AI & LLM SECURITY TESTING

GROUNDED IN EXPERIENCE. **FOCUSED ON WHAT'S NEXT.**

The rapid evolution of AI and large language models (LLMs) is transforming industries, unlocking new capabilities while introducing novel and significant security risks. Whether you're launching your first AI-powered feature or managing a production-scale model pipeline, understanding your exposure is essential to building secure and resilient systems.

Bishop Fox brings over two decades of offensive security experience and industry-leading expertise to help you navigate this emerging threat landscape. Our AI & LLM Security Testing services are designed to meet you where you are, offering both flexibility and technical depth.

This document outlines what you can expect from a typical engagement. Each assessment is tailored to your environment, maturity level, and risk profile. The phases described below can be delivered independently or combined with other services for a comprehensive, end-to-end evaluation of your AI infrastructure. This playbook explains what happens at each stage so you can choose the services that align with your priorities and development lifecycle.

SCOPING & KICKOFF

Before testing begins, we work closely with your team to understand your objectives, identify the most pressing AI-related security risks, and establish rules of engagement. By prioritizing strategic alignment and information sharing early in the process, we focus our testing on what truly threatens your business and set the stage for a successful engagement.

During this phase, we collect required information, such as the target application, access requirements, credentials, and architecture diagrams. Key stakeholders from your organization meet with a Bishop Fox Technical Engagement Manager and AI/LLM security expert to determine project goals, compliance drivers, and threat concerns, which can include AI-specific threat categories, such as:

- PROMPT INJECTION
- MODEL EXTRACTION
- DATA POISONING
- RESOURCE EXHAUSTION
- SUPPLY CHAIN COMPROMISE
- TRUST BOUNDARIES
- ISOLATION PATTERNS (E.G., DUAL LLM CONTROLLERS, SANDBOXING)
- SECRETS MANAGEMENT

Based on these conversations, we will define the test scope and create a structured communication plan, which typically includes a shared Slack or Teams channel.

After completing about 10% of the fieldwork, we review the test plan and adjust it as needed to ensure the assessment is highly relevant to your organization. This plan is shared with your team for feedback before it is finalized. Throughout the subsequent engagement phases, we provide weekly status reports to track progression against the test plan and project goals, along with optional interim walkthroughs of identified critical-risk findings.

HYBRID APPLICATION PENETRATION TESTING

This core testing phase combines focused simulation of LLM-specific threats with traditional application security testing. We conduct hands-on exploitation of the running software, target applications, and LLM endpoints.

LLM-specific attack simulation examines real-world adversary behaviors against your models. We test for data exfiltration via context leak chains and secrets extraction, jailbreak-style policy bypasses that ignore system instructions, and cost amplification or flooding attacks that may abuse your infrastructure. Techniques like Unicode obfuscation and Base64-encoded payloads help us probe your content moderation capabilities.

Traditional application and API testing complements the LLM-specific assessment by identifying foundational security weaknesses in the broader application ecosystem. These include classic web and API vulnerabilities, as well as novel issues arising from the intersection of traditional business logic and AI-driven workflows.

FOCUS	REPRESENTATIVE TESTING/TACTICS	GOAL
Data Exfiltration	Prompt and secret extraction, context leak chains	Verify confidentiality controls
Policy Bypass	“Ignore previous instructions” jailbreaks, role confusion payloads	Test guardrail enforcement
Resource Abuse	Cost amplification loops, oversized input flooding	Stress usage throttles & cost alarms
Stealth & Evasion	Unicode/zero width obfuscation, Base64-wrapped commands	Evaluate robustness of content moderation
AI logic flaws	Chain-of-thought (CoT) exposure, insecure function/tool invocation	Identify unique logic errors in LLM-driven workflows
Web/API Vulns	IDOR, XSS, CSRF, SSRF, and authentication/authorization flaws	Identify traditional input and access control flaws

CLOUD PENETRATION TESTING (OPTIONAL)

For organizations leveraging cloud platforms in their AI stack, Bishop Fox will assess cloud-specific risks. Our consultants look beyond basic misconfigurations and vulnerabilities to uncover deeper weaknesses and defensive gaps, from unguarded entry points to overprivileged access and vulnerable internal pathways.

