versions.

4.2 Framework Selection Rationale

Three regulatory frameworks were selected on specific, documented grounds. NIST AI RMF [2] was chosen for its comprehensive AI lifecycle coverage and its status as the de facto United States federal AI governance standard. The framework provides 72 subcategories across 19 categories and four core functions, and its four-function structure of Govern, Map, Measure, and Manage maps naturally to the six assessment domains of this framework, with Govern subcategories informing the Governance and

Accountability domain, Map subcategories informing the Risk Assessment domain, Measure subcategories informing the Technical Reliability and Transparency domains, and Manage subcategories informing the Monitoring domain.

ISO/IEC 42001 [3] was selected as the first internationally recognized AI management system certification standard, providing complementary coverage particularly in data governance, AI lifecycle management, and supplier relationships. Its clause-and-control format, modeled on ISO/IEC 27001, makes its requirements directly translatable into specific audit questions with named evidence artifacts, which is the operationalization approach this framework requires.

ISO/IEC 27001 [19] was incorporated selectively to address access control requirements for AI assets, including model weights, API keys, and fine-tuning datasets, that are not adequately covered by the other two frameworks. Annex A 5.15 of ISO/IEC 27001:2022 focuses on establishing and applying access control rules for information and associated assets, and it provides the standard citation for the access control domain control in this framework, specifically for role-based access control applied to LLM APIs and model infrastructure.

A fourth framework, NIST SP 800-53A [17], is not a primary source standard but provides the evidence hierarchy that anchors the scoring model. Its three-tier distinction between examine, interview, and test maps directly to the lower, middle, and upper ranges of the 0-to-5 maturity scale, as documented in Section 4.6.

4.3 The Six-Domain Structure

The 30 controls are organized across six domains of five controls each. This balanced allocation enables direct domain-level comparison in reporting and ensures that each of the major risk categories identified in the LLM security literature receives structured coverage.

Table 4.1. Six-domain structure of the compliance framework.

Domain	Primary Risk Area	Primary Standards
1. Governance and Accountability	Policy gaps, shadow AI, accountability vacuums	NIST GOV 1.2, 2.2, 5.2; ISO 42001 A.2.2, A.3.2
2. Data Integrity and Provenance	Training data poisoning, PII leakage, IP risk	ISO 42001 A.7.1-7.4; NIST MAP 2.2
3. Risk Assessment and Supply Chain	Hallucination, third-party model risk	NIST MAP 1.1, 3.1, 4.1; ISO 42001 A.5.1, Cl.8.4

Domain	Primary Risk Area	Primary Standards
4. Technical Reliability and Security	Adversarial attacks, insecure deployment	NIST MEA 2.7, 2.2; ISO 42001 A.6.3, A.6.4; ISO 27001 A.5.15
5. Transparency and Documentation	AI deception, liability gaps, explainability failures	ISO 42001 A.8.1-A.8.3, A.10.2; NIST MEA 2.11
6. Monitoring and Continuous Improvement	Model drift, undetected failures, zombie models	NIST MAN 4.1, 3.2; NIST GOV 2.3/MEA 2.6; ISO 42001 A.6.6, A.6.7

The Governance and Accountability domain (D1) addresses the organizational structures and policies that govern LLM deployment. Its five controls cover acceptable use policies, accountability role assignment, system prompt confidentiality, resource and cost governance, and vector store security, drawing primarily on the NIST GOVERN function subcategories that address organizational AI risk culture and the ISO 42001 clauses covering AI management system scope and leadership accountability.

The Data Integrity and Provenance domain (D2) addresses the quality, lineage, and protection of data used to train, fine-tune, and operate LLM systems. Its controls draw on ISO 42001 Annex A clauses 7.1 through 7.4, which specifically address data acquisition, data quality, and data documentation requirements for AI systems, areas where data provenance failures create direct LLM risks including training data poisoning and sensitive data leakage.

The Risk Assessment and Supply Chain domain (D3) addresses how organizations identify, document, and manage the specific risks introduced by their LLM deployments, including risks from third-party model providers and base model hallucination. Its controls draw on NIST MAP subcategories that cover context establishment, risk likelihood and impact documentation, and third-party AI component risk management.

The Technical Reliability and Security domain (D4) addresses the technical mechanisms that protect LLM systems from adversarial inputs, insecure code generation, and unauthorized access. This is the domain most directly aligned with OWASP LLM Top 10 operational risks, drawing on NIST MEA 2.7 for adversarial testing, ISO 42001 A.6.3 and A.6.4 for output handling and verification, and ISO 27001 A.5.15 for access control of LLM infrastructure.

The Transparency and Documentation domain (D5) addresses the documentation organizations produce to communicate how their LLM systems work, what they can and cannot do, and how to report failures. ISO 42001 Annex A clauses 8.1 through 8.3 specifically require AI system documentation, user

notification of AI interaction, and incident communication, making them the primary sources for this domain.

The Monitoring and Continuous Improvement domain (D6) addresses the ongoing operational practices that detect model degradation, security incidents, and governance failures in deployed LLM systems. Its controls draw on NIST MANAGE subcategories covering post-deployment monitoring, incident response, and change management, as well as ISO 42001 clauses A.6.6 and A.6.7 covering AI system operation and performance tracking.

4.4 The 30-Control Library

The control library consists of 30 controls with five controls per domain. Each control contains seven attributes: the domain number, the control ID referencing the source standard subcategory, the control name, the LLM-specific control objective, an example safeguard, the expected evidence types, and the OWASP LLM Top 10 mapping.

Why exactly 30 controls? NIST AI RMF 1.0 alone contains 72 subcategories, and its companion Playbook adds hundreds of suggested actions. ISO/IEC 42001 adds further controls across its Annex A clauses. Including all of them would make the tool impractical to use, require multiple working sessions to complete, and produce a report too voluminous to act on. On the other hand, using too few controls would leave important risk areas uncovered. The number 30 was chosen to satisfy two constraints simultaneously: every risk category in the OWASP LLM Top 10 (2025) must be addressed by at least one control, and a full assessment must be completable in a single working session of approximately two to three hours, which is important for practical adoption in organizational settings.

The five-controls-per-domain structure serves a specific reporting purpose: it ensures that all six domain scores share the same denominator of five times five, making them directly comparable in the compliance report without normalization. A domain with three controls and one with seven would produce scores on incompatible scales, obscuring meaningful comparisons between governance and technical maturity.

Controls were selected by going through each of the ten OWASP LLM risk categories and identifying the governance, data, technical, and monitoring controls in NIST AI RMF and ISO/IEC 42001 that most directly address that risk. Controls that addressed only generic IT security concerns without a meaningful LLM-specific dimension were excluded. The five-controls-per-domain structure was applied after selection to finalize the library, with the domain boundaries drawn around natural thematic clusters in the selected control set.

It is important to note that OWASP coverage is not uniform in nature across the source standards. Seven of the ten OWASP LLM Top 10 (2025) risk categories map directly to dedicated, purpose-built subcategories already present in NIST AI RMF 1.0 or ISO/IEC 42001:2023. The remaining three risk categories, System Prompt Leakage (LLM07), Vector and Embedding Weaknesses (LLM08), and Unbounded Consumption (LLM10), have no dedicated native subcategory in any of the three source standards. For these three risks, composite controls were constructed as described in the following section.

4.5 The Three Composite Controls

During the control selection process, a structural gap was identified across the source standards that is critical to understanding three of the thirty controls in this library. To understand why composite controls are necessary, it helps to understand what each source standard was designed to do and when it was written.

NIST AI RMF 1.0 was published in January 2023. At that time, the widespread enterprise deployment of LLMs was in its earliest stages, and several attack patterns that are now well-documented in the security research community had not yet been formally identified or named. NIST AI RMF was also designed as a general-purpose AI risk management framework applicable to many categories of AI systems, not specifically to LLM-based products. ISO/IEC 42001:2023, while more recent, is similarly a general AI management system standard not designed around LLM-specific deployment risks. As a result, three of the ten OWASP LLM Top 10 (2025) risk categories have no dedicated, purpose-built subcategory in any of these standards. For each of the three composite controls below, the two component subcategories were selected because their combined scope, when applied to the specific LLM risk context, produces a control requirement that neither subcategory achieves independently.

System Prompt Confidentiality (OWASP LLM07) addresses the risk that an attacker manipulates the AI system into revealing the hidden instructions it was given, its system prompt. These instructions often contain sensitive business logic, safety rules, or API credentials. NIST AI RMF 1.0 has no subcategory that specifically addresses the confidentiality of system prompts as an architectural security concern. This framework addresses this through a composite control combining NIST MEA 2.7, which covers security evaluation and adversarial testing of the AI system including extraction-attack tests, with NIST GOV 4.3, which covers organizational procedures for engaging external actors and managing third-party risks, including the testing and oversight processes that validate confidentiality controls. Together, these two subcategories support a control that requires organizations to classify system prompts

as confidential, conduct prompt extraction-attack tests, and ensure the system is technically configured to never echo its system prompt in outputs.

Vector and Embedding Security (OWASP LLM08) addresses the risks that arise in LLM deployments using Retrieval-Augmented Generation (RAG), where the AI searches a vector database of documents before responding. Attackers can manipulate this architecture by poisoning the documents stored in the vector database, manipulating what the AI retrieves, or exploiting weak access controls on the database itself. This deployment pattern did not yet exist at mainstream scale when NIST AI RMF 1.0 was written in January 2023. The closest native controls treat vector databases as generic third-party components rather than as a distinct attack surface with its own vulnerability profile. This framework addresses this through a composite control combining NIST MAP 4.2, which covers internal risk controls for AI components including third-party technologies, with NIST MEA 2.7, the adversarial testing and security evaluation subcategory, requiring organizations to apply access controls to vector stores, validate all ingested documents before embedding, and monitor for anomalous retrieval patterns.

Resource and Cost Governance (OWASP LLM10) addresses the risk of an AI system consuming excessive computational resources, incurring runaway costs, or becoming unavailable because no usage limits were placed on the system. An attacker could, for example, repeatedly send very long inputs or deliberately trigger expensive operations to exhaust an organization's API quota or cause denial of service. NIST AI RMF 1.0 has no subcategory covering rate limiting, token quotas, cost governance, or denial-of-service protection for AI APIs. This framework addresses this through a composite control combining NIST MAN 1.3, which covers the identification and prioritization of AI risks with planned responses, with NIST GOV 1.3, which covers risk management activities calibrated to the organization's risk tolerance. Together, these two subcategories require organizations to set per-user rate limits, token quotas, cost alerts, and timeouts on all LLM inference calls as a documented, risk-tolerance-calibrated operational control.

It is worth emphasizing that NIST itself recognized these structural gaps after publishing AI RMF 1.0. In July 2024, NIST released NIST AI 600-1 [21], a companion document titled the Generative AI Profile, specifically designed to extend the original framework with supplementary guidance for generative AI systems that the 2023 standard did not adequately address. The same structural gap in these source standards is also acknowledged by NIST AI 600-1, confirming that operationalization work of the kind performed in this thesis is a necessary step between publishing a governance standard and applying it to real LLM deployments. The three composite controls in this library serve two purposes: they close the OWASP coverage gap so that the framework assesses all ten current LLM risk categories, and they

constitute a research finding in their own right, demonstrating that NIST AI RMF 1.0 alone is insufficient to cover the current LLM threat landscape.

4.6 The 0-5 Maturity Scoring Model

Each control is scored on a 0-to-5 maturity scale grounded in ISO/IEC 33001 [16] and CMMI [15], with scoring thresholds anchored to the NIST SP 800-53A evidence hierarchy [17]. The six levels are defined as follows.

Table 4.2. Maturity scale definition and evidence requirements.

Score	Level	Evidence Requirement
0	No Control Exists	No evidence of any kind; the control has never been implemented or considered.
1	Verbal Claim Only	The assessor verbally asserts the control exists but can produce no documented evidence of any kind.
2	Partial or Informal Implementation	A one-time or ad hoc effort exists but there is no formal policy, no repeatable process, and no evidence of consistent application.
3	Formal Documentation, Consistently Applied	A formal policy or procedure exists, is consistently applied, and documented evidence can be produced on request.
4	Quantitative Metrics, Tracked Over Time	All of score 3 plus quantitative metrics tracking the control's implementation over time, linked to a review or remediation process.
5	Continuously Improved, Automated	All of score 4 plus automated monitoring, longitudinal trend analysis across multiple time periods, and documented evidence of iterative improvement.

The three NIST SP 800-53A assessment methods map directly to the scoring bands. The interview method, corresponding to oral inquiry without documentary support, anchors scores 0 and 1. The examine method, corresponding to inspection of submitted documents and artifacts, anchors scores 2 and 3 depending on the formality and consistency of what is produced. The test method, corresponding to technical verification of implemented controls, anchors scores 4 and 5 when supplemented by longitudinal data demonstrating iterative improvement.

To illustrate how these levels apply in practice, consider the Adversarial Testing control (NIST MEA 2.7), which requires organizations to evaluate their LLM system's resilience against adversarial inputs such as prompt injection attempts. A score of 0 means the organization has never conducted any form of adversarial testing. A score of 1 means the assessor verbally states that the team periodically tries

to break the model but can produce no documented test cases, results, or schedule. A verbal claim, as used at Score 1, refers specifically to a textual assertion that a control exists or is practiced, submitted without any accompanying evidence file. It is distinguished from Score 2 solely by the absence of any artifact, no policy document, no screenshot, no data export. Verbal claims are fully included in the evaluation and are a legitimate compliance posture for early-stage organizations; the scoring model accommodates them rather than treating them as invalid inputs. A score of 2 means a partial record exists, for example a one-time informal red team exercise was conducted at launch, but there is no formal methodology and no evidence the exercise has been repeated. A score of 3 means the organization can produce a formal adversarial testing policy, a documented set of test cases covering prompt injection scenarios, and evidence the tests were conducted against the current production model. A score of 4 means all of the above exists and is supplemented by quantitative metrics such as an injection success rate tracked over time, with test results reviewed on a defined schedule and linked to remediation. A score of 5 means the organization has a fully automated adversarial testing pipeline integrated into the deployment workflow, with continuous monitoring and documented evidence of iterative improvement in resilience across successive test cycles.

This progression from no evidence through verbal claims, informal records, formal documentation, quantitative tracking, and continuous automated improvement applies consistently across all 30 controls in the library. The use of the NIST SP 800-53A evidence hierarchy ensures that scores at each level reflect a specific standard of demonstrable compliance rather than self-assessed organizational maturity.

4.7 The Audit Dialogue Flow

Each control is operationalized as a specific, evidence-requesting audit question. Questions are designed to surface absence as readily as presence by asking for named, concrete artifacts rather than general practices. For example, the governance roles control (NIST AI RMF GOV 2.1) asks: 'Who holds the title of AI Model Owner or AI Risk Officer for this deployment? Can I see the RACI chart or documented roles defining their authority and lines of communication? This framing ensures that inability to produce the named artifact is itself a finding, and that a verbal claim of having defined roles without a documented RACI chart produces a score of 1 rather than 3.

Each question is accompanied by its originating standard reference, the specific control objective, a list of expected evidence types, and the relevant OWASP LLM Top 10 mapping. The OWASP mapping serves an important communication function: it allows the assessor to understand why a governance or data control is part of a security-oriented framework. Every control in the library traces to a specific LLM

vulnerability class that it is designed to address, making the connection between governance practice and security outcome explicit rather than implicit.

