Red Hat AI

RH AI 3 Overview & Roadmap

Generative AI adoption challenges

Cost

Generative AI frontier model services are cost prohibitive at scale for many enterprise customer use cases.

Complexity

Integrating models with private enterprise data for customer use cases is too complex for non-data scientists.

Control

Increasing concerns with data privacy, security, and latency are compelling organizations to adopt hybrid strategies.

Red Hat AI

Trusted, Consistent and Comprehensive foundation

NVIDIA AMD intel Hardware Acceleration Google aws IBM

Physical Virtual Private Cloud Public Cloud Edge

Red Hat AI

Accelerate the development and delivery of AI solutions across hybrid-cloud environments

Increase efficiency with **fast, flexible and efficient inferencing**

Simplified and consistent experience for **connecting models to data**

Accelerate Agentic AI deployments

Flexibility and consistency when **scaling AI across the hybrid cloud**

Introducing RH AI 3

Flexible and Efficient Inference

- GA distributed inference (llm-d)
- New validated and optimized models
- vLLM enhancements
- LLM Compressor GA

Connecting Data to Models

- Modular and extensible approach for: data ingestion, synthetic data generation, tuning, evaluations.
- RAG enhancements & partner integrations
- Feature Store UI

Red Hat AI

Agentic AI

- AI experiences: AI hub and gen AI studio
- Model Context Protocol support & MCP Server access in gen AI studio
- Llama Stack API integration

AI Platform

- Model catalog and registry GA
- Model as a Service provider enhancements and API Mgt integration
- GPU as a Service enhancements

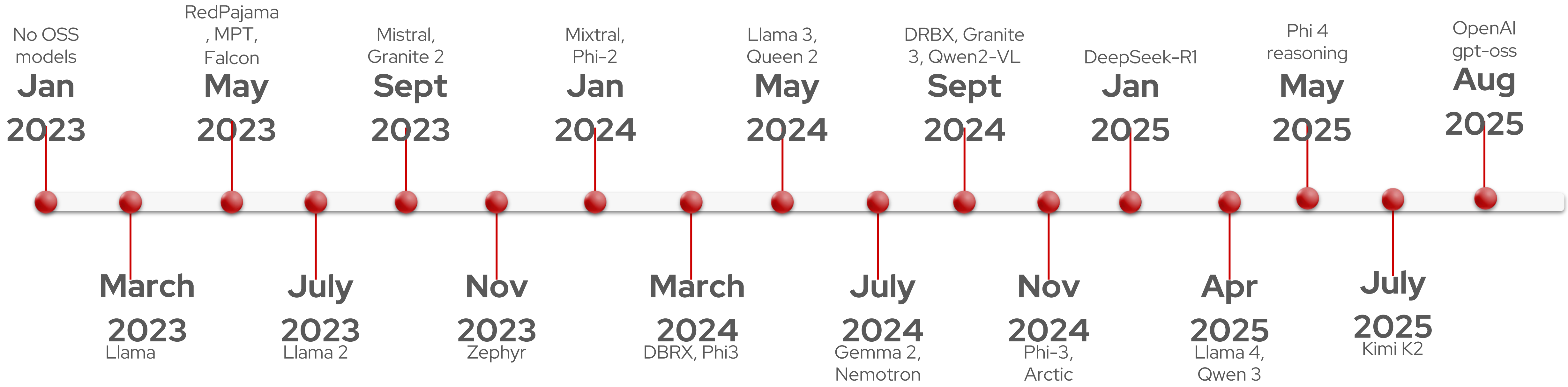
Single platform to run any model, on any accelerator, on any cloud

Red Hat AI

Fast, flexible, and efficient inference

Expanding choice of models

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



No OSS models
Jan 2023

March 2023
Llama

RedPajama, MPT, Falcon
May 2023

July 2023
Llama 2

Mistral, Granite 2
Sept 2023

Nov 2023
Zephyr

Mixtral, Phi-2
Jan 2024

March 2024
DBRX, Phi3

Llama 3, Queen 2
May 2024

July 2024
Gemma 2, Nemotron

DRBX, Granite 3, Qwen2-VL
Sept 2024

Nov 2024
Phi-3, Arctic

DeepSeek-R1
Jan 2025

Apr 2025
Llama 4, Qwen 3

Phi 4 reasoning
May 2025

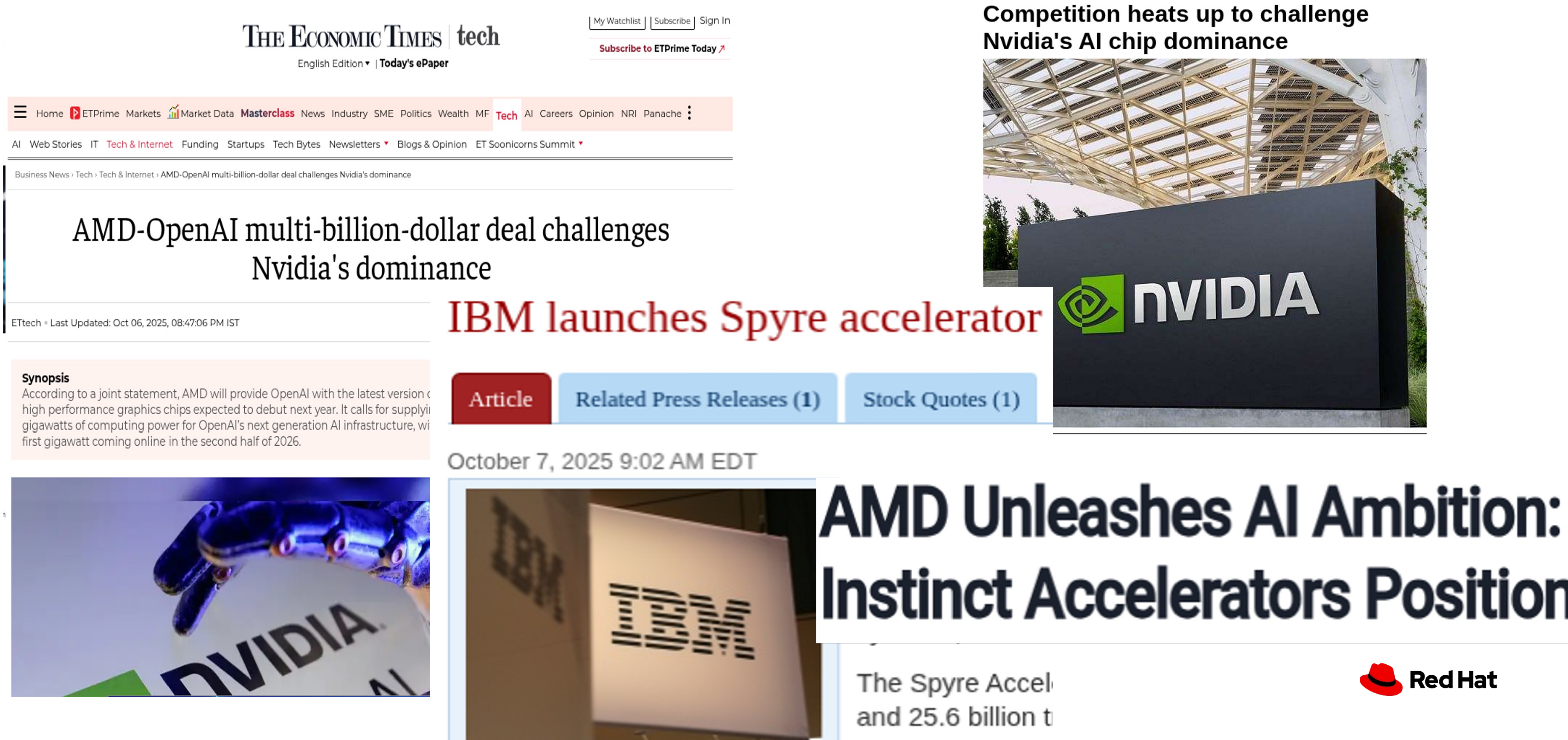
July 2025
Kimi K2

OpenAI gpt-oss
Aug 2025

Meta Ai2 MISTRAL AI_ NVIDIA Hugging Face databricks snowflake IBM Google Microsoft deepseek

Expanding choice of Accelerators

NVIDIA is still the major player but other entrants gaining ground



AMD-OpenAI multi-billion-dollar deal challenges Nvidia's dominance

ETtech • Last Updated: Oct 06, 2025, 08:47:06 PM IST

Synopsis
According to a joint statement, AMD will provide OpenAI with the latest version c high performance graphics chips expected to debut next year. It calls for supplyi gigawatts of computing power for OpenAI's next generation AI infrastructure, wi first gigawatt coming online in the second half of 2026.

