VIGIL GUARD

Vigil Guard Products
Data Sheet

Vigil Guard Enterprise

AI Detection & Response Platform for full control over LLM usage

CHALLENGES

Organizations adopting LLMs face a new attack surface that traditional security tools were never designed to handle. Data, control instructions, and user interaction all flow through a single shared channel, natural language. There is no protocol separation between "command" and "payload." Traditional WAFs, DLP, and EDR are blind to these language-based threats.

Key challenges:

- Prompt injection attacks manipulate AI behavior through crafted inputs
- Sensitive data (PII, credentials) leaks through model outputs
- Users bypass safety guardrails to extract harmful content
- No audit trail of what goes in and out of AI systems
- Shadow AI usage across LLM platforms is invisible to security
- **Assistants drift out of their intended role, inflating cost and regulatory exposure**

KEY BENEFITS

- **6-layer parallel detection** eliminates single points of failure
- PII protection for 20+ entity types including Polish identifiers (PESEL, NIP, REGON)
- Per-API-key policy segmentation across systems and user groups
- Browser extension monitors LLM platforms for Shadow AI
- Native SIEM integration (CEF/JSON) to Splunk, QRadar, Sentinel
- Full on-premise deployment, air-gap capable, no GPU required
- FP/FN learning system that improves with your team's feedback
- Built from the ground up for bilingual Polish + English operation
- **Semantic Drift Detection**: ON / NEAR / OFF_SCOPE classification per prompt, configured per API key

< 300 ms	6 layers	3 levels	20+
P99 latency	parallel detection	scope compliance	PII entity types

On-Premise · Air-Gapped · No GPU Required

WHAT IT IS

Vigil Guard Enterprise is an AI Detection and Response (AIDR) platform that gives security teams full observability and policy control over LLM usage across the organization. It sits between your applications and LLM providers, inspecting every prompt and every model response in real time, blocking prompt injection attacks, jailbreak attempts, PII leakage, harmful content, and off-scope prompts before they reach your models or leave your infrastructure.

For the first time, the CISO gets a unified view of an attack vector that was previously a black box: what goes into LLMs, what comes out, and the ability to enforce detection and response policies aligned with the organization's security posture.

THE PROBLEM

Organizations adopting LLMs face a new attack surface that traditional security tools were never designed to handle:

- **Prompt injection**: attackers manipulate AI behavior through crafted inputs
- **Data exfiltration**: sensitive data (PII, credentials, internal documents) leaks through model outputs
- **Jailbreaks**: users bypass safety guardrails to extract harmful content
- **Content safety**: toxic, hateful, or illegal content generated by models
- **Compliance gaps**: no audit trail of what goes in and out of AI systems
- **Scope compliance**: assistants getting prompts outside their intended role (cost, regulation, reputation)

Traditional WAFs and DLP tools were not designed for natural language threats. EDR protects endpoints but is blind to prompts. DLP protects data at rest but doesn't see generated content. WAF understands HTTP traffic but has no understanding of LLM intent.

WHY LANGUAGE MATTERS IN AI SECURITY

LLMs are fundamentally different from every other system in your stack. Data, control instructions, and user interaction all flow through a single shared channel, natural language. There is no protocol separation between "command" and "payload." An attacker's prompt injection and a legitimate user query look structurally identical, the difference is semantic and language-dependent. **Vigil Guard was built from the ground up for bilingual Polish + English operation.** The ML models were trained on bilingual datasets, and the PII engine includes Polish-specific entity recognizers with checksum validation.

WHEN PROMPT INJECTION ISN'T ENOUGH: THE SCOPE COMPLIANCE PROBLEM

Attack defense is one concern. Another, equally important, is the question: **is the user even talking about what the assistant is supposed to talk about?** A banking assistant asked for recipes, a tech-support bot asked to write poetry, an HR assistant asked for tax advice: these aren't attacks, but they are real exposures, wasted LLM cost, reputational risk, and loss of control over product boundaries. Vigil Guard addresses this with a dedicated module: **Semantic Drift Detection**, described in Key Capabilities.