In standard mode, the same generic question is presented for every assessment session regardless of deployment type. In adaptive mode, the question is rewritten by the LLM to be specific to the deployment context, incorporating the system name, deployment type, and a risk scenario relevant to that deployment category, while preserving the same compliance intent and standard reference. The question rewriting is scoped to language generation only. The adaptive layer does not alter which controls apply, what their priority tier is, or what scoring criteria the LLM uses to evaluate the response.

4.8 The Audit Report Design

The compliance report generated by the audit agent contains five primary sections, each serving a distinct purpose for the assessor or organizational stakeholder who reads it.

The session metadata section records the project name, deployment type, assessment mode, date, and the number of human overrides applied during the session. The override count provides an immediate session-level indicator of assessor-AI agreement and is relevant for RQ4 analysis.

The domain performance section presents each of the six domains with its arithmetic mean score expressed as a decimal out of 5.0, accompanied by a color-coded progress bar. The color thresholds progress from red at the lowest maturity through amber, yellow, and teal to green at the highest, enabling rapid visual identification of the domains with the most critical gaps.

The per-control table lists every active control with its LLM-assigned score, the assessor-confirmed final score, and the key finding extracted from the rationale. Displaying both scores separately makes the divergence between AI and human judgment visible at the control level, which is the primary data source for the RQ4 analysis in Chapter 7.

The RQ2 adaptive summary section is present only in adaptive mode. It documents the control count by priority tier (High, Medium, Low), the N/A exclusion list with the reason for each exclusion, and the RAG architecture override status indicating whether vector store controls were reinstated.

The remediation priority table lists all controls scoring 2 or below, sorted in ascending order of score, with the agent's specific actionable recommendation for each. Sorting ascending rather than descending ensures the most critical gaps, those at score 0 and 1, appear at the top of the table where they are most likely to drive immediate organizational action.

The overall compliance percentage is computed as the sum of all final scores divided by the product of the active control count and 5, expressed as a percentage. In standard mode this denominator is always 150 (30 times 5). In adaptive mode the denominator reflects the actual active control count, which may be less than 30 if some controls are excluded as N/A for the deployment type.

Chapter 5. System Design and Implementation

5.1 Tool Architecture Overview

The AI audit agent is implemented as two self-contained single-file web applications: audit_agent.html for standard mode and audit_agent_adaptive.html for adaptive mode with the RQ2 features. Both require no backend framework, no database, and no server-side rendering. The only server component is server.js, a lightweight Node.js CORS proxy running on localhost port 3001. Its sole function is to inject the Anthropic API key server-side and forward requests to api.anthropic.com/v1/messages, because browsers cannot call the Anthropic API directly from a local HTML file due to CORS restrictions. The proxy adds three headers before forwarding the request unchanged: x-api-key, anthropic-version 2023-06-01, and Content-Type application/json.

The data flow is: Browser to localhost port 3001 (Node.js proxy) to api.anthropic.com/v1/messages to response back to browser. This architecture was chosen to keep the implementation maximally portable and reproducible: any researcher with Node.js and a browser can run the complete evaluation environment without configuring a database, deploying to a server, or managing containerized infrastructure.

5.2 LLM Backend Selection

The LLM backend for the audit agent was finalized as Claude Haiku (claude-haiku-4-5-20251001). The selection was based on three criteria evaluated against the evaluation protocol's specific requirements.

Cost efficiency: the evaluation protocol requires 90 API calls across three scenarios of 30 controls each. Claude Haiku's per-token pricing at evaluation time made it the lowest-cost option for the required call volume without compromising response quality for structured compliance assessment tasks.

Native PDF document support: the Anthropic document API allows PDF files to be submitted directly as document-type content blocks rather than requiring client-side text extraction. This is particularly important for compliance assessment, where evidence artifacts frequently include scanned policy documents and signed reports in PDF format.

Scoring consistency: Claude Haiku was evaluated against GPT-4o on a subset of controls using identical evidence and scoring criteria. Claude Haiku demonstrated greater calibration stability, defined as lower variance in scores assigned to identical evidence across repeated calls, making it the more suitable choice for a compliance tool where reproducibility is a design requirement.

5.3 The Three-Screen Workflow

The audit agent presents a three-screen workflow designed to guide the assessor through a complete compliance session with no prior training required.

Screen 1 is the landing page. It collects the project or system name as free text, the deployment type from a dropdown of seven options (Public-Facing Chatbot, Internal Knowledge Assistant, Financial Advisory Tool, Healthcare Support System, Code Generation Assistant, Customer Support Agent, and Other), a RAG architecture toggle in adaptive mode indicating whether the system uses a vector store, and a standard/adaptive mode toggle. On Begin Audit Session, the tool captures all inputs, calls the adaptive control selection function if in adaptive mode, and renders the first control card.

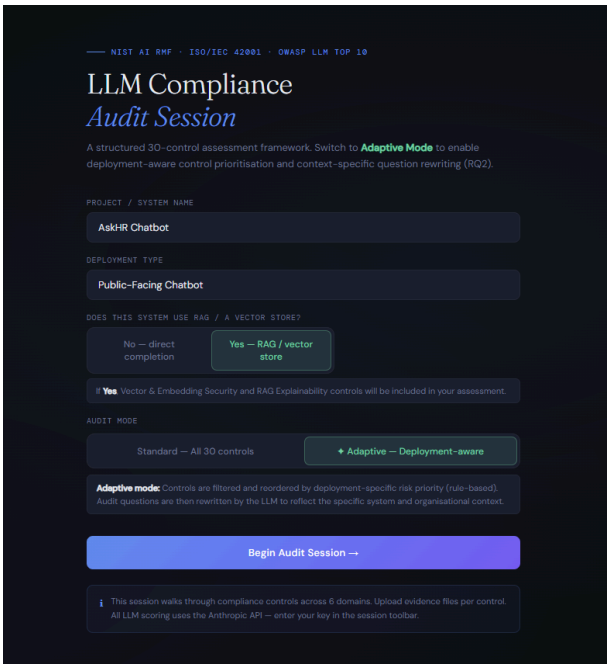

Figure 5.1. — Audit Session Configuration Landing Page

In Figure 5.1. the audit agent landing page in adaptive mode, showing the session configuration fields including project name, deployment type dropdown, RAG architecture toggle, and the Begin Audit Session button. These inputs are captured at session start and passed to the PRIORITY_MAP layer to filter and prioritize the 30-control set before the first control card is rendered.

Screen 2 is the audit chat interface. A scrollable message area presents one control at a time. Each control card shows the control number and domain badge, the control ID and name, the priority badge in adaptive mode, the OWASP tags, the audit question, the control objective in supporting text, and the

expected evidence types. After the assessor submits a text response and any evidence files, a thinking indicator appears while the API call is in progress, then a score card replaces it.

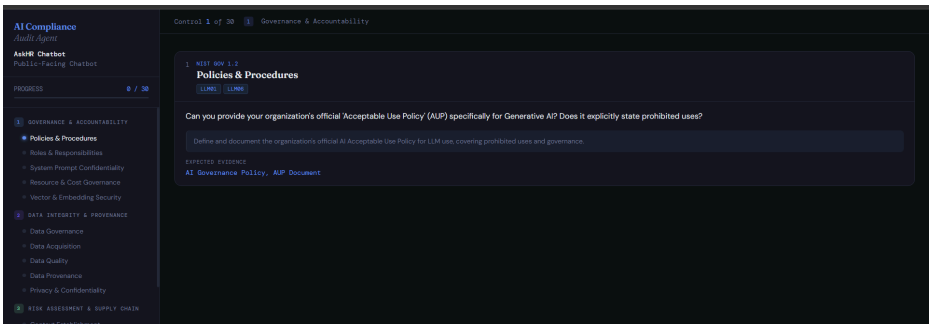

Figure 5.2. The Audit Chat Interface

In Figure 5.2. the audit chat interface displaying a control card in standard mode, showing the control ID, domain badge, OWASP mapping tags, audit question, control objective, and expected evidence types. The assessor enters a text response and optionally attaches evidence files before submission.

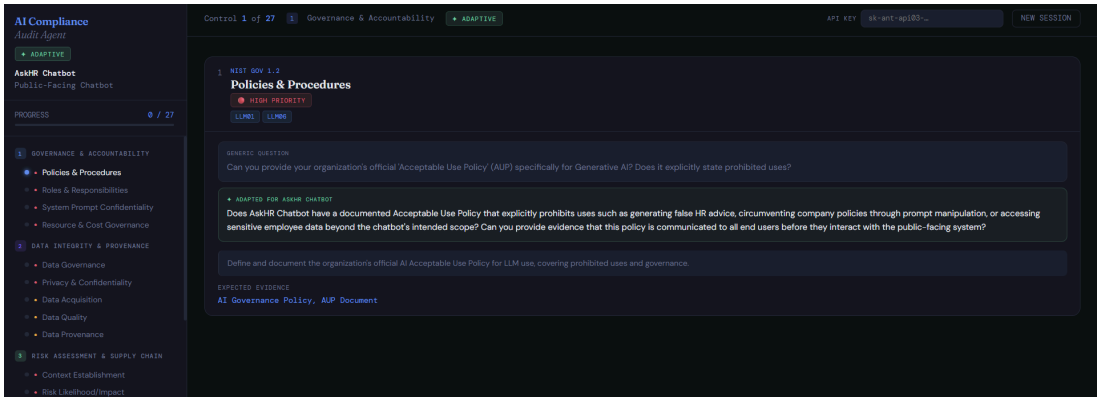

Figure 5.3. The Adaptive Mode Control

In Figure 5.3. the adaptive mode control card showing the generic question alongside the LLM-adapted question labeled with the project name, illustrating how the question adaptation feature produces deployment-specific questions while preserving the original compliance intent and standard reference.

Screen 3 is the compliance report. It is generated after all active controls are confirmed and contains all five sections described in Section 4.8. The report is rendered in the browser and can be printed or saved as a PDF.

5.4 Evidence File Handling Pipeline

The tool accepts 15 file formats. Processing happens client-side before the API call, ensuring that the LLM receives content in the format best suited to its native capabilities.

PDF files are converted to base64 and sent as type document using the Anthropic document API, enabling native document understanding without text extraction. PNG, JPG, JPEG, WEBP, and GIF files are converted to base64 and sent as type image for native vision processing, enabling the LLM to read screenshots, architecture diagrams, and scanned documents. DOCX and DOC files are parsed client-side using Mammoth.js to extract raw text, then sent as type text truncated to 8,000 characters. XLSX, XLS, CSV, JSON, XML, LOG, and TXT files are read as plain text via the FileReader API and sent as type text truncated to 8,000 characters.

A critical bug fix was applied during Winter 2026 testing. DOCX files were originally transmitted as raw binary, which caused garbled content to be sent to the API and produced artificially low scores for any control where DOCX evidence was submitted. The fix reads the file as an ArrayBuffer first, then passes it to Mammoth.js for text extraction before sending.

5.5 The Scoring LLM Call

For every control assessment in standard mode, the tool sends a structured prompt to Claude Haiku with the following system prompt:

You are an expert AI compliance auditor assessing an organization's LLM deployment against a structured 30-control compliance framework based on NIST AI RMF 1.0, ISO/IEC 42001:2023, and ISO/IEC 27001:2022, aligned with OWASP LLM Top 10 (2025). The session context includes the project name, the deployment type, the control number and ID, the control name, the control objective, the OWASP mapping, the expected evidence types, and the audit question that was asked. Score using the 0 to 5 maturity scale grounded in ISO/IEC 33001, CMMI, and NIST SP 800-53A. Respond ONLY with valid JSON in the specified format, nothing else.

The JSON response format requires five fields: score as an integer from 0 to 5, rationale as two to three sentences citing specific aspects of the response and evidence quality, strength as one specific strength or None identified if score is 1 or below, gap as one specific gap or missing evidence item, and recommendation as one specific actionable improvement recommendation.

In adaptive mode, two additions are injected into the session context: the audit mode is labeled Adaptive and the control priority for this deployment type is included. The audit question field uses the adapted question rather than the generic one, so the LLM scores against the deployment-specific question.

5.6 The Human Override Mechanism

After the LLM assigns a score, the assessor sees the numeric score from 0 to 5 with a colored badge, a two to three sentence rationale citing specific evidence aspects, three pills displaying one strength, one gap, and one recommendation, six override buttons pre-selecting the LLM score, and a Confirm and Continue button.

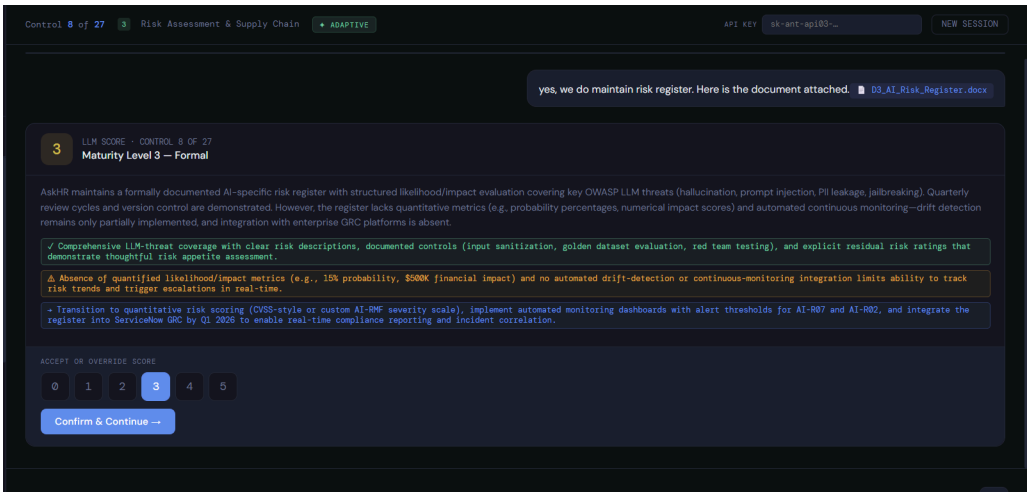

Figure 5.4 - LLM Maturity Scoring Output with Evidence Citation

In Figure 5.4. The score card displayed after evidence submission, showing the LLM-assigned maturity score with colored badge, the rationale paragraph, strength, gap, and recommendation pills, and the six override buttons with the LLM score pre-selected. The assessor can accept the LLM score or select any other value before confirming.

The assessor can accept the LLM score or select any other value. Both are recorded in the session state. Three design decisions actively discourage automation bias. The override interface presents all six score options as equal-weight buttons, the LLM's score is pre-selected but visually identical to the alternatives, avoiding anchoring. The rationale paragraph surfaces specific evidence gaps and weaknesses, not only strengths, ensuring the assessor sees reasons to consider overriding. The compliance report displays both the LLM-assigned and assessor-confirmed scores in separate columns for every control, making any pattern of uncritical acceptance visible and auditable after the fact. The session stores the LLM score and the final score separately for every control, enabling the divergence between them to be

analyzed across all 90 data points in the formal evaluation. The report header shows the override count for the session. In the formal evaluation, the operator accepted every LLM score across all three scenarios, producing an override rate of 2 out of 90 total assessments, both corrections applied to controls affected by the pre-fix DOCX binary failure in Scenario 1.

