LDC

YOUR CLOUD POWERHOUSE.

Aurai

Enterprise Sovereign AI Solution

Issue Date: 1/1/2026

Table of Contents

Executive Summary

Aurai is a high-performance **Large Language Model (LLM) platform** designed to provide the same advanced generative and conversational intelligence as global leaders like **ChatGPT** and **Gemini**, but within a strictly **Sovereign** and **Enterprise-Grade** framework.

While global AI platforms operate on public clouds, Aurai is purpose-built for organizations operating under strict regulatory, compliance, and data sovereignty requirements.

This is more than an AI tool—it's a **private AI ecosystem** that operates locally, allowing organizations to use advanced Large Language Model (LLM) features like intelligent document processing, and multilingual reasoning—while retaining full control over their data and AI operations.

The Value of Ready-to-Use Sovereignty

For many organizations, the challenge isn't just finding an AI model—it's the immense technical complexity of hosting, optimizing, and securing it locally. Building a sovereign AI stack from scratch requires specialized talent and months of experimentation, often distracting from the organization's core business objectives.

Aurai eliminates this friction. It provides a "Ready-to-Run" environment that serves as a foundation for innovation. By choosing Aurai, enterprises gain an immediate **Time-to-Market advantage**, bypassing the operational hurdles of AI infrastructure to not divert their focus away from their core business logic, domain innovation, or product differentiation, and focus entirely on building high-value applications and solving business problems.

Key Design Principles

- **GenAI-First by Design:** Built from the ground up to handle complex enterprise workloads, from creative generation to logical reasoning and data synthesis.
- **Accelerated Time-to-Market (TTM):** Aurai serves as a **Sovereign AI Accelerator**. Even for organizations with the resources to build their own solutions, Aurai provides a pre-optimized, production-grade foundation. This allows your team to stop worrying about low-level LLM orchestration and start delivering AI-driven value on day one.
- **Enterprise & Regulated Market Focus:** Tailored for sectors with compliance, privacy, and regulatory constraints.
- **Sovereign Alternative:** It delivers the conversational power and reasoning capabilities of models like GPT-4 or Gemini but ensures that **no data ever leaves** your control.
- **Domain Language-Aware Intelligence**: While global models are general-purpose, Aurai is fine-tuned to deeply understand **Modern Standard Arabic, regional dialects, and industry-specific terminology** (Legal, Finance, Government), providing a level of cultural and professional accuracy that global models lack.
- **Controlled Execution:** AI behavior is constrained to specific operational tasks to ensure predictability.
- **Enterprise Integration:** Native compatibility with existing enterprise systems, workflows, and authentication mechanisms, enabling seamless adoption without disrupting operational processes.
- **Production Accountability:** Unlike "experimental" AI, Aurai is designed for 24/7 production environments, offering stability, governance, and scale required by regulated markets.
- **Managed Service Delivery:** Aurai is delivered and operated as a managed service by LDC, ensuring continuous operation, updates, optimization, and enterprise-grade support—rather than leaving customers to manage complex AI stacks alone.

Core Architecture

Overview

Aurai is built on a modular, multi-layer architecture designed to provide both specialized expertise and massive operational scale.

Rather than relying on a single general-purpose model, Aurai utilizes a **Curated Expert Crew** approach, where specialized models are fine-tuned to master specific enterprise domains.

Development Philosophy

At the foundation of Aurai is a rigorous process of model selection and enhancement. Our engineering team continuously evaluates the global open-source landscape to identify the most capable base models for specific tasks. We then apply **LoRA (Low-Rank Adaptation)** and advanced alignment techniques to inject deep, specialized knowledge:

- **The Linguistic Expert:** Fine-tuned for **Advanced Arabic Reasoning**. Beyond Modern Standard Arabic (MSA), it is trained on regional dialects and industry-specific terminology (Legal, Finance, Healthcare, Government) to ensure it understands the cultural and professional context of the MENA region.
- **The Vision & IDP Expert:** A specialized Vision-Language Model (VLM) fine-tuned specifically for **Identity and Regulatory Documents**. It is engineered to process ID cards, commercial registries, and complex structured forms with high precision in both Arabic and English.
- **The Reasoning & Logic Expert:** A model optimized for complex chain-of-thought tasks, code generation, and structured data analysis, ensuring high accuracy in technical and mathematical prompts.

