quinta.

SOUVERÄNE ON-PREMISE-KI · VON TWENTY5AI

TECHNISCHES WHITEPAPER

Ihre KI. Ihre Hardware. Kein Byte verlässt das Haus.

Wie Quinta aus eigener Hardware einen privaten, OpenAI-kompatiblen KI-Dienst macht — mit Enterprise-Verwaltung, lückenlosem Audit-Trail und ohne Datenabfluss.

GDPR-KONFORM | EU-AI-ACT-READY | 100 % EUROPÄISCH | OPENAI-KOMPATIBLE API

twenty5ai · Stand: Juli 2026 · quinta.twenty5ai.com

01 — Kurzfassung

Souveränität kostet keine Leistung — sie liefert sie.

Quinta ist keine weitere Inferenz-Engine, sondern die komplette Betriebsschicht, die ein reguliertes Unternehmen braucht, um KI produktiv zu betreiben: Zugriffskontrolle, Multi-GPU-Orchestrierung, Governance und ein lückenloser Prüfpfad — vollständig auf der eigenen Hardware.

- OpenAI-kompatible API: bestehende Anwendungen migrieren, indem sie `base_url` und API-Key ändern — sonst nichts.
- Läuft on-premise, von der Workstation bis zur KI-Appliance; skaliert horizontal über selbstregistrierende GPU-Knoten.
- Zwei Inferenz-Engines eingebaut: ein leichtgewichtiger Modell-Runner für den Einstieg, vLLM für Hochleistung und den gesamten HuggingFace-Katalog.
- Enterprise-Verwaltung: Dashboard, RBAC, vier Login-Methoden, Multi-Organisation, Audit-Trail, Fine-Tuning und MCP.
- Kein Datentransfer, kein externer Auftragsverarbeiter für die Inferenz — die Grundlage für DSGVO, EU AI Act und NIS2.

02 — Das Problem

Cloud-KI heißt: fremde Infrastruktur, fremde Regeln.

Daten verlassen Europa
Sensibelste Daten gehen an US-Cloud-APIs — ohne Audit-Trail, mit GDPR-, EU-AI-Act- und NIS2-Risiko.

Kosten ohne Obergrenze
Token-Abrechnung macht Budgets unkalkulierbar. Jeder neue Anwendungsfall verteuert den Betrieb.

Abhängigkeit von Anbietern
Modelle, Preise und Verfügbarkeit bestimmt ein Dritter. Ein API-Wechsel bedeutet Umbau statt Umzug.

Kein Nachweis
Der Prüfpfad endet an der Systemgrenze des Anbieters — genau dort, wo Regulierung ihn verlangt.

03 — Die Plattform

Kein Model-Runner. Die komplette Betriebsschicht.

Jede Anfrage nimmt denselben Weg: durch das Gateway (Pförtner), zum Daemon auf dem GPU-Rechner, in das Modell — verbunden über eine zentrale Registry. Jeder weitere Rechner mit dem Quinta-Daemon wird automatisch Teil der Plattform.

Ihre Anwendung
OpenAI-SDK, unverändert

→

Gateway
Auth · RBAC · Audit · Dosierung

→

Daemon
je GPU-Rechner, selbstregistrierend

→

Modell
vLLM oder Modell-Runner

Zwei Inferenz-Engines

vLLM — Hochleistung für Produktion, ganzer HuggingFace-Katalog (Llama, Mistral, Qwen, Gemma, DeepSeek). Vollautomatischer Lebenszyklus: Download → Warmup → ready.

Leichter Modell-Runner — einfacher Einstieg, von der Workstation bis zur DGX-Spark-Klasse. GPU-Erkennung für NVIDIA, AMD und Intel, automatisches Routing.

Migration in zwei Zeilen

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

Chat, Embeddings, Rerank, Audio-Transkription und die Responses-API — kein Umprogrammieren, kein neues SDK.

04 — Leistung

Gleiche Engine. Ein anderes Verhalten am Limit.

