masterskill
education through innovation

aci LEARNING | SKILL LABS

CompTIA SEC AI+ (CY0-001), Skill Labs

Course Specifications

Course Number: ACI76-029SL_rev1.0
Lab Length: Approximately 17 hours

CompTIA SEC AI+ (CY0-001)

Introduction
Objective

This lab supports the CompTIA AI Essentials certification objectives and prepares students for the SecAI+ career path. The table below maps each task to the corresponding learning domains.

Task/Major Concept	Description	AI Essentials Domain	SecAI+ Relevance
Task 1: AI Fundamentals	Define AI, ML, DL, and generative AI; understand AI types and capabilities	AI Concepts & Terminology	1.1 Compare and contrast types of AI
Task 2: How AI Models Work	Explore training data, parameters, inference, and model architectures	AI Concepts & Terminology	1.1 Compare and contrast types of AI 1.3 AI system life cycle
Task 3: Prompt Engineering Basics	Master purpose, context, persona, and constraint-based prompting	Generative AI & Prompt Engineering	2.1 Secure AI development
Task 4: Advanced Prompt Techniques	Implement few-shot, chain-of-thought, and system prompt strategies	Generative AI & Prompt Engineering	2.1 Secure AI development 3.1 Prompt injection awareness
Task 5: Output Verification	Detect hallucinations, verify factual accuracy, cross-reference outputs	AI Output Evaluation	2.5 Monitoring AI systems
Task 6: Conversational AI	Build multi-turn conversations, manage context, implement chat loops	Practical AI Applications	2.1 AI application security
Task 7: Responsible AI & Ethics	Explore bias, fairness, transparency, and ethical AI principles	Responsible AI Use	4.1 AI governance structures
Task 8: Data Privacy & Security	Handle PII, implement data governance, understand AI data risks	Balancing Innovation & Privacy	1.2 Data security 4.2 AI risks
Task 9: Practical AI Use Cases	Apply AI to business, productivity, and real-world scenarios	Practical AI Applications	Cross-domain application

Task/Major Concept	Description	AI Essentials Domain	SecAI+ Relevance
Task 10: AI in Cybersecurity	Use AI for threat detection, security analysis, and the SecAI+ path	Future Trends & Cybersecurity	3.1 Attack vectors 4.1 Governance

Overview

Artificial intelligence (AI) is transforming every aspect of technology, business, and cybersecurity. The CompTIA AI Essentials certification validates foundational AI knowledge that is critical for anyone entering the AI-enhanced cybersecurity field. This lab provides a comprehensive, hands-on learning experience using local large language models (LLMs) via Ollama, giving you practical experience with the same types of AI systems you will encounter in production environments. By completing this lab, you will build the foundational knowledge required to pursue the CompTIA SecAI+ (CY0-001) certification, which focuses on securing and defending AI systems.

The lab is organized into ten progressive tasks that mirror the CompTIA AI Essentials exam domains. You will begin with foundational AI concepts and model mechanics, progress through prompt engineering and output verification, explore responsible AI and data privacy, and conclude with a capstone task that applies AI to real cybersecurity scenarios. Each task uses the qwen2.5:1.5b and SmolLM2 family of models running locally on Ubuntu via Ollama, ensuring all data stays on your machine—a key principle in AI security.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	Artificial Intelligence (AI)	The broad field of computer science focused on creating systems capable of performing tasks that typically require human intelligence, such as reasoning, learning, and problem-solving
2	Machine Learning (ML)	A subset of AI where systems learn patterns from data to make predictions or decisions without being explicitly programmed for each scenario
3	Deep Learning (DL)	A subset of ML that uses artificial neural networks with multiple layers to model complex patterns in large datasets (e.g., image recognition, language understanding)
4	Generative AI (GenAI)	AI systems that create new content (text, images, code, audio) based on patterns learned from training data, rather than simply classifying or predicting
5	Large Language Model (LLM)	A type of generative AI model trained on massive text datasets that can understand and generate human language (e.g., GPT, LLaMA, Qwen)
6	Prompt	The input text or instruction given to an AI model to generate a response. The quality of the prompt directly affects the quality of the output.
7	Prompt Engineering	The practice of crafting effective prompts to elicit accurate, relevant, and useful responses from AI models through techniques like context setting, persona assignment, and constraint specification
8	Hallucination	A phenomenon where an AI model generates plausible-sounding but factually incorrect, fabricated, or nonsensical information with apparent confidence
9	Inference	The process of running a trained AI model to generate predictions or outputs from new input data (as opposed to training)

	Key Term	Description
10	Training Data	The dataset used to teach an AI model patterns, relationships, and knowledge during the training phase
11	Parameters	The internal variables of an AI model that are learned during training and determine the model's behavior (e.g., 1.5 billion parameters in qwen2.5:1.5b)
12	Token	The basic unit of text processing for LLMs—typically a word, subword, or character that the model reads and generates
13	Context Window	The maximum amount of text (measured in tokens) that an AI model can process in a single interaction, including both input and output
14	Temperature	A parameter controlling the randomness of AI output: lower values (0.1) produce more deterministic responses, higher values (1.0) produce more creative/varied responses
15	Few-Shot Learning	A prompting technique where examples are provided within the prompt to guide the model's behavior without retraining
16	Chain-of-Thought (CoT)	A prompting technique that instructs the model to show its reasoning process step-by-step, improving accuracy on complex tasks
17	System Prompt	A special instruction that sets the model's role, behavior, and constraints for an entire conversation, typically hidden from the end user
18	Shadow AI	The unauthorized use of AI tools by employees without organizational approval, creating uncontrolled data exposure and compliance risks
19	AI Governance	The frameworks, policies, and practices that guide the responsible development, deployment, and use of AI systems within an organization
20	Responsible AI	The practice of designing, developing, and deploying AI systems in an ethical, fair, transparent, and accountable manner

CompTIA AI Prompting Essentials—Mastering AI Interaction for the Workplace

Introduction
Objective

CompTIA AI Prompting Essentials is organized into five modules with seven core competencies. The table below maps each lab task to the corresponding module and competency area.

Task	Description	Module	Competency Area
Task 1: AI Foundations	Understand AI tools and identify appropriate workplace tasks.	Module 2: Foundations of AI Prompting	AI Task Identification
Task 2: Crafting Effective Prompts	Apply the 5-element framework (Task/Role/Format/Tone/Scope).	Module 3: AI Prompting Basics	Prompt Crafting
Task 3: The Power of Context	Feed reference materials, data, and background into prompts.	Module 3: AI Prompting Basics	Prompt Crafting
Task 4: Output	Detect hallucinations, bias, and quality	Module 3: AI	Critical

Task	Description	Module	Competency Area
Verification	issues.	Prompting Basics	Evaluation
Task 5: Iterative Refinement	Refine AI responses through multi-turn conversation.	Module 3: AI Prompting Basics	Interactive AI Collaboration
Task 6: Ethical and Responsible Prompting	Apply privacy, copyright, transparency, and accountability principles.	Module 2: Foundations of AI Prompting	Ethical AI Use
Task 7: Task Automation	Create reusable templates and batch processing pipelines.	Module 4: Advanced AI Prompting Skills	Task Automation
Task 8: Multi-Step Projects	Chain prompts across a complete project workflow.	Module 4: Advanced AI Prompting Skills	Workplace Integration
Task 9: AI as Tutor, Coach, and Critic	Assign specialized roles to transform AI into different collaborators.	Module 5: Apply AI Skills in Real World Contexts	Workplace Integration

Overview

The ability to communicate effectively with AI tools is rapidly becoming a foundational workplace skill. CompTIA AI Prompting Essentials is a competency-based credential that validates practical skills in crafting effective prompts, evaluating AI outputs, automating tasks, and using AI responsibly in professional settings. Unlike traditional certifications, the Competency Certificate (CompCert) is earned by demonstrating hands-on proficiency through interactive exercises and a 30-minute competency assessment.

This lab provides a comprehensive, hands-on learning experience using local large language models (LLMs) via Ollama. Across nine progressive tasks, you will master the complete AI prompting life cycle: from understanding when to use AI tools, through crafting precise prompts with structured frameworks, to verifying outputs, refining responses iteratively, automating workflows, and applying AI in real-world professional roles. Every exercise runs locally on Ubuntu using the qwen2.5:1.5b model, ensuring all data stays on your machine—a key principle in responsible AI use.

This lab is designed as a companion to the CompTIA AI Essentials lab and a prerequisite for the seven SecAI+ Theory Labs (labs 01–15). The prompting skills developed here directly apply to cybersecurity AI use cases covered in the SecAI+ certification.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	Prompt	The natural language input or instruction given to an AI model to generate a response; the quality of the prompt directly determines the quality of the output
2	Prompt Engineering	The practice of designing, structuring, and refining prompts to elicit accurate, relevant, and useful responses from AI models
3	Large Language Model (LLM)	A type of AI model trained on massive text datasets that can understand and generate human language (e.g., GPT, LLaMA, Qwen)
4	Hallucination	A phenomenon where an AI model generates plausible-sounding but factually incorrect, fabricated, or nonsensical information with apparent

	Key Term	Description
		confidence
5	Context Window	The maximum amount of text (measured in tokens) that an AI model can process in a single interaction, including both input and output
6	Token	The basic unit of text processing for LLMs—typically a word, subword, or character that the model reads and generates
7	System Prompt	A special instruction that sets the model's role, behavior, and constraints for an entire conversation, typically hidden from the end user
8	Few-Shot Learning	A prompting technique where examples are provided within the prompt to guide the model's behavior without retraining
9	Chain-of-Thought (CoT)	A prompting technique that instructs the model to show its reasoning process step by step, improving accuracy on complex tasks
10	Iterative Refinement	The process of progressively improving AI output through a series of follow-up prompts that clarify, constrain, redirect, or combine previous responses
11	Temperature	A parameter controlling the randomness of AI output: lower values produce more deterministic responses, higher values produce more varied/creative responses
12	Inference	The process of running a trained AI model to generate predictions or outputs from new input data
13	Prompt Template	A reusable prompt structure with variable placeholders that can be filled with different data for consistent, repeatable AI interactions
14	Personally Identifiable Information (PII)	Any data that can be used to identify a specific individual, such as names, email addresses, Social Security numbers, or phone numbers
15	Shadow AI	The unauthorized use of AI tools by employees without organizational approval, creating uncontrolled data exposure and compliance risks
16	AI Agent	An AI system that can autonomously plan, execute, and adapt multistep tasks using tools and decision-making capabilities beyond simple prompt response
17	VERIFY Framework	A structured approach to evaluating AI outputs: Validate facts, Evaluate relevance, Recognize hallucinations, Identify bias, check Format, apply Your judgment
18	Responsible AI	The practice of using AI tools in a manner that is ethical, fair, transparent, privacy respecting, and accountable
19	Copyright (AI Context)	The legal framework governing ownership of AI-generated content; purely AI-generated works are generally not copyrightable under current US law
20	Human-in-the-Loop	A design principle requiring human review and approval of AI-generated outputs before they are used, published, or acted upon

Compare and Contrast Various AI Types Used in Cybersecurity

Introduction

Objective

This table maps the major concepts and learning objectives of this lab to the corresponding CompTIA SecAI+ (CY0-001) exam objectives.

