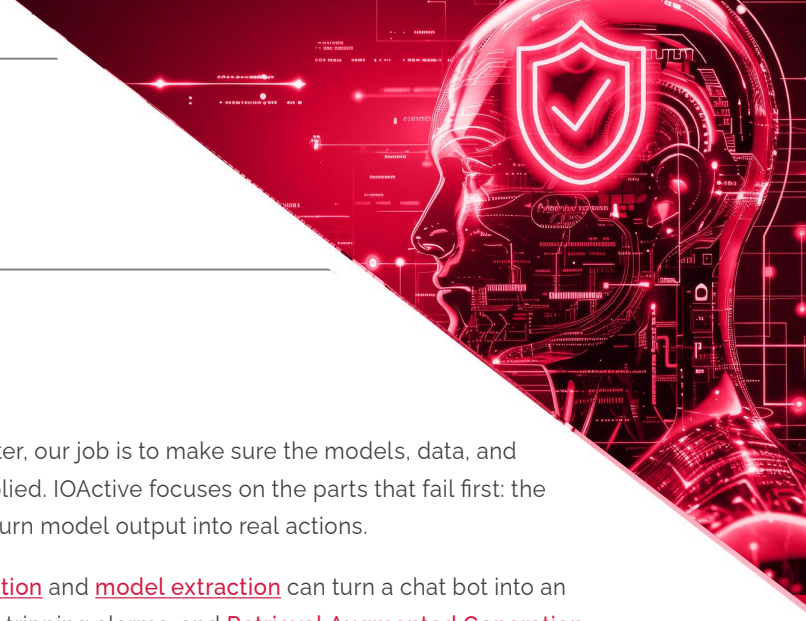
IOActive®

Research-fueled Security Services



AI/ML Security Services

Now that AI is inside products, workflows, and devices that matter, our job is to make sure the models, data, and systems around them behave as intended when pressure is applied. IOActive focuses on the parts that fail first: the model interface, the data it learns from, and the pathways that turn model output into real actions.

Attackers do not need shell access to cause harm. Prompt injection and model extraction can turn a chat bot into an exfiltration channel, data poisoning can steer outcomes without tripping alarms, and Retrieval Augmented Generation (RAG) and tool-calling can extend a single request and response into calls against files, APIs, Model Context Protocol (MCP) servers, and backend production systems. When AI security breaks, the impact is immediate: loss of IP, privacy violations, safety failures, and fraud. The business cost is rework, incident response, regulatory exposure, reputational damage, and slower release trains.

IOActive brings independent validation and verification to the AI space through adversarial testing and engineering discipline. We prove exploitability, measure confidence, and deliver fixes that fit how teams build. Results land as model, pipeline, and application changes with clear owners and success signals. That is how AI ships safely at scale.

As governance pressure and platform changes increase, the teams that instrument models, secure pipelines, run adversarial evaluation, and operationalize security as part of standard delivery will move faster, not slower. They will cut rework, retain customers, and meet emerging expectations for provenance and transparency.

Services and Methodology

IOActive secures AI systems as a whole: from the design, model, and data and evaluation pipeline, to the applications and devices that call the model and the infrastructure that moves and stores weights. We test like an attacker and report like an engineer, delivering reproducible evidence, prioritized fixes, and measurable risk reduction that product and platform owners can ship.

Model adversarial evaluation. We pressure models to behave against policy and interest, then measure how reliably they resist. That includes prompt injection and extraction, sensitive-information disclosure, safety policy evasion, and membership inference. Where access allows, we use white-box influence analysis to confirm whether specific facts are memorized or likely inferred. Every issue is written up with prompts and replay harnesses so teams can reproduce, regression-test, and track closure.

Beyond text models, we also test vision, speech, and decision systems for robustness and misuse, using image manipulation and object-detection edge cases, audio injection and distortion handling for voice interfaces, and decision-path and bias testing against diverse scenarios.