Competition heats up to challenge Nvidia's AI chip dominance

IBM launches Spyre accelerator

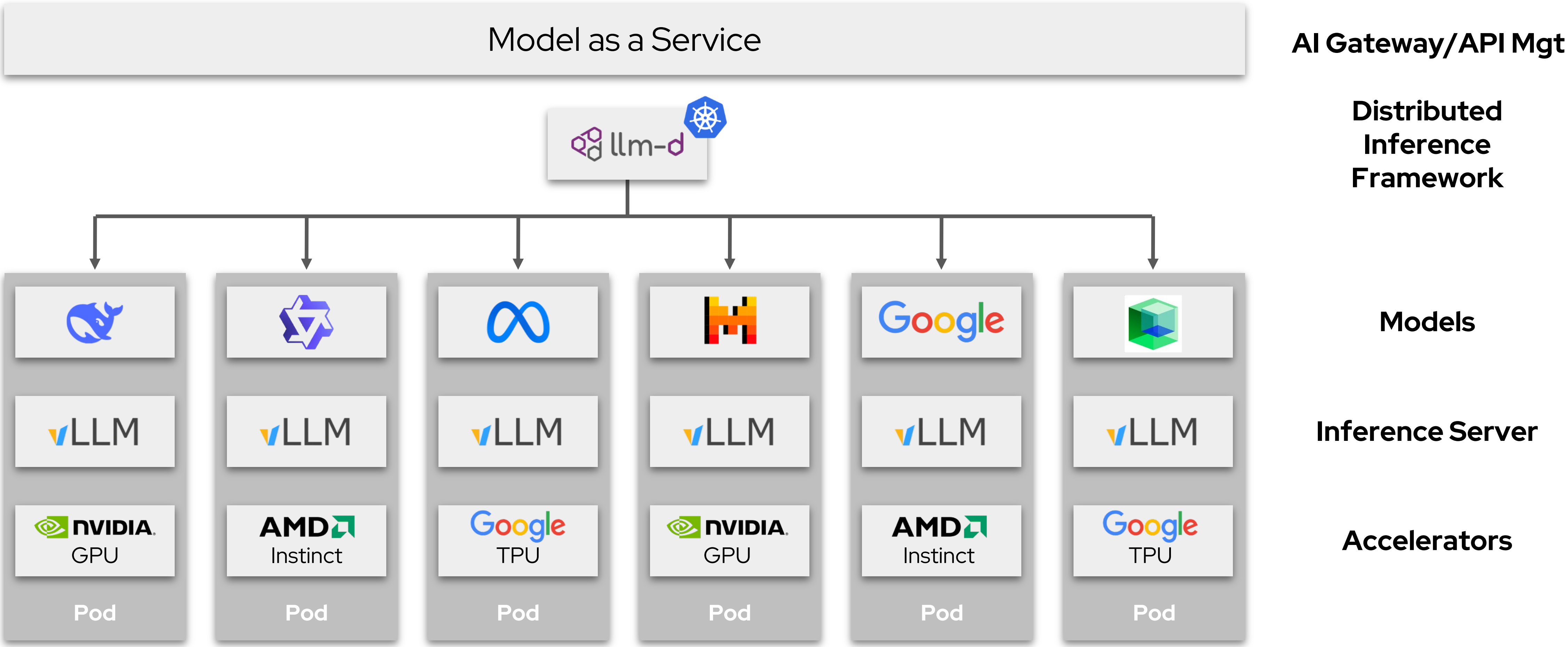
October 7, 2025 9:02 AM EDT

The Spyre Accel
and 25.6 billion t

AMD Unleashes AI Ambition: Instinct Accelerators Position

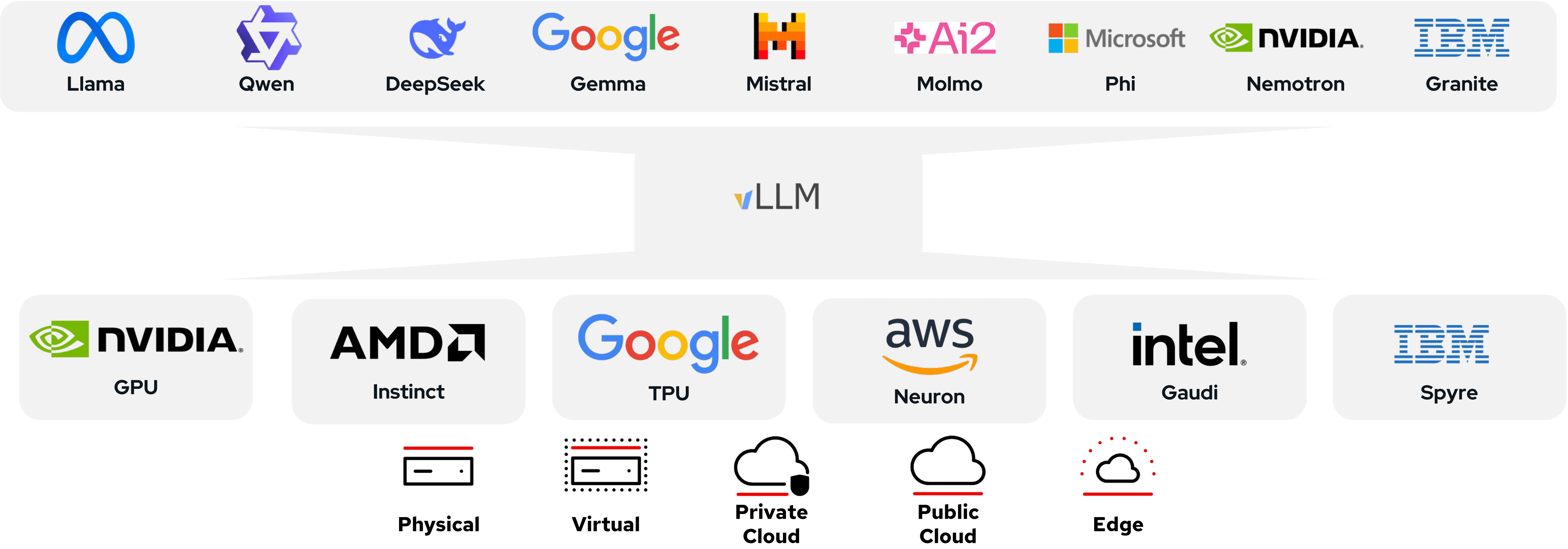
Enterprise GenAI Inference Platform

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



vLLM Inference Server

vLLM v1 with enhanced performance, expanded model and hardware support



Llama Qwen DeepSeek Gemma Mistral Molmo Phi Nemotron Granite

vLLM

GPU Instinct TPU Neuron Gaudi Spyre

Physical Virtual Private Cloud Public Cloud Edge

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

Red Hat AI Model Repository

New validated and optimized models and LLM Compressor now GA

Broad Collection of models

Llama	Qwen	Gemma
Mistral	DeepSeek	Phi
Molmo	Granite	Nemotron

Choice of Models

- Transformers (Dense, MOE), Multi-modal LLMs, Embeddings Models, Hybrid / Novel Attention, Vision
- Hugging Face compatible (safe tensors), OCI-compatible containers