Lab Concept/Task	CompTIA SecAI+ (CY0-001) Objective
Compare and Contrast ML, DL, and GenAI Types	1.1: Compare and contrast various types of AI used in cybersecurity
Analyzing Supervised, Unsupervised, and RL Techniques	1.1: Compare and contrast various types of AI used in cybersecurity
Understanding Data Poisoning and Adversarial Examples	2.6: Given a scenario, analyze an attack and implement compensating controls
Evaluating Prompt Engineering for Security Task Automation	3.1: Given a scenario, utilize AI tools for security tasks
Understanding Defensive Prompt Engineering (System Prompts, Sandboxing)	2.2: Given a scenario, implement security controls for AI systems
Analyzing Model Inversion and Model Extraction Attacks	4.2: Explain risks associated with AI

Overview

The rapid evolution of cyber threats necessitates equally advanced defensive mechanisms. Artificial intelligence (AI) has emerged as the cornerstone of modern cybersecurity, offering capabilities for automated threat detection, behavioral analysis, and proactive defense that far surpass traditional methods. This lab provides a comprehensive overview of the AI landscape within cybersecurity, focusing on the distinct types of AI, the critical techniques used to train them, and the emerging discipline of prompt engineering. By the end of this lab, the student will be able to:

- Compare and contrast the core AI types—machine learning (ML), deep learning (DL), and generative AI (GenAI)—and their specific applications in cybersecurity.
- Analyze the security implications of various model training techniques, including supervised, unsupervised, and reinforcement learning, and understand the risks of adversarial attacks like data poisoning and adversarial examples.
- Evaluate the dual role of prompt engineering as both a defensive measure (secure prompting) and an offensive tool (security task automation) in the context of large language models (LLMs), specifically noting the efficiency of the SmolLM2 family of models.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	Artificial Intelligence (AI)	A broad field of computer science concerned with building smart machines capable of performing tasks that typically require human intelligence, such as learning, problem-solving, and decision-making
2	Machine Learning (ML)	A subset of AI that provides systems with the ability to automatically learn and improve from experience without being explicitly programmed, often used for pattern recognition in security data

	Key Term	Description
3	Deep Learning (DL)	A specialized sub-field of ML that uses artificial neural networks with multiple layers (deep neural networks) to analyze complex, unstructured data like raw network traffic or system logs
4	Generative AI (GenAI)	A type of AI that can create new content, such as text, images, or code, often powered by models like large language models (LLMs) and used for both defensive simulation and offensive social engineering
5	Supervised Learning	An ML technique where the model is trained on a labeled dataset, meaning the input data is paired with the correct output, commonly used for malware classification
6	Unsupervised Learning	An ML technique where the model is trained on unlabeled data, tasked with finding hidden patterns or intrinsic structures, primarily used for anomaly detection
7	Reinforcement Learning (RL)	An ML technique where an agent learns to make decisions by interacting with an environment, receiving rewards for desired actions and penalties for undesired ones, often applied to autonomous defense systems
8	Data Drift	The phenomenon where the statistical properties of the target data change over time, causing a trained ML model's predictions to become less accurate, a critical challenge in dynamic threat environments
9	Zero-Day Threat Detection	The process of identifying and mitigating a vulnerability or threat that is unknown to security vendors and has no existing patch or signature, often relying on unsupervised learning for anomaly detection
10	False Positives	A security alert or detection that incorrectly identifies a benign or legitimate activity as malicious, a common issue with anomaly detection systems that can lead to operational disruption
11	Data Poisoning Attacks	An adversarial attack where an attacker injects malicious, mislabeled data into the training set of an ML model, causing the model to learn incorrect associations and potentially creating a backdoor
12	Adversarial Examples	Subtly modified inputs designed to intentionally fool a trained ML model, causing it to misclassify the input (e.g., classifying malicious code as benign) while remaining imperceptible to humans
13	Model Inversion Attacks	A type of attack where an adversary probes a deployed ML model to reconstruct or infer sensitive information about the data used to train the model
14	Model Extraction Attacks	A type of attack where an adversary attempts to steal the intellectual property of a deployed ML model by querying it repeatedly to reconstruct its parameters and logic
15	Prompt Engineering	The discipline of crafting precise, structured input (prompts) to guide a GenAI model to produce a desired, relevant, and safe output, especially in security-related tasks
16	Prompt Injection	A significant threat where an attacker attempts to override the original instructions or safety guidelines of a large language model by injecting a malicious or manipulative prompt
17	System Prompts	Hidden, high-level instructions provided to a GenAI model that define its persona, constraints, and safety guidelines, acting as a defense layer against

	Key Term	Description
		prompt injection
18	Chain-of-Thought (CoT) Prompting	An advanced prompting technique that instructs a model to break down a complex problem into intermediate, logical steps before providing the final answer, improving accuracy and transparency in security analysis
19	Behavioral Analytics	The use of AI and ML to monitor and analyze user and entity activity patterns to establish a baseline of "normal" behavior, allowing for the detection of deviations that may indicate a security threat
20	Convolutional Neural Networks (CNNs)	A class of deep neural networks primarily used for image processing, but applied in cybersecurity for analyzing visual representations of malware code or network traffic flow data

Importance of Data Security relating to AI

Introduction

Objective

This lab directly supports the following CompTIA SecAI+ (CY0-001) exam objectives by providing foundational knowledge and practical context for securing AI systems and their data.

Lab Section/Major Concept	CompTIA SecAI+ (CY0-001) Exam Objective
The Foundational Role of Data Security in the AI Life Cycle (CIA Triad, Model Poisoning, Data Leakage)	1.2: Explain the importance of data security as it relates to AI
Secure Data Processing in AI Pipelines (Ingestion, Training, Inference, Confidential Computing)	1.3: Explain the importance of security in the AI life cycle
Securing Diverse Data Types (PII, IP, Unstructured Data, Least Privilege)	2.4: Given a scenario, implement data security controls for AI systems
Watermarking for Authenticity and Integrity	2.2: Given a scenario, implement security controls for AI systems
Security in Retrieval-Augmented Generation (RAG) Systems (Data Leakage, Prompt Injection, Data Poisoning)	4.2: Explain risks associated with AI

Overview

This lab explores the critical importance of data security in the context of artificial intelligence (AI). As AI systems become increasingly integrated into core business and governmental functions, the volume and sensitivity of the data they process have grown exponentially. The objective of this lab is to explain the multifaceted nature of data security as it relates to the entire AI life cycle, from data ingestion and model training to deployment and inference. We will specifically examine the security implications across key areas: data processing, securing various data types, the role of watermarking, and the unique security challenges posed by retrieval-augmented generation (RAG) systems. A robust understanding of these concepts is fundamental to building trustworthy, responsible, and compliant AI solutions.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	Confidentiality, Integrity, and Availability (CIA Triad)	A foundational model for data security, defining the three core goals: ensuring data is accessible only to authorized parties (confidentiality), that it is accurate and protected from unauthorized modification (integrity), and that authorized users can access it when needed (availability)
2	Model Poisoning	A security attack where an adversary injects malicious, mislabeled, or corrupted data into an AI model's training dataset, causing the model to learn incorrect or harmful behaviors
3	Data Leakage	The unintentional exposure of sensitive information, often occurring when a model's outputs inadvertently reveal details about the private data used in its training or knowledge base
4	Differential Privacy	A system for publicly sharing information about a dataset by adding a controlled amount of noise to the data, which prevents the identification of individual records while preserving the dataset's statistical utility
5	Confidential Computing	A cloud computing technology that isolates sensitive data in a hardware-based trusted execution environment (TEE) during processing, ensuring the data remains encrypted even while in use.
6	Trusted Execution Environment (TEE)	A secure area within a main processor that guarantees code and data loaded inside are protected with respect to confidentiality and integrity
7	Model Inversion Attack	A type of privacy attack where an adversary attempts to reconstruct or infer the sensitive training data used by an AI model based on its outputs or parameters
8	Membership Inference Attack	A privacy attack where an adversary attempts to determine whether a specific data record was included in the AI model's training dataset
9	Personally Identifiable Information (PII)	Any data that could potentially identify a specific individual, such as names, addresses, social security numbers, and biometric data
10	Tokenization	The process of replacing sensitive data elements with a non-sensitive equivalent, or "token," that has no extrinsic or exploitable meaning
11	Data Masking	A technique used to obscure specific data elements within a dataset, often by replacing them with realistic but false data, primarily for non-production environments like testing or training
12	Digital Rights Management (DRM)	A set of access control technologies used to restrict the use, modification, and distribution of proprietary digital content and copyrighted works
13	Data Loss Prevention (DLP)	A set of tools and processes designed to ensure that sensitive data is not lost, misused, or accessed by unauthorized users, often by monitoring and controlling data in use, in motion, and at rest
14	Principle of Least Privilege	A security concept that requires that every user, process, or program be granted only the minimum access rights necessary to perform its job or function
15	Watermarking (AI)	The process of embedding a recognizable, often imperceptible, signal or marker into AI-generated content (text, images, audio) to indicate its artificial origin or ownership
16	Removal Attack (Watermarking)	An adversarial attempt to erase or destroy the embedded watermark signal in digital content without significantly degrading the content's quality

	Key Term	Description
17	Forgery Attack (Watermarking)	An adversarial attempt to embed a false or misleading watermark into content to falsely attribute its origin or ownership.
18	Retrieval-Augmented Generation (RAG)	An AI architecture that enhances large language models (LLMs) by retrieving relevant information from an external, proprietary knowledge base to ground its responses, thereby improving accuracy and reducing hallucination
19	Hallucination (AI)	A phenomenon where a generative AI model produces outputs that are factually incorrect, nonsensical, or unfaithful to the source data, often presented with high confidence
20	Prompt Injection	A security vulnerability where an attacker manipulates an LLM's behavior by inserting malicious instructions or data into the user prompt, often overriding the system's intended instructions

The Importance of Security in the AI Life Cycle

Introduction
Objective

This lab focuses on the foundational concepts of securing the AI life cycle. The following table maps the key concepts and sections of this lab to the corresponding CompTIA SecAI+ (CY0-001) exam objectives.

