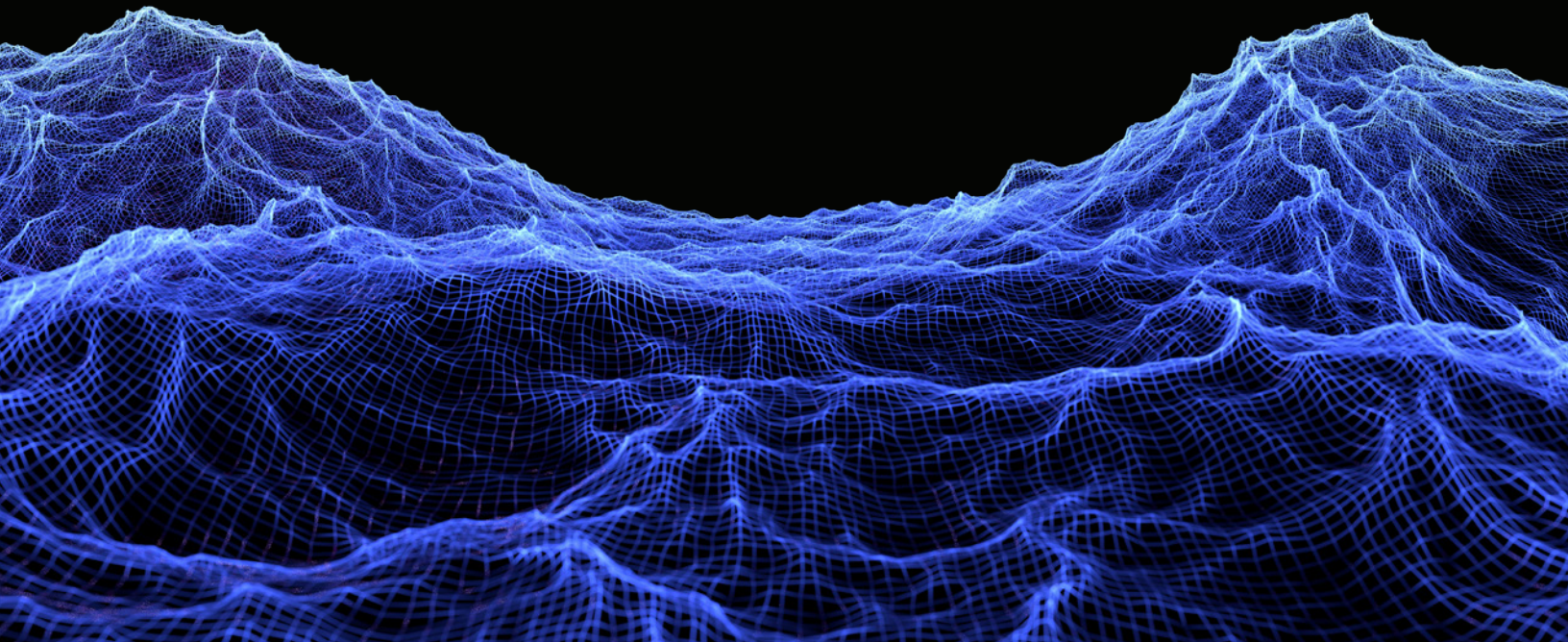


Peraton Labs

SECURITY AND RISK ASSESSMENT OF LARGE LANGUAGE MODEL (LLM)-BASED AI SYSTEMS

Assessing the security risks of LLM-based AI enterprise systems requires extension of classic cybersecurity controls and penetration testing to evaluate AI interfaces, integration, model instructions, and guardrails through a white-box threat analysis approach with adversarial testing.