Agentic Orchestration Layer (The "Brain")

The power of Aurai lies in its **Agentic Orchestration Layer**, which manages the model pool to ensure they work as a unified team. It operates with two primary functions depending on the deployment configuration:

A. The Heterogeneous Approach (Intelligent Routing & Decomposition)

In this mode, the Orchestrator acts as an **Agentic Router**. It doesn't just pass text; it performs "Prompt Interception":

- **Classification:** It analyzes the prompt in real-time to identify intent, language, and complexity.
- **Task Decomposition:** For complex requests, the Orchestrator can break a single prompt into sub-tasks, distributing them to the relevant "experts" (e.g., sending an image to the OCR expert and the extracted text to the Linguistic expert).
- **Response Aggregation:** It synthesizes the outputs from different models into a single, coherent, and high-quality response for the user.

B. The Homogeneous Approach (Elastic Scaling & Load Balancing)

For clients who require extreme throughput on a specific task (e.g., a high-volume customer service bot), Aurai can be configured to run **Multiple Concurrent Instances** of a single specialized model.

- **Traffic Control:** The Orchestration Layer acts as a high-performance **Load Balancer**, distributing hundreds of concurrent requests across the model pool.
- **Failover & High Availability:** It ensures zero downtime by instantly rerouting traffic if one instance becomes unresponsive, maintaining 24/7 production stability.

Ready-Made Task Suite

Aurai provides a set of **ready-to-use, production-grade AI capabilities** exposed as secure enterprise APIs, enabling organizations to embed AI features directly into existing applications without building custom AI pipelines or agentic systems.

These are not just prompts or prompt templates; these capabilities are the result of **targeted fine-tuning and model alignment**, ensuring that each task delivers **consistent,** predictable, and enterprise-ready outputs.

To ensure maximum flexibility, these tasks are exposed through two distinct interfaces:

RESTful APIs (For Traditional Integration)

Designed for developers building standard applications (Web, Mobile, ERP) who need to inject specific AI capabilities without building complex agentic logic.

- **Use Case:** A banking app that needs to automatically analyze customer feedback (Sentiment Analysis) or a legal firm that needs to summarize 100-page contracts via a simple API call.

MCP Server (For the Agentic Ecosystem)

Aurai acts as a **Sovereign Tool Provider** via the Model Context Protocol (MCP). This allows any AI Agent (whether running on Aurai or another orchestration framework) to "plug in" Aurai's specialized skills instantly.

- **Aurai as a Skill Provider:** If an organization uses a third-party agentic framework, they can grant their agents "Aurai Skills"—such as high-precision Arabic OCR or Dialectical NER—simply by connecting to the Aurai MCP Server.
- **Native Agentic RAG:** When building autonomous agents on the Aurai platform, the LLM can "discover" and use these tasks as native tools to solve complex, multi-step problems (e.g., "Analyze this image, extract the ID number, and check if it matches the sentiment of the attached voice note").

Key Categories of the Task Suite:

- **Linguistic Intelligence:** Advanced summarization, context-aware translation, and professional tone adjustment in Arabic & English.
- **Document Cognition (IDP):** High-fidelity extraction from specialized documents (National IDs, Commercial Registries, Invoices) using fine-tuned Vision-Language Models.
- **Behavioral Analytics:** Real-time Named Entity Recognition (NER) and deep Sentiment/Emotion analysis calibrated for regional nuances and dialects.

Pre-loaded Embedder & Tokenizer

Aurai platform comes pre-loaded with Native Embedding and tokenization engines to support retrieval-augmented generation (RAG) and advanced NLP tasks.

OpenAI-Compatible API Layer

OpenAI-compatible APIs to provide a seamless "drop-in" replacement for existing AI integrations, supporting secure enterprise authentication (JWT/OAuth2).

DEPLOYMENT & OPERATIONAL MODELS

Aurai is designed to meet organizations wherever their journey into Generative AI begins. We offer three primary delivery models, each optimized for different levels of sovereignty, cost, and infrastructure maturity.

Aurai AI Box (Dedicated Private Infrastructure)

Overview

The AI Box is a self-contained, high-performance hardware appliance. It is the premier choice for organizations requiring **single-tenant isolation** and maximum throughput.

