XINITY

xinity.ai

FROM GOOGLE GEMINI 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 Google's Gemini AI services (Gemini 3.1 Pro, Gemini 2.5 Pro/Flash, Vertex AI, Google AI Studio) who need to transition AI workloads to a sovereign, on-premise infrastructure. Whether you are evaluating alternatives due to data residency concerns within GCP's ecosystem, regulatory pressure from EU data sovereignty requirements, or strategic desire to reduce dependency on a single hyperscaler, 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 GEMINI STACK, REBUILT ON XINITY (MAPPED & READY)

This section establishes a clear capability-mapping framework for organizations migrating from Google's Gemini ecosystem to Xinity's on-premise platform. The goal is to help you translate every Gemini service you currently depend on -- multimodal inference, embeddings, Vertex AI pipelines, and Google AI Studio workflows -- into functionally equivalent or superior Xinity capabilities.

Google's Gemini ecosystem is deeply integrated with GCP services, so migration requires both API-level translation and architectural decoupling. This guide addresses both dimensions.

Core Inference & Multimodal AI

Source Service	Xinity Equivalent	Migration Notes
Gemini 3.1 Pro (Latest flagship)	Xinity Runtime (Mistral Large 3 / Nemotron-Ultra)	OpenAI-compatible API endpoint. Context windows up to 128K tokens. For 1M+ context: chunking + RAG pipeline.
Gemini 2.5 Pro (1M token context)	Xinity Runtime (Nemotron-Ultra / Qwen3.5 72B)	Complex reasoning and coding. Adaptive thinking capabilities. Local inference at fixed cost.
Gemini 2.5 Flash / Flash-Lite	Xinity Runtime (Qwen3.5 8B / Mistral Small 3)	Fast, cost-optimized inference. Ideal for high-throughput classification and summarization tasks.
Gemini Nano (On-device)	Xinity Edge Runtime (Qwen3.5 3B / Mistral Small 3)	Lightweight models for edge deployment. Runs on CPU-only nodes. Ideal for branch office / factory floor.

Vertex AI Platform Services

Source Service	Xinity Equivalent	Migration Notes
Vertex AI Model Garden	Xinity Model Registry	Curated open-weight model catalog. One-click deployment to inference cluster. Version management and rollback.
Vertex AI Pipelines	Xinity + Kubeflow / MLflow	On-premise ML pipeline orchestration. Training, evaluation, deployment automation. Full pipeline observability.
Vertex AI Feature Store	Xinity + Feast / Hopsworks	Self-hosted feature store. Real-time + batch feature serving. Data stays on-premise.
Vertex AI Workbench	Xinity + JupyterHub	On-premise notebook environment. GPU-accelerated experimentation. Direct access to Xinity Runtime APIs.

Embeddings & Search

Source Service	Xinity Equivalent	Migration Notes
Gemini text-embedding-004	Xinity Runtime (BGE-M3, E5-Mistral)	Local embedding generation. No per-token fees. Multilingual support included.
Vertex AI Vector Search	On-Prem Vector DB (Qdrant / Milvus / Weaviate)	Self-hosted similarity search. Billions of vectors at local latency. Full data sovereignty.
Google Cloud Search	Xinity + Elasticsearch / Typesense	On-premise enterprise search. Combine vector + keyword search. Index proprietary documents locally.

Audio, Vision & Multimodal

Source Service	Xinity Equivalent	Migration Notes
Gemini Vision (Image)	Xinity Runtime (LLaVA / CogVLM / Qwen-VL)	On-premise image understanding. Process sensitive images locally. No cloud upload required.
Cloud Speech-to-Text	Xinity Runtime (Whisper large-v3)	Local speech recognition. 99 languages supported. Unlimited transcription at fixed cost.
Cloud Text-to-Speech	Xinity Runtime (Bark / XTTS-v2)	On-premise speech synthesis. Custom voice models available.
Cloud Vision API	Xinity Runtime (YOLO / SAM / DINOv2)	Local image classification, OCR, object detection. Process at network speed.

Platform, Security & Governance

Source Service	Xinity Equivalent	Migration Notes
Google Cloud IAM	Xinity Admin Console (LDAP / SAML / OIDC)	On-premise identity integration. Granular RBAC per model/endpoint. No cloud IAM dependency.
Vertex AI Model Monitoring	Xinity + Prometheus / Grafana	Real-time model performance tracking. Drift detection, latency alerts. All metrics stay on-premise.
Cloud Audit Logs	Xinity Audit Module	Complete inference audit trail. Compliance reporting for GDPR, ISO 27001. No log data leaves your infrastructure.
Data Loss Prevention (DLP)	Xinity + On-Prem DLP	Data never leaves your network. Architectural DLP by design. No policy-based DLP required.

3. MIGRATION PROCESS

3.1 Assessment & Discovery

Audit Gemini & Vertex AI Usage

Export your GCP billing and usage reports to identify all active Gemini and Vertex AI services. Catalog every application calling Gemini APIs, every Vertex AI pipeline, and every integrated GCP service. Document request volumes, model versions, and GCP service dependencies per workload.

