SecureIQLab

Commissioned Report

F5 AI Guardrails Security Efficacy Validation Test

Q1 2026

Last Revision: 19 March, 2026

Language: English

Table of Contents

1 EXECUTIVE SUMMARY 2
2 INTRODUCTION 2
2.1 F5 AI GUARDRAILS SOLUTION 3
2.2 PURPOSE 3
3 TEST ARCHITECTURE 3
4 TEST SCOPE AND OBJECTIVES 4
5 VALIDATION METHODOLOGY 5
5.1 TEST ENVIRONMENT PREPARATION 5
5.2 BASELINE CONFIGURATION 6
5.3 INITIAL SMOKE TESTING 6
5.4 ADVERSARIAL SECURITY TESTING 6
6 SECURITY VALIDATION MAPPING AND SCORING 7
6.1 SCORING AND SCORECARD COMPILATION 7
6.2 PROMPT INJECTION 8
6.3 INDIRECT INJECTION 9
6.4 SENSITIVE DATA LEAK 9
6.5 INPUT LEAKAGE 10
6.6 UNSAFE OUTPUTS 10
6.7 EXCESSIVE AGENCY 11
6.8 SYSTEM PROMPT LEAKAGE 12
6.9 MODEL EXTRACTION 12
6.10 HARMFUL CONTENT AND BIAS 13
6.11 TOXIC OUTPUT 14
7 F5 GUARDRAILS DIFFERENTIATORS 14
7.1 FLEXIBLE CUSTOM SCANNER FRAMEWORK 14
7.2 SCANNER VERSIONING 15
7.3 NATIVE SIEM INTEGRATION (SPLUNK) 15
8 CONCLUSION 15
9 COPYRIGHT AND DISCLAIMER 16

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

1 EXECUTIVE SUMMARY

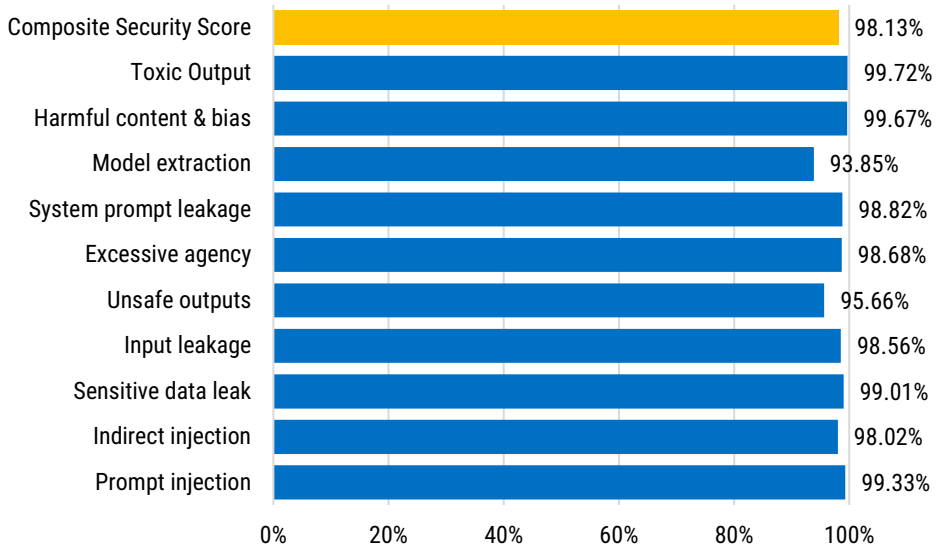


Figure 1. Composite and Detailed Category Block Rate Results

F5 commissioned SecureIQLab to test the ability of the F5 AI Guardrails-based cloud solution to detect and prevent threats targeted towards GenAI, both models and applications that use GenAI components. Figure 1 provides SecureIQLab's findings. The higher the score, the better. The test methodology measured the block rate of the tested GenAI solution against attack scenarios that were both inbound (prompt and context filtering) and outbound (response filtering). The Composite Security Score was calculated by averaging the block scores for each of the ten attack categories validated. Nearly 20,000 attack payloads were used during testing in conjunction with benign false positive traffic. The volume and quality of samples highlight F5's strong ability to secure GenAI applications within a representative full-stack environment that extends beyond simple traffic simulation to include real applications, RAG pipelines, enterprise data, and integrated security workflows under realistic adversarial and operational conditions.

This report represents the first independent industry validation designed specifically to measure the effectiveness of AI security technologies against adversarial GenAI attack techniques. These results demonstrate F5 Guardrails' protection against real-world AI-targeted attack scenarios. From these results, SecureIQLab concludes F5's approach against tested GenAI threat provides superior protection.

2 INTRODUCTION

Generative Artificial Intelligence (GenAI) systems, particularly Large Language Models (LLMs), are rapidly becoming integral components of enterprise technology environments. Organizations increasingly deploy these systems within customer support platforms, enterprise applications, software development workflows, and decision-support tools. By processing natural language inputs and generating contextually relevant outputs, these models enable automation, knowledge access, and improved operational efficiency. In many deployments, GenAI systems also interact with external tools, internal APIs, proprietary data sources, and automated workflows to perform complex tasks.

GenAI guardrails are the controls that keep Generative artificial intelligence (GenAI) systems operating safely, responsibly and within defined boundaries. Guardrails define trust boundaries between user and GenAI system.

While these capabilities deliver significant business value, they simultaneously introduce a new and evolving attack surface that differs fundamentally from traditional application architectures. Unlike conventional web applications

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

that operate on structured inputs and deterministic logic, GenAI systems process unstructured natural language prompts, dynamically generate responses, and may initiate actions or retrieve data based on interpreted user intent.

This dynamic and probabilistic behavior introduces unique security risks that are not adequately addressed by traditional security controls alone. Emerging threats include prompt injection attacks, unauthorized data access or exfiltration, insecure output handling, and unintended model behavior or misuse. As GenAI systems become more tightly integrated with enterprise systems and sensitive data sources, the potential impact of these vulnerabilities increases significantly.

Consequently, organizations must extend beyond traditional perimeter and application security mechanisms to adopt specialized security controls and validation methodologies designed specifically for AI-enabled systems. These controls must address the unique interaction patterns, trust boundaries, and operational risks introduced by GenAI deployments.