It can be deployed in two ways:

- **Option A: On-Premise Installation** Physical deployment within the client's own data center. This is the ultimate choice for air-gapped environments and national security use cases where data must never leave the building.
- **Option B: Managed Colocation (Hosted by LDC)** The dedicated hardware is housed within LDC's Tier-certified, PCI-GDPR compliant data centers. The client benefits from **isolated, non-shared hardware** performance and security, while LDC handles all physical environmental factors (power redundancy, cooling, and physical security). ***Benefit****:* Hardware-level privacy with zero infrastructure overhead for the client.

The "Family" Architecture

Rather than forcing a single operating pattern, and to match specific business needs, the Box is configured into two architectural families based on **how models are used**, not just hardware size.

Aurai-H Family (Single Model Architecture)

Designed for organizations that have a specific, high-frequency use case and workload which is concentrated within a single capability domain (e.g., a massive customer service bot or deep legal research).

- **Scale-Up Mode (Deep Reasoning):** Uses a single, high-parameter model. The entire GPU capacity is dedicated to this model to allow for lower quantization levels, **massive context windows**, and the highest level of reasoning for complex prompts.
- **Scale-Out Mode (High Throughput):** Uses a single, mid-sized model instance optimized for speed. Because the hardware capacity exceeds the model's needs, it can handle a significantly higher number of concurrent requests (**High RPM**) for standard tasks.

Aurai-X Family (Multi-Model Architecture)

The default and most versatile Aurai configuration, designed for multi-departmental use.

- **Specialization Mode (Standard):** It is a **Heterogeneous** Multi-Model architecture which deploys a "Crew" of different specialized models (Linguistic Expert, Vision/OCR Expert, Reasoning Expert). The Orchestration Layer routes or decomposes prompts to the right expert.
- **Concurrency Mode (Ultra-Scale):** It is a **Homogeneous** Multi-Instance architecture for environments requiring both specialization and massive volume. Multiple identical instances of the same expert model are deployed, using the Orchestrator as a **Load Balancer** to maintain speed during peak loads.

Hardware Sizing & Form Factors

The physical hardware is sized based on the required **Concurrency Level (RPM)** and the **complexity** of the model "Crew.":

- Single vs multi-model requirements.
- Required RPM and concurrency.
- Context window size.
- Degree of specialization vs throughput.

Aurai AI Boxes are available in multiple form factors:

- **Edge/SME Tier:** Compact **Tower Servers or Workstations**. Ideal for small datacenter rooms, smaller departments, or specialized local tasks.
- **Enterprise Tier:** High-density **Rack-Mounted Servers**. Designed for data center integration to serve the entire organization with massive GPU clusters.

Hardware selection is always aligned to **business workload—not arbitrary SKUs**.

Sovereign SaaS Platform (The National Cloud)

Overview

The **Sovereign SaaS** is a shared GenAI platform-as-a-service, hosted **within national borders** (e.g., LDC's Egypt-based data centers). The platform is operated within **PCI-GDPR certified sovereign datacenters**. It offers a Pay-As-You-Go (PAYG) model while strictly adhering to local data residency laws.

- **Regional Vision:** This model is designed for rapid expansion across the **GCC, Levant, Africa and North Africa**, allowing each nation to host its own "Sovereign AI Hub."
- **Private Cloud (Aurai Stack):** For very large organizations (e.g., Government Ministries), we offer a **Private Stack**—a dedicated, private version of our SaaS platform that acts like a local "GEN-AI Stack," providing the ease of SaaS with the isolation of a private cloud.

Operating Architecture

It Runs the full **Aurai-X Multi-Model Crew** by default, ensuring the highest level of intelligence and task decomposition.

Regional Cloud SaaS (Global Scale)

A fully managed shared GenAI platform-as-a-service (SaaS), designed for enterprises already operating on public cloud infrastructure (e.g., Microsoft Azure). While the infrastructure resides in global regions (e.g., Europe), it provides the most cost-effective and elastic way to access Aurai's intelligence.

- **Maximum Elasticity:** Built for the highest possible scale, combining multi-model specialization with high-concurrency instances.
- **Cost-Efficiency:** The best entry point for non-regulated workloads that require Aurai's superior Arabic linguistic and OCR capabilities at a competitive price point.

