Overview

Policy & risk detection

Every outbound AI request is scanned **before** it leaves, by layered detectors. Findings are recorded on the request's Activity row; depending on your settings they warn, report silently, or block.

Format detectors

High-precision pattern detectors: cloud keys (AWS, GitHub, Slack, Google, OpenAI, Anthropic), private keys, JWTs — plus a context detector for credential-looking values after words like "password:" that have no recognizable format.

Personal data (PII)

Emails, international phone numbers, IBANs (checksum-validated) and payment cards (Luhn-validated). Default action is **report** — PII in prompts is often legitimate; raise it to warn or block per your policy.

AI risk classifier

A local classifier plus an EU-hosted LLM tier evaluate prompt-injection and related risks with a per-tenant **sensitivity dial** (low / medium / high). A blocking verdict flags the conversation: the next request in that conversation is rejected pre-flight, and the block is visible (and clearable) in the console.

Dependency safety

Package references in prompts and AI responses are checked against registries and the OSV vulnerability database — nonexistent (slopsquat) packages and critical CVEs raise findings before `npm install` happens.

Actions per detector

Each detector has a per-tenant action: **allow** (off), **report** (silent record), **warn**, or **block**. Configure them on the Policy page; changes take effect immediately, no client changes.

Revision #1
Created 2026-08-01 13:06:24 UTC by Sentilai Docs
Updated 2026-08-01 13:06:24 UTC by Sentilai Docs