18×
höhere Erfolgsquote unter Extremlast (bounded admission)

~56 ms
zusätzliche Erst-Latenz — Token-Rate pro Nutzer identisch

512
gleichzeitige Anfragen im Test

Benchmark der zugrunde liegenden Engine, gemessen auf NVIDIA DGX Spark (128 GB): 31.200 Requests, bis 512 gleichzeitige Anfragen, Quinta-Gateway gegen vLLM pur. Das Gateway dosiert Anfragen und fängt Fehler ab, bevor Nutzer sie sehen — bei identischer Inferenz.

05 — Enterprise-Funktionen

Gebaut für IT-Leitung, nicht nur für Entwickler.

Deployments per Klick
Modelle ausrollen, skalieren und zurückziehen — ohne Terminal.

API-Key-Verwaltung
Keys je Team und Anwendung, jederzeit widerrufbar.

Multi-Organisation (RBAC)
Mandantenfähig, Rollen & Rechte, Team-Einladungen.

4 Login-Methoden
Passwort, 2-Faktor (TOTP), Passkeys, SSO/SAML (Entra ID, Okta).

Verbrauch & Metriken
Tracking je Einheit plus Prometheus-Metriken.

Lückenloser Audit-Trail
Jeder Zugriff, jede Änderung nachvollziehbar — revisionssicher.

Fine-Tuning
Eigene Modelle anpassen, direkt aus der Oberfläche.

MCP-Schnittstelle
KI-Assistenten verwalten die Infrastruktur selbst.

06 — Compliance & Souveränität

Regulatorik ist kein Anhang. Sie ist der Grund.

GDPR / DSGVO
Verarbeitung bleibt im Haus: kein Drittstaaten-Transfer, kein Auftragsverarbeiter für die Inferenz.

EU AI Act
Volle Kontrolle über Modelle und Versionen; Audit-Trail als Dokumentationsgrundlage.

NIS2
Betrieb in der eigenen Sicherheitszone: SSO/RBAC, nachvollziehbare Administration.

Kein AVV für die Inferenz
Ohne externen Auftragsverarbeiter entfällt das Übermittlungs- und AVV-Risiko der Cloud-KI.

> „Kein Byte verlässt das Haus" heißt datenschutzrechtlich: kein Transfer, kein fremder Auftragsverarbeiter, kein Zugriff nach fremdem Recht.

07 — Einführung

Drei Schritte bis zum eigenen KI-Dienst.

01 · Planen 30-Minuten-Gespräch: Anwendungsfälle, Hardware, Compliance-Rahmen. Danach steht der Deployment-Plan.

02 · Deployen Daemon auf die GPU-Rechner, Gateway und Dashboard dazu — Knoten registrieren sich selbst, Modelle starten automatisch. Stunden, nicht Monate.

03 · Betreiben Deployments per Klick, API-Keys, Verbrauch pro Einheit, Prometheus-Metriken und lückenloser Audit-Trail im Dashboard.

08 — Editionen

Business und Enterprise.

Business Für den produktiven Einsatz — mehrere Rechenknoten, RBAC, API-Keys, Modell-Management. Preis auf Anfrage (Lizenz je Standort).

Enterprise Zusätzlich Multi-Standort, SSO/SAML, 2FA & Passkeys, Audit-Trail, Fine-Tuning-UI & MCP, fester Ansprechpartner. Individueller Vertrag & SLA.

quinta.
Holen Sie Ihre KI nach Hause. Demo buchen (30 Min.).

hello@twenty5ai.com
quinta.twenty5ai.com

Benchmark-Zahlen beziehen sich auf die Messung der zugrunde liegenden Engine (Quinta-Gateway vs. vLLM pur, NVIDIA DGX Spark 128 GB) und werden unverändert wiedergegeben. Aussagen zu Compliance beschreiben die Architektur, keine Zertifizierungen. Die Inferenz-Engine basiert auf Open-Source-Technologie (Apache 2.0); die Verwaltungsebene ist eine Eigenentwicklung von twenty5ai. Dieses Dokument ist keine Rechtsberatung.