



Oops, AI Build It Again

Let AI build. Keep your hands on the wheel.

September 28, 2026

Sebastian Scherl

Cloudflare

Jonas Wagner

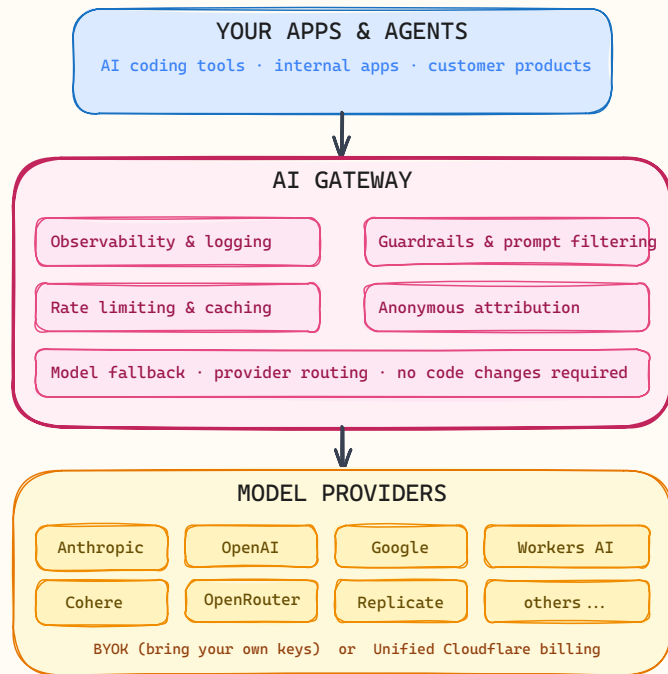
Cloudflare

AI Gateway

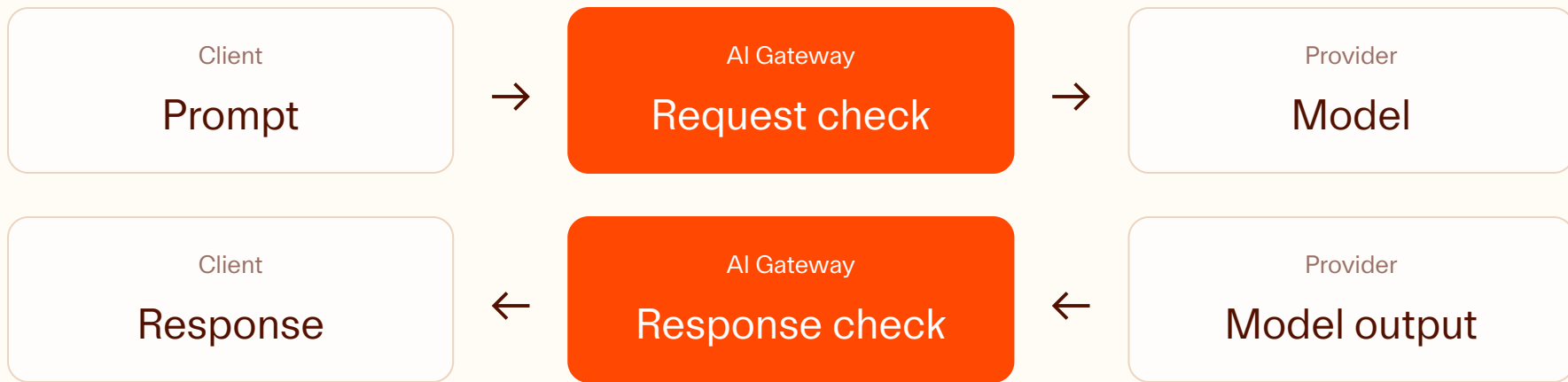


Observe, control, and secure every AI request — across every provider

- **Unified observability** — cost, latency, and usage across all providers in one dashboard
- **Security guardrails** — prompt injection detection and sensitive data filtering, per route
- **Cost control** — rate limiting, response caching, and model fallback
- **Anonymous attribution** — per-user analytics without exposing identities to providers



AI Gateway Guardrails



Identify unsafe or inappropriate content in prompts and responses based on selected categories, then flag or block as needed.



NOTE: Guardrails uses Workers AI, and usage will be reflected in the Workers AI dashboard [📄](#)

Prompt hazard categories

Privacy Hate Self-Harm Violent Crimes Sex Crimes
Non-Violent Crimes Child Exploitation

Response hazard categories

Violent Crimes Non-Violent Crimes Sex Crimes Child Exploitation
Privacy Hate Self-Harm

[Change](#)

AI Gateway DLP



Data Loss Prevention (DLP)



Scan content for sensitive data and prevent data leakage by selecting DLP profiles.

