PaletteAI

Model as a Service with Spectro Cloud PaletteAI

Deploy pre-trained or custom AI models on demand. PaletteAI's Model as a Service lets platform teams offer governed, self-service access to models from Hugging Face and NVIDIA NIMs — so AI teams can go from selection to running inference in minutes, not weeks.

Why Model as a Service matters

Every organization investing in AI faces the same bottleneck: getting models into the hands of the people who need them.

Data scientists and ML engineers want fast, repeatable access to pre-trained models for experimentation, prototyping, and production inference. What they get instead is a queue — tickets for infrastructure, manual configuration, and weeks of lead time before a single prompt gets a response.

Model as a Service changes this. It turns model deployment into a governed, on-demand capability that AI teams can consume through a simple interface or API, without needing to understand the underlying GPU infrastructure, inference engines, or Kubernetes configuration.

For platform teams, it means offering a curated, approved catalog of models as a shared service — with the controls, quotas, and visibility needed to keep costs predictable and governance intact.

Common use cases

Chatbots and copilots: Deploy conversational AI models quickly for internal tools, customer-facing assistants, or developer copilots.

Semantic search: Stand up embedding and retrieval models to power knowledge bases, document search, and RAG pipelines.

Content generation: Provide teams with governed access to generative models for text, code, and image generation workflows.

Rapid prototyping: Let data science teams experiment with new models from the Hugging Face catalog without waiting for infrastructure provisioning.

Why it’s hard to get right

Deploying a model sounds simple. In practice, it involves selecting the right inference engine, matching it to compatible GPU hardware, configuring networking and storage, managing container images, and enforcing resource limits across teams. The challenges stack up fast:

- **Infrastructure compatibility:** Different models require different inference frameworks, GPU types, and driver versions. Mismatches lead to failed deployments, wasted cycles, and frustrated teams.
- **Manual configuration:** Without automation, every model deployment is a bespoke exercise. Platform teams become a bottleneck, fielding tickets for each new model request.
- **Resource contention:** Multiple teams competing for the same GPU pools without proper quotas or scheduling leads to monopolization, idle hardware, or both.
- **Governance gaps:** If anyone can deploy anything, you lose visibility and control. If nobody can deploy without approval, you lose speed. Finding the balance is the hard part.
- **Scaling complexity:** What works for one model on one cluster breaks when you need to serve dozens of models across teams, environments, and hardware configurations.

How PaletteAI delivers Model as a Service

PaletteAI makes model deployment a managed, repeatable capability within your AI platform. It integrates directly with Hugging Face and NVIDIA NGC, automatically matches models to the right inference engine and infrastructure, and gives AI teams a self-service path to deploy — all within the guardrails your platform team defines.

Capability	What it means
One-click model deployment	Browse and deploy models from Hugging Face or NVIDIA NIMs through a graphical interface or API. No manual infrastructure configuration required.
Automatic Profile Bundle matching	PaletteAI evaluates model attributes and automatically selects the right inference engine and infrastructure profile, removing guesswork from deployment
Built-in model catalog	Search and filter models by name, framework (e.g. vLLM, Ollama), popularity, or source. Your teams see what’s available and deploy what they need.
GPU quota enforcement	Tenant-level and project-level GPU quotas prevent over-allocation and ensure fair resource distribution across teams.
Inference engine validation	PaletteAI validates that compute infrastructure is compatible with the model being deployed, catching mismatches before they cause failures.
Full PaletteAI platform integration	Model as a Service is part of PaletteAI, so all governance, RBAC, multi-tenancy, and lifecycle management capabilities apply to model deployments. Users get personalized access to the model capabilities they need for their role.

How it works

Model as a Service in PaletteAI is built around a simple concept: platform teams define the rules, AI teams deploy within them. Here's the workflow:

1. Configure model integrations.

Platform teams connect PaletteAI to Hugging Face (via API token) and/or NVIDIA NGC (via API key) in the Project settings. This enables model browsing and deployment for the project.

2. Define Model as a Service mappings.

Mappings link a model source and optional filters to a specific Profile Bundle. Each Profile Bundle defines the inference engine, infrastructure configuration, and compute requirements. This is where platform teams encode their standards.

3. AI teams browse and select.

Users browse the integrated model catalog, search by name, filter by inference framework, and sort by popularity or recency. For Hugging Face models, frameworks like vLLM and Ollama are supported. For NVIDIA NIMs, users select directly from the NGC catalog.

4. PaletteAI matches and deploys.

When a user selects a model, PaletteAI evaluates the configured mappings and automatically selects the Profile Bundle that matches the model's source and attributes. The correct inference engine and infrastructure are applied without manual intervention.

5. Deploy to a Compute Pool.

The model is deployed to an existing Compute Pool in running status, or a new one can be created during the deployment workflow using an Infrastructure or Fullstack Profile Bundle.

Results you'll see

Faster time to value

AI teams deploy models in minutes instead of waiting days or weeks for infrastructure provisioning and configuration.

Less operational overhead

Automatic Profile Bundle selection and inference engine validation remove manual steps and reduce the risk of misconfiguration.

Better GPU utilization

GPU quotas and shared Compute Pools prevent over-allocation and ensure expensive hardware is used efficiently across teams.

Governance without friction

Platform teams define approved infrastructure and policies. AI teams deploy freely within those boundaries.

Why Spectro Cloud

Model as a Service is one piece of the PaletteAI platform. When you choose Spectro Cloud, you get the full picture — not a point solution.

- **Part of a complete AI platform:** Model as a Service works alongside GPU as a Service, training environments, and full-stack infrastructure management. One platform for the entire AI lifecycle.
- **Open ecosystem:** PaletteAI integrates with Hugging Face, NVIDIA NIMs, NeMo, Triton, Run:ai, ClearML, Kubeflow, and more. Your teams use what works best, not what a vendor mandates.
- **Enterprise-grade governance:** SSO, RBAC, multi-tenancy, namespace isolation, and resource quotas are built in. PaletteAI VerteX adds FIPS-140-3 compliance for regulated industries.
- **Proven at scale:** Spectro Cloud's architecture is trusted by some of the most demanding organizations in the world, from T-Mobile and GE HealthCare to the US Air Force and Airbus Defence and Space.
- **Trusted NVIDIA partner:** Spectro Cloud is an NVIDIA Preferred Partner, validated to deploy and manage infrastructure in accordance with NVIDIA's Enterprise AI Factory designs.

Ready to offer models as a service?

Talk to our team about how PaletteAI can help you deliver AI models on demand — with less waste, less friction, and the governance your organization requires. Schedule a demo at **spectrocloud.com/get-started**

Spectro Cloud