XINITY

xinity.ai



# VON ANTHROPIC CLAUDE 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 Anthropics Claude API-Dienste (Claude Opus 4.6, Claude Sonnet 4.6, Claude Haiku 4.5, Messages API, Tool Use) nutzen und KI-Workloads auf eine souveraene On-Premise-Infrastruktur migrieren muessen. Ob Sie Claude wegen seiner Sicherheitseigenschaften gewaehlt haben und nun regulatorische Anforderungen vollstaendige Datensouveraenitaet verlangen, oder ob Sie die Cloud-API-Abhaengigkeit fuer Geschaeftskontinuitaet eliminieren moechten -- dieses Whitepaper liefert die technischen Zuordnungen und Migrationsprozesse.

# CONTENTS

# 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 ANTHROPIC-STACK, NEU AUFGEBAUT AUF XINITY (ZUGEORDNET & BEREIT)

Dieser Abschnitt stellt ein Faehigkeiten-Mapping fuer die Migration von Anthropics Claude API zu Xinitys On-Premise-Plattform bereit. Anthropic verwendet ein proprietaeres API-Format (Messages API), das sich vom OpenAI-Format unterscheidet. Xinity stellt einen OpenAI-kompatiblen Endpunkt bereit, sodass die Migration SDK- und Prompt-Format-Uebersetzung umfasst.

## Kern-Inferenz & Konversation

| Source Service | Xinity Equivalent | Migration Notes |
|---|---|---|
| **Claude Sonnet 4.6 (Neuestes Balanciert)** | Xinity Runtime (Mistral Medium 3 / Qwen3.5 32B) | Ausgewogene Leistung und Geschwindigkeit. OpenAI-kompatibler Endpunkt. Exzellent fuer Code, Analyse, Text. |
| **Claude Opus 4.6 (Leistungsstaerkstes)** | Xinity Runtime (Mistral Large 3 / Nemotron-Ultra-253B) | Maximale Faehigkeit fuer komplexe Aufgaben. Keine Token-basierte Abrechnung. |
| **Claude Haiku 4.5 (Schnellstes)** | Xinity Runtime (Qwen3.5 8B / Mistral Small 3) | Ultra-schnelle Inferenz. Ideal fuer Klassifizierung und Extraktion. |

## Tool Use & Funktionsaufrufe

| Source Service | Xinity Equivalent | Migration Notes |
|---|---|---|
| **Claude Tool Use** | Xinity Runtime (OpenAI-Format Tool Calling) | Anthropic-Tool-Schemas ins OpenAI-Format uebersetzen. Lokal ausfuehren. |
| **Claude Computer Use** | Xinity + Open-Source Agents | On-Premise Computer-Interaktions-Agenten. Keine Screenshots an Cloud-APIs. |
| **Claude MCP** | Xinity + Lokaler MCP Server | MCP-Server On-Premise bereitstellen. Interne Datenbanken und APIs verbinden. |

## Dokumentenverarbeitung & Analyse

| Source Service | Xinity Equivalent | Migration Notes |
|---|---|---|
| **Claude PDF-Analyse** | Xinity Runtime + PDF-Vorverarbeitungspipeline | Lokale Dokumentenaufnahme. Vertrauliche Dokumente sicher verarbeiten. |
| **Claude Vision** | Xinity Runtime (LLaVA / CogVLM / Qwen-VL) | On-Premise Bildverstaendnis. Medizinische Bilder lokal verarbeiten. |
| **Anthropic Prompt Caching** | Xinity Runtime (KV-Cache + Prompt-Prefix-Caching) | Lokales Kontext-Caching. Keine Cloud-Caching-Gebuehren. |

## Sicherheit & Guardrails

| Source Service | Xinity Equivalent | Migration Notes |
|---|---|---|
| **Claude Constitutional AI** | Xinity Runtime<br>+ Guardrails (NeMo / LLM Guard) | Konfigurierbare Sicherheitsebenen.<br>Guardrails nach Anwendungsfall anpassen. |
| **Anthropic Content Policy** | Xinity Custom Policy Engine | Eigene Inhaltsrichtlinien definieren.<br>Volle Kontrolle ueber Modellverhalten. |

## Plattform & Betrieb

| Source Service | Xinity Equivalent | Migration Notes |
|---|---|---|
| **Anthropic Console** | Xinity Admin Console<br>(RBAC, SSO, Audit Logs) | Enterprise Identity-Integration.<br>Vollstaendiger Audit-Trail On-Premise. |
| **Anthropic Usage API** | Xinity Monitoring<br>(Prometheus / Grafana) | Echtzeit-Inferenz-Metriken.<br>Keine Token-Abrechnung. |
| **Batches API** | Xinity Batch Processor | On-Premise Batch-Inferenz.<br>Prioritaets-Warteschlange. |

# 3. MIGRATIONSPROZESS

## 3.1 Bestandsaufnahme & Discovery

### Anthropic API-Nutzung auditieren

Ueberpruefen Sie Ihre Anthropic Console, um alle aktiven Workloads zu katalogisieren. Dokumentieren Sie, welches Claude-Modell jeder Workload nutzt und welche spezifischen Faehigkeiten benoetigt werden.