Policies

Add Policy

Policy ID: llm 



DLP Profiles

AI Prompt: AI Security, AI Prompt: Customer, AI Prompt: Financial Information, AI Prompt: PII, AI Prompt: Technical



The AI requests/responses will be checked against each of the selected profiles. [Manage profiles](#)

Action

Flag



If any of the selected profiles match a request, this is the action that will be taken

Check

Response, Request



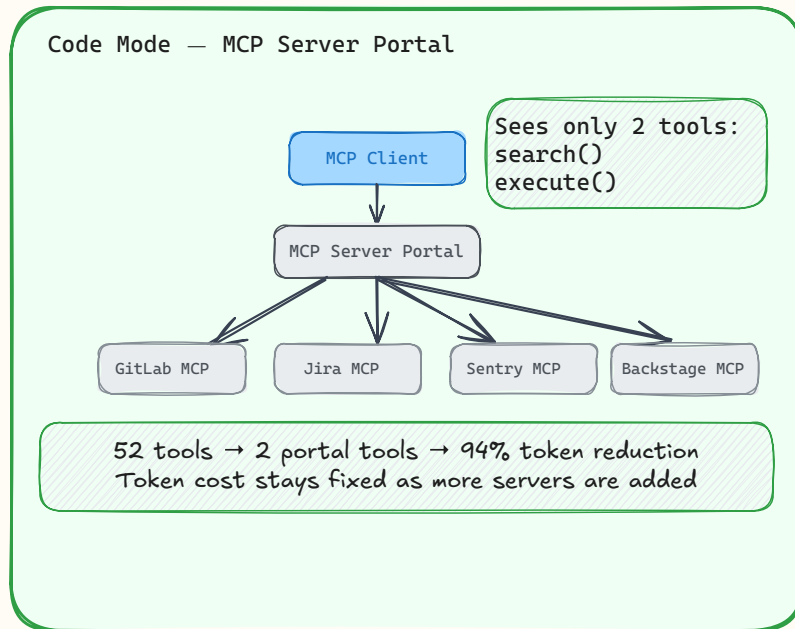
Choose whether to check the request, response, or both

MCP on Cloudflare

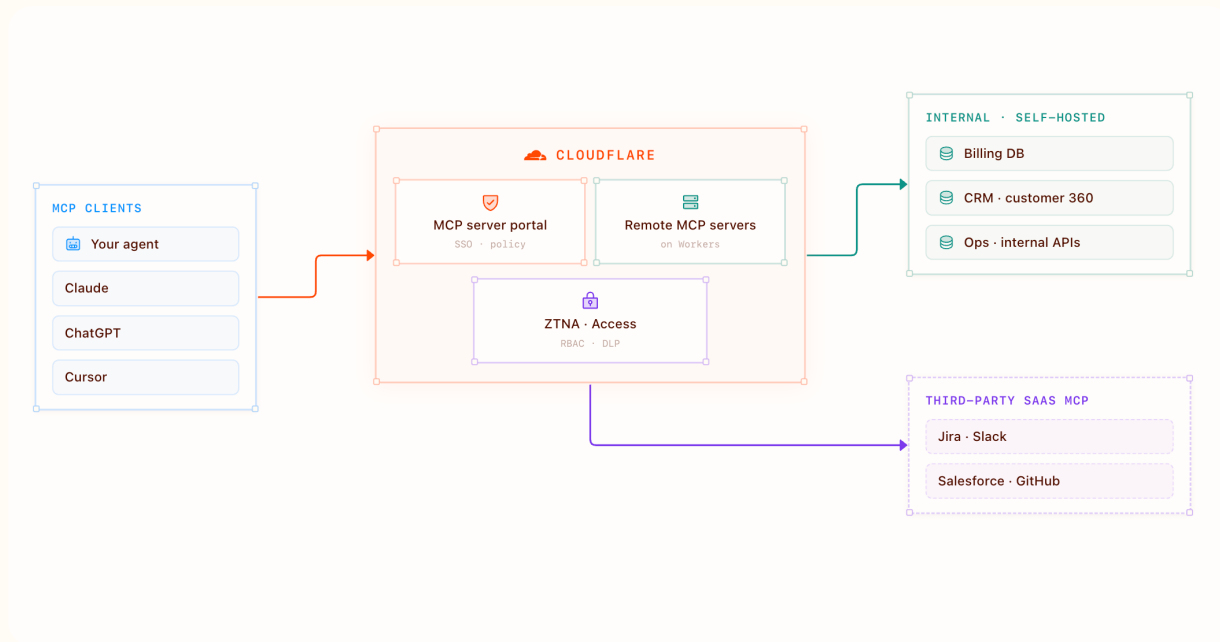


Move from unmanaged local connections to governed, remote deployments

- **Eliminate ungoverned local connections** — get full visibility into what tools are being called
- **Identity-aware AI access** — Cloudflare Access acts as the OAuth provider for every MCP server. Agents inherit the identity and permissions.
- **Least privilege access** — scope agent access to specific MCP servers using granular policies with one portal
- **Data loss prevention** — apply DLP across agent activity before sensitive data leaves your environment



MCP Server Portal



Access at the door

Tool selection

Tool-call logs

Let's build together