Validated models

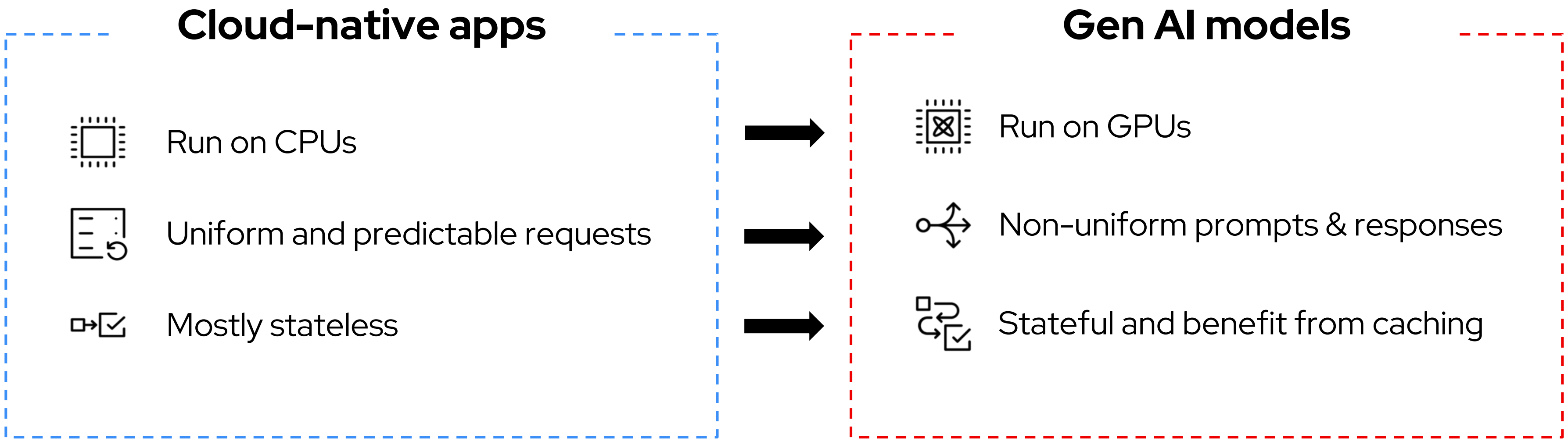
- Tested using realistic scenarios
- Assessed for performance across a range of hardware
- Done using GuideLLM benchmarking and LM Eval Harness

Optimized models

- Compressed for speed and efficiency
- Designed to run faster, use fewer resources, maintain accuracy
- Done using LLM Compressor with latest algorithms

Overcoming the generative AI challenges

Running LLMs efficiently



Cloud-native apps

- Run on CPUs
- Uniform and predictable requests
- Mostly stateless

Gen AI models

- Run on GPUs
- Non-uniform prompts & responses
- Stateful and benefit from caching

Inference at scale everywhere

Distributed, scalable gen AI inference for Enterprise AI

Red Hat AI

Now includes llm-d

llm-d reimagines how LLMs run on Kubernetes

- Lower infrastructure spend for AI
- Great performance at larger scale
- Seamless scaling for unpredictable AI demand

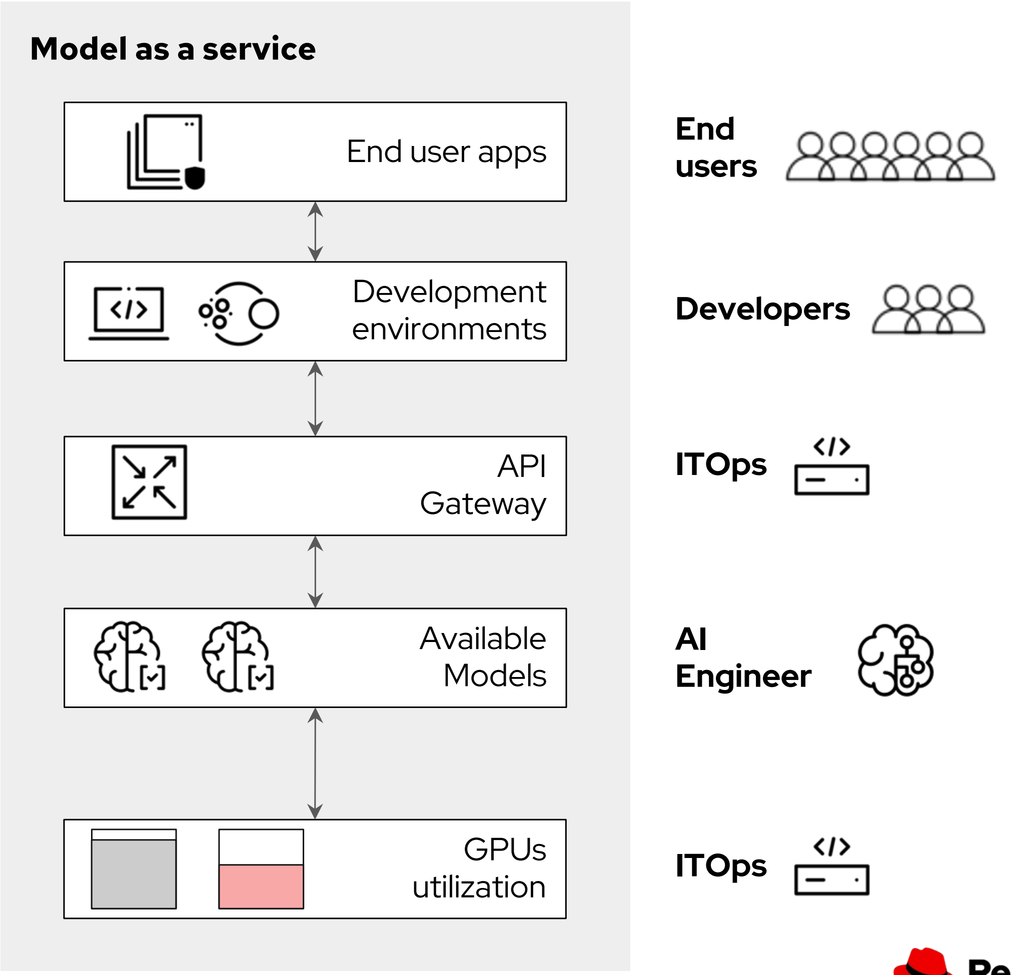
Deliver faster, cheaper, and more manageable AI systems for enterprise production

llm-d is GA in RHOAI 3.0

Model-as-a-Service in OpenShift AI (Dev Preview)

Offering AI models as the service to a larger audience

- IT serves common models centrally
  - Generative AI focus, applicable to any model
  - Centralized pool of hardware
  - Platform Engineering for AI
- Models available through the RHOAI console
- Developers consume models, build AI applications
  - For end users (private assistants, etc)
  - To improve products or services through AI
- Shared Resources business model keeps costs down

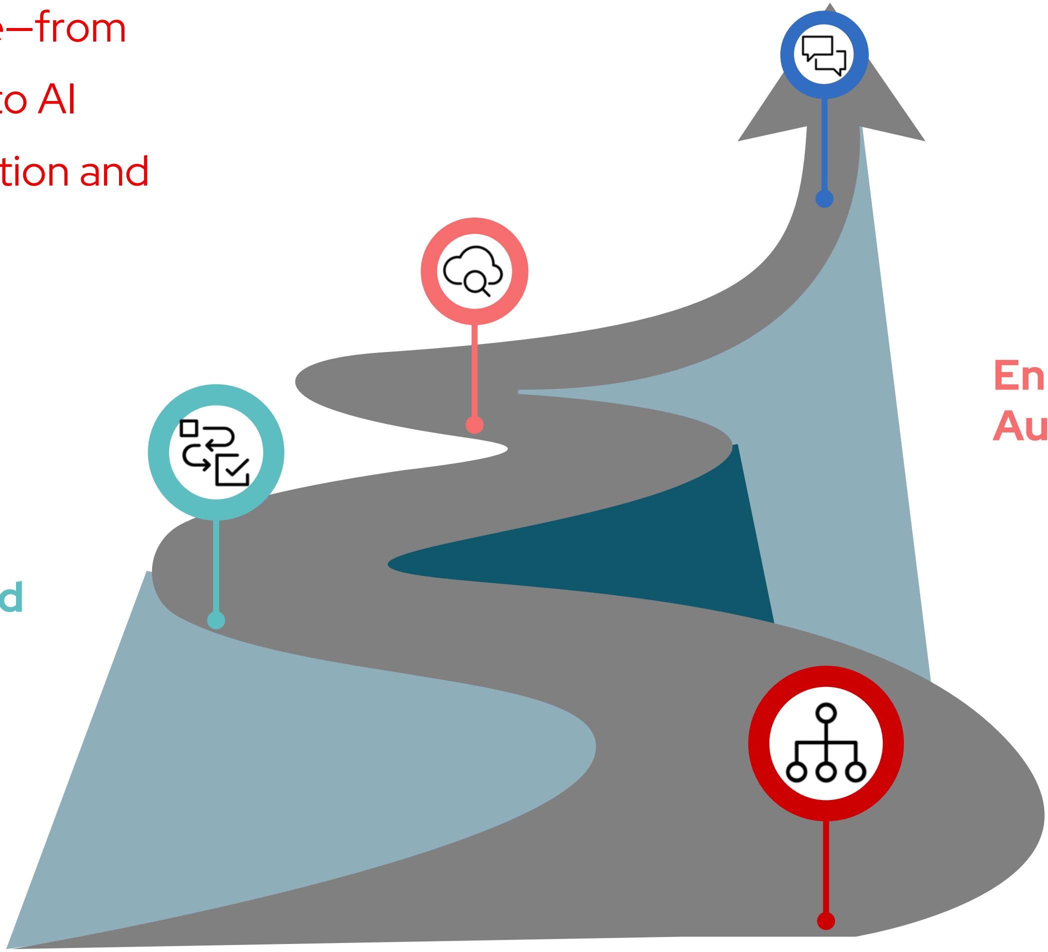


Red Hat AI

Connecting models to data

The evolution of Red Hat AI's tuning and alignment tools

Access to AI tooling that caters to different levels of AI expertise—from developers to data scientists to AI engineers—ensuring collaboration and frictionless interaction.



