

LAYA ROUTER GATEWAY

Governed AI model routing for teams that need one stable client contract across many providers.

A control plane for choosing the right model, provider, endpoint, credential path, and fallback route without making every application understand the whole provider landscape.

OPENAI CLIENT ONE STABLE MODEL ID	LAYA CONTROL PLANE HEALTH · POOL · POLICY · METADATA	PROVIDER TRANSPORT ENDPOINT · AUTH · FAILOVER
--------------------------------------	---	--

THE BUSINESS CASE

Model choice is usually hidden inside application code. A provider outage, a pricing change, a protocol mismatch, or a missing key then becomes a product incident. LAYA makes that decision an inspectable routing layer.

1,000	9	4	77
MODEL / PROVIDER PAIRS	PROVIDER RECORDS	ROUTING PRESETS	TESTS PASSED

LOCAL OPERATOR PILOT · SOURCE-VERIFIED · DRAFT FOR OWNER REVIEW

The business problem

Most AI applications are not short of models. They are short of a dependable way to choose between models.

- Different providers expose different catalogs, prices, context limits, modalities, and endpoint protocols.
- A configured provider is not necessarily a healthy provider. Credits run out, rate limits arrive, and a provider can return an error-shaped HTTP 200.
- Copying provider keys into every application expands the attack surface and makes revocation, rotation, and ownership harder.
- A model name in a prompt is not a routing policy. It does not tell another system why a cheap model was acceptable, why a paid model was necessary, or which fallback was tried.

LAYA Router Gateway turns preset names into concrete, healthy, cost-aware model routes. A client sends one stable model name to one OpenAI-compatible endpoint. LAYA evaluates the task, applies policy, checks pool and provider health, resolves the route, dispatches the request, and records the attempt.

One router. Every model.

Free means explicit zero pricing or NVIDIA's current public "Free Endpoint" label. A label is not a permanent-expiry promise; end dates, deprecation, and rate limits are shown when NVIDIA publishes them. Models with an unknown price stay marked unknown rather than being guessed as free.

This is the live LAYA gateway running on 127.0.0.1:4021. Provider credentials stay in the gateway; the UI only handles virtual gateway keys and non-secret metadata.

ROSTER

1000

discovered model/provider pairs

ENABLED POOL

6

available to automatic routing

PRESETS

4

cheap → maximum brain

VIRTUAL KEYS

1

gateway-issued, provider-safe

Model pool

Disabled models disappear from automatic routing immediately.

Enable all

Disable all

Search model, provider, owner...

All providers

All modalities

All status

Any price

Clear

ON	MODEL ↓	PROVIDER ↓	CONTEXT ↓	MODALITIES ↓	CREATED ↓	CAPABILITY ↓	COST / EVIDENCE ↓	ENDPOINT / LIMITS ↓
☒	Anthropic: Claude Fable 5.1 openrouter/anthropic/claude-fable-5.1	openrouter	1000000	text/image/file → text	2026-09-01	53.4	\$10 in · \$50 out	catalog
☒	GPT-6 Astra commandcode/gpt-6-astra	commandcode	1050000	→ text	2026-09-24	52.7	unknown · no price field unknown	/chat/completions · /responses
☒	gpt-6-astra openai/gpt-6-astra	openai	—	→ text	2026-08-27	52.7	unknown · no price field unknown	catalog
☒	gpt-6-astra openai/gpt-6-astra	openai	—	→ text	2026-08-27	52.7	unknown · no price field unknown	catalog

LIVE CONTROL CENTER / 1600 × 1200 / ROSTER, ACTIVE POOL, PRESETS, VIRTUAL KEYS, AND MODEL-POOL CONTROLS

What is actually wired

Stable client contract

The gateway exposes POST /v1/chat/completions, GET /v1/models, GET /health, and GET /router/explain. A client can use a concrete model or a stable preset such as auto-router.

Model pool governance

Operators can search, filter by provider, modality, status, and price, sort the roster, toggle individual pairs, and use Enable all or Disable all. Discovery and routing eligibility remain separate.

Provider-aware dispatch

Provider ID, base URL, endpoint protocol, authentication scheme, model ID, credential source, health, and pool state are independent dimensions. Endpoint metadata wins over model-name prefixes.

Safe credential boundary

Clients receive virtual gateway keys. Provider credentials remain on the server side. The browser never receives an existing provider key back, and gateway key material is stored as a hash.

