XINITY

xinity.ai

VON OPENAI ZU XINITY

KI-PLATTFORM MIGRATIONS-WHITEPAPER SERIE 2026

LEGAL NOTICES

Xinity reminds you to carefully read through and completely understand all content in this section before you read or use this document. If you read or use this document, it is considered that you have identified and accepted all contents declared in this section.

1. This document is published by Xinity for informational purposes. The contents are intended for legal and compliant business activities. You shall not use or disclose all or part of the contents to any third party without written permission from Xinity.

2. This document may be subject to change without notice due to product upgrades, adjustment, and other reasons. Xinity reserves the right to modify the contents without notice.

3. This document is only intended for product and service reference. Xinity provides this document for current products and services with current functions, which may be subject to change.

4. All content including images, architecture design, page layout, and description text is owned by Xinity. You shall not use, modify, copy, or publish the content without written permission.

5. If you discover any errors or mistakes within this document, please contact Xinity directly.

THE AUTHORS

CORE AUTHORS

Alexander Zehetmaier (CEO & Co-Founder, Xinity)

TECHNICAL REVIEW

Jonas (CTO & Co-Founder, Xinity)

EDITING AND DESIGN

Xinity Marketing Team

TARGET AUDIENCE

Dieser Leitfaden richtet sich an Engineering-Teams, CTOs und IT-Entscheidungstraeger, die derzeit OpenAI's API-Dienste (GPT-5.4, GPT-4.1, o3, o4-mini, Whisper, Embeddings, Assistants API) nutzen und KI-Workloads auf eine souveraene On-Premise-Infrastruktur migrieren muessen. Ob Sie eine schrittweise Migration zur Reduzierung der Cloud-Abhaengigkeit planen oder einen vollstaendigen Plattformwechsel aufgrund regulatorischer Anforderungen durchfuehren -- dieses Whitepaper liefert die technischen Zuordnungen, Migrationsprozesse und Werkzeuganleitungen fuer eine sichere Umsetzung.

CONTENTS

1. Enterprise AI ohne Kompromisse: Warum Xinity die bessere Wahl ist

2. Ihr OpenAI-Stack, neu aufgebaut auf Xinity (Zugeordnet & Bereit)

2.1 Kern-Inferenz & Chat Completions

2.2 Embeddings & Vektorsuche

2.3 Audio & Sprache

2.4 Plattform & DevOps

3. Migrationsprozess

3.1 Bestandsaufnahme & Discovery

3.2 Infrastrukturplanung & Design

3.3 Pilot-Migration

3.4 Vollstaendige Migration & Optimierung

4. Migrations-Werkzeuge & Beschleuniger

4.1 API-Kompatibilitaetsschicht

4.2 Observability & Betrieb

5. Naechste Schritte: Starten Sie Ihre Migration mit Xinity

1. ENTERPRISE AI OHNE KOMPROMISSE: WARUM XINITY DIE BESSERE WAHL IST

Wenn Ihr Unternehmen KI-Workloads in der Produktion betreibt, bietet die Migration von Cloud-gehosteten KI-APIs zur On-Premise-Plattform von Xinity etwas, das kein Cloud-Anbieter liefern kann: vollstaendige architektonische Souveraenitaet ueber Ihre Daten, Modelle und Inferenz-Infrastruktur. Dies ist kein einfacher Anbieterwechsel -- es ist ein fundamentaler Wandel vom Mieten von KI-Kapazitaet zum Besitzen.

-- Architektonische Souveraenitaet statt Richtlinien-Versprechen

Cloud-KI-Anbieter bieten vertraglichen Datenschutz durch Nutzungsbedingungen und Auftragsverarbeitungsvertraege. Xinity liefert architektonische Souveraenitaet: Ihre Daten verlassen niemals Hardware, die Sie physisch besitzen und kontrollieren. Fuer regulierte Branchen -- Gesundheitswesen, Recht, Finanzdienstleistungen, Medien und Fertigung -- ist diese Unterscheidung nicht akademisch. Es ist der Unterschied zwischen Compliance-Risiko und Compliance-Sicherheit. Keine auslaendische Regierungsvorladung, keine Aenderung der Cloud-Anbieter-Richtlinien und keine geopolitische Verschiebung kann Daten beeinflussen, die ausschliesslich auf Ihren Raeumlichkeiten existieren.

-- Planbare Wirtschaftlichkeit im Enterprise-Massstab

Cloud-KI-Preise skalieren mit dem Verbrauch: Jeder API-Aufruf, jedes Token, jede GPU-Stunde wird gemessen und abgerechnet. Xinity's On-Premise-Modell wandelt variable OPEX in planbare CAPEX um. Kunden, die Xinity Runtime auf ASUS Ascent GX10 Servern einsetzen, berichten von ca. 80% Kostenersparnis gegenueber vergleichbarer Cloud-Kapazitaet. Im Massstab bedeutet das ca. 320 EUR/Jahr Stromkosten gegenueber 18.600 EUR/Jahr fuer vergleichbare Cloud-Rechenleistung.