Lab Section/Concept	CompTIA SecAI+ Objective	Description
Overall Lab Focus	1.3: Explain the importance of security in the AI life cycle	The entire lab content is dedicated to detailing security considerations across all phases of the AI life cycle
4.1. Business Use Case and Scoping	2.1: Given a scenario, use AI threat-modeling resources	Focuses on the initial threat modeling and risk assessment required before development
4.2. Data Collection	2.4: Given a scenario, implement data security controls for AI systems	Covers data poisoning, source integrity, anonymization, and access control for raw data
4.3. Data Preparation and Feature Engineering	2.4: Given a scenario, implement data security controls for AI systems	Covers poisoning detection, label integrity, and feature leakage prevention during data processing
4.4. Model Development/Selection	2.2: Given a scenario, implement security controls for AI systems	Discusses model IP protection, adversarial robustness training, and supply chain security
4.5. Model Evaluation and Validation	2.2: Given a scenario, implement security controls for AI systems	Covers adversarial testing, bias/fairness audits, and explainability (XAI) as security controls
4.6. Monitoring and Maintenance	2.5: Given a scenario, implement monitoring and auditing for an AI system	Focuses on drift detection (Data and Model), real-time anomaly detection, and secure update pipelines
4.7. Feedback and Iteration	2.5: Given a scenario,	Covers feedback integrity and

Lab Section/Concept	CompTIA SecAI+ Objective	Description
	implement monitoring and auditing for an AI system	maintaining detailed audit trails for accountability
4.8. Human-Centric AI Design Principles	4.1: Explain AI governance structures	Relates to the principles of accountability and governance, which form the structure for secure AI
Glossary (Data Poisoning, Model Evasion)	2.6: Given a scenario, analyze an attack and implement compensating controls	Defines key attacks that require compensating controls throughout the life cycle

Overview

Artificial intelligence (AI) systems are rapidly becoming integral to critical business operations, national security, and daily life. As their deployment increases, so does the attack surface and the potential for malicious exploitation. This lab explores the critical importance of integrating security throughout the entire AI system life cycle, from the initial business case definition to continuous monitoring and maintenance. Unlike traditional software, AI systems introduce unique vulnerabilities, such as data poisoning, model evasion, and intellectual property theft, which necessitate a "security by design" approach. The objective of this lab is to explain the importance of security at every stage of the AI life cycle, ensuring robustness, trustworthiness, and adherence to human-centric design principles.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	AI Life Cycle	The iterative process of developing, deploying, and maintaining an AI system, from the initial business case to continuous monitoring and retirement
2	Data Poisoning	A security attack where an adversary injects corrupted or misleading data into the training set to compromise the model's integrity and performance
3	Adversarial Example	A subtly modified input that is intentionally designed to cause an AI model to misclassify or make an incorrect prediction
4	Model Evasion	A type of adversarial attack where the attacker manipulates the input data at inference time to bypass the deployed model's security controls
5	Model Inversion	An attack that attempts to reconstruct the sensitive, private training data from the model's outputs or parameters
6	Model Extraction (Theft)	An attack where an adversary queries a deployed model to steal its intellectual property by creating a functional copy (a "surrogate model")
7	Threat Modeling	A structured process for identifying potential threats, vulnerabilities, and attack vectors in an AI system, typically performed early in the life cycle
8	Secure-by-Design	A principle that mandates integrating security considerations into every phase of the AI life cycle, starting from the initial design and requirements
9	Data Drift	A change in the statistical properties of the live input data compared to the training data, which can degrade model performance
10	Model Drift	The degradation of a model's predictive performance over time due to changes in the real-world environment or data distribution

	Key Term	Description
11	Provenance Tracking	The process of recording and verifying the origin, history, and integrity of data and models to ensure trustworthiness and auditability
12	Adversarial Robustness	The ability of an AI model to maintain its performance and integrity when subjected to adversarial attacks
13	Human-Centric AI	An approach to AI development that prioritizes the needs, values, and well-being of humans, ensuring ethical and responsible outcomes
14	Bias Audit	A systematic review of an AI system's data and model to identify and mitigate unfair or discriminatory outcomes against specific demographic groups
15	Explainable AI (XAI)	Techniques that allow human users to understand the output and decision-making process of AI models, crucial for debugging and security analysis
16	Feature Leakage	The accidental inclusion of features in the training data that contain information about the target variable, leading to overly optimistic and misleading performance metrics
17	Supply Chain Security (AI)	Ensuring that all components used in the AI system (e.g., open-source libraries, pre-trained models) are free from vulnerabilities or malicious code
18	Inference Attack	A broad category of attacks that target the model during its deployment phase (inference time), such as evasion or model inversion
19	Confidentiality	The security principle ensuring that sensitive data and model intellectual property are protected from unauthorized access
20	Integrity	The security principle ensuring that data and models are accurate, complete, and have not been tampered with throughout the life cycle

Utilizing AI Threat-Modelling Resources

Introduction
Objective

This lab is designed to provide hands-on experience that directly maps to several objectives of the CompTIA SecAI+ (CY0-001) certification exam. The table below details how each major task or concept covered in this lab aligns with the official exam objectives.

Task/Concept	CompTIA SecAI+ (CY0-001) Exam Objective
Task 1: OWASP Top 10 for LLM (Prompt Injection)	2.6: Given a scenario, analyze an attack and implement compensating controls
	3.1: Given a scenario, utilize AI tools for security tasks
Task 2: MIT AI Risk Repository (Algorithmic Bias)	2.4: Given a scenario, implement data security controls for AI systems
	4.2: Explain risks associated with AI
Task 3: MITRE ATLAS (Adversarial Evasion)	2.1: Given a scenario, use AI threat-modeling resources
	2.6: Given a scenario, analyze an attack and implement compensating controls

Task/Concept	CompTIA SecAI+ (CY0-001) Exam Objective
Task 4: CVE/CWE Investigation	1.3: Explain the importance of security in the AI life cycle
	2.2: Given a scenario, implement security controls for AI systems
Task 5: STRIDE Threat Modeling	2.1: Given a scenario, use AI threat-modeling resources
	2.2: Given a scenario, implement security controls for AI systems

Overview

This lab provides a comprehensive, hands-on experience in using industry-leading resources for artificial intelligence (AI) threat modeling. Students will learn to navigate and apply frameworks such as the OWASP Top 10 for Large Language Model Applications, the MIT AI Risk Repository, and the MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS). The primary objective is to equip students with the practical skills necessary to analyze a given AI scenario and effectively apply these threat-modeling resources to identify, classify, and mitigate potential risks.

Learning Objective: Given a scenario, use AI threat-modeling resources.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	Adversarial Example	An input to an AI model that has been intentionally perturbed to cause the model to make an incorrect prediction
2	Data Poisoning	An attack where an adversary introduces malicious data into the training dataset, corrupting the model's integrity and performance.
3	Model Inversion	An attack that attempts to reconstruct the sensitive training data used to train a machine learning (ML) model
4	Prompt Injection	An attack that involves manipulating a LLM by providing malicious input (a "prompt") to override its original instructions
5	System Prompt	Hidden instructions given to an LLM before user interaction that define its behavior, personality, and constraints; often contains sensitive business logic or security controls
6	Jailbreaking	Techniques used to bypass an LLM's safety controls or content restrictions, making it produce outputs it was designed to refuse
7	Direct Prompt Injection	An attack where the user directly provides malicious instructions in their input to manipulate the LLM
8	Indirect Prompt Injection	An attack where malicious instructions are hidden in external data sources (documents, websites) that the LLM processes
9	OWASP Top 10 for LLM	A list of the top 10 most critical security risks specific to LLM applications, published by the Open Worldwide Application Security Project (OWASP).
10	MITRE ATLAS	A knowledge base of adversary tactics, techniques, and mitigations based on real-world observations of attacks against AI/ML systems; organized similar to

	Key Term	Description
		MITRE ATT&CK, ATLAS provides standardized IDs (AML.*) for tactics (why attackers act), techniques (how they act), and mitigations (how to defend).
11	MIT AI Risk Repository	A comprehensive, living database of categorized AI risks, providing a structured vocabulary for risk assessment
12	Threat Modeling	A structured process of identifying potential threats, vulnerabilities, and attack vectors to an application or system
13	STRIDE	Microsoft's threat modeling methodology using six categories: Spoofing (authentication), Tampering (integrity), Repudiation (non-repudiation), Information Disclosure (confidentiality), Denial of Service (availability), and Elevation of Privilege (authorization); each category maps to a security property being violated.
14	Evasion Attack	An attack where the adversary manipulates the input data at inference time to cause a misclassification
15	Model Extraction (Theft)	An attack where an adversary probes a model to steal its underlying parameters or architecture, often by observing its outputs
16	Supply Chain Vulnerability (AI)	A risk associated with the components, data, or services used in the development and deployment pipeline of an AI system
17	Insecure Output Handling	A risk where an LLM's output is accepted without sufficient scrutiny, potentially leading to cross-site scripting or remote code execution
18	Denial of Service (Model DoS)	An attack that consumes excessive computational resources of an AI model, making it unavailable to legitimate users
19	Confidentiality	The security principle that prevents the unauthorized disclosure of information
20	Integrity	The security principle that ensures data has not been modified or tampered with
21	Availability	The security principle that ensures systems and data are accessible when needed
22	Common Vulnerabilities and Exposures (CVE)	A dictionary of publicly known information-security vulnerabilities and exposures
23	Adversarial Robustness	The ability of an AI model to maintain its performance even when subjected to adversarial examples
24	Red Teaming (AI)	A simulated attack exercise performed to test the security and safety of an AI system
25	Taint Analysis	A technique used to track the flow of untrusted data (taint) through an application to prevent security vulnerabilities

Implement Security Controls for AI Systems

Introduction
Objective

Upon completion of this lab, the student will be able to:

- Implement basic defenses against adversarial attacks on machine learning models.
- Configure an API gateway to enforce security policies for AI service access.

