AI Red Teaming Eğitimi

BİLGİNÇ
IT Academy
info@bilginc.com
bilginc.com

Süre: **2 Gün**

Bu Eğitim Size Ne Kazandırır?

- **Büyük dil modelleri (LLM)** ve üretken yapay zekâ uygulamalarındaki güvenlik açıklarını sistematik olarak değerlendirmeyi öğrenebilirsiniz.
- **Prompt injection, jailbreak ve veri sızdırma** gibi saldırı senaryolarını kontrollü laboratuvar ortamlarında inceleyerek uygulamalı deneyim kazanabilirsiniz.
- **AI ajanları, RAG ve MCP tabanlı sistemlerde** ortaya çıkan yeni tehditleri ve güvenilmeyen verilerden kaynaklanan riskleri analiz edebilirsiniz.
- **PyRIT, Inspect AI ve AutoGen** gibi araçlarla otomatikleştirilmiş AI güvenlik değerlendirmeleri gerçekleştirmeye yönelik becerilerinizi geliştirebilirsiniz.
- **Responsible AI, tehdit modelleme ve risk azaltma** yaklaşımlarını kullanarak yapay zekâ sistemlerinin güvenliğini ve sorumlu kullanımını değerlendirebilirsiniz.
- **Tinycode Cyber Range** üzerinde uygulamalı senaryolar ve **CTF tarzı yarışmalı laboratuvarlar** aracılığıyla öğrendiklerinizi pekiştirebilirsiniz.

AI Red Teaming Eğitimi, modern yapay zekâ sistemlerinin güvenliğini saldırgan bakış açısıyla değerlendirmek isteyen güvenlik profesyonelleri için hazırlanmış **2 günlük, uygulama odaklı bir eğitim programıdır**.

Eğitim, **AI Red Teaming** çalışmalarının sistematik biçimde yürütülmesine odaklanır. Katılımcılar, büyük dil modelleri ve yapay zekâ destekli uygulamalardaki güvenlik açıklarını araştırmayı, farklı saldırı yaklaşımlarını değerlendirmeyi ve tespit edilen risklere karşı savunma yöntemlerini incelemeyi öğrenir.

Program kapsamında **adversarial machine learning**, **Responsible AI ihlalleri** ve **otonom AI ajanlarına yönelik yeni güvenlik tehditleri** ele alınır. Geleneksel makine öğrenmesi, derin öğrenme ve Transformer mimarilerinden başlayarak **Retrieval-Augmented Generation (RAG)** ve **Model Context Protocol (MCP)** gibi modern AI bileşenlerinin güvenlik açısından önemli özellikleri incelenir.

Eğitimin saldırı odaklı bölümlerinde **doğrudan ve dolaylı prompt injection**, **jailbreak**, **veri sızdırma**, **bellek zehirleme (memory poisoning)** ve **araç ele geçirme (tool hijacking)** senaryoları değerlendirilir. Ayrıca **Crescendo, Skeleton Key, GCG, PAIR ve TAP** gibi ileri düzey tekniklerin çalışma yaklaşımları ele alınır.

Katılımcılar, **PyRIT, Inspect AI ve AutoGen** gibi araçlarla otomatik AI güvenlik testlerini tanır.**Çok modlu (multimodal) modellerin** metin, görsel, ses ve video girdilerinden kaynaklanan riskleri; yanlış bilgilendirme, önyargı ve zararlı içerik senaryoları üzerinden incelenir.

Eğitim boyunca kullanılan **Tinycode Cyber Range** platformu, kontrollü saldırı senaryoları ve uygulamalı alıştırmalar için bir çalışma ortamı sağlar. **Puanlama ve liderlik tablosu** içeren CTF tarzı görevler, katılımcıların öğrendikleri teknikleri pratikte değerlendirmelerine yardımcı olur.

Programın son bölümlerinde **AI güvenlik kontrolleri, savunma teknikleri, yönetişim ve risk azaltma stratejileri** üzerinde durulur. Böylece katılımcılar, AI sistemlerindeki güvenlik zafiyetlerini belirlemenin

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

yanı sıra bu risklerin azaltılmasına yönelik yaklaşımları da değerlendirebilir.

Önkoşullar

Bu eğitime katılmak için **makine öğrenmesi (Machine Learning)** veya **veri bilimi (Data Science)** alanında önceden deneyim sahibi olmak gerekmez.

