XINITY

xinity.ai

FROM OPENAI 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 OpenAI's API services (GPT-5.4, GPT-4.1, o3, o4-mini, Whisper, Embeddings, Assistants API) who need to transition AI workloads to a sovereign, on-premise infrastructure. Whether you are planning a phased migration to reduce cloud dependency or a complete platform switch driven by regulatory requirements, 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 OPENAI STACK, REBUILT ON XINITY (MAPPED & READY)

This section establishes a clear capability-mapping framework for organizations migrating from OpenAI's cloud API to Xinity's on-premise platform. The goal is to help you translate every OpenAI service you currently rely on -- chat completions, embeddings, audio transcription, fine-tuning, and orchestration -- into functionally equivalent or superior Xinity capabilities.

Because Xinity Runtime exposes OpenAI-compatible API endpoints, many migrations are as simple as changing the base URL. For services with deeper architectural differences, we provide detailed migration paths and alternative approaches.

Core Inference & Chat Completions

Source Service	Xinity Equivalent	Migration Notes
GPT-5.4 / GPT-5.4-mini (Flagship)	Xinity Runtime (Mistral Large 3 / Nemotron-Ultra, or customer-selected model)	OpenAI-compatible /v1/chat/completions endpoint. Change base_url and api_key only. No application code changes required.
GPT-4.1 (1M token context, coding)	Xinity Runtime (Mistral Large 3 / Nemotron-Ultra with 128K context)	Long-context coding and instruction following. Local inference eliminates token-based pricing. For 1M+: chunking + RAG pipeline.
GPT-5.4-nano / GPT-4.1-mini	Xinity Runtime (Qwen3.5 8B / Mistral Small 3)	Cost-optimized small models for high-throughput. Run on CPU or entry-level GPU.
o3 / o3-pro (Reasoning)	Xinity Runtime (Qwen3.5 72B-Reasoning / Qwen3.5 72B)	On-premise reasoning models with chain-of-thought. No per-token reasoning surcharge.
o4-mini (Cost-effective reasoning)	Xinity Runtime (Qwen3.5 8B / Mistral Small 3 with reasoning)	80% cheaper than o3 with strong STEM perf. Local reasoning at fixed cost.

Embeddings & Vector Search

Source Service	Xinity Equivalent	Migration Notes
text-embedding-3-small/large	Xinity Runtime (BGE-M3, E5-Mistral, or Nomic-Embed)	Local embedding generation. No per-token embedding fees. Integrates with local vector DBs (Qdrant, Milvus, pgvector).
OpenAI Vector Store (Assistants API)	On-Prem Vector DB (Qdrant / Milvus / Weaviate)	Self-hosted vector storage. Data never leaves your infrastructure. Full control over indexing and retrieval.

Audio & Speech

Source Service	Xinity Equivalent	Migration Notes
Whisper API	Xinity Runtime (Whisper large-v3 local)	Identical Whisper model running locally. Unlimited transcription without per-minute billing. Supports 99 languages.
TTS (Text-to-Speech)	Xinity Runtime (Bark / XTTS-v2 / Piper)	On-premise speech synthesis. No usage-based pricing. Custom voice cloning available.

Fine-Tuning & Model Customization

Source Service	Xinity Equivalent	Migration Notes
OpenAI Fine-Tuning API	Xinity Training Module (LoRA / QLoRA on-prem)	Full fine-tuning on your own hardware. No data upload to third parties. Supports LoRA, QLoRA, full fine-tune. Training data stays 100% on-premise.
Custom GPTs / Assistants	Xinity Orchestration (LangChain / LlamaIndex local)	Build custom AI assistants with local tool calling. RAG pipelines on proprietary data. No data shared with model providers.

Image Generation & Vision

Source Service	Xinity Equivalent	Migration Notes
GPT-Image-1 / DALL-E 3	Xinity Runtime (Stable Diffusion XL / Flux)	On-premise image generation. No content policy restrictions on your own data. Unlimited generations at fixed cost.
GPT-5.4 Vision (Native multimodal)	Xinity Runtime (LLaVA / CogVLM / Qwen-VL)	Multimodal vision-language models. Process sensitive images without cloud upload. Ideal for medical imaging, document analysis.

Platform & DevOps

Source Service	Xinity Equivalent	Migration Notes
OpenAI API Keys / Org Management	Xinity Admin Console (RBAC, SSO, Audit Logs)	Enterprise identity integration (LDAP/SAML). Granular role-based access control. Complete audit trail on-premise.
Rate Limits & Quotas	Xinity Resource Manager	No external rate limits. Allocate GPU resources by team/project. Priority queuing for critical workloads.
Usage Dashboard & Billing	Xinity Monitoring (Prometheus / Grafana)	Real-time GPU utilization, inference latency, and throughput metrics. No per-token billing -- flat infrastructure cost.

3. MIGRATION PROCESS

3.1 Assessment & Discovery

Audit OpenAI API Usage

Export your OpenAI usage dashboard to identify all active endpoints (chat, embeddings, whisper, fine-tuning). Catalog every application, microservice, and workflow that calls the OpenAI API. Document request volumes, peak loads, average token counts, and latency requirements per endpoint.

Classify Workload Sensitivity

Map each workload to a data classification level: public, internal, confidential, or regulated. Identify which workloads are sovereignty-blocked (must move) versus sovereignty-preferred (should move). Regulated workloads under GDPR, banking secrecy, attorney-client privilege, or healthcare data laws are migration priorities.

Baseline Performance Metrics

Record current OpenAI API latency (p50, p95, p99), throughput (requests/second), and error rates. These become your migration success criteria. Xinity on-premise should match or exceed these baselines for each workload.

