# FEDML Nexus AI
## Your GenAI Platform at Scale

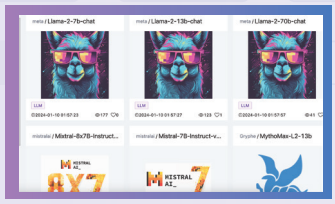
# FEDML Nexus AI - On Your Cloud

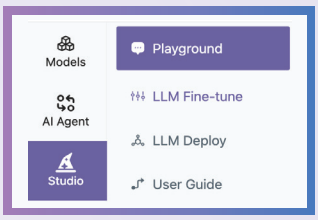
**Elevate your GPU Cloud into a cutting-edge Generative AI Cloud with FEDML's Nexus AI on your cloud.** This solution enables GPU cloud providers to create a bespoke generative AI platform, complete with custom domains and full ownership rights. Using this platform, providers can then offer expansive services to their users, including access to cutting-edge foundational GenAI models, and the ability to deploy, serve, fine-tune, and train these models, alongside developing advanced AI agents with them, all within their own GPU cloud. **FEDML Nexus AI** not only enriches your service portfolio but also aligns your platform with industry leaders like AWS's Bedrock, positioning you at the forefront of the GenAI evolution.

## Empowers you with key features:

### GenAI Model Hub with Playground and API Access

Start offering leading generative AI foundational models (like Llama, Mistral, Stable Diffusion) on your GPU cloud. FEDML simplifies your journey by managing the intricacies of API access, model deployment, serving, backend management, and payment processing. This enables you to concentrate on expanding your offerings, driving innovation and accelerating business growth.



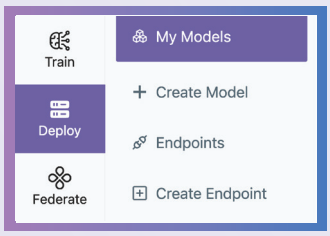


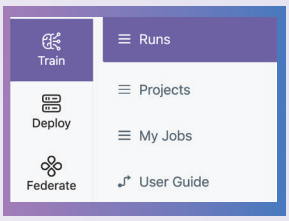

### GenAI Studio

Empower your users to create advanced AI agents using GenAI models from your model hub. GenAI studio offers a seamless RAG (Retrieval-Augmented Generation) pipeline and connectivity to external APIs, streamlining the development of customized agents. Additionally, GenAI Studio facilitates the deployment of models/agents on dedicated servers and enables users to fine-tune or train models with their private data, all within your GPU cloud infrastructure.

### FEDML Deploy

This suite allows your users to create dedicated or serverless end-points for their own proprietary models, or any model of their choice from Model Hub or repositories like HuggingFace. Users can further access advanced features including: (1) autoscaling to meet the traffic demand with any number of GPUs, ensuring dynamic and cost-effective scalability; (2) batch-job scheduling for operation of complex workflows or multi-model applications; (3) failover mechanisms for high SLA compliance and continuous service; and (4) comprehensive endpoint monitoring for insights into deployment logs, API usage, and interactions with their end-points.




### FEDML Train

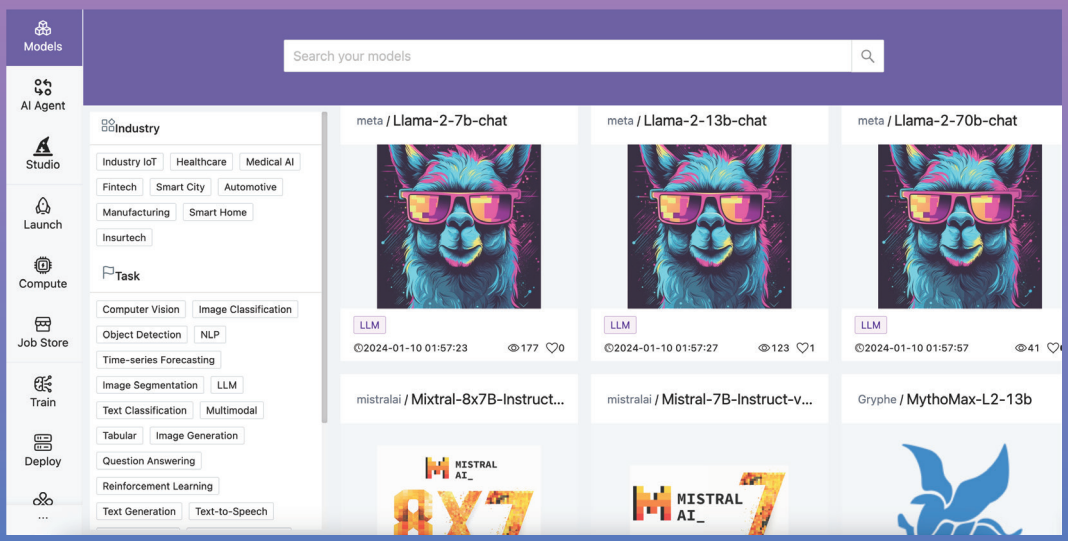
This part of the platform provides integrated support for advanced distributed training frameworks, such as Microsoft/DeepSpeed and Nvidia/Megatron-LM, for your users to launch scalable distributed model training on any number of GPUs in your cloud. It also provides advanced observability, system monitoring, and debugging tools for complex foundational model training jobs requiring many GPUs and several months of effort.