- Develop and execute systematic tests to validate the effectiveness of AI guardrails.
- Understand the requirements for a comprehensive, layered security approach for AI systems.
- Apply security controls to real-world AI deployment scenarios across multiple industry verticals.

This lab provides hands-on experience and conceptual understanding that directly maps to the following CompTIA SecAI+ (CY0-001) exam objectives.

Lab Task/Concept	CompTIA SecAI+ Objective	Description
Task 1: Input Validation	2.2: Given a scenario, implement security controls for AI systems	Implementing input sanitization is a fundamental security control to protect the model from malicious input.
Task 1: Input Validation	2.6: Given a scenario, analyze an attack and implement compensating controls	Input validation acts as a compensating control against adversarial inputs (e.g., SQL injection, prompt injection).
Task 2: Rate Limiting	2.2: Given a scenario, implement security controls for AI systems	Rate limiting is a control implemented to protect the availability and intellectual property of the AI service.
Task 2: Rate Limiting	4.2: Explain risks associated with AI	Model stealing (extraction) is a significant intellectual property risk mitigated by rate limiting.
Task 3: AI Gateway Authentication	2.3: Given a scenario, implement access controls for AI systems	The AI gateway enforces authentication and authorization, which are core access controls for AI services.
Task 3: AI Gateway Authentication	2.5: Given a scenario, implement monitoring and auditing for an AI system	Gateways are the central point for logging and auditing all access requests to the AI system.
Task 4 & 5: Guardrail Testing	2.2: Given a scenario, implement security controls for AI systems	Guardrails are a critical security control specifically designed to manage the behavior and output of generative AI models.
Task 4 & 5: Guardrail Testing	2.5: Given a scenario, implement monitoring and auditing for an AI system	Systematic testing and validation are forms of auditing to ensure the security controls (guardrails) are functioning as intended.
Task 4 & 5: Guardrail Testing	3.1: Given a scenario, utilize AI tools for security tasks	Using Ollama and custom Python scripts to test the security of an AI system is an example of utilizing AI tools for security.
Overall Lab	1.3: Explain the importance of security in the AI lifecycle	The lab covers controls applied at the input (Task 1), service layer (Task 2 & 3), and model output (Task 4 & 5), demonstrating a layered approach across the AI life cycle.

Overview

Artificial Intelligence (AI) systems, although offering transformative capabilities, introduce a unique and complex set of security challenges that traditional cybersecurity measures often fail to address. This practical lab is designed to provide hands-on experience in implementing and validating essential security controls across the AI life cycle. The focus will be on three critical areas: model controls, which protect the integrity and confidentiality of the core AI artifact; gateway controls, which secure the

communication layer between users and the AI service; and guardrail testing and validation, which ensures the safe and ethical behavior of generative models.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	Adversarial Attack	Malicious input crafted to cause a machine learning model to make an incorrect prediction or decision
2	Data Poisoning	An attack where an adversary injects corrupted or malicious data into the training set to compromise the model's integrity and performance
3	Model Inversion	A privacy attack that attempts to reconstruct sensitive training data points from the model's outputs or parameters
4	Model Stealing (Extraction)	An intellectual property attack where an adversary queries a target model to create a functionally equivalent, unauthorized copy.
5	Model Controls	Security measures applied directly to the AI/ML model life cycle, including training, deployment, and inference, to ensure integrity and confidentiality
6	AI Gateway	A centralized proxy or service that manages, secures, and monitors API traffic to and from one or more AI models
7	Prompt Injection	A type of attack where malicious or manipulative input is used to override or manipulate the model's pre-defined instructions or guardrails
8	Guardrails	Pre-defined rules, policies, and filters implemented to constrain the behavior, output, and safety of a generative AI model
9	Guardrail Testing	The systematic process of evaluating the effectiveness and robustness of AI guardrails against various adversarial inputs and boundary conditions
10	Validation Set	A subset of data used to tune hyperparameters and provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters
11	Inference	The process of using a trained machine learning model to make predictions or decisions on new, unseen data
12	Federated Learning	A decentralized machine learning approach where models are trained on local data samples, and only aggregated model updates are shared, enhancing data privacy
13	Homomorphic Encryption	An advanced encryption method that allows computations to be performed on encrypted data without the need for decryption
14	Differential Privacy	A system for sharing datasets publicly by describing the patterns of groups within the dataset while mathematically limiting the disclosure of individual records
15	Explainable AI (XAI)	A set of tools and techniques that allow users to understand and interpret the predictions and decisions made by machine learning models
16	Confidential Computing	A cloud computing technology that protects data in use by performing computation within a hardware-based trusted execution environment (TEE)
17	Rate Limiting	A gateway control that restricts the number of requests a user or client can make

	Key Term	Description
		in a given time period to prevent abuse, excessive costs, or denial-of-service (DoS) attacks
18	Input Sanitization	The process of cleaning, filtering, and validating user input before it is passed to the AI model to mitigate prompt injection and other input-based attacks
19	Red Teaming	A security practice where a dedicated team simulates adversarial attacks and exploits to test the resilience and security posture of an AI system
20	Taint Analysis	A technique used to track the flow of untrusted data (taint) through a program to identify potential security vulnerabilities, often used in code analysis

Implement Access Controls for AI Systems

Introduction
Objective

Learning Objectives: Upon completion of this lab, the student will be able to:

- Differentiate between access control requirements for AI models, data, and agents.
- Implement role-based access control (RBAC) to manage access to AI model inference endpoints (Use Case 3: Multi-Tenant AI Platforms).
- Configure fine-grained access policies for securing sensitive AI training and inference data (Use Case 4: Research Data Protection).
- Define and enforce attribute-based access control (ABAC) policies for AI agents interacting with external resources (Use Case 5: Autonomous AI Agents).
- Secure AI service APIs using industry-standard authentication and authorization protocols like API Keys.
- Establish logging and monitoring for auditing access control events within an AI system (Use Case 6: Regulatory Compliance).

This lab covers the following CompTIA SecAI+ (CY0-001) exam objectives:

Task/Concept	CompTIA SecAI+ Objective	Description
Exercise 1: RBAC for Model Inference	2.3: Given a scenario, implement access controls for AI systems	Directly implements role-based access control (RBAC) to restrict who can use the AI model's prediction endpoint
Exercise 2: Fine-Grained Data Access	2.4: Given a scenario, implement data security controls for AI systems	Focuses on using Linux permissions to simulate fine-grained access control (FGAC) and the principle of least privilege for sensitive training data and model artifacts
Exercise 3: ABAC for AI Agent	2.3: Given a scenario, implement access controls for AI systems	Implements attribute-based access control (ABAC), a dynamic access control model essential for securing AI agents
Exercise 4: API Key Security	2.2: Given a scenario, implement security controls for AI systems	Implements a common security control (API key/model name) to secure the network interface (API) of the AI service

Task/Concept	CompTIA SecAI+ Objective	Description
Exercise 5: Auditing and Monitoring	2.5: Given a scenario, implement monitoring and auditing for an AI system	Covers the essential practice of logging access attempts and generating an audit report to ensure compliance and detect policy violations
Introduction/Glossary	1.3: Explain the importance of security in the AI life cycle	The entire lab reinforces the need for security controls across the AI system components (model, data, agent, API).
Exercise 2 (Principle of Least Privilege)	1.2: Explain the importance of data security as it relates to AI	Emphasizes protecting sensitive training data and model integrity through least privilege, a core data security concept

Overview

Artificial intelligence (AI) systems, including machine learning models, data pipelines, and autonomous agents, present unique challenges for traditional access control mechanisms. The complexity arises from the distributed nature of AI components, the sensitivity of the data used for training, and the potential for agents to act on behalf of users with elevated privileges.

Real-World Context: Consider a healthcare AI system that predicts patient diagnoses. Data scientists need full access to anonymized training data, doctors need inference access to make predictions, and auditors need read-only access to model decisions. Without proper access controls, a malicious actor could steal sensitive patient data, manipulate model predictions, or access the system without authorization. In 2023, multiple organizations reported AI-related data breaches costing millions in damages and regulatory fines (Use Case 1: Healthcare AI Security).

Similarly, financial institutions deploying fraud detection AI must ensure that only authorized personnel can query transaction models, that customer data remains protected, and that all access attempts are logged for compliance (Use Case 2: Financial AI Compliance). The 2024 AI Security Report found that 67% of AI breaches involved inadequate access controls.

This lab provides a comprehensive, hands-on experience in implementing appropriate access controls across the critical components of an AI system: the model itself, the underlying data, the AI agents, and the network APIs that expose the service. By completing the tasks in this lab, students will gain practical skills in applying modern access control models, such as role-based access control (RBAC) and attribute-based access control (ABAC), to secure the AI life cycle and mitigate risks associated with unauthorized access and misuse.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	Access Control	A security technique that regulates who or what can view or use resources in a computing environment
2	AI Agent	A software entity that perceives its environment and takes actions that maximize its chance of successfully achieving its goals
3	Attribute-Based Access Control (ABAC)	An authorization model that grants access based on attributes (characteristics) of the user, the resource, and the environment

	Key Term	Description
4	Authorization	The function of specifying access rights to resources
5	Authentication	The process of verifying the identity of a user, process, or device
6	Confidentiality	The principle that prevents the unauthorized disclosure of information
7	Data Access	The process of retrieving or manipulating data, often governed by policies to ensure privacy and security
8	Fine-Grained Access Control (FGAC)	A method of restricting access to a resource at a very detailed level, such as individual rows or columns in a database
9	Inference Endpoint	A network address (API) where a deployed machine learning model can be queried to make predictions
10	Least Privilege	A security principle that requires that a user or process be given only the minimum levels of access necessary to perform its job functions
11	Model Access	The control mechanisms governing who can deploy, update, or query a machine learning model
12	Multi-Factor Authentication (MFA)	An authentication method that requires the user to provide two or more verification factors to gain access to a resource
13	Network/API Access	The security measures applied to the network interfaces and APIs that expose AI services to internal or external consumers
14	Ollama	A lightweight, open-source framework for running large language models (LLMs) locally
15	Policy Enforcement Point (PEP)	The component in an access control system that enforces the access decision made by the policy decision point (PDP)
16	Policy Decision Point (PDP)	The component in an access control system that evaluates the access request against the defined policies and makes an access decision
17	Principle of Separation of Duties	A security principle that ensures that no single individual has control over all critical functions of a process
18	Role-Based Access Control (RBAC)	An authorization model that grants access based on the roles users have within an organization
19	Tokenization	The process of replacing sensitive data with a non-sensitive equivalent, or token, that has no extrinsic or exploitable meaning
20	Zero Trust	A security concept centered on the belief that organizations should not automatically trust anything inside or outside its perimeters and must verify anything and everything trying to connect to its systems

Implement Data Security Controls for AI Systems

Introduction
Objective

Learning Objectives:

- Implement data encryption for AI training datasets at rest using industry-standard tools.