5.7 The Adaptive Control Selection Layer

The adaptive control selection layer is the core RQ2 contribution. Adaptivity in this framework has a specific, bounded definition: the deterministic reordering and scoping of the 30-control set based on the deployment type declared at session start, with no LLM involvement in the selection logic. The framework does not use the LLM to decide which controls apply or in what order, doing so would introduce non-determinism incompatible with a compliance instrument requiring reproducible, auditable outputs. Instead, it is implemented as a static JavaScript lookup table called PRIORITY_MAP that encodes, for each deployment type, a priority tier of High, Medium, Low, or N/A for each of the 30 controls. The LLM's involvement begins only after this deterministic filtering step, limited to generating adapted question wording (Section 5.8) and evaluating submitted evidence (Section 5.5).

The structure is PRIORITY_MAP indexed by deployment type and then by control ID, returning an object containing the priority field (high, medium, low, or na) and a na_reason string when applicable. Entries exist for all three deployment types used in the evaluation: Public-Facing Chatbot, Internal Knowledge Assistant, and Financial Advisory Tool.

The decision to make the priority mapping rule-based rather than LLM-generated is deliberate and architecturally important. Deterministic, reproducible control selection is a hard requirement for a compliance tool. If the set of controls selected for an assessment varied across sessions for the same deployment type, the tool's output would not be comparable across organizations or over time. The LLM's role is appropriately scoped to evidence analysis and scoring, which are language tasks requiring semantic understanding. Control selection is a compliance logic task requiring consistency and auditability, not creativity.

The priority counts before the RAG override are: for Public-Facing Chatbot, 16 High, 9 Medium, 2 Low, and 3 N/A (Vector and Embedding Security, RAG Explainability, and Agentic Tool Registry); for Internal Knowledge Assistant, 11 High, 15 Medium, 4 Low, and 0 N/A since RAG is in scope by default; and for Financial Advisory Tool, 25 High, 1 Medium, 0 Low, and 3 N/A matching the chatbot exclusions plus Agentic Tool Registry.

The RAG architecture intake question was added during the evaluation phase after identifying a structural false-negative: both the Public-Facing Chatbot and Financial Advisory Tool categories assumed no RAG architecture, incorrectly marking Vector and Embedding Security and RAG Explainability as N/A for systems that do use vector stores. The fix adds a binary toggle on the landing page in adaptive mode asking whether the system uses a RAG or vector store. When set to Yes, the tool re-enables both RAG controls at Medium priority and displays a RAG architecture detected label in those controls' priority badges. This fix was applied before the formal evaluation runs and validated across all three scenarios, all of which used Pinecone as their vector store.

5.8 The Question Adaptation Feature

A separate, smaller API call fires once per control at render time in adaptive mode, before the assessor can respond. It is cached in the adapted questions object so it only fires once per control per session. The model used is Claude Haiku with a maximum of 300 tokens.

The question adaptation system prompt instructs the model to rewrite a generic compliance audit question so it is specifically tailored to the organization and deployment context, keeping the same compliance intent and standard reference, making the question concrete and specific to the system using the system name and deployment type, adding one specific risk scenario relevant to this deployment type that the generic question does not mention, keeping it to two to three sentences, not changing what standard or control is being assessed, and returning only the rewritten question text with no preamble.

The adaptation call is deliberately scoped to language generation only. The LLM rewrites the question wording but does not decide which controls apply or what their priority is. That decision is made by the deterministic PRIORITY_MAP layer. This boundary is intentional and preserves the reproducibility of control selection while enabling the natural language improvement that makes the audit experience more useful for the assessor.

5.9 The Compliance Report

The overall compliance score is computed as the sum of all final scores divided by the product of the active control count and 5, expressed as a percentage. For standard mode this is always out of 150, meaning 30 controls times 5. For adaptive mode the denominator reflects the actual active control count. In the formal evaluation, Scenario 1 and Scenario 3 each had 29 active controls after N/A exclusions, giving a denominator of 145.

Domain scores are the arithmetic mean of final scores for all active controls in that domain, displayed as a decimal out of 5.0 with a color-coded progress bar. The color thresholds are: green at 80

percent and above, teal at 60 percent, yellow at 40 percent, amber at 20 percent, orange-red at 8 percent, and red below 8 percent.

The RQ2 adaptive summary section shows the count of controls assessed versus 30 total, the count breakdown by priority tier, the N/A count with an exclusion table showing control ID, name, domain, and reason, the RAG override notice if applicable, and three before-and-after question comparison examples using the first three high-priority controls.

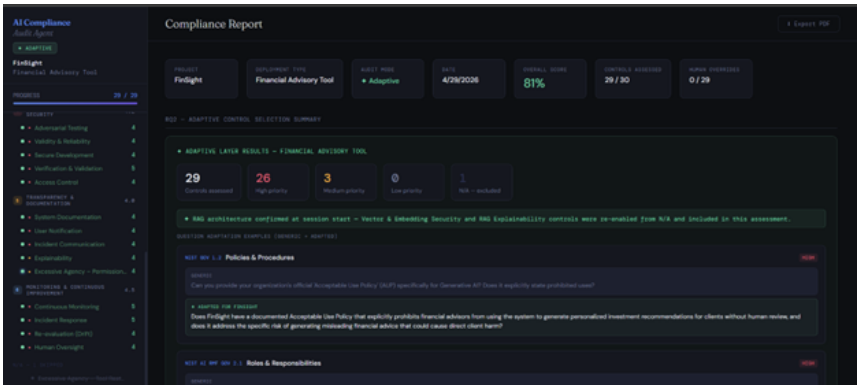

Figure 5.5 — Adaptive Compliance Report Header: FinSight/Meridian Capital Bank

In Figure 5.5. the compliance report dashboard showing the overall compliance percentage, domain performance scores with color-coded progress bars, and the remediation priority table listing all controls scoring 2 or below sorted in ascending order of score. The report header displays the session metadata including the number of human overrides applied during the assessment.

5.10 Key Engineering Decisions and Bug Fixes

Three engineering decisions warrant explicit documentation because they affected the evaluation results or represent design choices with methodological implications.

The DOCX binary parse failure and fix is documented in Section 5.4. It is classified as an infrastructure finding rather than a model calibration finding: the LLM's behavior when receiving garbled binary content, capping the score at 2 and noting unreadable content, was technically correct given its inputs. The failure was in the client-side file processing pipeline, not in the scoring logic.

The chat history persistence fix addressed a separate issue identified during Scenario 1 testing: each new control card wiped the previous conversation content from the visible interface, preventing the assessor from reviewing prior scores during the session. This was corrected by implementing persistent

chat with control dividers and correct scroll behavior. It had no effect on scoring or on the data recorded in session state.

The scoring system prompt calibration adjustments documented in the Winter 2026 progress report addressed two edge cases. An internal red team report was scoring 3 when the expected score given CISO sign-off and remediation tracking was 3 to 4, so the system prompt rules were tightened to weight documented sign-off authority more heavily. A monitoring CSV with basic metrics was scoring 3 when the expected score given the absence of alert configuration was 2, so a scoring threshold note was added specifying that the presence of monitoring data without alert configuration should not exceed score 2. Both fixes held correctly during the formal evaluation runs.

Chapter 6. Evaluation Methodology

6.1 Evaluation Design Overview

The empirical evaluation of the framework applies the AI audit agent to three synthetic deployment scenarios representing graduated organizational maturity. The evaluation is designed to generate evidence for all four research questions simultaneously. RQ1 is evidenced by whether the tool produces differentiated, coherent scores across three scenarios with materially different compliance postures. RQ2 is evidenced by the adaptive control selection outcomes across three deployment types. RQ3 is evidenced by the proportion of controls where the agent produces coherent, evidence-grounded rationale without assessor correction and by the eight calibration findings documented in Section 7.6. RQ4 is evidenced by the comparison between AI-assigned scores and independent human assessor scores.

The evaluation protocol applies the tool in adaptive mode to all three scenarios, using the Anthropic document API for PDF evidence and Mammoth.js for DOCX evidence. Each scenario is run as a single uninterrupted session from the landing page through the compliance report. Evidence files are submitted one control at a time in the sequence specified in the scenario evidence pack inventory. No evidence files are resubmitted for controls where the assessor has already confirmed a score.

The three scenarios were evaluated sequentially rather than in parallel, with Scenario 1 completed first to allow calibration findings from that session to inform the assessment approach for Scenarios 2 and 3. Specifically, the DOCX binary parse failure identified during Scenario 1 was resolved before Scenarios 2 and 3 were run, ensuring that the file handling pipeline was fully operational for the higher-maturity scenarios where DOCX evidence is more prevalent.

6.2 The Three Deployment Scenarios

Three deployment scenarios were designed to represent distinct organizational maturity profiles and deployment contexts. The selection of three scenarios spanning public-facing, internal, and regulated financial contexts was deliberate: each represents a categorically different risk profile, evidence quality level, and regulatory environment, ensuring that the evaluation tests the framework's behavior across the full range of conditions it is designed to handle rather than within a narrow band of similar organizations.

Table 6.1. Three evaluation scenario overview.

Scenario	System	Organization	Deployment Type	Intended Maturity
S1	AskHR	Nexora Inc. (~4,200 employees)	Public-Facing HR Chatbot	LOW/UNEVEN (0-4): strong in some domains, 7 controls with no evidence
S2	HelpdeskAI	Vertex Global Corp (~12,000 employees)	Internal Knowledge Assistant	UNIFORM MODERATE (2-3): all documents exist but each has exactly one documented gap
S3	FinSight	Meridian Capital Bank (~8,000 employees)	Financial Advisory Tool	UNIFORMLY HIGH (4-5): Board-approved policies, external red team, automated controls

Scenario 1, AskHR at Nexora Inc., represents a public-facing HR chatbot at a mid-sized organization that has made uneven governance investments. This scenario was chosen because public-facing chatbots are among the most common LLM deployment types and carry significant user-facing risk, yet mid-sized organizations without dedicated AI governance teams frequently exhibit exactly this uneven pattern of strong controls in some areas and complete absence in others. Seven controls have no evidence at all, reflecting areas where the organization has not yet begun governance work. The remaining controls span a range from partial informal implementation to formal documentation, with particular strength in risk and supply chain controls where the procurement process required third-party attestations. This scenario tests the tool's ability to correctly score a mix of absent, verbal, and documented evidence within a single session.

Scenario 2, HelpdeskAI at Vertex Global Corp, represents an internal knowledge assistant at a large enterprise where governance documentation exists for every control but each document contains exactly one identifiable gap. This scenario was chosen to stress-test the 2-versus-3 maturity boundary, which is the most consequential threshold in the scoring model: it is the boundary between informal partial implementation and formal, consistently applied controls. An organization that has documentation for every control but with systematic gaps represents a realistic enterprise maturity profile and a genuine test of the LLM's ability to identify specific deficiencies rather than rewarding the presence of documentation alone.

Scenario 3, FinSight at Meridian Capital Bank, represents a financial advisory LLM subject to regulatory oversight from financial services authorities. This scenario was chosen because financial

services deployments face the most stringent governance expectations of any common LLM deployment type and therefore provide the best test of the tool's ceiling behavior: whether it correctly withholds score 5 for evidence that is high-quality but point-in-time, and whether it correctly assigns score 4 to evidence that satisfies all formal documentation and quantitative metric requirements but lacks the longitudinal trend data required for the highest maturity level. The organization has Board-approved policies, external red team reports, automated monitoring dashboards, and quantitative metrics for most controls.

6.3 Evidence Pack Design Principles

Synthetic evidence was used for three reasons documented in Section 3.5. Beyond those reasons, several additional design principles guided the construction of each evidence pack.

Realism of format: evidence files use the formats that real organizational compliance artifacts take. Policy documents are in DOCX format with named signatories and approval dates, reflecting how governance documents appear in practice. Risk registers are in DOCX with structured tables mapping risk categories, likelihood, impact, and owner. Monitoring data is in CSV format with timestamped metric columns, reflecting how dashboard exports and log data are typically provided during audits. Architecture summaries are in TXT format. Scenario overview files, which document the intended maturity profile and expected score range for each control, are in PDF format. This format diversity exercises all major file processing pathways in the audit agent and ensures that the evaluation tests the full evidence pipeline rather than a subset of supported formats.

Intentional gaps: every scenario contains at least one control where evidence is absent or mismatched. In Scenario 1, seven controls have no evidence submitted at all, reflecting the intended low-maturity posture of the organization. In Scenario 2, Control 10 (Privacy and Confidentiality, NIST MEA 2.10) is intentionally submitted with a wrong-type evidence file to test the tool's evidence relevance discrimination capability, specifically whether it correctly identifies a document type mismatch and applies partial credit rather than binary zero scoring. These intentional gaps are documented in the scenario overview files so that the expected behavior, a zero score for absent evidence and a partial credit score for wrong-type evidence, can be verified against the tool's actual output.

Independence from tool behavior: all expected score ranges were specified in the scenario overview files before the evaluation runs began. The tool's actual scores are compared against these pre-specified ranges rather than against post-hoc assessor judgment, preserving the independence of the evaluation design from the tool's behavior. This design choice is the primary mechanism for ensuring that

the evaluation cannot be unconsciously biased by knowledge of how the tool scored during the assessment.

6.4 The Scoring Protocol

Each scenario is assessed in a single session. For each control, the assessor reads the control card, types a text response describing the organization's posture, and attaches any relevant evidence file from the evidence pack. The text response for controls with no evidence uses a standardized verbal claim format: a bare yes or no for controls where no documentation exists, or a brief description of what stage the implementation is at for controls where partial work has been done. For controls with uploaded evidence, the text response briefly summarizes what the file contains and what compliance-relevant content it demonstrates for the specific control being assessed.

After the LLM assigns a score, the assessor reviews the rationale and either confirms the score or selects an override value. All override decisions are documented in the evaluation log with a reason code. Infrastructure-related overrides, meaning overrides applied because a file was not processed correctly rather than because the LLM's judgment was wrong, are tagged separately from judgment-based overrides to preserve the integrity of the RQ4 calibration analysis. The session proceeds through all active controls in PRIORITY_MAP order, from High priority to Medium to Low, ensuring that the highest-risk controls for the deployment type are assessed first.

6.5 Human Assessor Design

Two independent human assessors provide the ground-truth baseline for the RQ4 analysis. The first assessor is a graduate peer with prior experience conducting security audits in academic research contexts. The second is an industry professional with hands-on compliance auditing experience in organizational settings. Using assessors with different backgrounds serves a specific analytical purpose: it allows the study to detect whether score divergence from the AI tool is systematic across all assessors or concentrated in assessor-specific judgment patterns that reflect individual training and experience rather than tool calibration errors.

Before comparing assessor judgments against the AI tool's scores, the two assessors' classifications are compared against each other to establish the reliability of the human judgment baseline. High inter-assessor agreement on the same evidence confirms that the scoring criteria and maturity scale definitions are sufficiently precise to support consistent human judgment, which is a methodological prerequisite for using human scores as a validity benchmark for the tool. Agreement is reported as percentage agreement across all scored controls, with disagreements by maturity boundary documented

separately to identify whether inter-assessor disagreement concentrates at the same 2-versus-3 and 4-versus-5 thresholds that are expected to challenge the AI tool.

