

AI gateway guide

DOCUMENT	VERSION	STATUS
SB-GD-121	1.0, October 2026	Public. Checked against the product, October 2026

Who this is for

Developers and workspace admins who want their own apps to call AI models through SynapBridge, using an OpenAI or Anthropic client.

Before you start

- Your workspace needs the Team plan or above (Team, Business or Enterprise).
- You need the owner or admin role to create or change a gateway key.
- Connect at least one model provider under AI providers in Admin. The gateway uses your own provider connections. A call to a provider with no connection is refused.
- Call the gateway from a server. The gateway sends no CORS headers, so a browser cannot call it, and a gateway key must never be placed in browser code.

What the gateway is

The AI gateway is an endpoint your own apps call in place of a provider. You keep using the OpenAI or Anthropic client libraries and change two things: the base URL and the key. Each call runs through your workspace policies before it reaches the provider, and the call is recorded in the request log and in spend tracking against the key that made it.

The gateway offers three routes:

- POST `/api/gateway/v1/chat/completions`, in the OpenAI format.
- POST `/api/gateway/v1/messages`, in the Anthropic format.
- GET `/api/gateway/v1/models`, which lists the models the calling key may use. A provider with no model restriction appears as `provider/*`, meaning any model that provider serves.

Note There is no model routing on the gateway. You name the model you want on each call.

Create a gateway key

- 1 Sign in as an owner or admin.
- 2 Open Admin center, then Applications, then AI gateway.
- 3 Choose Create gateway key.
- 4 Enter a name (up to 80 characters) and, if you want, the limits described below.
- 5 Copy the key from the confirmation panel. It is shown once. If you lose it, rotate the key.

A key made on this page carries only the gateway scope. You can also allow an existing key to call the gateway from the same page, using the Allow gateway calls action on its row. Stop gateway calls removes gateway access and leaves the key in place.

Note Below the Team plan, the page opens so you can see it, but creating a key is not available and the server refuses it with `plan_not_eligible`.

Limits on a key

Each gateway key has three optional limits. You set them when you create the key and can change them later with Edit limits.

- Monthly limit in USD. A number above 0 and up to 1,000,000, rounded to cents. It is counted per calendar month in UTC. Leave it blank for no limit on that key.
- Allowed models. Up to 50 entries, one per line, written as `provider/model`, or `provider/*` for any model from one provider. Leave it blank to allow any model from your connected providers.
- Requests per minute. A whole number from 1 to 6,000. The default is 120.

A call to a model that is not on the key's list returns 403 `model_not_allowed`.

Base URL and model names

The base URL is your workspace address followed by the gateway path. The AI gateway page shows the exact value for your workspace and has a Copy button. In the examples below, `YOUR_WORKSPACE_HOST` stands for that address.

- OpenAI clients use `https://YOUR_WORKSPACE_HOST/api/gateway/v1`
- Anthropic clients use `https://YOUR_WORKSPACE_HOST/api/gateway`, because the Anthropic client adds `/v1/messages` itself.

Name a model as `provider/model`, for example `openai/gpt-4o-mini`. A bare model id also works when its provider is clear from its start, such as `claude-`, `gpt-`, `chatgpt-`, `o` followed by a digit (for example

o3), gemini-, grok-, deepseek- or sonar. On the Anthropic route, any other bare id is treated as an Anthropic model.

Supported provider names are anthropic, openai, azure_openai, bedrock, google, xai, databricks, nvidia, perplexity, together, fireworks, groq, cerebras, openrouter, deepseek, qwen and custom_openai_compatible. A model can be called only if your workspace has that provider connected.

Send a call

Send the key as a Bearer token, or in the x-api-key header. OpenAI clients send the first and Anthropic clients send the second.

These are the examples the AI gateway page gives. The key is read from an environment variable, so it never sits in source code.

```
curl https://YOUR_WORKSPACE_HOST/api/gateway/v1/chat/completions \
  -H "Authorization: Bearer $SYNAPBRIDGE_API_KEY" \
  -H "Content-Type: application/json" \
  -d '{"model": "openai/gpt-4o-mini", "messages": [{"role": "user", "content": "Hello"}]}'

import OpenAI from 'openai';

const client = new OpenAI({
  baseURL: 'https://YOUR_WORKSPACE_HOST/api/gateway/v1',
  apiKey: process.env.SYNAPBRIDGE_API_KEY,
});

const reply = await client.chat.completions.create({
  model: 'openai/gpt-4o-mini',
  messages: [{ role: 'user', content: 'Hello' }],
});

import Anthropic from '@anthropic-ai/sdk';

const client = new Anthropic({
  baseURL: 'https://YOUR_WORKSPACE_HOST/api/gateway',
  apiKey: process.env.SYNAPBRIDGE_API_KEY,
});

const reply = await client.messages.create({
  model: 'claude-sonnet-5',
  max_tokens: 1024,
  messages: [{ role: 'user', content: 'Hello' }],
});
```

Set SYNAPBRIDGE_API_KEY to your gateway key, for example YOUR_GATEWAY_KEY.

