

Building Fully Private AI Foundation for a Leading Bank in Saudi Arabia

Using agentic AI for real-time drift detection with
automated remediation workflows



AI ready
architecture

Regulatory
compliance

Proactive
protection

Comprehensive
audit trails

About Client

A leading Saudi Arabian financial institution, headquartered in Riyadh and established in the 1950s, has grown into one of the Kingdom's major banks. It offers personal, commercial, investment, and Islamic banking services through a nationwide branch network and secure digital channels. The bank continues to strengthen its market position through digital innovation and strategic global partnerships.

Key Challenges

Operating in a highly regulated environment, the bank needed stronger control over configuration changes to reduce compliance risk, prevent security gaps, and avoid business disruption. Manual audits, limited visibility, and unmanaged configuration drift made it difficult to maintain continuous compliance and operational resilience.



Undocumented Configuration Changes

Increased the risk of compliance violations and security breaches



Manual Audit Processes

Created delays, inefficiencies, and higher operational effort

**Limited Real-Time Visibility**

Made it difficult to detect and address risks proactively

**Inefficient Processes**

Increased the likelihood of costly disruptions and control failures

The customer also required a fully private AI foundation, where every architectural decision including hardware selection, software stack design, automation, operations tooling, and integration had to be evaluated, engineered, and stitched together for enterprise-grade performance, governance, resilience and at optimal cost.

Solution

Microland assessed the customer's requirements and developed a range of solutions, including an Agentic AI based approach which could test the end to end fully private, on-premises AI Infrastructure environment that the customer wanted Microland to design, build and operate with enterprise grade governance and LLM Ops. While Microland has previously enabled AI infrastructure deployments on Hyperscalers platforms such as Azure AI Studio and Google Vertex AI, this engagement was distinct in its complexity and strategic importance.

Key Solution

The solution was delivered across three integrated layers:

**AI Infrastructure Layer**

A purpose-built, on-premises AI infrastructure foundation designed for secure, high-performance model workloads

**LLMOps Layer**

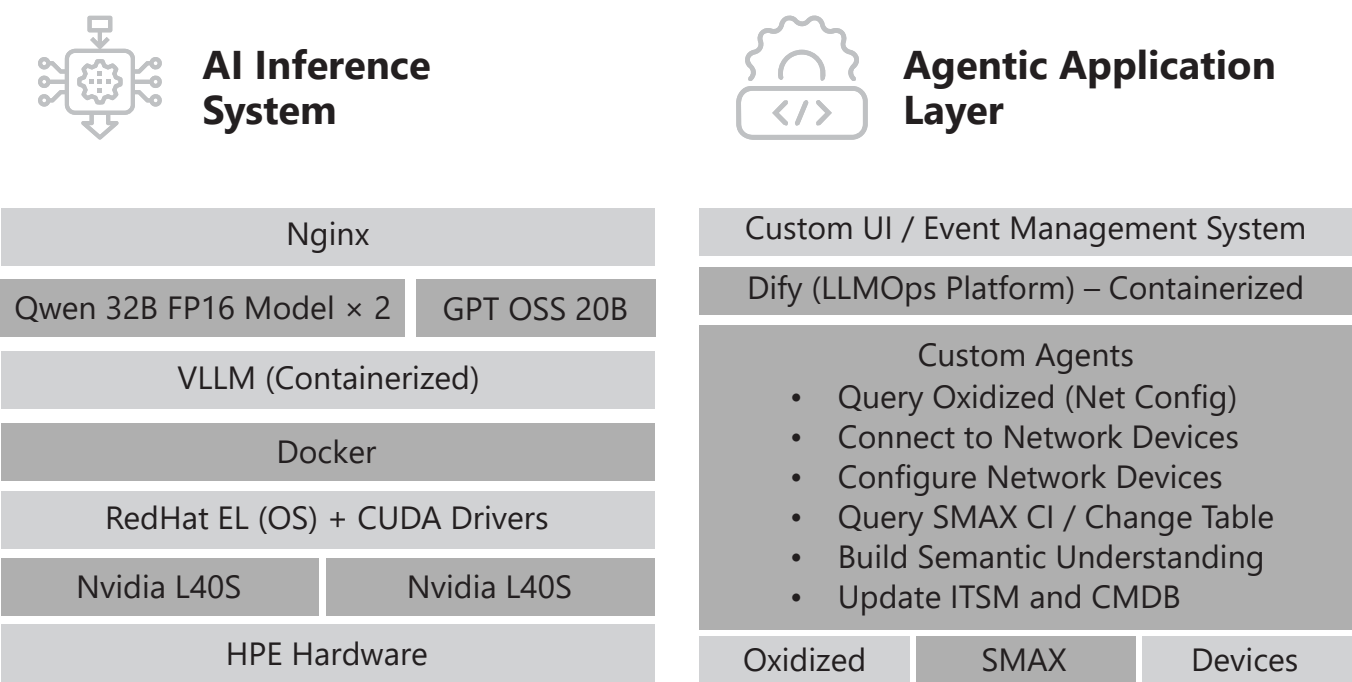
An operational layer to manage, monitor, govern, and scale large language model workloads across the environment

