

Inference at market prices.

And much more.

An AI inference marketplace and a workspace for building modern products. Connect models, run managed agents, host your tools and put spare capacity to work.

MODEL CHOICE



DeepSeek



Qwen



Llama



Gemma



Mistral



MiniMax

One platform. Four value propositions.

01

Inference as a service

Fit your workflow. Connect applications, automations and agent harnesses through an OpenAI-compatible API.

Control routing. Choose providers, regions, context and latency limits. Configure retries or rerouting.

Follow market prices. Compare provider-set offers, track usage and review each request's settled cost.

02

Hosting and storage

Managed agents. Run them in parallel, each on a harness from the catalog. Follow progress, pause tasks and review history.

Run a regular VM. Choose compute and storage for tools, files and services. Connect via SSH or browser terminal.

Rent GPU containers. H200s by the hour, up to 32 GPUs in a single container.

03

Team operations

Manage access. Bring colleagues into a shared workspace with owner, admin and member roles.

Track productivity. Follow agent activity, task completion and execution history.

Share resources. Share compute, inference and connections. Review activity alongside token usage, spending and earnings.

04

Efficient hardware utilization

Sell on the open market. Offer inference, choose your GPU, model and pricing, then monitor delivery and earnings.

Power your team. Share local inference with colleagues with zero exchange fee.

Choose your setup. Use hosted resources, your own device, or both. Operating costs remain yours.

Spawn a managed agent.

A virtual machine with the coding agent installed and pointed at the exchange — pi, Claude Code, opencode, Cline and more, each on a model we have tested for the work. It keeps running when you close the tab.

FROM \$6 / MONTH
Small · Standard · Large

2–8 vCPUs / 2–8 GB memory
Persistent workspace disk across restarts

01

Choose a VM size

Pick the compute and memory your agent needs. Keep its working files on a persistent disk.

02

Pick a vetted harness

Choose from the harness catalog. Each one is signed in and pointed at a model we run for that task.

03

Tools already wired up

GitHub, qrios research and ix MCP are connected. Open its terminal from any device.

Or connect your project through the API.

Choose a live model. Compare providers, token prices and context size. **Create an API key.** Keep it in your application's environment. **Make requests.** Review token usage, provider and settled cost in the console.

OPENAI-COMPATIBLE API

https://api.inference.exchange/v1

POST /chat/completions

Know what you pay for.

Example starting rates in USD per million tokens

MODEL	INPUT / 1M TOKENS	OUTPUT / 1M TOKENS
GLM-5.3-Flash	\$0.075	\$0.25
DeepSeek-V4-Flash-0731	\$0.06	\$0.18
GLM-5.3	\$0.5625	\$2.50

"From" rates observed 23 September 2026. Offers and availability change. Credit, service fee and applicable VAT are shown before payment; only credit is added to the inference balance.

Pay for usage. Each request reserves its maximum cost, settles against metered consumption and returns the unused amount.

Hosting and storage. Managed VMs start at \$6/month, and H200 GPU containers scale to 32 GPUs. Contact us for current configurations and pricing.

A few practical answers.

What happens to prompts? API logs record usage metadata, not prompt or response content. Console Chat retains chats until deleted. Serving providers apply their own retention policies.

Uncensored models? Available depending on live supply. Review the model's license and intended use.

How do I get access? Access is invitation-based. Email us for onboarding or support; include the request ID for an existing request. The platform is in beta with no availability commitment.

No eligible provider? The request is refused. Use routing preferences to choose eligible providers and regions.