Rafay-Powered AI Token Factory

Turn GPU Infrastructure into Scalable AI Services

AI Token Factory extends the Rafay Platform to deliver AI services through APIs and token-metered consumption.

Production-ready AI APIs run on GPU infrastructure while maintaining governance, multi-tenancy, and operational control. Token-metered consumption provides visibility into usage and enables internal chargeback or monetization models.

From Infrastructure to AI Services

GPU infrastructure investment is accelerating to support AI workloads. Yet many environments still expose clusters, nodes, or infrastructure resources rather than consumable AI services.

Applications and developers increasingly expect:

- Self-service APIs they can call instantly
- Elastic scaling to handle unpredictable workloads
- Consumption models aligned with AI usage

AI Token Factory bridges this gap by enabling organizations to deliver AI capabilities as services rather than infrastructure. Through standardized inference APIs and token-metered consumption, teams can deploy, scale, and operate AI services while maintaining governance and platform control.

Designed For

Cloud & Service Providers: Expand beyond raw compute by offering production-grade AI APIs as a monetizable service platform.

Designed For (Continued)

Enterprises: Enable internal AI platforms where teams can consume governed AI services through standardized APIs.

Sovereign & Regional Clouds: Deliver compliant AI services with policy enforcement, auditability, and token-based consumption models.

Key Capabilities

Outcome	Impact
Self-Service AI APIs	Expose AI models as standardized inference APIs that developers and applications can consume directly.
Token-Metered Consumption	Measure AI usage at the token level to enable consumption visibility, chargeback models, and monetization.
Multi-Tenant AI Platforms	Operate shared AI infrastructure across teams or tenants with governance, isolation, and policy controls.
Elastic AI Service Scaling	Scale inference workloads dynamically across GPU clusters to support unpredictable AI demand.

Business Outcomes for AI Factories

Outcome	Impact
Accelerated Monetization of GPUs	Convert GPU infrastructure into revenue-generating AI services delivered through APIs.
AI Services Platform Experience	Deliver self-service AI APIs with predictable performance and ecosystem compatibility.
Enterprise & Sovereign AI Platforms	Enable compliant AI service delivery with governance, auditability, and operational control.