Red Hat Summit

Connect

Agentic AI in Action

Red Hat & Intel Shaping the Future of Enterprise AI

Zurich 14 January 2026

Red Hat | intel.

Can Yurtseven

AI / Cloud Senior Solution Architect, EMEA CTO
Intel

Codrin Bucur

Principal AI Specialist Solution Architect, EMEA
Red Hat

Intel - RH Partnership

Open source software: Intel is committed

Intel® has a long history with Linux®, actively participating in open source development and collaboration with the Linux community, to ensure hardware is well-supported and delivers optimal performance on Linux-based systems.

Intel contributes to more than 100 different open source projects, from the Linux kernel to cloud orchestration and plugins for Kubernetes.

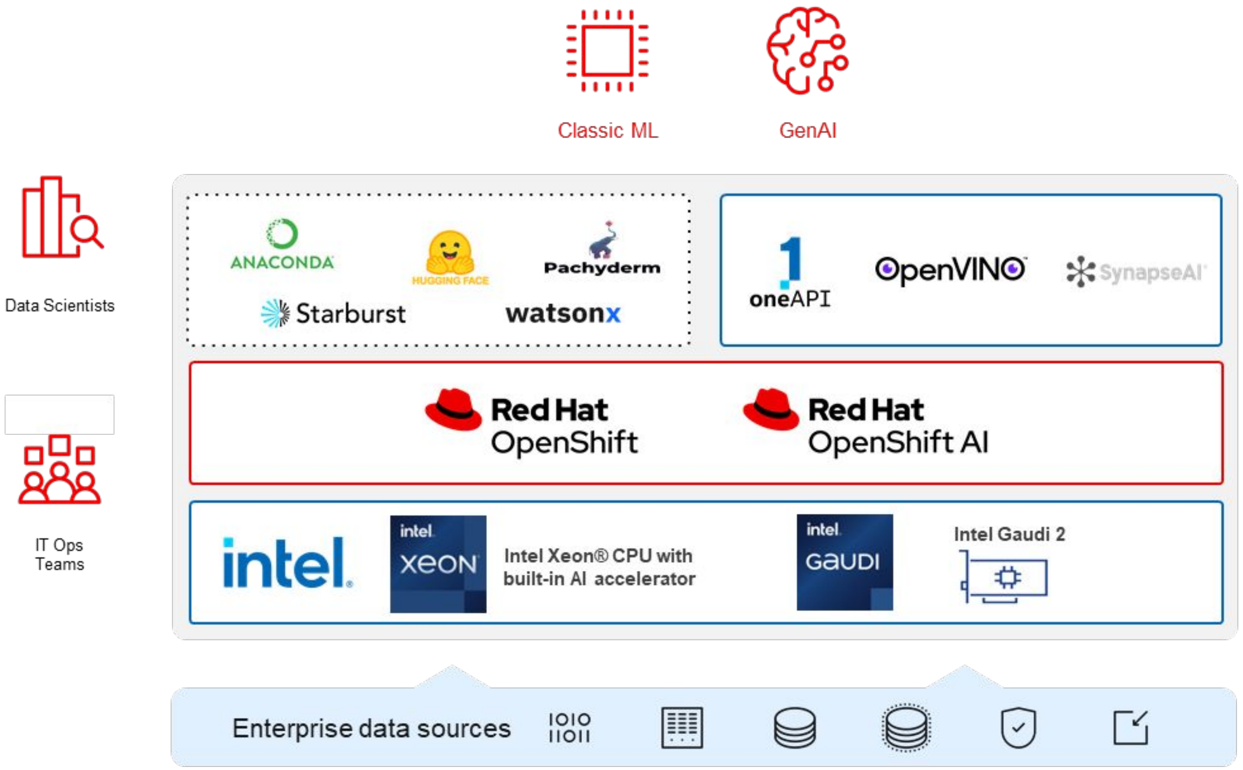
Over **25** years
of collaboration

Intel is one of the key contributors
Linux Kernel Corporate Contributor since 2007[1]

Source:

[1] Open source software: Intel engineers are engaged. Intel Innovation Digital 2021.

Real Customer Example: AI Sweden



Classic ML

GenAI

Data Scientists

IT Ops Teams

ANACONDA

HUGGING FACE

Pachyderm

Starburst

watsonx

oneAPI

OpenVINO

SynapseAI

Red Hat OpenShift

Red Hat OpenShift AI

intel

intel xeon

Intel Xeon® CPU with built-in AI accelerator

intel Gaudi

Intel Gaudi 2

Enterprise data sources

- Collaborating to deliver AI solutions
- Deeper, product collaboration focused on customer enablement with OpenShift AI, Intel Xeon, Gaudi 2 and the Intel AI Suite
- Testing, validation, and proof of concepts
- Receive support for building AI applications

Intel's AI Strategy and Capabilities

AI as a Strategic Priority

Massive transformation in AI

Redefining every layer of the stack deployed from AI PC, to edge, to data center

CPU relevance in AI ecosystem

Open heterogeneous strategy to deliver **systems,** software & GPUs

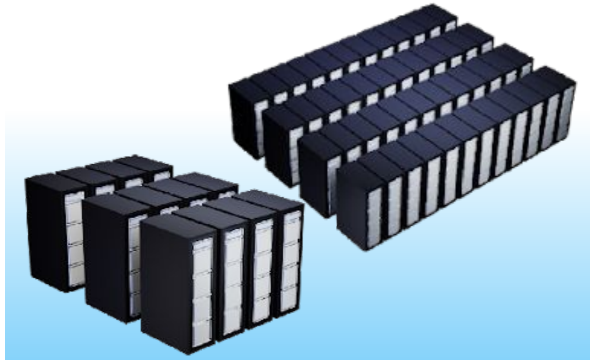
Bringing AI Everywhere

Intel’s AI Strategy

AI PC Node
AI Developer Productivity & Light Inference

Node
Fine-tuning, Inference

Cluster
Light Training, Tuning, Peak Inference



Super Cluster
Training, Tuning, Peak Inference

Mega Cluster
Large Scale Training & Inference

AI PC

Broadest AI SW Ecosystem

ENTERPRISE AI & EDGE AI

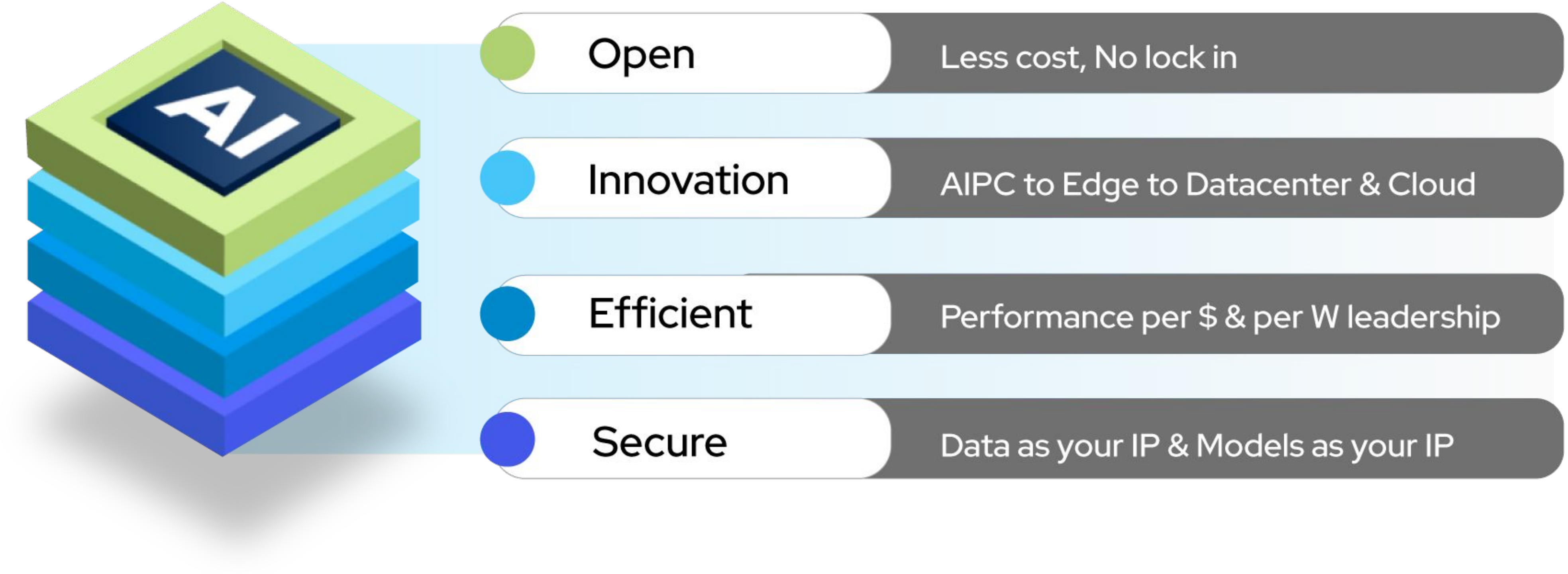
Open Standard, “Ready to Use”

DATA CENTER AI

AI Open, Scalable Systems & Reference Arch

intel CORE ULTRA | intel ARC GRAPHICS | intel CORE ULTRA | intel XEON | intel GAUDI | intel XEON | intel GAUDI | intel IPU

Intel's AI Strategy



Open Less cost, No lock in

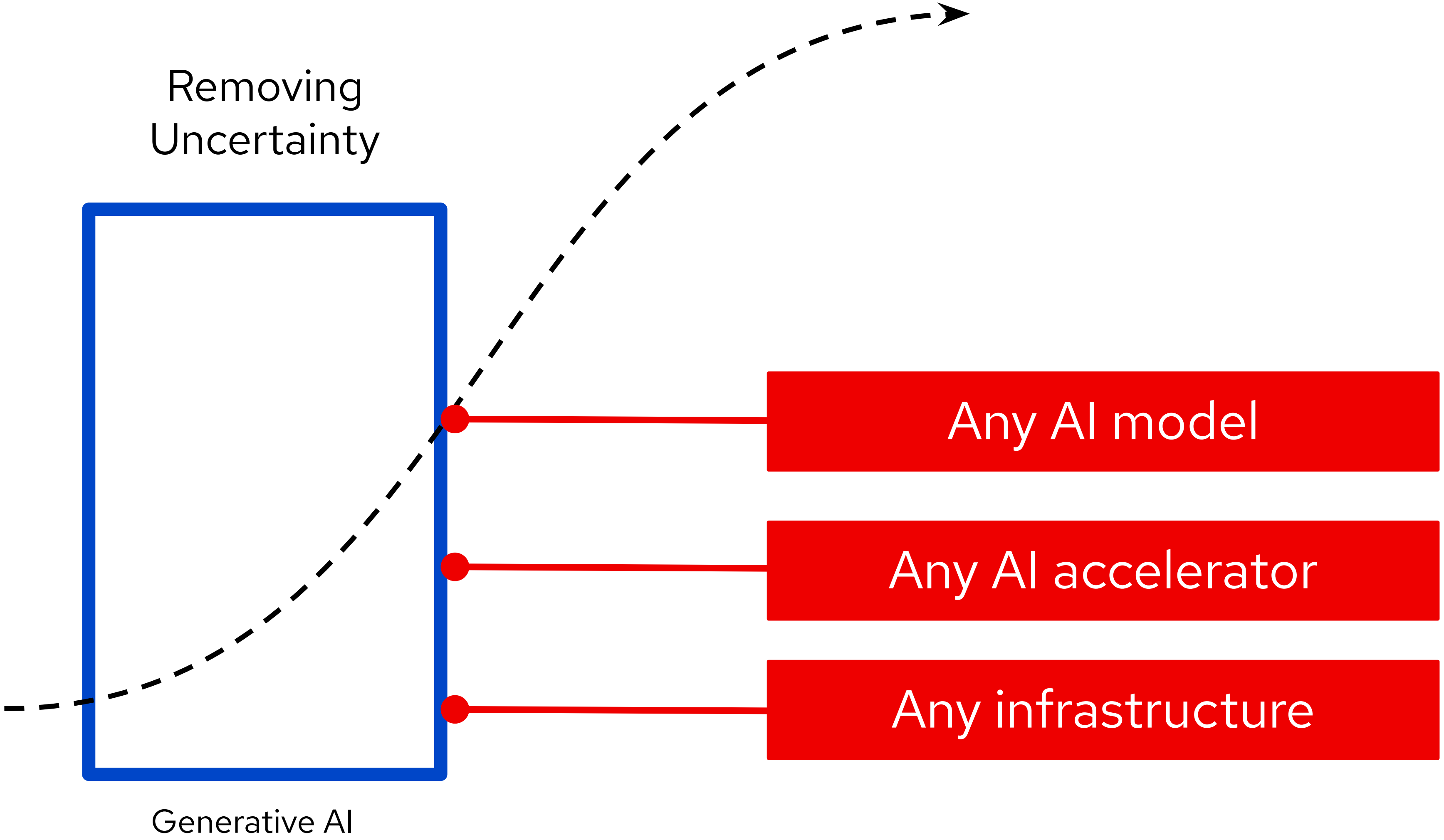
Innovation AIPC to Edge to Datacenter & Cloud

