

# Audi

GroupAI

November 2025

# What we'll discuss today

- ▶ [Motivational Speech](#)
- ▶ [Red Hat AI Inference Server Overview](#)
- ▶ [PoC takeaways](#)
- ▶ [What is next](#)
- ▶ [Next Steps](#)



<https://rhods-dashboard-redhat-ods-applications.apps.ocp.ocp-gm.de/projects/hello-world>

<https://maas.apps.prod.rhoai.rh-aiservices-bu.com/admin/applications/75149>





# What is what?

## Little tour on the buzzwords

### What is AI inference?

An AI inference server is the software that helps an AI model make the jump from training to operating. It uses machine learning to help the model apply what it's learned and put it into practice to generate inferences.

After successful training, the model can make inferences such as identifying a breed of dog, recognizing a cat's meow, or even delivering a warning around a spooked horse. Even though it has never seen these animals outside of an abstract data set before, the extensive data it was trained on allows it to make inferences in a new environment in real time.

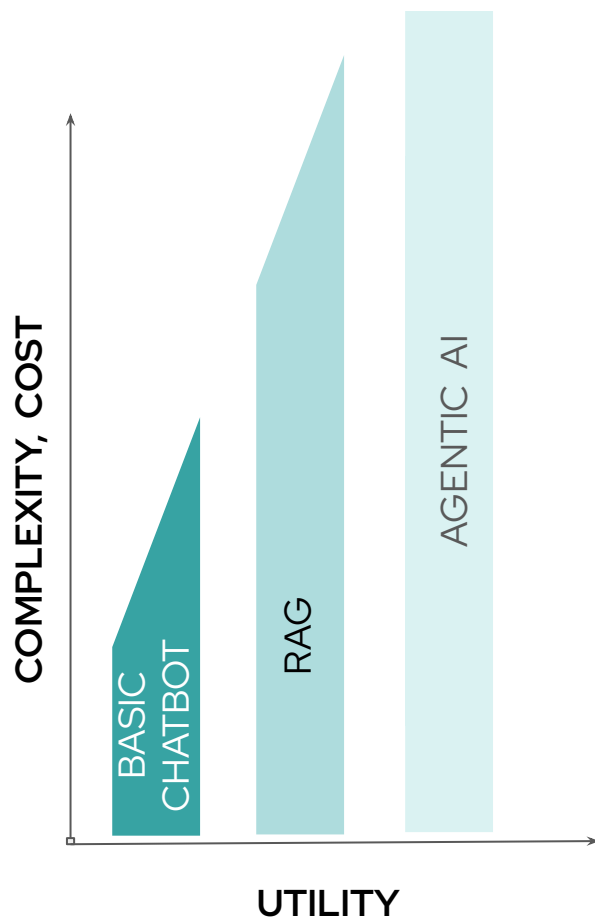
### What is vLLM?

vLLM, which stands for virtual large language model, is a library of open source code maintained by the [vLLM community](#). It helps [large language models \(LLMs\)](#) perform calculations more efficiently and at scale.

Specifically, vLLM is an [inference](#) server that speeds up the output of [generative AI](#) applications by making better use of the GPU memory.

# The Hidden Cost of Generative AI

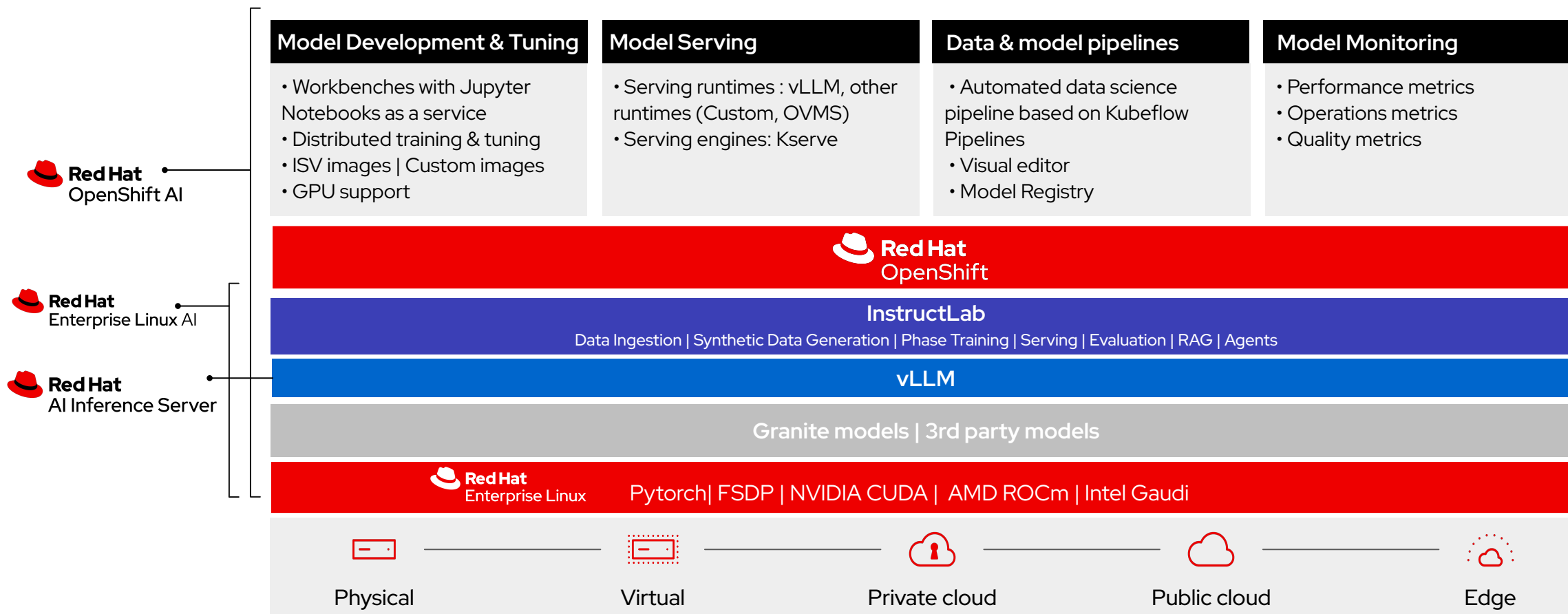
More Advanced Techniques Incur Higher Costs



- ▶ **Basic chatbots** that provide direct LLM querying are relatively cheap to operate but offer limited utility.
- ▶ **Retrieval augmented generation (RAG)** allows tapping into proprietary knowledge. The cost per query multiplies.
- ▶ **Agentic AI** provides arbitrary integration and query iteration capabilities. One user query may generate many high-cost LLM queries.

# Red Hat AI platform

Generative AI, Predictive AI & MLOps capabilities for building flexible, trusted AI solutions at scale



## Dashboard Application

Data Science Projects

Admin Features

Model Registry

### Model Development, Training & Tuning

#### Workbenches

- Minimal Python
- PyTorch
- CUDA
- Standard Data Science
- TensorFlow
- VS Code
- RStudio
- TrustyAI

CodeFlare SDK

ISV images

Custom images

#### Distributed workloads

KubeRay

Kueue

CodeFlare

#### Models

Granite Models

Ecosystem models

Data and model Pipelines

### Model Serving

#### Serving Engines

Kserve

ModelMesh

#### Serving Runtimes

OVMS

vLLM, Caikit/TGIS

Custom

### Model Monitoring

Performance metrics

Operations metrics

Quality metrics

Object  
Storage



OpenShift  
Operators

OpenShift  
GitOps



OpenShift  
Pipelines



OpenShift  
ServiceMesh



OpenShift  
Serverless



Prometheus

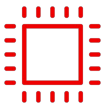


# OpenShift AI Feature Overview



## DataOps

- ▶ Connections
- ▶ S3 browser
- ▶ Feature store



## Hardware

- ▶ Hardware profiles
- ▶ Accelerators
  - Nvidia
  - Intel
  - AMD
  - IBM AIU (serving)
  - IBM Spyre AI (serving)
- ▶ CPU Architectures
  - x86
  - IBM Power & Z (serving)
  - ARM



## Model Development

- ▶ Workbenches
  - JupyterLab
  - VS Code
  - R Studio
  - LLM Compressor
- ▶ Data Science Pipelines
  - Experiment tracking
  - Pipeline versioning
  - Artifact tracking
- ▶ Distributed Workloads
  - Hyperparameter tuning
  - Kueue scheduling
  - KubeRay
  - Kubeflow Training Operator
  - LAB model customization

Retrain models



## Integrate models in app dev

- ▶ Model Serving Runtimes
  - OpenVINO Model Server
  - vLLM
  - TGIS
  - Caikit
  - Nvidia Triton\*
  - ML Server\*
  - Nvidia NIM\*\*
- ▶ Model Serving Modes
  - Single-model serving
  - Multi-model serving
  - Distributed serving
  - Disaggregated serving
  - Near edge deployments
  - Metrics-based autoscaling
- ▶ Agentic AI
  - Llama Stack
  - GenAI Studio



## Model monitoring and management

- ▶ OCI model support
- ▶ Model Registry
- ▶ Model Catalog
- ▶ Models-as-a-Service
- ▶ Monitoring
  - Operational metrics
  - Runtime metrics
  - Data drift detection
  - Bias detection
  - LM evaluation
  - LM evaluation UI
  - LLM guardrails

Preview  
Roadmap

This is an uncommitted roadmap and Red Hat reserves the right to change it.

\* certified and tested runtime, self-support  
\*\* requires Nvidia license



# Red Hat and [kubiya.ai](https://kubiya.ai) - better together

- **Faster Time-to-Production**

Kubiya automates pipelines and workflows; Red Hat OpenShift AI provides scalable model serving – helping enterprises move from idea to deployment in minutes.

- **Execution & Operations**

Kubiya adds an execution layer with RBAC, approvals, audit, and rollback; Red Hat handles model lifecycle and inference – making Day-2 operations safe and predictable.

- **Leverage Existing Investments**

Kubiya integrates with existing infrastructure, processes, and tools; Red Hat provides the enterprise hybrid-cloud backbone – avoiding new silos and extra vendor costs.

- **Enterprise Adoption**

Kubiya ensures secure, compliant workflows; Red Hat provides hardened infrastructure – reducing risk and cost at scale.

- **Joint GTM & Velocity**

Kubiya drives usage through workflow automation; Red Hat amplifies reach with OpenShift – together accelerating adoption and market impact.

Together: Kubiya + Red Hat deliver production-ready AI on OpenShift – faster, safer, and with accelerated customer adoption. **With Red Hat OpenShift AI as a core component**



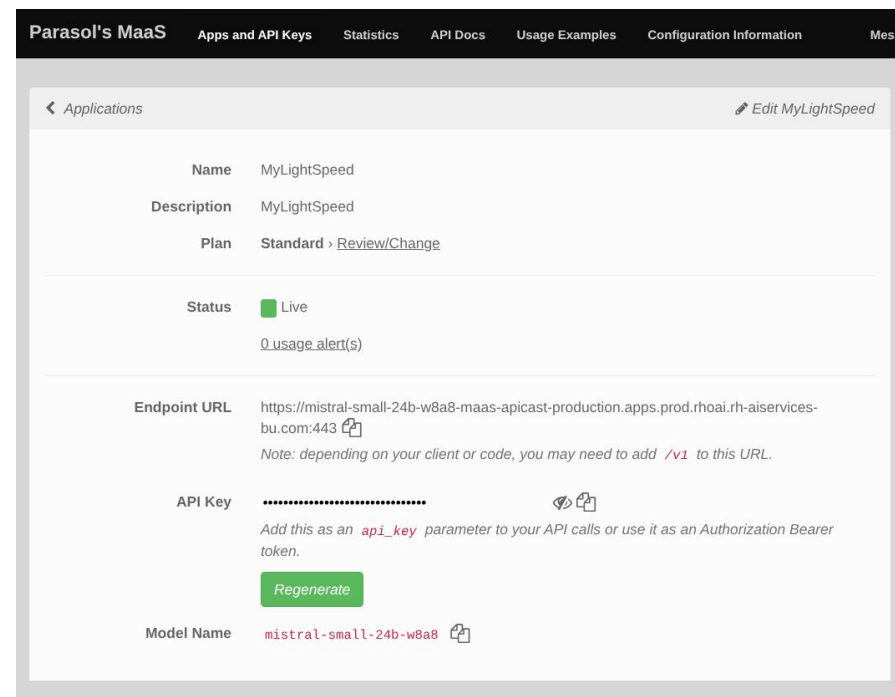
# Flavors of RHOAI

Supported deployment options		
Options available	Self-managed RHOAI	Cloud Service RHOAI
Bare metal	✓	
Virtual	✓	
Private cloud	✓	
Red Hat OpenShift on AWS (ROSA)	✓	✓
Azure Red Hat OpenShift (ARO)	✓	(future)
IBM Cloud	✓	
OSD-GCP/OSD-AWS	✓	✓
Edge (with SNO)	✓	

GroupAI

## Example which we use ourselves

<https://maas.apps.prod.rhoai.rh-aiservices-bu.com/admin/applications/75149>





# What we'll discuss today

- ▶ [Motivational Speech](#)
- ▶ [Red Hat AI Inference Server Overview](#)
- ▶ [PoC takeaways](#)
- ▶ [What is next](#)
- ▶ [Next Steps](#)

# What is what?

## Little tour on the buzzwords

### What are parameters

- Model Parameters: These are learned during training and include **weights and biases**. They are specific to the model's architecture and are adjusted based on the training data.
- Hyperparameters: These are set before training and influence the training process itself, such as the **learning rate or the number of training epochs**. Unlike model parameters, hyperparameters are **not learned from the data** but are manually configured.