2.1 F5 AI GUARDRAILS SOLUTION

F5 AI Guardrails is an enterprise AI security solution that provides runtime protection, governance, and policy enforcement for generative AI models, applications, and autonomous agents. It is designed to ensure that AI systems operate securely, safely, and in alignment with organizational security policies and regulatory requirements.

The platform functions as a security control layer surrounding AI systems, monitoring and filtering interactions between users, applications, AI models, and connected services or data sources. By inspecting both incoming prompts and outgoing model responses, F5 AI Guardrails helps prevent security risks such as sensitive data exposure, adversarial attacks (including prompt injection and jailbreak attempts), harmful or non-compliant outputs, and violations of defined security policies.

2.2 PURPOSE

The purpose of the F5 AI Guardrails Security Efficacy Validation is to systematically evaluate the effectiveness of F5 AI Guardrails in protecting generative AI systems from security threats, misuse, and policy violations during runtime operations. This validation methodology assesses whether the security controls provided by F5 AI Guardrails can successfully detect, prevent, and mitigate risks associated with real-world AI attack scenarios and unsafe AI behaviors.

The validation focuses on measuring the ability of the solution to enforce security policies and protect AI interactions across inputs, outputs, and AI-driven actions. This includes testing its effectiveness in identifying and blocking threats such as prompt injection attacks, jailbreak attempts, adversarial prompts, sensitive data exposure, and harmful or non-compliant outputs generated by AI models.

Overall, this security efficacy validation provides an objective assessment of how effectively F5 AI Guardrails reduces the attack surface of generative AI systems and safeguards enterprise AI deployments against emerging AI-specific threats.

3 TEST ARCHITECTURE

F5 AI Guardrail solution is available as a SaaS, which can be deployed either through an API gateway, through an SDK, or REST interface. Typical deployments will vary based on customers' specific integration. For testing purposes, SecureIQLab utilized an SDK deployment to facilitate volumetric test harness. The test harness leveraged build usage pattern to be protected by F5 AI Guardrail solution. Figure 2. below provides a high-level architecture for the test deployment.

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

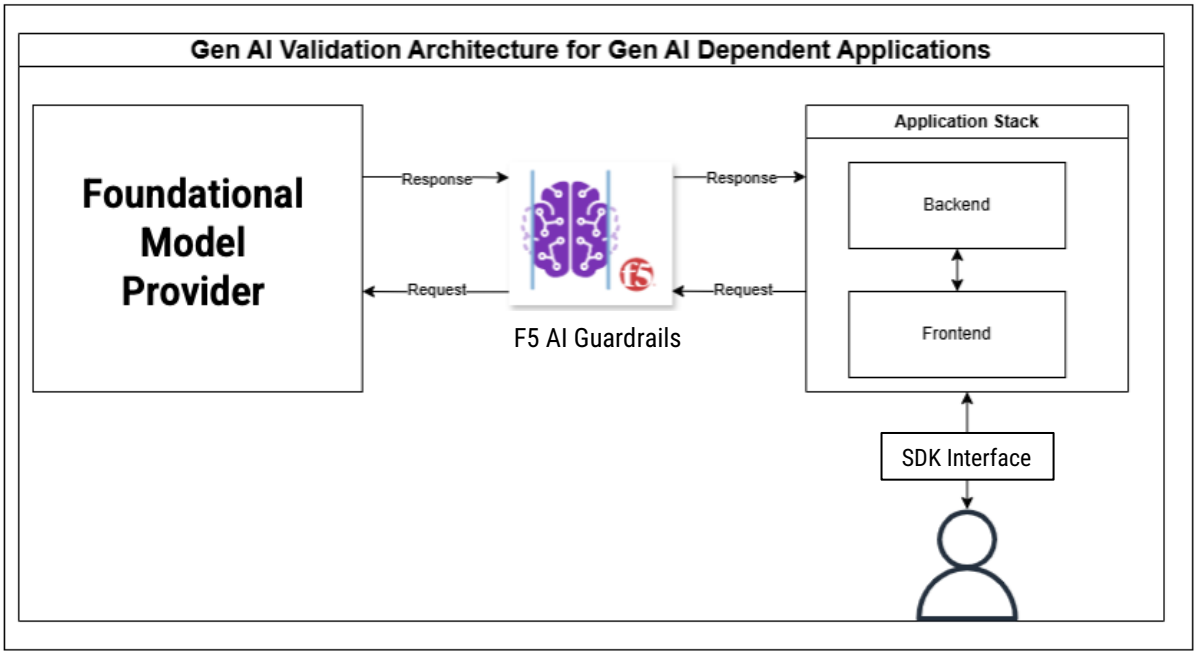


Figure 2. Gen AI Validation Test Architecture

The Validation Architecture includes validation steps at theF5 AI Guardrails solution layer:

- Step 1: Outbound Validation (Prompt & Context Filtering)
  - Backend sends tokens, query, and enhanced context to F5 AI Guardrails instead of directly to the LLM
  - F5 AI Guardrails inspects the outbound payload for:
    - PII and sensitive data leakage (e.g., names, credit cards, confidential context)
    - Prompt injection attempts designed to manipulate model behavior
  - Only validated requests are forwarded to the LLM as validated tokens, query, and enhanced context
- Step 2: Inbound Validation (Response Filtering)
  - LLM response is routed to F5 AI Guardrails before reaching the backend
  - F5 AI Guardrails inspects the response for:
    - PII and Sensitive information Leakage
    - Hallucinations/Toxicity (Ensuring the response is safe, on-topic, and non-toxic.)
    - Malicious code or executable code generation
  - Filtered response is forwarded to the backend for safe delivery to the end user

4 TEST SCOPE AND OBJECTIVES

F5 AI Guardrails is intended to act as a bidirectional security gateway between AI applications (chatbots, copilots, RAG pipelines) and the underlying models it inspects:

- Inbound: user prompts, tool inputs, retrieved documents
- Outbound: model responses, tool outputs

This is described as runtime protection at the inference layer, meaning the F5 AI Guardrails secures the moment when a model interacts with live user data and generates responses.

The scope of this validation is set forth in the F5 AI Guardrails Security Efficacy Validation Methodology and is intentionally focused on security control effectiveness, specifically evaluating how well F5 AI Guardrails protects

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

generative AI systems during runtime interactions. Clear testing boundaries ensure that the evaluation process remains objective, controlled, and repeatable.

