Rafay-powered Model as a Service

Deploy, scale, and monetize GPU-powered inference endpoints optimized for large language models (LLMs).

Offer tenants production-grade inference services with low latency, high throughput, and elastic scaling without the burden of manual infrastructure management.

From Bottleneck to Business Value

Demand for LLM-powered applications is rising fast. Yet traditional inference approaches create friction: static GPU allocation wastes capacity, idle costs add up, and runtime management strains operator resources.

Tenants, however, expect:

- Self-service APIs they can call instantly.
- Elastic scaling to handle unpredictable workloads.
- Predictable latency and performance for production use.

Rafay Inference as a Service delivers exactly that. Providers can expose LLM-ready inference endpoints that scale elastically across GPU clusters, powered by vLLM's optimized engine. With Hugging Face and OpenAI-compatible APIs, plus integrated governance and policy controls, operators deliver hyperscaler-grade inference services while retaining sovereignty and monetization.

Designed For

Cloud & Service Providers: Expand beyond compute by offering production-grade inference endpoints as a high-value service.

Enterprises: Accelerate application development by consuming governed, GPU-backed inference APIs with predictable performance.

Sovereign & Regional Clouds: Deliver compliant, in-region inference services with policy enforcement and auditability for regulated industries.

Key Capabilities

Capability	Description
Instant Deployment	Launch vLLM-based inference services in seconds via self-service, with no infrastructure setup required.
GPU-Optimized Inference	Leverage vLLM's memory-efficient architecture on GPU clusters with dynamic batching and offloading.
High-Performance Serving	Serve large models with low latency and high throughput using vLLM's optimized runtime engine.
Customizable & Scalable	Scale inference across GPUs and nodes with distributed vLLM, supporting Hugging Face and OpenAI-compatible APIs.

Business Outcomes

Outcome	Impact
Accelerated Monetization of GPUs	Expose inference endpoints as high-demand SKUs, increasing GPU ROI while reducing idle time.
Hyperscaler-Grade Tenant Experience	Deliver self-service APIs with elastic scaling, predictable performance, and ecosystem compatibility.
Sovereign & Enterprise Adoption	Provide compliant, in-region inference services with auditability and governance for regulated industries.