Both assessors are provided with access to all evidence pack files, the scenario overview with expected score ranges, the control definitions, and the scoring criteria. They are explicitly not provided with the AI agent's scores prior to completing their own independent evaluations, ensuring that human judgments are formed without anchoring to the tool's output. Each assessor independently scores all 30 controls per scenario using the same 0-to-5 maturity scale, with access to the scoring criteria documented in Chapter 4.

Controls where evidence submission failures occurred during the AI session are flagged in the evaluation log and analyzed as a separate category in the RQ4 comparison. For these controls, the human assessors' scores reflect what the correct score should be given the evidence in the pack, while the AI's score reflects what it assigned based on the incomplete or incorrectly processed submission. This distinction is methodologically important: the AI's lower score at these controls reflects the correct behavior of an evidence-based audit instrument, not a calibration error, and conflating them with genuine calibration divergences would systematically overstate the tool's disagreement with human judgment.

6.6 Validation Metrics

The validation metrics are organized by research question. Each metric is stated with its success criterion and the analytical implication of failure to meet it.

For RQ1, the primary metric is whether the tool produces correctly stratified overall compliance scores across the three scenarios. The success criteria are: Scenario 1 below 30 percent, Scenario 2 between 30 and 50 percent, and Scenario 3 above 60 percent. These thresholds reflect the intended maturity design of the three scenarios and are the minimum acceptable stratification for the tool to demonstrate useful discrimination across the low, moderate, and high maturity range. Failure to meet these thresholds would indicate a scoring calibration failure, not a framework design failure, since the thresholds are derivable from the scenario design independently of the tool's behavior. Secondary RQ1 metrics include whether domain scores reflect the intended maturity profile within each scenario, specifically whether domains designed to have stronger evidence (such as Risk and Supply Chain in Scenario 1) score higher than domains designed to have weaker evidence, and whether the tool assigns score 0 for all controls with no submitted evidence.

For RQ2, the primary metrics are the accuracy of N/A exclusions for each deployment type, specifically whether excluded controls are genuinely not applicable to that deployment context, the

appropriateness of High versus Medium versus Low priority assignments relative to the documented risk profile of each deployment type, and the correctness of the RAG override behavior when the RAG toggle is set to Yes. Success for RQ2 is defined as zero incorrect N/A exclusions, meaning no control is marked N/A that should be assessed for the deployment type, and correct reinstatement of RAG-specific controls when the RAG architecture toggle is activated.

For RQ3, the primary metric is the proportion of controls across all three scenarios where the agent produces a coherent, evidence-grounded rationale without requiring assessor correction of the scoring logic itself. Infrastructure-related corrections such as those caused by the DOCX binary failure are excluded from this count. Secondary RQ3 metrics include the verbal specificity gradient consistency across all verbal claim responses, the evidence relevance discrimination capability demonstrated by the wrong-type evidence submission in Scenario 2 Control 10, and file format reliability across all 15 supported formats.

For RQ4, the primary metric is the divergence rate between AI-assigned scores and human assessor scores, analyzed by domain, evidence type, and maturity level. Secondary metrics include the direction of divergence, specifically whether the AI over-scores or under-scores relative to human judgment and whether this directional pattern is consistent across assessors, and whether divergence concentrates at specific maturity boundaries, particularly the 2-versus-3 and 4-versus-5 thresholds identified as the most analytically challenging in the scoring model design.

6.7 Threats to Validity

The threats to validity for the evaluation methodology are documented in Section 3.7. Three additional evaluation-specific threats are noted here.

Single-session design: each scenario is run as a single uninterrupted session by a single assessor. With a single session, it is not possible to estimate the variance in scores that would result from different assessors making different evidence submission choices for the same scenario. The evaluation results reflect one specific set of evidence submission decisions and should be interpreted as a specific instantiation of the evaluation protocol rather than a distributional estimate of typical session behavior. The seven assessor behavior artifacts documented in Chapter 7, controls where evidence existed in the pack but was not submitted, illustrate the practical significance of this threat.

Synthetic evidence ceiling: the Scenario 3 evidence packs are point-in-time snapshots representing October 2025. They cannot by design satisfy the longitudinal evidence requirement for score 5, which requires multi-period trend data demonstrating continuous improvement across successive

cycles. This design choice means that the theoretical maximum achievable score in Scenario 3 under correct scoring is 4 for most controls, not 5. The evaluation's ability to assess whether the tool correctly assigns score 5 to genuinely score-5-quality evidence is therefore limited to future work with evidence packs designed to include multi-period longitudinal data. The 4-not-5 pattern documented in Chapter 7 must be interpreted with this ceiling in mind: it confirms correct threshold application rather than constituting evidence that the tool systematically withholds score 5.

Sequential evaluation and assessor learning: the three scenarios were run sequentially, with Scenario 1 first and the DOCX binary fix applied before Scenarios 2 and 3. This means the assessor's familiarity with the tool interface and the evidence submission protocol increased across sessions. Evidence submission quality in Scenarios 2 and 3 may therefore be higher than in Scenario 1 not only because those scenarios have stronger evidence packs but also because the assessor had more practice navigating the tool. This confound cannot be fully eliminated in a single-assessor design but is partially mitigated by the standardized verbal claim format specified in Section 6.4, which reduces variation in text response quality across sessions.

Chapter 7. Results

7.1 Results Overview

The full evaluation produced 90 control assessments across three deployment scenarios, covering all 30 controls per scenario. Two independent human assessors evaluated all three scenarios using the same 0-to-5 maturity scale, providing the ground-truth baseline for the RQ4 analysis. The overall results are summarized in Table 7.1.

Table 7.1. Cross-scenario results overview.

Scenario	LLM Score	LLM Avg /5.0	H1 Avg /5.0	H2 Avg /5.0	Consensus Avg /5.0	Operator Overrides	Scenario
S1 - AskHR	23%	0.80	0.97	1.00	0.98	2	S1 - AskHR
S2 - HelpdeskAI	33%	2.20	2.13	2.20	2.17	0	S2 - HelpdeskAI
S3 - FinSight	71%	4.03	3.87	3.93	3.90	0	S3 - FinSight

Note: LLM Score is the tool-reported overall compliance percentage. Averages are per-control means out of 5.0. Consensus is the mean of H1 and H2 per-control averages.

Two operator overrides were applied, both in Scenario 1 and both correcting DOCX binary failure artifacts, producing an operator override rate of 2.2 percent across 90 assessments. The three overall LLM scores of 23 percent, 33 percent, and 71 percent satisfy all three RQ1 stratification thresholds established in Section 6.6: Scenario 1 below 30 percent, Scenario 2 between 30 and 50 percent, and Scenario 3 above 60 percent. The human assessor consensus averages of 0.98, 2.17, and 3.90 track within 0.13 points of the LLM averages across all three scenarios, confirming that maturity stratification is consistent between the AI tool and human judgment.

The inter-assessor agreement rate between the two human assessors is 72.2 percent (65 of 90 controls), with only 2 large disagreements where the delta exceeds 1. This establishes that the scoring criteria are sufficiently precise to support consistent independent human judgment, which is a methodological prerequisite for using human scores as a validity benchmark. When the LLM score is compared against the consensus of the two human assessors, defined as agreement within 0.5 score points, the agreement

rate is 81.1 percent (73 of 90 controls). The LLM over-scored relative to human consensus in 8 controls and under-scored in 9, indicating no systematic directional bias across the full evaluation dataset.

7.2 Scenario 1 Results: AskHR / Nexora Inc.

Scenario 1 produced an overall LLM score of 23 percent, consistent with the intended low and uneven maturity profile. The final score after two DOCX-failure overrides was approximately 25 percent. The human assessor consensus average of 0.98 is close to the LLM average of 0.80, confirming broadly consistent scoring at the low-maturity range.

Table 7.2. Scenario 1 domain scores.

Domain	LLM Score	Final Score	H1 Avg	H2 Avg	Consensus	Target Range
1 - Governance and Accountability	0.8 / 5.0	0.8 / 5.0	0.2	0.2	0.20	0-1
2 - Data Integrity and Provenance	1.4 / 5.0	1.6 / 5.0	1.2	1.4	1.30	2-3
3 - Risk Assessment and Supply Chain	2.0 / 5.0	2.0 / 5.0	1.6	1.4	1.50	3-4
4 - Technical Reliability and Security	1.0 / 5.0	1.0 / 5.0	1.4	1.4	1.40	2-3
5 - Transparency and Documentation	0.8 / 5.0	1.2 / 5.0	0.8	1.0	0.90	3-4
6 - Monitoring and Continuous Improvement	0.8 / 5.0	0.8 / 5.0	0.6	0.6	0.60	1-2

Scenario 1 established three foundational calibration findings. First, the DOCX binary parse failure affected Controls 6, 14, and 21, producing artificially low scores that were corrected after the Mammoth.js fix was applied. The corrected scores reflected the genuine compliance content of the documents, which were substantially stronger than the garbled binary submission allowed the LLM to

assess. Second, three assessor behavior artifacts at Controls 17, 20, and 22 produced scores of 0 when relevant evidence existed in the pack but was not submitted during the session. The tool correctly scored based on submitted evidence only, which is the appropriate behavior for an evidence-based audit instrument. Third, the verbal specificity gradient was first observed consistently: an explicit denial produced score 0; a bare yes with no further description produced score 1; a yes with a named artifact type produced score 1; and a yes with both a named artifact type and an implementation stage description produced score 1 to 2. Seven intentionally absent controls all scored 0, consistent with the score 0 definition of no control existing and no evidence whatsoever.

The most significant human assessor divergence in Scenario 1 occurs at four controls where both assessors independently assigned the same score and diverged from the LLM. Both assessors scored Control 10 (Privacy and Confidentiality) at 3 versus the LLM's score of 1, and both scored Control 16 (Adversarial Testing) at 3 versus the LLM's score of 1. In both cases the submitted evidence was a verbal claim with no documentation. The LLM applied the strict verbal claim ceiling of score 1, while both assessors credited the organizational context described verbally as sufficient for score 3. Both assessors also scored Control 12 (Risk Likelihood and Impact) at 2 versus the LLM's score of 1, and Control 2 (Roles and Responsibilities) at 0 versus the LLM's score of 1, with the latter representing the one instance in Scenario 1 where both assessors were more conservative than the LLM. These four controls constitute the highest-confidence calibration findings for this scenario and are analyzed as part of Finding 4 in Section 7.6.

Two domains fell notably short of their target ranges. Domain 3 (Risk and Supply Chain) scored 2.0 against a target of 3 to 4, and Domain 5 (Transparency and Documentation) scored 0.8 against a target of 3 to 4. Both shortfalls are attributable to assessor behavior artifacts: the controls expected to anchor these domains at score 3 had evidence in the pack that was not submitted. Both human assessors also scored these domains below the target range, confirming the shortfall reflects genuine evidence submission gaps rather than LLM calibration errors.

7.3 Scenario 2 Results: HelpdeskAI / Vertex Global Corp

Scenario 2 produced an overall LLM score of 33 percent with zero operator overrides. The human assessor consensus average of 2.17 is nearly identical to the LLM average of 2.20, confirming strong calibration at the moderate maturity range.

Table 7.3. Scenario 2 domain scores.

Domain	LLM Score	Final Score	H1 Avg	H2 Avg	Consensus	Target Range
1 - Governance and Accountability	1.8 / 5.0	1.8 / 5.0	2.2	2.2	2.20	2-3
2 - Data Integrity and Provenance	2.2 / 5.0	2.2 / 5.0	2.0	2.2	2.10	2-3
3 - Risk Assessment and Supply Chain	2.6 / 5.0	2.6 / 5.0	2.6	2.4	2.50	2-3
4 - Technical Reliability and Security	2.4 / 5.0	2.4 / 5.0	2.2	2.4	2.30	2-3
5 - Transparency and Documentation	2.0 / 5.0	2.0 / 5.0	2.0	2.2	2.10	2-3
6 - Monitoring and Continuous Improvement	2.2 / 5.0	2.2 / 5.0	1.8	1.8	1.80	2-3

Scenario 2 was designed as a stress test of the 2-versus-3 maturity boundary. Every document in the evidence pack was designed to score 2 to 3: formally documented but with exactly one identifiable gap preventing score 3 in most cases. The verbal specificity gradient was confirmed across all verbal-claim controls without exception. Domain 1 (Governance) shows the largest gap between the LLM (1.8) and the human consensus (2.2): both assessors credited governance documentation that the LLM treated as incomplete. Domain 6 (Monitoring) shows the reverse, with both assessors (1.8) scoring lower than the LLM (2.2), where the LLM applied more generous partial credit to monitoring evidence than human assessors considered warranted.

Control 12 (Risk Likelihood and Impact) used the document D3_ConOps_and_Risk_Assessment.docx, which contained Board approval language, ServiceNow GRC platform integration, and OWASP-mapped quantitative risk classifications. The LLM scored it 4, above the pre-specified design target of 3. Both human assessors independently also scored it 4, confirming the LLM's judgment reflects genuine evidence quality rather than a calibration error. Control 11 (Context Establishment) used the same document and scored 3 from the LLM, matching both human assessors at 3, consistent with the design target. This demonstrates the LLM correctly differentiated between the two

controls using the same document, applying score 4 to Control 12 where quantitative risk metrics were the primary requirement and score 3 to Control 11 where the ConOps scope definition was the primary requirement.

Control 10 (Privacy and Confidentiality) was intentionally submitted with wrong-type evidence. The LLM scored 3, applying partial credit for privacy-relevant signals in the submitted governance document. Both human assessors scored it at 2, indicating the LLM applied more generous partial credit for wrong-type evidence than human assessors considered warranted. This is documented as Finding 6 in Section 7.6.

Eight controls in Scenario 2 produced scores where both assessors agreed with each other and diverged from the LLM, constituting the highest-confidence calibration findings for this scenario. The LLM under-scored on three of these controls: Control 6 (Data Governance), Control 13 (Improper Output Handling), and Control 21 (System Documentation), each scoring 2 from the LLM versus 3 from both assessors. The LLM over-scored on five: Control 10 (wrong-type evidence as noted above), Control 14 (Supplier Relationships), Control 18 (Secure Development), Control 26 (Continuous Monitoring), and Control 30 (Human Oversight), all scoring 3 from the LLM versus 2 from both assessors.

7.4 Scenario 3 Results: FinSight / Meridian Capital Bank

Scenario 3 produced an overall LLM score of 71 percent with zero operator overrides. The score reflects 29 controls scoring 4 and 1 control (Control 30, Human Oversight) scoring 5. The human assessor consensus average of 3.90 is slightly below the LLM average of 4.03, indicating the LLM was marginally more generous than the human assessors at high maturity overall.

Table 7.4. Scenario 3 domain scores.