BACKGROUND

Artificial intelligence (AI) is a transformative new technology that is quickly establishing a presence in enterprise computing infrastructures. From providing human-like assistance to autonomously acting and automating enterprise tasks, AI is powering a wide variety of applications across business, government, and critical infrastructure operations:

- **Energy and utilities:** Grid monitoring, demand forecasting, and customer support
- **Healthcare:** Imaging, diagnostics, and patient scheduling
- **Defense and national security:** Real-time threat detection, and autonomous operations
- **Transportation:** Autonomous vehicles, route planning, and safety optimization
- **Finance:** Fraud detection, credit scoring, and market analytics
- **Telecommunications:** Outage prediction and bandwidth management

Introduced as a productivity tool or autonomous agent, AI-based systems present novel cybersecurity risks and their own vulnerability surface that extends beyond traditional information technology (IT) controls defenses.

How AI Transforms the Security Landscape

Many AI implementations leverage LLMs, which analyze unstructured data and generate human-like responses.

These models, typically sourced from third-party vendors due to the high computational cost for their creation, come with pre-trained knowledge that is fine-tuned, augmented with proprietary data, and constrained for enterprise applications.

A typical LLM-based enterprise application system architecture, shown in figure 1, introduces new computing components that need to be individually analyzed to understand system security and risk.

- **Model user interfaces and application programming interfaces (API):** Determines how users and systems interact with LLM-enabled applications
- **LLM context and memory stores:** Models instructions to facilitate dialogue flow, remain on topic, shape responses, and take user input
- **Retrieval-augmented generation (RAG):** Integrates LLMs with enterprise knowledge stores and dynamic data to assist LLMs reasoning a response
- **Input and output guardrails:** Implements policy restrictions on input questions and output responses to prevent data loss, privacy violations, and inappropriate responses
- **Model tuning pipelines:** Alters model behavior for enterprise application-specific purposes
- **Agentic actions:** Allows LLMs to perform real-world tasks (e.g., ticketing, device control), which can be exploited if the LLM is deceived

