WITNESS AI

DATASHEET

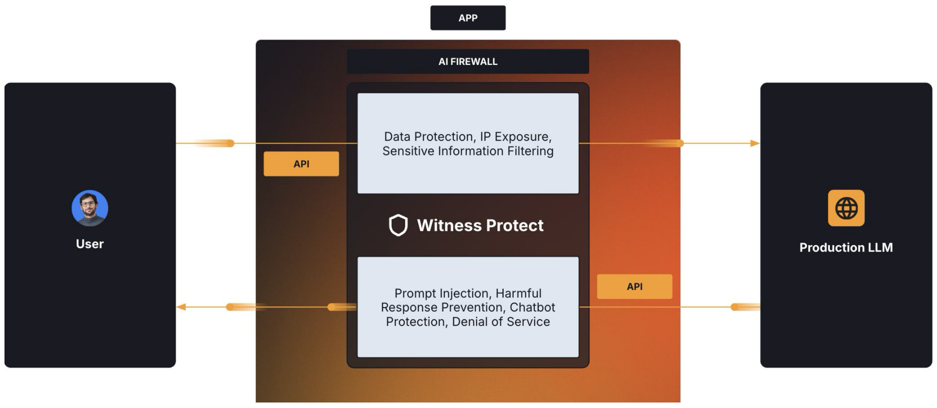
Witness Protect

The Enterprise AI Firewall for Models, Apps and Agents.

Your AI Apps and Agents Are Under Attack. Your Traditional Security Can't See It.

Every customer-facing chatbot. Every internal AI tool. Every model you've deployed. Every AI agent you have engineered. They're all exposed to threats your existing security stack wasn't built to handle. Prompt injection attacks that bypass model defenses. Chatbots manipulated into recommending competitors or offering unauthorized discounts. Sensitive data flowing to AI models through thousands of shadow applications your IT team doesn't even know exist. Rogue AI agents exceeding their scope.

Global 2000 enterprises face a security crisis that demands purpose-built AI protection. Traditional firewalls inspect packets, not prompts. DLP tools block everything or nothing because they can't understand intent. And while model providers promise security, their protection varies wildly across platforms, leaving you exposed when attacks evolve faster than their updates. Witness Protect is architected for this problem.


APP
AI FIREWALL
User
API
Data Protection, IP Exposure, Sensitive Information Filtering
Witness Protect
Prompt Injection, Harmful Response Prevention, Chatbot Protection, Denial of Service
API
Production LLM

Enterprise Scale, Proven Results

4,000+	AI applications detected
350,000+	employees secured
40+	operating across 40+ countries
Millions	of daily AI interactions monitored and secured
99.3%	true positive guardrails for model protection

Protection Across Your Entire AI Ecosystem

Advanced AI Attack Prevention

Stop AI attacks before they compromise your agents and models

Prompt Injection & Jailbreaking: Block sophisticated attacks that trick models or agents into revealing data or executing harmful instructions

Many-Shot & Role-Playing Exploits: Detect conversation manipulation attempts designed to wear down defenses

Invisible Character & Emoji Attacks: Catch obfuscated threats your provider defenses miss

AI Model Protection

Ensure every AI interaction maintains brand integrity and compliance

Brand Identity Enforcement: Ensure every AI response aligns with your brand voice and compliance requirements

Competitive Protection: Prevent AI chatbots from recommending competitors or making unauthorized commitments

Harmful Response Filtering: Block discriminatory, medical, or illegal content before it reaches users

Enterprise Data Protection

Protect sensitive information flowing through every AI touchpoint

Automatic PII Detection: Identify and protect sensitive information across various data types

Bidirectional Protection: Secure both employee prompts and model responses

Complete Audit Trail: Maintain immutable logs of every interaction for compliance proof

Global Airline

> The ability to see every AI interaction across our global workforce has transformed our security posture. WitnessAI helps us maintain compliance while enabling our teams to leverage AI for competitive advantage.
>
> **VP of Cybersecurity,** Global Top 5 Airline

Capabilities for Global 2000 Scale

Bidirectional Protection That Actually Works: Traditional tools look at traffic in one direction. Witness Protect handles AI prompts and responses by design. This is why we catch threats traditional tools miss, like attacks hidden in responses or multi-stage AI attacks.

Developer-Friendly Deployment: Deploy protection through multiple pathways: network-level integration with your existing proxy, our Witness Anywhere solution for agentless coverage, or a simple developer API for direct application integration. Get enterprise-grade protection without disrupting your architecture.

Real-Time Data Tokenization: Protect sensitive information with real-time data redaction that automatically identifies and tokenizes PII, credentials, and IP before they reach AI model or agent.

Comprehensive Threat & Toxicity Detection: Provide continuous threat detection alongside toxicity filtering to identify harmful content, inappropriate responses, and potential security risks in real-time.

Standardized Protection Across 100+ LLMs: Unlike point solutions that only work with specific models, Witness Protect provides consistent, standardized protection across over 100 types of LLMs including your custom models.

