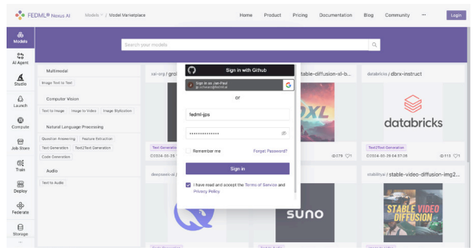
GenAI Platform



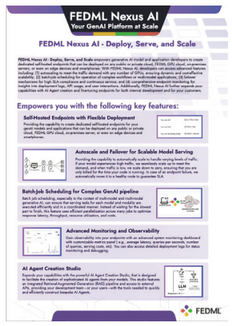
GenAI Cloud Platform for AI Developers

FEDML Nexus AI (https://fedml.ai/home) providing next-gen cloud service for LLMs and Generative AI. It helps developers or AI/ML teams to launch complex model training, deployment, and federated learning anywhere on decentralized GPUs, multi-clouds, edge servers, and smartphones easily, economically, and securely.

Starter	Advanced	Enterprise
$ 0/month	$ 199/month	Contact us

On-Premise Deployment for Enterprises

Deploying FEDML Nexus AI on-premise enables your company to operate your own Generative AI platform on your infrastructure, ensuring maximum privacy, ownership, and control. This solution provides a GenAI Model Hub and playground, a no-code GenAI Studio for model fine-tuning and AI Agent creation, and advanced MLOps for scalable model deployment and serving, model training, and federated learning. Leveraging FEDML Nexus AI on-premise, you can deliver state-of-the-art GenAI innovation to your employees and customers while maintaining full ownership and control.

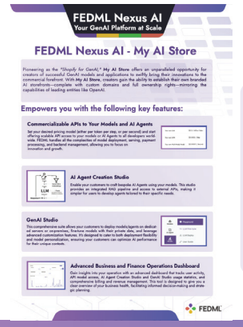


Deploy, Serve, and Scale for GenAI App Developers

FEDML Nexus AI - Deploy, Serve, and Scale empowers Generative AI model and application developers to create dedicated self-hosted endpoints that can be deployed on any public or private cloud, FEDML GPU cloud, on-premises servers, or even on edge devices and smartphones. With FEDML Nexus AI, developers can 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 user interactions. Additionally, FEDML Nexus AI further expands your capabilities with AI Agent creation and fine-tuning endpoints for both internal development and for your customers.

FEDML Nexus AI "On Your Cloud" for GPU Providers and AI Hardware Manurfacturers

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.



My AI Store for GenAI Model Developers

Pioneering as the "Shopify for GenAI," My AI Store offers an unparalleled opportunity for creators of successful GenAI models and applications to swiftly bring their innovations to the commercial forefront. With My AI Store, creators gain the ability to establish their own branded AI storefronts—complete with custom domains and full ownership rights—mirroring the capabilities of leading entities like OpenAI.

FEDML®

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

FEDML Nexus AI

Your GenAI Platform at Scale

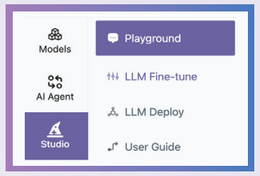
All-in-one foundations you need to build and commercialize your own generative AI applications easily, scalably, privately, and economically

Empowering you with the following key features:

GenAI Model Hub

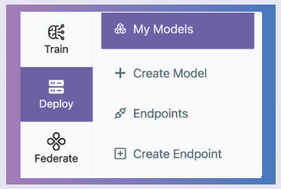
A hub for your company to safely access and use leading Generative AI foundational models (like Llama, Mistral, Stable Diffusion) on-premise. FEDML simplifies the journey by managing the intricacies of API access, model deployment, serving, and backend management. This enables your company to concentrate on expanding your GenAI offerings, driving innovation and accelerating productivity and business growth.

GenAI Studio



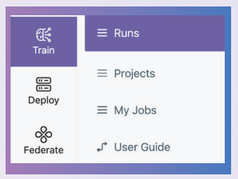
Empower your company to create advanced AI agents using GenAI models from your model hub. GenAI studio offers a seamless RAG (Retrieval-Augmented Generation) pipeline and connectivity to external APIs, streamlining the development of customized agents. Additionally, GenAI Studio facilitates the deployment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

Deploy, Serve, and Scale



This suite allows you to create dedicated or serverless end-points for your own proprietary models, or any model of your choice from Model Hub or repositories like HuggingFace. Further, you 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.

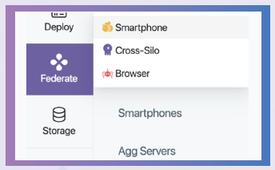
Train



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

Federate

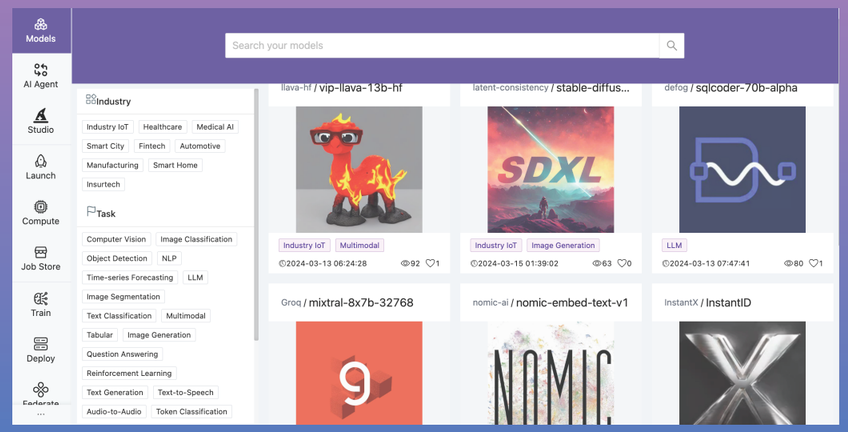
Enable zero-code, lightweight, cross-platform, and provably secure federated learning and analytics in your own infrastructure. It enables training of machine learning models from decentralized data at various users/silos/nodes without requiring data centralization, thus providing maximum privacy and efficiency.