-- Latenzfreie Inferenz fuer kritische Anwendungen

On-Premise-KI eliminiert Netzwerk-Roundtrips zu entfernten Cloud-Regionen. Fuer latenzsensitive Anwendungen -- Echtzeit-Dokumentenanalyse, Qualitaetskontrolle in der Produktion, klinische Entscheidungsunterstuetzung -- liefert lokale Inferenz konsistente Sub-Millisekunden-Antwortzeiten ohne Abhaengigkeit von Internetverbindung, Cloud-Region-Verfuegbarkeit oder grenzueberschreitenden Datentransfervorschriften.

-- Regulatorischer Rueckenwind beschleunigt die Adoption

Der EU Digital Networks Act (vorgeschlagen Januar 2026) mit Compliance-Fristen im August 2026, die 20 Milliarden EUR InvestAI-Foerderinitiative und aufkommende 'Buy European'-Beschaffungsregeln validieren die These der souveraenen KI-Infrastruktur. Organisationen, die jetzt auf On-Premise-KI migrieren, positionieren sich vor den Regulierungen statt spaeter hektisch reagieren zu muessen.

-- OpenAI-kompatible APIs -- migrieren ohne Neuentwicklung

Xinity Runtime stellt OpenAI-kompatible API-Endpunkte bereit. Das bedeutet: Ihr bestehender Anwendungscode, SDKs, Prompt-Bibliotheken und Orchestrierungsframeworks funktionieren mit minimalen Aenderungen weiter. Sie aendern die Base-URL und den API-Key; Ihre Anwendungen bemerken keinen Unterschied.

2. IHR OPENAI-STACK, NEU AUFGEBAUT AUF XINITY (ZUGEORDNET & BEREIT)

Dieser Abschnitt stellt ein klares Faehigkeiten-Mapping fuer Organisationen bereit, die von OpenAIs Cloud-API zu Xinitys On-Premise-Plattform migrieren. Das Ziel ist es, Ihnen zu helfen, jeden OpenAI-Dienst, den Sie derzeit nutzen -- Chat Completions, Embeddings, Audio-Transkription, Fine-Tuning und Orchestrierung -- in funktional aequivalente oder ueberlegene Xinity-Faehigkeiten zu uebersetzen.

Da Xinity Runtime OpenAI-kompatible API-Endpunkte bereitstellt, sind viele Migrationen so einfach wie das Aendern der Base-URL.

Kern-Inferenz & Chat Completions

Source Service	Xinity Equivalent	Migration Notes
GPT-5.4 / GPT-5.4-mini (Flaggschiff)	Xinity Runtime (Mistral Large 3 / Nemotron-Ultra)	OpenAI-kompatibler /v1/chat/completions Endpunkt. Nur base_url und api_key aendern. Kein Anwendungscode-Umbau.
GPT-4.1 (1M Token Kontext, Coding)	Xinity Runtime (Mistral Large 3 / Nemotron-Ultra 128K Kontext)	Long-Context Coding und Instruktionen. Lokale Inferenz eliminiert Token-Preise. Fuer 1M+: Chunking + RAG-Pipeline.
GPT-5.4-nano / GPT-4.1-mini	Xinity Runtime (Qwen3.5 8B / Mistral Small 3)	Kostenoptimierte kleine Modelle fuer Hochdurchsatz-Aufgaben.
o3 / o3-pro (Reasoning)	Xinity Runtime (Qwen3.5 72B-Reasoning / Qwen3.5 72B)	On-Premise Reasoning-Modelle. Kein Token-basierter Aufpreis.
o4-mini (Kosteneffizientes Reasoning)	Xinity Runtime (Qwen3.5 8B / Mistral Small 3 mit Reasoning)	80% guenstiger als o3. Lokales Reasoning zum Festpreis.

Embeddings & Vektorsuche

Source Service	Xinity Equivalent	Migration Notes
text-embedding-3-small/large	Xinity Runtime (BGE-M3, E5-Mistral)	Lokale Embedding-Generierung. Keine Token-Gebuehren. Integration mit lokalen Vektor-DBs.
OpenAI Vector Store	On-Prem Vektor-DB (Qdrant / Milvus / Weaviate)	Selbst gehosteter Vektorspeicher. Daten verlassen nie Ihre Infrastruktur.

Audio & Sprache

Source Service	Xinity Equivalent	Migration Notes
Whisper API	Xinity Runtime (Whisper large-v3 lokal)	Identisches Whisper-Modell lokal. Unbegrenzte Transkription ohne Minuten-Abrechnung. 99 Sprachen.

Source Service	Xinity Equivalent	Migration Notes
TTS (Text-to-Speech)	Xinity Runtime (Bark / XTTS-v2 / Piper)	On-Premise Sprachsynthese. Keine nutzungsbasierte Abrechnung.

