Mend AI

Secure your AI's core instructions

Harden system prompts with the industry's first AIWE standard

The Challenge

As organizations rapidly integrate AI, a new class of "behind-the-scenes" vulnerabilities is emerging within system prompts. These instructions guide AI behavior but remain hidden from traditional security tools, leaving a dangerous gap in the security stack.

- **Invisibility**: Legacy tools cannot detect or remediate flaws within hidden governing logic.
- **Manipulation**: Without visibility, prompt injections can override instructions to leak data or trigger unauthorized actions.
- **Lack of Metrics**: Security teams cannot prioritize risks or justify resource allocation without standardized ways to quantify the threat.

This lack of monitoring leaves the core rules governing AI behavior exposed to adversarial manipulation.

The Solution

By providing complete visibility into the logic governing your AI, Mend AI allows you to detect and remediate vulnerabilities at scale before they can be exploited.

Standardize defense with AIWE risk scoring

Mend.io leverages a proprietary AI Weakness Enumeration (AIWE) standard to quantify risk.

- **Proprietary Scoring:** A clear 1–100 severity scale for objective prioritization.
- **Proven Framework:** Modeled on the industry-approved CWSS approach for custom code.
- **Broad Coverage:** Initial support for the first 10 defined AI weaknesses out of the box.

Mend AI system prompt hardening…

Detects hidden system prompts

Gain visibility into your AI's core instructions to proactively monitor and control "behind-the-scenes" behavior.

Fortifies system prompt logic

Automatically patch vulnerabilities and gaps to block prompt injections and data leakage, ensuring resistance to adversarial attacks.

Standardizes risk with AIWE

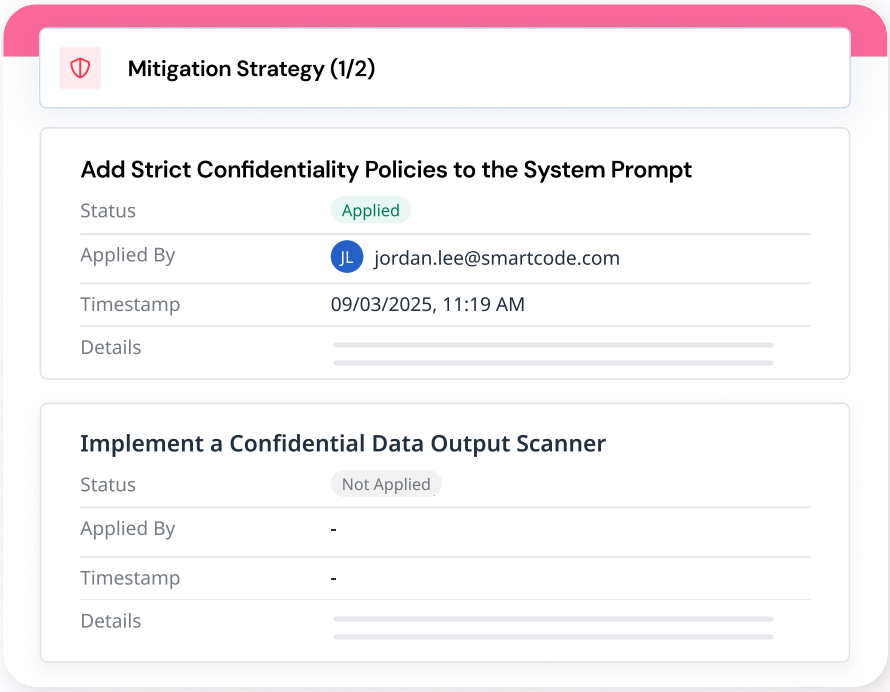
Translate weaknesses into a clear 1–100 score, allowing teams to instantly prioritize and remediate the most critical security flaws.

Identifies attack vectors instantly

Automatically categorize prompts as conversational to provide the immediate context needed to triage vulnerabilities efficiently.

Comprehensive protection for the AI era

From hardening core system prompt logic to identifying shadow AI and conducting advanced red-teaming, Mend AI provides the tools needed to detect, prioritize, and remediate vulnerabilities at scale.



Why Mend AI?

System Prompt Hardening: Improve AI reliability by refining internal prompts to ensure secure, consistent, and aligned responses.

AIWE Standardization: Leverage the first formal standard for quantifying the severity of AI system prompt weakness.

AI Supply Chain Management: Maintain real-time inventory of all models and frameworks, including shadow AI, while leveraging risk insights to mitigate licensing, vulnerability, and malicious package risks.

AI Red-Teaming: Utilize prebuilt and customizable tests to uncover behavioral risks like data exfiltration and hallucinations.

Proactive Governance: Enforce policies and workflows to manage AI security and compliance risk throughout the SDLC.

AI Agent Configuration Scanning (Coming Soon): Bring visibility and CI-friendly enforcement to the new AI control plan by treating “agents as code”.

AI Runtime Protection (In-Development): Defend against unpredictable behavioral threats by applying real-time safety filters and in-app guardrails for live AI interactions.

Mend.io helps organizations fix less and reduce application risk faster. By prioritizing real risk across modern and traditional applications on a single platform, Mend.io delivers measurable security impact with less effort.