Efficient Performance per $ & per W leadership

Secure Data as your IP & Models as your IP

Red Hat's AI Strategy and Capabilities

Red Hat AI - Enabling AI Success



Red Hat AI

Red Hat AI
Inference Server

Red Hat
Enterprise Linux AI

Red Hat
OpenShift AI

Trusted, Consistent and Comprehensive foundation

NVIDIA AMD intel Hardware Acceleration Google aws IBM

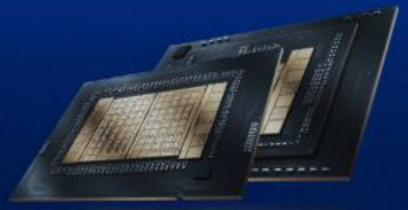
Physical **Virtual** **Private Cloud** **Public Cloud** **Edge**

* NVIDIA, AMD, Intel, Google TPU supported in Red Hat AI. AWS Inferentia/Neuron IBM AIU are on our roadmap

Intel Data Center AI

Intel AI Data Center Portfolio

intel xeon



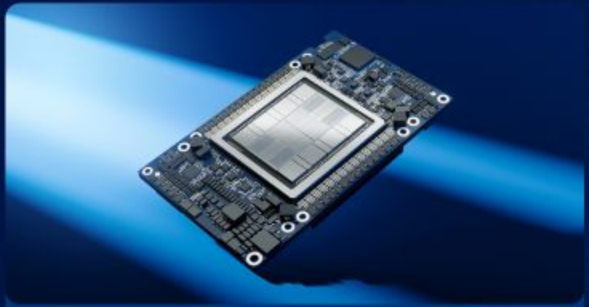
Data Center Accelerators

intel ARC GRAPHICS

Open heterogeneous strategy to deliver **systems, software** & **GPUs**

Intel® Gaudi® 3 AI Accelerator

Delivered to Meet Diverse Needs



Accelerator Card

OAM-Compliant
(HL-325L)

Universal Baseboard

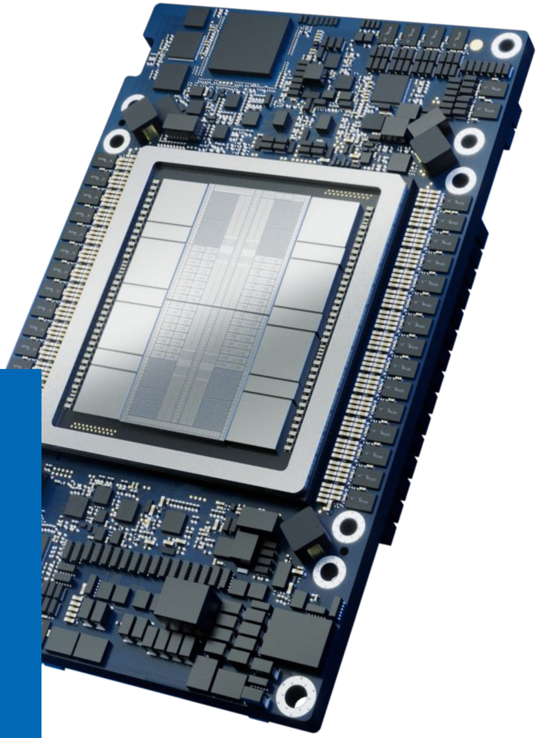
HLB-325

PCIe CEM

Add-In Card
(HL-338)

Intel® Gaudi® 3 AI Accelerator: AI Inferencing

Price Performance Advantage



Up to
43%

Higher throughput
(tokens per second)

on IBM Granite-3.1-8B-Instruct

vs. leading GPU competitor
with small context sizes

Up to
120%

More cost efficient
(tokens per dollar)

on Mixtral-8x7B-Instruct-v0.1

vs. leading GPU competitor
with long input and short output sizes

Up to
92%

More cost efficient
(tokens per dollar)

on Llama-3.1-405B-Instruct-FP8

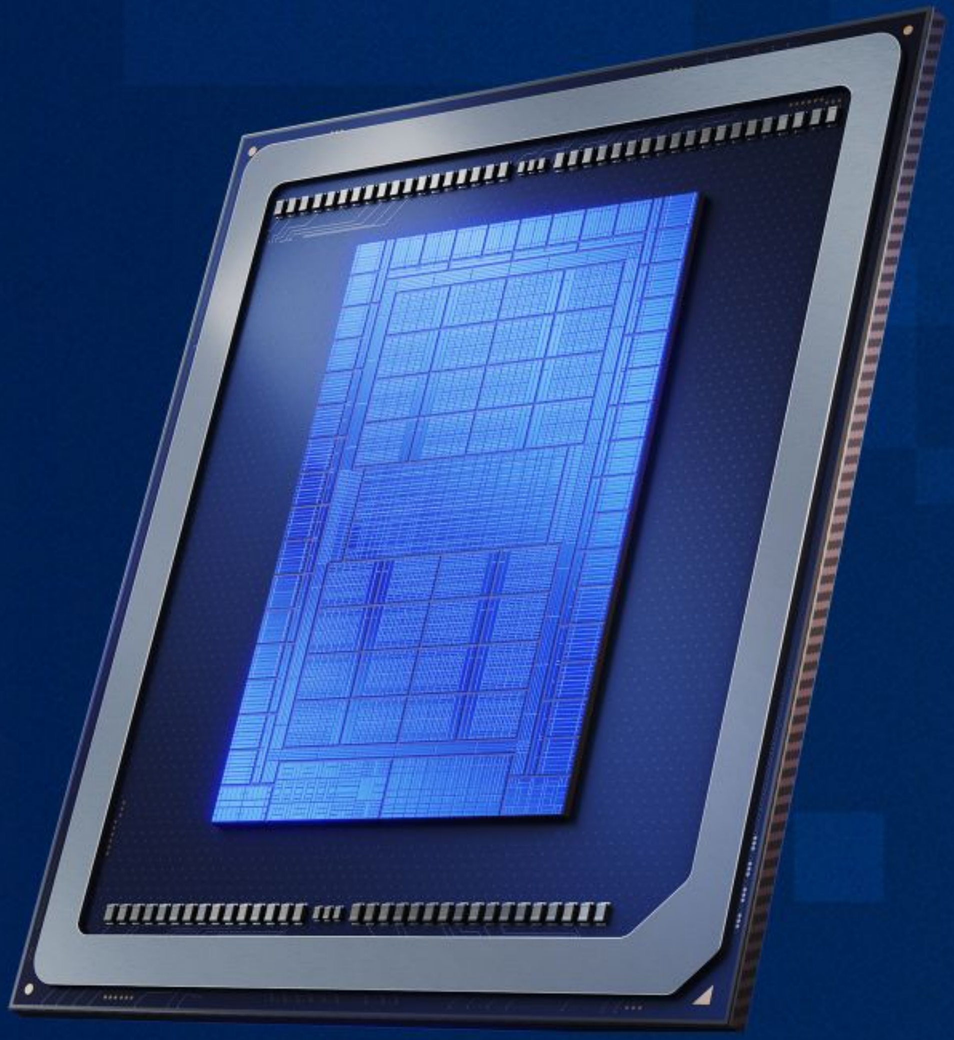
vs. leading GPU competitor
with large context sizes

Source: Signal65 Lab Insight Whitepaper - Intel Gaudi 3 AI Accelerator at Scale on IBM Cloud, Intel-commissioned study by Signal65, published April 22, 2025. Reported numbers are inferencing results on Intel Gaudi 3 vs. NVIDIA H200. See source for workloads and configurations. Results may vary.

Next Gen

Intel® Data Center GPU

Designed for tokens/watt built on a reliable open software stack



160GB LPDDR5x
to meet the needs of memory intensive AI workloads

Latest Generation X^e 3 IP
General purpose GPU architecture for competitive perf/watt

Widest Range AI datatypes
Support from FP4/MXP4 to FP32 and FP64

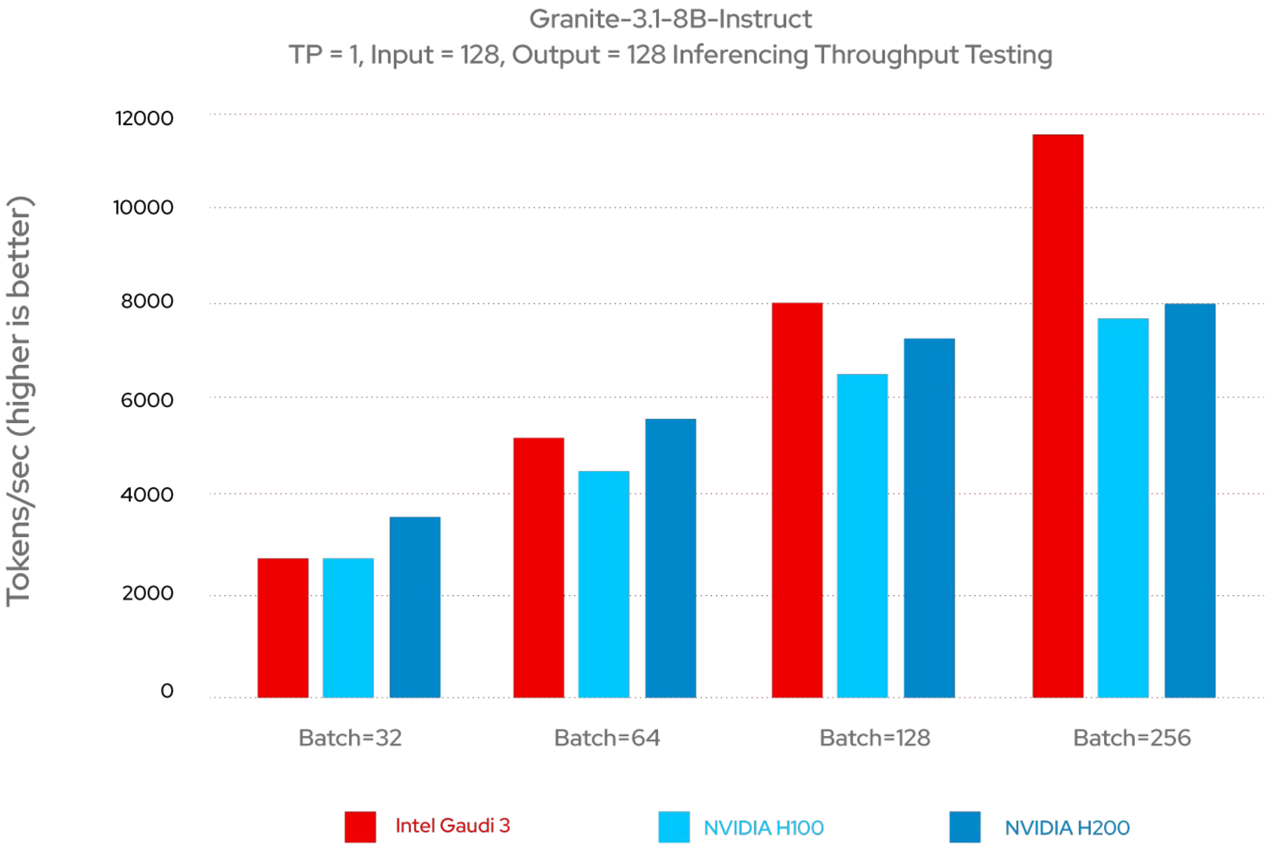
Up to **43% higher**

throughput than NVIDIA H200

Up to **52% higher**

throughput than NVIDIA H100

For lightweight AI Use Cases



*Source: NV H100 and H200 comparisons based on Signal65 Lab Insight: Intel Gaudi 3 Accelerates AI at Scale on IBM Cloud. April 2025.
Reported numbers are inferencing results for IBM Granite-3.1-8B-Instruct on Intel® Gaudi® 3 vs NVIDI H100 GPU and NVIDIA H200 GPU. Refer to this link for the latest published Gaudi3 performance https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html
Pricing estimates based on publicly available information and Intel internal analysis.
Results may vary.