What the gateway accepts

- Text content only. Image and other non-text parts are refused with a 400.
- One answer per call. n above 1 is refused.
- max_tokens (or max_completion_tokens) from 1 to 200,000. The Anthropic route requires max_tokens.
- On the OpenAI route: temperature from 0 to 2, up to 4 stop sequences, and reasoning_effort of minimal, low, medium or high.
- On the Anthropic route: temperature from 0 to 1, stop_sequences, and extended thinking with a budget_tokens of at least 1,024 and below max_tokens.
- Streaming with stream set to true. On the OpenAI route, usage is sent in the stream only when stream_options.include_usage is true.
- Tool use for Anthropic models only, up to 64 tools, with tool choice auto or none. Other tool-choice values and structured response formats are refused.
- Up to 1,000 messages and a request body of up to 2 MB.
- An optional Idempotency-Key header, 1 to 200 characters from letters, numbers and . _ : - .

Policies that run on a call

Before a call reaches the provider, the gateway checks the key, its scope, its rate limit, the workspace status and plan, the model and the budgets. The call is then sent through the same controls SynapBridge applies to its own AI features: data loss prevention (pattern-based), data residency and sovereignty rules, model entitlements, content policy, spend caps, output scanning and the kill switch. A call that one of these refuses returns an error with a specific code, listed below.

Budgets and spend limits

- Budgets set to block on breach stop calls once the workspace, or this key, has spent the budget. The call returns 402 budget_exceeded.
- A key's monthly limit is checked before each call by reserving a worst-case cost estimate, so calls that run at the same time count against the limit together. When the key has reached its limit, the call returns 402 key_budget_exceeded. The limit is a check, not an exact ceiling: a long answer with no max_tokens can run past its estimate, so the total can exceed the limit by the overrun of calls that were already in flight. The next call is refused.
- Workspace spend caps return 402 spend_cap_exceeded.
- If budgets cannot be checked, the call is refused and not sent. The gateway returns 503 budget_check_unavailable with a Retry-After of 5 seconds. If the workspace status cannot be read, it returns 503 governance_unavailable.

Note Budget stops use 402, not 429, because many client libraries retry a 429 on their own and a spent budget does not recover by retrying.

Responses and errors

A successful call returns the normal OpenAI or Anthropic response body, with the model name you sent. These headers are sent:

- x-request-id on chat completions and messages calls, including errors.
- x-synapbridge-invocation-id and x-synapbridge-cost-usd (the cost in USD, six decimals) on non-streaming responses. They are not sent on streamed responses.
- cache-control: private, no-store.
- retry-after on 429 and 503 responses that carry a wait time.

Errors use the format of the route you called. On the OpenAI route the body is an error object with message, type, param and code. On the Anthropic route the body has type "error" and an error object with type and message, and no code field, so use the status code there.

- 400: invalid_json, invalid_request, unsupported_parameter, unsupported_content, model_not_found, tools_not_supported, provider_not_connected, provider_not_configured.
- 401: missing_api_key, invalid_api_key (also returned for a revoked key).
- 402: budget_exceeded, key_budget_exceeded, spend_cap_exceeded, workspace_suspended.
- 403: insufficient_scope (the key lacks the gateway scope), plan_not_eligible, model_not_allowed, model_not_entitled, blocked_by_dlp, blocked_by_content_policy, blocked_by_sovereignty_policy, blocked_by_residency_policy, blocked_by_data_class_policy, blocked_by_auto_block_rule, memory_sharing_not_permitted, approval_required.
- 413: request_too_large.
- 423: kill_switch_active. An operator paused the resource. Retry after it is resumed.
- 429: rate_limited, with a Retry-After header in seconds.
- 451: exfil_output_blocked.
- 202: approval_required_pending. The call was held for a human decision. Do not retry it, because a retry creates a second pending request.
- 502: provider_error. 504: provider_timeout.
- 503: gateway_not_configured, gateway_unavailable, governance_unavailable, budget_check_unavailable.
- 500: internal_error. Nothing was sent to the model.

If a streamed call is refused before the first token, you get a real HTTP status. A failure after the stream has started arrives as an error event inside the stream.

Rotate or revoke a key

- Rotate. On the AI gateway page, choose Rotate on the key's row. SynapBridge creates a new key with the same name, scopes and limits, shows its secret once, and revokes the old key immediately. The old key stops working at once.
- Zero-downtime swap. Create a second gateway key, deploy it to your app, then revoke the first key.
- Revoke. Choose Revoke on the row. A revoked key returns 401 invalid_api_key. A revoked key cannot be restored or rotated.
- Rotating, revoking and changing limits require the owner or admin role, and each change is written to the audit log.

Limits and notes

- The gateway requires the Team plan or above. A plan change can take up to a minute to reach the gateway, because workspace status is cached for 60 seconds.
- The requests-per-minute limit is counted in memory on each running server instance, so with several instances running, the total a key can send may be higher than its setting.
- Gateway keys cannot be used in a browser. Keep them on your server.
- Detection by data loss prevention is pattern-based and finds known patterns only.
- This guide describes the gateway as built today. Image input, structured output formats and forced tool choice are not supported yet.
- Compliance: SynapBridge is architected to align with the SOC 2 Trust Services Criteria, and with HIPAA and PCI-DSS requirements. The HIPAA BAA program is planned.