

Gateway security

Evaluate the authority of each interface: application inference, administration, tool execution and outbound networking. A single API-key setting rarely describes all four.

Provider / product: _____ Version / artifact: _____

Deployment / topology: _____

Enabled features / configuration: _____

Reviewer / date: _____

01 Which identities can invoke inference, change routes and install tools?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

02 Can an application request reach a private or unapproved destination?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

03 Is tool policy applied to the exact forwarded message?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

04 Which patch and configuration evidence establishes the boundary?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

Evidence status: verified / partial / missing / not applicable (record the reason).

[Read the guide and vendor sources online](#)

OneQuill develops OneVir. Worksheet supports evaluation; it does not certify a product.

Published 2026-10-07

Privacy and data handling

Privacy is a property of the complete route and its retained copies. Gateway processing, upstream inference, observability and backups can have different terms.

Provider / product: _____ Version / artifact: _____

Deployment / topology: _____

Enabled features / configuration: _____

Reviewer / date: _____

01 Where do prompts, outputs, files and metadata travel?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

02 Do all upstream endpoints and fallback routes meet approved terms?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

03 What remains in logs, caches, exports and backups, and for how long?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

04 How are deletion and retention changes verified?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

Evidence status: verified / partial / missing / not applicable (record the reason).

[Read the guide and vendor sources online](#)

OneQuill develops OneVir. Worksheet supports evaluation; it does not certify a product.

Published 2026-10-07

Budgets and scaling

A rate limit is not a global spend ceiling. Budgets require well-defined counters, concurrent admission, final settlement and a plan for incomplete usage.

Provider / product: _____ Version / artifact: _____

Deployment / topology: _____

Enabled features / configuration: _____

Reviewer / date: _____

01 Which counters are shared across replicas, regions and tenants?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

02 Is maximum admitted spend reserved atomically before dispatch?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

03 How are retries, partial streams and missing usage settled?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

04 Which workload and latency target establish usable capacity?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

Evidence status: verified / partial / missing / not applicable (record the reason).

[Read the guide and vendor sources online](#)

OneQuill develops OneVir. Worksheet supports evaluation; it does not certify a product.

Published 2026-10-07

Reliability

A reliable gateway has a defined response to overload, partial output and upstream failure. Fallback must preserve policy and the client's understanding of the result.

Provider / product: _____ Version / artifact: _____

Deployment / topology: _____

Enabled features / configuration: _____

Reviewer / date: _____

01 Where are size, concurrency and deadline limits enforced?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

02 How does the client distinguish complete and partial output?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

03 Do retries and fallback preserve privacy, cost and tool policy?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

04 What recovery evidence covers gateway, dependency and worker failure?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

Evidence status: verified / partial / missing / not applicable (record the reason).

[Read the guide and vendor sources online](#)

OneQuill develops OneVir. Worksheet supports evaluation; it does not certify a product.

Published 2026-10-07

Supply chain

Security claims should identify the artifact, dependency set and model assets actually executed. A project name and a version label alone do not establish provenance.

Provider / product: _____ Version / artifact: _____

Deployment / topology: _____

Enabled features / configuration: _____

Reviewer / date: _____

01 Can the running artifact be traced to an approved build and digest?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

02 Are dependencies, plugins and model assets pinned and reviewed?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

03 Which distribution channels and configurations does each advisory cover?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

04 Who owns incident containment, secret rotation and safe restoration?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

Evidence status: verified / partial / missing / not applicable (record the reason).

[Read the guide and vendor sources online](#)

OneQuill develops OneVir. Worksheet supports evaluation; it does not certify a product.

Published 2026-10-07

Inference security and operations

An inference engine has its own API, internal networks, model loader, GPU allocator and scheduler. Gateway policy supports access control but cannot repair a vulnerable kernel or guarantee backend capacity.

Provider / product: _____ **Version / artifact:** _____

Deployment / topology: _____

Enabled features / configuration: _____

Reviewer / date: _____

01 Which API endpoints, input features and internal ports are exposed?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

02 Who approves model assets, and how are tenants isolated?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

03 What workload demonstrates acceptable TTFT, token latency and cache pressure?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

04 How do worker recovery and cold or warm starts behave on this topology?

Evidence / source / result: _____

Owner: _____ Action / due date: _____

Evidence status: verified / partial / missing / not applicable (record the reason).

Read the guide and vendor sources online

OneQuill develops OneVir. Worksheet supports evaluation; it does not certify a product.

Published 2026-10-07