Szkolenie AI 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.

Kod szkolenia: AI-RED-TEAM-01

Kategoria: AI, Red Team

Trenerka: Beata Zalewa

Czas trwania: 2 dni / 16 godzin (9:00 – 17:00)

Poziom zaawansowania: Średniozaawansowany

Język wykładowy: język polski

Forma szkolenia: zdalne. Po wcześniejszym uzgodnieniu możliwe szkolenie w siedzibie klienta.

Materiały: 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.

Efekty szkolenia:

- Uczestnik zna zaawansowaną metodologię AI Red Teaming oraz architekturę i specyfikację techniczną narzędzi firmy Microsoft, takich jak MS PyRIT oraz AI Red Teaming Agent.
- Uczestnik potrafi samodzielnie skonfigurować dedykowane środowisko testowe i zintegrować je z interfejsami API badanych modeli w celu pełnej automatyzacji testów penetracyjnych.

- Uczestnik potrafi projektować i przeprowadzać złożone symulacje ataków (m.in. jailbreaking, prompt injection czy manipulacja kontekstem) oraz oceniać odporność systemów AI za pomocą automatycznych ewaluatorów.
- Uczestnik rozumie naturę specyficznych podatności oraz wektorów zagrożeń charakterystycznych dla generatywnej sztucznej inteligencji (LLM).
- Uczestnik rozumie kluczową rolę ciągłego testowania i monitorowania modeli w procesie bezpiecznego oraz zgodnego z zasadami Responsible AI wdrażania technologii sztucznej inteligencji w organizacji.

Co otrzymasz?

- Materiały szkoleniowe.
- Szkolenie kończy się certyfikatem uczestnictwa.

Plan szkolenia

Godzina	Czas trwania	Moduł	Forma
Dzień 1			
9:00 – 9:30	30 minut	Moduł 1: Wprowadzenie i cele szkolenia	Prezentacja, dyskusja
9:30 – 10:30	60 minut	Moduł 2: Wprowadzenie do zagrożeń AI i podejścia "shift-left"	Teoria, przykłady
10:30 – 10:45	15 minut	Przerwa na kawę	—
10:45 – 12:30	105 minut	Moduł 3: AI Red Teaming – podstawy i strategie ataku	Teoria, analiza przypadków, demo
12:30 – 13:00	30 minut	Przerwa obiadowa	—
13:00 – 14:45	105 minut	Moduł 4: Testowanie odporności modeli z Microsoft PyRIT (Python Risk Identification Tool)	Pokaz, konfiguracja środowiska, praktyka
14:45 – 15:00	15 minut	Przerwa na kawę	—

Godzina	Czas trwania	Moduł	Forma
15:00 – 16:45	105 minut	**Moduł 5: Symulacja cyberataków za pomocą AI Red Teaming Agent**	Pokaz, konfiguracja środowiska, praktyka
16:45 – 17:00	15 minut	**Sesja Q&A**	Pytania
Dzień 2			
9:00 – 9:15	15 minut	**Podsumowanie dnia pierwszego**	Podsumowanie
9:15 – 11:15	120 minut	**Moduł 6: Symulacja ataków – Jailbreak, Base64, Leetspeak, Unicode**	Laboratorium praktyczne
11:15 – 11:30	15 minut	Przerwa na kawę	—
11:30 – 12:30	60 minut	**Moduł 7: Ocena ryzyka i analiza odpowiedzi modeli AI**	Laboratorium praktyczne, analiza wyników
12:30 – 13:00	30 minut	Przerwa obiadowa	—
13:00 – 15:00	120 minut	**Moduł 8: Automatyzacja testów i integracja z cyklem DevSecOps**	Warsztat, dyskusja
15:00 – 15:15	15 minut	Przerwa na kawę	—
15:15 – 16:45	90 minut	**Moduł 9: Przyszłość AI jako wektor zagrożeń i przewaga strategiczna**	Prezentacja
16:45 – 17:00	15 minut	**Moduł 10: 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: Testowanie odporności modeli z Microsoft PyRIT (Python Risk Identification Tool)

Praktyczne zapoznanie się z otwartym frameworkiem Microsoft PyRIT, służącym do identyfikacji ryzyk w modelach generatywnej sztucznej inteligencji (LLM). Uczestnicy skonfigurują środowisko programistyczne w Pythonie i poznają możliwości automatyzacji zaawansowanych testów penetracyjnych oraz oceny (scoringu) odpowiedzi modeli.