### Anthropic-spezifische Features kartieren

Identifizieren Sie Workloads, die Anthropic-spezifische Features nutzen: Constitutional AI, XML-Tag-Konventionen, Messages API Schema.

### Datensensitivitaet klassifizieren

Identifizieren Sie Workloads, bei denen Claudes Cloud-API Compliance-Risiken unter DSGVO oder nationalen Datenschutzgesetzen erzeugt.

## 3.2 Infrastrukturplanung

### Hardware-Dimensionierung

Claude Opus 4.6-Nutzer benoetigen typischerweise 253B-Parameter-Modelle (Nemotron-Ultra) auf Xinity. Claude Haiku 4.5-Nutzer koennen effiziente 8B Modelle nutzen.

### API-Uebersetzungsstrategie

Planen Sie die Uebersetzung von Anthropics Messages API zu OpenAIs Chat Completions API. Xinity bietet eine Uebersetzungs-Middleware.

### Prompt-Engineering-Anpassung

Claude-optimierte Prompts muessen getestet und angepasst werden. Planen Sie 1-2 Wochen fuer Prompt-Tests ein.

## 3.3 Pilot-Migration

### Xinity Runtime bereitstellen

Installation und Konfiguration mit den abgestimmten Open-Weight-Modellen.

### SDK-Migration

Von Anthropics SDK zum OpenAI SDK wechseln, das auf Xinity zeigt: from openai import OpenAI client = OpenAI( base_url='https://your-domain.com/v1', api_key='your-xinity-key' )

### Qualitaetsvalidierung

Validieren Sie, dass Xinitys Modelle aequivalente Qualitaet fuer Ihre spezifischen Aufgaben liefern.

## 3.4 Vollstaendige Migration

### Phasenweiser Rollout

Souveraenitaetsblockierte Anwendungsfaelle zuerst, dann Hochvolumen-Workloads, dann verbleibende.

### Guardrails-Konfiguration

Claudes Constitutional AI durch Xinitys konfigurierbare Guardrails ersetzen. Vorteil: Sie kontrollieren, was gefiltert wird.

### Anthropic-Dienste dekommissionieren

API-Keys widerrufen, Abrechnungskonten schliessen. 90 Tage Rollback-Faehigkeit beibehalten.

# 4. MIGRATIONS-WERKZEUGE & BESCHLEUNIGER

## 4.1 API-Uebersetzung

### Anthropic-zu-OpenAI Uebersetzungs-Middleware

Transparenter Proxy, der Anthropic Messages API Format ins OpenAI Chat Completions Format konvertiert. Null-Code-Migration.

### Prompt-Migrations-Toolkit

Analysiert Claude-optimierte Prompts und passt sie fuer Open-Weight-Modelle an.

## 4.2 Observability & Betrieb

### Xinity Dashboard

Umfassendes Monitoring mit vorkonfigurierten Dashboards.

### Compliance & Audit-Modul

Vollstaendiger Audit-Trail mit Compliance-Berichtgenerierung. DSGVO, ISO 27001.

# 5. NAECHSTE SCHRITTE: STARTEN SIE IHRE MIGRATION MIT XINITY

Die Migration von Anthropic Claude zu Xinity erfordert API-Format-Uebersetzung und Prompt-Anpassung, aber der Prozess ist klar definiert und Xinity bietet automatisierte Werkzeuge.

So starten Sie:

1. Discovery-Gespraech vereinbaren -- Xinitys Solutions-Architekten analysieren Ihre Claude-API-Nutzung und identifizieren optimale Open-Weight-Modell-Matches.

2. Proof of Concept anfordern -- Testen Sie Xinity mit der Anthropic-Uebersetzungs-Middleware. Null Code-Aenderungen noetig.

3. Prompt-Anpassung planen -- Optimieren Sie Prompts fuer Open-Weight-Modelle.

4. Go Live mit voller Kontrolle -- Besitzen Sie Ihre KI-Infrastruktur, Ihre Daten und Ihre Inhaltsrichtlinien.

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