Local MCP decision cache

Policy ? Local MCP decision cache (seconds), from 0 to 3600.

What it controls

When an AI tool calls a local MCP server, the Sentilai shim on the developer's machine asks the Gateway whether to allow it. This setting is how long the shim may reuse that answer for the same server and tool before asking again.

The trade-off

Higher means fewer round trips and a snappier agent — MCP tool calls happen in tight loops, and a network hop on each one is noticeable.

Lower means a policy change reaches developers' machines faster. At 0, every call asks the Gateway.

A reasonable setting

A few minutes suits most teams. It keeps agents responsive while ensuring that a rule you change at the start of a meeting is in force by the end of it.

If you are actively tuning MCP rules and want to see the effect immediately, drop it to 0 for the session and put it back afterwards.

What it does not affect

Only local stdio MCP servers behind the shim. Routed remote endpoints and the chat-request side-channel are evaluated by the Gateway on every request.

Revision #7
Created 2026-08-02 10:20:04 UTC by Sentilai Docs
Updated 2026-09-29 14:39:41 UTC by Sentilai Docs