XINITY

xinity.ai

FROM ANTHROPIC CLAUDE TO XINITY

AI PLATFORM MIGRATION
WHITEPAPER SERIES 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

This guide is intended for engineering teams, CTOs, and IT decision-makers currently using Anthropic's Claude API services (Claude Opus 4.6, Claude Sonnet 4.6, Claude Haiku 4.5, the Messages API, and tool use capabilities) who need to transition AI workloads to a sovereign, on-premise infrastructure. Whether you chose Claude for its safety characteristics and now face regulatory requirements demanding full data sovereignty, or you are seeking to eliminate cloud API dependency for business continuity, this whitepaper provides the technical mappings, migration processes, and tooling guidance to execute with confidence.

CONTENTS

1. ENTERPRISE AI WITHOUT COMPROMISE: WHY XINITY BECOMES THE BETTER FIT

If your organization runs AI workloads in production, migrating from cloud-hosted AI APIs to Xinity's on-premise platform delivers something no cloud provider can: complete architectural sovereignty over your data, models, and inference infrastructure. This is not just a vendor switch -- it is a fundamental shift from renting AI capacity to owning it.

-- Architectural sovereignty, not policy promises

Cloud AI providers offer contractual data protection through terms of service and data processing agreements. Xinity delivers architectural sovereignty: your data never leaves hardware you physically own and control. For regulated industries -- healthcare, legal, financial services, media, and manufacturing -- this distinction is not academic. It is the difference between compliance risk and compliance certainty. No foreign government subpoena, no cloud provider policy change, and no geopolitical shift can affect data that exists solely on your premises.

-- Predictable economics at enterprise scale

Cloud AI pricing scales with consumption: every API call, every token, every GPU-hour is metered and billed. Xinity's on-premise model transforms variable OPEX into predictable CAPEX. Customers deploying Xinity Runtime on ASUS Ascent GX10 servers report approximately 80% cost savings compared to equivalent cloud capacity. At scale, this means paying roughly EUR 320/year in electricity versus EUR 18,600/year for comparable cloud compute. The economics become more favorable as usage increases -- the opposite of cloud pricing.

-- Zero-latency inference for critical applications

On-premise AI eliminates network round-trips to distant cloud regions. For latency-sensitive applications -- real-time document analysis, production-line quality inspection, clinical decision support -- local inference delivers consistent sub-millisecond response times without dependency on internet connectivity, cloud region availability, or cross-border data transfer regulations.

-- Regulatory tailwinds accelerating adoption

The EU Digital Networks Act (proposed January 2026) with compliance deadlines in August 2026, the EUR 20 billion InvestAI funding initiative, and emerging 'Buy European' procurement rules all validate the sovereign AI infrastructure thesis. Organizations migrating to on-premise AI now position themselves ahead of regulations rather than scrambling to comply later.

-- OpenAI-compatible APIs -- migrate without rewriting

Xinity Runtime exposes OpenAI-compatible API endpoints. This means your existing application code, SDKs, prompt libraries, and orchestration frameworks continue to work with minimal modification. You change the base URL and API key; your applications do not notice the difference.

2. YOUR ANTHROPIC STACK, REBUILT ON XINITY (MAPPED & READY)

This section establishes a clear capability-mapping framework for organizations migrating from Anthropic's Claude API to Xinity's on-premise platform. The goal is to help you translate every Claude capability you currently rely on -- long-context conversations, tool use, document analysis, and structured outputs -- into functionally equivalent or superior Xinity capabilities.

Anthropic uses a proprietary API format (Messages API) that differs from OpenAI's format. Xinity provides an OpenAI-compatible endpoint, so migration involves SDK and prompt format translation. This guide covers both the API translation and the capability mapping.

Core Inference & Conversation

Source Service	Xinity Equivalent	Migration Notes
Claude Sonnet 4.6 (Latest balanced)	Xinity Runtime (Mistral Medium 3 / Qwen3.5 32B)	Balanced performance and speed. OpenAI-compatible /v1/chat/completions. Excellent for coding, analysis, writing.
Claude Opus 4.6 (Most capable)	Xinity Runtime (Mistral Large 3 / Nemotron-Ultra-253B)	Maximum capability for complex tasks. On-premise inference at full precision. No per-token pricing.
Claude Haiku 4.5 (Fastest)	Xinity Runtime (Qwen3.5 8B / Mistral Small 3)	Ultra-fast inference for high-throughput. Ideal for classification, routing, extraction. Runs efficiently on smaller GPUs.
Claude Opus/Sonnet 4.6 (1M token context)	Xinity Runtime (Mistral Large 3 / Nemotron-Ultra 128K context)	Extended context for document analysis. For 1M: RAG pipeline approach. No context window surcharges.

