AI Red Teaming Training

BİLGİNÇ
IT Academy
info@bilginc.com
bilginc.com

Duration: **2 Day**

What Can You Achieve with This Training?

- Learn to systematically assess **Large Language Models (LLMs)** and generative AI applications for security vulnerabilities.
- Gain hands-on experience investigating **prompt injection, jailbreak, and data exfiltration** scenarios in controlled laboratory environments.
- Analyse emerging security threats involving **AI agents, RAG, and MCP-based systems**, including risks introduced by untrusted data.
- Develop practical knowledge of automated AI security assessments using **PyRIT, Inspect AI, and AutoGen**.
- Explore **Responsible AI, threat modelling, and mitigation strategies** to evaluate the security and responsible use of AI systems.
- Reinforce your skills through hands-on exercises and **competitive CTF-style challenges** using the **Tinycode Cyber Range** platform.

The **AI Red Teaming Training** is a **two-day, hands-on programme** designed for security professionals who want to systematically evaluate the security of modern artificial intelligence systems from an adversarial perspective.

The course introduces practical approaches to **AI Red Teaming**, enabling participants to investigate vulnerabilities in large language models and AI-powered applications, assess different attack methods, and examine strategies for reducing identified risks.

Key topics include **adversarial machine learning**, **Responsible AI violations**, and **emerging threats to autonomous AI agents**. The programme begins with an overview of classical machine learning, deep learning, and Transformer architectures before exploring security considerations in modern AI components such as **Retrieval-Augmented Generation (RAG)** and the **Model Context Protocol (MCP)**.

Attack-focused modules examine **direct and indirect prompt injection, jailbreak techniques, data exfiltration, memory poisoning, and tool hijacking**. Participants also explore advanced approaches including **Crescendo, Skeleton Key, GCG, PAIR, and TAP**.

The training introduces open-source tools such as **PyRIT, Inspect AI, and AutoGen** for automated AI security assessments. It also examines security risks in **multimodal AI systems** that process text, images, audio, and video, including misinformation, bias, and harmful content scenarios.

Throughout the programme, participants use the **Tinycode Cyber Range** to investigate realistic security scenarios in a controlled environment. **CTF-style challenges**, scoring, and leaderboards provide opportunities to assess practical understanding.

The final modules focus on **AI security controls, defensive techniques, governance, and risk mitigation**. Participants develop a broader understanding of how to identify AI vulnerabilities and

BILGINÇ IT Academy

info@bilginc.com

bilginc.com

evaluate approaches for reducing their impact.

Prerequisites

No previous experience in **Machine Learning** or **Data Science** is required.

Participants should have:

- A solid understanding of **general cybersecurity principles**
- Basic scripting knowledge using **Python or Bash**

Who Should Attend

This course is suitable for professionals who want to assess and improve the security of AI-powered systems.

It is particularly relevant to:

- **Cybersecurity professionals**
- **Security consultants**
- **Red Team operators**
- Teams looking to add **AI vulnerability assessment** to their existing security capabilities
- Security professionals seeking to understand the **logic, data flows, and unique risks of LLM-integrated systems**

AI Red Teaming Training

BILGINC IT Academy
info@bilginc.com
bilginc.com

Duration: **2 Day**

What You Will Learn

By the end of this course, participants will be able to:

- Systematically assess **LLM applications** for prompt injection, data extraction, and jailbreak vulnerabilities.
- Explain and test **direct and indirect prompt injection** techniques in controlled environments.
- Evaluate advanced attack methods, including **Crescendo, Skeleton Key, GCG, PAIR, and TAP**.
- Identify security risks in **RAG, MCP, and AI agent** data flows.
- Analyse agentic AI threats such as **memory poisoning and tool hijacking**.
- Use **PyRIT, Inspect AI, and AutoGen** to support automated AI security assessments, benchmarking, and agent development.
- Compare **open-box and closed-box** approaches and explain the differences between **semantic and token-level attacks**.
- Assess risks related to **bias, misinformation, and harmful content** in multimodal AI systems.
- Identify **Responsible AI violations** involving bias, privacy, misinformation, and harmful content.
- Contribute to planning AI Red Team activities using the **Map, Measure, Mitigate** framework.
- Explain defensive approaches such as **Defensive Prompting, Spotlighting, and NVIDIA NeMo Guardrails**, including their limitations.
- Gain practical experience conducting security assessments within the **Tinycode Cyber Range**.
- Relate identified AI vulnerabilities to appropriate **risk mitigation and defensive strategies**.

Training Outline

1. Welcome and Lab Environment

- Introducing the training schedule, **laboratory exercises**, and platform
- Accessing the **Tinycode Cyber Range**
- Running controlled scenarios involving **direct prompt injection** and **data exfiltration**

2. Introduction to Machine Learning and AI

- Reviewing **classical machine learning, deep learning, and Transformer** architectures
- Understanding the **data-versus-code problem** in Large Language Models

BILGINC IT Academy

info@bilginc.com

bilginc.com

- Exploring the modern AI technology stack:
  - **Retrieval-Augmented Generation (RAG)**
  - **Model Context Protocol (MCP)**
  - Common **failure modes and security weaknesses**

3. Introduction to Attacking Generative AI

- Exploring **direct and indirect prompt injection** techniques
- Examining jailbreak approaches, including **Crescendo** and **Skeleton Key**
- Understanding **encoding tricks** and **persuasion-based attack patterns**
- Hands-on laboratory exercises involving:
  - **System prompt extraction** scenarios
  - Risks associated with **generating prohibited or harmful content**
  - **Chaining multiple attack techniques**

4. AI Agents and Agentic Systems

- Examining how **agentic AI** changes existing **threat models**
- Exploring **memory poisoning**, **tool hijacking**, and **OpenClaw-related failures**
- Completing laboratory exercises involving **AI agents processing untrusted data**

5. Open-Source Tools for AI Red Teaming

- Introducing automated AI security testing tools:
  - **PyRIT**
  - **Inspect AI**
  - **AutoGen**
- Building and running **automated attack workflows**
- Exploring applications in **security assessment, benchmarking, and agent development**

6. Advanced Attack Techniques

- Examining **algorithmic jailbreaking** approaches
- Comparing **open-box and closed-box** attack methods
- Exploring advanced techniques:

  - **Greedy Coordinate Gradient (GCG)**
  - **Prompt Automatic Iterative Refinement (PAIR)**
  - **Tree of Attacks with Pruning (TAP)**
- Evaluating trade-offs between **semantic and token-level attacks**
- Implementing **PAIR from scratch** in a practical exercise

7. Multimodal Models

- Exploring the evolution of AI models that process **text, images, audio, and video**
- Examining **security challenges and compounded biases** arising from multiple input types
- Completing laboratory exercises involving **multimodal misinformation and propaganda** scenarios

8. Responsible AI and the AIRT Process

- Understanding the principles of **fairness, transparency, and accountability**
- Developing an **AI Red Team programme** using the **Map, Measure, Mitigate** framework
- Discussing the **human impact of AI development**

9. Mitigations

- Reviewing **technical and procedural security controls** for AI systems
- Exploring defensive approaches:
  - **Defensive Prompting**
  - **Spotlighting**
  - **NVIDIA NeMo Guardrails**
- Examining the **limitations of technical controls** against prompt injection

10. AI, Science, and Open Discussion

- Evaluating high-stakes AI applications in **scientific research and national security**
- Reviewing a case study involving **Anthropic's Claude Mythos** and **Project Glasswing**
- Discussing emerging challenges and developments in **AI security**