The validation is strictly focused on security efficacy, measuring how effectively F5 AI Guardrails enforces runtime protection and governance controls across common AI attack scenarios and misuse patterns. Subjects that are covered and not covered by this test are listed below:

In Scope	Out of Scope
• Prompt-Injection • Indirect Injection • Sensitive Data Leak • Input Leakage • Unsafe Outputs • Excessive Agency • System Prompt Leakage • Model Extraction • Harmful Content Bias • Toxic Output	• GenAI Governance • Data Pipeline Security • Operational Efficiency • Performance Efficiency • Compliance Validation • Underlying Model Security • Model Alignment and Safety Training • Infrastructure Security • Application-Specific Business Logic

5 VALIDATION METHODOLOGY

The F5 AI Guardrails Security Efficacy Validation Methodology focuses on assessing how effectively the solution protects generative AI applications during runtime by enforcing policies, preventing malicious inputs, controlling unsafe outputs, and safeguarding sensitive data.

The validation process evaluates the runtime protection and threat mitigation capabilities of F5 AI Guardrails using a combination of automated security testing, human-led adversarial testing aka GenAI red-teaming, and enterprise application workflow validation.

5.1 TEST ENVIRONMENT PREPARATION

A controlled and production-representative test environment is established to ensure that validation results accurately reflect real-world enterprise AI deployments where F5 AI Guardrails is used to protect generative AI applications.

Environment preparation includes:

- Deployment of foundational model provider powered LLM-enabled enterprise applications protected by F5 AI Guardrails
- Configuration of representative enterprise datasets, including:
    - Public documents and knowledge sources
    - Internal enterprise-style documentation
    - Synthetic sensitive data and PII records
    - Canary values designed to detect unauthorized data leakage
- Integration with simulated enterprise infrastructure, including:
    - Identity and access management systems
        - Note: Since IDP sits outside the main traffic path, its integration could be considered outside the core request and response flow
    - Logging and monitoring platforms
    - Security telemetry and alerting systems
- Isolation of the test environment to ensure controlled and repeatable evaluation conditions

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

All data used during testing is synthetic, anonymized, or purpose-built, ensuring that no real customer or sensitive operational data is exposed during the validation process.

5.2 BASELINE CONFIGURATION

F5 AI Guardrails is deployed using vendor-recommended baseline configurations to reflect how organizations typically implement the solution in production environments.

Baseline configuration principles include:

- Enablement of all generally available F5 AI Guardrails runtime protection capabilities relevant to the defined validation scope
- Configuration of policies related to:
  - Prompt injection protection
  - Data protection and leakage prevention
  - Output filtering and safety controls
  - Governance and policy enforcement
- Alignment with publicly available vendor documentation and deployment guidance.
- Verification of successful deployment through functional checks prior to adversarial testing

Any configuration adjustments required for the testing environment are documented to ensure transparency and reproducibility. For our testing, we utilized standard deployment settings with minimal tuning.

5.3 INITIAL SMOKE TESTING

Before conducting adversarial security testing an initial validation phase confirms that F5 AI Guardrails is properly integrated and functioning as expected.

Smoke testing verifies:

- Successful interception and inspection of LLM inputs and prompts
- Monitoring and filtering of model outputs
- Enforcement of basic policy controls for high-risk scenarios such as:
  - Simple prompt injection attempts
  - Direct sensitive data requests
- Generation of logs, alerts, and telemetry for AI interactions
- System stability under normal operational conditions

This phase ensures the environment is operational before executing advanced attack simulations.

5.4 ADVERSARIAL SECURITY TESTING

The core validation phase involves structured adversarial testing designed to simulate realistic AI attack scenarios targeting generative AI systems protected by F5 AI Guardrails.

Testing activities are aligned with the defined validation scope, which includes:

- Prompt-Injection: Attempts to manipulate the model's behavior by embedding malicious or adversarial instructions within user prompts.
- Indirect Injection: Injection of malicious instructions through external or retrieved content sources (e.g., documents, web data, or plugins) rather than direct user input.

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

- Sensitive Data Leak: Exposure of confidential or restricted information through model responses, including training data, system data, or user inputs.
- Input Leakage: Unintended disclosure or retention of user-provided inputs that may be accessible across sessions, contexts, or outputs.
- Unsafe Outputs: Generation of responses that violate safety policies, including harmful, illegal, or policy-restricted content.
- Excessive Agency: Uncontrolled or over-permissive model actions, where the AI performs tasks or accesses resources beyond its intended scope or authorization.
- System Prompt Leakage: Disclosure of hidden system instructions or underlying prompt engineering that governs the model's behavior.
- Model Extraction: Attempts to reverse-engineer or replicate the model's internal logic, parameters, or capabilities through repeated querying.
- Harmful Content Bias: Generation of biased, discriminatory, or unfair outputs that could lead to reputational or ethical risks.
- Toxic Output: Production of offensive, abusive, or inappropriate language that may harm users or violate acceptable use standards.

Testing combines:

- Automated attack execution for consistency and large-scale evaluation
- Human-led red team exercises to test evasive, creative, and adaptive attack strategies

Each test case is executed multiple times to ensure consistency, reliability, and repeatability of results.

6 SECURITY VALIDATION MAPPING AND SCORING

6.1 SCORING AND SCORECARD COMPILATION

In this inaugural AI security validation, scoring was performed on a block/allow basis. Each test category score was calculated using the following format in equation 1. The Composite Security Score was calculated by averaging the scores for all 10 categories tested.

$$\begin{matrix} Percentage\ of \\ attacks\ blocked \end{matrix} = \left(\frac{blocked\ \ attacks}{total\ \ attacks} \right) \times 100\%$$

Equation 1. The Formula for Computation of Percentage of Attacks Blocked

The evaluation emphasizes security effectiveness and control coverage rather than product performance metrics or operational characteristics. Scoring also excludes out-of-scope capabilities

The Security Validation Criteria define the evaluation dimensions used to assess the effectiveness of F5 AI Guardrails in protecting generative AI systems from security threats, misuse, and policy violations. These criteria focus on measurable security outcomes under realistic enterprise deployment scenarios and adversarial testing conditions. Each criterion evaluates the ability of F5 AI Guardrails to detect, prevent, mitigate, or contain AI-specific risks during runtime interactions.