HOW IT WORKS

Your application sends a prompt to Vigil Guard via REST API (single call, < 300 ms P99). The platform runs six independent detection branches in parallel:

DETECTION LAYER	FUNCTION
Language Detection	Identifies the prompt's language (PL / EN / other) to route downstream layers into language-specific models and dictionaries
Heuristic Analysis	Pattern matching for obfuscation, encoding tricks, and code injection
Semantic Analysis	Vector similarity against known attack patterns (with continuous learning from feedback)
ML Classification	Dedicated injection detection model fine-tuned for prompt injection detection (Polish + English)
Content Moderation	Toxicity, hate speech, and harmful content detection across 9 safety categories
Semantic Drift Detection	Assistant scope compliance classification (ON / NEAR / OFF_SCOPE), configured per API key, runs as *late enrichment* (*fail-open*)

An intelligent arbiter aggregates all signals into a single weighted score and delivers a clear verdict: **ALLOW**, **BLOCK**, or **SANITIZE**. If PII is detected in allowed content, it is automatically redacted before forwarding. Every decision is logged to an analytics database with full audit trail.

KEY CAPABILITIES

Semantic Drift Detection

Independent analytics module checking whether a prompt falls within the assistant's configured scope. 3-level label per request (**ON / NEAR / OFF_SCOPE**), per-level action (ALLOW / BLOCK), custom block message, AES-encrypted Scope Definition per API key, three sensitivity levels (Relaxed / Balanced / Strict). Proprietary bilingual classifier, runs alongside core layers, *fail-open* on delay. Dashboard shows distribution and per-key drift ranking; events flow to your SIEM.

PII Protection (Polish + English)

Automatic detection and sanitization of 20+ entity types including PESEL, NIP, REGON, SSN, credit cards (Luhn-validated), IBAN, email, phone numbers, personal names. Three sanitization levels: light, heavy, block. Configurable redaction modes: replace, hash, or mask.

False Positive / False Negative Learning

When your team reports a false positive or false negative through the dashboard, the system indexes that prompt into vector memory. Future similar prompts receive automatic score adjustments; the system gets smarter over time without retraining models.

Custom Detection and Response Policies

Every API key maps to its own rule set, the core mechanism for policy segmentation across systems and user groups. Each rule set controls which detection branches are active, their weights, PII types and redaction modes, content moderation thresholds, custom DSL redaction patterns, and the block threshold.

SIEM Integration

Built-in SIEM forwarder exports detection events over TCP or TLS in CEF or JSON format. Targets: Splunk, IBM QRadar, Microsoft Sentinel, Elastic SIEM, or any syslog-compatible collector. Configuration via web dashboard, connection testing built in.

Content Moderation

Classifies prompt and response content across 9 unified safety categories (hate speech, toxicity, self-harm, violence, and others). Per-category action rules (ALLOW / SANITIZE / BLOCK). Dual-classifier architecture covers Polish natively plus 7+ languages. Runs in parallel and never blocks the core decision on delay (*fail-open*).

COVERAGE ACROSS AI DEPLOYMENT MODELS

Organizations don't use LLMs in one way. Vigil Guard covers every common deployment pattern, giving the security team a single enforcement point regardless of how AI is consumed:

DEPLOYMENT MODEL	HOW VIGIL GUARD COVERS IT	INTEGRATION
Browser-based LLMs	Chrome extension intercepts prompts before they leave the browser	Manifest v3 (GPO / Intune / Jamf)
Workflow automation (n8n, Make)	Input/output guard nodes wrap every LLM call in the workflow	n8n community node package
Central LLM proxy (LiteLLM)	Guardrail backend inspects all traffic routed through the proxy	HTTP adapter for LiteLLM guardrail API
Custom applications	REST API call before and after every LLM interaction	Python SDK + REST endpoint
Batch processing	Up to 100 prompts per call	POST /v1/guard/batch

Every integration point feeds into the same detection pipeline, the same rule engine, the same SIEM export. The security team gets a **unified view of all LLM traffic** across the organization. Semantic Drift Detection works across every deployment pattern; scope is defined per API key, with its own sensitivity and per-level actions.