Tool Use & Function Calling

Source Service	Xinity Equivalent	Migration Notes
Claude Tool Use (Function Calling)	Xinity Runtime (OpenAI-format tool calling)	Translate Anthropic tool schemas to OpenAI format. Functionally equivalent capability. Local tool execution without cloud roundtrips.
Claude Computer Use (Beta)	Xinity + Open-Source Agents (Browser Use / Playwright)	On-premise computer interaction agents. No screenshots sent to cloud APIs. Full control over agent permissions.
Claude MCP (Model Context Protocol)	Xinity + Local MCP Server	Deploy MCP servers on-premise. Connect to internal databases, APIs, tools. No external data exposure.

Document Processing & Analysis

Source Service	Xinity Equivalent	Migration Notes
Claude PDF Analysis (Direct upload)	Xinity Runtime + PDF preprocessing pipeline	Local document ingestion and analysis. PDF parsing + OCR + chunking + LLM. Process confidential documents safely.
Claude Vision (Image analysis)	Xinity Runtime (LLaVA / CogVLM / Qwen-VL)	On-premise image understanding. Medical images, legal documents, manufacturing QC -- all processed locally.
Anthropic Prompt Caching	Xinity Runtime (KV-cache + prompt prefix caching)	Local context caching for repeated prompts. No cloud caching fees. Faster response for repeated queries.

Safety, Alignment & Guardrails

Source Service	Xinity Equivalent	Migration Notes
Claude Constitutional AI (Built-in safety)	Xinity Runtime + Guardrails (NeMo / LLM Guard)	Configurable safety layers. Customize guardrails for your use case. No over-filtering of legitimate queries.
Anthropic Content Policy	Xinity Custom Policy Engine	Define your own content policies. No external provider blocking valid requests. Full control over model behavior boundaries.
Claude System Prompts	Xinity System Prompt Manager	Centralized system prompt management. Version control and A/B testing. No system prompt inspection by provider.

Platform & Operations

Source Service	Xinity Equivalent	Migration Notes
Anthropic Console (API Key Management)	Xinity Admin Console (RBAC, SSO, Audit Logs)	Enterprise identity integration. Granular access control per model. Complete audit trail on-premise.
Anthropic Usage API	Xinity Monitoring (Prometheus / Grafana)	Real-time inference metrics. GPU utilization, latency, throughput. No per-token billing.
Anthropic Workbench	Xinity Playground + JupyterHub	On-premise prompt engineering environment. Test models with real data safely. No data sent to external APIs.
Batches API (Async processing)	Xinity Batch Processor	On-premise batch inference. Process large datasets overnight. Priority queuing and scheduling.

3. MIGRATION PROCESS

3.1 Assessment & Discovery

Audit Anthropic API Usage

Review your Anthropic Console usage to catalog all active workloads: Messages API calls, tool use patterns, document uploads, and batch processing jobs. Document which Claude model tier each workload uses (Opus 4.6, Sonnet 4.6, Haiku 4.5) and the specific capabilities relied upon (1M context, extended thinking, tool use, vision).

Map Anthropic-Specific Features

Identify workloads using Anthropic-specific features: Constitutional AI behaviors, Claude's particular response formatting, XML tag conventions in prompts, and the Messages API schema. These require explicit translation during migration.

Classify Data Sensitivity

Map workloads to sovereignty requirements. Claude's cloud API processes all data on Anthropic's infrastructure (primarily US-based). Identify every workload where this creates compliance risk under GDPR, national data residency laws, or industry regulations.

3.2 Infrastructure Planning & Design

Hardware Sizing

Size your Xinity deployment to handle your Claude workload profile. Key consideration: Claude Opus 4.6 users typically need 253B-parameter models (Nemotron-Ultra) on Xinity, requiring multi-GPU configurations. Claude Haiku 4.5 users can often use efficient 8B models on minimal GPU setups.

API Translation Strategy

Plan the Anthropic Messages API to OpenAI Chat Completions API translation. Key differences to address: message format (Anthropic uses role/content pairs with different schema), tool use format (different JSON schema), and system prompt handling (Anthropic uses a separate system parameter). Xinity provides a translation middleware for automated conversion.

