Secure, control and observe your LLM traffic and AI agents

An AI gateway holds the provider keys, picks the provider, limits tokens, applies prompt rules and records every call. G8 AI Gateway does this in your own infrastructure, on the same Routes, roles and Environments as your APIs, and serves the APIs you already run to AI agents as MCP tools. Included in every edition, Community included; Enterprise AI adds the G8 Copilot, not the AI Gateway.

Capabilities

Providers and fallback
OpenAI, Anthropic, Azure OpenAI, Amazon Bedrock, Google Gemini, DeepSeek and any OpenAI-compatible endpoint, including models you host. Each request goes to the provider with the highest priority and weight and falls back to the next when one fails.

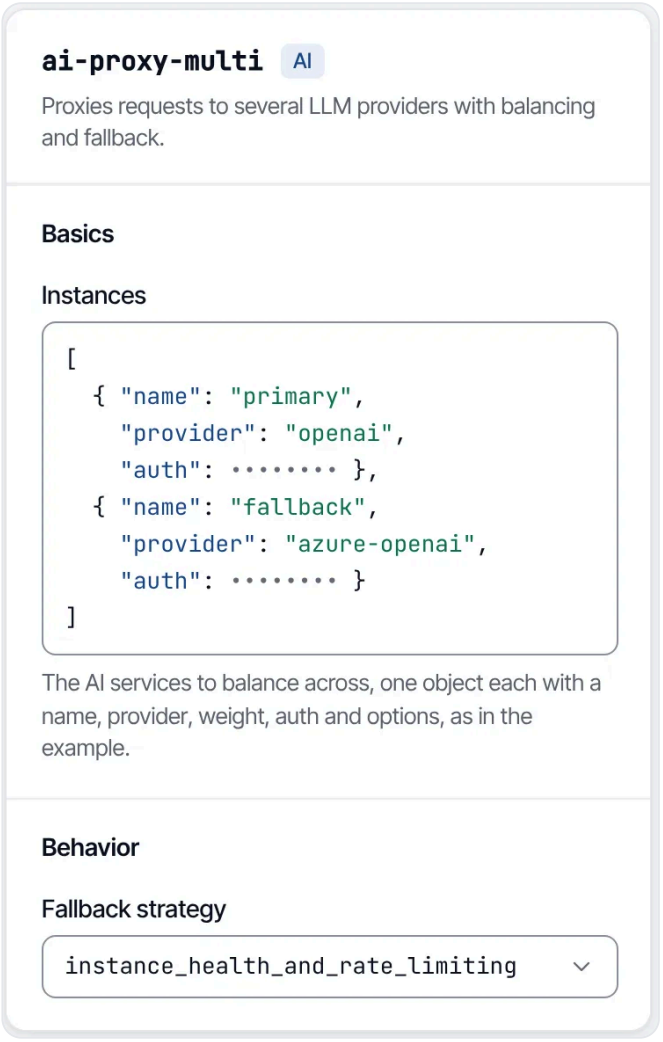
Token budgets
Prompt, completion or total tokens per window, per Consumer or per request variable, with a separate budget per provider instance.

Prompt rules
Allow or deny prompts by pattern, add the system prompt you choose, and fix the shape of a request with a template.

Caching
A repeated prompt is answered from the cache, matched exactly or by meaning, without a second call to the model.

Rewrite and retrieval
A model can rewrite the request before it is sent; a search of your documents can be added to the prompt.

Moderation
Content moderation services and Lakera Guard screen prompts and responses; block what they flag, or let it through and log it. Each of these connects to an external service and is marked so in the console.



