VIGIL GUARD

Vigil Guard Products
Karta Produktu

Vigil Guard Enterprise

Platforma Wykrywania i Reagowania na Zagrożenia AI

WYZWANIA

Organizacje wdrażające modele LLM stają przed nowymi zagrożeniami, na które tradycyjne narzędzia bezpieczeństwa nie są przygotowane. Dane, instrukcje sterujące i interakcja użytkownika przepływają przez wspólny kanał: język naturalny. Nie ma wyraźnego rozdzielenia między „poleceniem" a „ładunkiem". Tradycyjne zapory WAF, systemy DLP i EDR nie są zaprojektowane do ochrony przed zagrożeniami opartymi na języku naturalnym.

Główne wyzwania:

- Ataki prompt injection manipulują zachowaniem AI
- Wycieki danych wrażliwych (PII, poświadczenia) w odpowiedziach modelu
- Użytkownicy omijają zabezpieczenia (jailbreaki)
- Brak śladu audytowego interakcji z systemami AI
- Niekontrolowane użycie AI na platformach LLM (Shadow AI)
- **Asystenci odchodzą od zdefiniowanej roli, zwiększając koszty i ryzyko regulacyjne**

KLUCZOWE KORZYŚCI

- **6-warstwowe równoległe wykrywanie** eliminuje pojedyncze punkty awarii
- Ochrona PII dla 20+ typów encji, w tym polskich (PESEL, NIP, REGON)
- Segmentacja polityk per-klucz API między systemami i grupami użytkowników
- Rozszerzenie przeglądarki monitoruje platformy LLM (Shadow AI)
- Natywna integracja z SIEM (CEF/JSON) do Splunk, QRadar, Sentinel
- Pełne wdrożenie lokalne, gotowe do pracy w sieci odizolowanej, bez GPU
- System uczenia FP/FN, poprawia się z feedbackiem zespołu
- Zaprojektowany od podstaw do obsługi języka polskiego i angielskiego
- **Wykrywanie Dryfu Semantycznego**: klasyfikacja ON / NEAR / OFF_SCOPE dla każdego promptu, osobna reguła per klucz API

< 300 ms	6 warstw	3 poziomy	20+
opóźnienia P99	równoległe wykrywanie	klasyfikacji zgodności z zakresem	typów encji PII

On-premise · Air-gapped · Bez wymogu GPU

CZYM JEST VIGIL GUARD ENTERPRISE?

Vigil Guard Enterprise to platforma wykrywania i reagowania na zagrożenia związane z AI (AIDR), która daje zespołom bezpieczeństwa pełną kontrolę nad użyciem modeli LLM w organizacji. Działa jako pośrednik między aplikacjami a dostawcami LLM, analizując każde zapytanie i odpowiedź w czasie rzeczywistym. Zapobiega atakom prompt injection, wyciekom danych wrażliwych, generowaniu szkodliwych treści, omijaniu zabezpieczeń przez użytkowników oraz pytaniom wykraczającym poza zakres roli asystenta.

Po raz pierwszy dyrektor ds. bezpieczeństwa informacji (CISO) zyskuje kompleksowy wgląd w wektory ataków, które wcześniej były „czarną skrzynką": co trafia do modelu LLM, co z niego wychodzi, oraz możliwość egzekwowania polityk bezpieczeństwa zgodnych z wymaganiami organizacji.

PROBLEM

Organizacje wdrażające modele LLM stają przed nowymi zagrożeniami, na które tradycyjne narzędzia bezpieczeństwa nie są przygotowane:

- **Prompt injection**: atakujący manipulują zachowaniem AI za pomocą spreparowanych danych wejściowych
- **Wycieki danych**: wrażliwe informacje (PII, poświadczenia, dokumenty wewnętrzne) są ujawniane w odpowiedziach modelu
- **Jailbreaki**: użytkownicy omijają zabezpieczenia, aby uzyskać dostęp do szkodliwych treści
- **Bezpieczeństwo treści**: generowanie toksycznych, nienawistnych lub nielegalnych treści
- **Brak śladu audytowego**: brak rejestracji tego, co wchodzi i wychodzi z systemów AI
- **Zgodność z zakresem**: asystenci otrzymują prompty poza swoją zdefiniowaną rolą (koszt, regulacje, reputacja)

Tradycyjne zapory aplikacji webowych (WAF), systemy zapobiegania utracie danych (DLP) i systemy wykrywania i reagowania na zagrożenia (EDR) nie są zaprojektowane do ochrony przed zagrożeniami opartymi na języku naturalnym.

