.devmy

CORSO AVANZATO {AI}

AI Security & Red Teaming

Ogni applicazione AI apre una superficie di attacco nuova, che gli strumenti di sicurezza tradizionali non coprono. Prompt injection, esfiltrazione di dati attraverso il contesto, abuso degli strumenti esposti agli agenti, jailbreak: sono rischi concreti, e con la diffusione degli agenti autonomi diventano critici.

Questo corso adotta il doppio punto di vista dell'attaccante e del difensore. I partecipanti imparano a individuare le vulnerabilità tipiche dei sistemi LLM in un ambiente di laboratorio controllato (red teaming) e, soprattutto, a progettare le difese: guardrail, validazione, isolamento e principio del privilegio minimo per agenti e MCP server.

Livello

Avanzato

Durata

Corso da 16 ore

Lingua

Italiano

.devmy

A chi si rivolge

- Sviluppatori e architetti che costruiscono applicazioni o agenti AI
- Security engineer e professionisti AppSec che si affacciano sull'AI
- Tech lead responsabili della sicurezza dei prodotti AI

Cosa faremo

Corso pratico con laboratori offensivi e difensivi in ambiente controllato: analizzeremo applicazioni AI deliberatamente vulnerabili e costruiremo le relative contromisure. Approccio "capture the flag" guidato, seguito dalla progettazione delle difese.

Takeaway:

- Conoscere le principali vulnerabilità dei sistemi LLM (OWASP LLM Top 10)
- Comprendere prompt injection e jailbreak in ambiente di laboratorio controllato
- Proteggere dati e strumenti esposti ad agenti e MCP server
- Progettare guardrail, validazione e isolamento efficaci
- Impostare un processo di red teaming continuo

Numero massimo di partecipanti: 20

Skill minime necessarie

- Esperienza di sviluppo con almeno un linguaggio
- Familiarità con applicazioni web/API e concetti base di sicurezza
- Utile esperienza, anche minima, con LLM API o agenti

.devmy

Argomenti Trattati

- Modulo 1 — Il panorama delle minacce AI
  - Perché l'AI introduce una nuova superficie di attacco
  - OWASP Top 10 per applicazioni LLM: una mappa
  - Threat modeling per sistemi AI
  - Supply chain AI: modelli, dataset e dipendenze come vettori di attacco (model e data poisoning)
- Modulo 2 — Comprendere le tecniche offensive
  - Prompt injection diretta e indiretta
  - Jailbreak e aggiramento delle istruzioni di sistema
  - Esfiltrazione di dati tramite contesto e strumenti
  - Abuso degli strumenti e degli agenti
- Modulo 3 — Sicurezza di agenti e MCP
  - Rischi specifici degli agenti autonomi
  - MCP server sicuri: superfici esposte e isolamento
  - Identità, privilegio minimo e confini di accesso per gli agenti
  - Trattare i contenuti non attendibili come dati, non come istruzioni
- Modulo 4 — Difese e processo
  - Guardrail di input/output e validazione
  - Sandboxing, allow-list e human-in-the-loop per le azioni rischiose
  - Logging, detection e risposta agli incidenti AI
  - Red teaming continuo e integrazione nel ciclo di sviluppo
  - Obblighi normativi collegati: incident reporting nell'AI Act e raccordo con la direttiva NIS2

Cosa è necessario

- Il proprio computer portatile
- Tanta buona volontà e voglia di imparare
- Connessione ad Internet
- Account gratuiti su piattaforme di terze parti (indicati prima del corso)
- Una API key per un provider LLM

.devmy

Cosa è incluso

- Corso pratico con live coding
- Supporto setup environment
- Slide in formato PDF
- Repository del Progetto
- Attestato di Partecipazione
- Follow Up di fine corso
- Canale Slack dedicato ai partecipanti

Dove si svolge

Full-Remote
È possibile svolgere il corso in modalità *full-remote* con gli strumenti messi a disposizione da Devmy suddividendo, se lo si desidera, il tutto in sessioni da 4h.

Note

Le attività offensive si svolgono esclusivamente su ambienti di laboratorio forniti da Devmy. Le tecniche apprese devono essere utilizzate solo su sistemi propri o per i quali si dispone di autorizzazione esplicita e documentata.