Agentic AI Security Eğitimi

Süre: **3 Gün**

BİLGİNÇ
IT Academy
info@bilginc.com
bilginc.com

Overview

Large Language Models (LLMs) temeli üzerine geliştirilen Agentic AI sistemlerinin gelişmesiyle birlikte, agent'ların ürettiği aksiyonların güvenlik açısından öngörülemez şekilde evrimleştiğini söyleyebiliriz. LLM'lerin ürettiği text, ses, video ve görsellerin yanı sıra LLM App'lerin de kendine özgü güvenlik zafiyetleri bulunmaktadır.

AI Agent'lar ise LLM'lerden farklı olarak aksiyon seviyesinde güvenlik riskleri oluşturabilir. Bir agent'ın zehirlenmesi, hedefinin değiştirilmesi, kullandığı araçların manipüle edilmesi veya agent'ın kötü niyetli aksiyonlar alması bu riskler arasında yer almaktadır.

Bu eğitimde LLM ve Agentic AI zafiyetlerine uygulamalı olarak odaklanıyoruz. Özellikle yazılım ve mobil güvenlik alanlarında bir standart haline gelen OWASP'ı, LLM ve Agentic AI güvenliği alanında da referans alarak **OWASP Machine Learning Security Top 10, OWASP LLM Top 10 ve OWASP Agentic AI Security Top 10** kapsamındaki güvenlik zafiyetlerini derinlemesine inceliyoruz.

Who Should Attend

Bu eğitim; **AI Security, LLM Security ve Agentic AI Security** konularını öğrenmek isteyen, yapay zeka sistemlerinin güvenlik zafiyetlerini inceleyen veya bu sistemlerin güvenliği üzerinde çalışan katılımcılara yöneliktir.

Özellikle AI sistemleri, LLM App'ler ve AI Agent'lar üzerinde çalışan; AI Blue Teaming ve AI Red Teaming konularıyla ilgilenen; **OWASP Machine Learning Security, OWASP LLM ve OWASP Agentic AI Security** zafiyetlerini incelemek isteyen yazılım, yapay zeka ve siber güvenlik alanındaki teknik katılımcılar için uygundur.

Agentic AI Security Eğitimi

Süre: **3 Gün**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

What You Will Learn

Bu eğitimde yapay zeka güvenliğinin ne olduğunu, yapay zeka güvenliği ile güvenlikte yapay zeka kavramlarını ve AR-GE ile production ortamlarında kullanılan farklı AI sistemlerini inceleyeceksiniz.

Bir yapay zeka sisteminin güvenli olmasının ne anlama geldiğini**AI Robustness ve AI Security** kavramları üzerinden ele alacak; siber güvenlikten bildiğimiz Blue Teaming ve Red Teaming yaklaşımlarını AI perspektifinden inceleyeceksiniz. AI Blue Teaming ve AI Red Teaming takımlarının görev ve sorumluluklarının yanı sıra yapay zekanın kötü niyetli siber saldırılarda kullanımını ve genel AI Security ortamını ele alacaksınız.

OWASP Machine Learning Security Top 10 kapsamında uygulamalı Machine Learning ve Deep Learning projelerinin güvenliğini inceleyeceksiniz.

OWASP LLM Top 10 kapsamında Prompt Injection, Sensitive Information Disclosure, Supply Chain, Data and Model Poisoning, Improper Output Handling, Excessive Agency, System Prompt Leakage, Vector and Embedding Weaknesses, Misinformation ve Unbounded Consumption zafiyetlerini ele alacaksınız.

OWASP Agentic AI Security Top 10 kapsamında ise Agent Goal Hijack, Tool Misuse and Exploitation, Identity and Privilege Abuse, Agentic Supply Chain Vulnerabilities, Unexpected Code Execution (RCE), Memory & Context Poisoning, Insecure Inter-Agent Communication, Cascading Failures, Human-Agent Trust Exploitation ve Rogue Agents konularını inceleyeceksiniz.

Outline

AI Security

What is AI Security?