[Data Poisoning bei LLMs: Feste Zahl Gift-Dokumente reicht für Angriff | heise online](#)

[...]

**Sofern sich die Ergebnisse bestätigen**, wäre die Ansicht, dass das Vergiften von KI-Daten wie "ins Meer pinkeln" sei, **wissenschaftlich widerlegt**.

Ein einzelner Akteur benötigt keine riesigen Ressourcen, um Schaden anzurichten

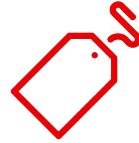
[...]



(Bild: busliq/Shutterstock.com)

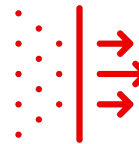


# The Operational Challenges in the Inference Era



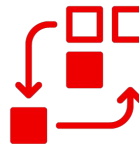
## Infrastructure cost

Requires substantial compute power to deliver expected experience



## Operational complexities

Non- standardized approach creates inefficiencies



## Deployment constraints

Inference across hybrid environments can lack flexibility



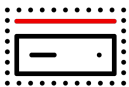


Trusted, Consistent and Comprehensive foundation

 **NVIDIA**  **AMD**  **intel** Hardware Acceleration  **aws**  **Google**  **IBM**



**Physical**



**Virtual**



**Private  
Cloud**



**Public  
Cloud**



**Edge**

\* NVIDIA GPUs fully supported in Red Hat AI. AMD Instruct & Intel Gaudi in Preview, AWS Inferentia/Neuron, Google TPU, IBM AIU are on our roadmap

**Fast**

**Cost-effective**

**Optimized**



# **Red Hat AI Inference Server**

**Any gen AI model**

**Any AI Accelerator**

**Any environment**

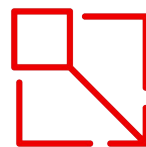
# Red Hat AI Inference Server

Gain consistent, fast and cost-effective inference at scale



## Inference runtime for the hybrid cloud

Run your models of choice across any accelerator and any environment



## Compress Models

Reduce compute and costs while preserving accuracy



## Red Hat AI Hugging Face repository

Access a collection of third-party validated and optimized models ready for inference.

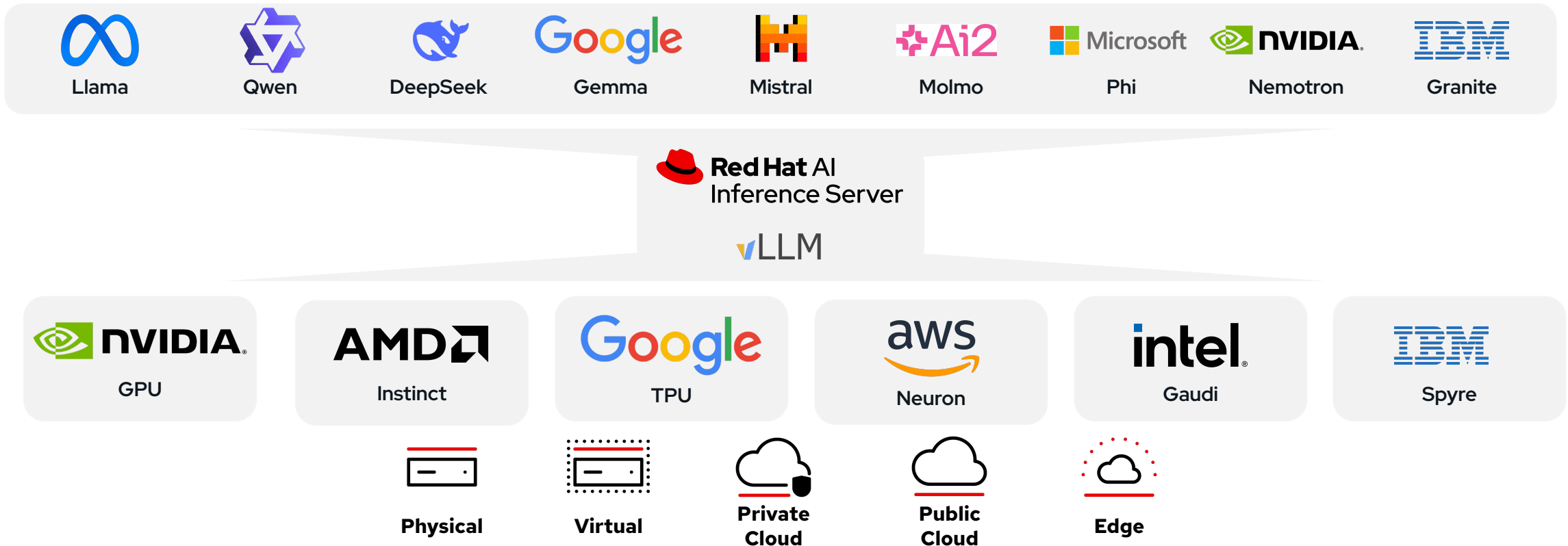


## Certified for all Red Hat products

Deployable across non-Red Hat Linux and Kubernetes platforms

# Red Hat AI Inference Server

vLLM connects model creators to accelerated hardware providers



Single platform to run any model, on any accelerator, on any cloud



# Red Hat AI Inference Server

vLLM is emerging as the Linux of GenAI Inference

## HIGH PERFORMANCE

- Advanced algorithms for high QPS serving
- Single server/GPU to distributed/multi GPU
- Already comparable to Nvidia (TRT-LLM)

## EASY TO USE CAPABILITIES DRIVING DEVELOPER AND IT PRODUCTIVITY

- Native Hugging Face integration
- Simple APIs for online and offline inference
- OpenAI-compatible API protocol

Scalable inference across the hybrid cloud

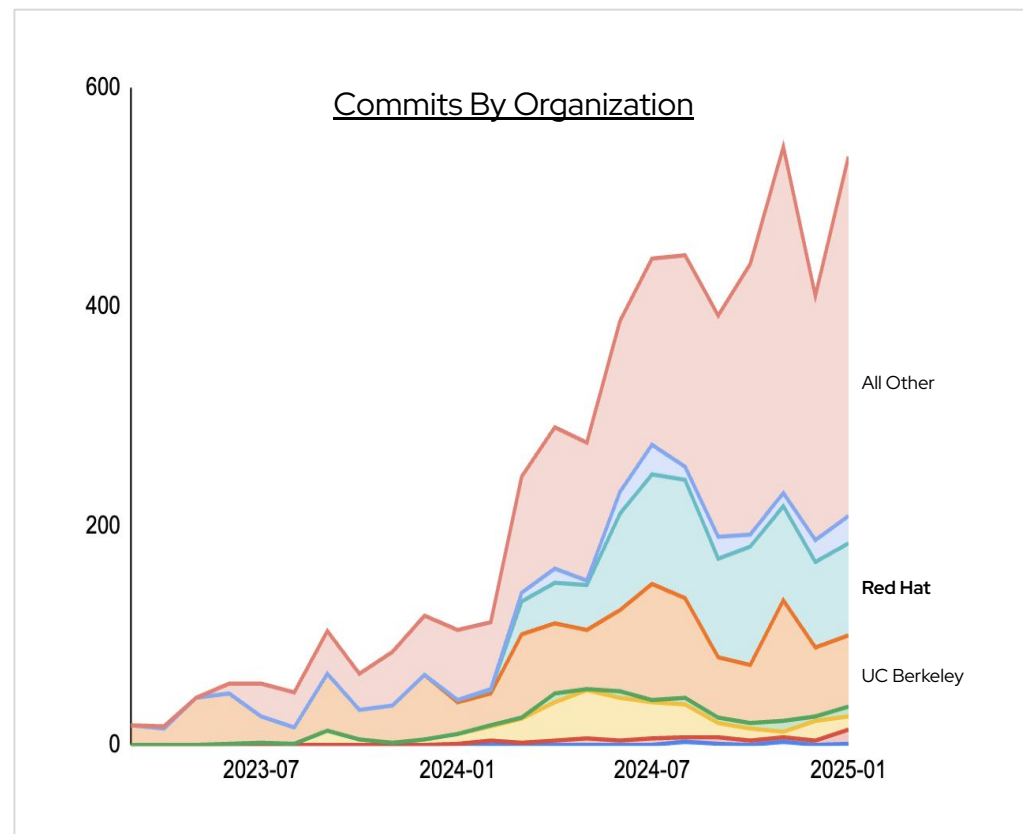
# Red Hat: Leaders in OSS GenAI Inference

Expertise across high performance inference and SOTA model optimizations

## Core Developers of vLLM

- HPC engineering team dedicated to vLLM, 7 core vLLM committers on staff
- Work on key subsystems, with a particular emphasis on fast model execution
- ML engineering team builds vLLM's optimization library llm-compressor
- ML research team create pre-optimized models for deployment with vLLM

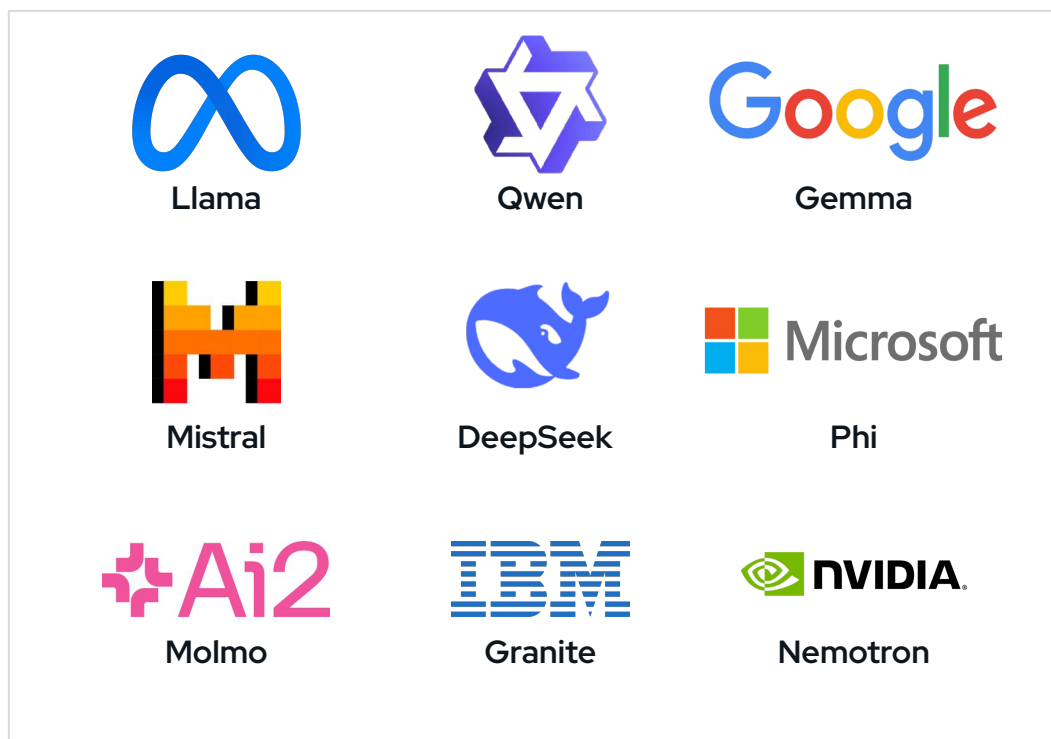
## Red Hat Community Contribution



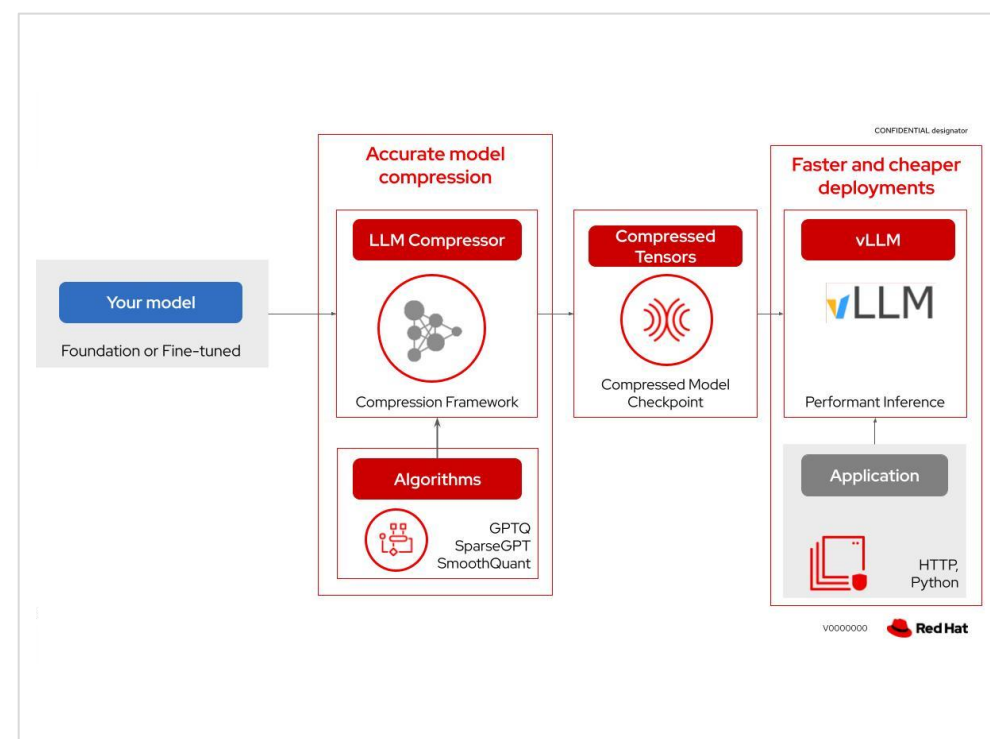
# Red Hat: Leaders in OSS GenAI Inference