Modular and extensible approach for data ingestion, synthetic data generation, model tuning, and evaluation

Enhance gen AI apps with Retrieval Augmented Generation (RAG)

Continual fine-tuning via Online Supervised Fine Tuning and rapid domain adaptation via LoRA/QLoRA

InstructLab allows customers to align models by adding new knowledge and skills

New model customization approach offers a modular extensible architecture

Data processing

Simplifies document processing and parsing into AI-readable data for model customization and RAG applications

Synthetic data Generation hub

Generate high-quality data, with dynamic parameters, run-time visibility, and multilingual support

Training hub

An algorithm-focused interface for common llm training, continual learning, and reinforcement learning techniques

Evaluations

Enables large language model inference evaluations.

Connecting models to data

Build customized AI solutions that address domain specific business cases

Coming soon

Prompt design

Prompt tuning and engineering

Design and engineer the prompts to enhance GenAI model responses and achieve more specific and accurate outcomes.

Enhanced

RAG

Retrieval Augmented Generation

Enhance Gen AI model generated text by retrieving relevant information from external sources, improving accuracy and depth of model's responses.

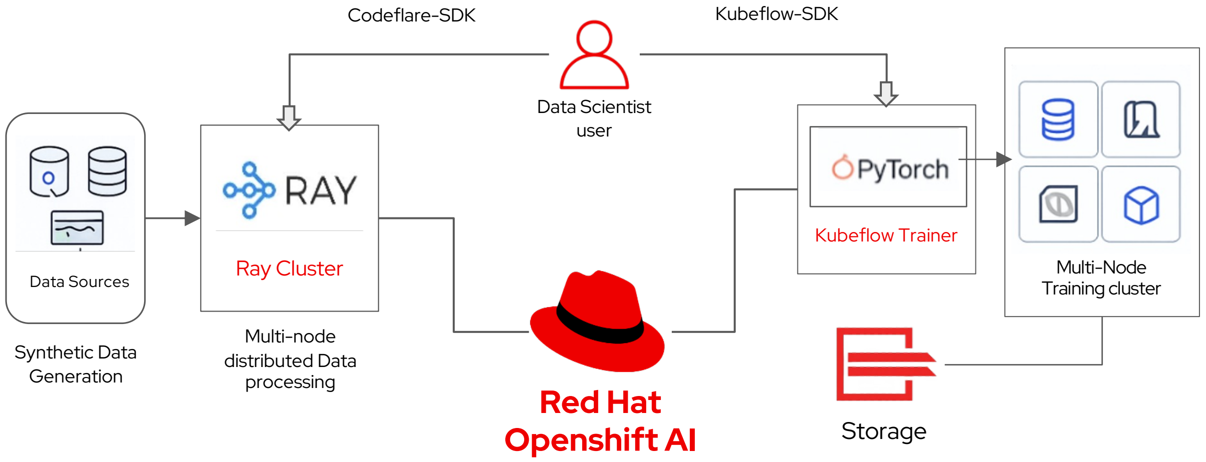
Enhanced

Fine tuning

InstructLab, LoRa and QLora

Adjust a pre-trained model on specific tasks or data, improving its performance and accuracy for specialized applications without full retraining.

Ray + Kubeflow Training : End to End ML Pipeline



NVIDIA AMD intel Hardware Acceleration aws Google IBM

Physical

Virtual

Edge

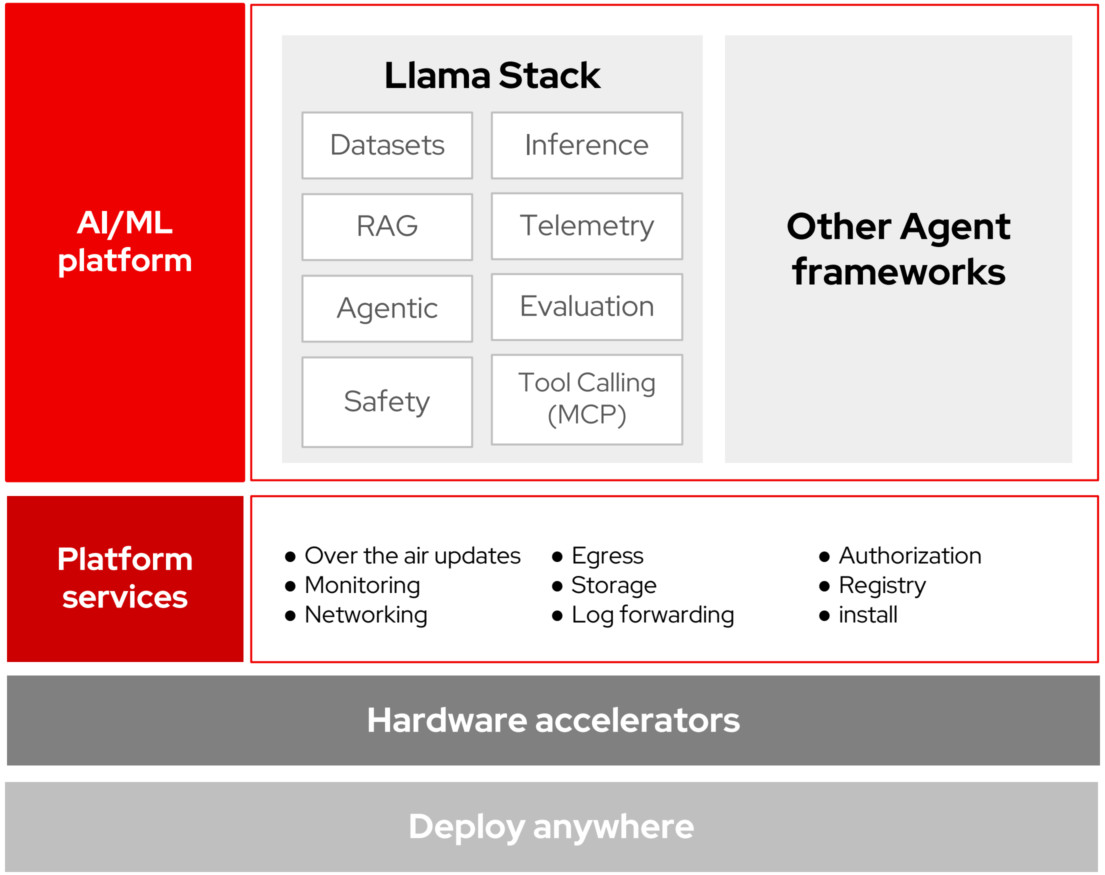
Red Hat AI

Accelerate Agentic AI

Red Hat AI provides an agile, stable foundation to accelerate the development and deployment of AI agentic workflows.

- Allows running and managing agents as microservices.
- Simplifies production deployment by managing LLM serving and scaling.
- Offers native capabilities to build and manage agents with Llama Stack, and standardized communication protocols (MCP).
- Provides the flexibility to integrate preferred tools like LangChain and Crew AI.