Honest failover

The router tries the same normalized model on another provider before escalating. It classifies credit, auth, rate-limit, upstream, and bad-request states, and never treats an error-shaped 200 as a completion.

A ROUTE, NOT A GUARANTEE

The route simulator explains what the policy would select. It does not promise that a provider will complete every future request. Health, credits, quotas, and upstream behavior can change after the decision.

1000

discovered model/provider pairs

6

available to automatic routing

4

cheap → maximum brain

1

gateway-issued, provider-safe

Model pool

Disabled models disappear from automatic routing immediately.

Enable all

Disable all

ON	MODEL :	PROVIDER	CONTEXT	MODALITIES	CREATED	CAPABILITY	COST / EVIDENCE :	ENDPOINT / LIMITS :
☐	NVIDIA: Nemotron 3 Super (free) openrouter/nvidia/nemotron-3-super-128b-a12b:free	openrouter	262144	text → text	2026-03-11	23"	\$0 in - \$0 out	catalog
☐	nvidia/nemotron-3-super-120b-a12b nvidia/nemotron-3-super-128b-a12b	nvidia	1000000	Text → Text	—	23"	\$0 - NVIDIA free endpoint explicit_nvidia_model_card	/chat/completions 40 rpm 10,000/day 39 benchmark rows expiry not published
☐	NVIDIA: Nemotron 3 Super openrouter/nvidia/nemotron-3-super-128b-a12b	openrouter	262144	text → text	2026-03-11	23"	\$0.08 in - \$0.45 out	catalog
☐	nemotron-3-super ollama-cloud/nemotron-3-super	ollama-cloud	—	— → text	2026-03-11	12"	unknown - no price field unknown	catalog

Route simulator

Ask the actual judge + policy path without spending a completion.

TASK

Investigate an unknown production failure, inspect the repository, and propose a verified fix.