Prompt Engineering Adaptation

Claude-optimized prompts may use Anthropic-specific conventions (XML tags, particular instruction formatting). While these often work with other models, testing and adaptation ensures optimal performance. Budget 1-2 weeks for prompt testing and optimization with Xinity's open-weight models.

3.3 Pilot Migration

Deploy Xinity Runtime

Install and configure Xinity Runtime with the open-weight models matched to your Claude model usage. Enable the Anthropic-to-OpenAI translation middleware if needed for legacy integration support.

SDK Migration

Migrate from Anthropic's SDK to OpenAI SDK pointing at Xinity:

```
# Before (Anthropic Claude)
import anthropic
client = anthropic.Anthropic(api_key='sk-ant-...')
message = client.messages.create(
    model='claude-sonnet-4-6',
    messages=[{'role':'user','content':'...'}]
)
```

```
# After (Xinity)
from openai import OpenAI
client = OpenAI(
    base_url='https://your-domain.com/v1',
    api_key='your-xinity-key'
)
response = client.chat.completions.create(
    model='qwen3.5-32b',
    messages=[{'role':'user','content':'...'}]
)
```

Quality Validation

Claude's outputs have distinctive characteristics (structured thinking, nuanced safety responses). Validate that Xinity's models produce equivalent quality for your specific tasks. Focus on: accuracy, response format compliance, tool use reliability, and edge case handling.

3.4 Full-Scale Migration & Optimization

Phased Rollout

Migrate workloads starting with sovereignty-blocked use cases, then high-volume workloads benefiting from fixed-cost pricing, then remaining workloads. Each phase: deploy, parallel-run, validate quality, cut over.

Guardrails Configuration

Replace Claude's built-in Constitutional AI with Xinity's configurable guardrails (NVIDIA NeMo Guardrails or LLM Guard). Advantage: you control what gets filtered. No more valid queries blocked by a cloud provider's safety policy.

Decommission Anthropic Services

Revoke Anthropic API keys, close billing accounts, and archive Anthropic Workbench configurations. Maintain rollback capability for 90 days post-migration.

4. MIGRATION TOOLS & ACCELERATORS

4.1 API Translation

Anthropic-to-OpenAI Translation Middleware

Transparent proxy that converts Anthropic Messages API format to OpenAI Chat Completions format. Handles: message schema translation, tool use format conversion, system prompt extraction, and streaming response adaptation. Deploy between your application and Xinity Runtime for zero-code migration.

Prompt Migration Toolkit

Automated tool that analyzes Claude-optimized prompts and adapts them for open-weight models. Identifies Anthropic-specific conventions, suggests equivalent formulations, and runs A/B tests to validate output quality.

4.2 Safety & Guardrails

Guardrails Configuration Framework

Pre-built guardrail profiles that approximate Claude's safety behaviors using NVIDIA NeMo Guardrails or LLM Guard. Customizable per use case: strictest for customer-facing, permissive for internal research, industry-specific for regulated environments.

Content Policy Engine

Define and enforce your own content policies without external provider restrictions. Create allowlists, blocklists, and context-aware rules that match your organization's specific needs rather than a cloud provider's one-size-fits-all policy.

4.3 Observability & Operations

Xinity Dashboard

Comprehensive monitoring with pre-built dashboards: inference performance, model health, guardrail trigger rates, and cost tracking vs. Anthropic cloud baseline.

Compliance & Audit Module

Full audit trail with compliance report generation. Demonstrates complete data sovereignty for regulators. Supports GDPR, ISO 27001, and sector-specific frameworks.

5. NEXT STEPS: START YOUR MIGRATION WITH XINITY

Migrating from Anthropic Claude to Xinity requires API format translation and prompt adaptation, but the process is well-defined and Xinity provides automated tooling for the most complex translation tasks.

Here is how to get started:

1. Schedule a Discovery Call -- Xinity's solutions architects will analyze your Claude API usage patterns, identify the optimal open-weight model matches for each workload, and provide a detailed migration timeline.

2. Request a Proof of Concept -- Deploy Xinity Runtime with the Anthropic translation middleware. Test your most critical workload with zero code changes to validate output quality and performance.

3. Plan Your Prompt Adaptation -- Work with Xinity's AI team to optimize prompts for open-weight models, ensuring equivalent or better output quality.

4. Go Live with Full Control -- Own your AI infrastructure, your data, and your content policies. No external provider can change the rules.

Contact Xinity: Web: xinity.ai Email: contact@xinity.ai Location: Vienna, Austria