A modular approach to building AI agents



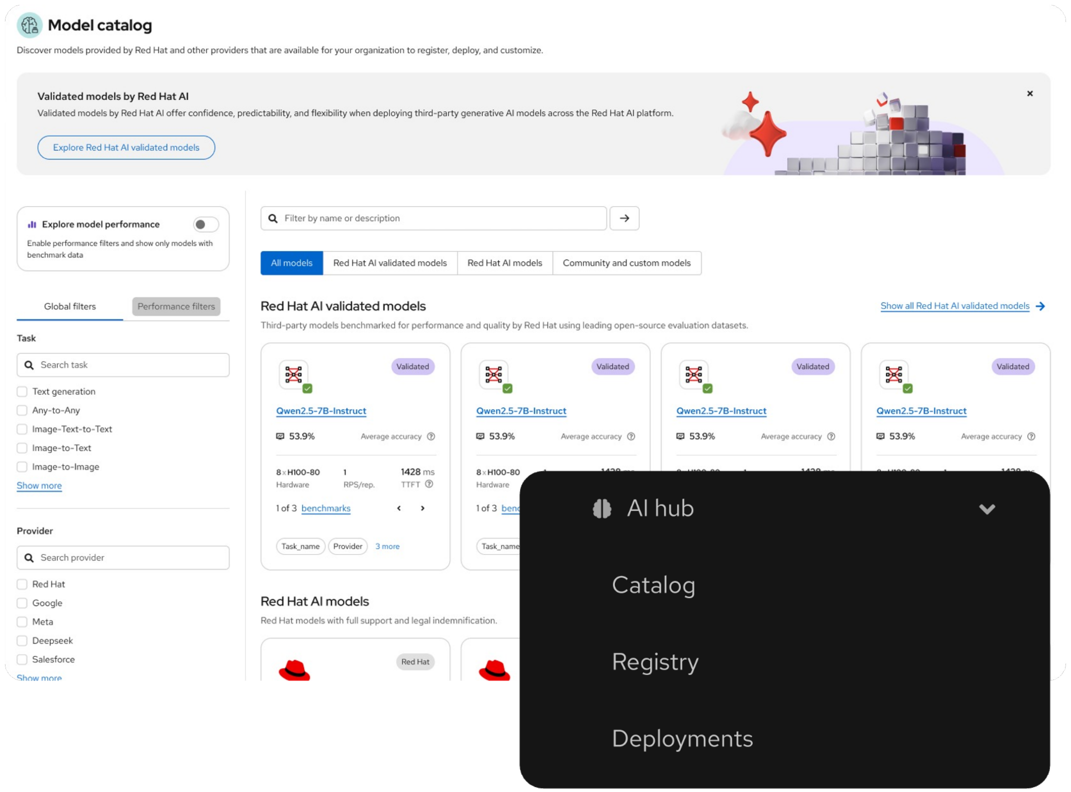
Red Hat AI allows to:

- Build agents using **Llama Stack's native capabilities and implementations**.
- **Bring compatible Llama Stack implementations** to OpenShift AI.
- **Use your own agent framework** and selectively incorporate Llama Stack APIs.
- **Build with Core Primitives** and manage your own agent framework as a standard workloads.

AI Dedicated Experiences

Dedicated dashboard experiences provide a seamless experience to platform and AI engineers

AI hub



Model catalog

Discover models provided by Red Hat and other providers that are available for your organization to register, deploy, and customize.

Validated models by Red Hat AI

Validated models by Red Hat AI offer confidence, predictability, and flexibility when deploying third-party generative AI models across the Red Hat AI platform.

Explore Red Hat AI validated models

Explore model performance

Enable performance filters and show only models with benchmark data

Global filters | Performance filters

Task

Search task

□ Text generation
□ Any-to-Any
□ Image-Text-to-Text
□ Image-to-Text
□ Image-to-Image

Show more

Provider

Search provider

□ Red Hat
□ Google
□ Meta
□ Deepseek
□ Salesforce

Show more

Filter by name or description

All models | Red Hat AI validated models | Red Hat AI models | Community and custom models

Red Hat AI validated models

Show all Red Hat AI validated models

Third-party models benchmarked for performance and quality by Red Hat using leading open-source evaluation datasets.

Validated

Qwen2.5-7B-Instruct

53.9% Average accuracy

8 x H100-80 Hardware | 1 RPS/rep. | 1428 ms TTFT

1 of 3 benchmarks

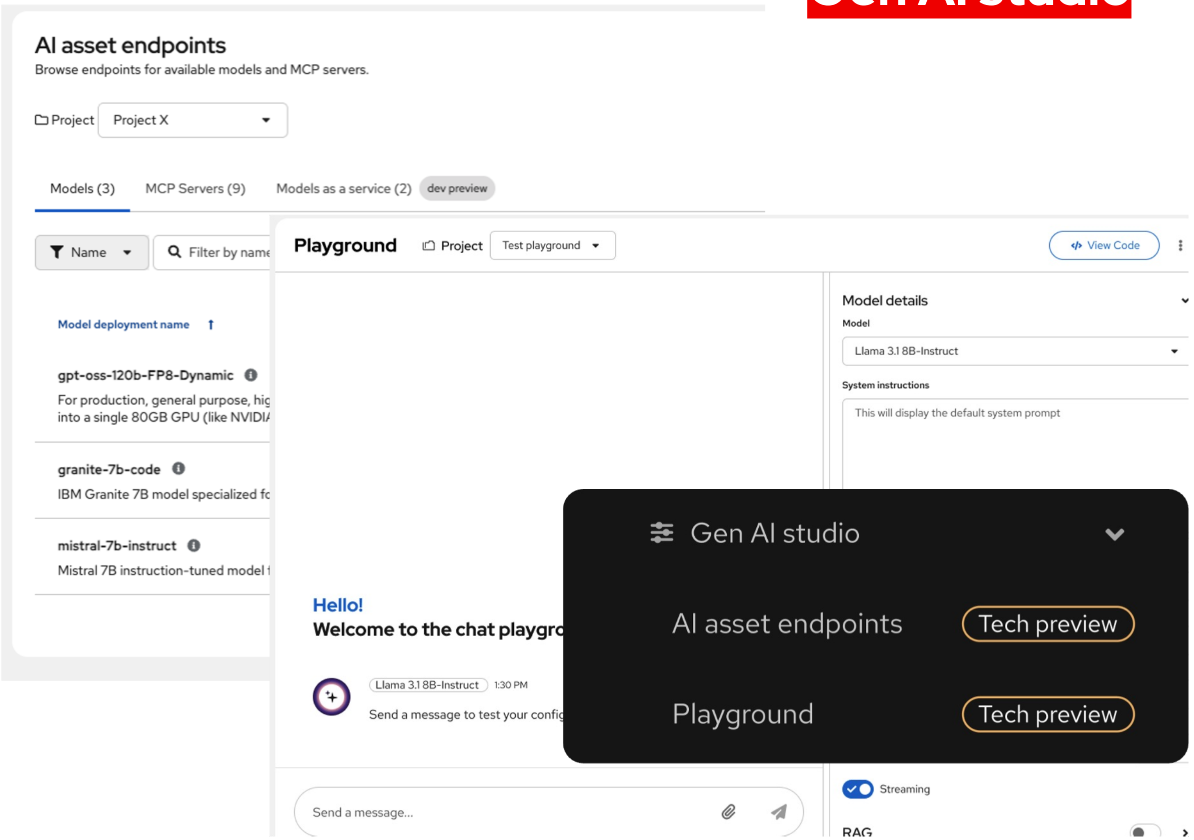
Task_name | Provider | 3 more

Red Hat AI models

Red Hat models with full support and legal indemnification.

Red Hat

Gen AI studio



AI asset endpoints

Browse endpoints for available models and MCP servers.

Project Project X

Models (3) | MCP Servers (9) | Models as a service (2) dev preview

Name | Filter by name

Model deployment name

gpt-oss-120b-FP8-Dynamic

For production, general purpose, hig
into a single 80GB GPU (like NVIDI/

granite-7b-code

IBM Granite 7B model specialized fc

mistral-7b-instruct

Mistral 7B instruction-tuned model f

Playground

Project Test playground

View Code

Model details

Model

Llama 3.1 8B-Instruct

System instructions

This will display the default system prompt

Hello!
Welcome to the chat playgro

Llama 3.1 8B-Instruct 1:30 PM

Send a message to test your config

Send a message...