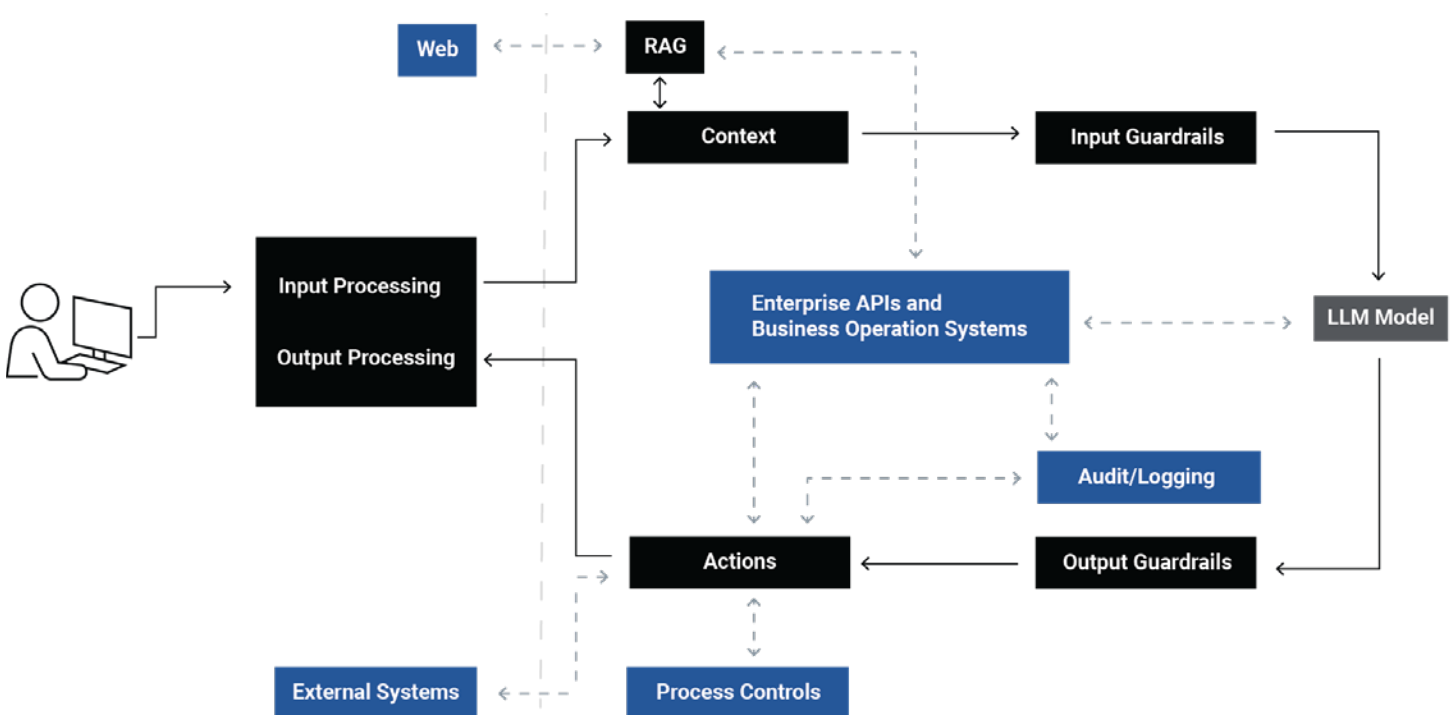


Figure 1 Typical LLM-based Enterprise System Architecture

REAL-WORLD AI THREATS

Attackers use a variety of deception, input injection, context manipulation, language ambiguities, and data poisoning techniques to alter and compromise AI behavior. Threats include evading safety guardrails, triggering unauthorized actions, and corrupting outputs that can result in serious consequences:

- **Privacy breaches:** Exposes personally identifiable information (PII), sensitive, or proprietary data
- **Operational disruption:** Incorrect outputs that mislead users or disrupt business processes
- **Security compromise:** Unauthorized access, privilege escalation, or automated malicious activity

Because AI systems operate above the plane of traditional IT cybersecurity controls for user, system, network, and application security controls, a new methodology is needed to assess their vulnerabilities and risks.

PERATON LABS AI SECURITY ASSESSMENT METHODOLOGY

As a leader in AI cybersecurity research, Peraton Labs innovated a new white-box AI testing approach that starts with a comprehensive review of the AI system architecture and then decomposes AI systems into functional components for independent evaluation. The architecture review identifies adversary goals, high-risk assets, and potential attack strategies that establish objectives for validation and offensive testing. The architecture review also identifies security and safety strategy, security enforcement points, design-driven controls, and monitoring and logging capabilities for process visibility. Functional component analysis evaluates the controls and policy implementations based on component role to help identify sources of weakness within the AI system. Offensive testing then attempts to validate and circumvent controls, deceive, and compromise the AI system. Peraton Labs applies an iterative testing approach with successive refinement that employs adversarial LLMs, dialogue review, and crafted input informed by policy controls. Finally, Peraton Labs conducts traditional IT penetration testing on the computing resources and networks used by AI systems.

Our methodology:

- Builds on and extends foundational IT/operations technology (OT) security assessment practices to AI systems
- Assesses weaknesses within AI systems not only on an input/output level, but at the functional component level
- Analyzes key AI system components including architecture, memory, retrieval, and enterprise interfaces
- Applies layered testing to uncover AI-native risks

Architecture Assessment

Peraton Labs performs an in-depth analysis of the AI system architecture and its integration within broader IT infrastructure and business workflows. This assessment includes:

- **System architecture:** End-to-end analysis of infrastructure, including model pipelines, APIs, RAG, and Model Context Protocol (MCP)
- **Hosting environment:** Determination of hosting model—on-premises, cloud-based, or accessed via third-party APIs
- **Network topology:** Mapping of data flows within the AI system and to/from external sources
- **Use cases:** Analysis of user and system interactions with the AI, including input/output interfaces, and business workflow integrations
- **Policy and governance:** Review of AI guardrail and policy definitions
- **System capabilities:** Assessment of the AI's functional scope, such as chat interfaces, prediction, and information retrieval
- **Data sources:** Identification of underlying training data, live data feeds, and knowledge bases
- **Agentic behaviors:** Evaluation of the AI's ability to invoke actions, access systems, or operate autonomously
- **Logging and monitoring:** Review of logging practices and the system's ability to detect and record anomalous activity