### Advanced Business and Finance Operations Dashboard

Gain insights into your operation with an advanced dashboard that tracks user activity, API model access, AI Agent and GenAI Studio usage statistics, and comprehensive billing and revenue management. This tool is designed to give you a clear overview of your business health, facilitating informed decision-making and strategic planning.

# FEDML Nexus AI On Your Cloud

## GenAI Model Hub with Playground and API Access




Start offering leading generative AI foundational models on your GPU cloud, and empower developers to use, deploy, fine-tune, and create AI Agents with them.

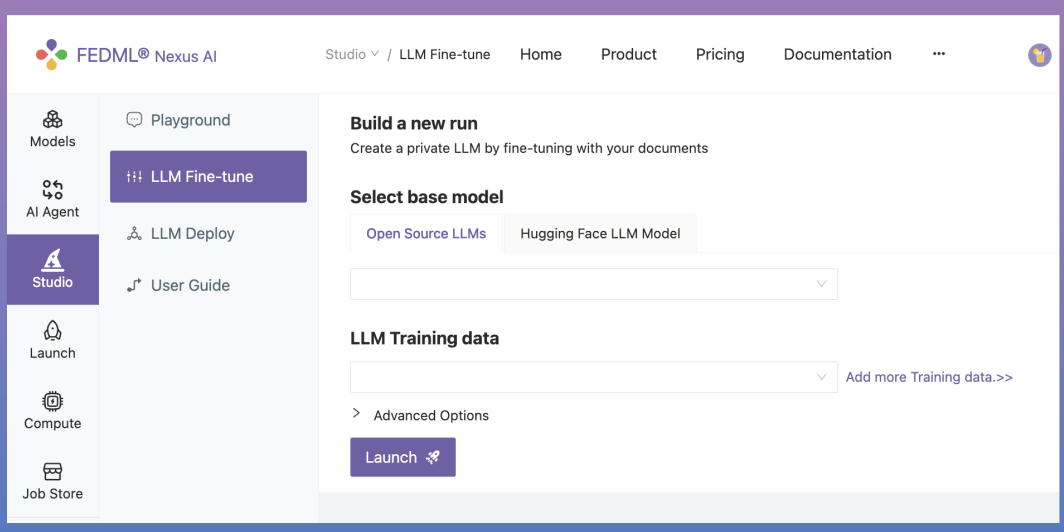
### Large Language Models (Text)
Access to all popular open-source models including Llama 2, Mistral, Mixtral, Gemma, etc.

### Image Models
Stable Diffusion, InstantID, LLava, CLIP

### Video Models
Stable Diffusion, Stable Diffusion XL, DiT

### Voice Models
Whisper, Canary, Wav2vec

## GenAI Studio




Empowers users to build advanced AI agents using your model hub, with a streamlined development pipeline, deployment options, and fine-tuning capabilities.

### LLM Fine-tune
Fine-tune your own LLM in 2 clicks by selecting a base model (via GenAI Model Hub or Hugging Face) and uploading your dataset.