Table 1 maps each vulnerability to a corresponding OWASP LLM Top 10 and MITRE ATLAS.

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

Vulnerability	OWASP LLM Top 10	MITRE ATLAS Technique (ID)	Percentage Attacks Blocked (%)
Prompt injection	LLM01 – Prompt Injection	AML.T0051 – Prompt Injection	99.33%
Indirect injection	LLM01 – Prompt Injection	AML.T0051 – Prompt Injection (Indirect / Retrieval Injection)	98.02%
Sensitive data leak	LLM06 – Sensitive Information Disclosure	AML.T0040 – Exfiltration via ML Model	99.01%
Input leakage	LLM06 – Sensitive Information Disclosure	AML.T0040 – ML Model Inference Leakage	98.56%
Unsafe outputs	LLM02 – Insecure Output Handling	AML.T0022 – Model Misbehavior	95.66%
Excessive agency	LLM08 – Excessive Agency	AML.T0035 – Abuse of Autonomous Decision Making	98.68%
System prompt leakage	LLM07 – System Prompt Leakage	AML.T0052 – Prompt Extraction	98.82%
Model extraction	LLM10 – Model Theft	AML.T0044 – Model Extraction	93.85%
Harmful content & bias	LLM09 – Overreliance / Harmful Content Generation	AML.T0022 – Model Misbehavior	99.67%
Toxic output	LLM09 – Overreliance / Harmful Content Generation	AML.T0022 – Model Misbehavior	99.72%

Table 1. Vulnerabilities Mapped to OWASP and MITRE

6.2 PROMPT INJECTION

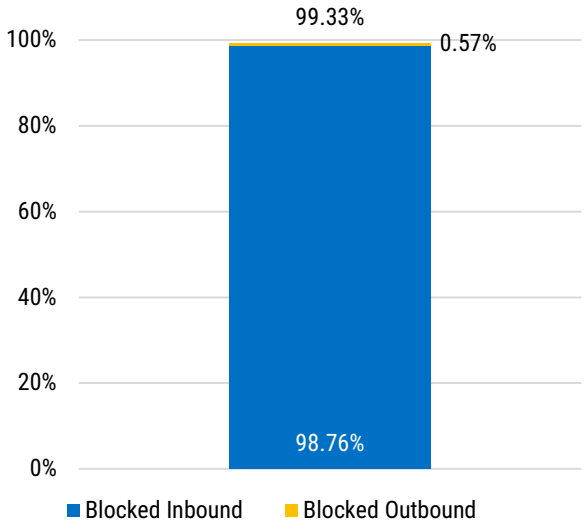


Figure 3. Prompt Injection Total, Inbound, and Outbound Block Rate

A Prompt Injection vulnerability occurs when user-supplied input manipulates the behavior of a generative AI model in unintended or unauthorized ways. These inputs may attempt to override system instructions, bypass guardrails, extract protected information, or influence the model to perform actions that violate defined policies or operational constraints.

Prompt injection attacks do not always appear as obvious malicious text. Any content that is parsed and interpreted by the model can potentially influence its behavior. This includes hidden instructions embedded within structured data, HTML comments, metadata fields, markdown content, encoded strings, documents, or external information retrieved through mechanisms such as APIs, files, or retrieval-augmented generation (RAG) systems.

The objective of this test case is to evaluate whether F5 AI Guardrails can effectively detect, block, or neutralize prompt injection attempts before they impact model behavior. This includes testing the system's ability to identify both direct adversarial prompts submitted by users and indirect prompt injections delivered through external content

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

sources. Figure 3 demonstrates that F5 AI Guardrails blocked 99.33% of the Prompt Injection attacks. 98.76% of these attacks were blocked inbound and 0.57% outbound.

6.3 INDIRECT INJECTION

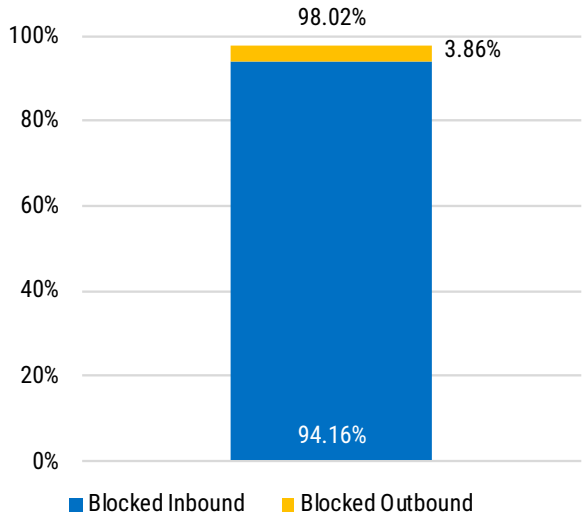


Figure 4. Indirect Injection Total, Inbound, and Outbound Block Rate

An Indirect Injection vulnerability occurs when malicious instructions are embedded within external content that is later processed by a large language model (LLM). Unlike direct prompt injections submitted by a user, indirect injections originate from third-party or external data sources such as documents, web pages, APIs, databases, or retrieval-augmented generation (RAG) systems that the application uses to provide context to the model.

These hidden instructions may be embedded in formats such as HTML comments, metadata fields, markdown blocks, encoded text, or structured data. When the model retrieves and processes this external content, it may interpret the embedded instructions as legitimate prompts, causing it to override system rules, expose sensitive information, or perform unintended actions.

The objective of this test case is to evaluate whether F5 AI Guardrails can detect and neutralize malicious instructions contained within externally sourced content before it is processed by the AI model. The validation assesses the system's ability to inspect, sanitize, and control third-party or retrieved data to prevent hidden adversarial instructions from influencing model behavior. Figure 4 shows F5 AI Guardrails blocked 98.02% of Indirect Injection attacks. 94.16% were blocked inbound, and 3.86% outbound.