Up to **36% higher**

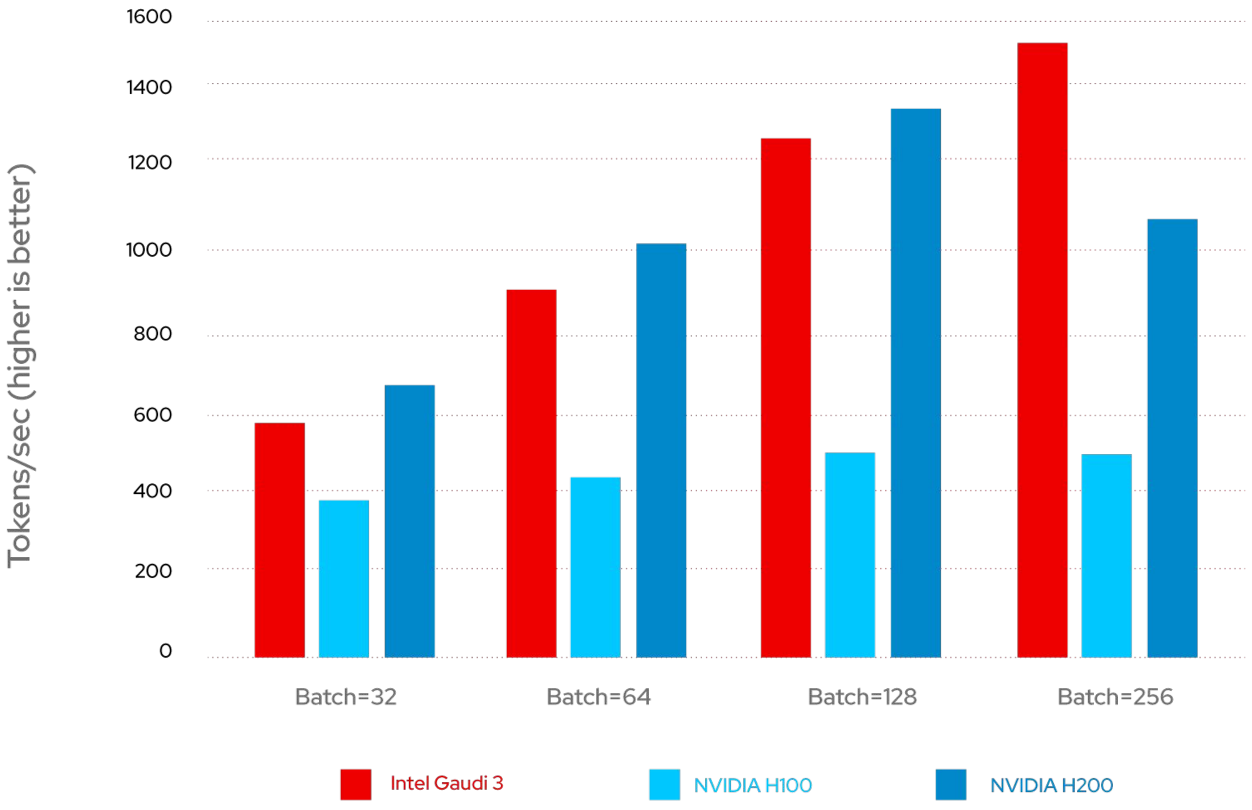
throughput than NVIDIA H200

Up to **200% higher**

throughput than NVIDIA H100

For Large AI Workloads

GLlama-3.1-405B-Instruct-FP8
TP = 8, Input = 4096, Output = 2048 Inferencing Throughput Testing



*Source: NV H100 and H200 comparisons based on Signal65 Lab Insight: Intel Gaudi 3 Accelerates AI at Scale on IBM Cloud. April 2025.
Reported numbers are inferencing results for IBM Granite-3.1-8B-Instruct on Intel® Gaudi® 3 vs NVIDI H100 GPU and NVIDIA H200 GPU. Refer to this link for the latest published Gaudi3 performance https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html
Pricing estimates based on publicly available information and Intel internal analysis.
Results may vary.

Intel® Xeon® 6 Processor

1.9x

higher performance per watt at a typical 40% server utilization vs. prior generation

Designed for Efficiency

2.5x

higher HPC performance vs. prior generation

Significant Performance Leaps

5.5x

higher AI Inferencing performance vs. AMD EPYC

Unmatched Performance

See [9G2, 9H9, 9A3] at intel.com/processorclaims: Intel Xeon 6. Results may vary

Intel® AMX Accelerates DEEP LEARNING Use Cases

Intel® Advanced Matrix Extensions (AMX)

BF16, INT8, and FP16 precision

- Recommender Systems
- Natural Language Processing
- Image Recognition Object Detection

Intel® Advanced Vector Extensions (AVX-512)

FP32 and FP64 precision

- Data Analytics
- Classical Machine Leaming

Many DL workloads are "mixed precision" and
5th Gen Xeon can seamlessly transition between AMX and AVX-512 as needed

Resolve Customer Queries Faster with More Concurrent Users in Your LLMs and Agents

- Get superior performance for batch, real-time inference, and training for small and medium language models with Intel® Xeon® processors.
- Use your CPU for cost-effective model updates.

Large language models (LLMs)

Intel Xeon 6 vs. AMD EPYC Turin

Llama2-7B

Up to

1.38x

higher throughput

with Intel Xeon 6980P
vs. AMD EPYC 9965'

Intel Xeon 6 vs. 5th Gen Intel Xeon

GPTJ-6B

Up to

2x

Higher performance

Intel Xeon 6980P
vs. Intel Xeon 8592+[2]

Llama-13B

Up to

2x

Higher performance

Intel Xeon 6980P
vs. Intel Xeon 8592+[2]

Llama2-7B

Up to

2.3x

Higher training performance

Intel Xeon 6980P
vs. Intel Xeon 8592+[3]'

5th Gen Intel Xeon vs. 3rd Gen Intel Xeon

Llama2-13B

Up to

2.1x

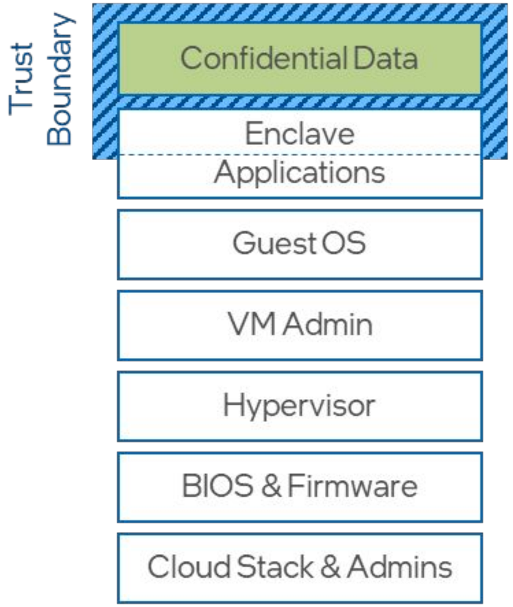
real-time inference performance speedup

5th Gen Intel Xeon vs.
3rd Gen Intel Xeon[4]

Intel Confidential Computing

App Isolation

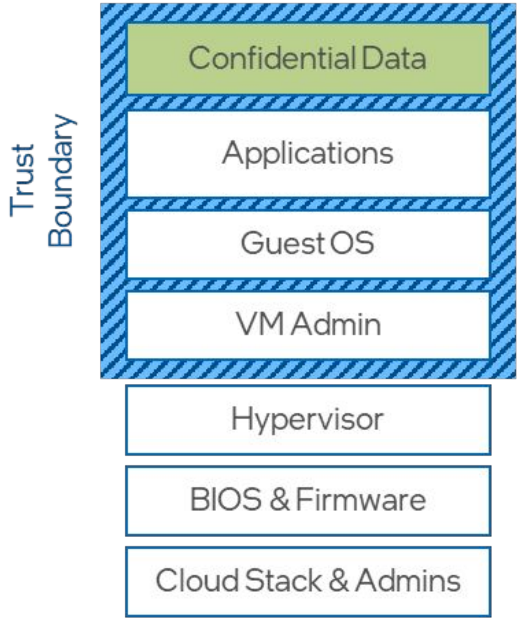
Intel® SGX


Trust Boundary
Confidential Data
Enclave
Applications
Guest OS
VM Admin
Hypervisor
BIOS & Firmware
Cloud Stack & Admins

Smallest trust boundary for greatest data protection & code integrity

VM Isolation

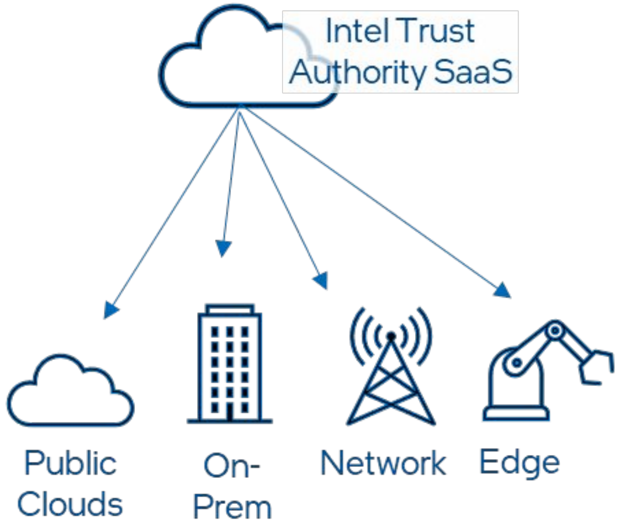
Intel® TDX


Trust Boundary
Confidential Data
Applications
Guest OS
VM Admin
Hypervisor
BIOS & Firmware
Cloud Stack & Admins

Most straightforward path to greater security for legacy apps

Trust Services

Intel® Tiber™ Trust Authority


Intel Trust Authority SaaS
Public Clouds
On-Prem
Network
Edge

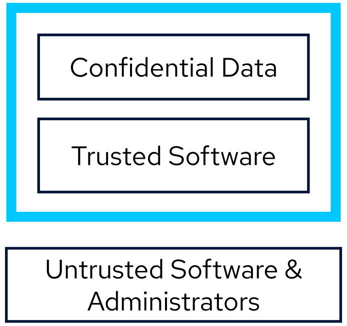
Uniform, independent attestation of trustworthy environments