- Configure secure communication channels (data in transit) for data ingestion into an ML platform.
- Apply role-based access control (RBAC) to limit access to sensitive AI assets, including data and trained models.
- Use data masking and anonymization techniques to protect personally identifiable information (PII) within datasets.
- Securely deploy an AI model endpoint and protect the data exchanged during the inference phase.

Lab Task/Concept	CompTIA SecAI+ Objective	Description
Task 1: Encryption at Rest	2.4: Given a scenario, implement data security controls for AI systems	Securing sensitive training data using AES-256 encryption with OpenSSL
Task 2: Encryption in Transit	2.4: Given a scenario, implement data security controls for AI systems	Securing data ingestion using SCP (SSH-based secure transfer)
Task 3: Role-Based Access Control (RBAC)	2.3: Given a scenario, implement access controls for AI systems	Implementing Linux file permissions to enforce least privilege access to data
Task 4: Data Masking/Anonymization	2.4: Given a scenario, implement data security controls for AI systems	Using Python/Pandas to mask PII in a dataset before training
Task 4: LLM Code Review	3.1: Given a scenario, utilize AI tools for security tasks	Using a SmolLM model to review a script for data leakage vulnerabilities
Task 5: Endpoint Security (TLS/SSL)	2.4: Given a scenario, implement data security controls for AI systems	Generating and using self-signed certificates to secure the model inference endpoint
Task 5: Adversarial Prompt Check	2.6: Given a scenario, analyze an attack and implement compensating controls	Testing the LLM's resistance to a simple prompt injection attack

Overview

Artificial intelligence (AI) systems, particularly those based on Machine Learning (ML), rely heavily on vast amounts of data for training and operation. The security of this data—both the raw input and the resulting models—is paramount to maintaining privacy, compliance, and system integrity. This lab is designed to provide a comprehensive, hands-on experience in implementing essential data security controls across the AI life cycle. The primary focus will be on meeting encryption requirements for data at rest and in transit, and establishing robust data safety protocols to prevent unauthorized access, leakage, and tampering.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	Adversarial Example	A subtle, intentional perturbation of an input designed to cause an AI model to make a mistake, often leading to misclassification.

	Key Term	Description
2	AI Data Security	The practice of protecting the data used by and generated from AI systems from unauthorized access, corruption, or theft throughout its life cycle
3	Anonymization	The process of removing or modifying personally identifiable information (PII) from a dataset so that the data subject cannot be identified
4	Confidential Computing	A technology that protects data in use by performing computation in a hardware-based, attested, and verifiable trusted execution environment (TEE)
5	Data Leakage	The unintentional exposure of sensitive information from a training dataset, often through poorly secured storage or model outputs
6	Data Masking	A technique where sensitive data is obscured or replaced with realistic, but nonsensitive, data to protect privacy while maintaining data utility for testing or training
7	Data Minimization	The principle that only the minimum amount of personal data necessary to achieve a specified purpose should be collected and processed
8	Data Provenance	The record of the origin and history of a piece of data, including where it came from, what transformations it underwent, and who accessed it
9	Data Safety	A broad term encompassing the measures and controls implemented to ensure the confidentiality, integrity, and availability of data, especially in the context of AI systems
10	Differential Privacy	A system for publicly sharing information about a dataset by describing the patterns of groups within the dataset while withholding information about individuals
11	Encryption at Rest	The process of encrypting data when it is stored on a physical medium, such as a hard drive or cloud storage bucket, to prevent unauthorized access
12	Encryption in Transit	The process of encrypting data as it moves from one location to another, typically over a network, often using protocols like TLS/SSL
13	Fully Homomorphic Encryption (FHE)	An advanced encryption method that allows computations to be performed on encrypted data without decrypting it first, enabling secure data processing
14	Inference Security	The security measures applied to the deployed AI model and the data it processes during the prediction or decision-making phase
15	Key Management System (KMS)	A system for managing cryptographic keys, including their generation, storage, usage, and destruction
16	Model Poisoning	A type of adversarial attack where an attacker injects malicious data into the training set to corrupt the integrity of the resulting AI model
17	Role-Based Access Control (RBAC)	A security mechanism that restricts system access to authorized users based on their role within an organization
18	Secure Enclave	A protected area of a processor that is isolated from the rest of the system, designed to protect sensitive data and code from unauthorized

	Key Term	Description
		access
19	Secure ML Pipeline	A set of automated processes for building, training, and deploying ML models that incorporates security checks and controls at every stage
20	Transport Layer Security/Secure Sockets Layer (TLS/SSL)	Transport layer security/secure sockets layer, cryptographic protocols designed to provide communication security over a computer network

Implement Monitoring and Auditing for an AI System

Introduction
Objective

This lab directly supports the following CompTIA SecAI+ (CY0-001) exam objectives. The table below maps the major concepts and hands-on tasks in this lab to the corresponding exam objectives, providing a clear link between the practical skills learned and the required certification knowledge.

Task/Major Concept	Description	CompTIA SecAI+ (CY0-001) Objective
Overall Lab Focus	Implementing a comprehensive monitoring and auditing framework for an AI system	2.5: Given a scenario, implement monitoring and auditing for an AI system.
Task 1: Prompt and Response Monitoring	Tracking inputs (prompts) and outputs (responses) and calculating a response confidence level	2.5: Given a scenario, implement monitoring and auditing for an AI system.
Task 2: Log Monitoring and Analysis	Using shell tools and scripting to filter, analyze, and count log entries for errors and warnings	2.5: Given a scenario, implement monitoring and auditing for an AI system.
Task 3: Log Sanitization (PII Masking)	Implementing controls to remove or mask sensitive data (PII) from logs	2.4: Given a scenario, implement data security controls for AI systems.
Task 3: Log Protection (Encryption/Permissions)	Applying security controls like encryption and file permissions to protect log data integrity	2.4: Given a scenario, implement data security controls for AI systems.
Task 4: Rate and Cost Monitoring	Implementing rate limiting and tracking token usage for resource management and cost control	2.5: Given a scenario, implement monitoring and auditing for an AI system.
Task 5: Compliance Audit	Simulating an audit to verify adherence to data governance policies (e.g., PII handling)	4.3: Explain the impact of compliance on the business use and development of AI.

Overview

The deployment of artificial intelligence (AI) and large language models (LLMs) into production environments introduces unique challenges related to performance, reliability, security, and compliance. Unlike traditional software, AI systems can exhibit model drift, data drift, and hallucinations, which necessitate specialized monitoring and auditing practices. This lab provides a practical, hands-on approach to implementing a robust observability and auditing framework for an AI system, focusing on

key areas such as prompt and response tracking, log management, cost control, and compliance checks. By the end of this lab, you will be able to implement essential monitoring components to ensure the quality, security, and responsible operation of AI applications using a local, self-hosted LLM environment based on Ubuntu and Ollama. This lab utilizes the SmolLM2 family of models, which are specifically designed for high-speed, resource-efficient local deployment, allowing for rapid iteration and lower operational overhead in monitoring tasks.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	AI Observability	The practice of collecting, analyzing, and visualizing key metrics and signals from AI systems to understand their internal state and performance in production
2	AI Cost Monitoring	Tracking and analyzing the expenditure associated with running an AI system, including API usage, compute resources, and storage costs
3	Auditing for Quality	The process of systematically evaluating an AI system's performance metrics (e.g., accuracy, latency, fairness) against predefined quality standards
4	Auditing for Compliance	The process of verifying that an AI system adheres to relevant legal, regulatory, and internal policy requirements (e.g., GDPR, HIPAA, internal ethical guidelines)
5	Hallucination	A phenomenon where an LLM generates plausible-sounding but factually incorrect or nonsensical information
6	Token Usage	The measure of input and output data for LLMs, where a token is a unit of text (e.g., a word or part of a word), used for billing and rate limiting
7	Personally Identifiable Information (PII)	Information that can be used to directly or indirectly identify an individual, such as names, addresses, or social security numbers
8	Red Teaming	A structured process of testing an AI system by simulating adversarial attacks to find vulnerabilities, biases, or unsafe behaviors
9	Fine-tuning	The process of further training a pre-trained model on a smaller, task-specific dataset to improve performance for a particular use case
10	Explainability (XAI)	The set of techniques that allows human users to understand the output of AI models, crucial for auditing and trust [9]
11	Ground Truth	The actual, verifiable outcome or correct answer used to evaluate the performance of an AI model
12	Model Drift	A decline in the model's performance due to changes in the real-world data distribution compared to the training data
13	Prompt Monitoring	Tracking the input prompts sent to an LLM to detect malicious, inappropriate, or unexpected user behavior
14	Response Confidence Level	A metric, often a probability score or a derived value, indicating the model's certainty in its generated output
15	Log Monitoring	The systematic collection and analysis of logs generated by an AI

	Key Term	Description
		application and its underlying infrastructure to detect errors, anomalies, and performance issues
16	Log Sanitization	The process of removing or masking sensitive, PII or proprietary data from logs before storage or analysis
17	Log Protection	Implementing security controls, such as encryption and access control, to prevent unauthorized access, modification, or deletion of log data
18	Rate Monitoring	Tracking the frequency of requests (e.g., API calls per second) to an AI service to manage capacity, detect denial-of-service attacks, and enforce usage limits

Analyzing an Attack and Implementing Compensating Controls for an AI System

Introduction

Objective

This lab directly supports the preparation for the CompTIA SecAI+ (CY0-001) certification exam. The following table maps the major concepts and tasks covered in this lab to the corresponding exam objectives.

