Certified AI Security Engineer Eğitimi

Süre: **3 Gün**

BİLGİNÇ
IT Academy
info@bilginc.com
bilginc.com

- **Gerçek dünyadaki yapay zekâ güvenlik tehditlerini tanımlayın, test edin ve bunlara karşı savunma geliştirin.**
- **LLM tabanlı sistemler geliştiren veya koruyan güvenlik mühendisleri ve AI geliştiricileri için tasarlanmıştır.**
- **Prompt injection, jailbreak, model theft ve diğer AI güvenlik risklerini uygulamalı laboratuvarlarla inceleyin.**

Certified AI Security Engineer eğitimi, modern yapay zekâ sistemlerinde ortaya çıkan güvenlik risklerini kapsamlı biçimde ele alan uygulamalı bir programdır. Eğitim boyunca katılımcılar hem tahmine dayalı AI modellerinin hem de üretken yapay zekâ sistemlerinin saldırganlar tarafından nasıl hedef alınabileceğini inceler.

Program; prompt injection, model jailbreak, denial of service, model theft ve data poisoning gibi yaygın AI saldırı tekniklerini ele alır. Bununla birlikte yalnızca model katmanına değil, eğitim verilerinden API entegrasyonlarına, uygulama mimarisinden dağıtım altyapısına kadar yapay zekâ sistemlerinin tüm saldırı yüzeyine odaklanır.

Katılımcılar, saldırı yöntemlerini anlamanın yanında bu risklere karşı uygulanabilecek savunma tekniklerini de öğrenir. Güvenlik API'leri, yapılandırılmış prompt savunmaları, güvenli entegrasyon desenleri, veri bütünlüğü kontrolleri ve dayanıklı altyapı tasarımları eğitim boyunca uygulamalı örneklerle işlenir.

Gerçek dünya senaryoları ve uygulamalı laboratuvarlar sayesinde katılımcılar; güvenilir, sorumlu ve güvenli AI sistemleri geliştirmek, kurumsal varlıkları korumak ve AI destekli uygulamalarda güven seviyesini artırmak için gerekli yaklaşımı kazanır.

Önkoşullar

Katılımcıların aşağıdaki temel bilgi ve deneyimlere sahip olması önerilir:

- Sinir ağları ve model yaşam döngüsü gibi temel yapay zekâ kavramları hakkında başlangıç seviyesinde bilgi
- Siber güvenlik prensipleri ve yaygın saldırı türleri konusunda temel farkındalık
- AI veya LLM özellikleri kullanan uygulamalarla çalışma deneyimi
- AI entegrasyonlarını uygulamalı olarak test edebilecek bir geliştirme ortamına erişim

Son iki madde zorunlu olmamakla birlikte eğitimden daha fazla verim alınmasına yardımcı olur.

Certified AI Security Engineer Eğitimi

Süre: **3 Gün**

BİLGİNÇ
IT Academy
info@bilginc.com
bilginc.com

Kimler Katılmalı

Bu eğitim özellikle şu profesyoneller için uygundur:

- AI çözümlerinin devreye alınması, entegrasyonu veya güvenliğinden sorumlu teknoloji profesyonelleri
- Yapay zekâya özgü saldırılar hakkında daha derin bilgi edinmek isteyen güvenlik uzmanları
- Büyük dil modelleri veya generative AI kullanan uygulamalar geliştiren yazılım geliştiricileri
- AI kaynaklı risklere karşı kurumsal dayanıklılığını artırmak isteyen ekipler ve organizasyonlar

Neler Öğreneceksiniz

Eğitimin sonunda katılımcıların:

- Farklı yapay zekâ sistemi türlerini tanımlayabilmesi ve bunlara özgü güvenlik açıklarını açıklayabilmesi
- Prompt injection, model jailbreak, visual prompt manipulation ve denial of service gibi saldırıları tespit edip azaltabilmesi
- AI sistemlerini güçlendirmek için savunma yöntemleri ve security API araçları kullanabilmesi
- Eğitim verilerini, model bütünlüğünü ve tedarik zinciri bağımlılıklarını değerlendirebilmesi
- Large Language Model'leri uygulamalara güvenli şekilde entegre edebilmesi
- LLM trust boundary kavramını doğru şekilde uygulayabilmesi
- Responsible AI, etik kullanım, güvenilirlik ve açıklanabilirlik prensiplerini değerlendirebilmesi
- Model davranışını inceleyebilmesi ve kötüye kullanım belirtilerini tespit edebilmesi
- AI iş akışları için yapılandırılmış threat modelling uygulayabilmesi
- Hallucination, kötüye kullanım ve hassas veri sızıntısını azaltan güvenli insan-AI etkileşimleri tasarlayabilmesi

amaçlanmaktadır.