3.2 Infrastructure Planning & Design

Hardware Sizing

Work with Xinity's solutions team to size your on-premise deployment. Key inputs: concurrent users, peak inference requests/second, model sizes (8B, 72B, 253B parameters), and context window requirements. Typical deployments use ASUS Ascent GX10 servers with NVIDIA GPUs, providing the compute density needed for enterprise-scale inference.

Network Architecture

Design the network topology for Xinity Runtime integration. Xinity exposes standard REST/gRPC endpoints on your local network. Plan for: load balancing across multiple inference nodes, TLS termination, DNS resolution, and firewall rules allowing internal application access while blocking external traffic.

Model Selection & Licensing

Select open-weight models that match your current OpenAI model capabilities. Xinity supports all major open-weight model families. For each OpenAI model you currently use, identify the best-fit open-weight alternative based on benchmark performance, context window, and language support requirements.

3.3 Pilot Migration

Deploy Xinity Runtime

Xinity's deployment team installs and configures the Runtime on your hardware. Initial setup includes: OS provisioning, GPU driver installation, model downloading, API endpoint configuration, and integration testing. Typical deployment takes 2-5 business days from hardware delivery to first inference.

Parallel Running

Run your pilot workload against both OpenAI and Xinity simultaneously for 2-4 weeks. Compare: output quality (human evaluation + automated metrics), latency, throughput, and cost. Use Xinity's monitoring dashboard to track GPU utilization

and identify optimization opportunities.

Application Code Migration

For OpenAI SDK users, migration is typically a 3-line change:

```
# Before (OpenAI Cloud)
client = OpenAI(api_key='sk-...')
```

```
# After (Xinity On-Premise)
client = OpenAI(
    base_url='https://your-domain.com/v1',
    api_key='your-xinity-key'
)
```

All subsequent API calls remain identical.

3.4 Full-Scale Migration

Phased Rollout

Migrate workloads in priority order: regulated/sovereignty-blocked workloads first, then high-volume cost-sensitive workloads, then remaining workloads. Each phase follows the pilot pattern: deploy, parallel-run, validate, cut over.

Data Pipeline Migration

Migrate RAG pipelines, vector databases, and fine-tuning datasets to on-premise infrastructure. Xinity integrates with on-premise vector databases (Qdrant, Milvus, pgvector) and supports local fine-tuning with LoRA/QLoRA without data ever leaving your environment.

Monitoring & Alerting

Deploy comprehensive observability: Prometheus for metrics collection, Grafana for dashboards, and alerting for GPU utilization, inference latency, model health, and queue depth. Xinity provides pre-built dashboard templates for common monitoring scenarios.

3.5 Optimization & Decommission

Performance Tuning

Optimize model quantization (INT8, INT4), batch sizes, and KV-cache settings for your specific workload patterns. Xinity's runtime supports dynamic batching and continuous batching to maximize GPU utilization. Typical optimizations yield 30-50% additional throughput.

OpenAI API Decommission

After validating full migration success, revoke OpenAI API keys, close billing accounts, and update documentation. Maintain a rollback plan for 90 days post-migration, after which legacy cloud configurations can be fully archived.

4. MIGRATION TOOLS & ACCELERATORS

4.1 API Compatibility Layer

Xinity Runtime API Gateway

Drop-in replacement for OpenAI's API endpoint. Supports /v1/chat/completions, /v1/embeddings, /v1/audio/transcriptions, and all standard OpenAI SDK methods. Handles authentication, rate limiting, and request routing to local GPU clusters.

SDK Migration Scanner

Automated tool that scans your codebase for OpenAI SDK calls and generates a migration report: compatible endpoints (green), endpoints requiring model substitution (yellow), and endpoints requiring architectural changes (red).

4.2 Model Management

Xinity Model Registry

Centralized model lifecycle management. Download, version, deploy, and rollback open-weight models across your inference cluster. Supports automated model updates with canary deployments and A/B testing.

Fine-Tuning Pipeline

End-to-end on-premise fine-tuning with LoRA and QLoRA. Includes data preprocessing, training orchestration, evaluation benchmarks, and one-click deployment to production inference endpoints.

4.3 Observability & Operations

Xinity Dashboard

Pre-configured Grafana dashboards showing: GPU utilization per model, inference latency percentiles, request throughput, queue depth, model-level error rates, and cost-per-inference tracking vs. cloud baseline.

Audit & Compliance Module

Complete audit trail of every inference request: who, when, which model, input/output token counts (without logging actual content). Generates compliance reports for GDPR, ISO 27001, and industry-specific regulatory frameworks.

5. NEXT STEPS: START YOUR MIGRATION WITH XINITY

Moving from OpenAI's cloud API to Xinity's on-premise platform is the most straightforward migration path in this whitepaper series, thanks to Xinity Runtime's full OpenAI API compatibility.

Here is how to get started:

1. Schedule a Discovery Call -- Xinity's solutions architects will analyze your current OpenAI usage, estimate hardware requirements, and provide a TCO comparison showing projected savings over 12, 24, and 36 months.

2. Request a Proof of Concept -- Deploy Xinity Runtime on a single server with your most critical workload. Validate performance, output quality, and integration compatibility in your own environment with your own data.

3. Plan Your Phased Migration -- Work with Xinity's migration team to build a prioritized rollout plan. Sovereignty-blocked workloads move first; cost-optimization workloads follow.

4. Go Live with Confidence -- Xinity provides ongoing support, model updates, and performance optimization as part of the platform subscription.

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