stack8s – Sovereign LLM Inference Endpoints

G-Cloud 15 Service Definition

1. Service description

stack8s Sovereign LLM Inference Endpoints is an enterprise-grade large language model (LLM) inference platform delivering production-ready AI capabilities as a managed SaaS service. It provides scalable, sovereign-first LLM endpoints across EU and UK cloud providers, ensuring data residency and regulatory compliance while offering high-performance inference infrastructure for public sector and regulated organisations.

stack8s integrates with leading model ecosystems, including Hugging Face and NVIDIA AI model repositories, to provide access to a curated catalogue of state-of-the-art foundation and domain-specific models within sovereign EU/UK environments. Models can be sourced from public or private Hugging Face repositories, as well as NVIDIA NIM and other NVIDIA AI model offerings, then onboarded into the platform's managed hosting, monitoring, and scaling framework while preserving data residency and security controls. This approach allows organisations to rapidly adopt the latest advances in generative AI and specialised LLMs without having to engineer their own GPU infrastructure or model deployment pipelines, while maintaining full alignment with UK and EEA sovereignty requirements

The service exposes production-ready endpoints for leading foundation models, including GPT, Llama and Mistral, alongside customer fine-tuned variants. Customers integrate via RESTful APIs secured with token-based authentication and rate limiting, enabling applications to use advanced natural language capabilities such as summarisation, question answering, classification, and conversational interfaces. stack8s operates and optimises the underlying GPU-backed infrastructure, autoscaling, and multi-cloud deployment so that teams can consume LLM capabilities without building and managing their own model hosting stack.

The platform is designed as a sovereign-first AI service: model hosting and inference occur within UK and EEA regions, with configuration options for jurisdictional control over data storage and processing. It is suitable for organisations that must demonstrate where data is handled, how it is protected in transit and at rest, and how AI systems operate within clearly defined governance and assurance boundaries.

---

2. Service features and benefits

2.1 Service features

CONFIDENTIAL - DO NOT DISTRIBUTE WITHOUT PERMISSION

- Data sovereignty with EU and UK cloud hosting
  All LLM inference workloads, including model execution, transient prompts, and logs, are processed within UK and EEA cloud regions. Customers can specify allowed locations to ensure alignment with data sovereignty, departmental policies, and GDPR requirements for AI workloads.

- Scalable inference with automatic load balancing
  The platform automatically scales model instances based on traffic and routes requests across healthy replicas using built-in load balancing. This provides consistent latency and high availability even under variable or spiky workloads, without the customer needing to manage scaling policies or infrastructure.

- Production-ready LLM Open Source model endpoints (GPT, Llama,, Mistral)
  stack8s offers managed endpoints for a curated set of leading LLMs, including GPT, Llama, and Mistral families. These models are hosted and optimised for inference in sovereign environments, enabling customers to choose the most appropriate model for each use case while avoiding the complexity of provisioning and maintaining them.

- RESTful API access with token authentication and rate limiting
  Clients access models through RESTful APIs with token-based authentication, enabling straightforward integration from applications, services, and back-end systems. Rate limiting is enforced per customer and/or API key to protect platform stability, manage fair usage, and provide predictable behaviour under high load.

- Real-time inference monitoring with latency and usage metrics
  The service provides real-time metrics on request volumes, latency, error rates, and usage per endpoint. These metrics support operational monitoring, capacity planning, and cost optimisation, and can be accessed via dashboards or APIs and integrated into existing observability tooling.

- Support for fine-tuned custom models alongside foundation models
  Customers can deploy fine-tuned variants of supported foundation models, allowing domain-specific behaviour and improved performance on organisational data. The platform hosts both base and custom models under the same governance and sovereignty controls, enabling a mix of generic and specialised endpoints in production.

- Multi-cloud deployment and isolation
  The service runs across multiple EU/UK cloud providers but presents a unified control plane, with stack8s managing placement, capacity, and failover. Each customer's backend services are deployed in isolated environments with dedicated resource allocations enforced using Kubernetes namespaces and quotas to prevent noisy neighbour effects.

- Unified management console and APIs
  A web-based management console provides visibility of endpoint configurations, usage, performance, and access controls, and supports configuration of model selection, limits, and

CONFIDENTIAL - DO NOT DISTRIBUTE WITHOUT PERMISSION

integration settings. APIs complement the console by enabling automated provisioning, configuration changes, and retrieval of metrics and logs.

- Comprehensive onboarding, documentation, and data export
  The service includes online training, webinars, and extensive documentation (HTML and PDF) covering provisioning, API integration, and AI-specific best practices, including EU/UK sovereignty guidance. Data export tools via APIs and the console allow retrieval of logs, configuration, and associated data in standard formats (JSON, CSV, SQL), with a 30-day read-only grace period following contract termination.