Map GCP Service Dependencies

Gemini workloads often depend on other GCP services (Cloud Storage, BigQuery, Pub/Sub, Cloud Functions). For each dependency, identify the on-premise equivalent and plan the decoupling sequence. Critical path: data storage migration must precede model migration.

Classify Workload Sovereignty

Map each workload to data sovereignty requirements. Identify workloads processing personal data under GDPR, trade secrets, or regulated data. These workloads are migration priorities. Document which workloads have Google-specific SDK dependencies requiring code changes.

3.2 Infrastructure Planning & Design

Hardware Sizing

Size your Xinity deployment based on: current Gemini API usage patterns, peak concurrent requests, model sizes for equivalent open-weight alternatives, and growth projections. Account for the GCP services being replaced on-premise (vector search, feature store, etc.).

GCP Decoupling Architecture

Design the target architecture that replaces GCP-hosted services with on-premise equivalents. Key decisions: where to host vector databases, how to replace Vertex AI Pipelines, and which data stores to migrate. Xinity provides reference architectures for common patterns.

API Translation Layer

While Xinity Runtime is OpenAI-compatible, Gemini SDK calls require a translation step. Options: (a) refactor application code to use OpenAI SDK pointing at Xinity, (b) deploy a lightweight API gateway that translates Gemini API format to OpenAI format. Option (a) is recommended for new applications; option (b) for legacy systems.

3.3 Pilot Migration

Deploy Xinity Runtime

Install Xinity Runtime on your provisioned hardware. Configure API endpoints, authentication, and model loading. Deploy the initial set of open-weight models matched to your Gemini model usage.

SDK Migration

Migrate application code from Google's Gemini SDK to OpenAI-compatible SDK calls pointing at Xinity:

```
# Before (Gemini)
import google.generativeai as genai
model = genai.GenerativeModel('gemini-pro')
response = model.generate_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='mistral-large-3',
    messages=[{'role':'user','content':'...'}]
)
```

Parallel Validation

Run pilot workloads against both Gemini and Xinity for 2-4 weeks. Compare output quality, latency, and throughput. Validate that the Gemini-to-OpenAI API translation produces equivalent results for your specific use cases.

3.4 Full-Scale Migration & Optimization

Phased Workload Migration

Migrate workloads in priority order: sovereignty-blocked first, then GCP-dependent services (Vertex AI pipelines, feature stores), then remaining API consumers.

GCP Service Replacement

Deploy on-premise replacements for GCP services: Qdrant/Milvus for Vector Search, Kubeflow for Vertex AI Pipelines, Feast for Feature Store, JupyterHub for Workbench. Migrate data and configurations to each replacement service.

Decommission GCP Resources

After full validation, terminate GCP Vertex AI endpoints, delete Cloud Storage buckets containing training data, revoke service account keys, and close billing accounts. Maintain rollback capability for 90 days.

4. MIGRATION TOOLS & ACCELERATORS

4.1 API Translation & Compatibility

Gemini-to-OpenAI API Gateway

Lightweight proxy that translates Gemini API requests to OpenAI-compatible format. Enables migration of legacy Gemini SDK applications without code changes. Supports chat, embedding, and multimodal endpoints.

SDK Migration Toolkit

Automated refactoring tool that converts Google Generative AI SDK calls to OpenAI SDK format. Provides a migration report with line-by-line change recommendations and compatibility notes.

4.2 GCP Service Replacement

Vertex AI Pipeline Migrator

Converts Vertex AI Pipeline definitions to Kubeflow pipeline specifications. Maps GCP-specific components to open-source equivalents. Validates pipeline execution on-premise before cutover.

Data Migration Accelerator

Bulk data transfer from Cloud Storage to on-premise storage. Handles: training datasets, model artifacts, feature store data, and vector embeddings. Validates data integrity post-transfer.

4.3 Observability & Operations

Xinity Dashboard

Pre-configured monitoring for all on-premise AI services: inference latency, GPU utilization, model health, vector DB performance, and pipeline execution metrics.

Compliance & Audit Module

Generates compliance reports demonstrating full data sovereignty. Audit trails for every inference request. Ready-made report templates for GDPR, ISO 27001, and sector-specific regulations.

5. NEXT STEPS: START YOUR MIGRATION WITH XINITY

Migrating from Google Gemini to Xinity involves both API translation and GCP service decoupling, but Xinity's OpenAI-compatible endpoints and comprehensive on-premise stack make the transition systematic and predictable.

Here is how to get started:

1. Schedule a Discovery Call -- Xinity's solutions architects will map your entire Gemini + GCP ecosystem, identify migration priorities, and provide a detailed TCO comparison.

2. Request a Proof of Concept -- Deploy Xinity Runtime with your highest-priority workload. Test the Gemini-to-OpenAI API translation and validate output quality on your own data.

3. Plan Your GCP Decoupling -- Work with Xinity's migration team to sequence the replacement of each GCP service with on-premise equivalents.

4. Go Live with Full Sovereignty -- Complete the migration knowing your data, models, and inference infrastructure are 100% under your control.

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