Operating Architecture

It Runs the full **Aurai-X Multi-Model** combining the **Multi-model Specialization**, and the **Multi-instance concurrency** achieving extreme RPM handling.

Trade-off:

- Not a sovereign deployment model
- Infrastructure may reside outside national borders

Value Proposition:

- Lowest cost of entry.
- Massive scalability.
- Fastest time to market.

Comparison Summary

Feature	AI Box	Sovereign SaaS	Regional Cloud
Sovereignty control	Physical Isolation (On Site/Colo)	Jurisdictional (In-Country)	Global Infrastructure with goverence layer
Data Residency	Local, air- 100% gapped.	Local, in country sovereign cloud.	Global cloud regions.
Deployment Model	Single Tenant. Dedicated Hardware	Multi-Tenant or Dedicated Private SaaS.	Public Cloud Managed Multi-Tenant SaaS.
Operating Model Architecture	H or X Families	Full X-Crew (Multi-model Specialization)	Full X (Multi-model Specialization + multi-instance concurrency)
Primary Execution	Single Model or Multi-Model	Multi-Model Orchestration	Multi-Model Orchestration + High Concurrency
Scalability / RPM	High (Configurable at HW level)	Moderate to High (Shared pool)	Very High (Elastic hyperscale)
Hardware	Dedicated (Tower/Rack)	Shared (LDC / Privte Cloud)	Shared (Public Cloud Hyperscaler)
Primary Driver (Best Fit For)	Security & Air-gap	Compliance & Residency	Scale & Cost
Cost Model	CapEx	OpEx	OpEx
Internet Dependency	None (Offline / Air-Gapped)	None (Closed Sovereign Network)	Hosted on public cloud
Dependency on Public APIs	None	None	Not required – but a search tool is enabled.

SECURITY, GOVERNANCE & COMPLIANCE

Regardless of the deployment model, the Aurai ecosystem is governed by a **Zero-Data-Exfiltration** philosophy:

- **Privacy by Design:** Data, prompts, and outputs are never used to train global models.
- **Local Inference:** Processing occurs within the defined jurisdictional or physical boundary.
- **Governance:** Native support for Role-Based Access Control (RBAC), auditing logs, and JWT/OAuth2 authentication.
- **No Internet Dependency:** The system can operate in fully disconnected (air-gapped) modes.

Management & Support Scope

We do not just deliver a product; we operate the intelligence layer. Our AI Engineering Team manages the AI Box end-to-end:

- **Inference Optimization:** Continuous tuning of batch sizes, memory usage, and GPU parameters.
- **Model Lifecycle:** Deploying new model versions and hot-fixing behavioral regressions.
- **Software Maintenance:** Security patches for the AI runtime and API layer updates.
- **Technical Support:** Remote incident handling and specialized support for RAG and API integration.

CONCLUSION: THE FUTURE OF SOVEREIGN INTELLIGENCE

The adoption of Generative AI is no longer an experimental luxury; it is a fundamental shift in how enterprises operate, innovate, and compete. However, for organizations in regulated sectors and national industries, the path to AI integration has often been blocked by the risks of data exposure, lack of cultural nuance, and dependency on external providers.

Aurai was built to remove these barriers.

By combining the raw power of world-class Large Language Models with a strictly sovereign architecture and deep Arabic linguistic mastery, Aurai empowers your organization to lead in the AI era without compromising on your most valuable asset: **Your Data.**

Why Aurai is the Strategic Choice:

- **Ownership:** You own the environment, the models, and the outputs. No data ever leaves your control.
- **Precision:** Unlike general-purpose AI, Aurai is fine-tuned for your language, your industry, and your specific operational tasks.
- **Reliability:** Operated as a managed service, Aurai provides the stability and performance required for mission-critical production environments.
- **Scalability:** From compact edge devices to national-scale cloud platforms, Aurai grows alongside your organizational needs.

As we move toward a future where "AI Sovereignty" becomes the global standard, LDC remains committed to being your partner in this journey—providing the tools, the expertise, and the infrastructure to turn the promise of Generative AI into a secure, predictable, and high-impact reality.

Own your intelligence. Secure your future. Deploy Aurai.