Streaming

RAG

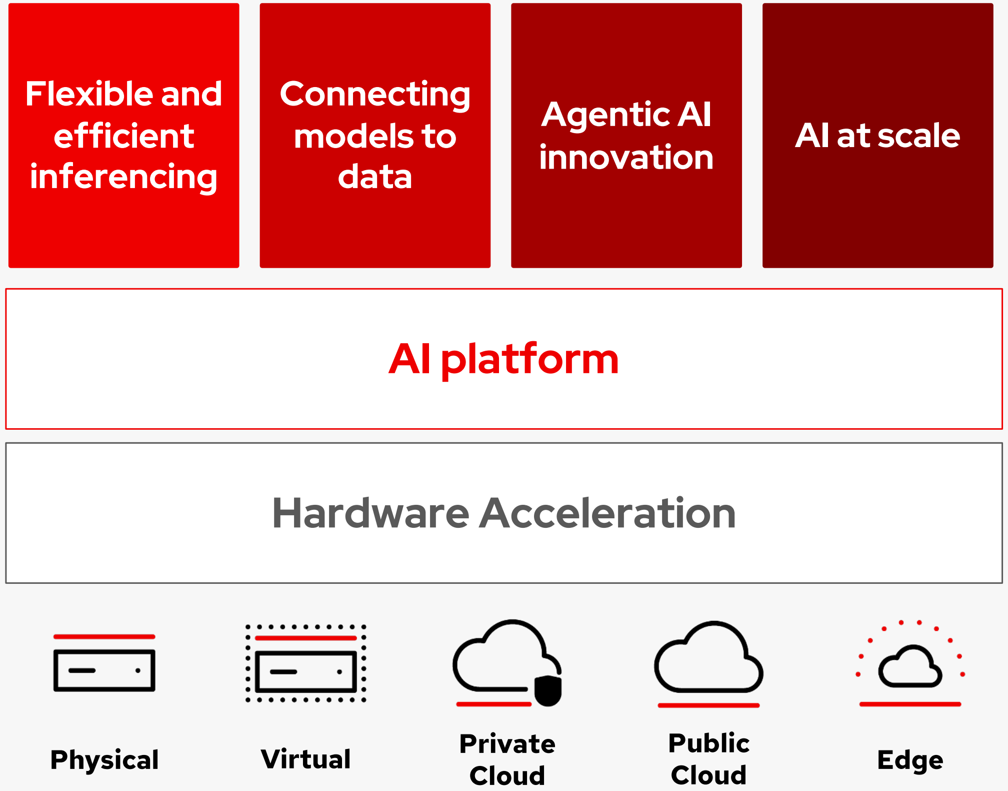
Red Hat AI

Scaling AI across the hybrid cloud

Red Hat AI provides a platform to consistently build, deploy and manage AI models, training and agentic applications at scale across the hybrid cloud

It includes:

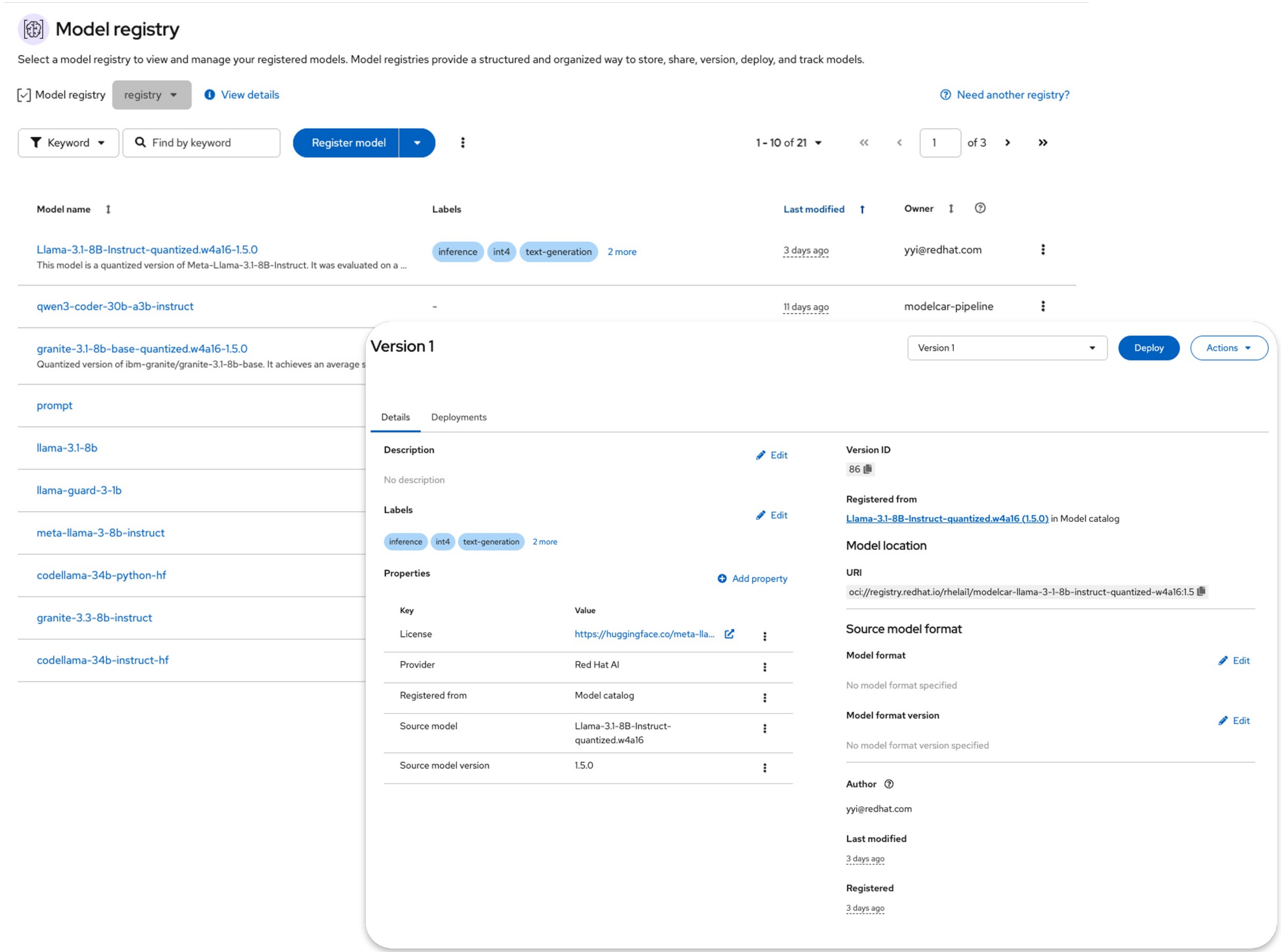
- Enterprise-grade, flexible and secure AI platform
- Private and sovereign AI capabilities and practices
- GPU-as-a-Service and customer-deployable Model-as-a-Service
- Full MLOps and Gen AI Ops lifecycle support from experimentation to production
- Model safety and trust with explainability, fairness, and guardrails



Registry GA

Registry (Models)

OpenShift AI users now have a centralized repository within the OpenShift AI platform designed to manage the lifecycle of machine learning models via the Registry.



Model registry

Select a model registry to view and manage your registered models. Model registries provide a structured and organized way to store, share, version, deploy, and track models.

Model registry: registry · View details · Need another registry?

Keyword · Find by keyword · Register model · 1 - 10 of 21 · 1 of 3

Model name	Labels	Last modified	Owner
Llama-3.1-8B-Instruct-quantized.w4a16-1.5.0 This model is a quantized version of Meta-Llama-3.1-8B-Instruct. It was evaluated on a ...	inference, int4, text-generation, 2 more	3 days ago	yyi@redhat.com
qwen3-coder-30b-a3b-instruct	-	11 days ago	modelcar-pipeline
granite-3.1-8b-base-quantized.w4a16-1.5.0 Quantized version of ibm-granite/granite-3.1-8b-base. It achieves an average s			
prompt			
llama-3.1-8b			
llama-guard-3-1b			
meta-llama-3-8b-instruct			
codellama-34b-python-hf			
granite-3.3-8b-instruct			
codellama-34b-instruct-hf			

Version 1

Version 1 · Deploy · Actions

Details · Deployments

Description

No description

Labels

inference, int4, text-generation, 2 more

Properties · Add property

Key	Value
License	https://huggingface.co/meta-lla...
Provider	Red Hat AI
Registered from	Model catalog
Source model	Llama-3.1-8B-Instruct-quantized.w4a16
Source model version	1.5.0

Version ID

86

Registered from

Llama-3.1-8B-Instruct-quantized.w4a16 (1.5.0) in Model catalog

Model location

URI

oci://registry.redhat.io/rhelai1/modelcar-llama-3-1-8b-instruct-quantized-w4a16:1.5