Founded on Intel's Security-First Development & Lifecycle Support

Confidential AI Helps Protect Data & Models In-Use

Utilizing Confidential Computing for Containers with Intel TDX

Hardware-Based Protection of Data In-Use
With Intel Trusted Domain Extensions (TDX)



+

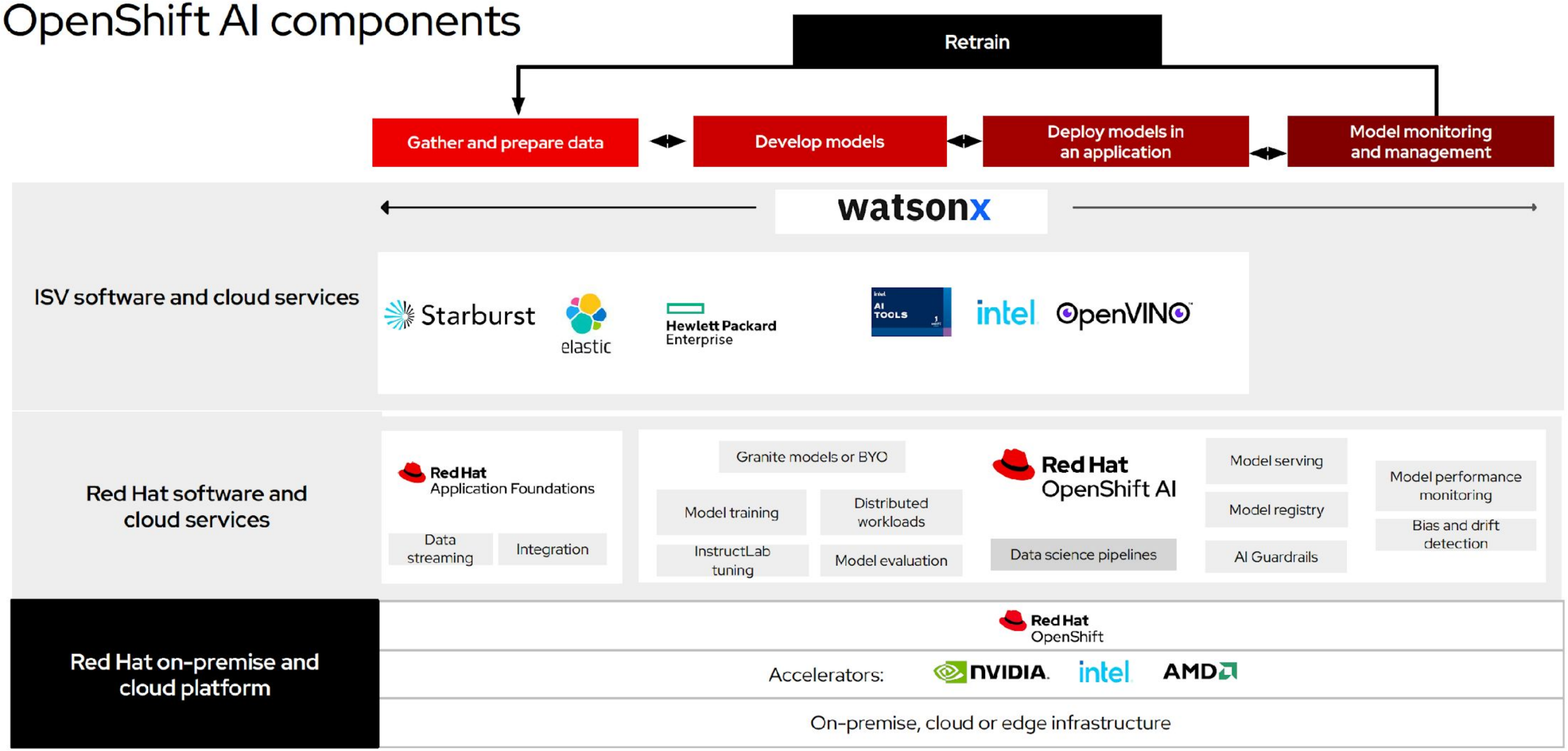
OpenShift / OpenShift AI sandboxed containers

The Confidential Containers project

Confidential Computing is about protecting data in-use.
You do not have to trust the system admins of the providers any longer.

Red Hat AI Platform + Intel AI Software

OpenShift AI components



Retrain

Gather and prepare data ↔ **Develop models** ↔ **Deploy models in an application** ↔ **Model monitoring and management**

watsonx

ISV software and cloud services

Starburst

elastic

Hewlett Packard Enterprise

intel AI TOOLS

intel

OpenVINO

Red Hat software and cloud services

Red Hat Application Foundations

- Data streaming
- Integration

Granite models or BYO

- Model training
- Distributed workloads
- InstructLab tuning
- Model evaluation

Red Hat OpenShift AI

- Data science pipelines
- Model serving
- Model registry
- AI Guardrails
- Model performance monitoring
- Bias and drift detection