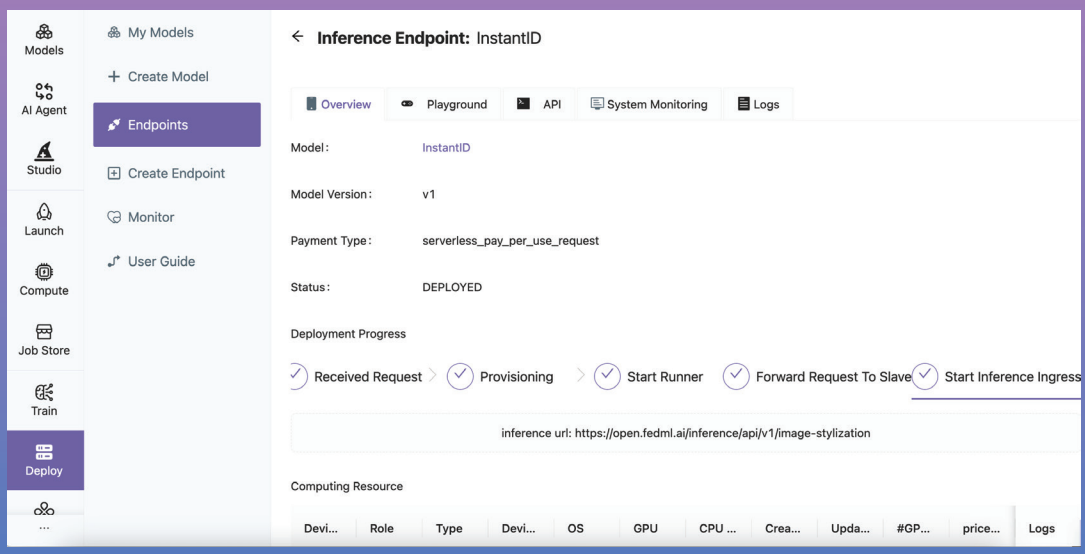
### LLM Deploy
Model deployment with full flexibility of dedicated or serverless GPUs, leveraging advanced features like autoscale for varying demands.

### AI Agents
Build custom AI agents using your models, simplifying development through an integrated pipeline and external API access.

### Playground
Interact and test with any model or agent before deployment.

# FEDML Nexus AI On Your Cloud

## Deploy




Enable dedicated and serverless deployment of any ML model (on Model Hub, HuggingFace, or private) on your cloud with advanced features like autoscaling, failover, and endpoint monitoring.

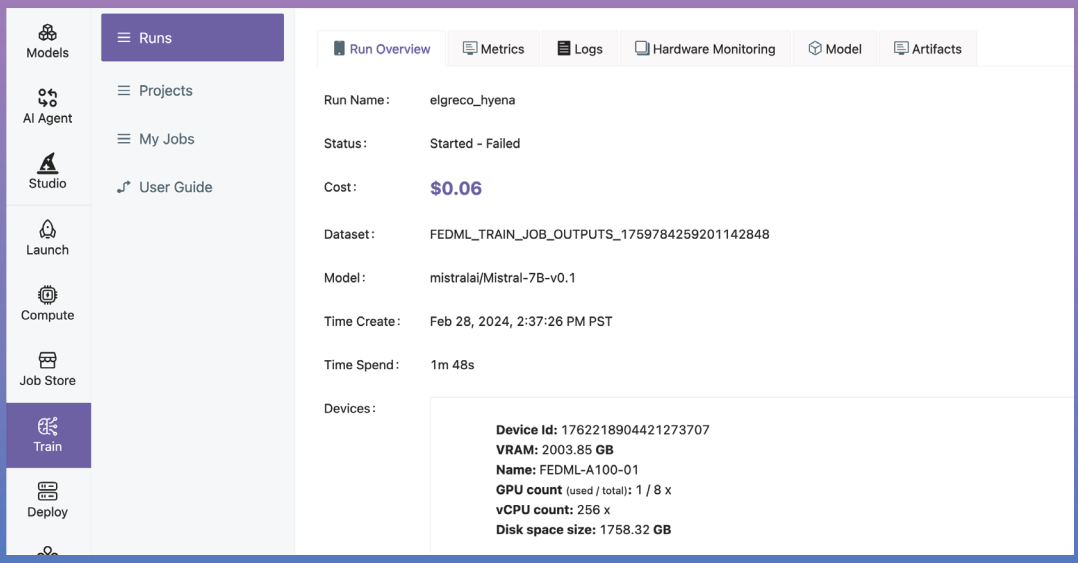
### My Models & Create Model
Easily import your own model to create model cards, including direct exports from Hugging Face and other libraries.