POLICY

AUTO / balanced

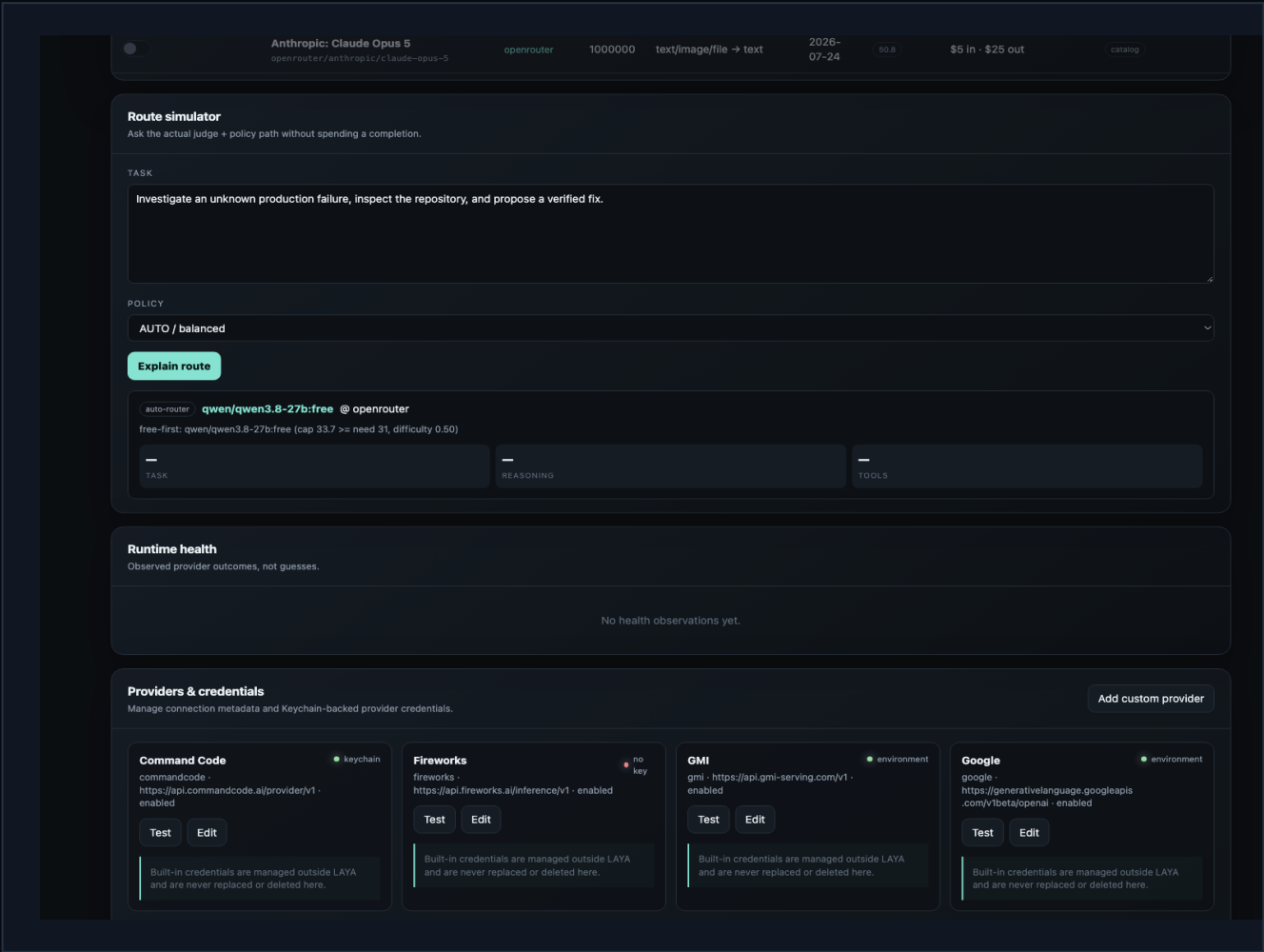
Explain route

MODEL POOL / SEARCH, FILTERS, SORTABLE ROSTER, COST PROVENANCE, AND ENDPOINT / LIMIT EVIDENCE

A route you can inspect

The route simulator runs the real decision path without spending a completion. The screenshot captures a completed request, not a loading state.

TASK	Investigate an unknown production failure, inspect the repository, and propose a verified fix.
POLICY	AUTO / balanced
SELECTED	qwen/qwen3.8-27b:free @ openrouter
REASON	free-first: qwen/qwen3.8-27b:free (cap 33.7 >= need 31, difficulty 0.50)