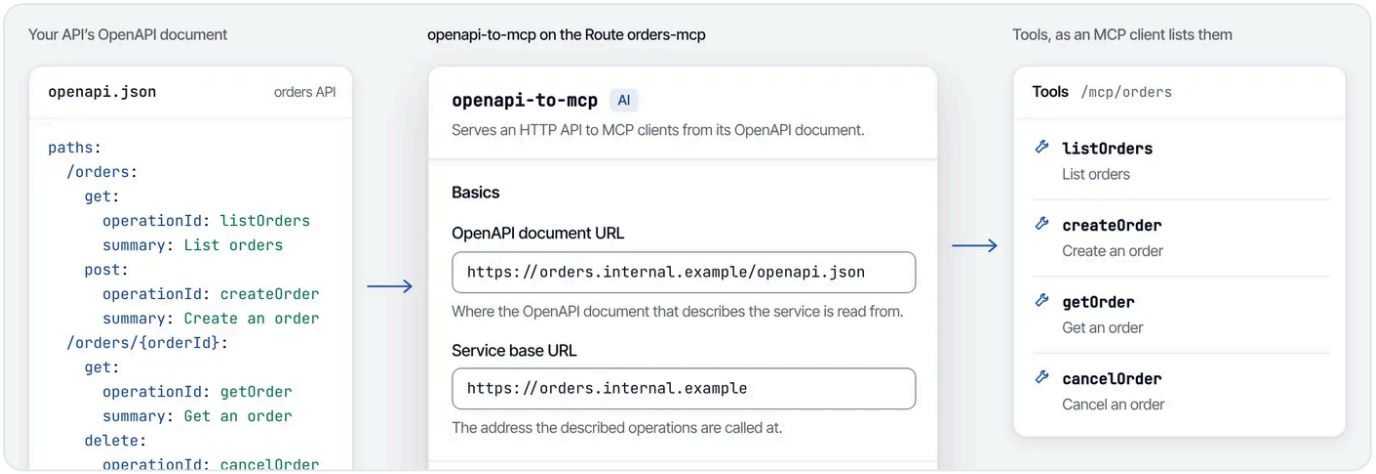
A Route to language models: the primary and the fallback provider, with the keys stored at the gateway and masked in the console.

THE AI PLUGINS

`ai-proxy` `ai-proxy-multi` `ai-rate-limiting` `ai-prompt-guard` `ai-prompt-decorator` `ai-prompt-template` `ai-cache` `ai-request-rewrite` `ai-rag` `ai-aws-content-moderation` `ai-aliyun-content-moderation` `ai-lakera-guard` `openapi-to-mcp` `mcp-bridge`

OpenAPI to MCP: serve your existing APIs to AI agents

Add the openapi-to-mcp Plugin to a Route and point it at an API's OpenAPI document. MCP clients, such as AI agents, list the API's operations as tools and call them; the G8 engine translates each tool call into the HTTP request the API already understands. The Route's authentication, rate limits and request records apply to every tool call, so an agent is a Consumer like any other.



From left to right: an OpenAPI document, the Plugin on a Route, and the tool list an MCP client sees.

mcp-bridge
Puts a Route in front of an MCP server you already have, so it gets the same authentication, limits and records.

Transport
Streamable HTTP or SSE; tool inputs flattened or kept as the document describes them.

Keys and data stay with you

Provider keys at the gateway
Applications never hold a provider key; keys are stored in the gateway configuration and masked in the console, in request records and in the audit log.

Records in your stores
Every call is a request record in your Elasticsearch or OpenSearch and a count in your Prometheus; the same Analytics, Traffic Explorer and Alerts as your APIs.

Governance
The same Environments and roles mapped to Active Directory groups. Every change is written as an audit record before it is applied; a refused change is a record too.

PROVIDERS

`OpenAI` `Anthropic` `Azure OpenAI` `Amazon Bedrock` `Google Gemini` `DeepSeek` `OpenAI-compatible endpoints` `Models you host`

NEXT STEP

Bring one application's LLM traffic and one OpenAPI document; we show both on the same Route. Request a demo at g8.dev/request-a-demo or write to hello@g8.dev. Read next: G8 API Security Datasheet (2 pages), at g8.dev/downloads.

g8.dev/request-a-demo · hello@g8.dev