Threat Modeling and Risk Identification

Threat modeling is a critical step to proactively identify

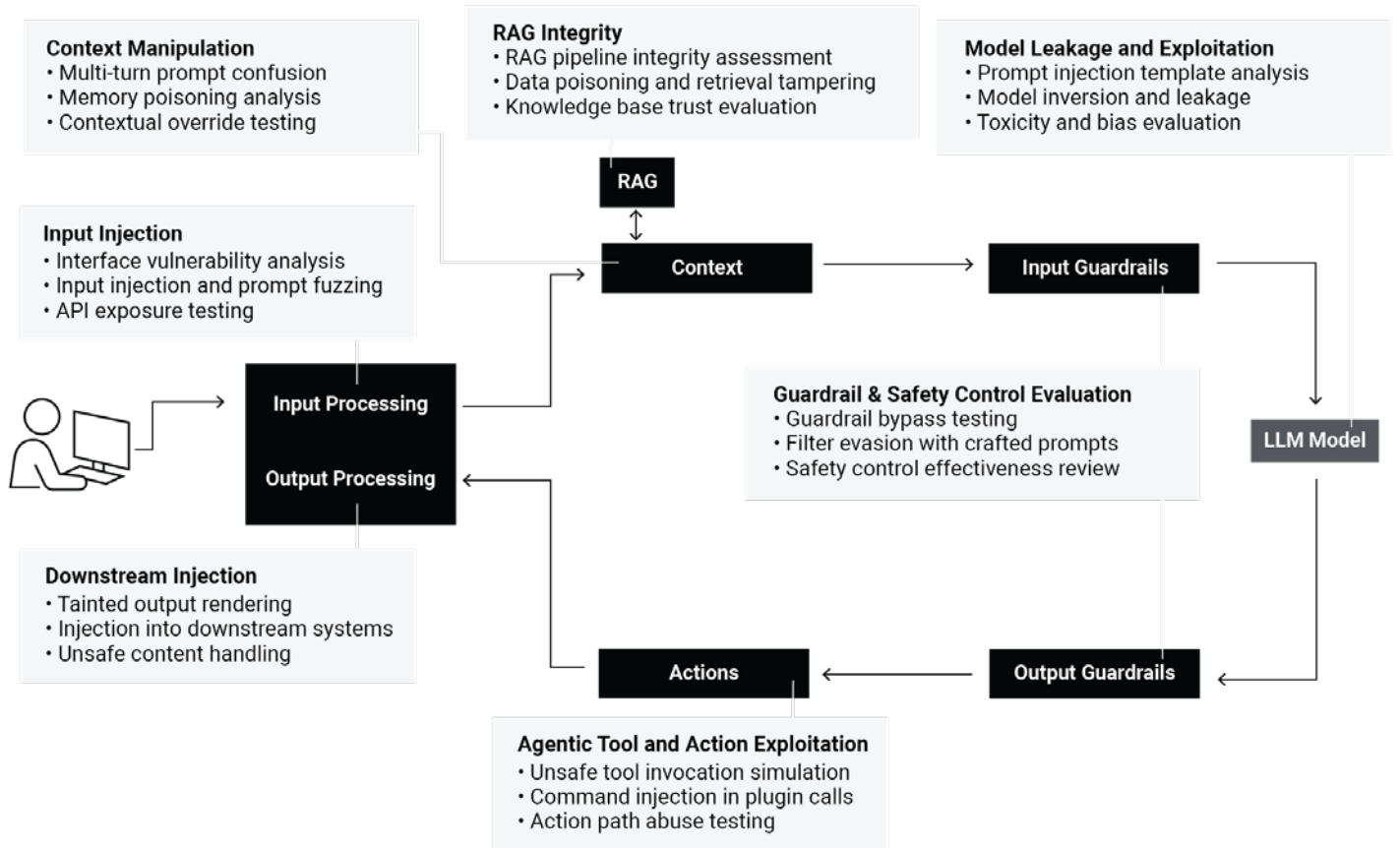


Figure 2 Peraton's Systematic Testing of AI Architecture Components

adversary goals, high-value assets, and the threats that pose the greatest business risks. Guided by architectural insights, Peraton Labs develops a threat model that informs offensive testing objectives. This phase includes:

- **Key assets:** Identification of sensitive elements in training data, prompt templates, system prompts, APIs, and credentials
- **Attacker goals:** Analysis of potential adversary motivations, including theft, misuse, misinformation, and system manipulation
- **Threats:** Mapping of threats using established frameworks such as MITRE ATLAS™ and Open Web Application Security Project (OWASP™) top 10 for LLMs
- **Attack paths:** Definition of realistic attack scenarios impacting confidentiality, integrity, and availability

The result is a set of clearly defined testing objectives to validate the effectiveness and attempt to circumvent security controls, policies, and guardrails in the hands-on testing phase.

VULNERABILITY ASSESSMENT

Test AI Components

Peraton Labs employs a layered, defense-in-depth methodology to assess AI components—combining traditional cybersecurity practices with advanced adversarial AI techniques. This comprehensive approach is designed to target vulnerabilities unique to LLMs and their integrations within enterprise workflows.

During the assessment, Peraton Labs uses an iterative, behavior-driven process to test each component. Inputs are systematically crafted and refined based on observed outputs, enabling validation of implemented controls, and identification of failure points. The primary goal is to uncover root sources of system weaknesses, where they are not adequately being handled or prevented later than anticipated in the LLM application and recommend effective mitigation strategies.

Peraton Labs tests AI systems with AI technology, combining adversarial LLMs with targeted manual testing. At a time of rapid technology evolution, Peraton Labs continually evaluates and integrates new tools.

Figure 2 and the following subsections describe how the principal components of an AI system are tested.

Input Processing

Peraton Labs evaluates how input is processed and sanitized by reviewing code and configurations, testing for input validation weaknesses, and attempting to bypass safeguards through malformed data input. We identify inputs that could propagate across multiple system layers, as these may increase the risk of injection vulnerabilities.

Context Handling

Peraton Labs analyzes how session state and conversational history are stored and managed, identifying potential leakage across sessions or users. We test for unintended exposure of sensitive information and attempt to manipulate context to produce unintended behavior.

RAG

Peraton Labs maps the RAG architecture, identifying APIs and implemented protections. We inject direct API calls to test for unauthorized data access or content injection and attempt to replicate these actions through the user interface to evaluate end-to-end protections.

Guardrails

Peraton Labs assesses guardrail effectiveness by reviewing policy definitions, using adversarial AI tools (such as Python Risk Identification Tool (PyRIT) and garak (Generative AI Red-Teaming and Assessment Kit) to generate test inputs, and monitoring application outputs and system logs to identify which inputs are blocked or permitted. These insights are used to iteratively refine further input to probe for filter bypasses.

LLM Model

Peraton Labs analyzes how the model is trained, fine-tuned, and updated, and attempts to perform unauthorized modifications. We evaluate the model's susceptibility to prompt injection and jailbreak attacks by querying it with adversarial prompts. These inputs are iteratively refined to bypass safeguards and extract sensitive or restricted information.

Actions

Peraton Labs examines how the AI system interfaces with external tools and APIs, identifying API parameters, access controls, and permissions. We attempt unauthorized API calls both directly and through carefully crafted prompts via the user interface. We identify weaknesses that may expose the system to vulnerabilities such as remote code execution.

Output Processing

Peraton Labs assesses how outputs are post-processed before being presented to end users and downstream systems. We identify and evaluate the filters and formatters in place and attempt to craft adversarial outputs to bypass these post-processing controls.