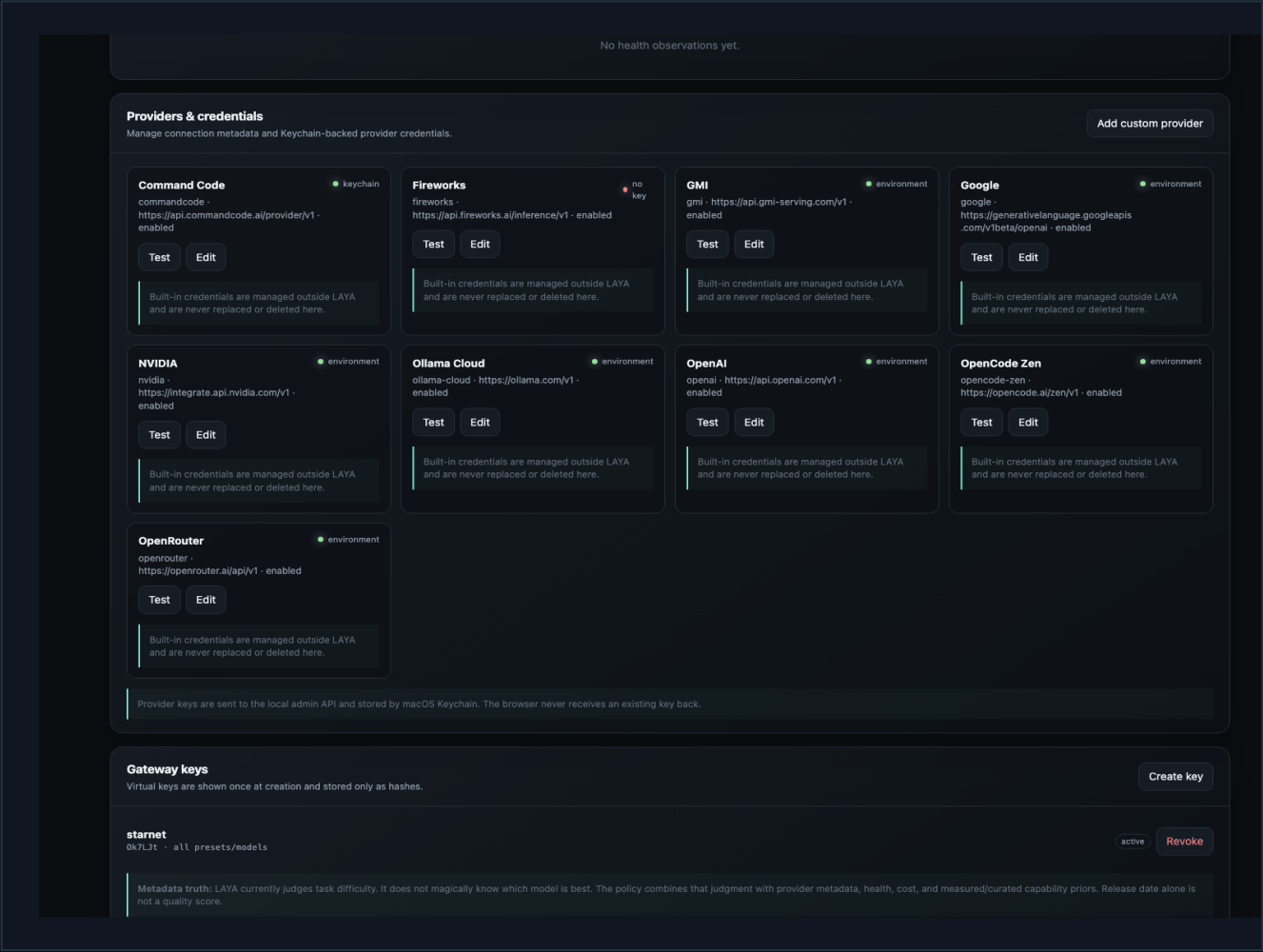


COMPLETED ROUTE / SELECTED MODEL, PROVIDER, AND POLICY REASON VISIBLE IN THE LIVE UI

Failure is part of the product

A provider failure should not automatically become an application failure. LAYA records the attempted route and keeps looking within explicit rules.

UPSTREAM SIGNAL	ROUTER TREATMENT
402 / 401 / 403	exhausted or unauthorized; skip that pair until its TTL lapses