Lab Task/Concept	CompTIA SecAI+ (CY0-001) Objective	Description
Task 2: Data Poisoning Analysis	2.5: Given a scenario, implement monitoring and auditing for an AI system	Analyzing training logs and model metrics to detect anomalies (loss spikes) indicative of a poisoning attack
Task 3: Evasion Attack Analysis	2.6: Given a scenario, analyze an attack and implement compensating controls	Investigating adversarial examples and calculating perturbation magnitude to understand the evasion attack vector
Task 4: Implementing Compensating Control	2.2: Given a scenario, implement security controls for AI systems	Implementing a compensating control (input filter) to mitigate the immediate threat of an evasion attack
Task 5: Final Reporting & Recommendation	2.6: Given a scenario, analyze an attack and implement compensating controls	Verifying the control's effectiveness and recommending a corrective control (model retraining) to address the root cause
General Lab Context	4.2: Explain risks associated with AI	Understanding the mechanics and impact of adversarial machine learning (AML) attacks (poisoning and evasion)

Overview

Artificial intelligence (AI) systems, particularly those based on machine learning (ML), are increasingly deployed in critical infrastructure, including security and defense applications. This reliance introduces a new class of security risks, primarily from adversarial machine learning (AML) attacks. These attacks aim to manipulate the behavior of the AI model, either during training (poisoning) or during inference (evasion), to cause a malfunction or a security breach.

This lab is designed to provide a simulation-based learning scenario where a security analyst must investigate evidence of an AML attack on a critical AI system—specifically, a computer vision model

used for object detection in a surveillance system. The analysis will focus on identifying the attack vector and the resulting impact through simulated outputs from Python scripts. Following the analysis, the student will be tasked with suggesting and implementing compensating controls to mitigate the identified risks, aligning with the objective: 2.6: Given a scenario, analyze the evidence of an attack and suggest compensating controls for AI systems.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	Adversarial Machine Learning (AML)	A field of study focused on the vulnerabilities of machine learning models to malicious inputs and the development of defensive techniques
2	Data Poisoning Attack	An attack where an adversary injects malicious data into the training dataset to corrupt the model's integrity and performance
3	Evasion Attack	An attack where an adversary crafts a subtly modified input (an adversarial example) to cause a trained model to misclassify it during inference
4	Adversarial Example	An input to a machine learning model that has been intentionally perturbed to cause the model to make an incorrect prediction, while remaining imperceptible to humans
5	Compensating Control	An alternative security control used to mitigate a risk when a primary control is not feasible or cannot fully address the threat
6	Preventative Control	A security control designed to stop an attack or security violation from occurring (e.g., input validation)
7	Detective Control	A security control designed to identify and alert on an attack or security violation that has occurred (e.g., anomaly detection)
8	Corrective Control	A security control designed to fix the effects of an attack or security violation (e.g., model retraining)
9	Model Robustness	The ability of a machine learning model to maintain its performance and accuracy when faced with noisy, corrupted, or adversarial inputs
10	Threat Model	A structured approach to identifying, analyzing, and prioritizing potential threats to a system, including the assets, vulnerabilities, and adversaries
11	Attack Surface	The sum of all points where an unauthorized user can try to enter data to or extract data from an environment
12	Feature Space	The multidimensional space where the data points used to train a machine learning model reside, with each dimension representing a feature
13	Perturbation	A small, often imperceptible, change applied to an input data point to create an adversarial example
14	Transferability	The phenomenon where an adversarial example crafted for one model can successfully cause misclassification in a different model
15	White-Box Attack	An attack where the adversary has full knowledge of the target model's architecture, parameters, and training data
16	Black-Box Attack	An attack where the adversary has no knowledge of the target model's internal workings, only access to its input/output interface

	Key Term	Description
17	Data Integrity	The assurance that data is accurate, consistent, and trustworthy throughout its lifecycle
18	Model Inversion Attack	An attack that attempts to reconstruct sensitive training data from the model's outputs
19	Model Extraction Attack	An attack that attempts to steal the intellectual property of a model by querying it and replicating its functionality
20	NIST AI RMF	The National Institute of Standards and Technology's Artificial Intelligence Risk Management Framework, a voluntary framework for managing risks associated with AI

Utilizing AI Tools for Security Tasks

Introduction

Objective

This lab directly supports the preparation for the CompTIA SecAI+ (CY0-001) certification exam by providing hands-on experience with key concepts. The table below maps the major tasks and concepts covered in this lab to the corresponding exam objectives.

Task/Concept Covered	CompTIA SecAI+ (CY0-001) Objective
Overall Lab Theme: *Utilizing AI tools for security tasks* (log analysis, threat intel, vulnerability prioritization, phishing detection)	3.1: Given a scenario, utilize AI tools for security tasks
Task 3.1: AI-Assisted Anomaly Detection in Web Logs	3.1: Given a scenario, utilize AI tools for security tasks
Task 3.2: AI-Assisted Threat Intelligence Summarization	3.1: Given a scenario, utilize AI tools for security tasks
Task 3.3: AI-Assisted Vulnerability Prioritization	3.1: Given a scenario, utilize AI tools for security tasks
Task 3.4: AI-Assisted Phishing Detection	3.1: Given a scenario, utilize AI tools for security tasks
Task 3.5: AI-Assisted Incident Response Triage (Generating SOAR actions)	3.2: Given a scenario, automate security tasks using AI

Overview

Artificial intelligence (AI) and machine learning (ML) have become transformative forces in the field of cybersecurity, moving beyond traditional signature-based detection to enable predictive threat intelligence, automated incident response, and behavioral anomaly detection. This lab is designed to provide a practical understanding of how AI-enabled tools are used to facilitate critical security tasks, thereby enhancing the efficiency and effectiveness of security operations. Instead of relying on external cloud APIs, this lab utilizes Ollama and Docker to run highly efficient small language models (SLMs) like SmolLM2 135M and SmolLM2 360M locally on your Ubuntu system, simulating a secure, on-premise AI environment.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	Artificial Intelligence (AI)	The simulation of human intelligence processes by machines, especially computer systems, including learning, reasoning, and self-correction
2	Machine Learning (ML)	A subset of AI that enables systems to automatically learn and improve from experience without being explicitly programmed, often used for pattern recognition in security data
3	Threat Detection	The process of identifying malicious activities or indicators of compromise (IoC) within a network or system, often significantly accelerated by AI algorithms.
4	Anomaly Detection	The identification of items, events, or observations that do not conform to an expected pattern or other items in a dataset, which is a core function of AI in security
5	Incident Response (IR)	The structured approach an organization takes to manage the aftermath of a security breach or cyberattack, with AI assisting in triage and containment
6	Security Operations Center (SOC)	A centralized function within an organization employing people, processes, and technology to continuously monitor and improve an organization's security posture.
7	Generative AI	A type of AI that can create new content, such as text, images, or code, with applications in security for generating threat intelligence summaries or simulating attacks
8	Adversarial Attack	A technique used to fool a machine learning model by supplying deceptive input, a key concern in the security of AI systems themselves
9	Endpoint Security	The practice of securing end-user devices like desktops, laptops, and mobile devices from malicious threats, often using AI for behavioral analysis
10	Vulnerability Assessment	The process of identifying, quantifying, and prioritizing the vulnerabilities in a system, with AI tools automating the scanning and analysis of results
11	Phishing Prevention	Security measures designed to stop social engineering attacks that attempt to steal sensitive information, with AI models analyzing email content and sender behavior
12	Security Orchestration, Automation, and Response (SOAR)	A stack of software that allows organizations to collect inputs from security products and define a workflow for automated response
13	Natural Language Processing (NLP)	A branch of AI that gives computers the ability to understand human language, used in security for analyzing threat reports and social media for intelligence
14	Deep Learning (DL)	A subset of ML that uses neural networks with multiple layers (deep neural networks) to analyze complex data, such as network traffic or malware code.

	Key Term	Description
15	Behavioral Analysis	The process of monitoring and analyzing user and entity behavior analytics (UEBA) to detect deviations from a baseline, indicating a potential compromise
16	Zero Trust	A security model based on the principle of "never trust, always verify," where no user or device is trusted by default, regardless of location
17	Cloud Security Posture Management (CSPM)	Tools that identify security risks and compliance violations in cloud environments, often leveraging AI for continuous monitoring.
18	Security Information and Event Management (SIEM)	A system that aggregates and analyzes data from various security devices and applications, with AI enhancing correlation and alerting
19	Extended Detection and Response (XDR)	A unified security incident detection and response platform that automatically collects and correlates data across multiple security layers
20	Red Teaming	The practice of simulating a real-world attack on an organization's security controls to test their effectiveness, with AI assisting in attack path generation

AI Enabled and Enhanced Attack Vectors

Introduction
Objective

This lab covers the fundamental concepts of how artificial intelligence (AI) is weaponized to enhance and enable offensive cyber operations. The table below maps the major concepts discussed in this lab to the corresponding CompTIA SecAI+ (CY0-001) exam objectives.

