

RUNTIME GUARDRAILS

Stop unsafe AI interactions the moment they happen

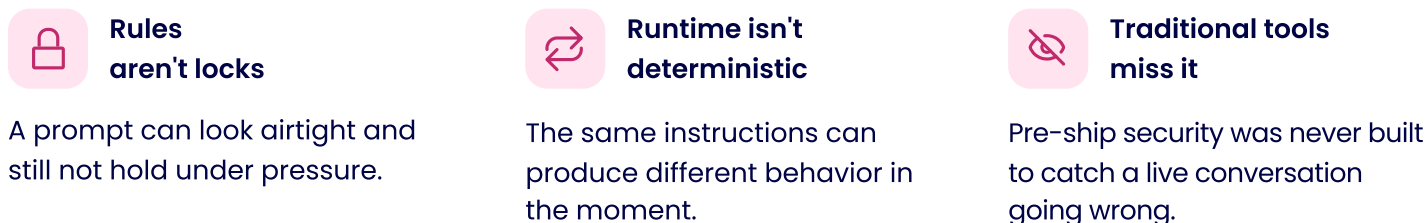
Real-time protection for AI in production – not just before deployment, but for as long as it's live.

— 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.

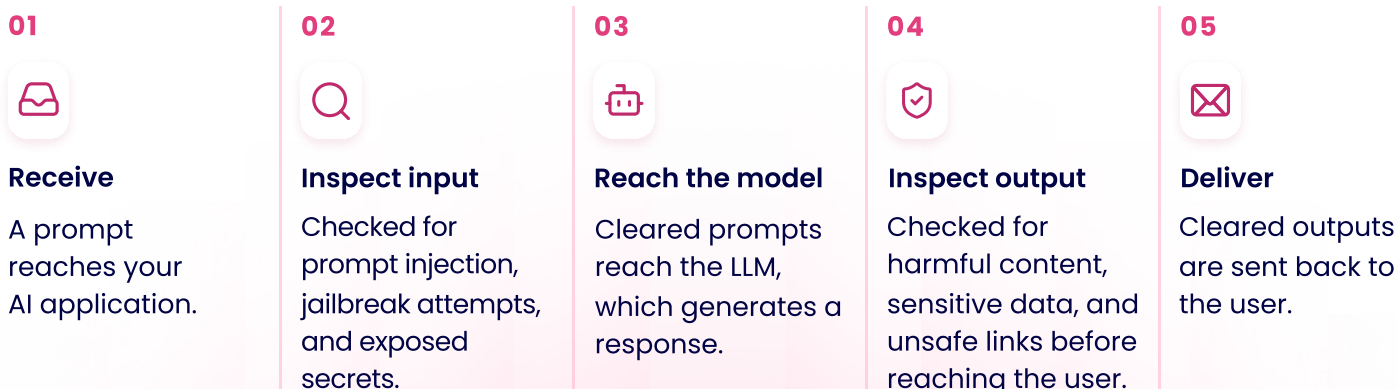
That means sealing the instructions isn't the same as sealing the risk. Traditional security tools, built to secure code before it ships, were never designed to catch a model being talked out of its own rules live, in the conversation itself.



— 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.



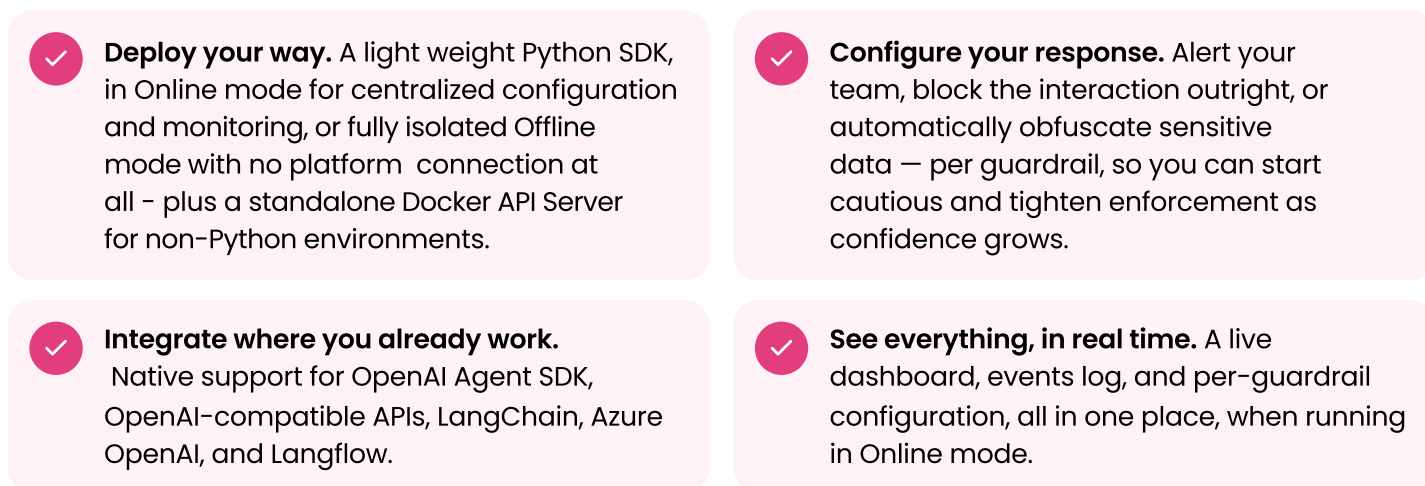
CORE CAPABILITIES

What Runtime Guardrails cover



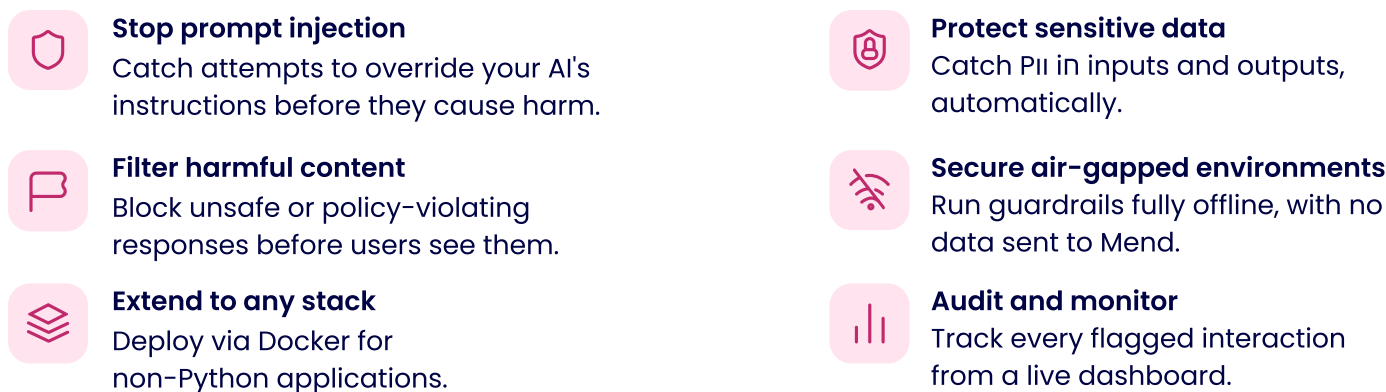
WHY RUNTIME GUARDRAILS ARE DIFFERENT

Flexible, configurable, built to fit your stack



WHY RUNTIME GUARDRAILS ARE DIFFERENT

Flexible, configurable, built to fit your stack



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.