Source model format

Model format

No model format specified

Model format version

No model format version specified

Author

yyi@redhat.com

Last modified

3 days ago

Registered

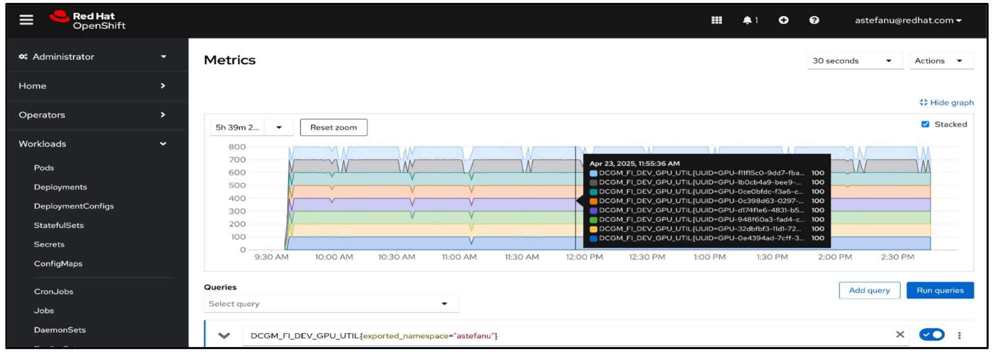
3 days ago

Registry is GA in OpenShift AI in the 3.0 release

GPUaaS - enhanced scheduling, management and observability

GPUaaS

- Enables efficient management and allocation of GPU resources for a variety of AI workloads: workbenches, training/tuning, model serving
- Supports both whole and fractional GPU allocation
- Observability tools for resource optimization and to facilitate chargeback scenarios



Apr 23, 2025, 11:55:36 AM

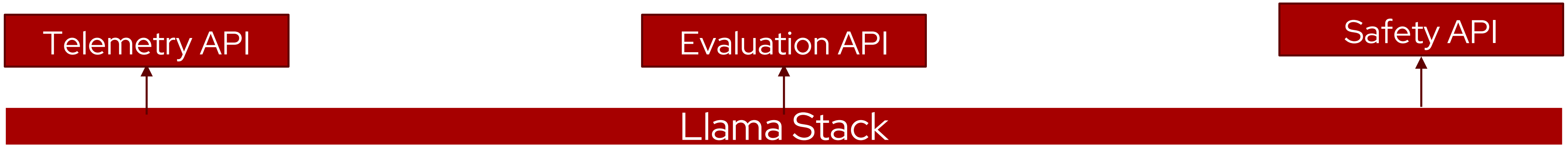
```
DCGM_FI_DEV_GPU_UTIL{UUID=GPU-f11f15c0-9dd7-fba...  100
DCGM_FI_DEV_GPU_UTIL{UUID=GPU-1b0cb4a9-bee9-...     100
DCGM_FI_DEV_GPU_UTIL{UUID=GPU-0ce0bfdc-f3a6-c...    100
DCGM_FI_DEV_GPU_UTIL{UUID=GPU-0c398d63-0297-...     100
DCGM_FI_DEV_GPU_UTIL{UUID=GPU-d174f1e6-4831-b5...   100
DCGM_FI_DEV_GPU_UTIL{UUID=GPU-948f60a3-fad4-c...    100
DCGM_FI_DEV_GPU_UTIL{UUID=GPU-32dbfbf3-11d1-72...   100
DCGM_FI_DEV_GPU_UTIL{UUID=GPU-0e4394ad-7cff-3...    100
```

```
DCGM_FI_DEV_GPU_UTIL{exported_namespace="astefanu"}
```

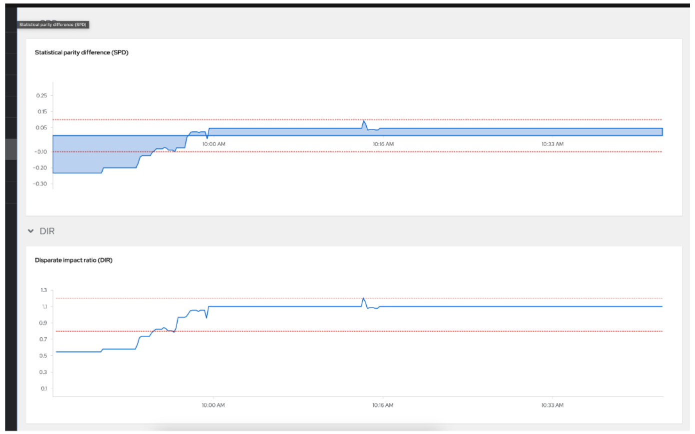
Why does it matter?

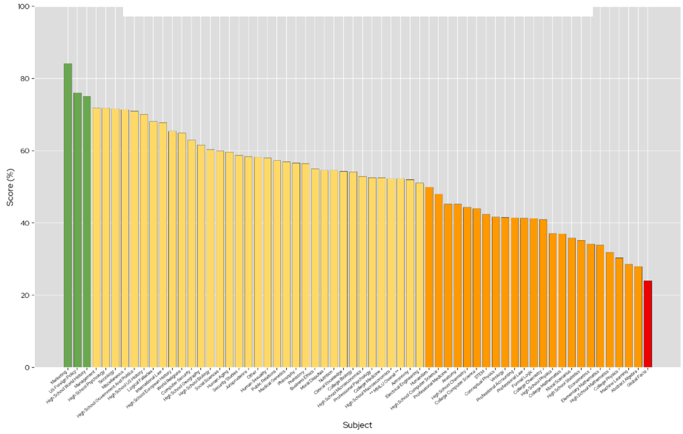
- ***Improves Resource Utilization***: Reclaiming idle GPUs and optimizing allocation to reduce waste
- ***Supports the complete AI Lifecycle***: Handling workloads from notebooks to model serving
- ***Provides Visibility***: Offering metrics for both data scientists and administrators

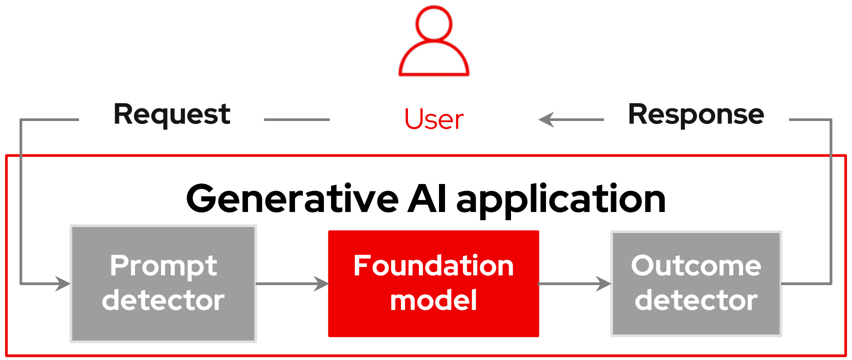
Responsible AI and governance with Red Hat AI



Red Hat AI







AI Monitoring

Monitors tabular model inferences with customizable metrics for bias (outcome disparities) and drift (deployment vs. training data differences)

AI Evaluation

Perform a huge variety of evaluation tasks over LLMs to understand and quantify their knowledge, capabilities, and behaviors

Guardrails for AI

Customizable guardrails framework to moderate interactions between users and generative AI models, ensuring secure, compliant, and efficient operations

OpenShift AI - What's Next

RHOAI 3.0

4Q 2025

Inference

- Distributed inference - llm-d GA
- Additional validated models
- LLM Compressor images GA
- MaaS (Dev preview)

Model development & Alignment

- Modular & extensible solns for data ingestion, SDG, training & evaluation
- RAG - OpenAI APIs, local Milvus vector DB
- RAGAS evaluation framework
- Docling for document processing (Dev Preview)
- Feature Store GA

Agentic

- Llama Stack APIs
- AI Hub UI & gen AI studio
- MCP support & MCP Server access in gen AI studio

AI platform

- Model registry & catalog GA
- Platform metrics foundation - centralized observability infrastructure
- Expand platform support - ARM / Grace Hopper, IBM Z (GA), Power (TP)
- Hardware Profiles GA

RHOAI NEXT

1H 2026

Inference

- Llm-d enhancements - eg. autoscaling, multi modal model support
- KEDA GA including UI
- OOTB ML Server runtime for expanded predictive model support
- MaaS GA - platform admin UI, external models mgmt