Domain	LLM Score	Final Score	H1 Avg	H2 Avg	Consensus	Target Floor	LLM Gap
1 - Governance and Accountability	2.8 / 5.0	2.8 / 5.0	3.8	3.8	3.80	4-5	-1.2
2 - Data Integrity and Provenance	4.0 / 5.0	4.0 / 5.0	3.6	3.8	3.70	4-5	0
3 - Risk Assessment and Supply Chain	3.2 / 5.0	3.2 / 5.0	4.2	4.2	4.20	4-5	-0.8

Domain	LLM Score	Final Score	H1 Avg	H2 Avg	Consensus	Target Floor	LLM Gap
4 - Technical Reliability and Security	4.0 / 5.0	4.0 / 5.0	4.0	3.8	3.90	4-5	0
5 - Transparency and Documentation	4.0 / 5.0	4.0 / 5.0	3.6	3.8	3.70	4-5	0
6 - Monitoring and Continuous Improvement	3.4 / 5.0	3.4 / 5.0	4.0	4.2	4.10	4-5	-0.6

Domain 1 shows the largest gap from the target floor at 2.8 versus 4 to 5, driven by two intentional scenario gaps: Control 3 (System Prompt Confidentiality) scored 1 because test reports were referenced verbally but not uploaded, and Control 5 (Vector and Embedding Security) scored 1 because the implementation was described as in progress. Notably, both human assessors scored these controls at 3, higher than the LLM's score of 1, consistent with the verbal claim calibration gap identified in Scenario 1 where both assessors apply more generous credit to specific verbal claims than the LLM. Without Controls 3 and 5, the domain LLM average for the remaining three controls would be approximately 3.67.

Domain 3 (Risk Assessment and Supply Chain) is the only domain where the human consensus (4.20) exceeds the LLM score (3.2 reported, 4.0 for the three correctly assessed controls). This is primarily driven by Control 15 (Third-Party Dependencies), which the LLM scored 3 due to the format penalty while the assessors diverged: H1 scored 5 and H2 scored 3. The human assessors disagreed on this control, reflecting genuine boundary ambiguity about whether comprehensive textual content without the specified visual format merits score 3 or 5.

Domain 6 (Monitoring and Continuous Improvement) shows the human consensus (4.10) slightly above the LLM (4.2 for the domain, pulled by the LLM's score of 4 on Control 26). Both human assessors scored Control 26 (Continuous Monitoring) at 5, while the LLM scored it at 4, representing a confirmed under-scoring instance where both assessors independently judged the continuous monitoring evidence as meeting the score 5 threshold.

7.5 Cross-Scenario Domain Score Comparison

The domain score comparison across all three scenarios demonstrates that the framework correctly stratifies maturity at both the overall and domain level.

Table 7.5. Domain score comparison across all three scenarios: LLM scores and human assessor consensus.

Domain	S1 LLM	S1 Consensus	S2 LLM	S2 Consensus	S3 LLM	S3 Consensus
1 - Governance and Accountability	0.4	0.20	1.8	2.20	4.0	3.80
2 - Data Integrity and Provenance	0.8	1.30	2.2	2.10	4.0	3.70
3 - Risk Assessment and Supply Chain	1.0	1.50	2.6	2.50	4.0	4.20
4 - Technical Reliability and Security	1.0	1.40	2.4	2.30	4.0	3.90
5 - Transparency and Documentation	0.8	0.90	2.0	2.10	4.0	3.70
6 - Monitoring and Continuous Improvement	0.8	0.60	2.2	1.80	4.2	4.10

Note: S3 LLM domain scores reflect the full 30-control LLM averages including controls affected by intentional gaps and assessor artifacts. Consensus is the mean of H1 and H2 domain averages.

Four observations from this comparison are analytically significant.

First, in Scenario 1, Domain 1 (Governance) shows the LLM scoring higher than human consensus (0.4 versus 0.20). Both human assessors scored the Roles and Responsibilities control at 0 rather than the LLM's score of 1, and they also scored Domain 6 (Monitoring) lower than the LLM (consensus 0.60 versus LLM 0.8). At the very lowest maturity levels, the LLM was marginally more generous than human assessors in some domains.

Second, in Scenario 2, Domain 1 (Governance) shows the largest gap between the LLM and human consensus, with the LLM scoring 1.8 and human consensus at 2.2. This is consistent with the verbal claim calibration gap: human assessors credit governance verbal descriptions more generously than the LLM at the 1-to-2 boundary. Domain 6 (Monitoring) shows the reverse pattern, with human consensus (1.80) below the LLM (2.2), where the LLM applied more partial credit to monitoring evidence.

Third, in Scenario 3, Domain 3 (Risk Assessment and Supply Chain) is the only domain where the human consensus (4.20) exceeds the LLM (4.0), consistent with the format penalty finding at Control 15 where at least one human assessor scored the textual architecture narrative more generously than the LLM.

Fourth, Domain 6 (Monitoring) shows the most consistent pattern across both tool and assessors: the LLM and human consensus are closely aligned in Scenarios 2 and 3, with Domain 6 monitoring evidence producing the most consistent scoring behavior of any domain across the evaluation.

One anomaly is worth noting: Domain 2 (Data Integrity) scores 0.8 from the LLM in Scenario 1 but drops slightly to the human consensus of 1.30, then the LLM scores 2.2 in Scenario 2 while human consensus is 2.10, before rising to 4.0 (LLM) and 3.70 (consensus) in Scenario 3. The non-monotonic pattern between S1 and S2 reflects that the Scenario 1 data governance evidence, while weak overall, included two controls with partial upload credit that were slightly stronger than corresponding Scenario 2 documents at the 1-to-2 boundary.

7.6 Calibration Findings

Eight distinct calibration findings were identified across the three scenarios. The human assessor data either confirms, refines, or revises each finding's interpretation relative to what was possible from the LLM-only evaluation. The findings are organized by analytical significance for the research questions.

Finding 1: The 4-Not-5 Pattern Resolved (Primary RQ4 Finding)

In 29 of 30 Scenario 3 controls, the LLM assigned score 4, withholding score 5 and citing one of three consistent gap categories: absence of multi-period trend analysis, a referenced artifact not provided, and outcome-level validation absent. Prior to human assessor evaluation, it was analytically unclear whether this pattern reflected correct application of the longitudinal evidence threshold or conservative over-restraint.

The human assessor data resolves this question. The industry assessor assigned score 5 to 2 controls in Scenario 3 and score 3 to 6 controls. The academic assessor assigned score 5 to 6 controls and

score 3 to 8 controls. Both assessors were willing to assign score 5 to high-quality point-in-time evidence where the LLM withheld it, applying a more permissive interpretation of the longitudinal evidence standard. The 4-not-5 pattern therefore reflects a conservative calibration finding: the LLM applies the longitudinal evidence threshold more strictly than experienced compliance assessors do in practice.

The Scenario 3 score distributions illustrate this directly. The LLM distribution is 29 scores of 4 and 1 score of 5. The industry assessor distribution is 6 scores of 3, 22 scores of 4, and 2 scores of 5. The academic assessor distribution is 8 scores of 3, 16 scores of 4, and 6 scores of 5. Importantly, both human assessors also assigned score 3 to several controls where the LLM scored 4, meaning the LLM was not uniformly conservative: it was more generous than the industry assessor on 6 controls and more generous than the academic assessor on 8 controls in Scenario 3. This produces a net LLM average (4.03) slightly above the human consensus (3.90), confirming that the 4-not-5 pattern describes the LLM's ceiling behavior but does not characterize its overall calibration direction at high maturity.

Finding 2: DOCX Binary Parse Failures (Tool Infrastructure Finding)

Three controls in Scenario 1 were assessed before the Mammoth.js fix was applied. Raw DOCX binary was transmitted to the LLM, which correctly identified unreadable content and applied a maximum score of 2. Post-fix manual review confirmed all three documents were substantially stronger: Control 6 contained a CTO-approved data classification policy (corrected to 3), Control 14 contained a CISO-approved SOC 2 vendor review (corrected to 3), and Control 21 contained a comprehensive Model Card with quarterly review cycle (corrected to 4). This finding demonstrates that file format handling is a material variable in AI audit tool reliability independent of the LLM's analytical capabilities.

Finding 3: Assessor Behavior Artifacts (RQ4 Finding)

Across all three scenarios, seven controls were scored below true maturity because the assessor failed to submit evidence that existed in the evidence pack. In each case the tool's score was technically correct given submitted evidence. The most analytically interesting instance is Scenario 3 Control 29 (Excessive Agency and Tools), where the wrong evidence file was submitted entirely. The monitoring pack was uploaded for a control requiring a plugin registry, and the LLM correctly scored 1 while noting the strong incident response content and the complete absence of relevant evidence for the control objective. This pattern confirms a fundamental characteristic of evidence-based audit instruments: accuracy is bounded by assessor compliance with submission requirements.

Finding 4: Verbal Specificity Gradient Confirmed but Calibration Gap Identified (RQ3 and RQ4 Finding)

The LLM applied a consistent specificity scale to verbal claims across all three scenarios with zero operator corrections: explicit denial produces 0; bare yes with no description produces 1; yes plus named artifact produces 1; yes plus stage description produces 1 to 2; yes plus both produces 2; uploaded documents produce 2 to 4.

The human assessor data introduces a significant qualification. In Scenario 1, both human assessors independently scored Control 10 (Privacy and Confidentiality) at 3 and Control 16 (Adversarial Testing) at 3, while the LLM scored both at 1. The submitted evidence for both controls was a verbal claim with no documentation. Both assessors credited the specific organizational context described verbally as sufficient for score 3, exercising practitioner judgment about the plausibility of the claimed practices. This divergence, confirmed independently by both assessors, is the clearest calibration gap across the entire evaluation: the LLM's strict verbal claim ceiling of score 1 is more conservative than experienced compliance assessors apply in practice when the verbal description is specific and contextually plausible. Both assessors also independently scored Control 12 (Risk Likelihood and Impact) at 2 versus the LLM's score of 1, further confirming the gradient applies more strictly in the tool than in human judgment.

Finding 5: Over-Scoring Claim Revised by Human Assessor Data (RQ4 Finding)

The pre-evaluation expectation was that Control 12 (Risk Likelihood and Impact) in Scenario 2 would score 3, reflecting the design target of formal documentation with one gap. The LLM scored it 4. Both human assessors independently also scored it 4. This finding revises the original characterization: the pre-specified design target of 3 was too conservative for the evidence submitted. The LLM and both human assessors agreed the evidence met the score 4 threshold. Control 11 (Context Establishment) used the same document and scored 3 from the LLM, matching both human assessors at 3, consistent with the design target. The LLM correctly differentiated between the two controls using the same document based on which requirement was primarily being assessed.

Finding 6: Evidence Relevance Discrimination Partially Confirmed (RQ3 Finding)

In Scenario 2, Control 10 was intentionally submitted with wrong-type evidence. The LLM scored 3, applying partial credit for privacy-relevant signals in the submitted governance document. Both human assessors scored it at 2. The LLM was therefore more generous than both human assessors for wrong-type evidence, suggesting the partial credit mechanism slightly over-weights extractable signals from mismatched document types. Human assessors applied stricter credit when the document type did not match the control requirement. The finding confirms the tool does not zero-score wrong-type

evidence, which is a positive characteristic, but also indicates the partial credit calibration at the document-type mismatch boundary requires refinement.

Finding 7: File Format Reliability

PDF files submitted via the Anthropic native document API were 100 percent reliable across all scenarios. DOCX files processed via Mammoth.js after the fix were 100 percent reliable across Scenarios 2 and 3. CSV and TXT files were 100 percent reliable. DOCX files submitted as raw binary before the fix in Scenario 1 were 0 percent reliable. Multi-control document parsing demonstrated context-aware evidence extraction with no cross-contamination between controls across all three scenarios.

Finding 8: Format Penalty Produces Assessor Disagreement (Engineering Finding)

Scenario 3 Control 15 (Third-Party Dependencies) was submitted as a text-based architecture narrative rather than a visual diagram as specified in the expected evidence types. The LLM scored 3. The two human assessors disagreed: the industry assessor scored 5 and the academic assessor scored 3. This is analytically important because it shows that even among human assessors, the format penalty question produces split judgment. The LLM's score of 3 aligns with one assessor's judgment while diverging from the other's. The finding confirms that the expected evidence type specification creates ambiguity at the format boundary, and that the LLM's format penalty behavior is not clearly wrong relative to human judgment, only more conservative than one of the two assessors. The recommendation stands that expected evidence type descriptions should explicitly accommodate text-based alternatives where the informational content is comprehensive.

7.7 Discussion of Findings

The evaluation results, incorporating both LLM-generated assessments and independent human assessor scores, collectively support four primary findings that directly address the research questions.

Finding A (RQ1): The 30-control library successfully operationalizes compliance requirements from NIST AI RMF 1.0, ISO/IEC 42001:2023, ISO/IEC 27001:2022, and OWASP LLM Top 10 (2025) into structured, evidence-requesting audit logic. The tool correctly stratified maturity at 23 percent, 33 percent, and 71 percent, satisfying all three pre-specified stratification thresholds. The human assessor consensus averages of 0.98, 2.17, and 3.90 across the three scenarios confirm that the maturity stratification reflects genuine differences in organizational compliance posture rather than tool-specific behavior. Zero hallucinated standard references or fabricated control requirements were identified across all 90 assessments.

Finding B (RQ2): The adaptive control selection layer correctly identifies N/A controls per deployment type, applies the RAG override when the RAG toggle is set to Yes, and reorders controls by risk priority tier. The question adaptation feature produces deployment-specific questions that preserve the compliance intent and standard reference of the generic question while incorporating context-specific risk scenarios. The deterministic PRIORITY_MAP design ensures these outcomes are reproducible across sessions.

Finding C (RQ3): The tool generated coherent, evidence-grounded rationales for all 90 control assessments. Positive completeness indicators include the verbal specificity gradient consistently applied with zero operator corrections across all three scenarios, context-aware document parsing with no cross-contamination between controls, zero hallucinated standard references, and 100 percent post-fix file format reliability. The primary completeness limitation is assessor behavior dependency: seven controls across three scenarios were scored below true maturity due to evidence submission failures. The tool correctly scores based on submitted evidence only, which is the appropriate design for a compliance audit instrument.

Finding D (RQ4): The LLM-to-human assessor agreement rate, measured against the consensus of two independent assessors, is 81.1 percent (73 of 90 controls within 0.5 score points). The inter-assessor agreement rate of 72.2 percent between the two human assessors confirms that the human baseline is itself not perfectly consistent, and that the LLM's 81.1 percent agreement with their consensus represents strong calibration for a first-pass compliance assessment instrument. The LLM shows no systematic directional bias overall, over-scoring on 8 controls and under-scoring on 9 relative to human consensus. Two specific calibration gaps are identified with high confidence, confirmed independently by both assessors. First, at the verbal claim boundary, the LLM applies a strict ceiling of score 1 for verbal claims without documentation, while human assessors credit contextually specific verbal descriptions at score 3 when the organizational context is plausible. Second, at the score 4 to 5 boundary, the LLM applies a stricter longitudinal evidence standard than human assessors, who accept high-quality point-in-time evidence as sufficient for score 5 in cases where the LLM withheld that score. A third finding of note is a partial over-scoring tendency in the LLM at the wrong-type evidence boundary and at several moderate-maturity governance and monitoring controls, where the tool applied more generous partial credit than both assessors considered warranted. All three gaps represent specific, targeted opportunities for scoring model refinement rather than fundamental calibration failures.

Chapter 8. Discussion

8.1 Answering the Research Questions

The four research questions posed in Chapter 1 are answered by this thesis with different levels of confidence and completeness, and each is addressed honestly rather than with uniform strength claimed across all four.