- Yapay zeka güvenliği nedir?
- Yapay zeka güvenliği ve güvenlikte yapay zeka kavramları.

Different Types of AI Systems

- AI sistemlerinin AR-GE ve production ortamlarındaki farklı türleri.

What Does "Secure AI" Mean?

- Yapay zekanın güvenli olması ne anlama gelir?
- AI Robustness ve AI Security.

Blue Teaming vs. Red Teaming

- AI odaklı Blue Teaming ve Red Teaming.
- AI Blue Teaming ve AI Red Teaming takımlarının görevleri ve sorumlulukları.

Agentic AI Security Eğitimi

Süre: **3 Gün**

BİLGİNÇ
IT Academy
info@bilginc.com
bilginc.com

Using AI for Malicious Intents

- Yapay zekanın kötü niyetli siber saldırılarda kullanılması.

The AI Security Landscape

- Yapay zeka güvenliği ortamı ve genel mimarisi.

OWASP Machine Learning Security Top 10

- Project reviews of applied Machine Learning and Deep Learning security.

OWASP LLM Top 10

LLM01: Prompt Injection

- Prompt'ların bir LLM için oluşturabileceği güvenlik riskleri ve Prompt Injection ile LLM manipülasyonu.

LLM02: Sensitive Information Disclosure

- LLM'lerin öğrendiği kritik veya hassas verilerin ifşa edilmesine yönelik güvenlik riskleri.

LLM03: Supply Chain

- AI altyapısındaki tedarik zinciri zafiyetleri ve güvenliği.

LLM04: Data and Model Poisoning

- Model ve veri zehirleme zafiyetleri ve bunlara karşı korunma.

LLM05: Improper Output Handling

- LLM output'larının oluşturabileceği güvenlik problemleri ve güvenli çıktı yönetimi.

LLM06: Excessive Agency

- LLM uygulamalarında güvenli geliştirme ve tasarım süreci.

LLM07: System Prompt Leakage

- System prompt'ların ifşa edilmesi ve korunması.

LLM08: Vector and Embedding Weaknesses

- Vector DB ve Embedding süreçlerindeki güvenlik zafiyetleri.

LLM09: Misinformation

- Hatalı bilginin LLM sonuçları üzerindeki etkileri.

LLM10: Unbounded Consumption

- AI altyapısının aşırı kullanılması sonucunda oluşabilecek finansal kayıplar ve alınabilecek önlemler.

OWASP Agentic AI Security Top 10

ASI01: Agent Goal Hijack

- Bir agent'ın önceden belirlenen görevinin veya hedefinin değiştirilmesi.

ASI02: Tool Misuse and Exploitation

- Agent aksiyonlarının ve kullandığı araçların manipüle edilmesi.

ASI03: Identity and Privilege Abuse

- Agent kimlikleri ve yetkilendirme mekanizmalarının güvenliği.

ASI04: Agentic Supply Chain Vulnerabilities

- Tedarik zinciri zafiyetlerinin Agentic AI perspektifinden incelenmesi.

ASI05: Unexpected Code Execution (RCE)

- Agent'ların araç çalıştırması veya zararlı kod üretmesi sonucunda oluşabilecek riskler.

ASI06: Memory & Context Poisoning

- LLM ve Agent hafıza mekanizmalarının zehirlenmesi.

ASI07: Insecure Inter-Agent Communication

- A2A gibi protokoller üzerinden gerçekleşen agent iletişiminin manipüle edilmesi veya saldırıya uğraması.

ASI08: Cascading Failures

- Multi-Agent sistemlerde bir agent'ın çıktı veya kararının diğer agent'ları zincirleme şekilde etkilemesi.

ASI09: Human-Agent Trust Exploitation

- Human-in-the-Loop süreçlerinde kötü niyetli veya zehirlenmiş agent'ların insan onayını manipüle etmeye çalışması.

ASI10: Rogue Agents

- Agent'ın tasarım amacı dışında hareket ederek arka planda kötü niyetli işlem, istek veya yönlendirmeler gerçekleştirmesi.

Agentic AI Security Eğitimi

Süre: **3 Gün**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com