Red Hat has built a comprehensive set of model optimization capabilities to drive operational efficiencies

## Third-party validated and optimized models



## LLM Compression Tools

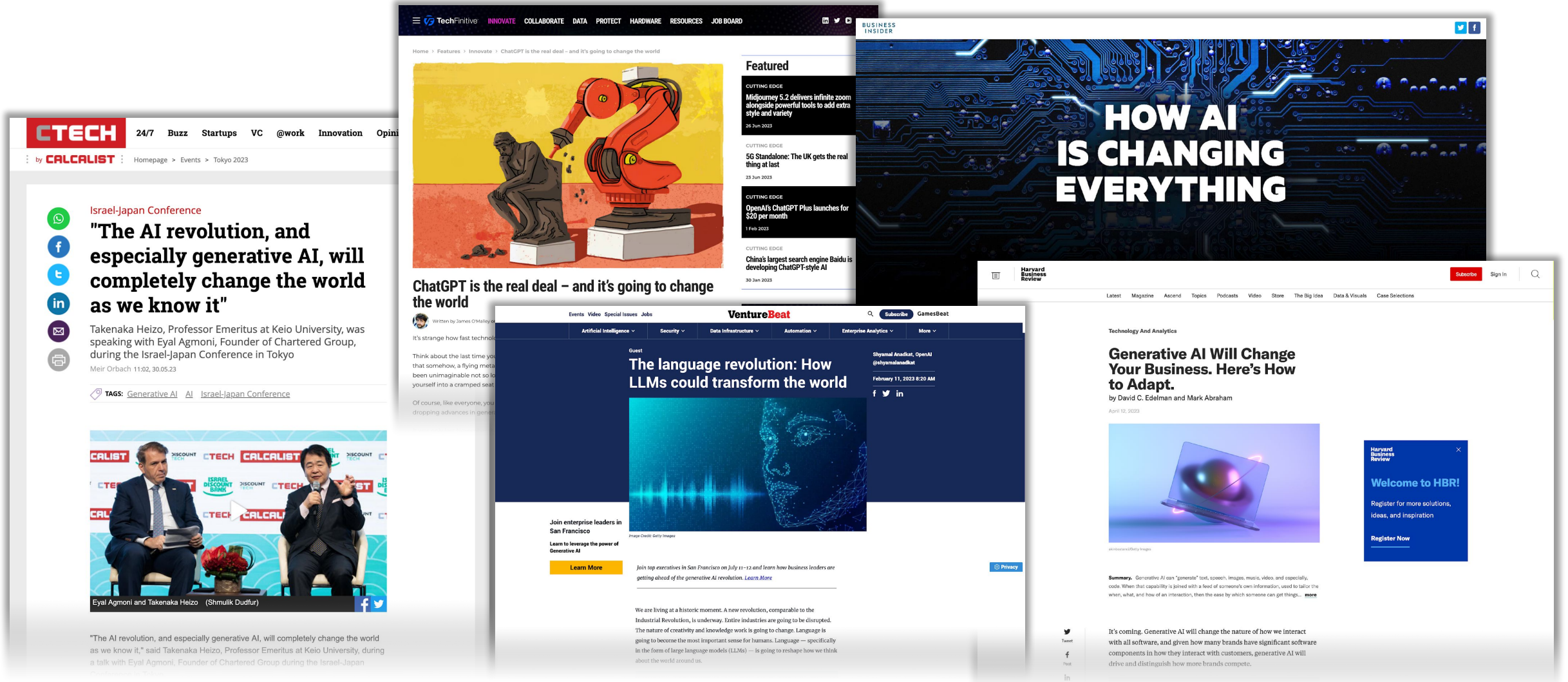


Hosted on the [Red Hat AI repository on Hugging Face](#)



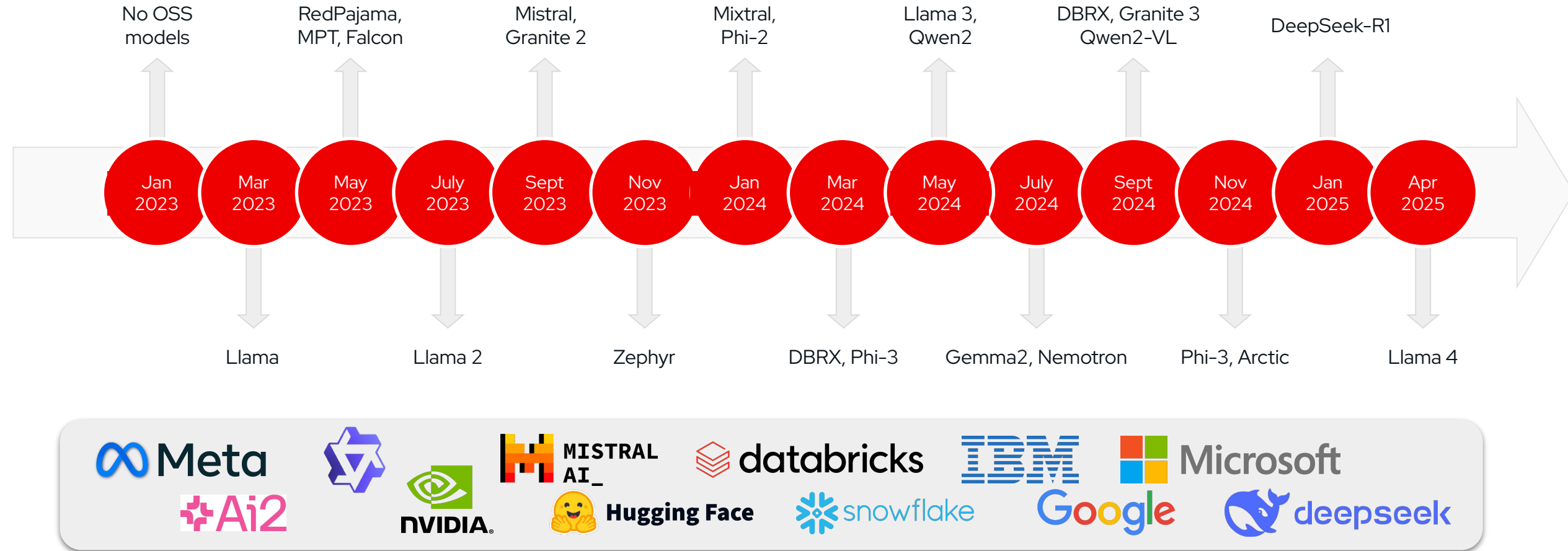
# The World Changed in November 2022

ChatGPT woke the world up to the power of generative AI



# The Power of Open

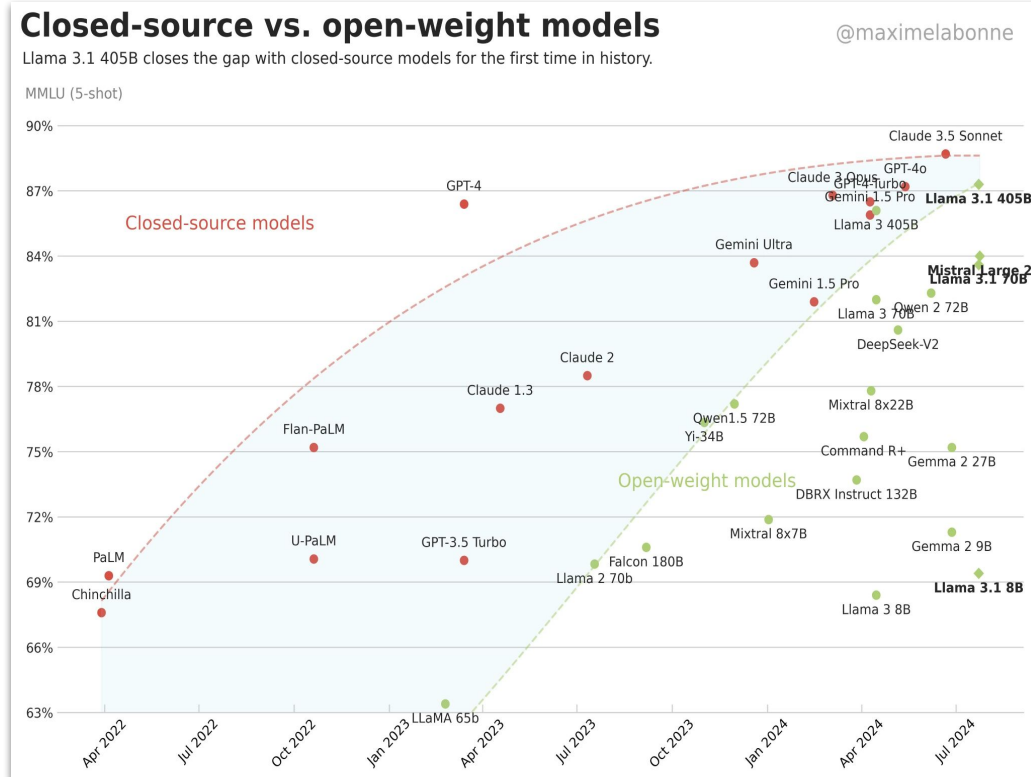
There has been an explosion of capability from open-source over the last 2 years.



# The Power of Open

Open models are deployment targets today. And the trend is not slowing down.

## Headlines



Llama

- 650M downloads in 2024
- 85,000 Llama derivative models
- 1B, 3B, 8B, 70B, 405B variants
- Multilingual, Multimodal, Mobile



R1

- First reasoning model on par in quality with OpenAI O1
- 1-70B parameter distilled versions
- Global market pandemonium?

Models are commoditizing → many options for diverse enterprise needs.



# Advantages of Open Source Models

Open-source models play an important role in the Enterprise AI landscape



## Customization

Improve accuracy and costs with task specific tuning



## Security

Complete data privacy (no 3rd party APIs)



## Control

Model lifecycle (no changes to the model in place) and Resources (no rate limits / API downtime)



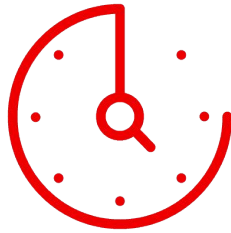
## Cost

Self managed infrastructure. 1B-405B size - match task difficulty to model





# Inference is becoming the gravity in AI because it is where the real world value happens



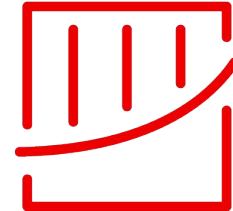
## Always-on intelligence

Inference powers every user interaction 24/7, making inference the constant cost and performance driver



## Latency defines experience

Low-latency and fast inference ensures optimal user experiences, essential for real-time applications and user retention

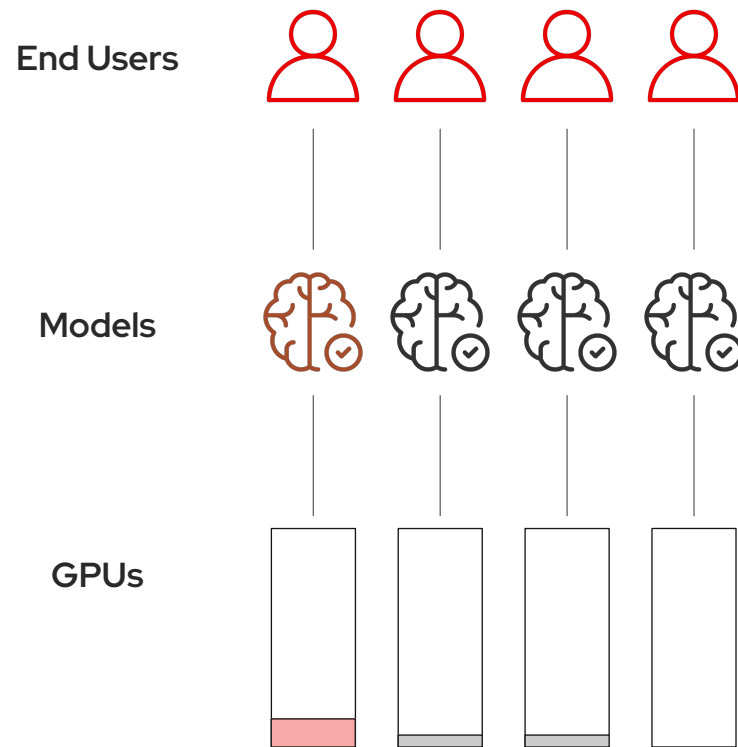


## Exponential market momentum

The AI inference market is forecast to grow from USD 106.15 billion in 2025 to USD 254.98 billion by 2030, underscoring its role



