Szkolenie Praktyczne testy penetracyjne i bezpieczeństwo AI i LLM

Opis szkolenia: To szkolenie to odwrócenie proporcji klasycznych szkoleń - ograniczamy czas poświęcony na "zwykłe, zabawne jailbreaki tekstowe" do niezbędnego minimum, traktując LLM jako kolejny komponent złożonej aplikacji webowej połączonej z bazą danych i systemem operacyjnym.

Dzień 3 jest dniem opcjonalnym, przy 2-dniowym szkoleniu cena jak za szkolenie 2-dniowe.

Kod szkolenia: SEC-AI-LLM-01

Kategoria: AI / RED TEAM

Trenerka: Beata Zalewa

Czas trwania: 3 dni / 24 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: Znajomość tematyki związanej z LLMami i podstaw Pythona. Warsztaty opierają się na symulacji systemów, które firmy wdrażają lokalnie lub w prywatnej chmurze (modele open-source, bazy on-premise), co pozwala na realne testowanie podatności sieciowych i konfiguracyjnych.

Grupa decelowa:

- Pentesterzy oraz członków zespołów Red Team, którzy chcą rozwinąć swoje kompetencje o zaawansowane, ofensywne testy systemów sztucznej inteligencji, wychodząc daleko poza proste, tekstowe manipulacje promptami.
- Inżynierowie bezpieczeństwa (Security Engineers), odpowiedzialni za audytowanie, projektowanie architektury systemów AI oraz ocenę ryzyka (Threat Modeling) w środowiskach korporacyjnych.
- Inżynierowie DevSecOps, wdrażający rozwiązania automatyzujące testy bezpieczeństwa oraz odpowiedzialnych za potoki CI/CD walidujące bezpieczeństwo artefaktów AI przed wdrożeniem produkcyjnym.
- Programiści wykorzystujący AI / LLM oraz architekci oprogramowania, twórcy aplikacji wykorzystujących frameworki takie jak LangChain czy LlamaIndex, którzy chcą zrozumieć podatności implementacyjne oraz metody skutecznej obrony swoich systemów.

Cel szkolenia: Głównym celem szkolenia jest przekazanie zaawansowanej, praktycznej wiedzy z zakresu ofensywnego testowania i zabezpieczania systemów opartych na sztucznej inteligencji oraz dużych modelach językowych (LLM). Warsztaty kładą silny nacisk na traktowanie modelu jako integralnego komponentu złożonej architektury IT, wychodząc daleko poza proste techniki manipulacji promptami. Podczas zajęć kursanci nauczą się identyfikować i eksploatować krytyczne luki bezpieczeństwa w warstwach orkiestracji, bazach wektorowych, autonomicznych

agentach oraz architekturze RAG. Istotnym elementem kursu jest również praktyczne wdrożenie rynkowych standardów oceny ryzyka, ze szczególnym uwzględnieniem OWASP Top 10 for LLM Applications. W ostatecznym rozrachunku szkolenie przygotowuje specjalistów do kompleksowego modelowania zagrożeń i projektowania skutecznych mechanizmów obronnych w nowoczesnych środowiskach produkcyjnych.

Efekty szkolenia:

- Uczestnik zna najnowsze wektory ataków na systemy sztucznej inteligencji i aplikacje oparte o LLM, w tym zaawansowane techniki obchodzenia zabezpieczeń typu Guardrails oraz zagrożenia ukryte w łańcuchu dostaw modeli.
- Uczestnik rozumie kompleksową anatomię nowoczesnych środowisk AI oraz specyfikę modelowania zagrożeń (Threat Modeling), co pozwala mu na precyzyjną ocenę ryzyka w architekturach on-premise oraz chmurowych.
- Uczestnik potrafi samodzielnie przeprowadzić głębokie testy penetracyjne, polegające m.in. na eskalacji uprawnień w architekturze RAG, zatruwaniu potoków danych, atakach na bramki API oraz wykonywaniu złośliwego kodu (RCE) poprzez luki w autonomicznych agentach.
- Uczestnik potrafi automatyzować procesy audytu bezpieczeństwa przy użyciu specjalistycznych narzędzi oraz skutecznie wdrażać strategie mitygacji luk zgodnie z dobrymi praktykami DevSecOps.

Co otrzymasz?

- Materiały szkoleniowe.
- Szkolenie kończy się certyfikatem uczestnictwa.

