

Limits and rate limits

DOCUMENT	VERSION	STATUS
SB-GD-123	1.0, October 2026	Public. Checked against the product, October 2026

Who this is for

Developers and admins planning how much traffic, data and usage their workspace can send to SynapBridge.

Before you start

- Plan names used here are Free, Personal, Pro, Team, Business and Enterprise.
- Rate limits are per API key. Usage allowances such as AI runs and memory storage are per workspace.
- SynapBridge does not send X-RateLimit-Limit, X-RateLimit-Remaining or X-RateLimit-Reset headers on the AI gateway or the public REST API. Read the status code and the Retry-After header instead. The headers that are sent are listed at the end of this page.

AI gateway limits

The AI gateway needs the Team plan or above. See the AI gateway guide for how to set these.

- Requests per minute, per gateway key. You set it per key. It is a whole number from 1 to 6,000, and the default is 120. The count uses a sliding one-minute window.
- Monthly limit per gateway key, in USD. Optional. Above 0 and up to 1,000,000, counted per calendar month in UTC.
- Allowed models per key. Up to 50 entries.
- Request body: up to 2 MB.
- Messages in one request: up to 1,000.
- Output tokens (max_tokens): 1 to 200,000.
- Tools in one request: up to 64, on Anthropic models only.
- Stop sequences: up to 4 on the OpenAI route.

When a key goes over its requests-per-minute limit, the gateway returns 429 with the code `rate_limited` and a `Retry-After` header in seconds. Wait that long, then retry.

When a key reaches its monthly limit, or a budget set to block on breach is used up, the gateway returns 402, not 429. Retrying does not help until the window resets or an admin raises the limit.

Note The per-minute count is kept in memory on each running server instance. With more than one instance running, the total a key can send per minute may be higher than its setting.

AI run limits

Free and Personal workspaces have a monthly pool of AI runs, counted across the whole workspace in the UTC calendar month. The size of the pool is shown in the app.

Other plans:

- Pro, Team, Business and Enterprise: no run cap

When the pool is used up, the call is refused with 429 `monthly_run_limit_reached`. The pool resets at the start of the next UTC month.

Free and Personal workspaces can also connect a limited number of AI providers: Free 2, Personal 4. Pro and above have no cap on providers.

Memory storage

Each plan has an allowance for Memory Graph storage, counted across the workspace:

- Free: 100 MB
- Personal: 2 GB
- Pro: 10 GB
- Team: 50 GB
- Business: 500 GB
- Enterprise: custom

When the allowance is full, a request to add a memory fact is refused with 413 `storage_limit_reached`. Archived facts do not count toward usage.

Memory import from another AI tool is limited to 5 imports per user in any rolling 24 hours. Past that, the request returns 429 `daily_quota_exceeded`.

Other plan limits

- Custom team personas: Free 0, Personal 0, Pro 1, Team 10, Business and Enterprise no cap.
- Workflow size: Free and Personal workflows can have up to 10 steps and 20 connections. Pro and above have no plan cap.

Response headers that are sent

- AI gateway: cache-control: private, no-store on all responses, and x-request-id on chat completions and messages calls, including errors.
- AI gateway, non-streaming successes: x-synapbridge-invocation-id and x-synapbridge-cost-usd.
- AI gateway, 429 and 503 responses with a wait time: retry-after, in seconds.

No other rate-limit headers are sent on these two surfaces.

Limits and notes

- Limits can change as plans change. The figures here are the ones enforced today.
- Rate limits and usage allowances are separate. A key can be within its rate limit and still be refused because a budget or allowance is used up.
- 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.