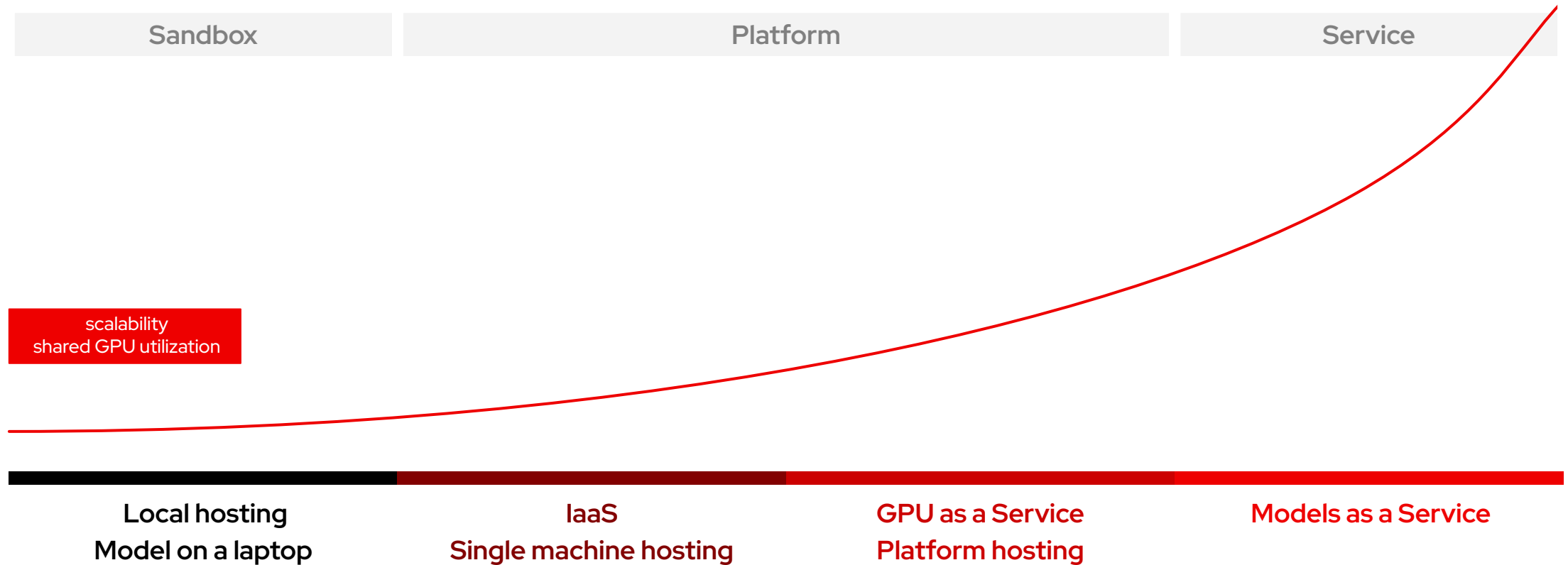
# Infrastructure as a Service can be costly



Self-Service AI could be good for small teams with ample resource, but can become risky and costly for wider audiences.

Most users primarily need an LLM endpoint, not simply direct GPU access.

# Model Hosting Stages



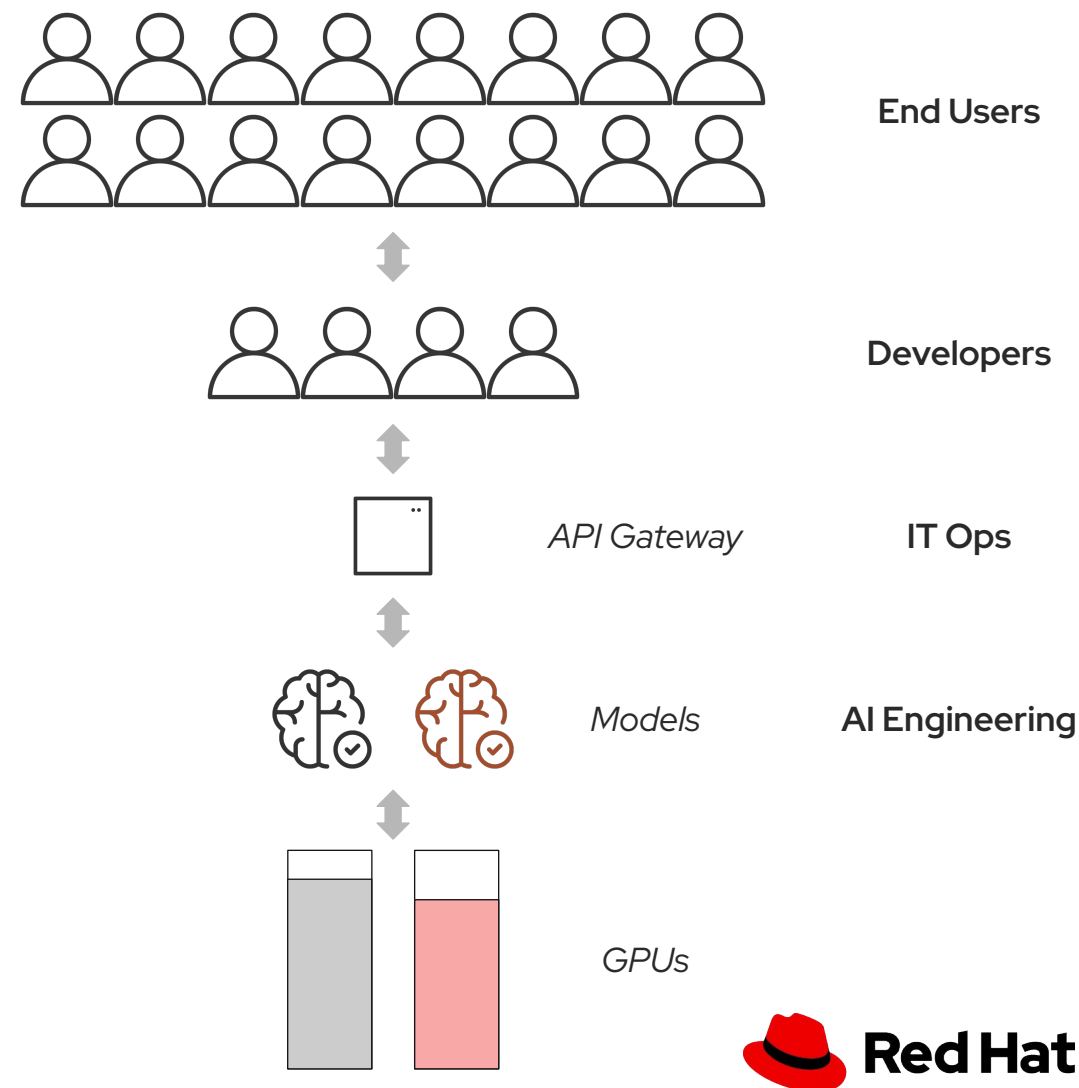
# What we'll discuss today

- ▶ [Motivational Speech](#)
- ▶ [Red Hat AI Inference Server Overview](#)
- ▶ [PoC takeaways](#)
- ▶ [What is next](#)
- ▶ [Next Steps](#)

# Introducing a Private-LLM-as-a-Service Pattern

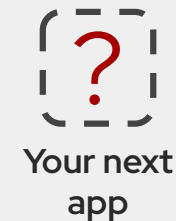
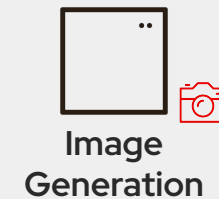
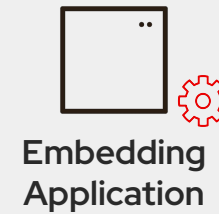
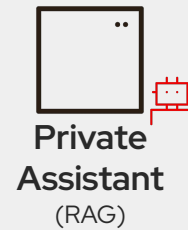
## Centralized AI Model Service for Broad Accessibility

- IT centrally manages and governs common models, ensuring compliance and efficiency
- Models available through an API Gateway
- Developers consume models and build AI applications
- Shared Resources business model keeps costs down by optimizing GPU utilization

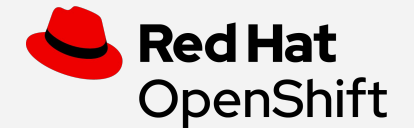
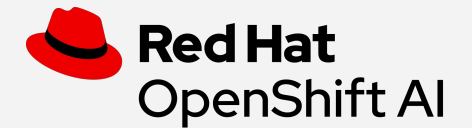


# Become the **Private AI Provider** for your organization

## Use cases



## Red Hat can help



+ **API Gateway Partners**

# Models as a Service – Requirements



As a **developer** I want to

- explore a catalog of available models with relevant descriptions and metadata,
- subscribe to individual models and receive credentials for programmatic access,
- rely on high quality of service.



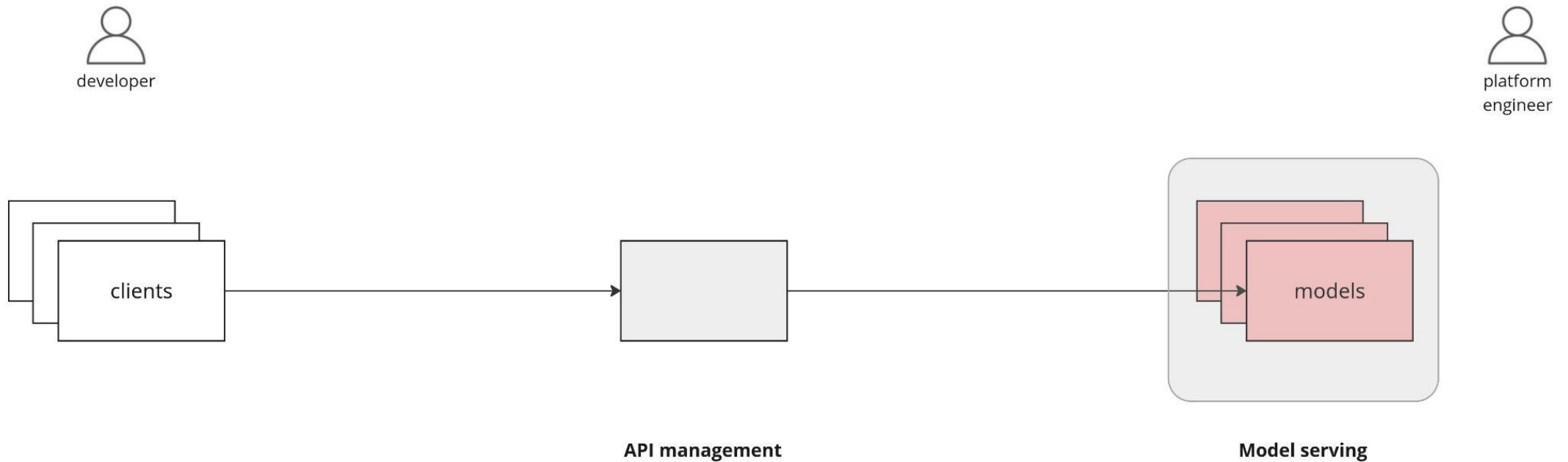
As a **platform engineer** I want to

- track usage of individual model endpoints and create reports for billing,
- enforce rate limits if needed,
- leverage existing identity providers for authentication.