6.4 SENSITIVE DATA LEAK

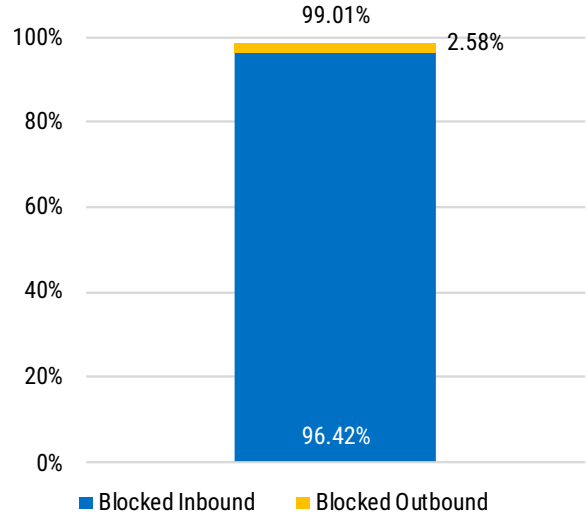


Figure 5. Sensitive Data Leak Total, Inbound, and Outbound Block Rates

A Sensitive Data Leak vulnerability occurs when a large language model (LLM) generates responses that reveal confidential, proprietary, or regulated information. This may include personally identifiable information (PII), financial records, internal business data, authentication credentials, or other sensitive enterprise information that should not be disclosed through AI interactions.

Sensitive data leakage can occur when attackers deliberately craft prompts to extract protected information from the model or when the model unintentionally exposes sensitive content due to contextual inference, training data artifacts, or integration with internal knowledge sources such as databases, documents, or retrieval-augmented generation (RAG) systems. Adversaries may also use indirect techniques such as prompt chaining, obfuscation, or gradual probing to bypass safeguards and obtain restricted information.

The objective of this test case is to evaluate whether F5 AI Guardrails can effectively detect and prevent attempts to extract sensitive data from AI systems. This includes assessing the solution's ability to identify data extraction

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

patterns, enforce data protection policies, and block responses that may expose confidential or regulated information during AI interactions. Figure 5 showcases F5 AI Guardrails blocked 99.01% of Sensitive Data Leak attacks. 96.42% were blocked inbound, and 2.58% outbound.

6.5 INPUT LEAKAGE

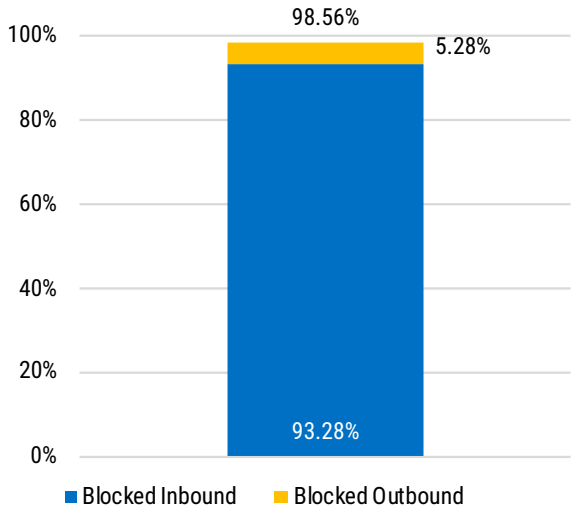


Figure 6. Input Leakage Total, Inbound, and Outbound Block Rate

An Input Leakage vulnerability occurs when sensitive or confidential information provided to a large language model (LLM) as part of user input unintentionally appears in subsequent responses or interactions. This issue arises when the model retains, references, or reproduces sensitive input data, potentially exposing private, confidential, or protected information.

In some cases, input leakage occurs through direct repetition of previously supplied data. However, more advanced scenarios may involve indirect disclosure, where fragments of sensitive input, contextual details, or inferred information are revealed through later outputs. Adversaries may also attempt to exploit this behavior by crafting prompts that encourage the model to reproduce or expose prior inputs.

The objective of this test case is to evaluate whether F5 AI Guardrails can detect and prevent the unintended exposure of sensitive input data during AI interactions. The validation assesses the solution's ability to identify attempts to retrieve previously submitted information, block unauthorized disclosure, and enforce data protection policies that prevent confidential inputs from appearing in model outputs. Figure 6 shows F5 AI Guardrails blocked 98.56% of Input Leakage attacks. 93.28% were blocked inbound, and 5.28% outbound.

6.6 UNSAFE OUTPUTS

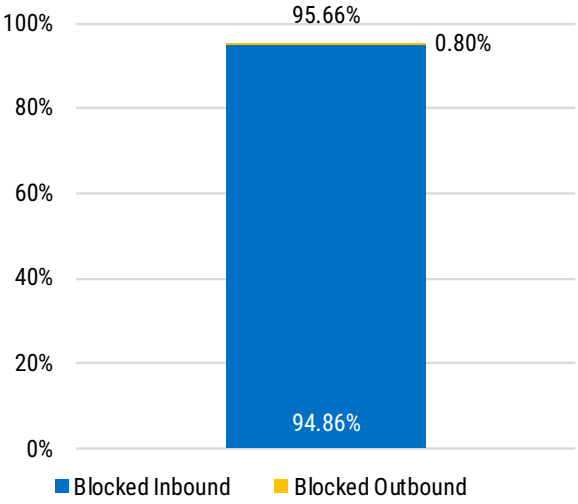


Figure 7. Unsafe Outputs Total, Inbound, and Outbound Block Rates

Unsafe Outputs in large language models (LLMs) refer to responses generated by the model that may introduce safety, security, or operational risks. These outputs generally fall into two primary categories: content-level risks and application-level risks.

Content-level risks occur when the model generates responses that may harm users directly. This includes dangerous advice, unsafe instructions, or other forms of content that may create ethical, legal, or safety concerns for organizations deploying AI systems.

Application-level risks occur when the model generates outputs that, if consumed or processed by downstream systems without proper validation, may introduce security vulnerabilities. Examples include outputs containing executable code, malicious scripts, or structured content that could trigger attacks such as Cross-Site Scripting (XSS), Server-Side Request Forgery (SSRF), command injections, or other input-based exploits in connected applications.

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

Unsafe outputs may therefore create risks not only for users but also for the surrounding application infrastructure, especially when AI-generated responses are automatically integrated into workflows, APIs, web applications, or automation systems.