FEDML®

FEDML Nexus AI

GenAI Model Hub with Playground and API Access



API access and playground for leading Generative AI foundational models. Further, empowering you to privately use and deploy them.

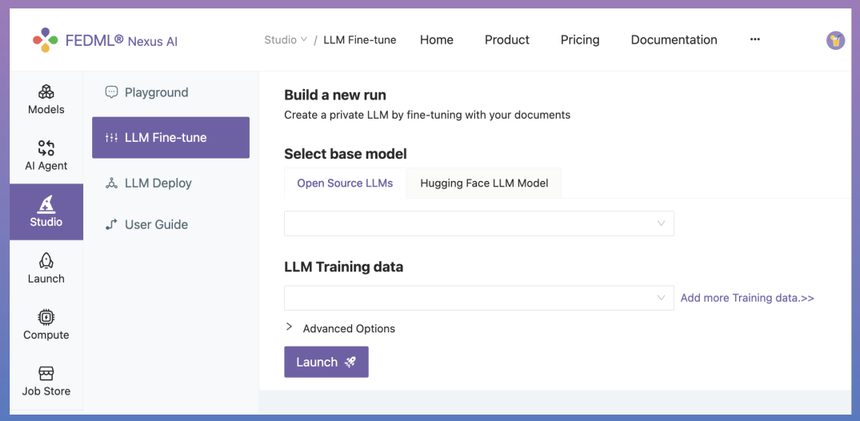
Large Language Models (Text)
Access to all popular open-source models including Llama 2, Mistral, Mixtral, Gemma, etc.

Video Models
Stable Diffusion, Stable Diffusion XL, DiT

Image Models
Stable Diffusion, InstantID, LLava, CLIP

Voice Models
Whisper, Canary, Wav2vec

GenAI Studio



GenAI Studio providing zero-code fine-tuning capabilities, as well as integrated support for building advanced AI agents with seamless RAG pipeline and connectivity to external APIs.

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.

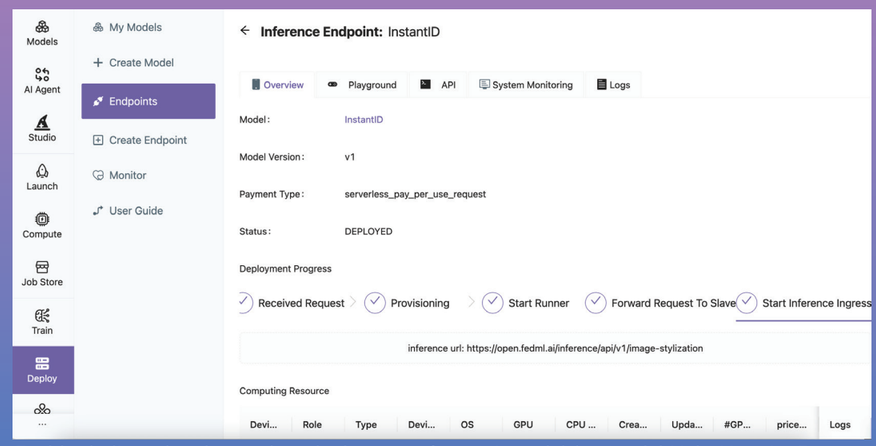
LLM Deploy
Model deployment with full flexibility of dedicated or serverless GPUs, leveraging advanced features like autoscale for varying demands.

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

Deploy



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

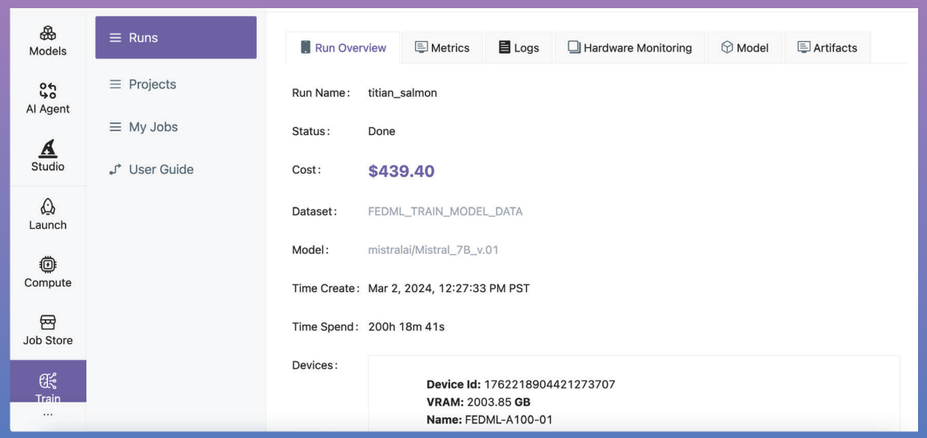
My Models & Create Model
Easily import your own model to create model cards, including direct exports from Hugging Face and other libraries.

Create Endpoints
Select model and version, name your endpoint, choose deployment type (dedicated/serverless), specify computing resources (private cloud, on-premise, edge devices), set autoscale limit 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 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.