DLACZEGO JĘZYK MA ZNACZENIE W BEZPIECZEŃSTWIE AI?

Modele LLM są fundamentalnie inne niż inne systemy w Twoim środowisku. Dane, instrukcje sterujące i interakcja użytkownika przepływają przez wspólny kanał: język naturalny. Nie ma wyraźnego rozdzielenia między „poleceniem" a „ładunkiem". Atak prompt injection i legalne zapytanie użytkownika wyglądają strukturalnie identycznie, różnica leży w semantyce i kontekście językowym. **Vigil Guard został zaprojektowany od podstaw do obsługi języka polskiego i angielskiego.** Nasze modele uczenia maszynowego są trenowane na dwujęzycznych zbiorach danych, a system wykrywania PII zawiera polskie narzędzia do identyfikacji encji, weryfikowane za pomocą sum kontrolnych.

KIEDY PROMPT INJECTION TO ZA MAŁO: PROBLEM ZGODNOŚCI Z ZAKRESEM

Ochrona przed atakami to jedno. Drugie, równie istotne, to pytanie: **czy użytkownik w ogóle rozmawia o tym, o czym ma rozmawiać?** Asystent bankowy pytany o przepisy kulinarne, bot wsparcia technicznego proszony o pisanie wierszy, asystent HR wypytywany o doradztwo podatkowe: to nie są ataki, ale realne zagrożenia, koszty LLM, ryzyko reputacyjne, utrata kontroli nad granicami produktu. Vigil Guard adresuje ten problem osobnym modułem: **Wykrywanie Dryfu Semantycznego**, opisanym w Kluczowych Funkcjonalnościach.

JAK TO DZIAŁA

Aplikacja wysyła zapytanie do Vigil Guard przez interfejs API REST (jedno wywołanie, średnio < 300 ms opóźnienia P99). Platforma uruchamia sześć niezależnych warstw wykrywania równolegle:

WARSTWA WYKRYWANIA	FUNKCJA
Wykrywanie języka	Identyfikacja języka promptu (PL / EN / inne) do routowania kolejnych warstw do modeli i słowników specyficznych dla danego języka
Analiza heurystyczna	Wykrywanie wzorców obfuskacji, sztuczek kodowania i wstrzykiwania kodu
Analiza semantyczna	Porównywanie wektorów z znanymi wzorcami ataków (z ciągłym uczeniem na podstawie feedbacku)
Klasyfikacja ML	Dedykowany model detekcji ataków dostrojony do wykrywania prompt injection (obsługa polskiego i angielskiego)
Moderacja treści	Wykrywanie toksyczności, mowy nienawiści i szkodliwych treści w 9 kategoriach bezpieczeństwa
Wykrywanie Dryfu Semantycznego	Klasyfikacja zgodności z zakresem asystenta (ON / NEAR / OFF_SCOPE), konfigurowana per klucz API, działa jako *late enrichment* (*fail-open*)

Inteligentny arbitraż łączy wszystkie sygnały w jeden wynik ważony i podejmuje decyzję: **ZEZWÓL**, **ZABLOKUJ** lub **OCZYŚĆ**. Jeśli dozwolona treść zawiera dane osobowe (PII), są one automatycznie redagowane przed przesłaniem. Każda decyzja jest rejestrowana w bazie danych analitycznych z pełnym śladem audytowym.

KLUCZOWE FUNKCJONALNOŚCI

Wykrywanie Dryfu Semantycznego

Niezależny moduł analityczny sprawdzający, czy prompt mieści się w skonfigurowanym zakresie asystenta. 3-poziomowa etykieta per zapytanie (**ON / NEAR / OFF_SCOPE**), akcja per poziom (ZEZWÓL / ZABLOKUJ), własny komunikat blokady, szyfrowany AES Scope Definition per klucz API, trzy poziomy czułości (Relaxed / Balanced / Strict). Autorski dwujęzyczny klasyfikator, równolegle z rdzennymi warstwami, *fail-open*. Dashboard pokazuje rozkład i ranking kluczy API; zdarzenia trafiają do SIEM.

Ochrona danych osobowych (PII)

Automatyczne wykrywanie i redagowanie ponad 20 typów encji: PESEL, NIP, REGON, SSN, numery kart (walidacja Luhn), IBAN, e-mail, numery telefonów, nazwiska. Trzy poziomy redakcji (lekki, ciężki, blokada). Konfigurowalne tryby: zamiana, hashowanie lub maskowanie.