Katılımcıların aşağıdaki konularda bilgi sahibi olması beklenir:

- **Genel siber güvenlik ilkeleri** konusunda sağlam bir temel
- **Python veya Bash** ile temel düzeyde betik yazma bilgisi

Kimler Katılmalı

Eğitim, yapay zekâ sistemlerinin güvenlik değerlendirmelerini gerçekleştirmek isteyen profesyoneller için uygundur.

Özellikle:

- **Siber güvenlik uzmanları**
- **Güvenlik danışmanları**
- **Red Team operatörleri**
- **AI güvenlik değerlendirmelerini** mevcut test ve analiz süreçlerine eklemek isteyen ekipler
- **LLM entegre sistemlerin mantığını, veri akışlarını ve güvenlik risklerini** anlamak isteyen güvenlik profesyonelleri

AI Red Teaming Eğitimi

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Süre: **2 Gün**

Neler Öğreneceksiniz

Bu eğitimi tamamladığınızda:

- **LLM tabanlı uygulamaları** prompt injection, veri sızdırma ve jailbreak zafiyetleri açısından sistematik olarak değerlendirebileceksiniz.
- **Doğrudan ve dolaylı prompt injection** tekniklerinin güvenlik üzerindeki etkilerini açıklayabilecek ve kontrollü ortamlarda test edebileceksiniz.
- **Crescendo, Skeleton Key, GCG, PAIR ve TAP** gibi saldırı yöntemlerinin kullanım yaklaşımlarını değerlendirebileceksiniz.
- **RAG, MCP ve AI ajanlarının** veri akışlarında ortaya çıkan güvenlik risklerini belirleyebileceksiniz.
- **Memory Poisoning** ve **Tool Hijacking** gibi agentic AI tehditlerini analiz edebileceksiniz.
- **PyRIT, Inspect AI ve AutoGen** araçlarını otomatik AI güvenlik testleri, değerlendirme çalışmaları ve ajan geliştirme senaryolarında kullanmaya yönelik bilgi edinebileceksiniz.
- **Open-Box ve Closed-Box** saldırı yaklaşımlarını karşılaştırabilecek ve **semantic/token-level** yöntemler arasındaki farkları açıklayabileceksiniz.
- **Çok modlu AI sistemlerinde** önyargı, yanlış bilgilendirme ve zararlı içerik risklerini değerlendirebileceksiniz.
- **Responsible AI** kapsamında **önyargı, gizlilik, yanlış bilgilendirme ve zararlı içerik** ihlallerini belirleyebileceksiniz.
- **Map, Measure, Mitigate** yaklaşımıyla AI Red Team çalışmalarının planlanmasına katkı sağlayabileceksiniz.
- **Defensive Prompting, Spotlighting ve NVIDIA NeMo Guardrails** gibi savunma yöntemlerini ve sınırlamalarını açıklayabileceksiniz.
- **Tinycode Cyber Range** ortamında güvenlik değerlendirmeleri gerçekleştirerek uygulamalı deneyim kazanabileceksiniz.
- AI sistemlerindeki güvenlik bulgularını **risk azaltma ve savunma stratejileriyle** ilişkilendirebileceksiniz.

Eğitim İçeriği

1. Eğitime Giriş ve Laboratuvar Ortamı

- Eğitim programının, **laboratuvar çalışmalarının** ve kullanılacak platformun tanıtılması
- **Tinycode Cyber Range** ortamına giriş yapılması

- Kontrollü senaryolarda **doğrudan prompt injection** ve **veri sızdırma (Data Exfiltration)** örneklerinin uygulanması

2. Makine Öğrenmesi ve Yapay Zekâya Giriş

- **Klasik makine öğrenmesi, derin öğrenme ve Transformer** mimarilerinin genel olarak incelenmesi
- Büyük dil modellerinde **veri ile kod arasındaki ayrımın** neden güvenlik sorunu oluşturabildiğinin açıklanması
- Modern AI teknolojilerinin değerlendirilmesi:
  - **Retrieval-Augmented Generation (RAG)**
  - **Model Context Protocol (MCP)**
  - Bu bileşenlerde ortaya çıkabilen **hata ve güvenlik zafiyeti türleri**