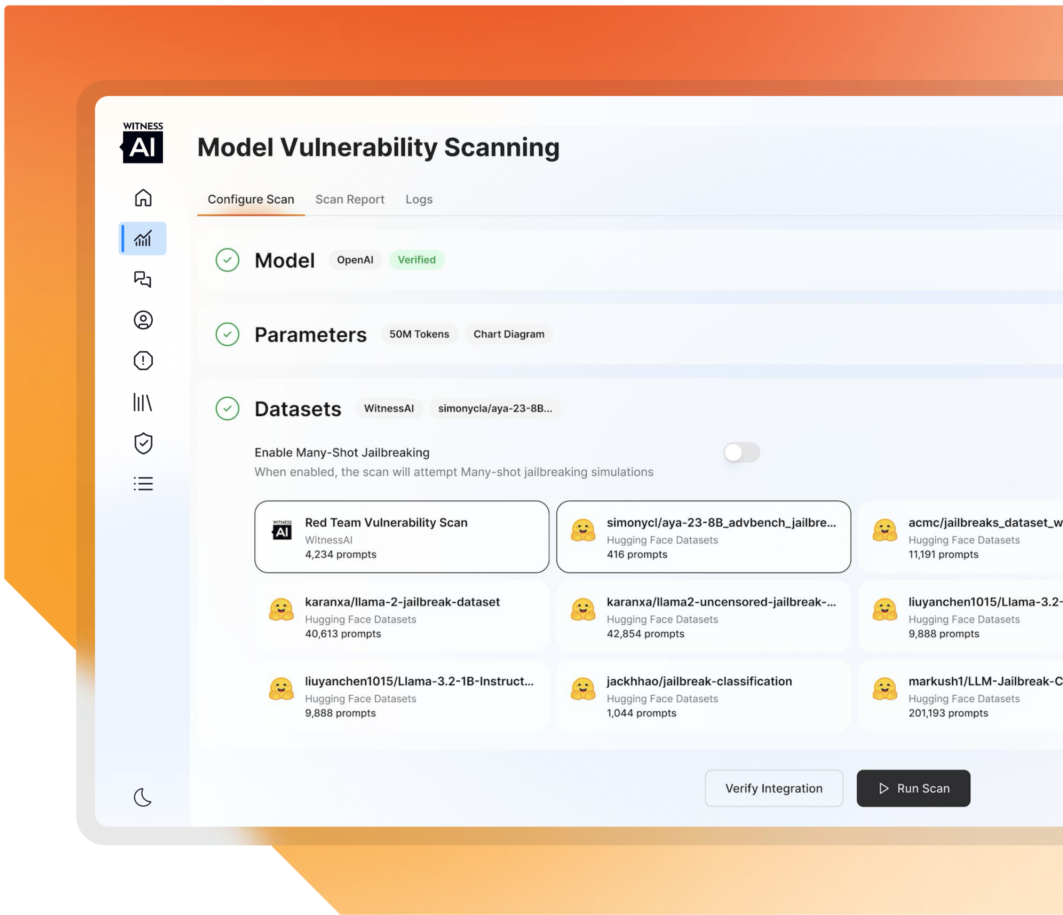
Model Identity Enforcement: Ensure your AI models maintain their assigned role and brand identity. Prepend identity instructions to every prompt and validate responses to ensure alignment, preventing manipulation attempts that could lead to off-brand responses, compliance violations, or reputational damage.

OWASP Top 10 for LLMs Compliance: Witness Protect addresses all critical vulnerabilities identified in the OWASP Top 10 for Large Language Model Applications, providing comprehensive protection against prompt injection (LLM01), sensitive information disclosure (LLM02), improper output handling (LLM05), excessive agency (LLM06), and more.

InComm Payments

We chose WitnessAI enabling compliance, data-loss prevention, and privacy teams to have total visibility and confidence in our AI security. We're reducing risk while maximizing our productivity because of WitnessAI.

CISO, InComm Payments



Model Vulnerability Scanning

Model OpenAI Verified

Parameters 50M Tokens Chart Diagram

Datasets WitnessAI simonycla/aya-23-8B...

Enable Many-Shot Jailbreaking
When enabled, the scan will attempt Many-shot jailbreaking simulations

Red Team Vulnerability Scan
WitnessAI
4,234 prompts

simonycl/aya-23-8B_advbench_jailbre...
Hugging Face Datasets
416 prompts

acmc/jailbreaks_dataset_w
Hugging Face Datasets
11,191 prompts

karanxa/llama-2-jailbreak-dataset
Hugging Face Datasets
40,613 prompts

karanxa/llama2-uncensored-jailbreak-...
Hugging Face Datasets
42,854 prompts

liuyanchen1015/Llama-3.2-
Hugging Face Datasets
9,888 prompts

liuyanchen1015/Llama-3.2-1B-Instruct...
Hugging Face Datasets
9,888 prompts

jackhhao/jailbreak-classification
Hugging Face Datasets
1,044 prompts

markush1/LLM-Jailbreak-C
Hugging Face Datasets
201,193 prompts

Witness Attack: Harden Your Models Before They're Exploited

Find Vulnerabilities Before Threat Actors Do

Witness Attack is our automated red-teaming engine that stress-tests your models before deployment. While Witness Protect secures models at runtime, Witness Attack ensures they're battle-tested from day one.

Pre-Deployment Validation That Scales

- **Multimodal Attack Simulation**: Test across text, images, and audio simultaneously with sophisticated attack chains
- **Adaptive Testing**: Reinforcement-learning attacks that evolve based on model responses to find novel vulnerabilities
- **Continuous Validation**: Run attack simulations that stress-test defenses with synthetic prompts and multi-step jailbreaks
- **Developer-Friendly**: Enable engineering teams to identify and fix vulnerabilities during development, not after deployment

Secure Your AI Adoption

Book a Demo

About Witness AI

WitnessAI enables safe and effective adoption of enterprise AI, through security and governance guardrails for public and private LLMs. The WitnessAI Secure AI Enablement Platform provides visibility of employee AI use, control of that use via AI-oriented policy, and protection of that use via data and topic security.

Learn more at witness.ai.