Uczenie się na podstawie fałszywych alarmów

Gdy zespół zgłosi fałszywy alarm (FP/FN) przez panel sterowania, system indeksuje ten prompt w pamięci wektorowej. Przyszłe podobne prompty otrzymują automatyczne korekty wyników; system staje się mądrzejszy bez ponownego trenowania modeli.

Własne polityki wykrywania i reakcji

Każdy klucz API ma przypisany własny zestaw reguł, główny mechanizm segmentacji polityk między systemami i grupami użytkowników. Zestaw reguł kontroluje aktywne warstwy i ich wagi, typy encji PII i tryb redakcji, progi moderacji, wzorce DSL i próg blokady.

Integracja z SIEM

Wbudowany forwarder SIEM eksportuje zdarzenia przez TCP lub TLS w formacie CEF lub JSON. Cele: Splunk, IBM QRadar, Microsoft Sentinel, Elastic SIEM lub dowolny kolektor syslog. Konfiguracja przez panel webowy, wbudowane testowanie połączenia.

Moderacja treści

Klasyfikacja treści promptu i odpowiedzi w 9 zunifikowanych kategoriach bezpieczeństwa (mowa nienawiści, toksyczność, samookaleczenie, przemoc i inne). Reguły akcji per kategoria (ZEZWÓL / OCZYŚĆ / ZABLOKUJ). Architektura z dwoma klasyfikatorami obsługuje natywnie polski oraz 7+ języków. Pracuje równolegle i nigdy nie blokuje decyzji rdzenia przy opóźnieniu (*fail-open*).

POKRYCIE MODELI WDROŻENIA AI

Organizacje nie używają LLM w jeden sposób. Vigil Guard obsługuje każdy popularny wzorzec wdrożenia, zapewniając zespołowi bezpieczeństwa jeden punkt egzekwowania polityk, niezależnie od sposobu korzystania z AI:

MODEL WDROŻENIA	JAK VIGIL GUARD TO OBSŁUGUJE	INTEGRACJA
LLM w przeglądarce	Rozszerzenie Chrome przechwytuje zapytania przed wysłaniem	Manifest v3 (GPO / Intune / Jamf)
Automatyzacja (n8n, Make)	Węzły ochronne na wejściu/wyjściu opakowują każde wywołanie LLM	Pakiet węzłów n8n community
Centralny proxy LLM (LiteLLM)	Backendowa zapora bezpieczeństwa inspekcjonuje cały ruch przez proxy	Adapter HTTP dla LiteLLM guardrail API
Aplikacje własne	Wywołanie API przed i po interakcji z LLM	Python SDK + REST endpoint
Przetwarzanie wsadowe	Do 100 zapytań na wywołanie	POST /v1/guard/batch

Każdy punkt integracji zasila ten sam pipeline wykrywania, ten sam silnik reguł i ten sam eksport do SIEM. Zespół bezpieczeństwa otrzymuje ujednolicony widok całego ruchu LLM w organizacji. Wykrywanie Dryfu Semantycznego działa dla każdego z tych wzorców; zakres jest zdefiniowany per klucz API, z własną czułością i polityką akcji.

SUWERENNOŚĆ DANYCH I KONTROLA (ON-PREMISE)

Nic nie opuszcza Twojej sieci

Cała platforma działa na Twojej infrastrukturze jako stos Docker Compose. Brak zewnętrznych wywołań API, brak telemetrii chmurowej, brak mechanizmów phone-home. Wszystkie modele uczenia maszynowego (klasyfikator ataków prompt injection, moderacja treści, wykrywanie PII, detekcja języka, dryf semantyczny) są dostarczane jako binaria ONNX wbudowane w obrazy kontenerów. Inferencja działa lokalnie na CPU, bez wymogu GPU. System działa w pełni w sieci odizolowanej (air-gapped).

Ty kontrolujesz politykę bezpieczeństwa

Zestawy reguł per klucz API definiują aktywne warstwy, typy encji PII i tryb redakcji, progi moderacji oraz własne wzorce DSL. Wszystkie progi, wagi i flagi są konfigurowalne przez panel webowy bez restartów; zmiany propagują się natychmiast.

Szyfrowanie definicji zakresu

Definicje zakresu (Scope Definitions) są szyfrowane AES w spoczynku, z obsługą rotacji kluczy. Żaden plaintext scope nie jest logowany; w audycie widnieje jedynie SHA-256 fingerprint. Identyczna polityka szyfrowania jak dla reguł DSL i list redakcji PII.