# Models as a Service – Architecture

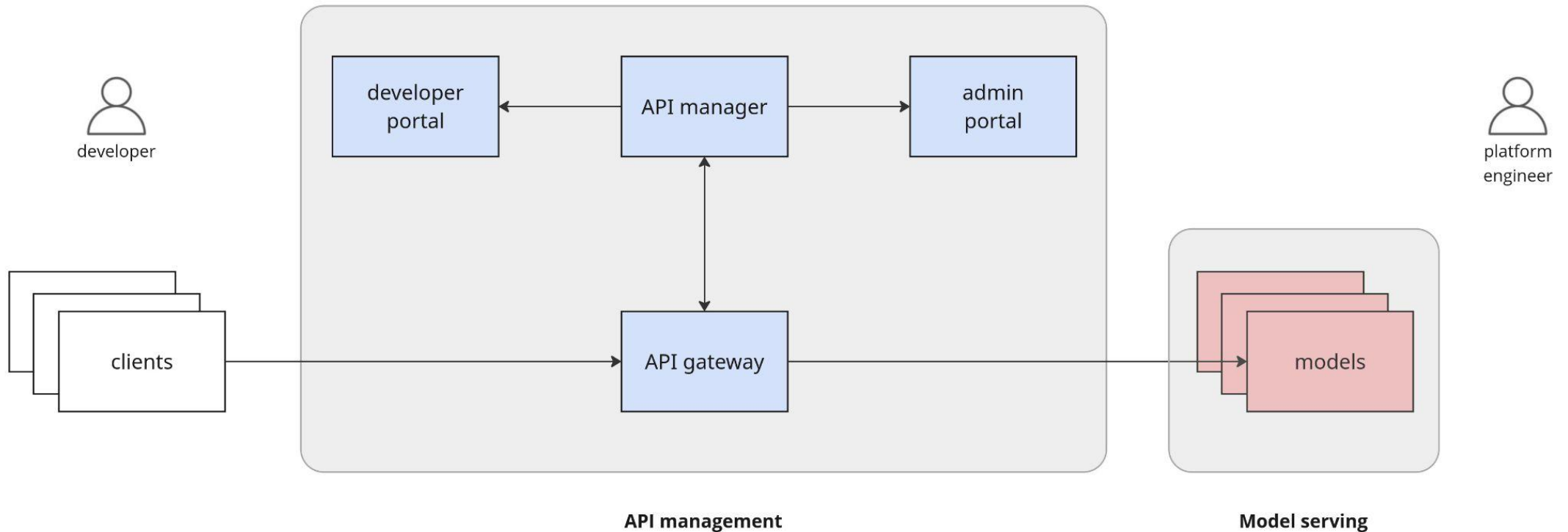


# Models as a Service – Architecture

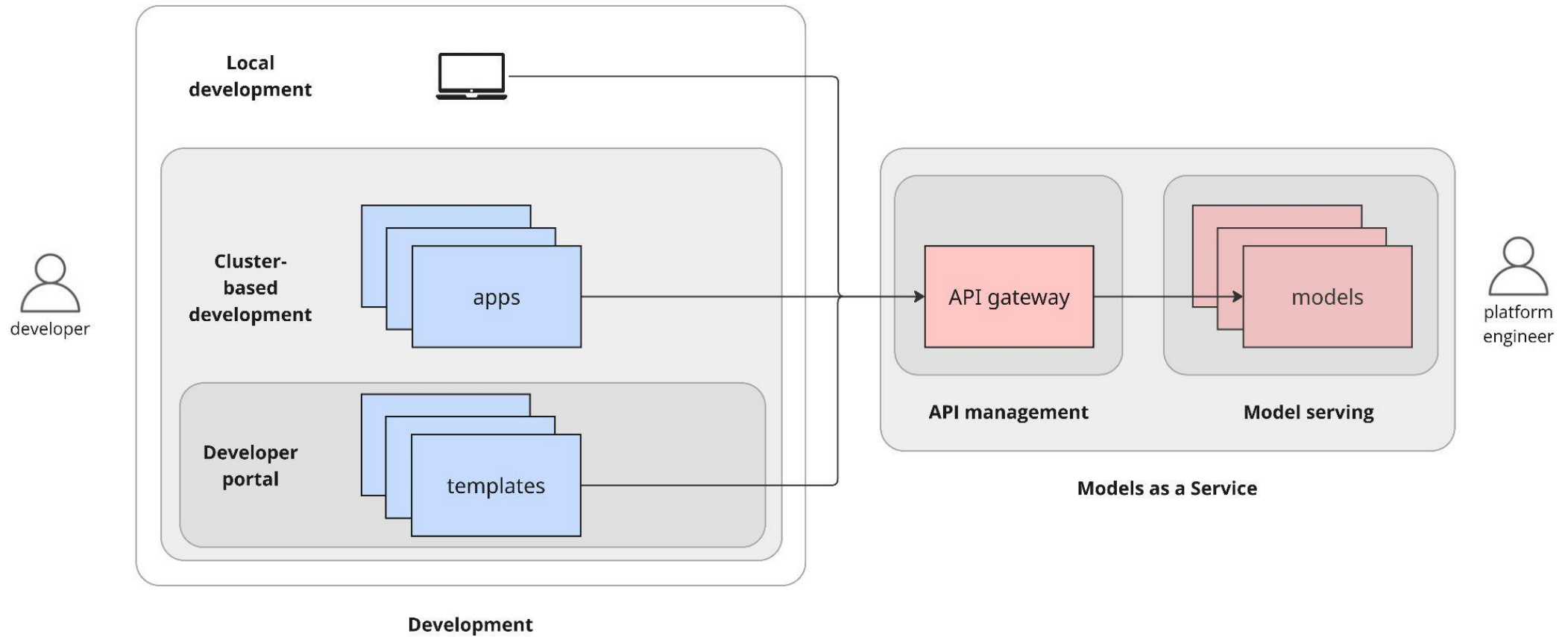




# Models as a Service – Architecture



# Models as a Service & Developers

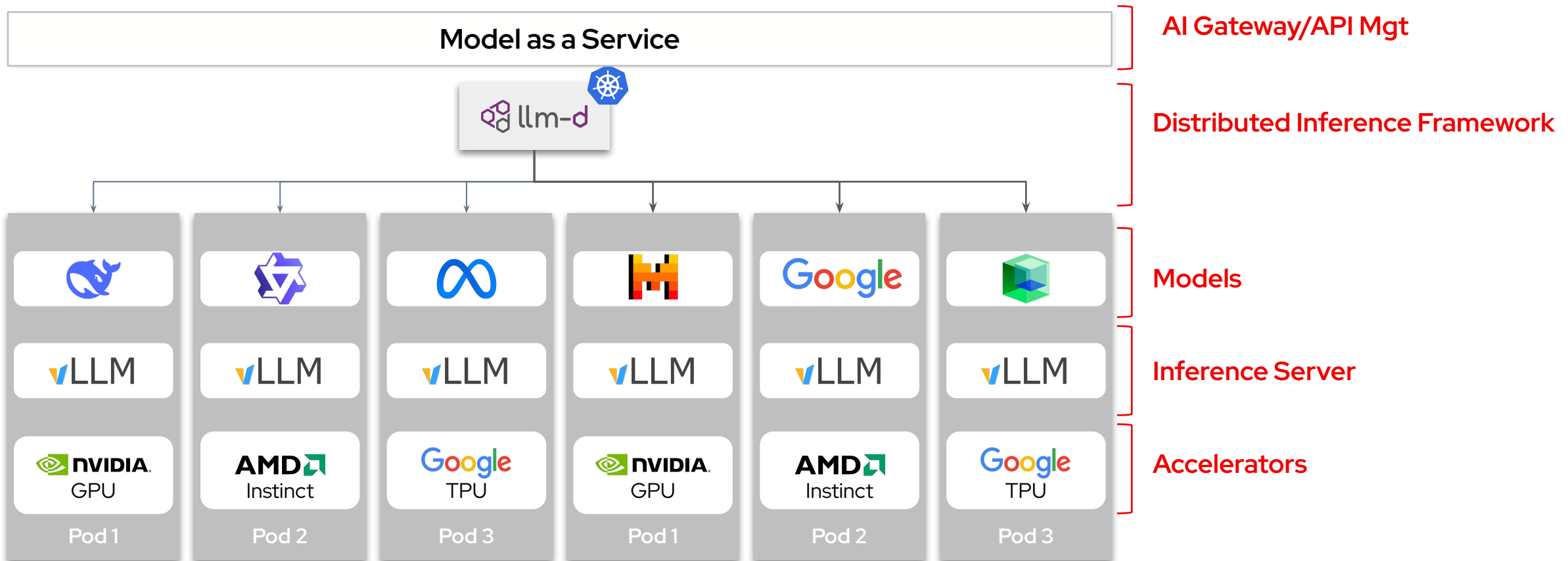


# What we'll discuss today

- ▶ [Motivational Speech](#)
- ▶ [Red Hat AI Inference Server Overview](#)
- ▶ [PoC takeaways](#)
- ▶ [What is next](#)
- ▶ [Next Steps](#)

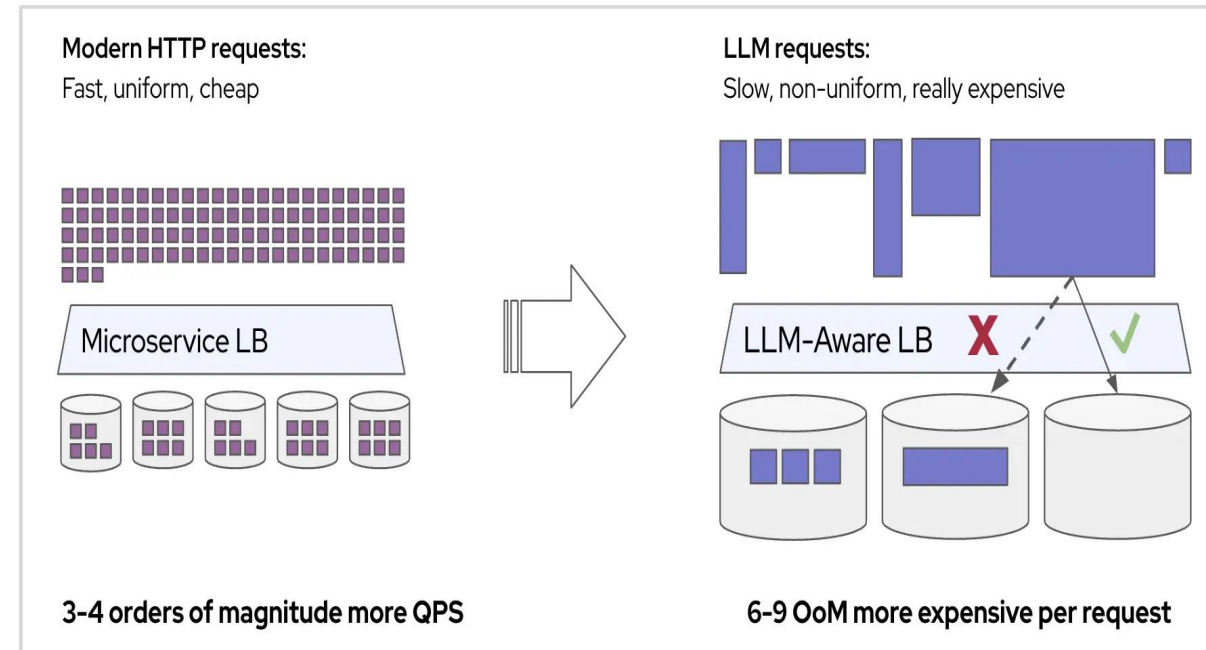
# Enterprise GenAI inference platform

Holistic approach to optimize and operationalize deployment and scaling of open-source LLMs



# Distributed Inference is essential, but introduces unique challenges

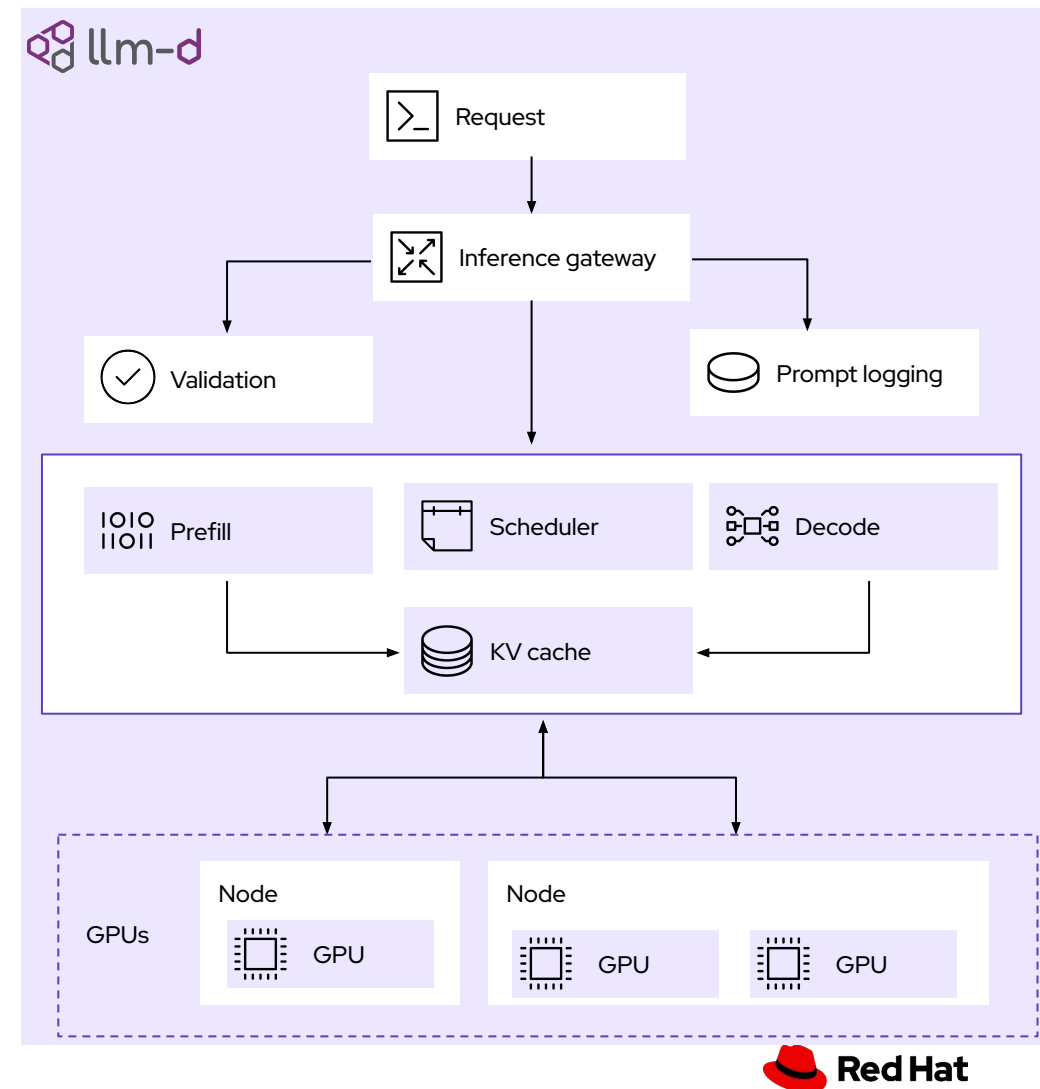
- LLM inference workloads break under traditional Kubernetes scaling
- **Ensuring SLO** (throughput, TTFT, latency) while **minimizing** resource utilization and operational complexity
- Leveraging and managing **heterogeneous hardware** for better cost-efficiency
- Low Inference efficiency