The objective of this test case is to evaluate whether F5 AI Guardrails can detect, filter, or control unsafe model outputs before they are delivered to users or consumed by downstream systems. The validation assesses the solution's ability to enforce output safety policies, sanitize potentially harmful responses, and prevent AI-generated content from introducing security vulnerabilities into enterprise applications. Figure 7 demonstrates F5 AI Guardrails blocked 95.66% of Unsafe Outputs attacks. 94.86% were blocked inbound, and 0.80% outbound.

6.7 EXCESSIVE AGENCY

Excessive Agency occurs when a generative AI system is able to perform actions, access resources, or invoke tools beyond its intended authority or operational scope. This risk arises when large language models (LLMs) are integrated with external tools, APIs, plugins, automation workflows, or enterprise systems that allow the model to initiate actions based on user prompts.

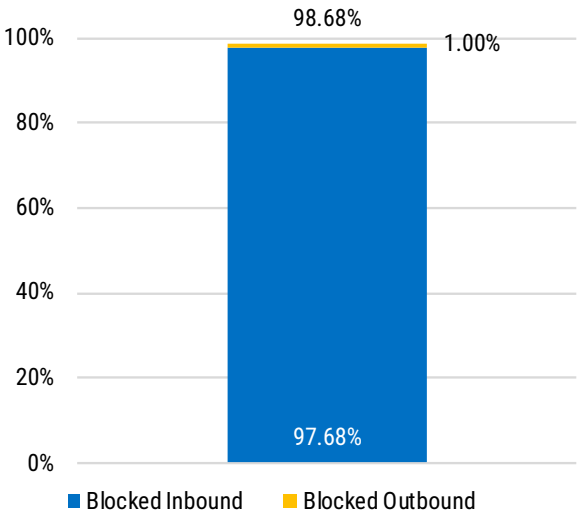


Figure 8. Excessive Agency Total, Inbound, and Outbound Block

In such environments, adversaries may craft prompts designed to manipulate the model into executing unauthorized operations, accessing restricted systems, retrieving sensitive information, or triggering automated workflows without proper authorization. These attacks may exploit weak permission boundaries, insufficient validation of tool usage, or inadequate policy enforcement within the AI system.

Excessive agency risks are particularly significant in AI agent architectures where models can autonomously interact with multiple services or perform multi-step operations. Without proper controls, malicious prompts may cause the AI system to bypass operational constraints, misuse connected tools, or perform actions that exceed its intended capabilities.

The objective of this test case is to evaluate whether F5 AI Guardrails can effectively enforce operational boundaries and prevent unauthorized actions initiated through AI interactions. This includes assessing the solution's ability to monitor AI-driven requests, validate tool and API usage, enforce access policies, and block attempts to trigger unauthorized operations or access restricted resources. Figure 8 showcases F5 AI Guardrails blocked 98.68% of Excessive Agency attacks. 97.68% were blocked inbound, and 1.00% outbound.

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

6.8 SYSTEM PROMPT LEAKAGE

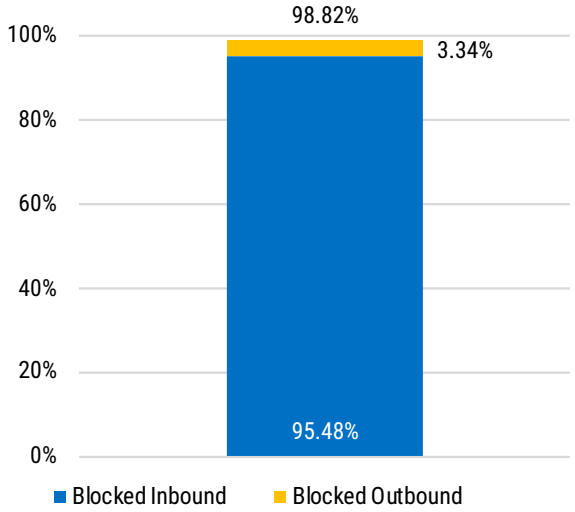


Figure 9. System Prompt Leakage Total, Inbound, and Outbound Block Rates

System Prompt Leakage occurs when a user is able to extract or infer hidden system instructions that govern the behavior of a large language model (LLM). System prompts typically contain internal configuration details, operational rules, safety policies, and guidance that instruct the model on how to respond to users and interact with external systems.

Adversaries may attempt to expose these hidden prompts by crafting specific queries that request the model to reveal its instructions, ignore previous directives, or disclose internal configuration details. In some cases, attackers may use iterative prompting, prompt chaining, or obfuscation techniques to gradually extract portions of the system prompt or infer its structure and constraints.

Exposure of system prompts can significantly weaken AI security controls. Once these internal instructions become known, attackers may use the information to design more effective prompt injection attacks, bypass safety guardrails, or manipulate model behavior to violate organizational policies.

The objective of this test case is to evaluate whether F5 AI Guardrails can detect and prevent attempts to expose system prompts or internal AI instructions. The validation assesses the solution's ability to identify prompt patterns associated with system prompt extraction and block responses that would reveal hidden instructions or internal configuration details. Figure 9 demonstrates F5 AI Guardrails blocked 98.82% of System Prompt Leakage attacks. 95.48% were blocked inbound, and 3.34% outbound.

6.9 MODEL EXTRACTION

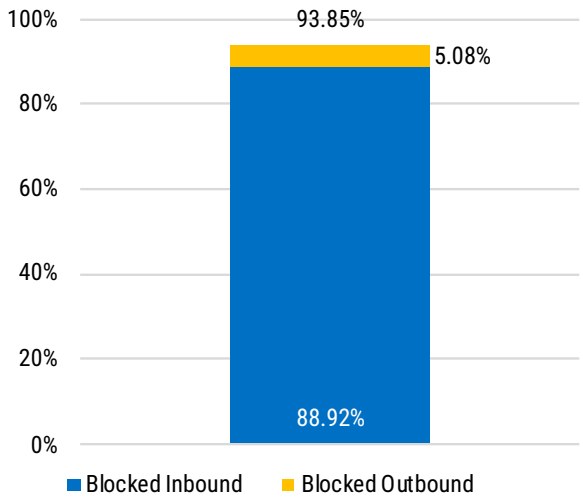


Figure 10. Model Extraction Total, Inbound, and Outbound Block Rates

Model Extraction occurs when an adversary attempts to reconstruct or infer the behavior, capabilities, or underlying logic of a large language model (LLM) through repeated and structured interactions. By systematically querying the model and analyzing its responses, attackers may attempt to approximate the model's decision boundaries, replicate its functionality, or infer proprietary prompts, configurations, or training characteristics.