3. Üretken Yapay Zekâ Sistemlerine Yönelik Saldırılara Giriş

- **Doğrudan (Direct)** ve **dolaylı (Indirect) Prompt Injection** tekniklerinin incelenmesi
- **Crescendo** ve **Skeleton Key** gibi jailbreak yöntemlerinin değerlendirilmesi
- **Kodlama (Encoding)** yöntemleri ve **ikna temelli saldırı kalıplarının** açıklanması
- Uygulamalı laboratuvar çalışmaları:
  - **Sistem promptlarının** açığa çıkarılmasına yönelik senaryolar
  - **Yasaklanmış veya zararlı içerik üretimi** risklerinin değerlendirilmesi
  - Farklı saldırı tekniklerinin **zincirleme kullanılması**

4. AI Ajanları ve Agentic Sistemler

- **Agentic AI** sistemleriyle birlikte değişen **tehdit modellerinin** incelenmesi
- **Memory Poisoning**, **Tool Hijacking** ve **OpenClaw** ile ilişkili güvenlik başarısızlıklarının değerlendirilmesi
- **Güvenilmeyen verilerle etkileşime giren AI ajanları** üzerinde laboratuvar senaryolarının gerçekleştirilmesi

5. AI Red Teaming için Açık Kaynak Araçlar

- Otomatik AI güvenlik değerlendirmelerinde kullanılan araçların tanıtılması:
  - **PyRIT**
  - **Inspect AI**

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

    - **AutoGen**
- **Otomatik saldırı iş akışlarının** oluşturulması ve çalıştırılması
- Araçların **güvenlik testi, değerlendirme ve AI ajanı geliştirme** süreçlerindeki kullanım alanlarının incelenmesi

6. İleri Düzey AI Saldırı Teknikleri

- **Algoritmik Jailbreaking** yaklaşımlarının incelenmesi
- **Open-Box** ve **Closed-Box** saldırı yöntemlerinin karşılaştırılması
- İleri düzey saldırı tekniklerinin değerlendirilmesi:
    - **Greedy Coordinate Gradient (GCG)**
    - **Prompt Automatic Iterative Refinement (PAIR)**
    - **Tree of Attacks with Pruning (TAP)**
- **Anlamsal (Semantic)** saldırılar ile **token seviyesindeki** saldırıların avantaj ve sınırlamalarının karşılaştırılması
- **PAIR yönteminin sıfırdan uygulanmasına** yönelik laboratuvar çalışması

7. Çok Modlu Yapay Zekâ Modelleri

- **Metin, görsel, ses ve video** işleyebilen multimodal modellerin gelişiminin incelenmesi
- Birden fazla girdi türünün oluşturduğu **güvenlik risklerinin ve birleşik önyargıların** değerlendirilmesi
- **Multimodal yanlış bilgilendirme ve propaganda** senaryolarına yönelik laboratuvar çalışmaları

8. Responsible AI ve AIRT Süreci

- **Adillik (Fairness)**, **şeffaflık (Transparency)** ve **hesap verebilirlik (Accountability)** ilkelerinin açıklanması
- **Map, Measure, Mitigate** çerçevesiyle bir **AI Red Team programı** oluşturulması
- Yapay zekâ geliştirme süreçlerinin **insanlar ve toplum üzerindeki etkilerinin** tartışılması

9. Güvenlik Önlemleri ve Risk Azaltma

- AI sistemlerini korumak için kullanılan **teknik ve prosedürel güvenlik kontrollerinin** incelenmesi
- Savunma yaklaşımlarının değerlendirilmesi:
    - **Defensive Prompting**
    - **Spotlighting**

BİLGİNÇ
IT Academy
info@bilginc.com
bilginc.com

    - **NVIDIA NeMo Guardrails**
- **Prompt Injection** saldırılarına karşı teknik kontrollerin **sınırlamalarının** ele alınması

10. Yapay Zekâ, Bilim ve Açık Tartışma

- **Bilimsel araştırmalar** ve **ulusal güvenlik** gibi yüksek riskli alanlarda yapay zekâ uygulamalarının değerlendirilmesi
- **Anthropic Claude Mythos** ve **Project Glasswing** üzerine vaka incelemesi
- AI güvenliğinin güncel ve gelişmekte olan boyutlarına ilişkin **açık tartışma**