Eğitim İçeriği

AI Security'ye Giriş

Bu bölümde yapay zekâ güvenliğinin kapsamı ve temel kavramları ele alınır.

Konular:

- AI ve security kavramlarının tanımlanması
- AI security kapsamı ve eğitimin sınırları
- AI sistem türleri: neural networks, models ve integrated systems
- AI sistemlerinin organizasyonlarda kullanım biçimleri
- Secure AI kavramı
- Responsible, reliable, explainable ve aligned model yaklaşımı
- İnsan-AI etkileşimlerinde oluşabilecek güvenlik riskleri
- Uncensored veya malicious modellerin oluşturduğu tehditler
- Deepfake, voice cloning ve social engineering örnekleri
- AI tarafından üretilen içeriklerle misinformation yayılımı

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

- Uncensored modeller ve image watermarking üzerine uygulamalar

AI Security Landscape

Bu bölüm, bir AI sisteminin yaşam döngüsü boyunca ortaya çıkabilecek saldırı yüzeylerini inceler.

Konular:

- AI model lifecycle boyunca attack surface analizi
- AI pipeline bileşenleri
- Supply-chain security'nin önemi
- API üzerinden erişilen modeller
- Modellerin eriştiği API'ler
- AI dışında kalan ancak AI sistemleri için hâlâ geçerli olan klasik saldırı vektörleri
- OWASP ML Top 10
- OWASP LLM Top 10
- Modern AI sistemlerinde bu risklerin uygulanması
- AI entegre uygulamalar için threat modelling yaklaşımları
- AI destekli iş akışlarında yaygın güvenlik bulguları

Uygulama: Gerçekçi bir veri akışı üzerinden LLM entegre bir uygulama için threat modelling çalışması.

Prompt Injection

Bu bölümde prompt injection saldırılarının nasıl gerçekleştiği ve etkilerinin nasıl ortaya çıktığı ele alınır.

Konular:

- Prompt injection saldırılarına genel bakış
- Direct prompt injection
- Indirect prompt injection
- Prompt tabanlı social engineering
- Phishing senaryoları
- Attack logic ifade etmek için SudoLang kullanımı
- LLM entegrasyon tercihlerinin güvenlik açıklarına etkisi

Uygulamalar:

- Prompt'ları SudoLang formatına dönüştürme
- Level 1 ve Level 2 senaryolarında parola elde etme teknikleri

Model Jailbreaks

Bu bölüm, güvenlik kısıtlamalarını aşmayı hedefleyen jailbreak yöntemlerini inceler.

Konular:

- Jailbreak mekanizmaları
- Yaygın jailbreak teknikleri
- DAN prompt örnekleri
- AutoDAN
- Tree of Attacks with Pruning (TAP)

Uygulamalar:

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

- Level 3
- Level 4
- Level 5

senaryolarında kısıtlanmış bilginin elde edilmesine yönelik çalışmalar.

Prompt Extraction

Bu bölüm, sistem prompt'larının ve özel bilgilerin çıkarılmasına yönelik saldırıları ele alır.

Konular:

- System prompt extraction
- Private data extraction
- Model boundary keşfi
- Yarışmalarda ve gerçek uygulamalarda kullanılan extraction teknikleri

Uygulamalar:

- Level 6
- Level 7

senaryolarında prompt ve boundary bilgilerini elde etme.

AI Sistemlerini Savunma

Bu bölümde AI saldırılarına karşı orta ve ileri seviye savunma yaklaşımları ele alınır.

Konular:

- Intermediate ve advanced defence strategies
- ReBuff
- Llama Guard
- Lakera
- Benzeri security API çözümleri
- Açık güvenlik challenge'larında görülen exploit örnekleri
- Reverse psychology
- Manipulation techniques
- Injection saldırılarının sınıflandırılması
- Sağlam koruma yöntemlerinin uygulanması
- Bergeron method gibi yapılandırılmış savunma modelleri

Uygulama:

- Level 8 ve Level 9 korumalarını aşmaya yönelik senaryolar

Visual Prompt Injection

Bu bölüm, multimodal AI sistemlerinin görsel içerikler üzerinden nasıl manipüle edilebileceğini inceler.