These attacks typically involve automated or iterative probing techniques, where carefully crafted prompts are submitted in sequence to observe patterns in the model's responses. Over time, an attacker may be able to approximate model behavior, extract valuable intellectual property, or identify weaknesses that can be exploited in subsequent attacks.

Model extraction poses both security and business risks. From a security perspective, it may reveal internal system behavior or expose vulnerabilities that can be leveraged in further attacks. From a commercial perspective, it can undermine intellectual property protections by enabling adversaries to replicate proprietary AI capabilities without authorization.

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

The objective of this test case is to evaluate whether F5 AI Guardrails can detect and mitigate model extraction attempts. This includes assessing the solution's ability to identify abnormal query patterns, rate-limit suspicious interactions, and restrict responses that could expose sensitive information about model behavior, internal configurations, or proprietary AI functionality. Figure 10 shows F5 AI Guardrails blocked 93.85% of Model Extraction attacks. 88.92% were blocked inbound, and 5.08% outbound.

6.10 HARMFUL CONTENT AND BIAS

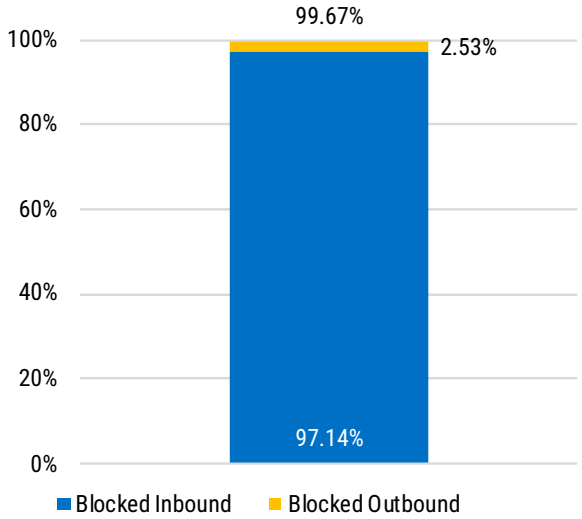


Figure 11. Harmful Content and Bias Total, Inbound, and Outbound Block Rates

Harmful Content and Bias occur when a large language model (LLM) generates responses that contain discriminatory, prejudiced, or unfair content related to sensitive attributes such as race, gender, religion, nationality, disability, or other protected characteristics. Such outputs may arise due to biases present in training data, contextual misinterpretation, or adversarial prompts intentionally designed to provoke biased or harmful responses.

Attackers or users may craft prompts that attempt to manipulate the model into producing discriminatory statements, offensive stereotypes, or content that violates ethical guidelines and organizational policies. Even when unintentional, these responses can create serious legal, reputational, and ethical risks for organizations deploying AI-powered systems.

Harmful or biased outputs can negatively impact user trust, expose organizations to regulatory scrutiny, and undermine responsible AI practices. As generative AI systems become integrated into customer-facing applications and decision-support tools, preventing such outputs becomes a critical governance and risk management requirement.

The objective of this test case is to evaluate whether F5 AI Guardrails can effectively detect and prevent the generation of harmful, discriminatory, or biased content during AI interactions. The validation assesses the solution's ability to enforce safety and fairness policies, filter inappropriate responses, and ensure that AI-generated outputs remain aligned with responsible AI principles and organizational standards. Figure 11 illustrates F5 AI Guardrails blocked 99.67% of Harmful Content and Bias attacks. 97.14% were blocked inbound, and 2.53% outbound.

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

6.11 TOXIC OUTPUT

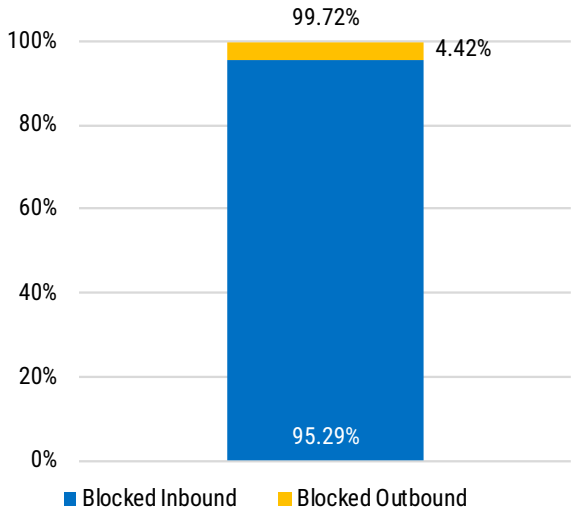


Figure 12. Toxic Output Total, Inbound, and Outbound Block Rates

Toxic Output occurs when a large language model (LLM) generates responses that contain abusive, offensive, hostile, or otherwise inappropriate language. Such outputs may include insults, harassment, hate speech, threats, or inflammatory content that can negatively affect users and degrade the safety of AI-powered interactions.

Toxic responses may arise when users intentionally submit abusive or provocative prompts designed to trigger offensive outputs. In other cases, toxicity may occur due to contextual misinterpretation, insufficient content filtering, or weaknesses in safety guardrails that fail to properly moderate harmful language.

The presence of toxic outputs in AI systems can create significant risks for organizations. These risks include damage to brand reputation, negative user experiences, potential regulatory or legal consequences, and violations of acceptable use or community standards—particularly in customer-facing applications.

The objective of this test case is to evaluate whether F5 AI Guardrails can effectively detect, block, or sanitize toxic content generated by AI systems. The validation assesses the solution's ability to identify abusive or harmful language patterns, enforce content moderation policies, and ensure that AI-generated responses remain safe, respectful, and appropriate for users. Figure 12 demonstrates F5 AI Guardrails blocked 99.72% of Toxic Output attacks. 95.29% were blocked inbound, and 4.42% outbound.

7 F5 GUARDRAILS DIFFERENTIATORS

7.1 FLEXIBLE CUSTOM SCANNER FRAMEWORK

One of F5 AI Guardrails' most powerful features is its custom scanner system, which enables organizations to create tailored detection logic for AI security.