DATA SOVEREIGNTY AND ON-PREMISE CONTROL

Nothing leaves your network

The entire platform runs on your infrastructure as a Docker Compose stack. There are no external API calls, no cloud telemetry, no phone-home mechanisms. All ML models (prompt injection classifier, content moderation, PII detection, language detection, semantic drift) are shipped as ONNX binaries baked into container images. Inference runs locally on CPU, no GPU required. The system works fully air-gapped.

You control the security policy

Per-API-key rule sets define which detection branches are active, what PII entities to redact, content moderation thresholds, and custom redaction patterns. All thresholds, weights, and feature flags are tunable via the web dashboard without service restarts. Changes propagate immediately across all workers. The Rule Engine supports DSL-based redaction patterns; your team defines exactly how sensitive data is handled.

Scope Definition encryption

Scope Definitions are AES-encrypted at rest, with key rotation support. Scope plaintext is never logged; the audit trail carries only the SHA-256 fingerprint. Same encryption policy as DSL rules and PII redaction lists.

Automatic scaling profiles

The installer detects host facts (CPU, RAM, disk) and picks the matching profile from a shipped scaling matrix, writing a deterministic lock file for reproducible deployments. Runtime budgets (ML thread counts, concurrency limits) adapt to actual host resources rather than being hardcoded. The selected profile can be changed after install through a dedicated admin script, without manual config edits.

Enterprise deployment options

- **On-premise**: single Linux host, Docker Compose, x86_64. Installer picks the scaling profile from host facts.
- **Air-gapped**: fully functional without internet access.

CISO CONCERN MAP

CISO CONCERN	HOW VIGIL GUARD ADDRESSES IT
LLM observability	Unified view of all LLM traffic: browser, automation, proxy, custom apps
Policy segmentation	Per-API-key rule sets with custom thresholds, redaction, and detection config
Data sovereignty	Fully on-premise, air-gap capable. No data leaves your network
Prompt injection	6 parallel detection branches with weighted scoring, no single point of failure
PII leakage	20+ entity types (incl. Polish: PESEL, NIP, REGON), auto-sanitization before LLM
Shadow AI	Browser extension for LLM platforms, deployed via GPO / Intune / Jamf
Assistant scope compliance	Semantic Drift Detection: 3-level classification, per-level action, AES-encrypted scope per API key
SIEM / SOC	Native CEF/JSON export to Splunk, QRadar, Sentinel, Elastic over TCP/TLS
Compliance / audit	Full event logging, rule engine audit trail, configurable data retention
False positives	FP/FN memory system that learns from your team's feedback, improves over time

OBSERVABILITY AND AUDIT

- **Web Dashboard**: real-time detection metrics, event history, FP/FN reporting, configuration management
- **SIEM export**: CEF/JSON forwarding over TCP/TLS to Splunk, QRadar, Sentinel, or any syslog collector
- **Scope Drift Analytics**: dedicated dashboard pane: ON / NEAR / OFF_SCOPE distribution and per-API-key drift ranking

LICENSING

7-day trial on first startup, all features unlocked, no license key required. Subscription license activated via dashboard. See pricing and plans at vigilguard.ai/pricing.

ON-PREMISE
Docker Compose, x86_64, auto-scale

AIR-GAPPED
No phone-home, no cloud calls

BILINGUAL
Polish + English, native ML

AES ENCRYPTED
Scope, DSL, redaction lists at rest

You can't secure what you don't see.

AI is already part of your environment. Vigil Guard makes it visible, controllable, and safe.

WHAT SETS US APART: SEMANTIC DRIFT DETECTION

The first AIDR platform that doesn't just block attacks, it also keeps your assistants within their defined role. 3-level ON / NEAR / OFF_SCOPE classification, policy configured per API key, proprietary bilingual model. Attack and drift in one pane.

START YOUR 7-DAY FREE TRIAL

All features unlocked. No license key required. Deploy in minutes.

This is not a research project. This is production-grade AI security.

Learn more: https://www.vigilguard.ai/

Start a free trial today: https://www.vigilguard.ai/pricing/

Security Teams	AI / Platform Teams	Risk & Compliance
Visibility, control, auditability	Security without slowing innovation	Lower AI risk, improved audit readiness