Red Hat on-premise and cloud platform

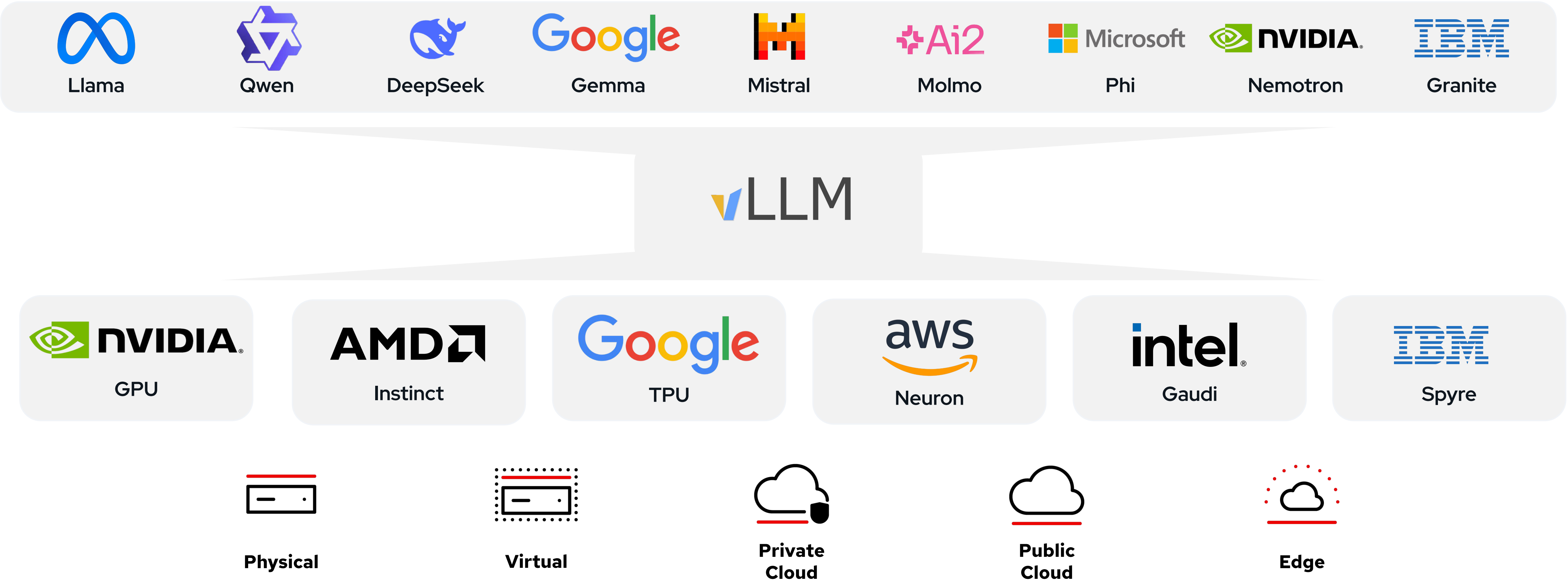
Red Hat OpenShift

Accelerators: NVIDIA intel AMD

On-premise, cloud or edge infrastructure

Red Hat AI the inference engine for the hybrid cloud

vLLM supports the key models on the key hardware accelerators



Llama
Qwen
DeepSeek
Gemma
Mistral
Molmo
Phi
Nemotron
Granite

vLLM

GPU
Instinct
TPU
Neuron
Gaudi
Spyre

Physical
Virtual
Private Cloud
Public Cloud
Edge

Red Hat AI repository on Hugging Face

A collection of third-party validated and optimized large language models

Broad Collection of models

Llama

Qwen

Gemma

Mistral

DeepSeek

Phi

Molmo

Granite

Nemotron

Validated models

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

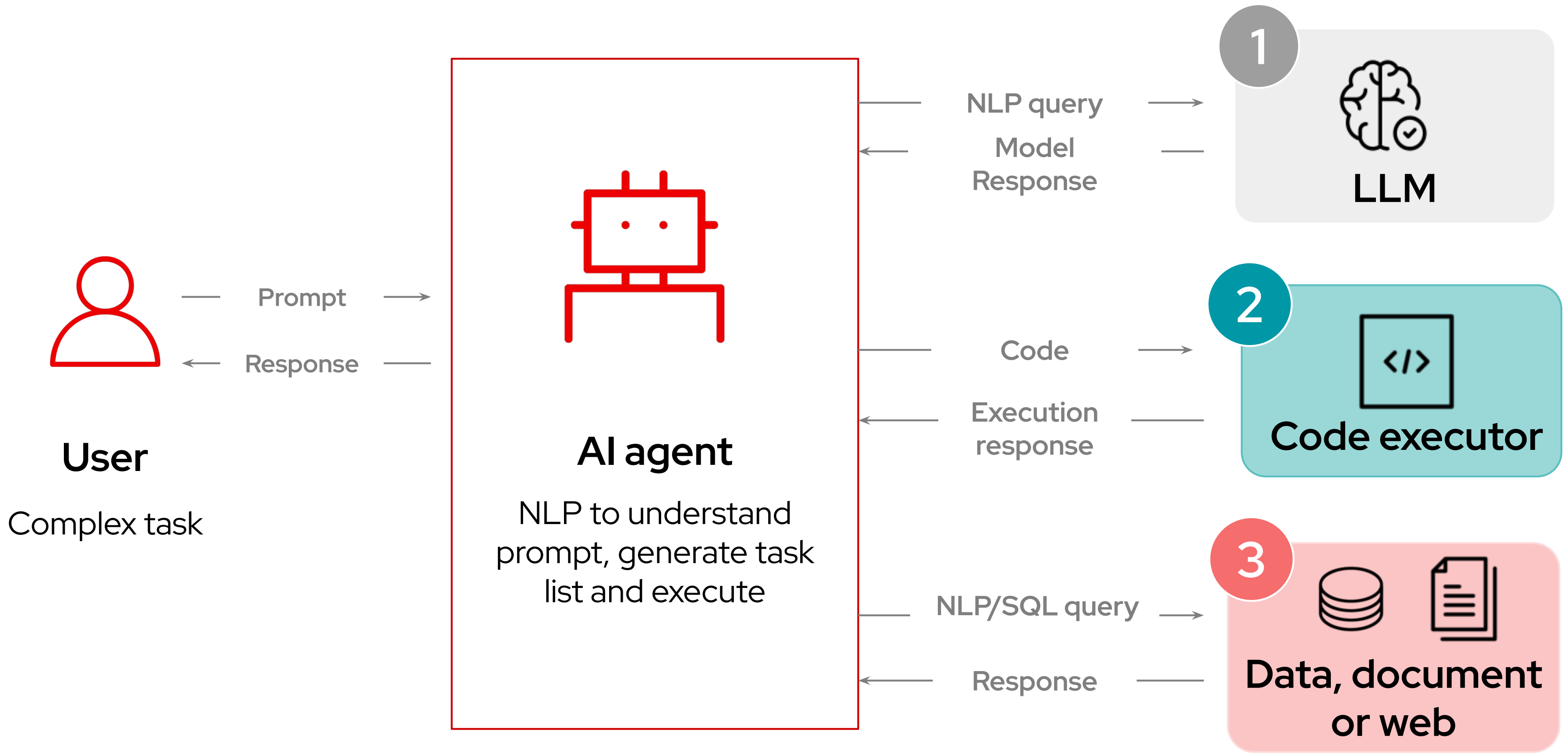
Optimized models

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

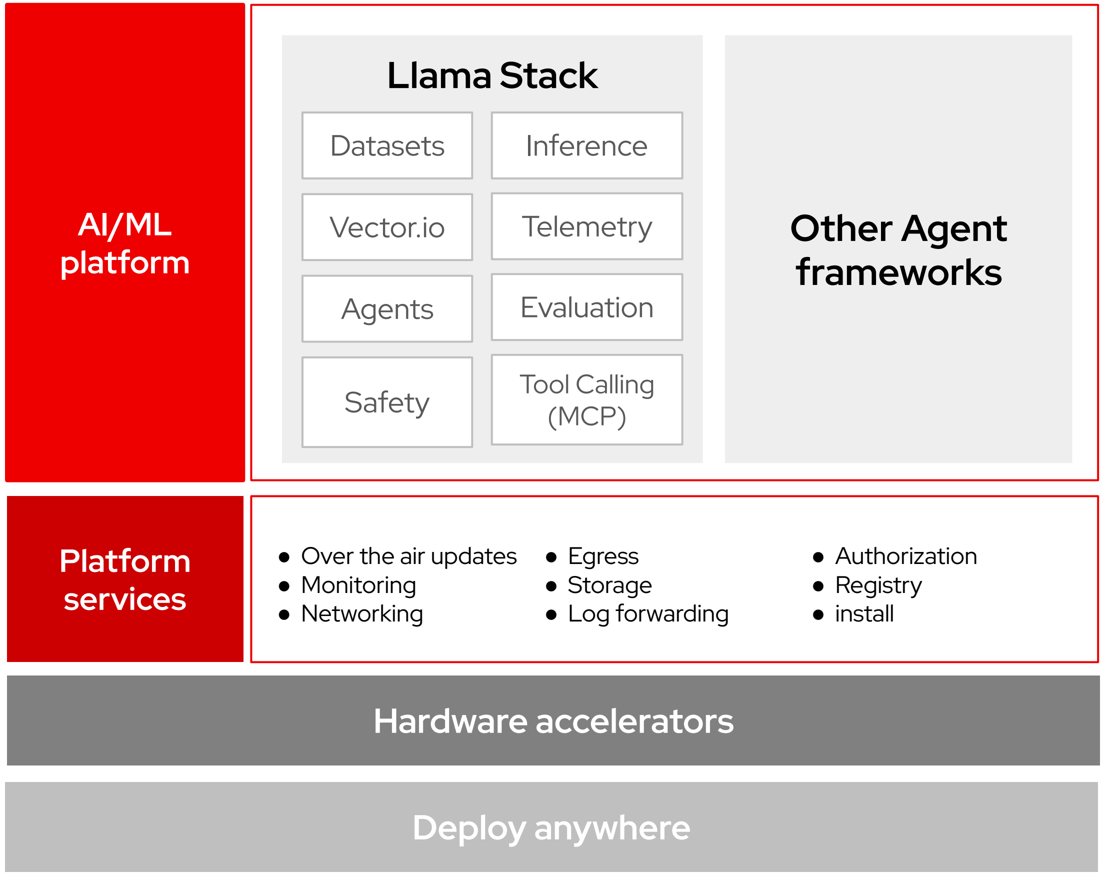
Agentic AI Demo

AI agents integrate models, functions & tools

Gen AI Models, Predictive AI Models, Code Functions, Search & more



A modular approach to building AI agents



AI/ML platform

Llama Stack

- Datasets
- Inference
- Vector.io
- Telemetry
- Agents
- Evaluation
- Safety
- Tool Calling (MCP)

Other Agent frameworks

Platform services

- Over the air updates
- Monitoring
- Networking
- Egress
- Storage
- Log forwarding
- Authorization
- Registry
- install

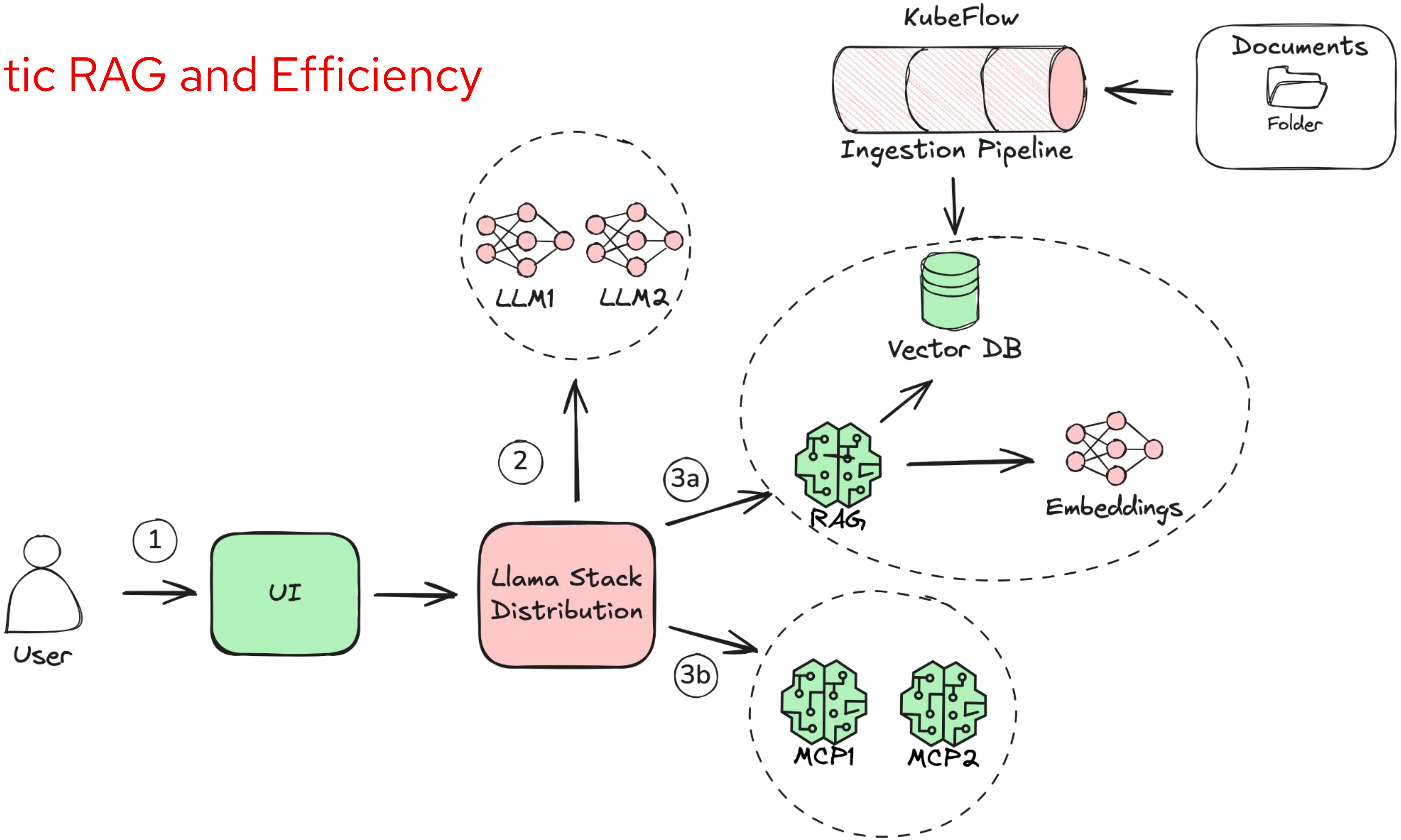
Hardware accelerators

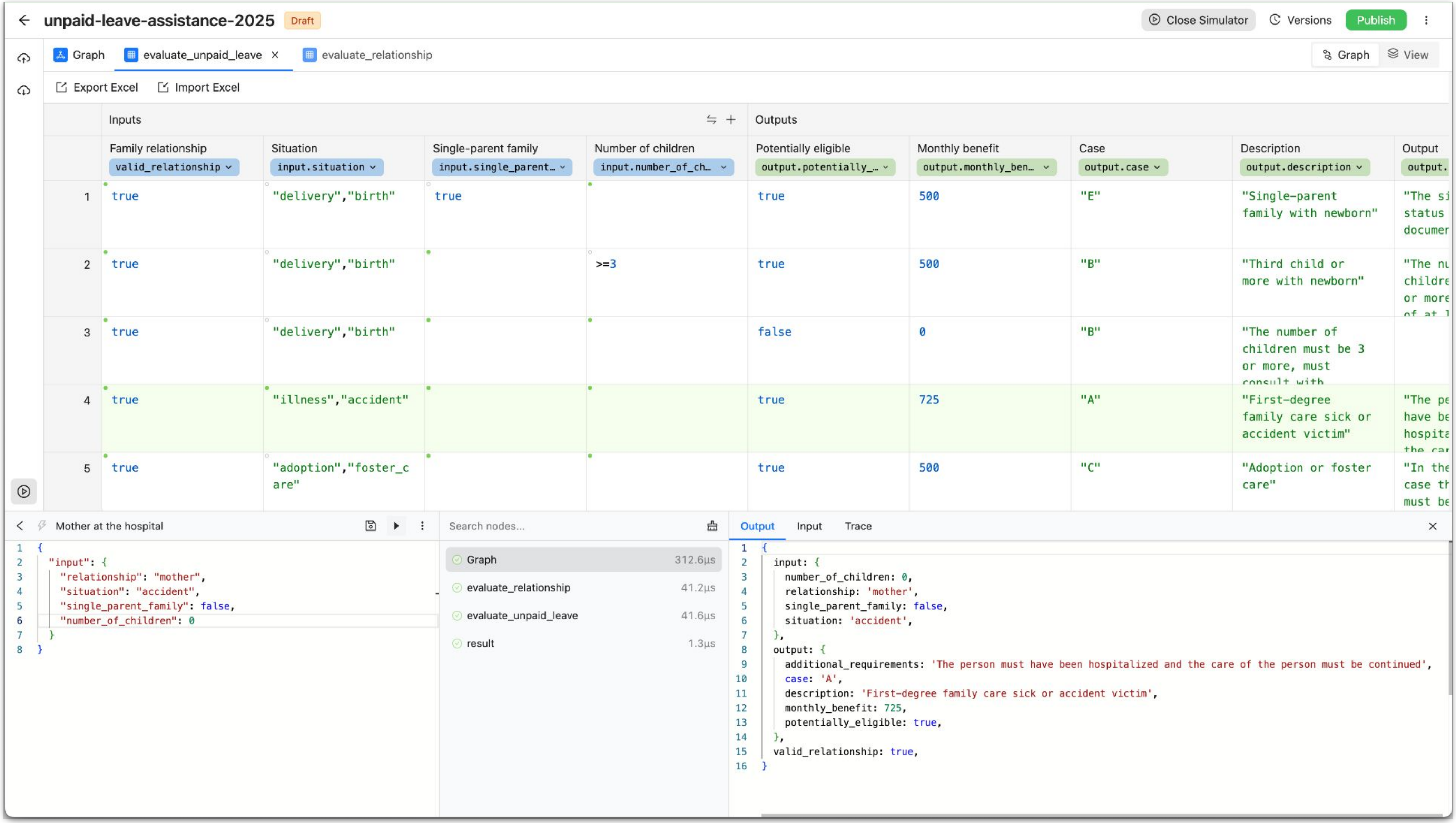
Deploy anywhere

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.