Data and pipeline assurance. We examine how training and evaluation data is sourced, labeled, transformed, and admitted into the pipeline. We look for poisoning opportunities, weak validation, and gaps in provenance and auditability and review artifact lineage and high-influence samples, then prescribe practical guardrails for collection, labeling, and CI/CD. The result is higher trust in the model's inputs and a cleaner signal in its outputs.

Application, tool-calling, and RAG safety. We test AI-enabled applications and devices end to end. That covers model gateway and inference API authorization, transport security, and the places where tool-calling expands risk from a chat box to production systems. We try to cross tenant boundaries through RAG, break file isolation, abuse function execution, and chain classic web or mobile flaws that AI features can amplify. Then we map all of our findings to concrete fixes in the application and the policy layer.

Infrastructure and model custody. We verify how models are stored, signed, moved, and deployed, and we test model-in-transit paths and deployment integrity (signatures, certificates, and access privileges). We test access boundaries around weights, evaluate DoS and scalability controls (resource limits and performance guardrails), and review observability for anomalous use. We call out integrity and isolation gaps that enable model theft, manipulation, or lateral movement between projects and tenants.

Architecture review and threat modeling. We trace attacker paths across SDKs, agents, plugins, and the surrounding supply chain, then align exposures to ATLAS-style TTPs so detections and guardrails fall out as engineering tasks. This is where we connect model behavior to real business impact in identity, payments, and data workflows.

Evidence, observability, and retest. Every engagement includes structured logs, replay scripts, and clear oracles for pass and fail. We verify results across seeds and decoding settings to separate durable failures from noise. After fixes, we retest with the same harness to confirm risk reduction and to seed ongoing evaluation.

End-to-end, full-stack AI security testing across model, pipeline, app, device, and infrastructure; adversarial methods with reproducible evidence; governance-aware recommendations; and delivery that integrates with how your teams build, deploy, and ship.

Governance and standards alignment. We test beyond checklists and show how results align with OWASP ML Top 10, MITRE ATLAS, and emerging governance needs such as training-data provenance and model transparency. Where findings touch policy or regulation, we flag them for legal review with precise technical context.

Outcomes

We favor measurable reductions over vanity metrics. Expect decreased prompt-injection success against protected tasks, lower model-extraction signal at exposed endpoints, reduced sensitive memorization confirmed by replay, higher provenance coverage in pipelines, and shorter time-to-fix for high-risk findings after the first test-and-retest cycle.

IOActive.

For more information about IOActive's Cybersecurity Services, email **info@ioactive.com** or visit **ioactive.com.**

Prepare for the Future Now

Companies trust IOActive because:

1. **Pioneering Research:** Renowned for ground-breaking cybersecurity research, uncovering vulnerabilities that shape industry standards.
2. **Expert Cybersecurity Assessments:** Drawing on extensive expertise, we uncover risks often overlooked by others, ensuring robust protection for your infrastructure.
3. **Customized Advice:** We deliver personalized cybersecurity strategies that address our client's specific business needs and threats.
4. **Global Industry Recognition:** Acknowledged by both peers and clients, our contributions to the cybersecurity community have earned a prestigious reputation.
5. **Innovative Cybersecurity Tools:** Leveraging state-of-the-art tools and techniques, we are at the forefront of cybersecurity technology.
6. **Dedicated Client Partnership:** We prioritize long-term client relationships, offering continuous support and strategic guidance to navigate the evolving security threatscape.

ABOUT IOACTIVE

IOActive, a trusted partner for Global 1000 enterprises, provides research-fueled security services across all industries. Our cutting-edge cybersecurity teams provide highly specialized technical and programmatic services including full-stack penetration testing, program efficacy assessments, and hardware hacking. IOActive brings a unique attacker's perspective to every engagement to maximize cybersecurity investments and improve the security posture and operational resiliency of our clients. Founded in 1998, IOActive is headquartered in Seattle with global operations, including state of the art hardware hacking labs in Seattle, WA, Madrid, Spain and Cheltenham, UK. For more information, visit ioactive.com.