# Distributed Inference with llm-d

Maximize GPU utilization and deliver on your SLOs with distributed inference

- Joint open source project by Red Hat, Google, NVIDIA, AMD, Hugging Face, and many more
- Kubernetes-Native Architecture for simple deployment and management of GenAI models
- Optimized GenAI Inference to accelerate LLM's and MoE
- Intelligent Resource Utilization to reduce inference costs
- High Performance and Scalability to meet demanding Service Level Objectives (SLOs).
- Supported on Heterogeneous Hardware like NVIDIA and AMD GPUs (and many more to come in the future)

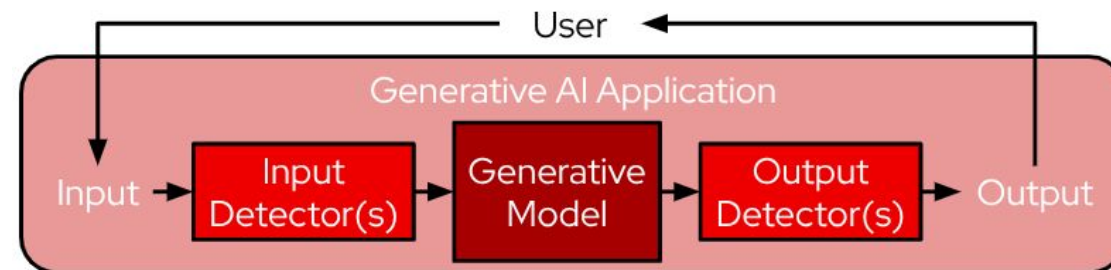


# Guardrails for Generative AI in Red Hat AI

- ▶ Ensure **secure, compliant, and efficient** AI operations with

These key features:

- **Customizable Input and Output Validators:** Tailor the AI's behavior to meet your business needs
  - **Request-Time Configuration:** Dynamically apply guardrails on a per-request basis
  - **Role-Specific Detection:** Design targeted validation pathways for different user groups
- 
- ▶ **Protect customer's Brand:** Prevent mentions of competitors to maintain focus on your products.
  - ▶ **Minimize Risk:** Restrict contract creation and negotiation to human oversight.
  - ▶ **Enhance Customer Experience:** Provide role-specific, meaningful interactions for every user group.
  - ▶ **Boost Efficiency:** Redirect technical queries to systems equipped with actionable insights for mechanics.



PS: signing and securing model artifacts is part of the Model Registry's OCI-compliance storage provided by RHOAI

# What we'll discuss today

- ▶ [Motivational Speech](#)
- ▶ [Red Hat AI Inference Server Overview](#)
- ▶ [PoC takeaways](#)
- ▶ [What is next](#)
- ▶ [Next Steps](#)



# Next steps

Some options to discuss

- Probe Red Hat OpenShift AI in ROSA
- Probe [kubiya.ai](https://kubiya.ai) on ROSA
- Probe Red Hat OpenShift AI on-prem with [kubiya.ai](https://kubiya.ai) to on-prem
- Probe ArgoCD in use cases of choice



# Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.



[linkedin.com/company/red-hat](https://linkedin.com/company/red-hat)



[youtube.com/user/RedHatVideos](https://youtube.com/user/RedHatVideos)

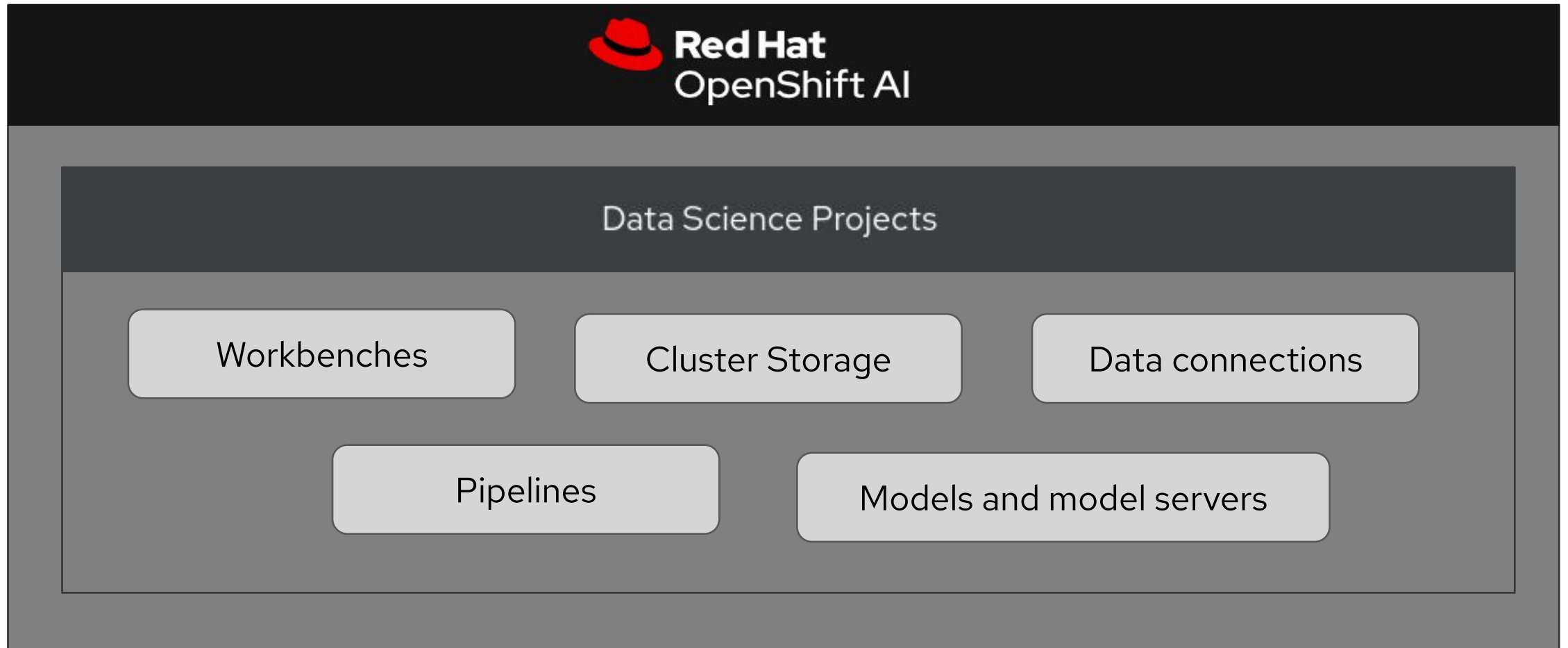


[facebook.com/redhatinc](https://facebook.com/redhatinc)

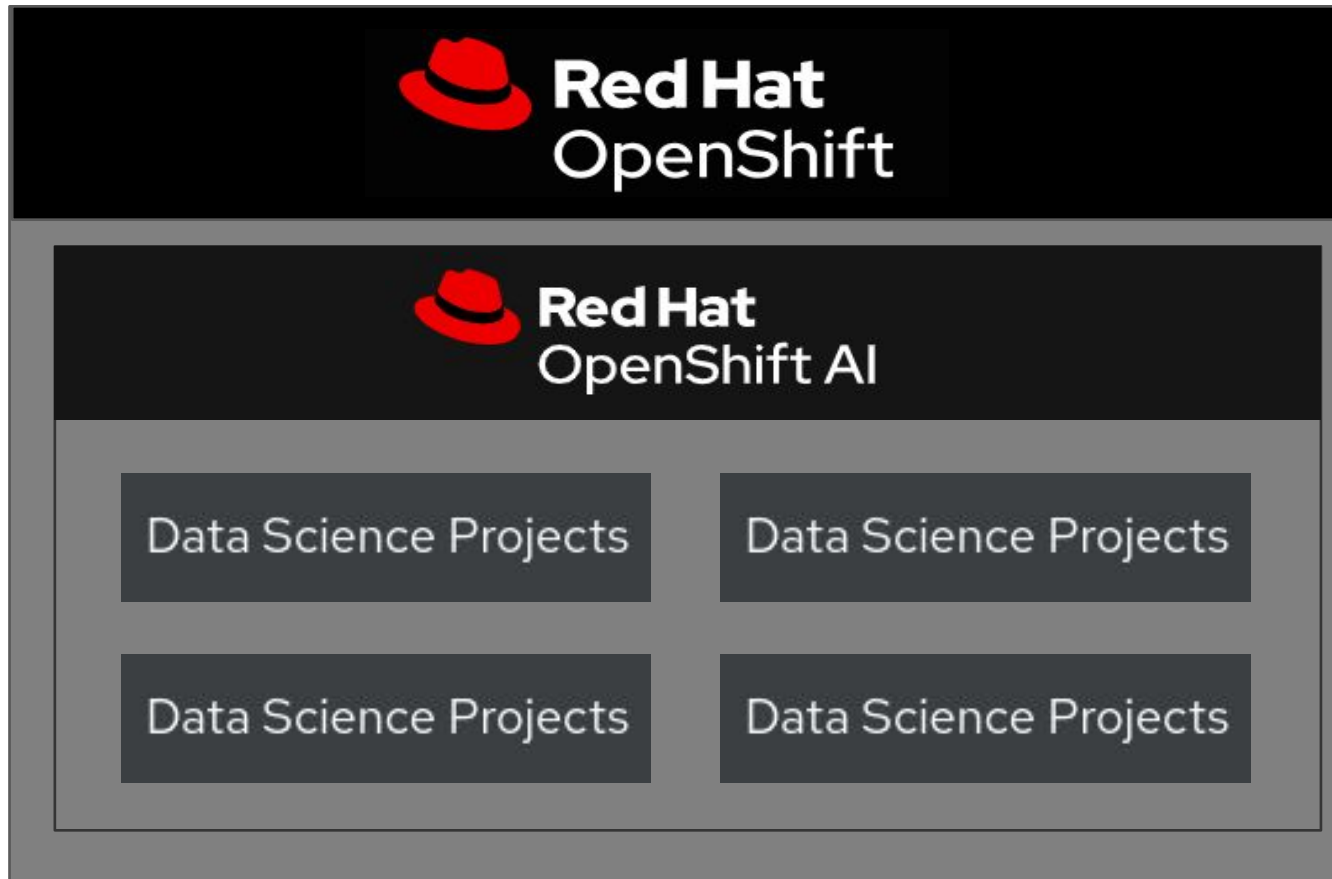


[twitter.com/RedHat](https://twitter.com/RedHat)

# Data Science Projects



# Data Science Projects



- Multiple data science projects.
- Isolation from other projects
- Created by admins or users
- User/Group access privileges

# Data Science Projects

Applications

Enabled

Explore

Data Science Projects

Data Science Pipelines

Model Serving

Resources

Settings

Data Science Projects

View your existing projects or create new projects

Data science projects

Name Find by name

Project	Created	Workbenches
Name		Name Status
project-1	5/31/2024, 9:06:22 AM	<a href="#">Create a workbench</a> to add a custom notebook.
project-2	5/31/2024, 9:06:32 AM	<a href="#">Create a workbench</a> to add a custom notebook.

Data Science projects allow users to **organize** and **manage** contents related to their AI/ML experiments in **isolation** from other projects

[Create a new data science project](#)

# Data Science Projects

Home

Data science projects

Models

Model catalog

Model registry

Model deployments

Data science pipelines


Pipelines

Runs

Experiments

Experiments and runs

Executions



Data science projects

View your existing projects or create new projects.

Name

▼

🔍

project

Name

project ✕

[Clear all filters](#)

Name	Created	Workbenches
<a href="#">project-1</a> ⓘ georg	27.10.2025, 14:36:58	▶ 0 ⓘ 0
<a href="#">project-2</a> ⓘ georg	27.10.2025, 14:37:10	▶ 0 ⓘ 0

1 - 2 of 2

« < 1 of 1 > »

Public workbench

> >>

Data Science projects allow users to **organize** and **manage** contents related to their AI/ML experiments in **isolation** from other projects

# Data Science Projects

**Red Hat OpenShift AI**

panbalag@redhat.com

**Data Science Projects**

View your existing projects or create new projects.

Data science projects

Name Find by name

[Launch Jupyter](#) [Create data science project](#)

Project	Workbenches
Name	Name Status
<a href="#">project-1</a> panbalag@redhat.com	
<a href="#">project-2</a> panbalag@redhat.com	

Resource names and types are used to find your resources in OpenShift.

