GCloud-15

Driving Secure Digital Transformation

Private LLM Service

Secure Local Large Language Model

Service Definition Document

January 2026

+44 (0) 20 3384 3350 | gcloud@viewdeck.com | www.viewdeck.com | Viewdeck Consulting Limited - 207 Regent St. London, W1B 3HH

Private LLM Service – Secure Local Large Language Model

Summary

Enables organisations to deploy large language models entirely within their controlled environment. It provides secure, compliant, high-performance AI capability without sending data to external providers. The platform supports private, hybrid and offline deployments, ensuring full data sovereignty, predictable performance and enterprise governance for modern AI workloads.

Features

- Local AI model hosting with full data control
- Secure processing for sensitive organisational workloads
- Predictable performance for critical AI operations
- Unified interface for all internal AI applications
- Centralised governance across all model interactions
- Compliance-ready deployment for regulated environments
- Flexible model selection for diverse needs
- Scalable architecture for team-to-enterprise growth
- Offline-capable for high-assurance environments
- Simple integration with existing workflows

Benefits

- Keeps all data within the organisation
- Enables safe AI adoption without external dependency
- Supports strict regulatory and sovereignty requirements
- Reduces operational cost through local inference
- Improves productivity with consistent low-latency responses
- Provides trustworthy, auditable AI behaviour
- Protects intellectual property during AI processing
- Ensures resilience through offline operation
- Accelerates AI rollout across business units
- Enhances decisions with reliable private intelligence

Service Description

Overview

The Private LLM Service provides a secure, self-contained environment for local deployment of modern large language models across private cloud, sovereign cloud, on-premise data-centres or entirely offline infrastructure. It is designed for organisations requiring full control over data processing, confidentiality, model behaviour and operational boundaries.

The platform supports Ollama, vLLM, or both, enabling flexible model execution aligned to performance, concurrency and workload patterns. Ollama offers lightweight, rapid model loading suitable for experimentation and multi-model scenarios. vLLM provides a high-throughput, optimised inference engine ideal for production workloads, larger models and parallel use at scale. When deployed together, teams can select the runtime that best matches each use case.

A user-friendly interface is provided through OpenWebUI for testing, evaluation and prompt engineering. For integration with internal systems, an optional LiteLLM gateway exposes a fully local OpenAI-compatible API endpoint with routing policies, usage controls and audit logging. This allows applications and automation tools to interact with local models securely, without any external data transmission.

The platform supports models from Ollama and Hugging Face, including compact, multilingual, domain-specific and high-parameter LLMs. Deployments are optimised for Linux, CUDA and NVIDIA GPU hardware, using containerised delivery to ensure portability, isolation and predictable performance. GPU resources—from a single evaluation node to multi-GPU production clusters—are configured for efficient batching and high-performance inference.

The Private LLM Service integrates naturally with internal agent frameworks, supporting retrieval-augmented generation, SQL querying and data-aware reasoning without exposing any information outside the organisational boundary. All data, embeddings, prompts and logs remain internal, supporting compliance with legal, contractual and sovereign requirements.

Managed updates, patching, optimisation and model lifecycle controls are included. Deployment options cover test, starter, production and resilient multi-node environments across private cloud, public cloud private tenancy, hybrid infrastructure or offline installations.

This service gives organisations complete control of model execution, governance, hardware and data, enabling safe and scalable adoption of generative AI while protecting confidentiality and ensuring compliance.

Architecture

The platform follows a modular, secure architecture designed for high-performance, local model inference with strict control of data flows. It avoids reliance on public AI services or external APIs.

Incident Management

Viewdeck classifies incidents raised at its service desk using the following P1 – P3 priority. Incidents raised with the Service Desk will be triaged and responded to accordingly in priority order.

- P1 Total loss of service to all users.
- P2 Some loss of critical service for some or all users.
- P3 Limited loss of service or work around possible limiting loss experienced.

Priority	Response type (during standard business hours)	Target resolution (Mon-Fri 9-5)
P1	30 minute response, sustained effort using all necessary and available resources until service is restored.	4 hours
P2	4 hour response to assess the situation, staff may be interrupted and taken away from lower priority jobs.	1 working day
P3	One working day response using standard procedures and operating within the normal frameworks.	3 working days