Szkolenie Bezpieczeństwo AI i Red Teaming z wykorzystaniem narzędzi Microsoft

Opis szkolenia: Szkolenie koncentruje się na zagadnieniach bezpieczeństwa sztucznej inteligencji oraz praktykach Red Teamingu z wykorzystaniem narzędzi Microsoft, takich jak Azure AI Red Teaming Agent i PyRIT. Uczestnicy poznają typowe zagrożenia związane z generatywną AI, nauczą się przeprowadzać testy odporności modeli oraz integrować je z cyklem DevSecOps. Szkolenie ma charakter praktyczny i nie wymaga wcześniejszego doświadczenia z AI.

Czas trwania: 8 godzin (9:00 – 17:00)

Kategoria: AI, Red Team

Język wykładowy: język polski

Materiały i narzędzia: w języku angielskim. Na życzenie klienta materiały w języku polskim.

Wymagania wstępne: Nie jest wymagana wcześniejsza wiedza z zakresu AI ani programowania.

Grupa decelowa:

- Specjaliści ds. bezpieczeństwa IT i DevSecOps
- Architekci rozwiązań AI i inżynierowie danych
- Liderzy zespołów odpowiedzialnych za wdrażanie AI
- Osoby odpowiedzialne za zgodność, ryzyko i audyt technologiczny
- Programiści i analitycy zainteresowani testowaniem odporności modeli AI

Cel szkolenia: Celem szkolenia jest przekazanie uczestnikom wiedzy i umiejętności w zakresie identyfikacji zagrożeń związanych z AI, przeprowadzania testów odporności modeli oraz integracji Red Teamingu z procesami DevSecOps.

Uczestnicy otrzymają zestaw danych testowych i materiałów (skrypty, linki). Podczas szkolenia będą korzystali z własnych komputerów i licencji.

Szkolenie kończy się certyfikatem uczestnictwa.

Plan szkolenia

Godzina	Czas trwania	Moduł	Forma
9:00 – 9:15	15 minut	**Moduł 1: Wprowadzenie i cele szkolenia**	Prezentacja, dyskusja

Godzina	Czas trwania	Moduł	Forma
9:15 – 10:15	60 minut	**Moduł 2: Wprowadzenie do zagrożeń AI i podejścia "shift-left"**	Teoria, przykłady
10:15 – 10:30	15 minut	Przerwa na kawę	—
10:30 – 11:30	60 minut	**Moduł 3: Red Teaming AI – podstawy i strategie ataku**	Teoria, analiza przypadków
11:30 – 12:30	60 minut	**Moduł 4: Narzędzia Microsoft – Azure AI Red Teaming Agent i PyRIT**	Pokaz, konfiguracja środowiska
12:30 – 13:30	60 minut	Przerwa obiadowa	—
13:30 – 14:30	60 minut	**Moduł 5: Symulacja ataków – Jailbreak, Base64, Leetspeak, Unicode**	Laboratorium praktyczne
14:30 – 15:15	45 minut	**Moduł 6: Ocena ryzyka i analiza odpowiedzi modeli AI**	Laboratorium praktyczne, analiza wyników
15:15 – 15:30	15 minut	Przerwa na kawę	—
15:30 – 16:15	45 minut	**Moduł 7: Automatyzacja testów i integracja z cyklem DevSecOps**	Warsztat, dyskusja
16:15 – 16:45	30 minut	**Moduł 8: Przyszłość AI jako wektor zagrożeń i przewaga strategiczna**	Prezentacja
16:45 – 17:00	15 minut	**Moduł 9: Podsumowanie, pytania i dalsze kroki**	Dyskusja, Q&A, materiały końcowe

Szczegółowy program szkolenia

Moduł 1: Wprowadzenie i cele szkolenia

Przedstawienie programu, celów i wartości szkolenia. Uczestnicy poznają kontekst rosnącego znaczenia bezpieczeństwa AI i roli Red Teamingu w nowoczesnych organizacjach.

- Przedstawienie programu, celów i wartości szkolenia.
- Kontekst rosnącego znaczenia bezpieczeństwa AI.