**Resource name** project-1

**Resource type** Project

Data science projects are 'Projects' in OpenShift identified by the label under 'Resource name'

# Data Science Projects

Home

Data science projects

Models

Model catalog

Model registry

Model deployments

Data science pipelines


Pipelines

Runs

Experiments

Experiments and runs

Executions



Data science projects

View your existing projects or create new projects.

Name

▼

🔍

project

Create project

1 - 2 of 2

«

<

1

of 1

>

»

Name

project

×

[Clear all filters](#)

Name

↑

[project-1](#)


georg

[project-2](#)

georg

Resource name

project-1



Resource type

Project

1 - 2 of 2

«

<

1

of 1

>


»

Start basic workbench

Resource names and types are used to find your resources in OpenShift.

×

Data science projects are 'Projects' in OpenShift identified by the label under 'Resource name'





# Data Science Projects

Applications

Enabled

Explore

Data Science Projects

Data Science Pipelines

Model Serving

Resources

Data Science Projects

View your existing projects or create new projects.

Data science projects

Name Find by name

Project

Name

project-1

panbalag@redhat.com

Filter

Name Search by name...

Name	Display name	Status
project-1	project-1	Active
project-2	project-2	Active

Projects

# Data Science Projects

Red Hat OpenShift AI

Home

Data science projects

Models

Model catalog

Model registry

Model deployments

Data science pipelines

Pipelines

Runs

Experiments

Experiments and runs

Executions

Artifacts

Distributed workloads

Data science projects

View your existing projects or create new projects.

Name

project

Create project

1 - 2 of 2

<<

<

1

Name

project

Clear all filters

project-1

georg

project-2

georg

Resource names and types are used to find your resources in OpenShift.

Resource name

project-1

Resource type

Project

Workbenches

0

0

0

1 - 2 of 2

<<

<

Red Hat OpenShift

Home

Overview

Projects

Search

API Explorer

Events

Favorites

Ecosystem

Software Catalog

Installed Operators

Helm

Workloads

Topology

Pods

Deployments

DeploymentConfigs

StatefulSets

<a href="#">user-workload-monitoring</a>						
PR <a href="#">openshift-vsphere-infra</a>	No display name	Active	No requester	-	-	
PR <a href="#">project-1</a>	project-1	Active	georg	-	-	
PR <a href="#">project-2</a>	project-2	Active	georg	-	-	
PR <a href="#">red-hat-build-of-keycloak</a>	No display name	Active	kube:admin	1.893,6 MiB	0,006 cores	
PR <a href="#">redhat-ods-applications</a>	No display name	Active	No requester	1.336,4 MiB	0,038 cores	
PR <a href="#">redhat-ods-monitoring</a>	No display name	Active	No requester	-	-	
PR <a href="#">redhat-ods-operator</a>	No display name	Active	No requester	628,7 MiB	0,005 cores	
PR <a href="#">rhdt-operator</a>	No display name	Active	No requester	28,2 MiB	0,001 cores	
PR <a href="#">rhoai-model-registries</a>	No display name	Active	No requester	40,8 MiB	0,001 cores	
PR <a href="#">rhods-notebooks</a>	No display name	Active	No requester	-	-	

## Collaborate within a project

- Users that create a data science project
  - become an admin of that project
  - can give access to a project to any user or group
- Users with access permissions can access all resources in the project, modify them, and create new ones.
- Limiting user level access to data science projects needs to be handled at an OpenShift level at the moment

## Collaborate between projects

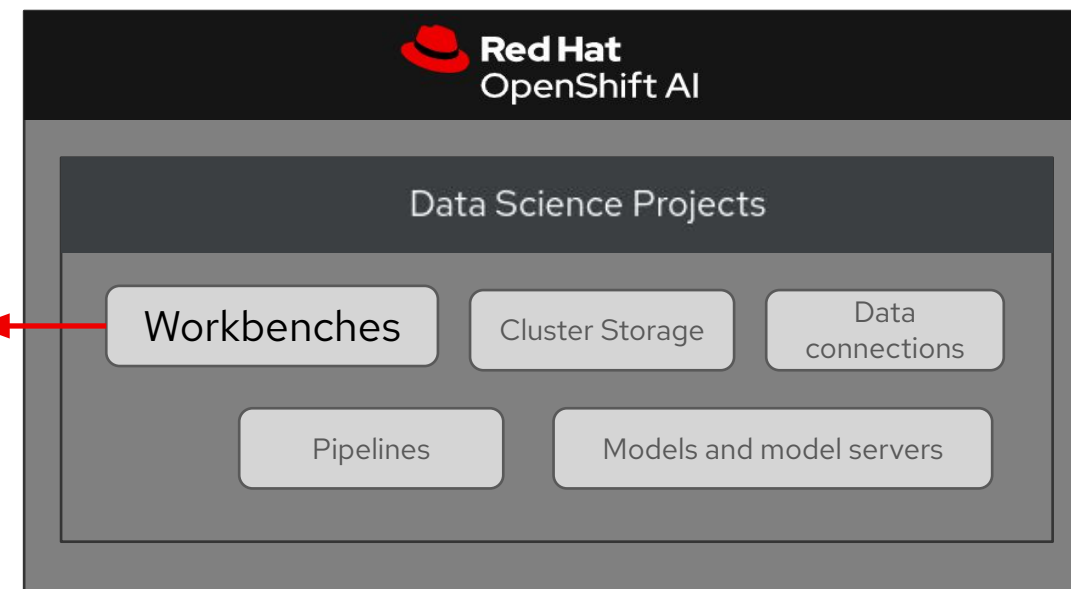
- Due to isolation of data science projects, resources need to be explicitly exposed in order to be shared between projects.
- A good way to do this is to have an external resource which the projects have access to.
  - Examples:
    - A git repository with shared code
    - An object storage with shared artifacts
    - A structured database with shared data

# What we'll discuss today

- ▶ [Motivational Speech](#)
- ▶ [Red Hat AI Inference Server Overview](#)
  - [What are Data Science Projects](#)
    - [What are workbenches](#)
    - [Model Registry](#)
- ▶ [PoC takeaways](#)
- ▶ [Next Steps](#)

# Workbenches

- **Notebook Image**
  - **Development environment** in the form of a container image
    - combination of IDE like Jupyter Notebook, VSCode, etc., and choice of AI/ML framework like Tensorflow, PyTorch etc.,
  - **Custom notebook images.**
- **Deployment size**
  - Container size → **# CPUs & Memory** size
  - Accelerator → Choice of **Accelerators/GPUs**
- **Environment variables**
  - Config Map
  - Secret
- **Cluster Storage**
  - **PVC** connected to the development environment to store code & related artifacts.
- **Data connections**
  - **Object store** for hosting models as well as storing pipeline artifacts.

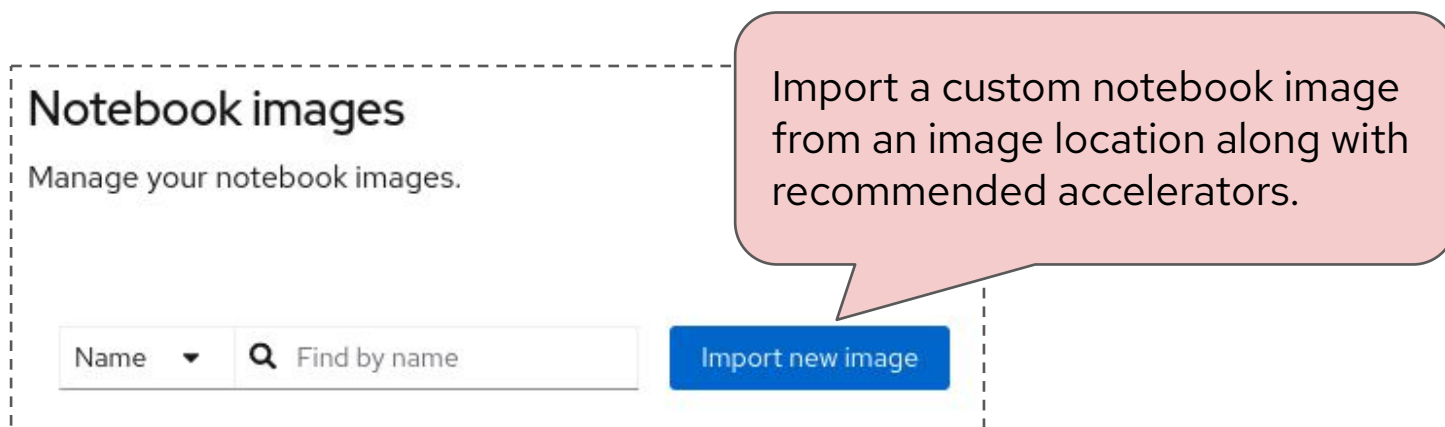
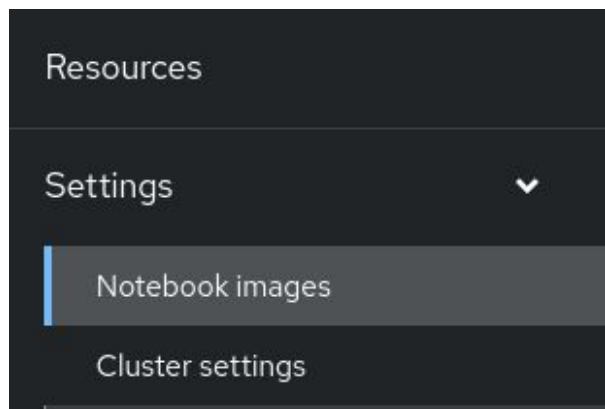


## Default Notebook Images

Image	Description
CUDA	For compute-intensive data science models that require GPU support, the Compute Unified Device Architecture (CUDA) notebook image provides <b>access to the NVIDIA CUDA Toolkit</b> with GPU-accelerated libraries and optimization tools.
Standard Data Science	Contains <b>commonly used libraries</b> to assist you in developing your machine learning models.
TensorFlow	<b>TensorFlow</b> , a popular open source machine learning platform. TensorFlow provides advanced libraries, data visualization features that allows users to build, monitor and track models.
PyTorch	<b>PyTorch</b> is another open source machine learning library optimized for deep learning like computer vision or natural language processing models.
Minimal Python	A <b>minimal environment with JupyterLab</b> for basic exploration.
Trusty AI	For AI/ML work with <b>model explainability, tracing, and accountability</b> , & runtime monitoring
Habana AI	For high-performance optimization of deep learning training workloads and maximize training throughput and efficiency with <b>Habana Gaudi devices</b> .
code-server (Technology Preview)	Provides you with a <b>VSCode</b> environment, allowing you to customize the environment through <b>extensions</b> .

## Customizing Workbenches

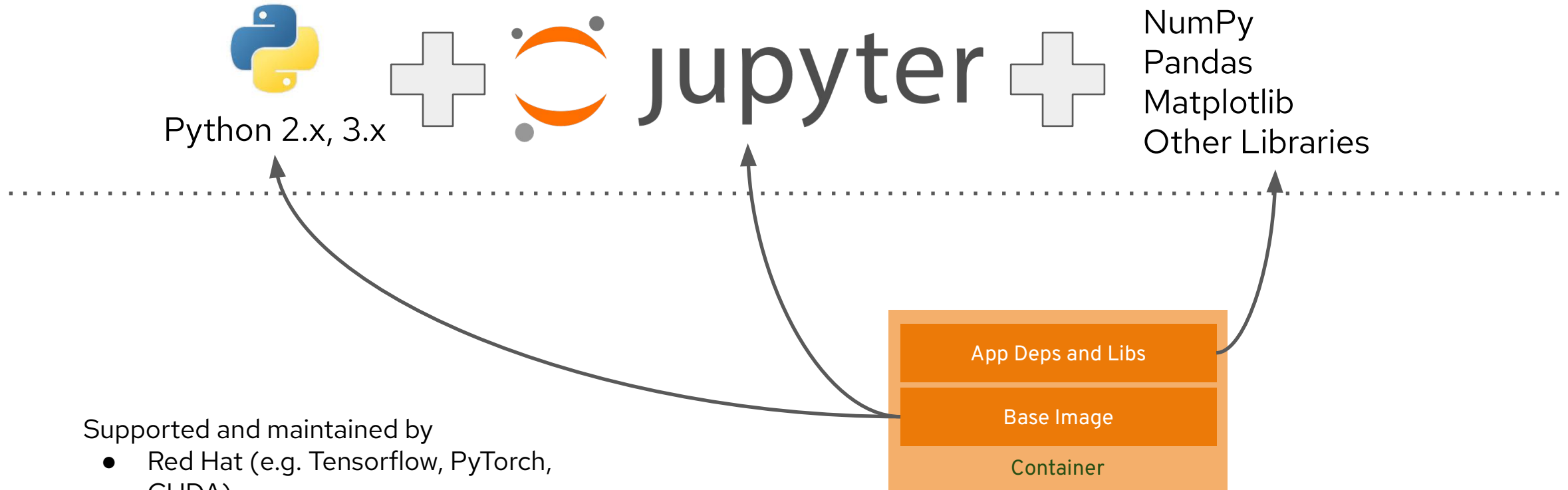
- To customize the workbench you can either:
  - Install dependencies on top of a workbench
  - Use a custom notebook image
- You can use package managers such as pip to add/remove dependencies in an existing workbench
  - Dependencies installed within the workbench are by default not saved to the persistent storage, this is by choice as restarting the workbench is an easy way to reset the environment if something caused an issue with the dependencies
- You can create and use custom notebook images to completely customize the environment