429	rate limited; temporary cooldown, then retry later
5xx	upstream error; brief skip and retry elsewhere
400	bad request; do not retry the same failure everywhere
200 with error body	failure, never a fake completion



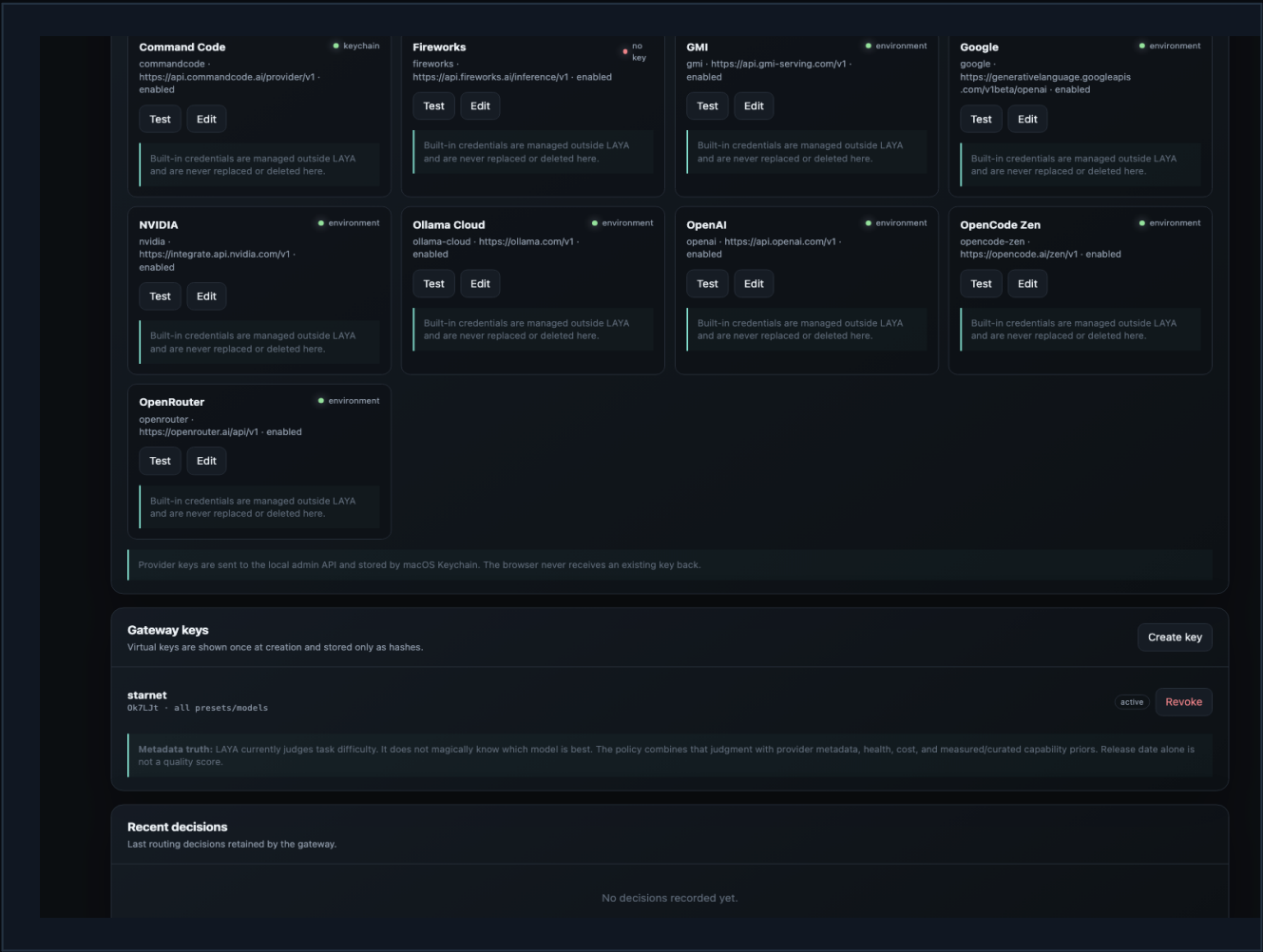
PROVIDER ADMINISTRATION / NINE PROVIDER RECORDS, CREDENTIAL SOURCES, BASE URLS, AND TEST / EDIT CONTROLS