Agenda szkolenia

Godzina	Czas trwania	Moduł	Forma
Dzień 1			
9:00 – 9:15	15 minut	**Moduł 1: Powitanie i omówienie**	Powitanie, interaktywna ankieta
9:15 – 10:30	75 minut	**Moduł 2: Wprowadzenie do architektury AI i modelowania zagrożeń**	Teoria, przykłady
10:30 – 10:45	15 minut	Przerwa	-

Godzina	Czas trwania	Moduł	Forma
10:45 – 11:45	60 minut	**Moduł 2: Wprowadzenie do architektury AI i modelowania zagrożeń**	Case studies
11:45 – 12:30	45 minut	**Moduł 3: Ofensywne testowanie modeli LLM**	Teoria, przykłady
12:30 – 13:00	30 minut	Przerwa obiadowa	-
13:00 – 15:00	120 minut	**Moduł 3: Ofensywne testowanie modeli LLM**	Laboratorium praktyczne, analiza wyników
15:00 – 15:15	15 minut	Przerwa	-
15:15 – 16:45	90 minut	**Moduł 3: Ofensywne testowanie modeli LLM**	Laboratorium praktyczne, analiza wyników
16:45 – 17:00	15 minut	**Sesja Q&A**	Pytania
Dzień 2			
9:00 – 9:15	15 minut	**Podsumowanie dnia pierwszego**	Podsumowanie
9:15 – 10:30	75 minut	**Moduł 4: Bezpieczeństwo danych AI – RAG, osadzenia i zatruwanie**	Teoria, laboratorium
10:30 – 10:45	15 minut	Przerwa	-
10:45 – 12:30	105 minut	**Moduł 4: Bezpieczeństwo danych AI – RAG, osadzenia i zatruwanie**	Laboratorium praktyczne, analiza wyników
12:30 – 13:00	30 minut	Przerwa obiadowa	-

Godzina	Czas trwania	Moduł	Forma
13:00 – 15:00	120 minut	**Moduł 5: Testowanie infrastruktury AI i bezpieczeństwo API**	Teoria, laboratorium
15:00 – 15:15	15 minut	Przerwa	-
15:15 – 16:45	90 minut	**Moduł 5: Testowanie infrastruktury AI i bezpieczeństwo API**	Laboratorium praktyczne, analiza wyników
16:45 – 17:00	15 minut	**Sesja Q&A**	Pytania
Dzień 3 (opcjonalny)			
9:00 – 9:15	15 minut	**Podsumowanie dnia drugiego**	Podsumowanie
9:15 – 10:30	75 minut	**Moduł 6: Warsztaty Red Team – scenariusz End-to-End**	Warsztaty
10:30 – 10:45	15 minut	Przerwa	-
10:45 – 12:30	105 minut	**Moduł 6: Warsztaty Red Team – scenariusz End-to-End**	Warsztaty
12:30 – 13:00	30 minut	Przerwa obiadowa	-
13:00 – 15:00	120 minut	**Moduł 6: Warsztaty Red Team – scenariusz End-to-End**	Warsztaty
15:00 – 15:15	15 minut	Przerwa	-
15:15 – 16:45	90 minut	**Moduł 6: Warsztaty Red Team – scenariusz End-to-End**	Warsztaty
16:45 – 17:00	15 minut	**Moduł 7: 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.

Moduł 2: Wprowadzenie do architektury AI i modelowania zagrożeń

- Anatomia nowoczesnej aplikacji LLM: Frontend, API Gateway, orkiestracja (LangChain/LlamaIndex), wektorowe bazy danych, modele hostowane lokalnie vs. zewnętrzne API.
- Filozofia i struktura projektów *OWASP Top 10 for LLM Applications* oraz *OWASP AI Testing Guide*.
- Szacowanie ryzyka i modelowanie zagrożeń (Threat Modeling) specyficznych dla systemów uczących się.
- Praktyka / Case Studies: Analiza podatności architektury referencyjnej aplikacji korporacyjnej. Mapowanie punktów styku danych i identyfikacja powierzchni ataku poza warstwą interfejsu użytkownika.

