Which threats we detect

The signals that can appear on an event, what each one means, and where you control it.

In the request content

Signal	What it finds
Secret scanning	AWS keys, GitHub tokens, private keys, Anthropic/OpenAI keys, Slack tokens, Google API keys, JWTs
Personal data (PII)	IBANs, payment cards, phone numbers, email addresses — each checksum-validated, so a random 16-digit number is not reported as a card
Credentials in context	phrasing that suggests a credential is being discussed even when no key matches
Prompt injection	our classifier's verdict, with its score
Dependency risk	packages the AI suggested that do not exist, are known-vulnerable, are suspiciously new, or match a known slopsquat
MCP rug-pull	an MCP tool that changed its definition after you approved it
Lethal trifecta	one MCP call combining private data, untrusted content and a way out

All of these are set on **Policy**, each to `allow`, `report`, `warn` or `block`.

On the device

The Endpoint Suite scans for ungoverned AI agents and reports what it finds: whether one is installed, whether it starts automatically, whether its gateway is reachable from the network, whether it requires authentication, whether it can run commands unsandboxed, whether anyone who can message it can drive it — and whether all three legs of the lethal trifecta are present.

We **detect** ungoverned agents; we do not govern their traffic. They talk to their providers directly and we are not in that path. What the `block` policy does is take away the developer's *governed* tools until the agent is gone.

What "blocked" tells you

When a request is blocked, the event names the cause:

- `classifier` — the prompt-injection classifier refused it
- `risk` — a content detector set to `block` fired
- `mcp_policy` — an MCP server or tool rule refused the call

Retention, and what it means for your SIEM

Events are kept for your organization's retention period and then deleted. The stream is gap-free **within that window**: a collector that catches up after an outage gets everything it missed, as long as the outage was shorter than your retention setting.

If your SIEM stops receiving events for longer than that, the gap is permanent. This is why the delivery status on the SIEM Export screen is worth an occasional look — a stalled audit pipeline is the failure that announces itself least.

Revision #2
Created 2026-08-04 08:39:24 UTC by Sentilai Docs
Updated 2026-09-29 14:39:21 UTC by Sentilai Docs