viewdeck

GCloud-15

Private LLM Service

Pricing 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

Private-LLM enables organisations to deploy and operate modern LLMs **entirely inside their own controlled environment** (private cloud, sovereign regions, hybrid, or fully offline on-prem). It provides a governed inference stack using Ollama and/or vLLM, plus optional OpenAI-compatible API access via LiteLLM and an interactive UI via OpenWebUI—so internal applications and internal agents can route inference privately without data egress.

Key Dependencies

- **Operating system**: Ubuntu 24.04 LTS (preferred)
- **Container runtime**: Docker CE/EE;
- **GPU**: NVIDIA CUDA-capable GPUs (minimum VRAM 32GB for predictable enterprise-grade performance)
- **Drivers/runtime**: NVIDIA proprietary drivers; supported CUDA stack
- **Optional integration**: SSO/RBAC controls when used with internal agents

Component & Options

Key component	Included (Base)		Dependencies / Notes
Inference Runtime	Yes	Ollama; vLLM; Hybrid (both)	Ollama: fast switching/dev; vLLM: High-throughput/large models
API & Interaction Layers	Optional	LiteLLM (OpenAI-compatible gateway); OpenWebUI; native APIs	LiteLLM adds routing, metering, controls
Model Support	Yes	Curated model packs; model lifecycle management; optimisation	Model files licensed by model providers; verify licensing
Security & Governance	Yes	Air-gapped operation; secret-management integrations; compliance hardening	Keeps prompts/outputs/logs inside boundary
Performance & Reliability	Yes	Performance tuning; batching/memory tuning; resilience enhancements; monitoring/telemetry	Performance depends on GPU/VRAM and runtime choice
PaaS	Additional (Required)	Depending on deployment pattern, either a Bare-Metal, IaaS, Container or Cloud service is needed to run the service. On-Prem, Private, Public or Sovereign solutions	Full range of platforms can be used. Bare Metal service will require provisioning. Support for IaC over Terraform.
GPU	Additional (Required)	NVidia CUDA Supported GPU.	16Gb->9Gb VRAM. All platforms supported by the current Nvida CUDA drivers.

Patterns

Variant	Description	Ideal use
Single / Test	Lightweight single-node for evaluation and experimentation	PoC, model evaluation, R&D
Dual / Starter	Two-node for redundancy and moderate concurrency	Departmental use; small-scale production
Multi / Production	Multi-node scalable environment	Enterprise inference; automation at scale
Resilient / HA	High-availability clustered config with failover/load balancing	Mission-critical, regulated continuity-driven environments

Price List

Line item	Unit	Setup & Installation	Service Per month	Pattern	Notes
LLM Private Hosted LLM	Per Container	£2500	£500	Single Container Environment - 1 Host	Single Instance. Model Loading/Management. Tuning for Performance. Upgrade/Patch Management. Ollama or VLLM. OpenWebUI for admin/testing. Includes Rev-Proxy. Any LLM(s) supported by platform
LLM Proxy	Per Instance	£1500	£200	Single Container	LLM resilience, Rev Proxy. OpenAi API. Configuration. Upgrade/Patch etc.
Container Environment	Per Host	£2500	£200	Single Docker Instance	Platform enabling GPU configuration. Enabling container environment to support GPU containers.
Bastion Host & Remote Support Access/VPN	Per Solution	£2,500	£250	3 Containers per site/Zone	Remote Access over VPN, with controlled end point access, logging, Admin Terminal, Container Management, IAC deployment. Build Desktop
Model Testing Onboarding & Tuning	Per Model	£2,500		Dev/Test Container	Model Testing against agreed baselines to confirm operation and parameters.

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