Automatyczne profile skalowania

Instalator wykrywa parametry hosta (CPU, RAM, dysk) i dobiera profil skalowania z gotowej matrycy, zapisując deterministyczny plik konfiguracyjny dla powtarzalnych wdrożeń. Runtime budgets (wątki ML, limity concurrency) dopasowują się do rzeczywistych zasobów; profil zmieniany po wdrożeniu dedykowanym skryptem, bez ręcznych edycji.

Opcje wdrożenia enterprise

- **On-premise**: jeden serwer Linux, Docker Compose, x86_64. Instalator dobiera profil skalowania na podstawie parametrów hosta.
- **Air-gapped**: pełna funkcjonalność bez dostępu do internetu.

MAPA OBAW CISO

OBAWA CISO	JAK VIGIL GUARD TO ROZWIĄZUJE
Widoczność LLM	Ujednolicony widok całego ruchu LLM: przeglądarka, automatyzacja, proxy, aplikacje własne
Segmentacja polityk	Zestawy reguł przypisane do kluczy API z własnymi progami, redakcją i konfiguracją wykrywania
Suwerenność danych	W pełni lokalne wdrożenie, gotowe do pracy w sieci odizolowanej. Żadne dane nie opuszczają sieci organizacji
Ataki prompt injection	6 równoległych warstw wykrywania z ważonym wynikiem, brak pojedynczych punktów awarii
Wycieki PII	20+ typów encji (w tym polskie: PESEL, NIP, REGON), automatyczna redakcja przed wysłaniem do modelu
Shadow AI	Rozszerzenie przeglądarki dla platform LLM, wdrażane przez GPO / Intune / Jamf
Zgodność z rolą asystenta	Wykrywanie Dryfu Semantycznego: 3-stopniowa klasyfikacja, akcja per poziom, definicja zakresu szyfrowana AES per klucz API
SIEM / SOC	Natywny eksport CEF/JSON do Splunk, QRadar, Sentinel, Elastic przez TCP/TLS
Zgodność / audyt	Pełne logowanie zdarzeń, ślad audytowy silnika reguł, konfigurowalna retencja danych
Fałszywe alarmy	System pamięci FP/FN uczący się z feedbacku zespołu, poprawia się z czasem

OBSERWOWALNOŚĆ I AUDYT

- **Panel webowy**: metryki wykrywania w czasie rzeczywistym, historia zdarzeń, raportowanie fałszywych alarmów, zarządzanie konfiguracją
- **Eksport do SIEM**: forwardowanie CEF/JSON przez TCP/TLS do Splunk, QRadar, Sentinel lub dowolnego kolektora syslog
- **Scope Drift Analytics**: dedykowany panel w dashboardzie: rozkład ON / NEAR / OFF_SCOPE oraz ranking kluczy API generujących dryf

LICENCJONOWANIE

7-dniowy okres próbny przy pierwszym uruchomieniu, wszystkie funkcje odblokowane, bez klucza licencyjnego. Licencja subskrypcyjna aktywowana przez panel sterowania. Cennik i plany dostępne na vigilguard.ai/pricing.

ON-PREMISE
Docker Compose, x86_64, auto-skalowanie

AIR-GAPPED
Bez phone-home, bez chmury

DWUJĘZYCZNY
Polski + angielski, natywny ML

SZYFROWANIE AES
Scope, DSL, listy redakcji w spoczynku

Nie możesz zabezpieczyć tego, czego nie widzisz.

AI jest już częścią Twojego środowiska. Vigil Guard sprawia, że staje się widoczne, kontrolowalne i bezpieczne.

CO NAS WYRÓŻNIA: WYKRYWANIE DRYFU SEMANTYCZNEGO

Pierwsza platforma AIDR, która nie tylko blokuje ataki, ale też trzyma Twoich asystentów w granicach ich zdefiniowanej roli. 3-stopniowa klasyfikacja ON / NEAR / OFF_SCOPE, polityka konfigurowana per klucz API, autorski dwujęzyczny model. Atak i dryf w jednym panelu.

ROZPOCZNIJ 7-DNIOWY DARMOWY TEST

Wszystkie funkcje odblokowane. Bez klucza licencyjnego. Wdrożenie w kilka minut.

To nie jest projekt badawczy. To produkcyjne zabezpieczenie AI.

Dowiedz się więcej: https://www.vigilguard.ai/

Rozpocznij darmowy test: https://www.vigilguard.ai/pricing/

Zespoły Bezpieczeństwa
Widoczność, kontrola, audytowalność

Zespoły AI / Platformowe
Bezpieczeństwo bez spowalniania innowacji

Zarządzanie Ryzykiem i Zgodnością
Niższe ryzyko AI, lepsza gotowość audytowa