- Instalacja MS PyRIT, integracja z API testowanych modeli oraz konfiguracja baz danych do logowania testów.
- Praca z targetami (Targets), orkiestratorami (Orchestrators) oraz automatycznymi ewaluatorami (Scorers).
- Tworzenie powtarzalnych strategii testowych (np. generowanie promptów adversarialnych) i automatyczna analiza podatności AI.

Moduł 5: Symulacja cyberataków za pomocą AI Red Teaming Agent

Praktyczne warsztaty z wykorzystaniem AI Red Teaming Agent (w środowisku Azure AI) do zautomatyzowanego testowania systemów AI. Uczestnicy skonfigurują autonomicznego agenta, który symuluje zachowanie ludzkiego hakera (adversary), i poznają techniki dynamicznego wykrywania luk w bezpieczeństwie.

- Uruchomienie i parametryzacja AI Red Teaming Agent w chmurze Microsoft pod kątem specyfiki testowanej aplikacji.
- Wykorzystanie agenta do wyszukiwania podatności na *jailbreaking*, *prompt injection* oraz manipulację kontekstem w trybie zamkniętej pętli (closed-loop).
- Przegląd wygenerowanych logów, ocena poziomu ryzyka systemu AI oraz automatyzacja testów regresyjnych.

Moduł 6: 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 7: 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 8: 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 9: 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.

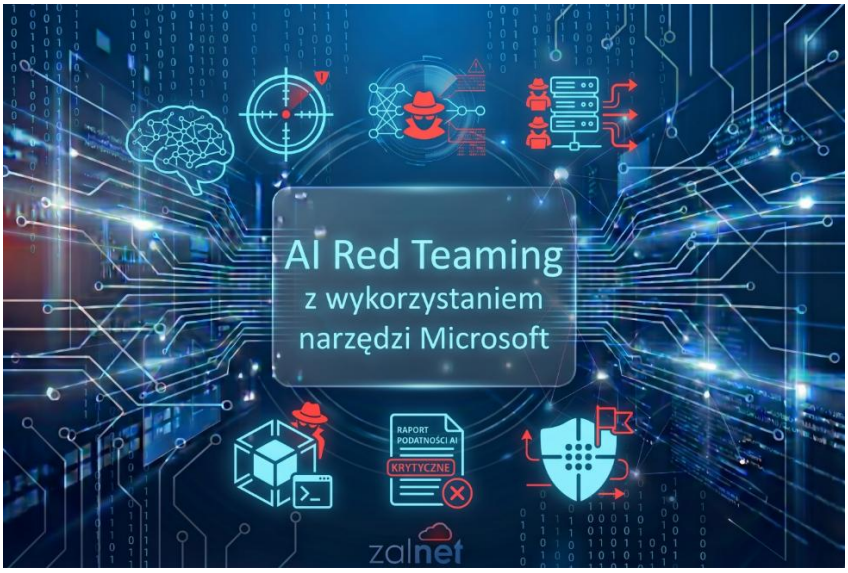
Moduł 10: Podsumowanie, pytania i dalsze kroki

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: szkolenia@zalnet.pl

https://zalnet.pl/edu/ai-red-teaming-z-wykorzystaniem-narzedzi-microsoft/



ZALNET
Beata Zalewa

+48 600 926 031
info@zalnet.pl
www.zalnet.pl

Wola Radzięcka 62
23-440 Frampol
NIP 717 111 99 64