Test AI Computing Infrastructure

AI systems rely on underlying IT infrastructure that shares the same security risks as traditional IT environments. Peraton Labs conducts a comprehensive assessment of the supporting systems and networks, including:

- Identity management and authentication
- Permissions and access management
- Firewall policy and network segmentation
- Network architecture and external exposure
- Data encryption, at rest and in transit
- Authentication and access control
- Web, API, and interface vulnerabilities
- DevOps continuous integration (CI)/continuous delivery (CD) pipeline risks

INTEGRATED ANALYSIS OF IT AND AI SYSTEMS WEAKNESSES

Peraton Labs combines traditional IT findings with AI-specific vulnerabilities to simulate hybrid attack scenarios, such as:

- **Data poisoning:** Exploiting system access to training data to inject misleading data or malicious content.
- **Toolchain exploits:** Using knowledge of internal APIs to inject malicious code into AI-generated scripts, enabling unauthorized remote access.
- **Privilege escalation:** Exploiting misconfigurations or overly permissive API calls to escalate privileges and gain administrative control of the AI system.
- **Data exfiltration:** Combining access to backend systems with AI prompt manipulation to extract sensitive information.

Peraton Labs documents findings, an interpretation of risk and impact, and provides recommendations resulting from the security assessment in a test assessment report. The report pulls individual findings into larger threats to assess big-picture risk and the combined risk of AI and IT weaknesses.

Using application knowledge and the assessment findings, Peraton Labs constructs an adversarial attack tree to model and analyze potential vulnerabilities and attack paths within the AI system to help identify weaknesses, understand how attackers might exploit them, and develop countermeasures to mitigate risks. This requires identifying:

- **Root goals:** Attacker's ultimate objectives (e.g., extract PII, gain system access)
- **Branches:** Logical steps an attacker can take to achieve intermediate goals
- **Breakpoints:** Defense-in-depth locations where a mitigation can be introduced to break the chain (e.g., input validation, output filtering, role-based access.)

Our report documents the artifacts and details of each finding, which are assigned a severity level based on a risk assessment model customized to our customer's environment using a four-tier rating system. We also include the NIST Common Vulnerability Scoring System version 4.0 (CVSSv4.0) score for the findings.

Our findings and recommendations not only provide the technical details but also the implications and risks to the enterprise. We focus on identifying the root cause of issues, not just technical symptoms, to help management make better and more informed decisions about how to focus their remediation plans and budgets. Our findings and recommendations are prioritized by their severity, potential impact, and risk to the enterprise. The findings recognize that adversaries will attempt to string together several lower severity vulnerabilities to create more divisive attacks.

WHY PERATON LABS?

Peraton Labs brings decades of experience in cybersecurity, vulnerability assessment, and advanced threat analysis to the fast-evolving AI domain. Leveraging our industry-recognized strength of helping customers secure new technology for four decades, our team uniquely applies a research discipline to not only practice the art but advance the science. Our expertise is grounded in industry-leading AI cybersecurity research and development in high-impact government programs, including:

- Finding trojans in AI
- Extensible catalog of adversarial machine learning
- Poisoning attack on ML in the cyber domain
- Disruptive techniques for machine vision
- AI defenses against adversarial ML attacks
- Radio frequency (RF) application lifecycle management—disrupting reactive jamming
- Protecting machine learning (ML) from privacy leakages

Whether securing traditional infrastructure, LLMs, or autonomous AI agents, Peraton Labs delivers actionable insight to help organizations stay ahead of evolving threats.

Resources

¹ ATLAS is a trademark of The MITRE Corporation

OWASP is a trademark of the trademark of the OWASP Foundation, Inc.

ABOUT PERATON

Peraton is a next-generation national security company that drives missions of consequence spanning the globe and extending to the farthest reaches of the galaxy. As one of the world's leading mission capability integrator and transformative enterprise IT provider, we deliver trusted, highly differentiated solutions and technologies to protect our nation and allies from threats across the digital and physical domains. Peraton supports every branch of the U.S. Armed Forces, and we serve as a valued partner to essential government agencies that sustain our way of life. Every day, our employees do the can't be done by solving the most daunting challenges facing our customers. Visit peraton.com to learn how we're safeguarding your peace of mind.