- Rola Red Teamingu w nowoczesnych organizacjach.

Moduł 2: Wprowadzenie do zagrożeń AI i podejścia "shift-left"

Omówienie typowych zagrożeń związanych z generatywną AI oraz znaczenia wczesnego wykrywania ryzyk w cyklu rozwoju. Przedstawienie koncepcji „świadomej AI" i jej wpływu na strategię bezpieczeństwa.

- Typowe zagrożenia związane z generatywną AI.
- Znaczenie wczesnego wykrywania ryzyk w cyklu rozwoju.
- Koncepcja „świadomej AI" i jej wpływ na strategię bezpieczeństwa.

Moduł 3: Red Teaming AI – podstawy i strategie ataku

Wprowadzenie do technik Red Teamingu stosowanych w kontekście AI. Uczestnicy poznają metody omijania zabezpieczeń, takie jak Jailbreak, manipulacja promptami, kodowanie i inne.

- Techniki Red Teamingu w kontekście AI.
- Metody omijania zabezpieczeń: Jailbreak, manipulacja promptami.
- Kodowanie danych wejściowych: Base64, Unicode, Leetspeak.

Moduł 4: Narzędzia Microsoft – Azure AI Red Teaming Agent i PyRIT

Praktyczne zapoznanie się z narzędziami Microsoft do testowania odporności modeli AI. Uczestnicy skonfigurują środowisko testowe i poznają możliwości automatyzacji testów penetracyjnych.

- Konfiguracja środowiska testowego.
- Przegląd funkcji narzędzi Microsoft.
- Automatyzacja testów penetracyjnych.

Moduł 5: Symulacja ataków – Jailbreak, Base64, Leetspeak, Unicode

Ćwiczenia praktyczne z wykorzystaniem zestawów danych i scenariuszy ataków. Uczestnicy przeprowadzą symulacje i nauczą się identyfikować luki w zabezpieczeniach modeli.

- Ćwiczenia praktyczne z wykorzystaniem zestawów danych.
- Symulacja ataków na modele językowe.
- Identyfikacja luk w zabezpieczeniach.

Moduł 6: Ocena ryzyka i analiza odpowiedzi modeli AI

Analiza wyników testów – jak ocenić, czy odpowiedzi modelu zawierają treści szkodliwe, nieetyczne lub niezgodne z polityką organizacji. Wprowadzenie do metryk skuteczności ataków.

- Analiza wyników testów.
- Identyfikacja treści szkodliwych i nieetycznych.
- Metryki skuteczności ataków.

Moduł 7: Automatyzacja testów i integracja z cyklem DevSecOps

Omówienie sposobów włączenia Red Teamingu AI do procesów CI/CD. Uczestnicy poznają dobre praktyki integracji testów bezpieczeństwa z pipeline'ami rozwojowymi.

- Integracja Red Teamingu z CI/CD.
- Dobre praktyki bezpieczeństwa w pipeline'ach.
- Przykłady automatyzacji testów.

Moduł 8: Przyszłość AI jako wektor zagrożeń i przewaga strategiczna
Dyskusja o roli AI jako zarówno narzędzia innowacji, jak i potencjalnego źródła zagrożeń. Jak organizacje mogą wykorzystać Red Teaming jako przewagę konkurencyjną.

- AI jako narzędzie innowacji i źródło zagrożeń.
- Red Teaming jako przewaga konkurencyjna.
- Trendy w zabezpieczaniu modeli AI.

Moduł 9: Podsumowanie, pytania i dalsze kroki
Zebranie kluczowych wniosków, sesja Q&A oraz przekazanie materiałów dodatkowych i rekomendacji do dalszego rozwoju kompetencji w zakresie bezpieczeństwa AI.

- Podsumowanie kluczowych zagadnień i wnioski.
- Sesja Q&A.
- Materiały dodatkowe i rekomendacje.

Zapisz się na szkolenie: https://zalnet.pl/edu/bezpieczenstwo-ai-i-red-teaming-z-wykorzystaniem-narzedzi-microsoft/