Konular:

- Visual prompt attack mantığı

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

- Basit görsel manipülasyon örnekleri
- Advanced adversarial attacks
- Self-driving sistemlerini etkileyen örnekler
- Image classifier saldırıları
- OpenAI vision capabilities ile uygulamalar
- Adversarial sample oluşturma
- Görsel saldırılara karşı koruma yöntemleri
- Dataset güvenliği

Denial of Service

Bu bölümde LLM ve chatbot sistemlerinin hizmet dışı bırakılması veya kaynak tüketiminin artırılması hedeflenen saldırılar ele alınır.

Konular:

- LLM sistemlerinde DoS saldırıları
- Chatbot sistemlerinde resource exhaustion
- Prompt routing problemleri
- Kaynak tüketimi riskleri
- Sistem seviyesinde savunma yaklaşımları

Uygulama: Model davranışını durdurmayı veya performansını düşürmeyi hedefleyen prompt tasarımları.

Model Theft

Bu bölüm, model davranışının veya model yeteneklerinin kopyalanmasına yönelik saldırıları inceler.

Konular:

- Model extraction threat landscape
- Dataset exploration riskleri
- Query-based model stealing
- Fine-tuned modellerin klonlanması
- API parameter kullanımının model replikasyonundaki rolü
- Model gizliliğinin korunması
- Rate limiting
- Advanced monitoring yaklaşımları

Uygulama: API parametreleri üzerinden model davranışını taklit etmeye yönelik çalışmalar.

LLM Integration

Bu bölüm, large language model'lerin uygulamalara güvenli entegrasyonuna odaklanır.

Konular:

- LLM trust boundary
- Yeni AI iş akışlarında klasik entegrasyon riskleri
- LLM output'un untrusted user input olarak değerlendirilmesi
- Exchange formats
- Secure function calling

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

- Custom GPT riskleri
- Identity flow
- Cross-application access
- SQL injection
- XSS payload generation
- Invalid parameter passing
- Privilege escalation
- AI sistemlerinde secure coding prensipleri
- Bishop, Saltzer ve Schroeder prensipleri
- AI bileşenleri için privilege boundary tasarımı

Uygulama: AI sandbox ortamından çıkmaya yönelik saldırı senaryosu.

Training Data Manipulation

Bu bölüm, eğitim verisinin bütünlüğünü ve güvenilirliğini etkileyen saldırıları ele alır.

Konular:

- Dataset integrity
- Dataset reliability
- Training data poisoning
- Dataset cards
- Model cards
- Dataset analizi
- Dataset objective değerlendirmesi

Uygulama: Zararlı dataset oluşturma ve analiz etme.

Secure Supply Chain

Bu bölüm AI sistemlerinin geliştirme ve dağıtım yaşam döngüsündeki supply chain risklerine odaklanır.

Konular:

- Model bütünlüğünün doğrulanması
- Emerging cryptographic methods
- Hardware-assisted attestation
- Model verification
- Model building lifecycle riskleri
- Deployment lifecycle riskleri

Human-AI Interaction

Bu bölüm, insanların AI sistemleriyle etkileşimi sırasında ortaya çıkabilecek güvenlik ve güvenilirlik sorunlarını ele alır.

Konular:

- LLM output'a aşırı güvenmenin riskleri
- Hallucination etkileri
- Bilgiyi doğrulama yöntemleri

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

- Sandboxing
- Safe API patterns

Uygulama: Gerçekçi senaryolarda LLM çıktısının doğrulanması.

Secure AI Infrastructure

Bu bölüm, güvenli AI sistemleri çalıştırmak için gereken altyapı özelliklerini inceler.

Konular:

- Monitoring
- Observability
- Traceability
- Confidentiality
- Integrity
- Availability
- Privacy
- Samsung data leak gibi gerçek olaylar
- LangSmith
- Güvenli değerlendirme süreçleri
- BlindLlama
- Emerging AI evaluation tools

Uygulama: LangSmith kullanılarak güvenli AI değerlendirme çalışmaları.

Sınav ve Değerlendirme

Certified AI Security Engineer sınavı eğitim sonrasında bağımsız olarak gerçekleştirilir.

Sınav, APMG tarafından sağlanan voucher code kullanılarak APMG proctor platformu üzerinden alınır.

Sınav detayları:

- **Süre:** 60 dakika
- **Soru Sayısı:** 60
- **Soru Tipi:** Çoktan seçmeli
- Her soru için 4 seçenek bulunur ve yalnızca 1 cevap doğrudur
- **Geçme Notu:** %50