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

RQ1 is answered with the highest confidence of the four. The 30-control library demonstrates a concrete, replicable methodology for translating governance standard requirements into structured, evidence-requesting audit questions with explicit standard citations. The operationalization process produced 30 controls covering all ten OWASP LLM Top 10 (2025) risk categories, including three composite controls constructed for risks that had no dedicated native subcategory in NIST AI RMF 1.0 or ISO/IEC 42001:2023. The empirical evaluation confirms that these controls produce differentiated, evidence-grounded assessments across three materially different deployment scenarios, satisfying all three pre-specified stratification thresholds. The composite control finding constitutes a research contribution in its own right: it demonstrates that NIST AI RMF 1.0 alone is insufficient to cover the current LLM threat landscape and that operationalization work of the kind performed in this thesis is a necessary step between publishing a governance standard and applying it to real LLM deployments.

RQ2: To what extent can a rule-based adaptive selection layer accurately filter and prioritize compliance controls based on LLM deployment type?

RQ2 is answered with strong design-level evidence and validated structural outcomes. The PRIORITY_MAP correctly identifies and excludes N/A controls per deployment type when RAG is not in use, and correctly reinstates those controls at Medium priority when the RAG toggle is activated. The priority tier assignments reflect documented risk profile differences across deployment types: public-facing chatbots prioritize user notification, output handling, and governance controls; internal knowledge assistants prioritize data governance and access control; financial advisory tools prioritize regulatory compliance, human oversight, and quantitative risk management. The question adaptation feature produces contextually specific questions while preserving the compliance intent and standard reference of the generic baseline, as confirmed by the before-and-after question comparisons documented in the RQ2 adaptive summary section of each compliance report. The honest qualification is that the priority tier assignments themselves have not been independently validated by domain experts who

specialize in compliance for each deployment category. The assignments reflect the researcher's judgment about relative risk priority, grounded in the source standards, and would benefit from external expert review as a direction for future work.

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

RQ3 is answered with strong empirical evidence from the calibration findings in Section 7.6. The tool generated coherent, evidence-grounded rationales for all 90 control assessments without manual correction of the scoring logic itself. The verbal specificity gradient was applied consistently across all three scenarios with zero operator exceptions. Evidence relevance discrimination was confirmed by the wrong-type evidence finding in Scenario 2, where the tool extracted partial credit proportional to the relevant content present rather than binary-scoring the submission as zero. Zero hallucinated standard references were identified across all 90 assessments. The human assessor comparison data further confirms that the tool's rationales were sufficiently well-grounded for both human assessors to produce independent scores that agreed with the tool at an 81.1 percent rate. The primary completeness limitation is assessor behavior dependency: seven controls across three scenarios were scored below true maturity due to evidence submission failures, which is a characteristic of evidence-based audit instruments generally and not specific to AI-powered assessment.

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

RQ4 is answered with a validated metric framework and complete empirical evidence from two independent human assessors. The evaluation establishes five primary metrics for RQ4 assessment: the operator override rate as a measure of session-level assessor-AI agreement; the domain-level score divergence as a measure of systematic directional bias; the boundary-specific divergence rate at the 2-to-3 and 4-to-5 thresholds as a measure of calibration precision; the assessor behavior artifact rate as a measure of the protocol dependency of AI assessment accuracy; and the format penalty rate as a measure of evidence type sensitivity.

The complete empirical results are as follows. The operator override rate of 2.2 percent, with both overrides attributable to infrastructure failure rather than LLM calibration errors, indicates strong session-level assessor-AI alignment. The inter-assessor agreement rate between the two human assessors is 72.2 percent (65 of 90 controls), establishing the reliability of the human baseline. The LLM-to-human consensus agreement rate is 81.1 percent (73 of 90 controls within 0.5 score points), which compares

favorably to the 72.2 percent inter-assessor baseline. Research on LLM-as-judge evaluation systems reports that strong LLM evaluators reach approximately 80 percent agreement with human evaluators, which is roughly the level of agreement humans reach with each other, indicating the tool's calibration is within the expected range for AI-assisted assessment instruments. The LLM shows no systematic directional bias overall, over-scoring on 8 controls and under-scoring on 9 relative to human consensus. Two specific calibration gaps are identified with high confidence, confirmed independently by both assessors: the strict verbal claim ceiling at score 1, which is more conservative than experienced compliance assessors apply for contextually specific verbal descriptions, and the conservative longitudinal evidence standard at the 4-to-5 boundary, where human assessors accept high-quality point-in-time evidence as sufficient for score 5 in cases where the LLM withholds that score.

8.2 Limitations

Six limitations of this research are significant enough to warrant explicit acknowledgment.

Limitation 1: Synthetic Evidence Environment. The evaluation was conducted against synthetic evidence designed within this research. Scoring behavior observed against designed policy documents and architecture files may not accurately predict behavior against real organizational evidence with its inherent complexity, ambiguity, and inconsistency. The evaluation results are framed as demonstrated capabilities within the designed test environment rather than population-level performance estimates. Validating the framework against real organizational evidence under authorized data-sharing agreements is the most important direction for future work.

Limitation 2: Single-Assessor Session Design. Each scenario was run as a single uninterrupted session by a single operator. With a single session, it is not possible to estimate the variance in scores that would result from different operators making different evidence submission choices for the same scenario. The assessor behavior artifact finding, where seven controls across three scenarios were scored below true maturity due to evidence submission failures, demonstrates that session outcomes are sensitive to operator navigation decisions. Multi-operator session designs would provide more robust estimates of tool reliability across a range of operator behaviors.

A related concern is selective evidence withholding: in a production setting, organizations may decline to upload directly relevant evidence, including incident reports, red team findings, vulnerability disclosures, or training data documentation, because submitting it to an external LLM API creates legal, regulatory, or competitive exposure. The seven assessor behavior artifacts in this evaluation, where evidence existed in the pack but was not submitted, model this dynamic even if unintentionally. A production deployment

should address this through a data handling disclosure at session start, guidance that allows the assessor to use the human override mechanism to reflect evidence they have reviewed but cannot upload, and as a longer-term direction, a local inference option for organizations with strict data residency requirements.

Limitation 3: Longitudinal Evidence Ceiling. The synthetic evidence packs are point-in-time documents. They cannot by design satisfy the score 5 requirement for longitudinal trend data and demonstrated improvement across multiple time periods. As a result, the LLM's score 5 threshold could only be assessed at one data point: Control 30 (Human Oversight) in Scenario 3 received score 5 from the LLM, and both the tool and one human assessor independently assigned that score, providing a single validation instance. The human assessor data confirms that the LLM's longitudinal standard for score 5 is more conservative than human assessors in practice, as documented in Finding 1 of Section 7.6. However, fully assessing whether the score 5 threshold is correctly calibrated requires evidence packs with multi-period trend data across multiple controls, which remains a gap for future evaluation work.

Limitation 4: Three Deployment Types. The PRIORITY_MAP currently covers three deployment types: Public-Facing Chatbot, Internal Knowledge Assistant, and Financial Advisory Tool. The LLM deployment landscape includes additional high-risk deployment contexts, including Healthcare Support Systems, Code Generation Assistants, and multi-agent agentic deployments, for which priority mappings have not been defined and validated. The adaptive selection feature's scope is currently limited to the three evaluated deployment types, and extending it to additional contexts requires both priority mapping design and empirical validation of the resulting control selections.

Limitation 5: Human Assessor Familiarity. Both human assessors were individuals known to the researcher, which introduces the potential for social desirability bias. Two design choices partially mitigate this risk. First, assessors followed a blind scoring protocol: they submitted all scores before seeing the AI-generated scores or each other's results. Second, the 0 to 5 maturity scale is anchored to concrete evidence criteria derived from NIST SP 800-53A, which constrains subjective interpretation and reduces the scope for generous scoring without evidentiary basis. Nevertheless, the possibility that assessor familiarity influenced scoring cannot be fully ruled out. Future evaluation should engage assessors with no prior relationship to the researcher, drawn from professional compliance auditing communities.

Limitation 6: Documentation Volume Bias. The scoring model rewards formally documented artifacts at Scores 2 to 3 and quantitatively tracked evidence at Scores 4 to 5. An organization that produces voluminous documentation that is procedurally correct in format but substantively insufficient, such as

policy documents that assert compliance without evidence of implementation or monitoring dashboards with metrics that do not reflect genuine operational controls, could receive scores that overstate actual maturity. The scoring system prompt partially mitigates this risk by requiring the LLM to identify specific gaps and by anchoring score thresholds to artifact type and content quality rather than artifact volume. The Scenario 2 evaluation stress-tested this boundary: every document was designed to contain exactly one identifiable gap, and the tool correctly scored these at 2 to 3 rather than crediting documentation presence alone. However, real organizational evidence may contain more subtle quality gaps that a scoring prompt calibrated on synthetic evidence does not reliably identify. Production use should include assessor guidance on this failure mode and should treat Score 2 to 3 controls with voluminous but vague policy documentation as candidates for deeper human review before a compliance finding is finalized.

8.3 Implications for Practice and Research

For practice, the primary implication of this research is that AI-powered compliance assessment is a viable and practically useful approach for organizations deploying LLMs. The 81.1 percent LLM-to-human consensus agreement rate, achieved against a human baseline that itself reaches 72.2 percent inter-assessor agreement, indicates that the tool produces compliance findings that are calibrated within the expected range for AI-assisted assessment instruments. The consistent verbal specificity gradient, applied without exception across 90 assessments, and the evidence relevance discrimination capability demonstrated in Scenario 2 further confirm that the tool produces findings that are evidence-grounded and not easily gamed by verbal claims alone. Organizations in the early stages of LLM governance can use the 30-control library as a structured starting point for understanding their compliance posture, even without deploying the audit agent itself.

For deployment beyond a research prototype, five control requirements would need to be addressed. First, data handling governance: organizations submitting real compliance evidence should have a data processing agreement with the LLM API provider, and the tool should display a clear disclosure at session start that submitted evidence is processed by an external API. Second, assessor access controls: the current CORS proxy architecture provides no authentication on the client side; a production version would require role-based access control restricting who can initiate and complete assessments. Third, session persistence and audit trail: the current in-browser session state is ephemeral; production use requires persistent storage of session records, override logs, and compliance reports to satisfy audit trail requirements. Fourth, scoring prompt versioning: the scoring system prompt directly determines assessment outputs, and a production deployment requires a formal versioning and change management process for it, so that score changes between versions can be identified, documented, and

disclosed to users. Fifth, validation against real evidence: the evaluation in this thesis used synthetic evidence only; production deployment should be preceded by a validation study using real organizational evidence under authorized data-sharing agreements, as identified in Section 9.2.

Two calibration gaps identified by the human assessor comparison have direct practical implications for tool deployment. The strict verbal claim ceiling means that organizations with strong governance practices but limited documentation will be systematically scored lower by the tool than by experienced compliance assessors. Practitioners using the tool should be aware that score 1 on verbal-claim controls may understate actual organizational maturity and should use the human override mechanism to apply practitioner judgment in those cases. The conservative longitudinal evidence standard means that high-maturity organizations operating a genuinely optimized control environment may receive score 4 from the tool even when score 5 would be appropriate, pending multi-period trend data. Future versions of the tool should incorporate a mechanism for explicitly submitting longitudinal data artifacts to unlock the score 5 threshold.

The composite control finding has direct practical implications for organizations using NIST AI RMF 1.0 as their primary governance standard. Organizations relying solely on AI RMF 1.0 without supplementary operationalization will have no structured assessment mechanism for System Prompt Leakage, Vector and Embedding Weaknesses, or Unbounded Consumption. NIST AI 600-1 addresses this gap at the guidance level, but practical operationalization remains the responsibility of individual organizations or framework implementers. This thesis provides one such operationalization, and the composite control construction methodology is replicable for future gaps as the LLM threat landscape continues to evolve.

For research, this thesis opens three specific directions. First, the framework and its empirical evaluation dataset provide a validated baseline for comparative evaluation of alternative compliance assessment approaches, including fully manual audits, LLM-assisted audits with different scoring architectures, and assessment using different underlying models. The 81.1 percent agreement rate and the two identified calibration gaps constitute specific, quantified targets against which alternative approaches can be benchmarked. Second, the two confirmed calibration gaps, the verbal claim ceiling and the longitudinal evidence standard at the 4-to-5 boundary, provide specific hypotheses for targeted scoring model refinement studies. A scoring model that incorporates few-shot examples of contextually specific verbal claims scored at 3 by human assessors, and examples of high-quality point-in-time evidence scored at 5, would directly address both gaps. Third, the PRIORITY_MAP architecture provides a replicable template for extending adaptive compliance assessment to additional deployment types and to emerging

LLM risk categories not yet covered in the OWASP LLM Top 10, as the threat landscape continues to evolve beyond the 2025 edition.

Chapter 9. Conclusion and Future Work

9.1 Summary of Contributions

This thesis set out to address a gap that sits at the intersection of two real and growing problems: the rapid deployment of LLM-integrated systems into organizational workflows, and the absence of practical, structured compliance assessment tooling that operationalizes existing AI governance standards into actionable audit criteria. The AI governance frameworks that exist, including NIST AI RMF 1.0 and ISO/IEC 42001, are intentionally nonprescriptive and use-case agnostic, leaving organizations to bridge the distance between high-level principles and the concrete, evidence-based audit checkpoints that compliance practice requires. This thesis bridges that distance through four equally significant contributions.

The 30-Control Compliance Library. The library operationalizes requirements from NIST AI RMF 1.0, ISO/IEC 42001:2023, ISO/IEC 27001:2022, and OWASP LLM Top 10 (2025) into structured, evidence-requesting audit questions organized across six risk domains. What distinguishes this library from prior compliance checklists is its dual grounding: every control traces to a named standard subcategory, and every OWASP LLM Top 10 (2025) risk category is covered by at least one control. The library is a research artifact in its own right, usable independently of the audit agent as a structured compliance baseline for organizations assessing their LLM deployments, even without an automated tool.

The Three Composite Controls and Their Justification. The identification of three OWASP risk categories with no dedicated native subcategory in NIST AI RMF 1.0 or ISO/IEC 42001:2023, and the construction of composite controls to address them, constitutes a finding about the current state of AI governance standards rather than simply a design decision about the framework. The finding is acknowledged by NIST AI 600-1 (2024), which extended the original AI RMF specifically to address the generative AI risks that the 2023 standard did not cover. Any organization relying on NIST AI RMF 1.0 alone should treat this finding as a prompt to supplement their compliance assessment with operationalization work of the kind documented in this thesis.

The Adaptive Control Selection Architecture. The PRIORITY_MAP provides a replicable template for deployment-aware compliance assessment that is deterministic, auditable, and extensible to additional deployment types. The design principle of rule-based control selection with LLM-powered question adaptation scoped to language generation only is a generalizable approach applicable to any compliance domain where both reproducibility and contextual specificity are required. This boundary between deterministic compliance logic and LLM-assisted language generation is itself a design contribution: it

ensures that the selection of which controls apply, and at what priority, remains auditable and reproducible across sessions, while still enabling the contextual specificity that makes each assessment relevant to the specific deployment being audited.

The Empirical Evaluation Dataset and Human Assessor Validation. The 90-control assessment dataset across three synthetic deployment scenarios, validated against two independent human assessors with different professional backgrounds, provides a quantified characterization of AI compliance tool behavior that researchers and practitioners can build on. The dataset produces eight calibration findings documented in Chapter 7, an 81.1 percent LLM-to-human consensus agreement rate measured against a human inter-assessor baseline of 72.2 percent, and two specific calibration gaps confirmed independently by both assessors: a strict verbal claim ceiling at score 1 that is more conservative than experienced practitioner judgment, and a conservative longitudinal evidence standard at the 4-to-5 boundary that withholds score 5 from high-quality point-in-time evidence that human assessors accept. Together these findings constitute a practical, quantified characterization of current AI compliance tool behavior that future research can use as a calibration baseline.

