The prompt-injection classifier

A model, rather than a pattern, judging whether a request is being manipulated.

How it runs

The classifier runs **in parallel with the upstream call**, so it does not add latency to your developers' requests. In Activity you can see how long it took — the millisecond badge in the Signals column — and the score it produced.

The sensitivity dial

Policy ? AI risk classifier sensitivity, three positions:

- **Low** — only the clearest cases.
- **Medium** (recommended) — the default balance.
- **High** — aggressive. At this setting a high-severity verdict can block the next turn of the conversation.

The dial governs this classifier. It does **not** govern secret scanning: detected secrets are handled by that detector's own action regardless of where the dial sits.

Block-next-turn

When the classifier returns a high-severity verdict under High sensitivity, the **conversation** is stopped rather than the individual request — the next turn is refused. This is the right granularity for injection, where the problem is the conversation's context rather than one message.

Blocked conversations are listed at the bottom of the Policy screen with the developer, tool, reason and severity, each with an **Unblock** button. They also clear by themselves when they age out of your retention window.

The paid classifier tier

A second, larger classifier hosted in the EU covers a wider risk taxonomy. It is a disclosed sub-processor — if your DPA review needs the details, ask us and we will send the current list.

Revision #6
Created 2026-08-02 10:20:02 UTC by Sentilai Docs
Updated 2026-08-04 08:39:46 UTC by Sentilai Docs