Moduł 3: Ofensywne testowanie modeli LLM

- Mechanizmy obronne modeli (Guardrails: np. LlamaGuard, NeMo Guardrails) – zasada działania i słabe punkty.
- Kwestia halucynacji i uprzedzeń (biases) jako wektor ataku na logikę biznesową.
- Ataki kradzieży intelektualnej: *Model Extraction* (odwracanie parametrów) oraz *Model Inversion*.
- Laboratoria praktyczne:
    - Laboratorium 3.1: Zaawansowany *Jailbreaking* i *Guardrail Bypass* – obchodzenie filtrów bezpieczeństwa za pomocą tokenów kontrolnych, technik wieloetapowych (multi-turn) oraz kodowania lingwistycznego.
    - Laboratorium 3.2: Wykorzystanie narzędzia promptfoo do automatyzacji testów penetracyjnych zachowania modelu w celu wykrywania podatności na manipulację aplikacją.
    - Laboratorium 3.3: Symulacja ataku *Model Extraction* – odpytywanie czarnej skrzynki (black-box API) w celu odtworzenia zachowania i funkcjonalności modelu bazowego.

Moduł 4: Bezpieczeństwo danych AI – RAG, osadzenia i zatruwanie

- Architektura RAG (Retrieval-Augmented Generation) i jej krytyczne punkty podatności.
- Ataki na dane i zatruwanie danych treningowych (*Training Data Poisoning*) vs zatruwanie danych kontekstowych w czasie rzeczywistym.
- Bezpieczeństwo osadzeń (*Embeddings*) i matematyczna manipulacja wektorami przestrzeni semantycznej.

- Laboratoria praktyczne:
    - Laboratorium 4.1: *RAG Data Leakage* – konstrukcja zapytań zmuszających LLM do nieautoryzowanego przeszukania i ujawnienia poufnych dokumentów z bazy danych, do których użytkownik nie powinien mieć dostępu.
    - Laboratorium 4.2: Zatruwanie potoku danych (*Embedding Poisoning*) – wstrzykiwanie złośliwych danych do procesów ETL, co skutkuje trwałym zniekształceniem wyników wyszukiwania semantycznego dla innych użytkowników.

Moduł 5: Testowanie infrastruktury AI i bezpieczeństwo API

- Podatności środowisk wykonawczych i hostingu modeli (np. vLLM, Ollama, Open WebUI).
- Architektura i bezpieczeństwo baz wektorowych (np. Milvus, Qdrant, Pinecone) – brak uwierzytelnienia, wstrzykiwanie poleceń bazodanowych.
- Zagrożenia w łańcuchu dostaw AI: Ukryte exploity w formatach plików modeli (np. podatności *pickle* w plikach .bin / .pth vs bezpieczniejsze formaty jak *Safetensors*).
- Podatności wtyczek (Plugins) i autonomicznych agentów AI wykonujących kod (RCE).
- Laboratoria praktyczne:
    - Laboratorium 5.1: Testy bramki API i autoryzacji (Model Gateway Security) – przełamywanie kontroli dostępu opartej na rolach (RBAC) skonfigurowanej za pomocą rozwiązań takich jak DreamFactory, Kong czy Tyk.
    - Laboratorium 5.2: Wykorzystanie agenta AI – ucieczka z piaskownicy (sandbox escape) i wykonanie zdalnego kodu (RCE) na serwerze hostującym infrastrukturę poprzez lukę we wtyczce systemowej.

Moduł 6: Warsztaty Red Team – scenariusz End-to-End (Opcjonalny)

- Praktyczne warsztaty podsumowujące:
    - Uczestnicy stają przed wyzwaniem typu Red Team w przygotowanym środowisku laboratoryjnym (skonteneryzowany stos: Open WebUI + vLLM + Milvus + DreamFactory API Gateway).
    - Zadanie: Przeprowadzenie pełnego łańcucha ataku (Kill Chain) – od rekonesansu infrastruktury API, przez zatrucie danych semantycznych, aż po obejście guardrailu modelu i eskalację uprawnień do hosta za pomocą wtyczki.
    - Krótkie omówienie strategii mitygacji w duchu DevSecOps: jak wdrożyć potoki CI/CD walidujące bezpieczeństwo artefaktów AI przed wdrożeniem produkcyjnym.

Moduł 7: 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 i modeli LLM.

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

Zarejestruj się na szkolenie: szkolenia@zalnet.pl

https://zalnet.pl/edu/praktyczne-testy-penetracyjne-i-bezpieczenstwo-ai-i-llm

ZALNET
Beata Zalewa

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

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