9.2 Future Work

The contributions of this thesis open several directions for future research and development. Four are identified as the most important.

Real Organizational Evidence Validation. The most important next validation step is to extend the framework to real organizational evidence under authorized data-sharing agreements. Real evidence introduces complexity, ambiguity, and incompleteness that synthetic evidence cannot replicate. Validating the framework's scoring behavior against real organizational policy documents, audit reports, and monitoring data would produce empirical findings that generalize beyond the synthetic evaluation environment and would identify calibration issues not visible in a designed evidence study. This validation would also allow the verbal claim calibration gap to be assessed in a real context where organizational verbal claims are accompanied by genuine organizational context that cannot be fully captured in synthetic scenarios.

Scoring Model Refinement Targeting Confirmed Calibration Gaps. The two calibration gaps identified by the human assessor comparison provide specific, targeted hypotheses for scoring model refinement. The strict verbal claim ceiling can be addressed by incorporating few-shot examples into the scoring system prompt that demonstrate contextually specific verbal claims scored at 3 by human assessors, specifically for cases where the organizational context is detailed and internally consistent. The

conservative longitudinal evidence standard can be addressed by adding a distinction in the score 5 criteria between evidence that explicitly documents multi-period trends and evidence that demonstrates comprehensive automation and continuous monitoring at a single point in time, with the latter qualifying for score 5 when the monitoring infrastructure is clearly self-sustaining. These refinements are directly motivated by the empirical findings and do not require redesigning the framework, only targeted iteration on the scoring prompt.

Expanded Deployment Type Coverage. The PRIORITY_MAP currently covers three deployment types. Extending it to cover Healthcare Support Systems, Code Generation Assistants, Customer Support Agents, and multi-agent agentic deployments would broaden the framework's applicability to the full range of LLM deployment contexts. Each extension requires a principled risk-profile analysis to determine the appropriate priority tier for each of the 30 controls in that deployment context, informed by the regulatory and operational characteristics of each deployment category. Healthcare deployments, for example, would require elevated priority for human oversight, validity and reliability, and incident communication controls relative to the current three deployment types.

Extending the PRIORITY_MAP to a new deployment type follows a three-step methodology derivable from the existing entries. First, the regulatory and operational risk profile of the new deployment type is characterized using the source standards and any domain-specific frameworks relevant to that context, for example HIPAA Security Rule considerations for Healthcare Support Systems or secure development lifecycle requirements for Code Generation Assistants. Second, each of the 30 controls is assigned a priority tier based on whether the risk it addresses is elevated, standard, low-salience, or structurally inapplicable for that deployment context. Healthcare deployments would elevate human oversight, validity and reliability, and incident communication controls; Code Generation Assistants would elevate insecure output handling and supply chain controls. Third, the new deployment type entry is added to the PRIORITY_MAP lookup table and validated against at least one representative synthetic scenario before inclusion. No changes to the control library, the scoring model, or the LLM scoring prompt are required; the PRIORITY_MAP is the only artifact that needs to be updated for deployment type extensions.

CI/CD Integration for Continuous Assessment. The current audit agent is designed for point-in-time assessment. Integrating the framework into a CI/CD pipeline would allow organizations to run a defined subset of controls automatically whenever a new model version, fine-tuning update, or system prompt change is deployed, catching compliance regressions before they reach production. This integration pattern would transform the framework from a periodic audit instrument into a continuous governance mechanism, which is the appropriate model for LLM deployments that are updated frequently and for

which continuous assurance of compliance posture is increasingly expected by regulators and customers. A CI/CD integration would also address the longitudinal evidence ceiling identified as a limitation in Chapter 8: repeated automated assessments over time would naturally produce the multi-period trend data that the score 5 threshold currently requires but that point-in-time synthetic evidence packs cannot provide.

Regulatory Extensibility. Adding controls in response to new or updated regulations requires changes to three components, in descending order of effort. The control library is the primary artifact to update: a new control requires a control ID, a control name, an LLM-specific objective, expected evidence types, an OWASP mapping, and a domain assignment. For regulations introducing entirely new risk categories not addressed by existing OWASP mappings, as NIST AI 600-1 introduced supplementary guidance for generative AI risks absent from AI RMF 1.0, the composite control construction methodology documented in Section 4.5 provides a replicable approach. For regulations adding requirements within existing OWASP-mapped risk categories, a new control can be inserted into an existing domain without restructuring the library. The PRIORITY_MAP then requires a new entry per deployment type for the added control, which is low-effort given the table's structure. The scoring system prompt does not need to change unless the new control introduces a materially different evidence type hierarchy. The overall extensibility posture is therefore low-effort for incremental additions within existing risk categories and moderate-effort for structural expansions that introduce new OWASP-unmapped risks. The modular separation of control library, selection layer, and scoring engine was designed specifically to allow regulatory updates without requiring full revalidation of the tool's calibration behavior.

9.3 Closing Remarks

The governance community has spent decades building mature frameworks for assessing the security and compliance of deterministic software systems. Auditors, compliance professionals, and standard-setting bodies have collectively produced an assessment ecosystem where the gap between the publication of a governance standard and the availability of practical compliance tooling is measured in years rather than decades. AI governance does not yet have that ecosystem. The NIST AI RMF and ISO/IEC 42001 have existed long enough to be taken seriously as governance standards, but the operationalized, tooling-supported compliance infrastructure needed to apply them consistently to real LLM deployments has not kept pace.

This thesis contributes to closing that gap in four specific ways: a 30-control library that operationalizes these standards into structured audit criteria with explicit standard citations; three composite controls that address specific gaps in those standards for the current LLM threat landscape; an

adaptive selection architecture that makes assessments context-aware without sacrificing reproducibility; and an empirical evaluation, validated by two independent human assessors, that characterizes how AI-powered compliance assessment behaves across the full range of organizational maturity and identifies the specific calibration boundaries where human judgment and AI scoring diverge. Together these contributions represent a step toward the kind of practical, standards-aligned AI compliance infrastructure that the pace of LLM deployment into consequential organizational systems makes increasingly necessary to build.

Bibliography

[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, doi: 10.56726/IRJMETS61241.

[2] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, Jan. 2023. [Online]. Available: https://doi.org/10.6028/NIST.AI.100-1

[3] International Organization for Standardization, "ISO/IEC 42001:2023 - Artificial Intelligence Management System," ISO/IEC, Geneva, 2023.

[4] 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.

[5] National Institute of Standards and Technology, "NIST AI RMF Playbook," 2023. [Online]. Available: https://airc.nist.gov/Docs/2

[6] 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.

[7] Organisation for Economic Co-operation and Development, "OECD Principles on Artificial Intelligence," OECD/LEGAL/0449, 2019, updated 2024. [Online]. Available: https://oecd.ai/en/ai-principles

[8] H. P. Kothandapani, "AI-Driven Regulatory Compliance: Transforming Financial Oversight through Large Language Models and Automation," Emerging Science Research, vol. 3, Jan. 2025.

[9] A. Mohammed, "AI in Cybersecurity: Enhancing Audits and Compliance Automation," Innovative Computer Science Journal, vol. 7, no. 1, pp. 1-4, 2023.

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

[11] OWASP Foundation, "OWASP Top 10 for Large Language Model Applications," v2025, 2025. [Online]. Available: https://owasp.org/www-project-top-10-for-large-language-model-applications/

[12] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1-38, Mar. 2023, doi: 10.1145/3571730.

[13] F. Perez and I. Ribeiro, "Ignore Previous Prompt: Attack Techniques for Language Models," NeurIPS ML Safety Workshop, 2022. [Online]. Available: https://arxiv.org/abs/2211.09527

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

[15] CMMI Institute, "CMMI V2.0 Model - Capability Maturity Model Integration," CMMI Institute, Pittsburgh, PA, 2018. [Online]. Available: https://cmmiinstitute.com

[16] International Organization for Standardization, "ISO/IEC 33001:2015 - Process Assessment Concepts and Terminology," ISO/IEC, Geneva, 2015.

[17] National Institute of Standards and Technology, "Assessing Security and Privacy Controls in Information Systems and Organizations," NIST SP 800-53A Rev. 5, Jan. 2022, doi: 10.6028/NIST.SP.800-53Ar5.

[18] National Institute of Standards and Technology, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," NIST SP 1270, Mar. 2022. [Online]. Available: https://doi.org/10.6028/NIST.SP.1270

[19] International Organization for Standardization, "ISO/IEC 27001:2022 - Information Security, Cybersecurity and Privacy Protection," ISO/IEC, Geneva, 2022.

[20] M. Weidinger et al., "Taxonomy of Risks posed by Language Models," in Proc. ACM Conf. on Fairness, Accountability, and Transparency (FAccT), ACM, 2022, pp. 214-229, doi: 10.1145/3531146.3533088.

[21] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," NIST AI 600-1, Jul. 2024, doi: 10.6028/NIST.AI.600-1.

[22] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," Journal of Management Information Systems, vol. 24, no. 3, pp. 45-77, 2007, doi: 10.2753/MIS0742-1222240302.

[23] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, "Experimentation in Software Engineering," Springer, 2012.

Appendix A. The 30-Control Library (Full Table)

Table A.1 presents all 30 controls with their domain, control ID, control name, LLM-specific objective, expected evidence types, and OWASP mapping. Controls marked with an asterisk (*) in the # column are composite controls combining two standard subcategories around a new LLM-specific control objective. The rationale for Controls 3, 4, and 5 is documented in Section 4.5.

Table A.1. Complete 30-control compliance library (corrected against tool code).

#	Control ID	Control Name	Control Objective (LLM-Specific)	Evidence Types	OWASP
1	NIST GOV 1.2	Policies and Procedures	Documented acceptable use policy covering LLM deployment scope, prohibited uses, and approval workflow	AUP, governance policy, sign-off record	LLM01, LLM06
2	NIST AI RMF GOV 2.1	Roles and Responsibilities	Named AI Model Owner or AI Risk Officer with documented RACI authority over LLM deployment decisions	RACI chart, job descriptions, org chart	LLM06
3 *	NIST AI RMF MEA 2.7 + GOV 4.3	System Prompt Confidentiality	System prompt classified as confidential; extraction attack tests conducted; system never echoes prompt in outputs	Test reports, prompt classification policy, red team findings	LLM07
4 *	NIST AI RMF MAN 1.3 + GOV 1.3	Resource and Cost Governance	Per-user rate limits, token quotas, cost alerts, and timeouts on all LLM inference calls	Rate limit config, cost monitoring dashboard, alert settings	LLM10
5 *	NIST AI RMF MAP 4.2 + MEA 2.7	Vector and Embedding Security	Access controls on vector stores, document validation before embedding, anomalous retrieval monitoring	Access control matrix, ingestion validation logs, retrieval audit logs	LLM08
6	ISO 42001 A.7.1	Data Governance	Data Bill of Materials covering all training and fine-tuning data sources with classification and lineage	DBoM, data classification policy, lineage documentation	LLM02, LLM04
7	ISO 42001 A.7.2	Data Acquisition	Legal basis confirmed for all training data; data use rights documented; copyright compliance verified	DPA, data licensing agreements, acquisition policy	LLM03, LLM04
8	ISO 42001 A.7.3	Data Quality	Multi-dimension quality assessment with documented metrics; poisoning detection mechanism in place	QA report, quality metrics dashboard, anomaly detection logs	LLM04, LLM03
9	ISO 42001 A.7.4	Data Provenance	Chain of custody documentation for all datasets; version-controlled data registry with owner attribution	Data registry, version history, chain-of-custody log	LLM04, LLM03
10	NIST AI RMF MEA 2.10	Privacy and Confidentiality	PII detection and redaction before training and inference; Privacy Impact Assessment completed and approved	PIA, DLP tool configuration, PII scan reports	LLM02

#	Control ID	Control Name	Control Objective (LLM-Specific)	Evidence Types	OWASP
11	NIST MAP 1.1	Context Establishment	Documented ConOps with defined scope, permitted uses, prohibited uses, and user population	ConOps document, use case registry, user documentation	LLM06
12	NIST AI RMF MAP 5.1	Risk Likelihood and Impact	Structured risk register with likelihood and impact scores for each identified LLM-specific risk	Risk register, risk assessment methodology, review meeting records	LLM01, LLM09
13	ISO 42001 A.6.3 (Output Validation)	Improper Output Handling	Output validation and sanitization controls; harmful output detection mechanism; scope enforcement	Output validation policy, content filter configuration, test results	LLM05
14	ISO 42001 Cl. 8.4	Supplier Relationships	Third-party AI vendors assessed against security requirements; vendor risk register maintained	Vendor assessment records, SLAs, third-party audit reports	LLM03
15	NIST MAP 4.1	Third-Party Dependencies	Architecture diagram showing all third-party AI components, APIs, and data flows with risk assessment	Architecture diagram, component inventory, dependency risk log	LLM03, LLM08
16	NIST MEA 2.7	Adversarial Testing	Formal adversarial testing program with documented test cases covering prompt injection; results tracked over time	Red team report, test case library, injection test results, schedule	LLM01, LLM04
17	NIST AI RMF MEA 2.5 + MEA 2.3	Validity and Reliability	Regular evaluation against verified benchmark datasets; hallucination rate measured and tracked	Evaluation reports, benchmark dataset, accuracy metrics, trend data	LLM09
18	ISO 42001 A.6.3	Secure Development	Security testing integrated into LLM deployment pipeline; SAST/DAST results documented	SAST results, DAST results, secure coding guidelines, scan history	LLM01, LLM05
19	ISO 42001 A.6.4	Verification and Validation	Formal V&V process with multi-stakeholder sign-off before production deployment	V&V report, sign-off records, test results, deployment checklist	LLM05, LLM06
20	ISO 27001 A.5.15	Access Control	RBAC enforced on LLM APIs, model weights, and fine-tuning infrastructure; MFA and least privilege applied	IAM policy, access control matrix, MFA configuration, access logs	LLM02, LLM07
21	ISO 42001 A.8.2	System Documentation	Model card or system card documenting capabilities, limitations, intended use, and known failure modes	Model card, system card, capability documentation, update history	LLM09
22	ISO 42001 A.8.1	User Notification	Users informed they are interacting with an AI system;	UI screenshot, disclosure policy,	LLM09

#	Control ID	Control Name	Control Objective (LLM-Specific)	Evidence Types	OWASP
			AI disclosure persistent in interface	user notification records	
23	NIST AI RMF MEA 2.9	Explainability	Explanation mechanism for model outputs; confidence scores or citation coverage tracked and audited	Explainability audit report, confidence score distribution, citation coverage metrics	LLM09, LLM08
24	ISO 42001 A.8.3	Incident Communication	Incident communication plan with stakeholder notification procedures and legally reviewed templates	Communication plan, notification templates, tabletop exercise record	LLM02
25	NIST AI RMF MAP 3.5 + GOV 1.3	Excessive Agency - Permission	Explicit allowlist of permitted LLM actions; violation monitoring in place; no autonomous financial or legal transactions	Permission allowlist, violation log, HITL enforcement documentation	LLM06
26	NIST MAN 4.1	Continuous Monitoring	Real-time monitoring dashboard covering performance, security, and fairness metrics; alerting configured	Monitoring dashboard, alert configuration, metric trend data, incident log	LLM09, LLM04
27	NIST AI RMF MAN 4.3 + MAN 2.3	Incident Response	AI-specific incident response plan with tested runbooks; killswitch capability documented	IRP, runbook, tabletop exercise record, killswitch documentation	LLM01, LLM09
28	ISO 42001 A.6.6	Re-evaluation and Drift Detection	Automated performance threshold monitoring; drift detection triggers re-evaluation with documented SLA	Drift detection config, re-evaluation reports, SLA documentation, threshold logs	LLM09, LLM04
29	NIST AI RMF GOV 6.1 + MAP 4.1	Excessive Agency - Tools	Registry of all tools and APIs accessible to the LLM; agentic action audit log maintained	Tool registry, permission allowlist, agentic action audit log	LLM06
30	NIST AI RMF MAP 3.5 + MAN 2.4	Human Oversight	HITL mechanism technically enforced for high-stakes decisions; override log maintained and reviewed	HITL configuration, override log, review records, audit trail	LLM06

Note: The asterisk (*) in the # column denotes composite controls. Control ID prefixes reflect the exact identifiers used in the audit agent tool code. OWASP mappings reflect primary and secondary coverage as coded in the tool.

