# SEC536: **Adversarial AI – Penetration Testing AI Systems™**

2 Day Course | 12 CPEs | Laptop Required

### You Will Be Able To

- Chain indirect prompt injections across agents to steal production weights
- Exploit RAG retrieval boundaries and outbound channels to exfiltrate sensitive documents
- Defeat facial recognition and identity checks with adversarial attacks and patches
- Attack LLM APIs through role confusion, path traversal, and missing authentication
- Compromise agentic systems via dynamic tool discovery, parameter abuse, and context injection
- Compromise MCP servers through techniques such as SQL injection, homoglyph tool shadowing, and name collision

### Business Takeaways

- Assess AI systems against the attacks they actually face
- Counter AI-specific threats beyond the OWASP top 10
- Reduce AI deployment risk before the first incident
- Validate vendor claims with attacker side evidence
- Protect model weights, training data, and AI supply chains
- Map attack techniques to architectural and detection controls

### Who Should Attend

- Security practitioners who interact with and need to assess AI systems the way they work, not the way a demo describes them
- Penetration testers and red teamers who want to learn techniques that work against LLM-integrated targets instead of recycled web app attacks with an AI label
- Application security engineers, architects who get the threat models and testing patterns they need as AI features hit production
- Data scientists and ML platform engineers who get an unfiltered look at how their pipelines and deployed models appear from an attacker's perspective
- Cyber defenders and blue teamers who get to anticipate attacks instead of reacting to the first incident
- Technical managers who want to learn the ground truth that cuts through vendor hype and shows where AI risk lives in the stack

### Prerequisites

- Students should be comfortable with HTTP APIs and familiar with general web and application security concepts. An AI or ML background is helpful but not required. Python experience is useful for customizing payloads and tooling but is not a prerequisite.

Every organization wants AI features. Very few are asking what those features look like from an attacker's view. Teams are skipping foundational security work and building systems that fail in ways traditional appsec tools were never designed to catch. SEC536 puts you in the attacker's seat so you can see exactly how those failures happen.

### Your Organization Is Deploying AI Systems, Whether Security Is Ready or Not

Most organizations are deploying AI faster than they are securing it. The attacks in this course map directly to vulnerabilities being found in production systems and exploited by attackers today. By emulating them accurately against a realistic target, you give your organization the only thing that informs a security investment: an honest measure of its real risk exposure.

You will learn to attack AI systems the way real adversaries do today through hands-on exploitation of RAG-backed assistants, prompt injection techniques, agent abuse, model weight theft and manipulation, model poisoning, and MCP server attacks. Each technique is paired with clear defensive insights, mapping what you just broke to the architectural decisions, controls, and detection opportunities that would have stopped it, so you leave not just knowing how attacks work, but where and how to stop them in real environments.

## Section Descriptions

**SECTION 1:**

### Foundations of Attack Techniques

In this section, we explore how AI systems are probed, manipulated, and exploited by adversaries. Each module highlights specific issues, moving from basic abuse patterns through reconnaissance, injection, jailbreaks, and adversarial attacks.

**TOPICS:** AI Components and Implications; Prompt Injection (Direct, Indirect, and Others!); Jailbreaking; Evasion and Defense Bypasses

**SECTION 2:**

### Infrastructure, Integrations, and Advanced Attacks

In this section, we explore how the components surrounding AI models create their own attack surface. From infrastructure weaknesses and API misconfigurations to alignment failures and the emerging risks of agentic architectures, this section focuses on exploiting the implementation layer: where and how AI meets the real world.

**TOPICS:** Infrastructure Flaws and Side Channel Attacks; API Security Issues; Alignment Problems; Agentic and MCP Attack Patterns