Blocked conversations

When the classifier stops a conversation, it appears at the bottom of the **Policy** screen — and this is where you let it continue.

What you see

Each entry names the developer, the tool, the reason the classifier gave, the severity, and when it happened.

Unblocking

Unblock lets that conversation continue from the next turn. Use it when you have looked at the reason and concluded it was a false positive — for example a developer legitimately working on prompt-injection defences, whose prompts look exactly like the thing being detected.

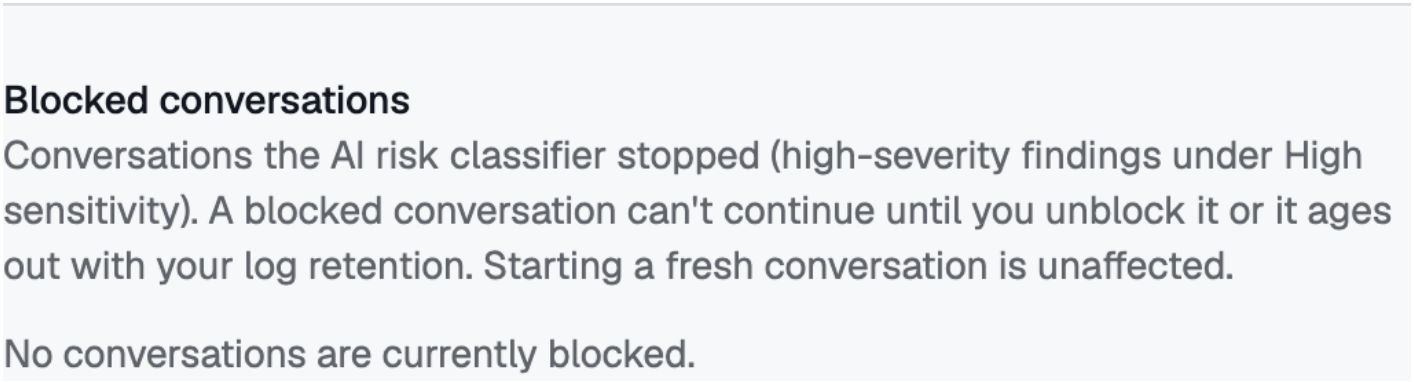
They also expire

Blocked conversations clear by themselves when they age out of your retention window. You do not have to maintain this list.

If it is happening a lot

Frequent blocks usually mean the sensitivity dial is one position too high for your team's work. Move it from High to Medium and see whether the genuine findings survive — they usually do, because the clearest cases are caught at every setting.

An empty list is the normal state and shows "No conversations are currently blocked."



Blocked conversations — sessions the AI risk classifier stopped, and where an admin unblocks one.

Revision #12
Created 2026-08-02 10:20:04 UTC by Sentilai Docs
Updated 2026-09-29 14:39:42 UTC by Sentilai Docs