Example - Five of the strongest widely used open(-weights) LLM families today are:

- Llama 3 (Meta) – Current Llama 3 70B and smaller variants are among the most capable openly available general-purpose models, with strong performance on reasoning and MMLU-style benchmarks.
- Gemma 2 (Google) – Gemma 2 (9B and 27B Instruct) delivers competitive quality with relatively efficient footprint, and regularly appears at or near the top of open-source leaderboards.
- Qwen2 / Qwen2.5 (Alibaba) – Qwen2 72B and related instruct/chat variants score very highly on benchmark leaderboards (including MMLU and MT-bench), and are strong choices for multilingual and reasoning-heavy tasks.
- Mixtral (Mistral AI) – Mixtral 8x7B and 8x22B sparse Mixture-of-Experts models offer very strong performance for their effective parameter counts and are widely used for high-throughput, high-quality inference.
- Nemotron-4 (NVIDIA) – Nemotron-4 340B Instruct is a large, high-performing open model family from NVIDIA that ranks near the top of open leaderboards and is designed to integrate well with NVIDIA's AI inference stack.

2.2 Service benefits

- Accelerate AI application development with ready-to-use model endpoints
  Development teams can integrate advanced language capabilities into applications quickly by calling managed endpoints rather than training or hosting models themselves. This reduces time-to-value for AI features and allows teams to focus on user experience and business logic instead of platform engineering.

- Maintain full data sovereignty with EU/UK hosted inference
  By ensuring all inference processing occurs in UK/EEA regions and providing clear control over data storage and export, the service supports compliance with GDPR and public sector data policies. This enables organisations to deploy generative AI features while retaining confidence that data is not transferred outside approved jurisdictions.

- Reduce infrastructure complexity with managed model hosting and scaling
  stack8s handles GPU provisioning, scaling, upgrades, and resilience for LLM hosting, as well as

operational concerns such as monitoring, logging, and security patching. Organisations avoid the cost and complexity of building their own LLM infrastructure while still benefiting from enterprise-grade performance and availability.

---

3. Use cases

stack8s Sovereign LLM Inference Endpoints supports a wide range of AI-driven use cases, including:

- Conversational assistants and chatbots
  Building chat interfaces for citizen services, internal helpdesks, or policy guidance, using sovereign LLM endpoints for natural language understanding and response generation.

- Document understanding and summarisation
  Processing long-form documents such as policy papers, reports, and case files to extract key points, generate summaries, or answer questions, with processing constrained to UK/EEA regions.

- Knowledge retrieval and Q&A over internal content
  Combining LLM endpoints with retrieval techniques to answer questions based on organisational knowledge bases, ensuring that both content and inference remain within sovereign infrastructure.

- Content drafting and templating
  Assisting staff with drafting emails, reports, briefings, and other written materials based on organisational templates and guidelines, with fine-tuned models aligning outputs to house style.

- Classification, routing, and triage
  Automatically classifying inbound requests, documents, or messages and routing them to appropriate teams or processes, enabling smarter triage in contact centres and back-office workflows.

---

4. Pricing overview

Pricing for stack8s Sovereign LLM Inference Endpoints is set out in the stack8s UK G-Cloud 15 Pricing Document in the Digital Marketplace. This document describes pricing for usage (for example, token or request-based charging), model tiers, support levels, and any optional professional services or enterprise features.

CONFIDENTIAL - DO NOT DISTRIBUTE WITHOUT PERMISSION

5. Service constraints

- Dependency on underlying EU/UK cloud providers
  Service availability and performance are influenced by the health of the sovereign cloud providers hosting the inference infrastructure. While the platform is architected for resilience, major provider incidents may impact endpoint availability until underlying services recover or traffic is rerouted.

- Scheduled maintenance windows with notice
  Planned maintenance windows are required for platform upgrades and security patches and are communicated to customers at least 48 hours in advance. Maintenance is typically performed outside standard UK business hours, but temporary reductions in service or brief interruptions may occur during these periods.

- SaaS delivery model and platform boundaries
  The service is delivered as SaaS and cannot be installed on-premises or in customer-managed cloud accounts. Customers integrate via the management console and APIs and do not have direct access to the underlying model hosting infrastructure or GPU clusters.

6. Technical requirements

- Client access requirements
    - Modern web browser (such as Edge, Chrome, Firefox, or Safari) with JavaScript enabled for access to the management console and dashboards.
    - HTTPS connectivity for secure access to APIs and the console, using TLS 1.2 or above between buyer and supplier networks and within the supplier network.

- Network and integration requirements
    - Customer environments must allow outbound HTTPS traffic to stack8s endpoints for API calls and management operations.
    - Applications integrating with LLM endpoints must support RESTful HTTP calls and handling of token-based authentication and rate limiting.

CONFIDENTIAL - DO NOT DISTRIBUTE WITHOUT PERMISSION

- Data export and integration capabilities
    - To make use of data export features, customers require the ability to receive and process standard formats such as JSON, CSV, and SQL dumps, either via APIs or through console-based downloads.

CONFIDENTIAL - DO NOT DISTRIBUTE WITHOUT PERMISSION