# Customizing Workbenches

## Base Notebook Images

Reproducible and shareable environments for building, training and serving



Supported and maintained by

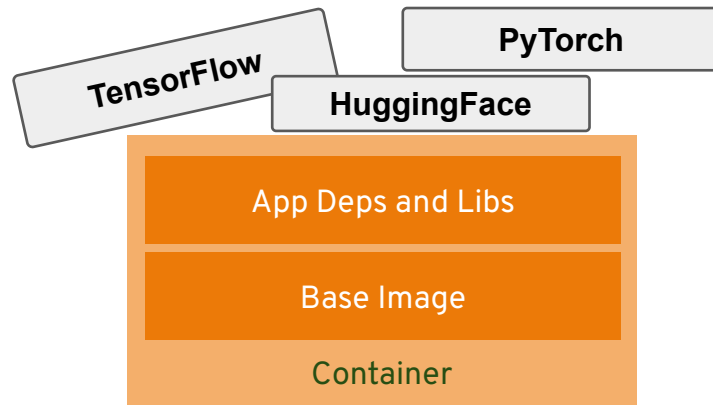
- Red Hat (e.g. Tensorflow, PyTorch, CUDA)
- partner (Anaconda, Intel)
- you (custom notebooks)



# Customizing Workbenches

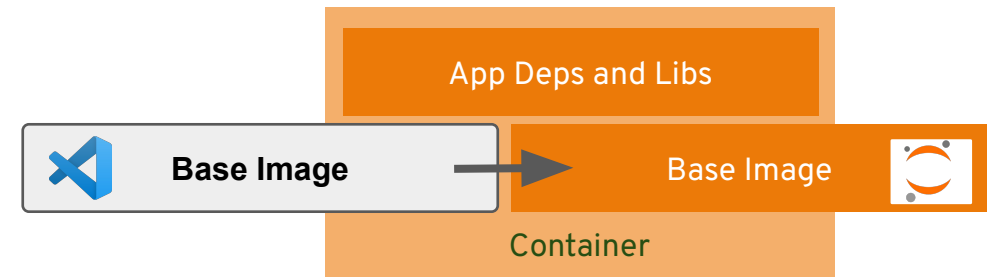
## Customizing the workbench

Adding packages on top of a good image



Just remember that they are removed when restarting the workbench\*

Creating your own custom image with all dependencies you need



You can now version and maintain it according to your preferences

\* This is on purpose so that you can un-mess-up your environment easily if you get into dependency issues.

# What we'll discuss today

- ▶ [Motivational Speech](#)
- ▶ [Red Hat AI Inference Server Overview](#)
  - [What are Data Science Projects](#)
  - [What are workbenches](#)
  - [Model Registry](#)
- ▶ [PoC takeaways](#)
- ▶ [Next Steps](#)

## How will it work?

- Can register a model along with properties such as name, tags, description, model type, dataset etc.
- Can edit the details of the model.
- Uses S3 as a default backend but can link to models in other storages as well, for example separate S3 or PVC.
- Can store artifacts such as generated files, sample data, text files, etc.

# Model Registry Preview

## List models

Applications

Enabled

Explore

Data Science Projects

Data Science Pipelines

Experiments

Experiments and runs

Artifacts

Executions

Model Registry

Model Serving

Resources

Settings

Cluster settings

Serving runtimes

Model Registry settings

User management

Registered models

View and manage your registered models.

finance-team-registry

Keyword

Filter by keyword

Register model

1 - 10 of 12

1 of 2

Model name	Labels	Last modified	Owner
<a href="#">Fraud Detection Model</a> A machine learning model trained to detect fraudulent transactions in financial data.	Fraud Detection Machine Learning Financial	Just now	Alice Smith
<a href="#">Customer Churn Prediction Model</a> Predicts the likelihood of a customer churning based on historical data and customer behavior.	Strategic Customer Churn Prediction and Retent... Predictive Analytics	5 minutes ago	Bob Johnson
<a href="#">Credit Risk Assessment Model</a> Assesses the credit risk of loan applicants using machine learning algorithms.	Credit Risk Machine Learning Financial	3 days ago	David Lee
<a href="#">Stock Price Prediction Model</a> Predicts future stock prices based on historical stock data and market trends.	Time Series Analysis Financial	2 months ago	Charlie Brown
<a href="#">Credit Scoring</a> Predicts the creditworthiness of individuals or businesses based on their financial history and other re...	Portfolio Management Credit Score Predictor Risk Assessment 2 more	2 months ago	Michael Johnson
<a href="#">Product Recommendation</a> Recommends products to customers based on past purchases and preferences.	Product Recommendation	3 months ago	Hannah Liu
<a href="#">Investment Portfolio Optimization Model</a> Optimizes investment portfolios to maximize returns and minimize risks.	Portfolio Optimization Investment Strategy Financial	1 year ago	Charlie Brown

# Model Registry Preview

## Model details and versions

Applications

Enabled

Explore

Data Science Projects

Data Science Pipelines

Experiments

Experiments and runs

Artifacts

Executions

Model Registry

Model Serving

Resources

Settings

Cluster settings

Serving runtimes

Model Registry settings

User management

Registered models - finance-team-registry > Fraud Detection Model

Fraud Detection Model

A machine learning model trained to detect fraudulent transactions in financial data.

Actions

Versions

Details

Keyword Filter by keyword Register new version

Version name	Last modified	Owner	Labels
<a href="#">v8.0 - Cross-domain</a> trained on data from multiple domains for enhanced generalization	1 minute ago	Joe Doe	-
<a href="#">v7.0 - Adaptive Learning</a> Version of the fraud detection model with adaptive learning capabilities to adapt to changing fraud patter...	3 days ago	Joe Doe	-
<a href="#">v6.0 - Explainable AI</a> using explainable AI techniques to provide insights into model predictions.	5 days ago	Bob Anderson	-
<a href="#">v5.0 - Ensemble</a> Ensemble version of the fraud detection model combining multiple base models for improved accuracy.	1 week ago	Joe Doe	Custom label 1 very very... 2 more
<a href="#">v4.0 - Advanced Features Version of Fraud Detection</a> incorporating advanced features and machine learning algorithms	2 months ago	Bob Anderson	-
<a href="#">v3.0 - Real-time</a> optimized for low-latency processing of transactions	2 months ago	Jack Smith	Real-time version
<a href="#">v2.0 - Enhanced</a> improve accuracy and performance	2 months ago	Jack Smith	-

Applications

Enabled

Explore

Data Science Projects

Data Science Pipelines

Experiments

Experiments and runs

Artifacts

Executions

Model Registry

Model Serving

Resources

Settings

Cluster settings

Serving runtimes

Model Registry settings

User management

Registered models - finance-team-registry > Fraud Detection Model

Fraud Detection Model

Our Advanced Fraud Detection Model represents the pinnacle of modern fraud detection technology, meticulously designed to safeguard businesses and financial institutions against the ever-evolving threat of fraudulent activities. Leveraging cutting-edge machine learning algorithm...

Actions

Versions

Details

Description

Our Advanced Fraud Detection Model represents the pinnacle of modern fraud detection technology, meticulously designed to safeguard businesses and financial institutions against the ever-evolving threat of fraudulent activities. Leveraging cutting-edge machine learning algorithms, statistical analysis, and behavioral analytics, our model offers unparalleled accuracy and efficiency in identifying fraudulent transactions, activities, and patterns.

At its core, our model utilizes a sophisticated ensemble approach, combining the strengths of multiple machine learning techniques to achieve superior performance. By aggregating insights from various algorithms, including decision trees, random forests, logistic regression, and neural networks, our model can effectively detect and mitigate a wide r... Show more

Model ID

124dsdk-jlaskw-dl32oa-2kjnf-d-sjkwer2

Owner

Haley Wang

Last modified at

1 minute ago

Created at

1 minute ago

Labels

Classification Transformers PyTorch ONNX distilbert generated\_from\_trainer English Transformers.js Eval Results 24 papers Predictive Analytics Real-time Detection Ensemble Learning Feature Engineering Enterprise Solution Cloud Deployment Documentation Anomaly Detection Supervised Learning 3 more

Properties

Key Value

team finance

code\_format def print\_prime(n):  
...  
Print all primes between 1 and n  
Show more



# Model Registry Preview

## Deploy and keep track

Applications

Enabled

Explore

Data Science Projects

Data Science Pipelines

Experiments

Experiments and runs

Artifacts

Executions

Model Registry

Model Serving

Resources

Settings

Cluster settings

Serving runtimes

Model Registry settings

User management

Registered models - Haley-private > my-model > v0.0.2

v0.0.2

improve accuracy and performance

DetailsRegistered deployments

Only partial of the model deployments showing in the list

This list only shows the model deployments that are deployed through Model Registry. To view the full list of deployments of this model and manage deployments, please navigate to the [Model Serving](#) section.

Model deployment name	Project	Serving runtime	Inference endpoint	Status
<a href="#">Invest-Portfolio-CD-4</a>	finance-prod	ONNX Runtime	<a href="http://invest-portfolio-cd-3.example.com/pr...">http://invest-portfolio-cd-3.example.com/pr...</a>	
<a href="#">Invest-Portfolio-CD-3</a>	finance-prod	ONNX Runtime	<a href="http://invest-portfolio-cd-3.example.com/pr...">http://invest-portfolio-cd-3.example.com/pr...</a>	
<a href="#">Invest-Portfolio-CD-2</a>	finance-dev	ONNX Runtime	<a href="http://invest-portfolio-cd-2.example.com/pr...">http://invest-portfolio-cd-2.example.com/pr...</a>	
<a href="#">Invest-Portfolio-CD-1</a>	finance-dev	OpenVINO Model Server (Supports GPUs)	<a href="http://invest-portfolio-cd-1.example.com/pr...">http://invest-portfolio-cd-1.example.com/pr...</a>	

# What we'll discuss today

- ▶ [Motivational Speech](#)
- ▶ [Red Hat AI Inference Server Overview](#)
  - [What are Data Science Projects](#)
  - [What are workbenches](#)
  - [Model Registry](#)
- ▶ [PoC takeaways](#)
- ▶ [Next Steps](#)

# Data Science Projects

Red Hat  
OpenShift AI

Applications

Enabled

Explore

Data Science Projects

Data Science Pipelines

Model Serving

Resources

Settings

Data Science Projects

View your existing projects or create new projects

Data science projects

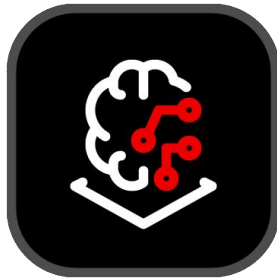
Name Find by name

Project	Created	Workbenches
Name		Name Status
<div>project-1</div> <div>panbalag@redhat.com</div>	5/31/2024, 9:06:22 AM	<div>Create a workbench to add a custom notebook.</div>
<div>project-2</div> <div>panbalag@redhat.com</div>	5/31/2024, 9:06:32 AM	<div>Create a workbench to add a custom notebook.</div>

Data Science projects allow users to **organize** and **manage** contents related to their AI/ML experiments in **isolation** from other projects

science project





### Objective #1]

Provide a high level description of finding #1

### [Objective #2]

Provide a high level description of finding #2

### [Objective #3]

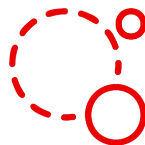
Provide a high level description of finding #2

These were our desired outcomes from this PoC.



[insert desired outcome]

Description



[insert desired outcome]

Description



[insert desired outcome]


Description

# What we'll discuss today

- ▶ [Motivational Speech](#)
- ▶ [Red Hat AI Inference Server Overview](#)
  - [What are Data Science Projects](#)
  - [What are workbenches](#)
  - [Model Registry](#)
- ▶ [PoC takeaways](#)
- ▶ [Next Steps](#)

## Data Science projects

- Your “toolbox”
- 

 **Data Science Projects**

[3scale](#)  
Red Hat Integration - 3scale  
**Created**  
8/15/2025, 11:17:43 AM  
**Owner**  
Unknown

[cert-manager](#)  
  
**Created**  
8/15/2025, 10:40:32 AM  
**Owner**  
Unknown

5 of 19 projects [Go to Data Science Projects](#)

- Workbenches: Where you can create and manage various development environments like JupyterLab, VSCode, or other custom Workbenches. It provides a user-friendly interface for data scientists to work with notebooks, libraries, and datasets.
- Pipelines: You may use pipelines to automate the process of processing data or training and deploying machine learning models.
- Models: Where you can manage and deploy machine learning models. You can create, update, and delete models, as well as monitor their performance and usage.
- Cluster storage: Here you can manage the storage resources used by your models and workbenches. You can create, update, and delete storage resources, as well as monitor their usage.
- Connections: This is where you can manage the connections between your workbenches or model runtimes and other services, such as storage (S3), databases or APIs. You can create, update, and delete connections, as well as see which environment is using them.
- Permissions: This is where you can manage the permissions for project. You can create, update, and delete permissions, as well as see which users or groups have access to which resources.

[Data Science Projects](#) > LLM Host



