quinta.

SOVEREIGN ON-PREMISE AI · BY TWENTY5AI

TECHNICAL WHITEPAPER

Your AI. Your hardware. Not a single byte leaves the building.

How Quinta turns your own hardware into a private, OpenAI-compatible AI service — with enterprise management, a complete audit trail and no data egress.

GDPR-COMPLIANT | EU-AI-ACT-READY | 100% EUROPEAN | OPENAI-COMPATIBLE API

twenty5ai · July 2026 · quinta.twenty5ai.com

01 — Executive summary

Sovereignty costs no performance — it delivers it.

Quinta is not another inference engine, but the complete operating layer a regulated enterprise needs to run AI in production: access control, multi-GPU orchestration, governance and a complete audit trail — entirely on your own hardware.

- OpenAI-compatible API: existing applications migrate by changing `base_url` and the API key — nothing else.
- Runs on-premise, from a workstation to an AI appliance; scales horizontally across self-registering GPU nodes.
- Two inference engines built in: a lightweight model runner for a simple start, vLLM for high performance and the entire HuggingFace catalogue.
- Enterprise management: dashboard, RBAC, four login methods, multi-organization, audit trail, fine-tuning and MCP.
- No data transfer, no external processor for inference — the basis for GDPR, the EU AI Act and NIS2.

02 — The problem

Cloud AI means someone else’s infrastructure, someone else’s rules.

Data leaves Europe
The most sensitive data goes to US cloud APIs — without an audit trail, with GDPR, EU AI Act and NIS2 risk.

Costs without a ceiling
Token billing makes budgets unpredictable. Every new use case raises the running cost.

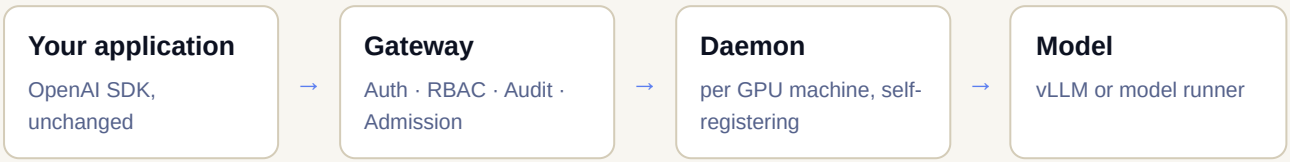
Dependence on vendors
Models, prices and availability are set by a third party. Switching an API means a rebuild, not a move.

No proof
The audit trail ends at the provider’s system boundary — exactly where regulation demands it.

03 — The platform

Not a model runner. The complete operating layer.

Every request takes the same path: through the gateway (the gatekeeper), to the daemon on the GPU machine, into the model — connected through a central registry. Every additional machine running the Quinta daemon becomes part of the platform automatically.



Two inference engines

vLLM — high performance for production, the entire HuggingFace catalogue (Llama, Mistral, Qwen, Gemma, DeepSeek). Fully automated lifecycle: download → warmup → ready.

Lightweight model runner — a simple start, from a workstation to the DGX Spark class. GPU detection for NVIDIA, AMD and Intel, automatic routing.

Migration in two lines

```
client = OpenAI(
  base_url="https://quinta.intern/v1",
  api_key="qnt-prod-…",
)
```

Chat, embeddings, rerank, audio transcription and the Responses API — no reprogramming, no new SDK.

04 — Performance

Same engine. A different behaviour at the limit.

18×
higher success rate under extreme load (bounded admission)

~56 ms
additional first-token latency — per-user token rate identical

512
concurrent requests in the test

Benchmark of the underlying engine, measured on an NVIDIA DGX Spark (128 GB): 31,200 requests, up to 512 concurrent requests, Quinta gateway against vLLM alone. The gateway meters requests and absorbs errors before users see them — at identical inference.

05 — Enterprise features

Built for IT leadership, not just for developers.

Deployments by click
Roll out, scale and retract models — without a terminal.

API key management
Keys per team and application, revocable at any time.

Multi-organization (RBAC)
Multi-tenant, roles & permissions, team invitations.

4 login methods
Password, 2-factor (TOTP), passkeys, SSO/SAML (Entra ID, Okta).

Usage & metrics
Tracking per unit plus Prometheus metrics.

Complete audit trail
Every access, every change traceable — tamper-evident.

Fine-tuning
Adapt your own models, directly from the interface.

MCP interface
AI assistants manage the infrastructure themselves.

06 — Compliance & sovereignty

Regulation is not an appendix. It is the reason.

GDPR
Processing stays in the house: no third-country transfer, no processor for inference.

EU AI Act
Full control over models and versions; the audit trail as documentation basis.

NIS2
Operation in your own security zone: SSO/RBAC, traceable administration.

No DPA for inference
Without an external processor, the transfer and DPA risk of cloud AI disappears.

> “Not a single byte leaves the building” means, in data-protection terms: no transfer, no external processor, no access under foreign law.

07 — Rollout

Three steps to your own AI service.

01 · Plan A 30-minute conversation: use cases, hardware, compliance framework. The deployment plan follows.

02 · Deploy The daemon onto your GPU machines, gateway and dashboard alongside — nodes register themselves, models start automatically. Hours, not months.

03 · Operate Deployments by click, API keys, usage per unit, Prometheus metrics and a complete audit trail in the dashboard.

08 — Editions

Business and Enterprise.

Business For production use — multiple compute nodes, RBAC, API keys, model management. Price on request (license per site).

Enterprise Additionally multi-site, SSO/SAML, 2FA & passkeys, audit trail, fine-tuning UI & MCP, a dedicated contact. Custom contract & SLA.

quinta.
Bring your AI home. Book a demo (30 min.).

hello@twenty5ai.com
quinta.twenty5ai.com

Benchmark figures refer to the measurement of the underlying engine (Quinta gateway vs. vLLM alone, NVIDIA DGX Spark 128 GB) and are reproduced unchanged. Compliance statements describe the architecture, not certifications. The inference engine builds on open-source technology (Apache 2.0); the management layer is developed in-house by twenty5ai. This document is not legal advice.