**Agentic Application Layer**

A business application layer built on top of the AI foundation to enable the customer's specific use case

The Architecture

The overall architecture was driven by four core enterprise priorities around compliance, data sovereignty, cost efficiency, and operational control. Rather than building a generic “AI factory” intended for any workload, Microland designed a fit-for-purpose AI environment aligned to the customer’s specific business and regulatory requirements and use cases.



AI Infrastructure Layer

At the infrastructure layer, Microland designed a private AI foundation optimized for inference-led GenAI workloads. The solution used a dual NVIDIA Ada Lovelace GPU architecture, delivering the right balance of cost per token, performance per watt, and workload alignment. The stack was built on Red Hat Enterprise Linux with CUDA, rather than a full DGX software stack, to keep the environment streamlined, cost-effective, and aligned to the pilot objectives. vLLM on Docker was used for model serving, with a clear path to Kubernetes for production-scale orchestration.

LLMOps Layer

The environment supported two open-source models - Qwen for multilingual capabilities, including Arabic, and GPT-OSS for broader GenAI experimentation. NGINX provided lightweight model routing for the pilot, with the option to evolve to NGINX Plus, LiteLLM, or Bifrost in production.

GPU management & Agentic Application

For GPU management, NVIDIA BCM Kickstart was used to simplify foundational GPU configuration and operations. Heavier NVIDIA software components were intentionally avoided, given the customer's need to control cost. Above this infrastructure layer, Microland implemented Dify as the LLMOps layer, enabling model integration, workflow orchestration, monitoring, and observability for agentic applications. Microland also built a purpose-specific Agentic application for network drift detection, remediation, and network configuration audit, demonstrating a tangible enterprise use case on the private AI platform.

Business Value Delivered

By aligning the AI hardware and open-source software stack to the customer's actual workload needs, Microland helped avoid an over-engineered architecture while creating a roadmap for an OEM-supported enterprise stack. This delivered a fully compliant solution at 40% lower cost than the customer's original in-house plan.

The Agentic AI solution for drift detection and auto-remediation reduced security gaps, lowered regulatory exposure, and strengthened operational control. It also demonstrated that the customer's private AI infrastructure could be enterprise-grade, scalable, and cost-effective from day one.

Agentic Drift Detection & Remediation

Agents detects drifts or changes captures and backs it up using semantic understanding, compares it to the previous version, verifies approval in the ITSM system, and autonomously fixes it and generates a detailed, fully traceable audit report

AI-Driven Change Intelligence

The AI system understands the intent behind changes, distinguishes authorized from unauthorized actions, and links them to approved requests improving accuracy, reducing false alerts, and strengthening compliance

Protection & Containment

The solution runs securely on private, on-premises infrastructure to meet 100% data privacy and regulatory requirements, while providing clear visibility into all network configuration changes

Recognized by Global Analysts



About Microland

Microland is a leading AI-first, platform-led, technology infrastructure services company. We have enabled enterprises to build intelligent, resilient, and future-ready operations and are a trusted partner to global enterprises. We bring over 35 years of expertise in digital networks, cloud, data centers, workplaces, and cybersecurity, and combine it with our commitment to customer centricity, delivery excellence, and continuous innovation. Our operations, currently in more than 100 countries, are supported by a strong global delivery model and our AIOps platform, **intelligeni**, powered by Agentic AI, which is shaping the future of autonomous technology operations across enterprises.

www.microland.com