Task/Major Concept	CompTIA SecAI+ (CY0-001) Objective
Understanding AI-Enabled Attack Vectors (General)	4.2: Explain risks associated with AI
Automated Reconnaissance and Data Correlation	3.1: Given a scenario, use AI tools for security tasks
AI-Enhanced Social Engineering and Deepfakes	1.1: Compare and contrast various types of AI used in cybersecurity
AI-Powered Obfuscation and Polymorphic Code	2.6: Given a scenario, analyze an attack and implement compensating controls
Adversarial Networks (GANs) and Automated Attack Generation	1.1: Compare and contrast various types of AI used in cybersecurity

Overview

The integration of AI and machine learning (ML) has fundamentally reshaped the landscape of offensive cyber operations. AI is no longer just a tool for defense; it has become a powerful enabler for malicious actors, allowing them to automate, accelerate, and personalize attacks with unprecedented sophistication. This lab explores the critical shift in the threat environment, detailing how AI technologies are being weaponized to create and enhance attack vectors.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	Artificial Intelligence (AI)	The simulation of human intelligence processes by machines, especially computer systems, including learning, reasoning, and self-correction
2	Machine Learning (ML)	A subset of AI that allows systems to automatically learn and improve from experience without being explicitly programmed, often through statistical models
3	Attack Vector	A path or means by which a hacker can gain unauthorized access to a computer or network in order to deliver a malicious payload or execute a cyberattack
4	Cyber Kill Chain	A model that describes the stages of a cyberattack, from reconnaissance to actions on the objective, which is often enhanced by AI automation
5	Reconnaissance	The initial phase of a cyberattack where the attacker gathers information about the target to identify vulnerabilities and potential entry points
6	Open-Source Intelligence (OSINT)	The collection and analysis of data gathered from publicly available sources, a process that is heavily automated and correlated by AI in modern attacks
7	Automated Data Correlation	The AI-driven process of linking disparate, seemingly unrelated pieces of information (e.g., social media posts, network logs, public records) to form a complete, actionable intelligence picture of a target
8	Vulnerability Prediction	The use of ML models to analyze a target's system configuration and patch history to predict which unpatched or zero-day vulnerabilities are most likely to succeed
9	Social Engineering	The psychological manipulation of people into performing actions or divulging confidential information, now highly personalized and scaled by AI
10	Spear-Phishing	A highly targeted phishing attempt that is personalized to a specific individual, often leveraging AI-gathered data to increase its credibility and success rate
11	Deepfake	Synthetic media (audio, video, or image) in which a person's likeness or voice is digitally manipulated or entirely generated using deep learning models, often used for impersonation
12	Vishing	A form of social engineering that uses voice communication (often via phone or VoIP) to trick victims; "deepfake vishing" uses AI-generated voice impersonation
13	Obfuscation	The intentional act of creating code or data that is difficult for humans or machines to understand, used by AI to make malware adaptive and evasive
14	Polymorphic Code	Malicious code that changes its internal structure and signature with every execution, making it difficult for signature-based security systems to detect
15	Adversarial Machine Learning	A field of study that explores the vulnerabilities of ML models, often used offensively to craft inputs (adversarial examples) that cause a model to misclassify data
16	Generative Adversarial Network (GAN)	A class of ML framework composed of a generator and a discriminator network, used offensively to rapidly evolve malware or create hyperrealistic

	Key Term	Description
		deepfakes
17	Generator (GAN)	The component of a GAN that creates synthetic data (e.g., new malware samples or deepfake content)
18	Discriminator (GAN)	The component of a GAN that attempts to distinguish between real and synthetic data, effectively acting as a simulated security system for the generator to bypass
19	Adversarial Example	A subtle modification to a legitimate input that is designed to cause a ML model (such as an IDS or spam filter) to make an incorrect prediction or classification
20	Automated Attack Generation	The use of AI to autonomously plan, execute, and adapt a multi-stage cyberattack without human intervention, moving beyond simple automation to dynamic, goal-oriented offensive operations

AI Automation of Security Tasks

Introduction
Objective

Upon completion of this lab, the student will be able to:

- Understand the role of AI and scripting in automating security workflows.
- Apply basic scripting techniques for log analysis and data summarization.
- Simulate the use of AI agents for incident triage and ticket management.
- Automate the security review process for configuration changes.
- Integrate automated security scanning into a continuous integration/continuous deployment (CI/CD) pipeline.

Objective Alignment:

This lab directly addresses the objective: *3.3* Given a scenario, use AI to automate security tasks.

VM Credentials

Username: student

Password: student

Overview

This practical lab is designed to provide hands-on experience with the principles and application of artificial intelligence (AI) in automating common cybersecurity tasks. As the volume and complexity of security data and threats continue to grow, the integration of AI and machine learning (ML) into security operations centers (SOCs) has become essential for efficient and effective defense. This lab will focus on practical scenarios involving scripting, document synthesis, incident response, change management, and CI/CD integration, demonstrating how AI agents and tools can augment human security analysts.

	Key Term	Description
1	Adversarial AI	The use of malicious inputs or techniques to deceive, manipulate, or exploit

	Key Term	Description
		AI models, often to bypass security defenses
2	Agentic AI	An AI system capable of autonomous decision-making and taking actions to achieve defined goals without constant human intervention
3	AI Agent	A software entity powered by artificial intelligence that performs tasks on behalf of a user or system, such as monitoring threats or responding to incidents
4	AI Attacks	Cyberattacks enhanced or automated by AI to increase speed, scale, and sophistication, including adaptive phishing and automated exploitation
5	AI Bias	Systemic errors in an AI model that produce unfair or skewed outcomes, which can lead to inequitable security decisions or false positives
6	AI Bill of Rights	A policy framework outlining principles to protect individuals from harmful AI use, addressing privacy, discrimination, transparency, and accountability
7	AI Cloud Security	The protection of AI workloads, models, and data hosted in cloud environments, safeguarding against data leakage, model theft, and adversarial attacks
8	AI Code Generation	The automated creation of source code using AI models, which requires security reviews to prevent the introduction of vulnerabilities
9	AI Compliance	The adherence of AI systems to relevant laws, regulations, and industry standards, including data protection and governance frameworks
10	AI Explainability (XAI)	The ability to interpret and understand how an AI model makes decisions, which is essential for trust, compliance, and debugging security systems
11	AI Intrusion	Unauthorized access, manipulation, or disruption of AI systems, models, or data pipelines, leading to compromised outputs or service outages
12	AI Jailbreaks	A method to bypass an AI system's safety constraints, enabling it to produce restricted or harmful outputs, a critical concern for AI security testing
13	AI Model	The algorithmic structure trained on data to perform tasks such as detection, classification, or generation, which must be evaluated for robustness
14	AI Red Teaming	A proactive security exercise that simulates adversarial attacks against AI systems to identify vulnerabilities and improve resilience
15	AI Risk Assessment & Management	The process of identifying, evaluating, and mitigating risks related to AI systems, covering model robustness and supply chain vulnerabilities
16	AI SecOps	The integration of AI into security operations to automate threat detection, incident response, and SOC workflows, enhancing speed and accuracy
17	AI Security Guardrails	Predefined safety and compliance boundaries that prevent AI from generating unsafe or unauthorized outputs, essential in enterprise AI deployments
18	AI Threat Hunting	The use of AI tools to proactively search for hidden cyber threats across networks and systems, often detecting patterns missed by human analysts
19	AI Transparency	The practice of making AI system operations, decision-making processes, and limitations understandable to stakeholders, supporting trust and auditing
20	AI TRiSM	Short for trust, risk, and security management in AI, a governance approach to ensure AI systems are safe, reliable, and compliant

AI Governance Structures

Introduction
Objective

This lab is designed to provide foundational knowledge in AI governance, directly supporting the following CompTIA SecAI+ (CY0-001) exam objectives:

Lab Concept/Task	CompTIA SecAI+ (CY0-001) Objective
AI Governance Definition & Importance	4.1: Explain AI governance structures
Three Lines of Defense Model	4.1: Explain AI governance structures
Governing Body & Accountability	4.1: Explain AI governance structures
Core Principles of Responsible AI	4.1: Explain AI governance structures
Model Drift & Algorithmic Bias	4.2: Explain risks associated with AI
AI Governance Best Practices (Monitoring, Audit Trails)	2.5: Given a scenario, implement monitoring and auditing for an AI system
Global Regulatory Landscape (EU AI Act, SR-11-7)	4.3: Explain the impact of compliance on the business use and development of AI
Second Line of Defense (Compliance, Legal)	4.3: Explain the impact of compliance on the business use and development of AI
First Line of Defense (Technical Controls)	2.2: Given a scenario, implement security controls for AI systems

Overview

Artificial intelligence (AI) governance is a critical discipline for organizations seeking to deploy AI systems responsibly, ethically, and legally. As AI technologies become increasingly integrated into core business functions, the need for robust organizational structures and clearly defined roles to manage associated risks has become paramount. This lab provides a comprehensive overview of the foundational concepts, organizational models, and key roles essential for establishing effective AI governance within an enterprise.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	AI Governance	The system of rules, policies, standards, and processes that guides the development, deployment, and monitoring of AI systems to ensure they are safe, ethical, compliant, and aligned with organizational values
2	Responsible AI (RAI)	A holistic approach to developing and deploying AI systems that prioritizes ethical principles, fairness, transparency, accountability, and human oversight
3	Three Lines of Defence Model	An organizational risk management framework adapted for AI, which divides responsibilities among the First Line (system owners), Second Line (governance and compliance), and Third Line (independent assurance/audit)

	Key Term	Description
4	Governing Body	The highest decision-making forum within an organization (e.g., Board of Directors, Executive Committee) ultimately accountable for the outcomes and adequacy of AI governance
5	Model Drift	The degradation of an AI model's performance over time due to changes in the real-world data distribution compared to the training data
6	Bias Control	The process of rigorously examining training data and model outputs to prevent the embedding of real-world biases into AI algorithms, ensuring fair and equitable outcomes
7	Explainability (XAI)	The ability to articulate how an AI system arrived at a particular decision or outcome in terms understandable to humans, crucial for transparency and accountability
8	Accountability	The principle that individuals and organizations must be responsible for the impacts and outcomes of AI systems, requiring clear assignment of roles and oversight
9	AI Ethics Board	A cross-functional committee within an organization responsible for reviewing AI initiatives to ensure alignment with ethical standards, societal values, and internal policies
10	First Line of Defence	The operational management and staff (e.g., product owners, data scientists) responsible for the day-to-day governance and risk management of the AI system
11	Second Line of Defence	Functions (e.g., risk management, compliance, legal) that establish AI governance policies, provide expertise, and monitor the effectiveness of the First Line's risk controls
12	Third Line of Defence	The independent assurance function, typically internal audit, which provides objective evaluation of the effectiveness of the AI governance and risk management framework across the first two lines
13	Formal Governance	The highest level of AI governance, involving a comprehensive, documented framework that aligns with organizational values, principles, and relevant laws and regulations
14	Ad Hoc Governance	A step up from informal governance, involving the development of specific policies and procedures in response to particular AI challenges or risks, often lacking a comprehensive, systematic approach
15	Informal Governance	The least intensive approach, relying primarily on the organization's values and principles with few or no formal structures, processes, or dedicated committees for AI oversight
16	NIST AI Risk Management Framework (AI RMF)	A voluntary framework developed by the US National Institute of Standards and Technology to help organizations manage the risks associated with AI systems
17	EU AI Act	The European Union's comprehensive, risk-based regulatory framework for artificial intelligence, considered the world's first such law
18	Data Governance	The overall management of the availability, usability, integrity, and security of data used in an enterprise, which is foundational for effective AI governance
19	Generative AI	A type of AI that can create new content, such as text, images, or code,

	Key Term	Description
		based on the data it was trained on, posing unique governance challenges
20	SR-11-7	A US regulatory guidance for banks on model risk management, which is often applied to AI models and requires strong governance, validation, and inventory management.