Appendix B. Scoring System Prompt (Verbatim)

The following is the exact system prompt sent to Claude Haiku for every control assessment in standard mode. In adaptive mode, two additional lines are injected into the session context block: 'Audit mode: Adaptive' and 'Control priority for this deployment: [HIGH/MEDIUM/LOW]'.

```
You are an expert AI compliance auditor assessing an organization's LLM
deployment
against a structured 30-control compliance framework based on NIST AI RMF 1.0,
ISO/IEC 42001:2023, and ISO/IEC 27001:2022, aligned with OWASP LLM Top 10
(2025).

Session context:
  Project: [session.project]
  Deployment type: [session.deployType]

You are assessing Control [N] of [total]:
  Control ID: [ctrl.id]
  Control Name: [ctrl.name]
  Control Objective: [ctrl.objective]
  OWASP Mapping: [ctrl.owasp]
  Expected Evidence Types: [ctrl.evidence]
  Audit question asked: "[ctrl.question]"

Score using this 0-5 maturity scale grounded in ISO/IEC 33001, CMMI, and NIST
SP 800-53A:
  0 = No control exists, no evidence whatsoever
  1 = Verbal claim only, no documented evidence
  2 = Partial/informal implementation, one-time effort, no repeatable process
  3 = Formal documentation exists, consistently applied, evidence provided
  4 = Quantitative metrics, tracked over time, linked to remediation processes
  5 = Fully automated, continuous improvement, trend analysis across versions

Respond ONLY with valid JSON in exactly this format, nothing else:
{
  "score": <integer 0-5>,
  "rationale": "<2-3 sentences citing specific aspects of the response and
evidence quality>",
  "strength": "<one specific strength or 'None identified' if score <= 1>",
  "gap": "<one specific gap or missing evidence item>",
  "recommendation": "<one specific, actionable improvement recommendation>"
}
```

Appendix C. Scenario Evidence Pack Inventory

Table C.1 lists all synthetic evidence files used in the evaluation, organized by scenario. Each row identifies the file name, format, and the primary controls for which it was submitted.

Table C.1. Evidence pack inventory by scenario.

Scenario	File Name	Format	Controls Submitted For
S1 - AskHR	Nexora_Scenario_Overview.pdf	PDF	All controls (scenario context)
S1 - AskHR	AskHR_Architecture.txt	TXT	C15 - Third-Party Dependencies
S1 - AskHR	AskHR_Monitoring_Dashboard.csv	CSV	C26 - Continuous Monitoring
S2 - HelpdeskAI	Vertex_Scenario_Overview.pdf	PDF	All controls (scenario context)
S2 - HelpdeskAI	D1_Governance_Pack.docx	DOCX	C1, C2, C4 - Governance controls
S2 - HelpdeskAI	D2_Data_Governance_Pack.docx	DOCX	C6, C7, C8, C9, C10 - Data controls
S2 - HelpdeskAI	D3_ConOps_and_Risk_Assessment.docx	DOCX	C11, C12 - Risk controls
S2 - HelpdeskAI	D4_Technical_Security_Validation.docx	DOCX	C16, C18, C19 - Technical controls
S2 - HelpdeskAI	D5_Transparency_Pack.docx	DOCX	C21, C22, C23 - Transparency controls
S2 - HelpdeskAI	D6_Monitoring_Pack.docx	DOCX	C26, C27, C28 - Monitoring controls
S3 - FinSight	Meridian_Scenario_Overview.pdf	PDF	All controls (scenario context)
S3 - FinSight	D1_Governance_AUP_RACI.docx	DOCX	C1, C2, C4 - Governance controls
S3 - FinSight	D2_Data_Management_Complete.docx	DOCX	C6, C7, C8, C9, C10 - Data controls
S3 - FinSight	D3_ConOps_and_Risk_Register.docx	DOCX	C11, C12, C14 - Risk controls
S3 - FinSight	D4_Security_Validation.docx	DOCX	C16, C17, C18, C19, C20 - Technical controls
S3 - FinSight	D5_Model_Card_and_Documentation.docx	DOCX	C21, C22, C23, C24, C25 - Transparency controls
S3 - FinSight	D6_Monitoring_Dashboard.docx	DOCX	C26, C27, C28, C30 - Monitoring controls
S3 - FinSight	D6_Monitoring_Metrics.csv	CSV	C26 - Continuous Monitoring (metrics data)

Appendix D. Full Control-Level Score Log (Scenario 3 Representative)

Table D.1 presents the complete control-level score log for Scenario 3 (FinSight / Meridian Capital Bank), the highest-maturity scenario, with LLM score and key finding per control. Controls marked with a dagger (†) are assessor behavior artifacts where evidence existed in the pack but was not submitted. The full 90-control dataset including Scenarios 1 and 2 is available in the evaluation session records.

Corrections applied: Control 2 standard ID corrected from NIST GOV 2.2 to NIST AI RMF GOV 2.1 to match tool code. Control 30 LLM score corrected from 4 to 5 to reflect the actual tool output (Human Oversight scored 5 as the one score-5 control in Scenario 3).

Table D.1. Scenario 3 (FinSight) complete control-level score log.

#	Control Name	Standard ID	LLM Score	Key Finding
1	Policies and Procedures	NIST GOV 1.2	4	Board-approved AUP v2.1, CCO and GC sign-off. Gap: no AUP violation frequency metrics or trend analysis across policy versions.
2	Roles and Responsibilities	NIST AI RMF GOV 2.1	4	Named RACI, Board-appointed CAIRO, HR-registered JDs. Gap: no quantitative role performance KPIs or escalation response time tracking.
3†	System Prompt Confidentiality	NIST AI RMF MEA 2.7+GOV 4.3	1	Verbal claim only: test reports referenced but not uploaded. Assessor behavior artifact.
4	Resource and Cost Governance	NIST AI RMF MAN 1.3+GOV 1.3	4	Full rate limiting table (5 dimensions), Azure auto-suspend, AOD monitoring. Gap: no trend analysis or post-incident root-cause documentation.
5†	Vector and Embedding Security	NIST AI RMF MAP 4.2+MEA 2.7	1	In-progress implementation claim, no evidence. Intentional scenario gap despite Pinecone RAG architecture.
6	Data Governance	ISO 42001 A.7.1	4	Complete DBoM (4 sources), 100% classification compliance via automated tag-checking. Gap: no classification enforcement at query/inference time.
7	Data Acquisition	ISO 42001 A.7.2	4	Azure DPA and Bloomberg MSA with AI/ML rights confirmed. Gap: no independent third-party copyright audit.
8	Data Quality	ISO 42001 A.7.3	4	5-dimension quantitative QA (97.8% accuracy), continuous poisoning scan, remediation documented. Gap: no multi-cycle trend analysis.
9	Data Provenance	ISO 42001 A.7.4	4	Structured DBoM with version control, CDO-approved. Gap: no formal chain-of-custody log for dataset mutations.
10	Privacy and Confidentiality	NIST AI RMF MEA 2.10	4	DPO-approved PIA, Microsoft Presidio DLP integrated, Azure DPA. Gap: no PII detection effectiveness metrics.
11	Context Establishment	NIST MAP 1.1	4	Board-approved ConOps, 5 prohibited uses, technically enforced via HITL and DLP. Gap: no prohibition-violation attempt rate metrics.

#	Control Name	Standard ID	LLM Score	Key Finding
12	Risk Likelihood and Impact	NIST AI RMF MAP 5.1	4	Archer GRC 5x5 risk register, OWASP-mapped, monthly review cycle. Gap: no closed-loop remediation metrics.
13 †	Improper Output Handling	ISO 42001 A.6.3	1	Verbal claim only: reports from last quarter referenced but not uploaded. Assessor behavior artifact.
14	Supplier Relationships	ISO 42001 Cl. 8.4	4	All vendors SOC 2 Type II and ISO 27001, Archer GRC monthly review. Gap: no continuous automated monitoring between annual cycles.
15	Third-Party Dependencies	NIST MAP 4.1	3	Text-based architecture narrative submitted; no visual diagram. Format gap penalized despite comprehensive textual content.
16	Adversarial Testing	NIST MEA 2.7	4	CrowdStrike external red team (132 cases, LOW RISK) and GARAK weekly (847 cases, 99.6% pass rate). Gap: no recurring red team schedule documented.
17	Validity and Reliability	NIST AI RMF MEA 2.5+MEA 2.3	4	Monthly golden dataset eval (94.1% accuracy, ECE 0.041), 6-month trend, CI/CD automated. Gap: no external benchmark comparison.
18	Secure Development	ISO 42001 A.6.3	4	Semgrep SAST and ZAP DAST on every deployment, 0 High/Critical over 6 deployments. Gap: actual scan reports not provided.
19	Verification and Validation	ISO 42001 A.6.4	4	6-signatory V&V sign-off (CISO, CAIRO, MRM, DPO, CCO, Ethics Committee), SR 11-7 compliant. Gap: V&V document content not provided.
20	Access Control	ISO 27001 A.5.15	4	SSO, MFA, VPN Zero Trust, HSM Key Vault 90-day rotation, PIM just-in-time. Gap: no real-time access anomaly detection metrics.
21	System Documentation	ISO 42001 A.8.2	4	Model Card v2.1, 99.7% citation coverage, ECE 0.041, monthly explainability audits (98% pass). Gap: no Model Card version control tied to KB refresh cycles.
22	User Notification	ISO 42001 A.8.1	4	Persistent UI banner, 87/87 RMs certified, MiFID II and EU AI Act compliance reviewed. Gap: no user comprehension effectiveness metrics.
23	Explainability	NIST AI RMF MEA 2.9	4	99.7% citation coverage, calibrated confidence scores (ECE 0.041), monthly MRM audits. Gap: no evidence confidence scores correlate with RM decision quality.
24	Incident Communication	ISO 42001 A.8.3	4	AICP document, legal-reviewed templates (4 types), October 2025 tabletop completed. Gap: no lessons-learned documentation.
25	Excessive Agency - Permission	NIST AI RMF MAP 3.5+GOV 1.3	4	APP document with explicit allowlist (3 operations), zero violations since August 2025. Gap: no technical implementation details showing enforcement mechanism.
26	Continuous Monitoring	NIST MAN 4.1	4	AOD 9-metric dashboard, PagerDuty/Slack alerting, 30-day CSV data. Gap: no toxicity metric, reactive not predictive drift detection.
27	Incident Response	NIST AI RMF MAN 4.3+MAN 2.3	4	IRP document, tested killswitch, 4 AI-specific scenarios, October 2025 tabletop. Gap: no post-incident RCA, no playbook version history.
28	Re-evaluation and Drift Detection	ISO 42001 A.6.6	4	REP document with automated thresholds (accuracy below 90%), 48-hour SLA, 6-month stable trend. Gap: no retraining completion logs.

#	Control Name	Standard ID	LLM Score	Key Finding
29 †	Excessive Agency - Tools	NIST AI RMF GOV 6.1+MAP 4.1	1	Wrong evidence submitted: monitoring pack irrelevant to plugin registry requirement. Assessor behavior artifact.
30	Human Oversight	NIST AI RMF MAP 3.5+MAN 2.4	5	Technically enforced HITL, rubber-stamp detection (>95% threshold), 7-year audit log. Gap: raw HITL override logs not provided.

Appendix E. PRIORITY_MAP (Adaptive Selection Table)

Table E.1 presents the complete rule-based priority mapping for all 30 controls across the three deployment types covered in the evaluation, as coded in audit_agent_adaptive.html. H = High, M = Medium, L = Low, N/A = Not Applicable. Controls marked N/A are excluded from the active control set for that deployment type unless the RAG override is activated.

Table E.1. PRIORITY_MAP: Control priority by deployment type (corrected against tool code).

#	Control Name	Public-Facing Chatbot	Internal Knowledge Assistant	Financial Advisory Tool
1	Policies and Procedures	H	M	H
2	Roles and Responsibilities	H	M	H
3	System Prompt Confidentiality	H	M	H
4	Resource and Cost Governance	H	M	H
5	Vector and Embedding Security	N/A (RAG override: M)	H	N/A (RAG override: M)
6	Data Governance	H	H	H
7	Data Acquisition	M	M	H
8	Data Quality	M	H	H
9	Data Provenance	M	M	H
10	Privacy and Confidentiality	H	H	H
11	Context Establishment	H	M	H
12	Risk Likelihood and Impact	H	M	H
13	Improper Output Handling	H	M	H
14	Supplier Relationships	M	H	H
15	Third-Party Dependencies	M	H	H
16	Adversarial Testing	H	M	H
17	Validity and Reliability	M	H	H
18	Secure Development	H	M	H
19	Verification and Validation	M	M	H
20	Access Control	H	H	H
21	System Documentation	M	L	H

#	Control Name	Public-Facing Chatbot	Internal Knowledge Assistant	Financial Advisory Tool
22	User Notification	H	M	H
23	Explainability	N/A (RAG override: M)	H	N/A
24	Incident Communication	H	L	H
25	Excessive Agency - Permission	L	H	M
26	Continuous Monitoring	H	H	H
27	Incident Response	H	M	H
28	Re-evaluation and Drift Detection	M	H	H
29	Excessive Agency - Tools	N/A	H	N/A
30	Human Oversight	L	M	H

RAG Override: Control 5 (Vector and Embedding Security, NIST AI RMF MAP 4.2 + MEA 2.7) is automatically upgraded from N/A to Medium priority when the RAG architecture toggle is set to Yes on the session landing page. This override applies to Public-Facing Chatbot and Financial Advisory Tool deployment types. Control 23 (Explainability, NIST AI RMF MEA 2.9) is similarly upgraded from N/A to Medium priority for Public-Facing Chatbot deployments when RAG is active. For Financial Advisory Tool, Control 23 remains N/A even with RAG override active, as explainability via source citations is addressed through the Validity and Reliability and System Documentation controls for that deployment type.

Corrections applied to this table versus the pre-correction version: Control 4 (Resource and Cost Governance) for Public-Facing Chatbot corrected from M to H. Multiple other priority cells updated to match the actual PRIORITY_MAP in audit_agent_adaptive.html. RAG Override note updated to name both affected controls specifically (C5 and C23).