Plattform & DevOps

Source Service	Xinity Equivalent	Migration Notes
OpenAI API Keys / Org	Xinity Admin Console (RBAC, SSO, Audit Logs)	Enterprise Identity-Integration. Granulare rollenbasierte Zugriffskontrolle. Vollstaendiger Audit-Trail.
Rate Limits & Quotas	Xinity Resource Manager	Keine externen Rate Limits. GPU-Ressourcen nach Team zuordnen.
Usage Dashboard	Xinity Monitoring (Prometheus / Grafana)	Echtzeit GPU-Nutzung und Latenz. Keine Token-Abrechnung.

3. MIGRATIONSPROZESS

3.1 Bestandsaufnahme & Discovery

OpenAI API-Nutzung auditieren

Exportieren Sie Ihr OpenAI-Usage-Dashboard, um alle aktiven Endpunkte zu identifizieren. Katalogisieren Sie jede Anwendung und jeden Workflow, der die OpenAI API aufruft. Dokumentieren Sie Anfragevolumen, Spitzenlasten und Latenzanforderungen.

Workload-Sensitivitaet klassifizieren

Ordnen Sie jeden Workload einer Datenklassifizierung zu: oeffentlich, intern, vertraulich oder reguliert. Identifizieren Sie souveraenitaetsblockierte Workloads (muessen migriert werden) gegenueber souveraenitaetsbevorzugten (sollten migriert werden).

Performance-Baseline erstellen

Erfassen Sie aktuelle OpenAI API-Latenz (p50, p95, p99), Durchsatz und Fehlerraten. Diese werden zu Ihren Migrations-Erfolgskriterien.

3.2 Infrastrukturplanung & Design

Hardware-Dimensionierung

Arbeiten Sie mit Xinitys Solutions-Team zusammen, um Ihre On-Premise-Bereitstellung zu dimensionieren. ASUS Ascent GX10 Server bieten die noetige Rechendichte fuer Enterprise-Inferenz.

Netzwerkarchitektur

Entwerfen Sie die Netzwerktopologie fuer die Xinity Runtime Integration. Planen Sie Load Balancing, TLS-Terminierung und Firewall-Regeln.

Modellauswahl

Waehlen Sie Open-Weight-Modelle, die Ihren aktuellen OpenAI-Modell-Faehigkeiten entsprechen.

3.3 Pilot-Migration

Xinity Runtime bereitstellen

Xinitys Deployment-Team installiert und konfiguriert die Runtime auf Ihrer Hardware. Typische Bereitstellung: 2-5 Werktage.

Parallelbetrieb

Fuehren Sie Ihren Pilot-Workload 2-4 Wochen parallel gegen OpenAI und Xinity aus. Vergleichen Sie Ausgabequalitaet, Latenz und Durchsatz.

Anwendungscode-Migration

Fuer OpenAI SDK-Nutzer ist die Migration typischerweise eine 3-Zeilen-Aenderung: client = OpenAI(base_url='https://your-domain.com/v1', api_key='your-xinity-key')

3.4 Vollstaendige Migration & Optimierung

Phasenweiser Rollout

Migrieren Sie Workloads nach Prioritaet: regulierte Workloads zuerst, dann kostenintensive, dann verbleibende.

Optimierung & Dekommissionierung

Optimieren Sie Modellquantisierung, Batch-Groessen und KV-Cache. Nach Validierung: OpenAI API-Keys widerrufen und Abrechnungskonten schliessen.

4. MIGRATIONS-WERKZEUGE & BESCHLEUNIGER

4.1 API-Kompatibilitaetsschicht

Xinity Runtime API Gateway

Drop-in-Ersatz fuer OpenAIs API-Endpunkt. Unterstuetzt alle Standard-OpenAI-SDK-Methoden.

SDK-Migrations-Scanner

Automatisiertes Tool, das Ihre Codebasis nach OpenAI-SDK-Aufrufen scannt und einen Migrationsbericht generiert.

4.2 Observability & Betrieb

Xinity Dashboard

Vorkonfigurierte Grafana-Dashboards: GPU-Nutzung, Inferenz-Latenz, Durchsatz und Kostenvergleich zur Cloud.

Audit & Compliance-Modul

Vollstaendiger Audit-Trail. Compliance-Berichte fuer DSGVO, ISO 27001 und branchenspezifische Frameworks.

5. NAECHSTE SCHRITTE: STARTEN SIE IHRE MIGRATION MIT XINITY

Die Migration von OpenAIs Cloud-API zu Xinitys On-Premise-Plattform ist dank der vollstaendigen OpenAI API-Kompatibilitaet der einfachste Migrationspfad.

So starten Sie:

1. Discovery-Gespraech vereinbaren -- Xinitys Solutions-Architekten analysieren Ihre aktuelle OpenAI-Nutzung und erstellen einen TCO-Vergleich.

2. Proof of Concept anfordern -- Testen Sie Xinity Runtime mit Ihrem kritischsten Workload in Ihrer eigenen Umgebung.

3. Phasenweise Migration planen -- Souveraenitaetsblockierte Workloads zuerst.

4. Go Live mit Vertrauen -- Xinity bietet laufenden Support, Modell-Updates und Performance-Optimierung.

Kontakt: Web: xinity.ai E-Mail: contact@xinity.ai Standort: Wien, Oesterreich