Agentic RAG and Efficiency





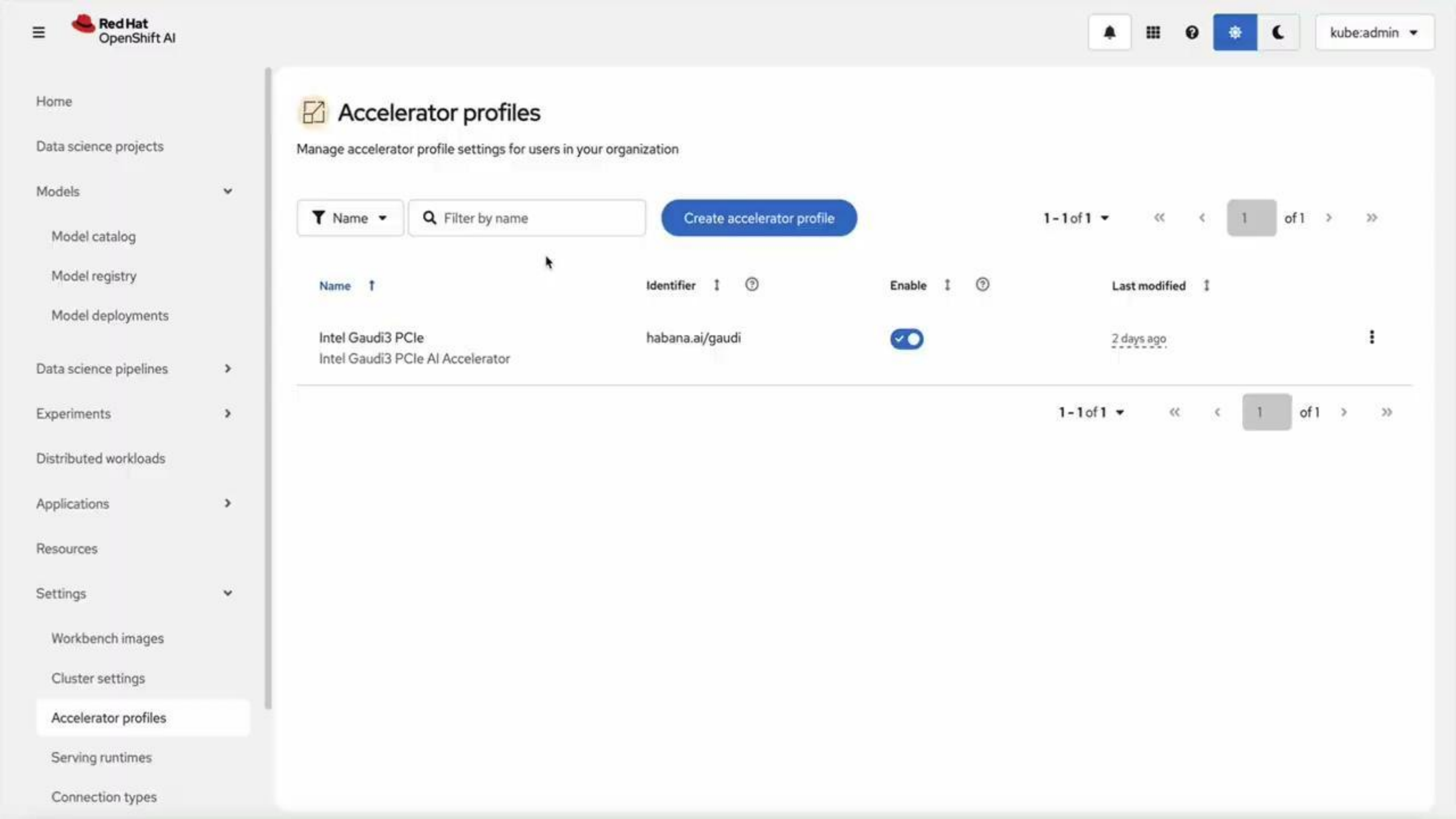
	Inputs				Outputs				
	Family relationship valid_relationship	Situation input.situation	Single-parent family input.single_parent…	Number of children input.number_of_ch…	Potentially eligible output.potentially_…	Monthly benefit output.monthly_ben…	Case output.case	Description output.description	Output output.
1	true	"delivery","birth"	true		true	500	"E"	"Single-parent family with newborn"	"The si status documer
2	true	"delivery","birth"		>=3	true	500	"B"	"Third child or more with newborn"	"The nu childre or more of at l
3	true	"delivery","birth"			false	0	"B"	"The number of children must be 3 or more, must consult with	
4	true	"illness","accident"			true	725	"A"	"First-degree family care sick or accident victim"	"The pe have be hospita the car
5	true	"adoption","foster_care"			true	500	"C"	"Adoption or foster care"	"In the case th must be

Mother at the hospital

```
{
  "input": {
    "relationship": "mother",
    "situation": "accident",
    "single_parent_family": false,
    "number_of_children": 0
  }
}
```

Graph	312.6µs
evaluate_relationship	41.2µs
evaluate_unpaid_leave	41.6µs
result	1.3µs

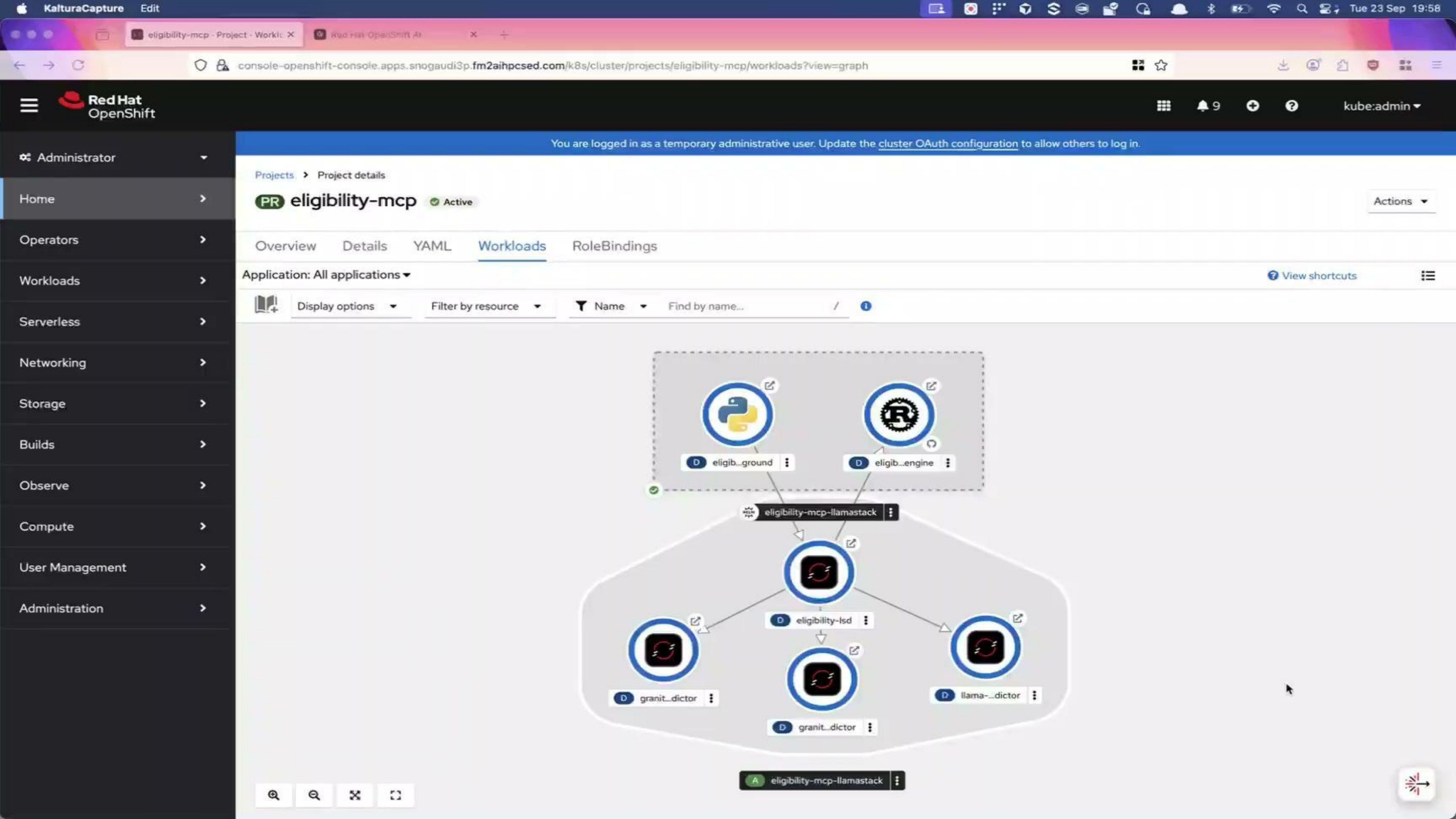
```
{
  input: {
    number_of_children: 0,
    relationship: 'mother',
    single_parent_family: false,
    situation: 'accident',
  },
  output: {
    additional_requirements: 'The person must have been hospitalized and the care of the person must be continued',
    case: 'A',
    description: 'First-degree family care sick or accident victim',
    monthly_benefit: 725,
    potentially_eligible: true,
  },
  valid_relationship: true,
}
```



Accelerator profiles

Manage accelerator profile settings for users in your organization

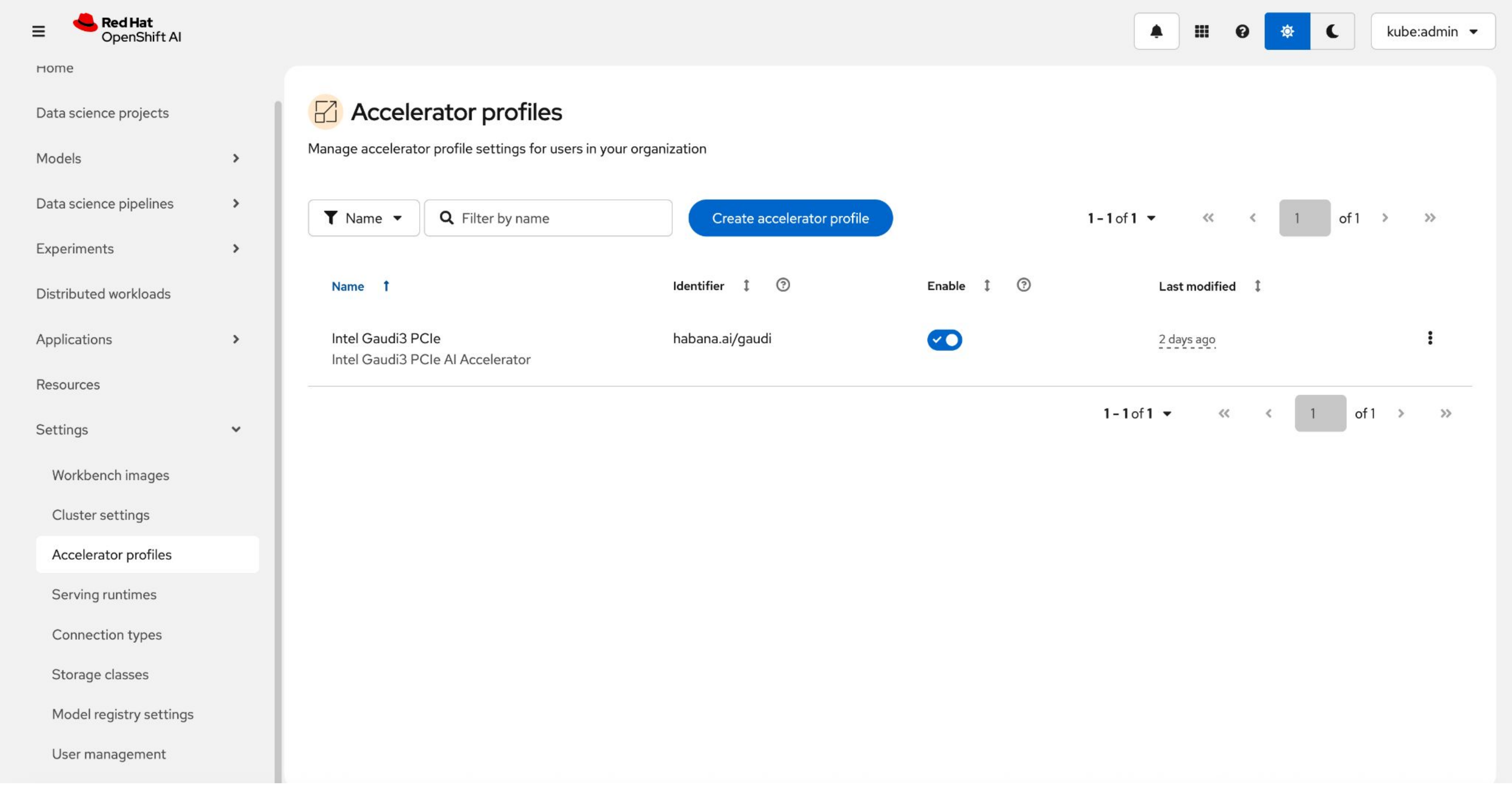
Name	Identifier	Enable	Last modified
Intel Gaudi3 PCIe Intel Gaudi3 PCIe AI Accelerator	habana.ai/gaudi	☑	2 days ago



You are logged in as a temporary administrative user. Update the cluster OAuth configuration to allow others to log in.

