Mend AI

DATASHEET

RUNTIME GUARDRAILS

Stop unsafe AI interactions the moment they happen

Real-time protection for AI in production - not just before deployment, but for as long as it's live.

THE PROBLEM

A written rule isn't a guarantee

AI models don't follow instructions the way traditional software does. A system prompt that says "never do X" is a rule, not a lock. Attackers are using conversational, social-engineering-style prompts to convince a model to override its own instructions, the same way someone might talk a person into breaking a rule they know they shouldn't.

That means sealing the instructions isn't the same as sealing the risk. Traditional security tools, built to secure code before it ships, were never designed to catch a model being talked out of its own rules live, in the conversation itself.

Rules aren't locks

A prompt can look airtight and still not hold under pressure.

Runtime isn't deterministic

The same instructions can produce different behavior in the moment.

Traditional tools miss it

Pre-ship security was never built to catch a live conversation going wrong.

WHAT RUNTIME GUARDRAILS DO

Inspect, detect, and enforce - in real time

Mend AI's Runtime Guardrails act as a deterministic backstop for a system that can't fully guarantee its own behavior. Running locally inside your application, guardrails inspect every input before it reaches your model and every output before it reaches your user.

01

Receive

A prompt reaches your AI application.

02

Inspect input

Checked for prompt injection, jailbreak attempts, and exposed secrets.

03

Reach the model

Cleared prompts reach the LLM, which generates a response.

04

Inspect output

Checked for harmful content, sensitive data, and unsafe links before reaching the user.

05

Deliver

Cleared outputs are sent back to the user.

CORE CAPABILITIES

What Runtime Guardrails cover

- **Prompt Injection & Jailbreak Detection**
- **Sensitive Data (PII) Protection**
- **Harmful Content Filtering**
- **Secret Keys Guardrail**
- **URL Filtering**

WHY RUNTIME GUARDRAILS ARE DIFFERENT

Flexible, configurable, built to fit your stack

- **Deploy your way.** A light weight Python SDK, in Online mode for centralized configuration and monitoring, or fully isolated Offline mode with no platform connection at all - plus a standalone Docker API Server for non-Python environments.
- **Configure your response.** Alert your team, block the interaction outright, or automatically obfuscate sensitive data — per guardrail, so you can start cautious and tighten enforcement as confidence grows.
- **Integrate where you already work.** Native support for OpenAI Agent SDK, OpenAI-compatible APIs, LangChain, Azure OpenAI, and Langflow.
- **See everything, in real time.** A live dashboard, events log, and per-guardrail configuration, all in one place, when running in Online mode.

WHY RUNTIME GUARDRAILS ARE DIFFERENT

Flexible, configurable, built to fit your stack

- **Stop prompt injection**
  Catch attempts to override your AI's instructions before they cause harm.
- **Protect sensitive data**
  Catch PII in inputs and outputs, automatically.
- **Filter harmful content**
  Block unsafe or policy-violating responses before users see them.
- **Secure air-gapped environments**
  Run guardrails fully offline, with no data sent to Mend.
- **Extend to any stack**
  Deploy via Docker for non-Python applications.
- **Audit and monitor**
  Track every flagged interaction from a live dashboard.

Mend.io is built for every risk, across AI and AppSec. By securing the code layer and the AI layer—and the interactions between them, where modern application risk now lives—**Mend.io** extends proven AppSec workflows to the models, prompts, and agents inside today's applications, delivering continuous protection across the entire AI application lifecycle.

Visit our website