This phase delivers a validated vulnerability list with proof-of-concept exploits and actionable remediation guidance.

FOCUS	REPRESENTATIVE TESTING/TACTICS	GOAL
Discovery & Enumeration	Configuration enumeration	Gather and analyze details about the cloud deployment
Cloud Misconfigurations	IAM escalation, exposed buckets	Uncover privilege and data exposure risks
IaC Security	Scan Terraform templates for known issues	Identify insecure infrastructure-by-design
Cost Exploitation	Excessive resource invocation or consumption (e.g., function abuse)	Reveal denial-of-wallet risks

AI-FOCUSED RED TEAM & READINESS (OPTIONAL)

In this phase, we emulate realistic, multistep adversary operations targeting your AI pipeline. Red team operations may execute scenarios such as OSINT reconnaissance followed by spear phishing of DevOps personnel, cloud pivots to access model artifacts, and eventual data exfiltration or extortion scenarios. We will also test across the full model lifecycle, injecting poisoned data during training and tampering with automated gates in your CI/CD pipeline to uncover trust boundary breakdowns.

Our Purple Teaming engagements help identify and resolve gaps in your detection and response capabilities using tailored test cases executed by our Red Team in collaboration with your Blue Team. You receive real-time feedback and actionable recommendations for immediate improvements.

Finally, we assess your incident response readiness by running tabletop drills and identifying runbook gaps. This ensures your team is prepared to not just prevent AI-centric attacks, but also to recover if they occur.

FOCUS	REPRESENTATIVE TESTING/TACTICS	GOAL
Red Team Operations	Recon → spear phish → cloud pivot → ransom extortion	Test layered detection and containment
	Poison data, tamper with promotion gates	Validate lifecycle security
Detection and Response	Purple Teaming	Identify gaps in detection and response capabilities
Incident Readiness	Tabletop drills	Assess recovery capability for AI threat

REPORTING & EXECUTIVE READOUT

At the end of testing, you'll receive a detailed technical report and a high-level executive summary. The technical report ranks issues by severity and includes proof-of-concept examples and remediation guidance. The executive summary connects findings to business risk and outlines forward-looking defense strategies. We also hold an interactive session where we present results to your team and help prioritize the path to remediation.

FOCUS	REPRESENTATIVE TESTING/TACTICS	GOAL
Technical Findings	Ranked issues, proof-of-concept examples, remediation guidance	Enable engineering teams to resolve findings efficiently
Executive Summary	Business risk & roadmap	Inform leadership decision-making
Interactive Readout	Walkthrough of results, Q&A	Accelerate understanding and prioritization

REMEDIATION VALIDATION (OPTIONAL)

If you choose, we can revisit previously identified issues to validate your successful remediation. This may include a targeted retest of vulnerabilities and an updated attestation letter for regulators, partners, or internal stakeholders. For teams seeking further insight, we can also provide a snapshot of your AI security toolchain, mapping your exposure based on categories like prompt injection, cloud misconfigurations, and model artifact hygiene.

FOCUS	REPRESENTATIVE TESTING/TACTICS	GOAL
Targeted Retest	Focused re-evaluation of remediation	Verify remediation and update risk register
Attestation Letter	Formal letter for external validation	Support stakeholder assurance
Toolchain Snapshot	Prompt fuzzers, cloud scanners, artifact audit tools	Benchmark security coverage

READY TO PUT YOUR DEFENSES TO THE TEST?

Bishop Fox is at the forefront of AI and LLM security, partnering with some of the world's most innovative organizations to safeguard their technologies in an ever-evolving threat landscape. As we continue to push the boundaries of offensive security, we welcome the opportunity to share our expertise. Contact us to learn how we can help your organization reduce risk while you innovate with confidence.

ABOUT BISHOP FOX

Bishop Fox is the leading authority in offensive security, providing solutions ranging from continuous penetration testing, red teaming, and attack surface management to product, cloud, and application security assessments.

LEARN MORE AT BISHOPFOX.COM