Demo Screenshots for Distributing the Deck (Unskip screenshot slides when generating PDF)

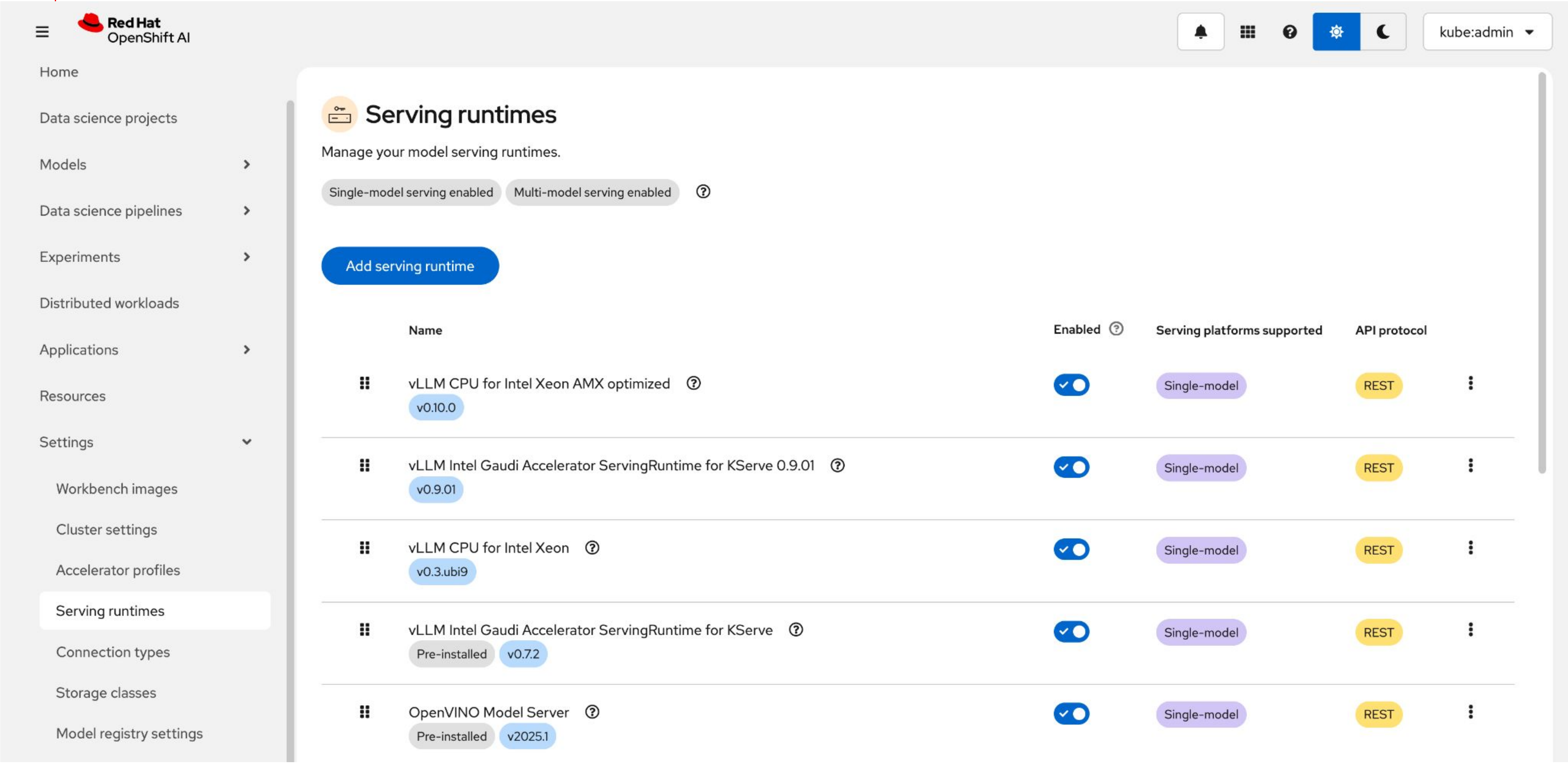
Agentic AI Demo
– Admin Video Slides



Accelerator profiles

Manage accelerator profile settings for users in your organization

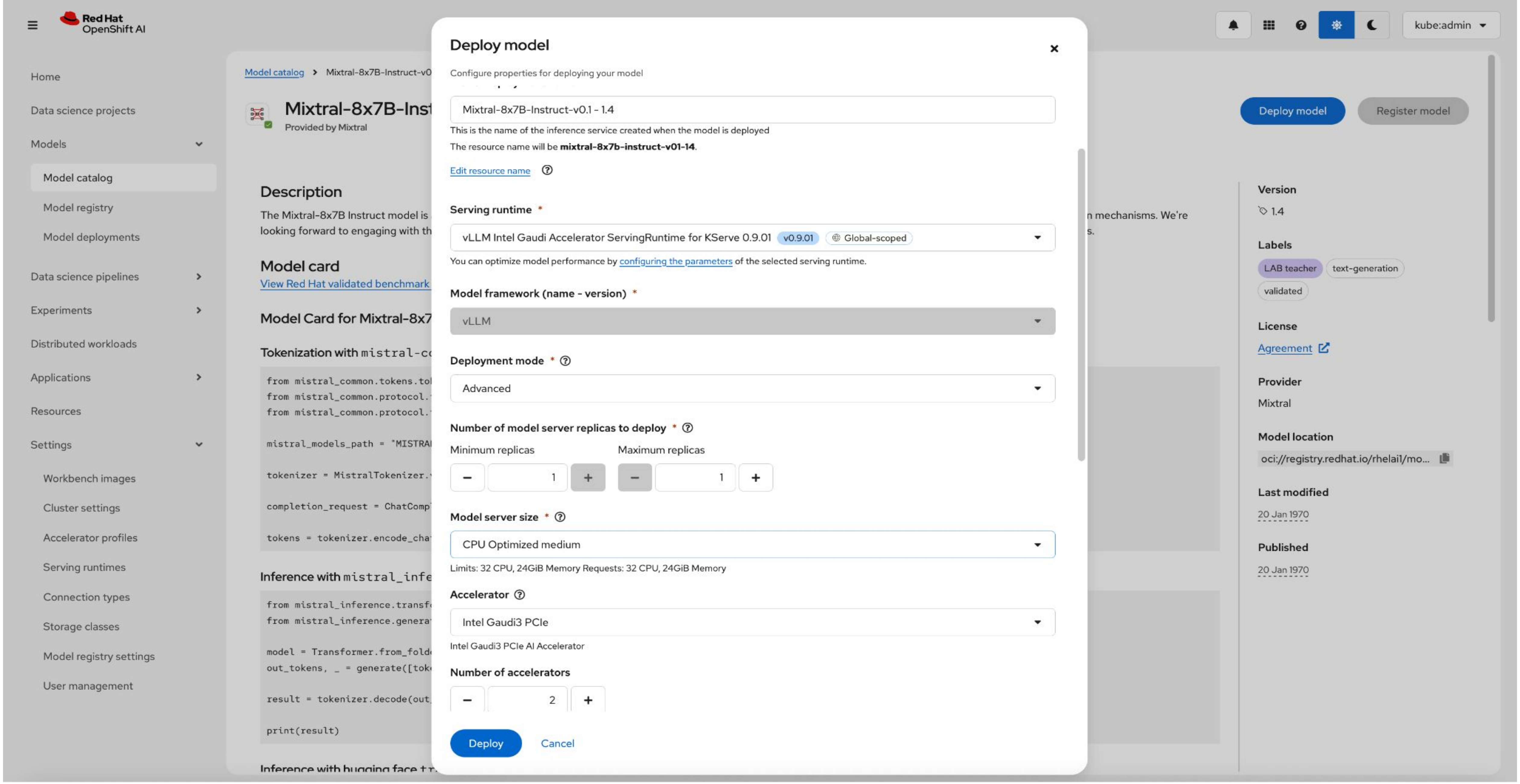
Name	Identifier	Enable	Last modified
Intel Gaudi3 PCIe Intel Gaudi3 PCIe AI Accelerator	habana.ai/gaudi	✓	2 days ago



Serving runtimes

Manage your model serving runtimes.

Name	Enabled	Serving platforms supported	API protocol
vLLM CPU for Intel Xeon AMX optimized v0.10.0	✓	Single-model	REST
vLLM Intel Gaudi Accelerator ServingRuntime for KServe 0.9.01 v0.9.01	✓	Single-model	REST
vLLM CPU for Intel Xeon v0.3.ubi9	✓	Single-model	REST
vLLM Intel Gaudi Accelerator ServingRuntime for KServe Pre-installed v0.7.2	✓	Single-model	REST
OpenVINO Model Server Pre-installed v2025.1	✓	Single-model	REST



Deploy model

Configure properties for deploying your model

Mixtral-8x7B-Instruct-v0.1 - 1.4

This is the name of the inference service created when the model is deployed

The resource name will be **mixtral-8x7b-instruct-v01-14**.

Edit resource name

Serving runtime *
vLLM Intel Gaudi Accelerator ServingRuntime for KServe 0.9.01 v0.9.01 Global-scoped

You can optimize model performance by configuring the parameters of the selected serving runtime.