Scanners act as automated inspection engines that analyze prompts, outputs, and interactions for risky or non-compliant content.

Capabilities include:

- Detection of sensitive data exposure
- Identification of prompt manipulation attempts
- Monitoring of policy violations
- Enforcement of organization-specific compliance requirements

Organizations can create scanners using multiple detection approaches:

- AI-driven contextual analysis scanners
- Keyword-based scanners
- Regular expression-based scanners

These scanners can be applied to:

- Prompts
- Responses
- Both

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

This layered scanning capability allows enterprises to implement policy-as-code security for AI systems, translating governance policies directly into enforceable runtime controls.

7.2 SCANNER VERSIONING

F5 AI Guardrails introduces version control for AI security policies, enabling security teams to continuously evolve detection logic without disrupting production systems.

Security teams can:

- Create multiple versions of the same scanner
- Test scanner versions simultaneously
- Roll back or deploy scanner updates safely
- Enforce different versions across different environments

This versioning capability allows organizations to rapidly adapt their AI defenses to new attack techniques, policy changes, or regulatory requirements while maintaining operational stability.

For large enterprises operating multiple AI applications, scanner versioning enables environment-specific security configurations while maintaining centralized governance.

7.3 NATIVE SIEM INTEGRATION (SPLUNK)

SecureIQLab validated F5 AI Guardrails integration with Splunk. This improves the operationalization of AI security telemetry.

Through this integration:

- AI security logs can be ingested directly into the SIEM
- AI threat events appear alongside network and endpoint telemetry
- Automated alerts and response workflows can be triggered
- Compliance logging can be centralized within existing retention frameworks

This integration allows AI security events to become part of the broader enterprise security monitoring ecosystem, improving detection, investigation, and incident response workflows.

8 CONCLUSION

Across 19,679 adversarial test payloads, the F5 AI Guardrails demonstrated 98.36% overall protection, successfully blocking 19,356 attacks across ten major AI security categories. The strongest results were observed in Toxic Output (99.72%), Harmful Content Bias (99.67%), Prompt Injection (99.33%), and Sensitive Data Leakage (99.01%), each exceeding 99% effectiveness. These results indicate highly effective baseline protection against the most common and operationally relevant AI misuse and prompt manipulation scenarios.

Performance remained consistently strong across most additional categories, including Excessive Agency (98.68%), System Prompt Leakage (98.82%), and Input Leakage (98.56%), demonstrating broad coverage across both inbound prompt manipulation and outbound model response controls. The most challenging attack class observed was Model Extraction, which achieved 93.85% protection, followed by Unsafe Outputs (95.66%). These categories represent emerging industry-wide challenges related to model behavior and information inference rather than simple prompt filtering.

Overall, the results show consistently high protection rates across all evaluated AI attack classes, with 4 above 99% and four more above 98% effectiveness. The testing indicates that the F5 AI Guardrails provides strong defensive coverage across prompt manipulation, data leakage, safety controls, and bias mitigation, while highlighting a small number of complex attack classes that remain active areas of research across the broader AI security ecosystem.

©SecureIQLab LLC, 2019 – 2026. All rights reserved.

9 COPYRIGHT AND DISCLAIMER

Copyright © 2026 SecureIQLab, LLC. All rights reserved. The content of this report is protected by United States and international copyright laws and treaties. You may only use this report for your personal, non-commercial, informational purposes. Without SecureIQLab's prior written consent, you may not: (i) reproduce, modify, adapt, create derivative works from, publicly perform, publicly display, or distribute this report; or (ii) use this report, the SecureIQLab name, or any SecureIQLab trademark or logo as part of any marketing, promotion or sales activities. THIS REPORT IS PROVIDED "AS IS," "AS AVAILABLE" AND "WITH ALL FAULTS." TO THE MAXIMUM EXTENT PERMITTED BY LAW, SECUREIQLAB EXPRESSLY DISCLAIMS ALL WARRANTIES AND REPRESENTATIONS, EXPRESS OR IMPLIED, INCLUDING: (a) THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE; AND (b) ANY WARRANTY WITH RESPECT TO THE QUALITY, ACCURACY, CURRENCY OR COMPLETENESS OF THE REPORT, OR THAT USE OF THE REPORT WILL BE ERROR-FREE, UNINTERRUPTED, FREE FROM OTHER FAILURES OR WILL MEET YOUR REQUIREMENTS. WITHOUT LIMITING THE GENERALITY OF THE FOREGOING SENTENCE, YOU ACKNOWLEDGE AND AGREE THAT THE QUALITY, ACCURACY, CURRENCY AND COMPLETENESS OF THE REPORT DEPEND UPON VARIOUS FACTORS, INCLUDING FACTORS OUTSIDE OF SECUREIQLAB'S CONTROL, SUCH AS: (1) THE QUALITY, ACCURACY, CURRENCY OR COMPLETENESS OF INFORMATION AND MATERIALS PROVIDED BY OTHER PARTIES THAT ARE RELIED UPON BY SECUREIQLAB IN PERFORMING PREPARING THE REPORT; AND (2) THE UNDERLYING ASSUMPTIONS MADE BY SECUREIQLAB IN PREPARING THE REPORT REMAINING TRUE AND ACCURATE. YOU ARE SOLELY RESPONSIBLE FOR INDEPENDENTLY ASSESSING THE QUALITY, ACCURACY, CURRENCY AND COMPLETENESS OF THE REPORT BEFORE TAKING OR OMITTING ANY ACTION BASED UPON THE REPORT. IN NO EVENT WILL SECUREIQLAB BE LIABLE FOR ANY LOST PROFITS OR COST OF COVER, OR DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, PUNITIVE OR CONSEQUENTIAL DAMAGES, INCLUDING DAMAGES ARISING FROM OR RELATING TO ANY TYPE OR MANNER OF COMMERCIAL, BUSINESS OR FINANCIAL LOSS, EVEN IF SECUREIQLAB HAD ACTUAL OR CONSTRUCTIVE KNOWLEDGE OF THE POSSIBILITY OF SUCH DAMAGES AND REGARDLESS OF WHETHER SUCH DAMAGES WERE FORESEEABLE.

For more information about SecureIQLab and the testing methodologies, please visit our website.

SecureIQLab (March 2026)