Evidence, not theater

CHECK	RESULT	MEANING
Health	HTTP 200 / ok	The isolated local gateway is serving its control plane.
Roster	1,000 pairs	Catalog ingestion and management are active.
Enabled pool	6 pairs	Automatic eligibility is a separate operational state.
Providers	9 records	Multiple upstream configurations share one surface.
Tests	77 passed / 0 failed	Policy, keys, metadata, pool, and transport behavior are covered.
Failover E2E	all checks passed	Credit exhaustion, same-model failover, and escalation are exercised.
Completion smoke	HTTP 200 / ROUTER_OK	The gateway can dispatch a real completion.
StarNet contract	15 assertions	The desktop client exposes a first-class LAYA provider.

Virtual keys stay separate

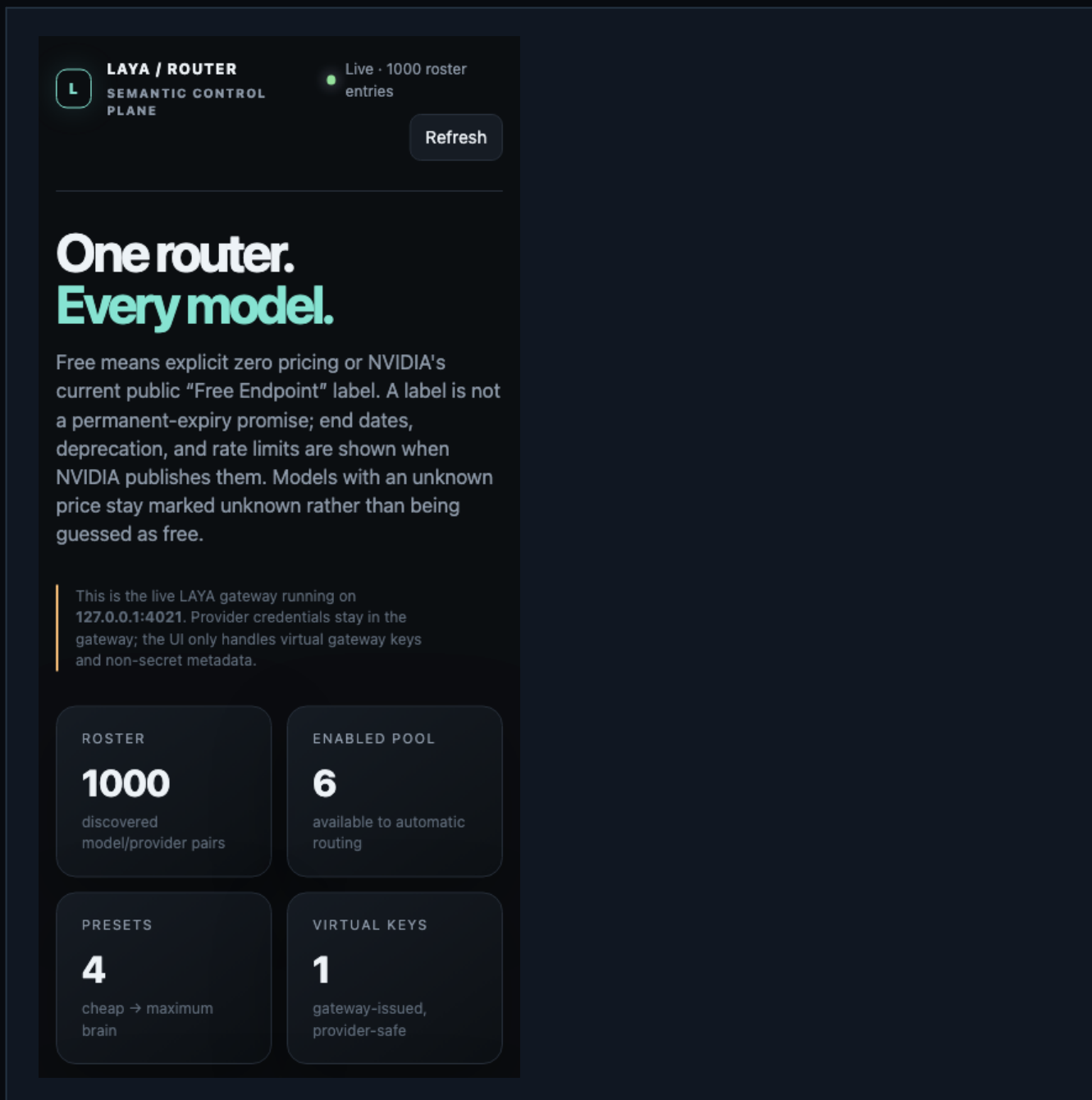
Clients receive a virtual gateway key rather than a provider key. The UI can show a safe fingerprint or identifier for recognition and lifecycle management without returning the usable secret.



GATEWAY KEYS / VIRTUAL-CLIENT KEY METADATA; NO USABLE PROVIDER SECRET IS SHOWN

The control plane holds together

The same operating model has to survive a narrow screen. At mobile width, the control center keeps the service status, headline metrics, routing model pool, presets, and quick navigation available without pretending the desktop layout fits everywhere.



RESPONSIVE CONTROL CENTER / 390 × 844 / CORE STATUS AND METRICS AT A NARROW WIDTH

What a business can build

One internal AI gateway

Give several teams one stable client contract while the underlying model roster changes behind it.

Cost-aware model choice

Send routine work to an eligible free route and reserve paid routes for work that clears a higher capability requirement.

Provider resilience

Keep serving a request when one provider runs out of credit or becomes temporarily unavailable.

Credential governance

Give applications virtual keys instead of distributing provider credentials across environments.

Operator visibility

Search, filter, sort, enable, disable, test, and inspect routes from one control surface.

Auditable decisions

Retain the selected model, provider, reason, and attempted routes for support and cost analysis.

The portfolio point: this is the layer that turns a model provider list into an operating system for AI work. It is useful before the first prompt is sent, and it becomes more valuable as the roster, provider contracts, and traffic patterns grow.

STATED LIMITS

- This is a local operator pilot on loopback, not a hosted multi-tenant production claim.
- Capability is measured for a subset of models and estimated by family priors for the rest; the interface marks the difference.
- Unknown price, free status, and endpoint expiration remain unknown when evidence is missing.
- The route simulator is a decision preview, not a guarantee of future provider health.
- The StarNet integration is source-only in this work and has not been committed, deployed, or installed into the signed app.

Build the routing layer your AI product can operate.

Supreme / GiveMePrompts

MAIN SITE givemeprompts.com

STORE givemeprompts.com/store

PORTFOLIO contra.com/supreme_dream_63yh320u

MUSIC / BEATS samuraiblaque.com/beats

INSTAGRAM instagram.com/givemeprompts

LINKEDIN linkedin.com/in/givemeprompts

CONTACT admin@givemeprompts.com