Risk classifier sensitivity

The risk classifier inspects outbound prompts for things that shouldn't leave — secrets, personal data, prompt-injection attempts, and exfiltration-shaped tool sequences.

Policy ? AI risk classifier sensitivity controls how aggressively it acts:

- **Low** — suppresses low-confidence signals, for a quieter feed.
- **Medium (recommended)** — records medium- and high-severity findings as warnings.
- **High** — the strictest setting.

Findings appear on the **Activity** screen either way; sensitivity changes what is acted on, not what is recorded.

Secrets are blocked regardless

A detected credential — an AWS key, a private key, a provider token — is blocked in the request path independently of this setting. Sensitivity governs the softer signals.

Reading the result

On **Activity**, a request with findings shows chips in the **Signals** column (`aws_access_key_id`, `email_address ×3`, and so on) and its outcome is `Blocked` if policy stopped it. Click through to the conversation when prompt capture is on.

Revision #8
Created 2026-08-01 13:06:38 UTC by Sentilai Docs
Updated 2026-09-29 14:39:31 UTC by Sentilai Docs