Model development & alignment

- Enhanced experiment tracking (MLflow)
- Docling support (GA)
- Prompt engineering capabilities
- Kubeflow Trainer v2 and checkpointing capabilities
- Enhanced model customization experience

Agentic

- MCP Server catalog & registry
- Agent catalog & registry
- Enhanced AI Playground - eg. model comparison, guardrails
- Knowledge source creation (Gen AI Studio UI)
- Inclusion of more Llama Stack providers across modules

AI platform

- Enhanced native observability - model/agent monitoring, GPUaaS
- Integration of NVIDIA Nemo Guardrails
- Support Kubeflow Spark operator
- Workbenches 2.0 - next-gen workbench experience
- RH Trusted Software Supply Chain integration
- GPU partitioning GA (based on DRA)

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

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.

in linkedin.com/company/red-hat

▶ youtube.com/user/RedHatVideos

f facebook.com/redhatinc

𝕏 twitter.com/RedHat

What’s New and Next in GPU-as-a-Service (GPUaaS)

Highlights

RHOAI 3.0

Q4 2025

- **Ray and Codeflare** updates:
  - **Architectural Simplification** – Complete CodeFlare Operator removal for cleaner, more maintainable platform
  - CodeFlare SDK providing a more aligned user experience by focusing on **RayJob**, while providing full access to RayClusters
- **Accelerator Slicing**: for all MIG-enabled devices, via the NVIDIA operator
- **Kueue**: supporting more workloads:
  - Ray Training Jobs
  - Training Operator-based Jobs
  - Inferences Services and Workbenches

RHOAI NEXT

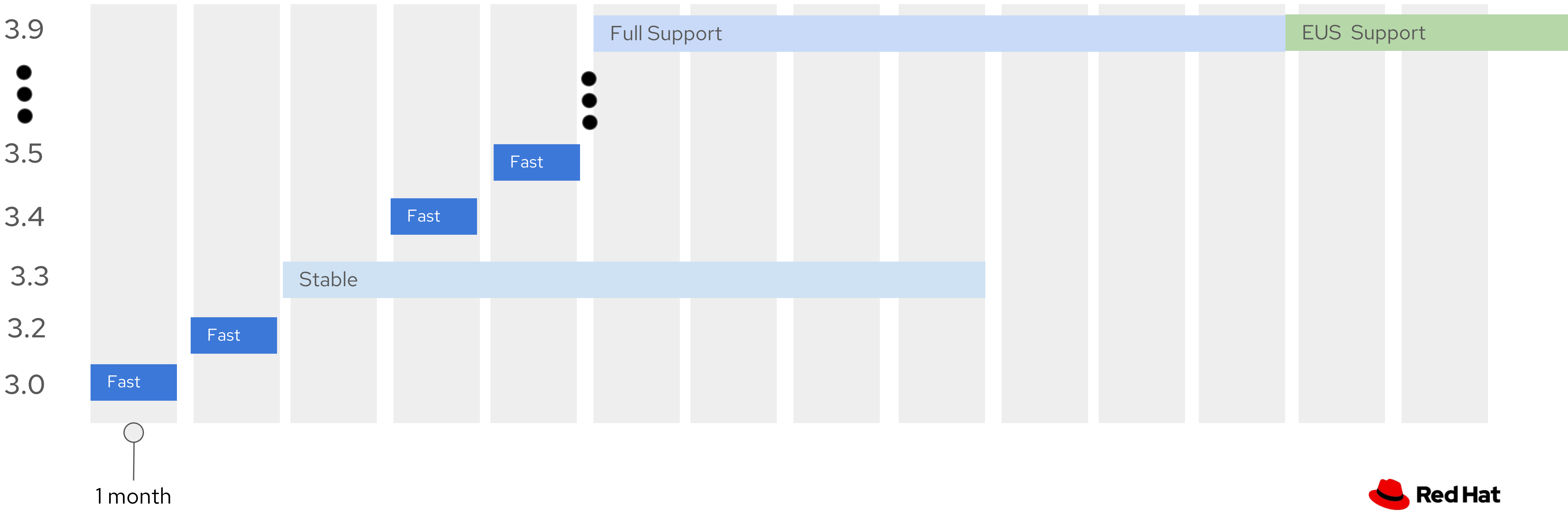
H1 2026

- **Kubeflow Trainer v2** and **Red Hat Build of Kueue 1.2**:
  - major technology updates for workload scheduling
  - converged API for Training Jobs
- **GPU and Workload specific Observability**: based on RHOAI observability and accompanied by data scientists- and admin-centric views
- **Model Checkpointing**: to support better workload management, like Job-preemption and resume, to increase GPU utilization
- **Dynamic Resource Allocation (DRA)**: focus on this technology for (not only) GPU partitioning, long lead-time, as a lot of OCP work is to be done!

Lifecycle Updates

Red Hat AI products life cycle

Fast: Monthly cadence, 1 month life cycle	**Stable**: 3 Month Cadence, 7 months life cycle	**EUS:** 12 month cadence, 18 month life cycle

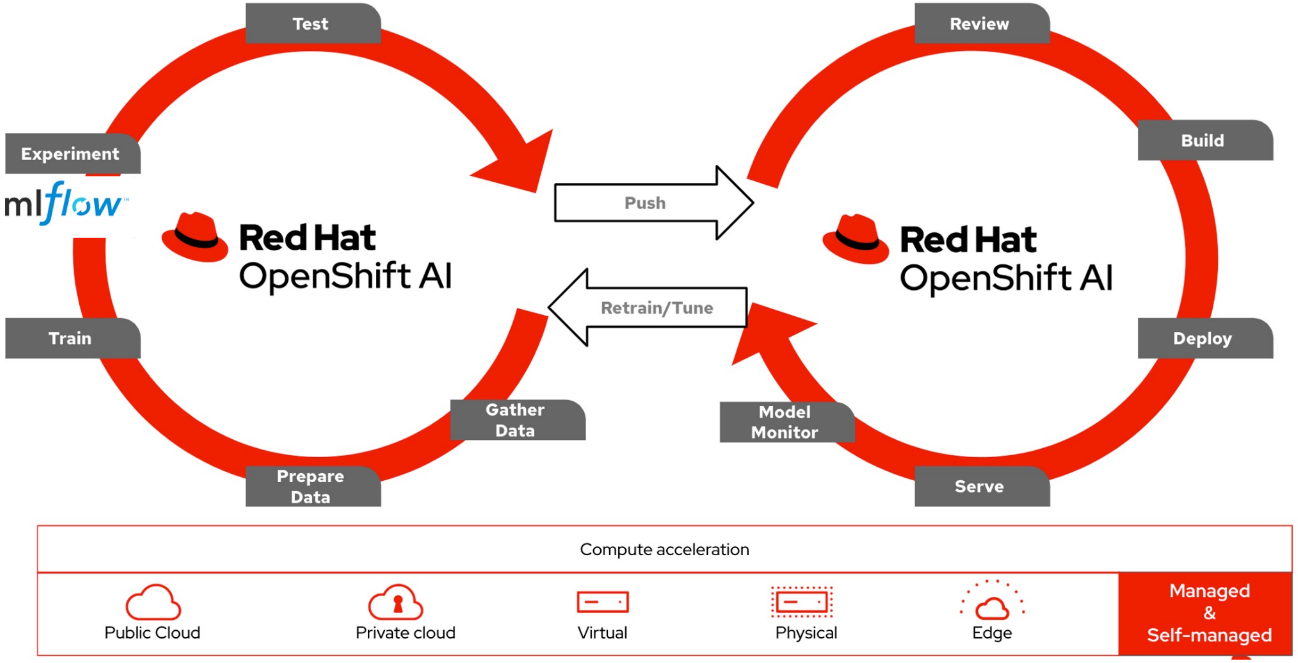


What’s New and What’s Next in Experiment tracking

MLFlow integration

What’s next?

- Experiment tracking for more than just pipelines
- Integration of MLFlow on RH AI for inner loop experiment tracking
- More GenAI visualizations and metrics
- AIP and MLFlow integration



Introducing multi-architecture

IBM Power/ Z

Increasing flexibility and choice with an open source approach

- **In other to have a true hybrid platform, we are enhancing our release process to support multiple architectures. Starting with IBM Z and ARM. IBM Power shortly after.**

Flexibility

Access to cutting-edge open source innovations to keep up with a fast moving market.

Choice

Access to an open ecosystem of communities, technology providers, ISVs and customers.