### Create Endpoints
Select model and version, name your endpoint, choose deployment type (dedicated/serverless), specify computing resources (public/private cloud, FEDML GPU cloud, on-premise, edge devices), set autoscale limi and deploy with one click.

### Endpoints
Dashboard for your deployed endpoints, includes a testing playground, API access information, and integrated system monitoring with logs.

### Monitor
Advanced monitoring dashboard with customizable metrics (average latency, queries per second, serving costs, etc. ) and deployment logs for status tracking and debugging.

## Train




Enable large-scale AI model training on your cloud with integrated support for popular frameworks, advanced monitoring, and debugging tools.

### Runs
Create and launch your distributed training job.

### System Monitoring
Statistics about GPU/CPU utilization, GPU/CPU memory allocated, etc.

### Observability
Collection of logs and artifact detect model performance, accuracy, error rate, etc. to monitor the performance.

### Model
Collection of all trained model files.

# Pricing

## Starter

### Contact us for pricing

**+ GenAI Model Hub with Playground and API Access**
- Access to popular GenAI models
- Playground for easy interaction
- Model APIs at your desired pricing served on your cloud

**+ GenAI Studio**
- Zero-code fine-tuning for a selected set of models in the Model Hub
- AI Agent creation with RAG and limited API integration
- Playground for AI Agents

**+ Deploy**
- Dedicated and serverless deployment on your cloud
- Basic monitoring dashboard (QPS, latency, logs)

**+ Train**
- Only supported in GenAI Studio

**+ Launch**
- Not supported

**+ Operations and Management**
- Payment and pricing management
- Basic operations management

**+ Support**
- 24/7 customer service through a dedicated Slack channel
- Limited on-site / on-zoom support
- Annual update for new features

## Advanced

### Contact us for pricing

**All features from Starter**

**+ GenAI Model Hub with Playground and API Access**
- Support for adding new GenAI models (2 months onboarding timeframe)
- Advanced playground for easy interaction and model comparison
- Optimized Model APIs

**+ GenAI Studio**
- Zero-code fine-tuning for any model in the Model Hub
- Optimized fine-tuning algorithms
- AI Agent creation with optimized RAG and customized API integration

**+ Deploy**
- Capability to import and deploy private models
- Autoscale to any number of GPUs
- Advanced monitoring dashboard

**+ Train**
- Support for popular distributed training libraries
- Support for launching distributed training jobs across any number of GPUs
- Experiment tracking, tracing APIs, and observability dashboard

**+ Launch**
- Not supported

**+ Operations and Management**
- Adding collaborators with team member management

**+ Support**
- Prioritized on-site / on-zoom support
- Quarterly update to new features

## Customized

### Contact us for pricing

**All features from Advanced**

**+ GenAI Model Hub with Playground and API Access**
- Support for adding new GenAI models (customized onboarding timeframe)
- Custom playground design
- Custom model APIs optimized for your cloud

**+ GenAI Studio**
- Zero-code fine-tuning for any model in the Model Hub
- Custom fine-tuning algorithms optimized for your cloud

**+ Deploy**
- Custom inference engine optimization
- Alert and audit trails for safety and compliance
- Advanced model monitoring to detect drift, and model refinement pipeline

**+ Train**
- Custom clusters & queue management
- Geo-distributed and federated training
- Custom dashboard and observability support

**+ Launch**
- Capability to launch any ML job
- Automatic queue management across jobs
- Autoscale & failover
- Resource monitoring dashboards

**+ Operations and Management**
- Payment and pricing management (custom integration with your own payment system)
- Custom operation management tools

**+ Support**
- Dedicated on-site / on-zoom support
- Prioritize feature requests in FEDML roadmap

## FEDML NexusAI
https://fedml.ai/

## Follow us
https://github.com/FedML-AI

FEDML®

**Contact us:**
contact@fedml.ai
165 University Avenue,
Palo Alto
California 94301