Risks Associated with AI

Introduction

Objective

This lab directly supports the following CompTIA SecAI+ (CY0-001) exam objectives by providing the foundational knowledge necessary to understand, govern, and mitigate the risks associated with AI systems.

Lab Concept/Task	CompTIA SecAI+ (CY0-001) Objective
Understanding the nature of AI risks (Bias, Model Drift, etc.)	4.2: Explain risks associated with AI
Responsible AI (RAI) Framework and Principles	4.1: Explain AI governance structures
NIST AI RMF (Govern, Map, Measure, Manage)	4.1: Explain AI governance structures
Shadow AI and Data Leakage	2.4: Given a scenario, implement data security controls for AI systems
Compliance and Regulatory Violations (GDPR, HIPAA)	4.3: Explain the impact of compliance on the business use and development of AI
Security and Resilience, Adversarial Attacks	1.3: Explain the importance of security in the AI life cycle
Accountability and Transparency	2.5: Given a scenario, implement monitoring and auditing for an AI system

Overview

This theory lab, Risks Associated with AI, is designed to provide a comprehensive understanding of the multifaceted risks inherent in the design, development, deployment, and use of Artificial Intelligence (AI) systems. The learning objective is to explain risks associated with AI, focusing on three critical areas: the principles and practices of responsible AI (RAI), the broad spectrum of AI Risks as defined by leading industry frameworks, and the specific, often hidden, dangers posed by Shadow IT (or Shadow AI). By exploring these topics, learners will gain the knowledge necessary to identify, assess, and mitigate potential harms, ensuring the trustworthy and ethical application of AI technologies in various organizational and societal contexts.

VM Credentials

Username: student

Password: student

	Key Term	Description

	Key Term	Description
1	Responsible AI (RAI)	A framework of principles and practices designed to ensure AI systems are developed and deployed in a manner that is fair, transparent, accountable, safe, and beneficial to society
2	AI Risk Management Framework (AI RMF)	A structured approach, such as the one developed by NIST, for identifying, assessing, and mitigating risks associated with AI systems throughout their life cycle
3	Shadow AI	The use of AI tools, services, or models by employees within an organization without the knowledge, approval, or oversight of the IT, security, or governance teams
4	Data Leakage	The unauthorized transmission of sensitive or confidential data from within an organization to an external destination, often a risk when using unapproved AI tools
5	Bias	Systematic error in an AI system's output due to flawed assumptions in the machine learning process, often stemming from unrepresentative or prejudiced training data
6	Fairness	A principle of RAI ensuring that AI systems do not perpetuate or amplify unjust or discriminatory outcomes against individuals or groups.
7	Accountability	The principle that organizations and individuals responsible for the design, development, and deployment of AI systems can be held responsible for their outcomes and impacts
8	Transparency	The degree to which an AI system's inner workings, data, and decision-making processes are understandable and accessible to relevant stakeholders
9	Explainability	The ability to articulate how an AI system arrived at a particular output or decision, which is crucial for building trust and enabling effective risk management
10	Model Drift	The phenomenon where the performance or accuracy of an AI model degrades over time due to changes in the real-world data it processes
11	Adversarial Attack	Maliciously crafted input data designed to intentionally cause an AI model to make an incorrect classification or decision
12	Systemic Risk	Risks that can cascade across multiple systems, sectors, or society, often associated with the widespread deployment of a single, flawed AI model.
13	P-FER	The four core functions of the NIST AI RMF: govern, map, measure, and manage
14	Governance	The organizational structures, policies, and processes put in place to direct and control the development and use of AI systems
15	Validation	The process of ensuring that an AI system meets the needs of its users and stakeholders in its intended operational environment
16	Resilience	The ability of an AI system to maintain its function and integrity despite internal or external disturbances, such as cyberattacks or data corruption
17	Inadvertent Leakage	The unintentional exposure of sensitive data when employees input proprietary information into public, unmonitored AI services

	Key Term	Description
18	Data Sovereignty	The concept that data is subject to the laws and governance structures of the nation in which it is collected and stored, a key concern with Shadow AI
19	Model Card	A short document accompanying a trained machine learning model that provides benchmarked evaluation metrics, intended uses, and ethical considerations
20	Human-Centric AI	An approach to AI development that prioritizes human values, well-being, and control, ensuring the technology serves human needs

The Impact of Compliance on Business Use and Development of AI

Introduction
Objective

This lab provides a theoretical foundation for understanding the critical impact of compliance on the business use and development of AI. The concepts covered directly align with the following CompTIA SecAI+ (CY0-001) exam objectives:

Lab Concept/Section	CompTIA SecAI+ (CY0-001) Objective
Introduction & Regulatory Landscape	4.3: Explain the impact of compliance on the business use and development of AI
EU AI Act Risk Categories	4.2: Explain risks associated with AI
OECD Principles (Transparency, Explainability)	1.1: Compare and contrast various types of AI used in cybersecurity
ISO/IEC 42001 (AIMS, Governance)	4.1: Explain AI governance structures
NIST AIRMF (Govern, Map, Measure, Manage)	4.1: Explain AI governance structures
Corporate Policies (Data Governance, Quality)	1.2: Explain the importance of data security as it relates to AI
Corporate Policies (Documentation, Life cycle)	1.3: Explain the importance of security in the AI life cycle
Third-Party Evaluations (Audits)	4.1: Explain AI governance structures
Data Sovereignty & Localization	1.2: Explain the importance of data security as it relates to AI
Data Sovereignty & Localization	4.2: Explain risks associated with AI

Overview

The rapid advancement of artificial intelligence (AI) has ushered in a new era of technological capability, offering unprecedented opportunities for business innovation, efficiency, and growth. However, this transformative power is not without risk. The deployment of AI systems, particularly those that interact with sensitive data or make decisions impacting human lives, introduces complex ethical, legal, and societal challenges. Consequently, a global consensus has emerged on the necessity of robust AI compliance—a framework of laws, regulations, standards, and internal policies designed to ensure that AI systems are developed and used in a trustworthy, transparent, and responsible manner.

AI compliance is no longer a peripheral concern; it is a core strategic imperative that fundamentally impacts the business use and development life cycle of AI. For businesses, compliance dictates everything from the initial design choices of an AI model to its final deployment and ongoing monitoring. Noncompliance carries severe consequences, including massive financial penalties, reputational damage, loss of consumer trust, and legal liabilities. For developers, compliance translates into concrete technical requirements, such as ensuring data quality, documenting system logic, conducting rigorous risk assessments, and implementing mechanisms for human oversight. This lab will summarize the profound impact of these compliance requirements, examining key global frameworks and the critical role of internal governance and data sovereignty.

VM Credentials

Username: student

Password: student

	Key Term	Description
1	AI Compliance	A comprehensive framework of laws, regulations, standards, and internal policies designed to ensure that AI systems are developed and used in a trustworthy, transparent, and responsible manner
2	EU AI Act	The world's first comprehensive legal framework for AI, established by the European Union (EU), which employs a risk-based approach to regulation
3	Risk-Based Approach	A regulatory strategy, central to the EU AI Act, that categorizes AI systems based on their potential to cause harm, applying stricter rules to higher-risk systems
4	High-Risk AI	AI systems used in critical areas such as employment, credit scoring, law enforcement, and critical infrastructure, which are subject to the most stringent compliance requirements under the EU AI Act
5	Unacceptable Risk AI	AI systems that pose a clear threat to fundamental rights (e.g., social scoring), which are explicitly prohibited by the EU AI Act
6	Conformity Assessment	A mandatory procedure under the EU AI Act for high-risk systems, requiring providers to prove that their AI system meets all legal requirements before being placed on the market
7	Compliance-by-Design	A development methodology where regulatory and compliance requirements are integrated into the AI system's design and development process from the very first stage
8	Quality Management System (QMS)	A formalized system that documents processes, procedures, and responsibilities for achieving quality policies and objectives, mandated for high-risk AI under the EU AI Act
9	OECD Principles on AI	The first intergovernmental standard for AI, providing a nonbinding ethical and moral compass for responsible AI development, adopted by over 40 countries
10	Explainable AI (XAI)	The ability to make the logic, process, and decisions of an AI system understandable and interpretable to human users, a key requirement in many compliance frameworks
11	ISO/IEC 42001	The first international standard for an artificial intelligence management system (AIMS), providing a framework for managing AI-related risks and opportunities

	Key Term	Description
12	AI Management System (AIMS)	A system of processes and procedures for an organization to establish, implement, maintain, and continually improve its management of AI-related risks and opportunities, as defined by ISO/IEC 42001
13	NIST AI Risk Management Framework (AI RMF)	A nonregulatory, voluntary framework from the US National Institute of Standards and Technology designed to improve the trustworthiness and responsible use of AI systems
14	Data Sovereignty	The concept that data is subject to the laws and governance structures of the nation in which it is collected and processed, impacting where and how AI models can be trained and deployed
15	Data Localization	Regulatory requirements in certain jurisdictions that mandate that specific types of data must be stored and processed exclusively within the national borders of that country
16	Corporate Policies	Internal rules and procedures created by a company to translate external laws and standards (like the EU AI Act or NIST AI RMF) into actionable, company-specific compliance steps
17	Third-Party Compliance Evaluation	An external assessment, often referred to as an AI audit, conducted by an independent body to objectively validate an AI system's adherence to regulatory standards and best practices
18	Responsible AI Principles	High-level ethical guidelines adopted by a corporation defining its fundamental stance on fairness, transparency, accountability, and human oversight in AI development and deployment
19	Federated Learning	A machine learning technique that trains an algorithm across multiple decentralized edge devices or servers holding local data samples, without exchanging the raw data itself, often used to address data sovereignty concerns
20	AI Governance	The system of rules, practices, and processes by which an organization manages its AI activities to ensure accountability, transparency, ethical outcomes, and regulatory compliance