Model framework (name - version) *
vLLM

Deployment mode *
Advanced

Number of model server replicas to deploy *
Minimum replicas: 1
Maximum replicas: 1

Model server size *
CPU Optimized medium

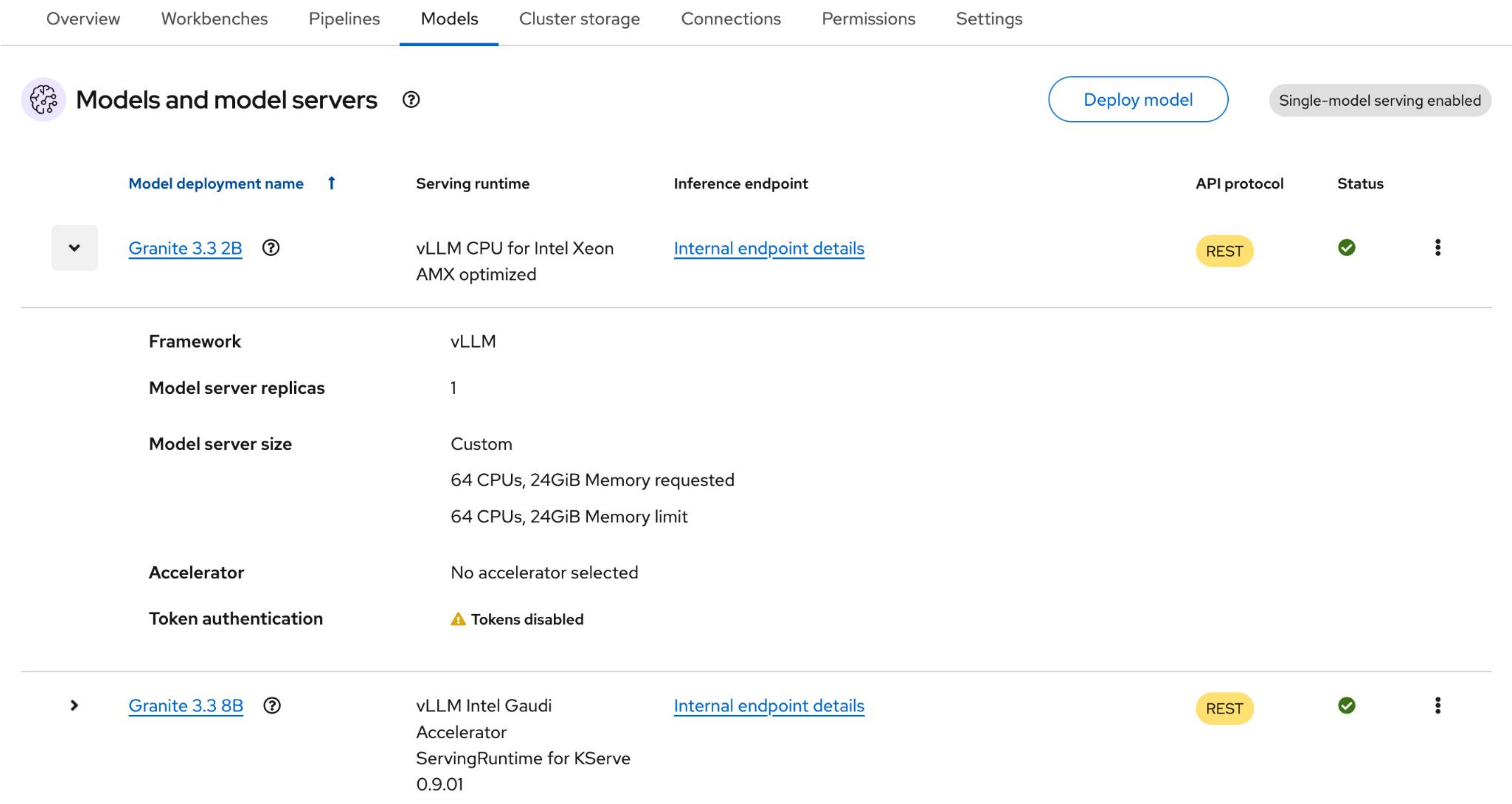
Limits: 32 CPU, 24GiB Memory Requests: 32 CPU, 24GiB Memory

Accelerator
Intel Gaudi3 PCIe

Intel Gaudi3 PCIe AI Accelerator

Number of accelerators
2

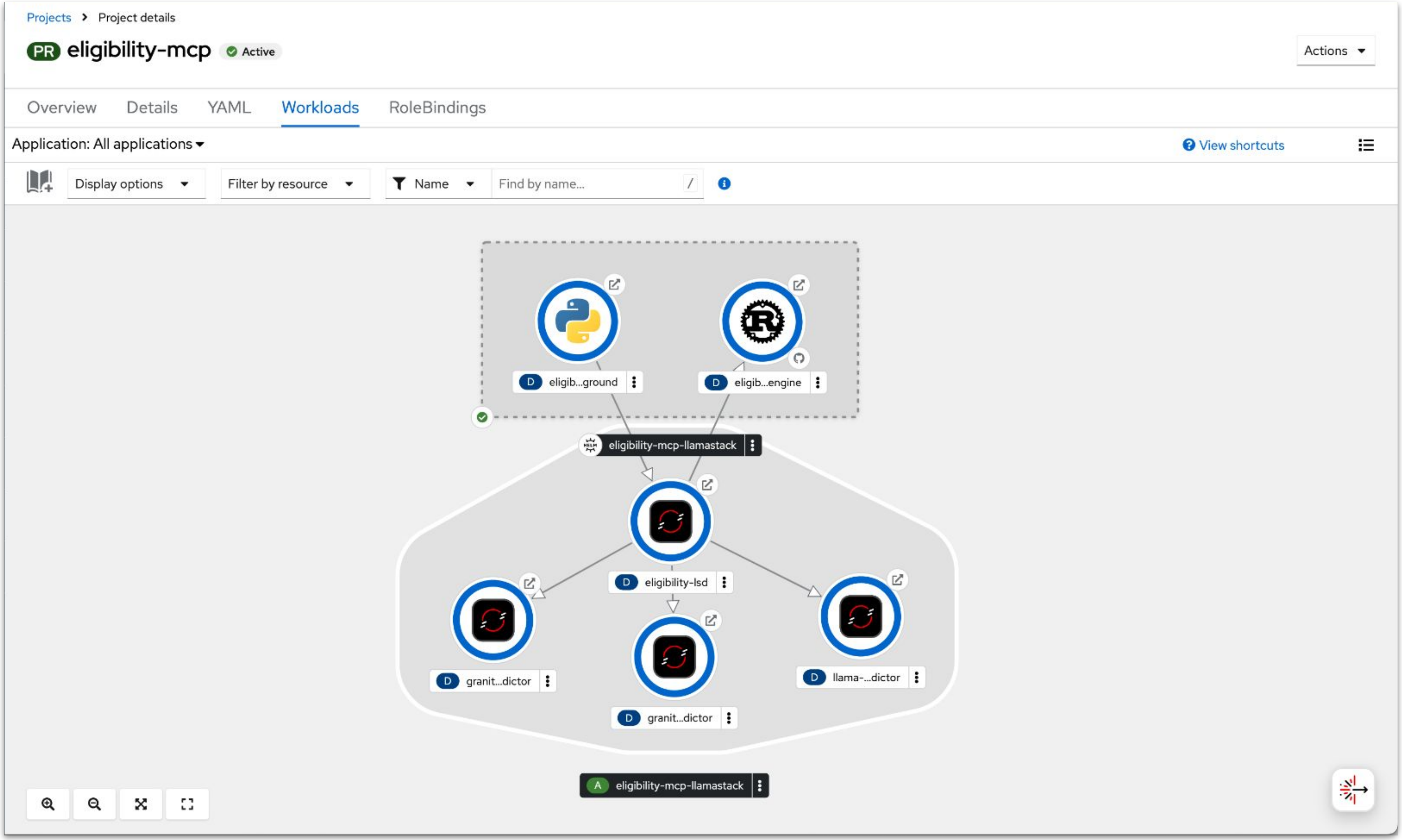
Deploy
Cancel

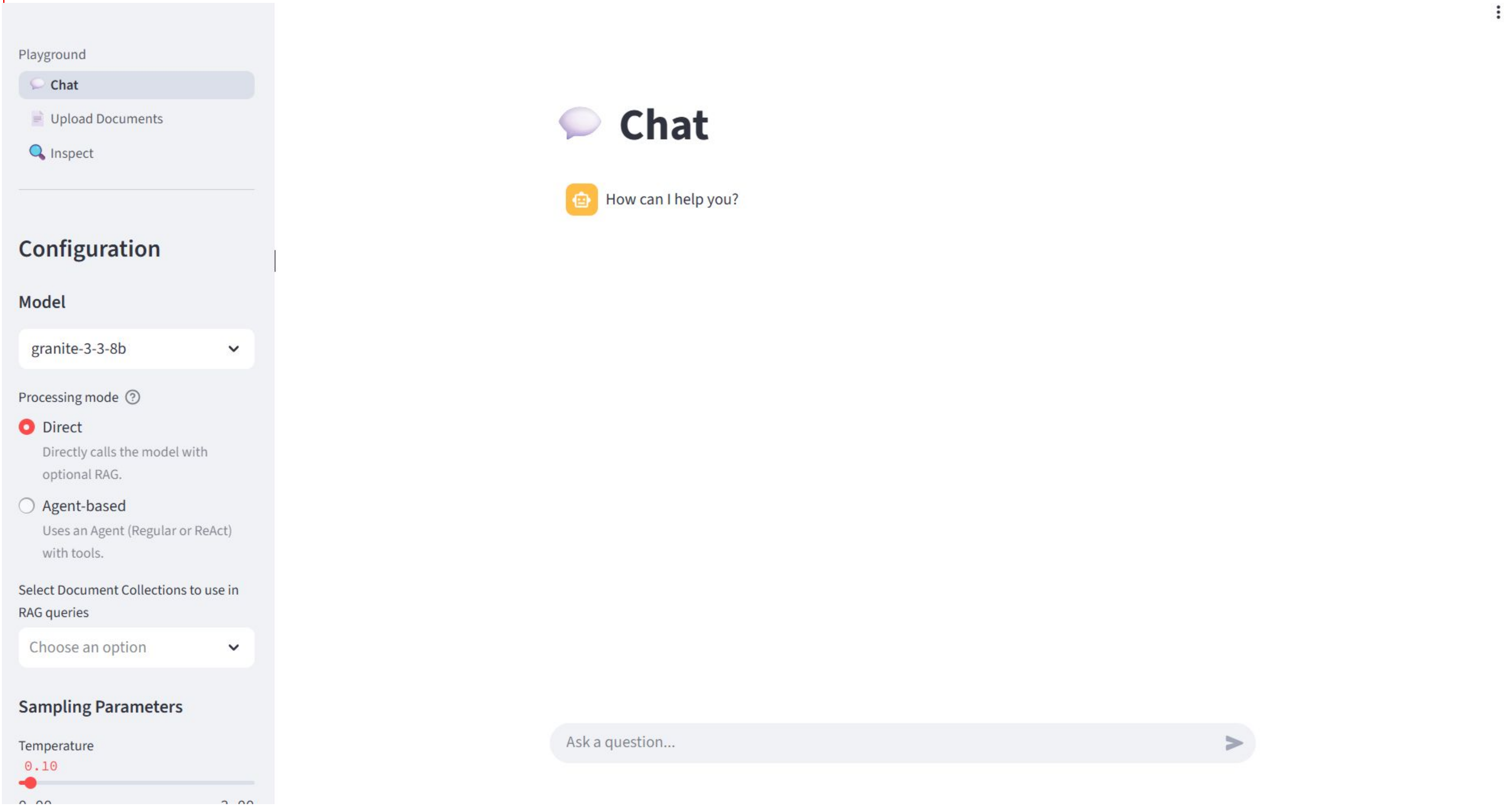


Model deployment name	Serving runtime	Inference endpoint	API protocol	Status
Granite 3.3 2B	vLLM CPU for Intel Xeon AMX optimized	Internal endpoint details	REST	✓
Granite 3.3 8B	vLLM Intel Gaudi Accelerator ServingRuntime for KServe 0.9.01	Internal endpoint details	REST	✓

Framework	vLLM
Model server replicas	1
Model server size	Custom 64 CPUs, 24GiB Memory requested 64 CPUs, 24GiB Memory limit
Accelerator	No accelerator selected
Token authentication	⚠ Tokens disabled

Agentic AI Demo – Agentic AI Application Video Slides


Projects > Project details
PR eligibility-mcp
Active
Actions
Overview
Details
YAML
Workloads
RoleBindings
Application: All applications
View shortcuts
Display options
Filter by resource
Name
Find by name...
/
D eligib...ground
D eligib...engine
eligibility-mcp-llamastack
D eligibility-lsd
D granit...dictor
D granit...dictor
D llama-...dictor
A eligibility-mcp-llamastack



Configuration

Model

granite-3-3-8b

Processing mode

● Direct

Directly calls the model with optional RAG.

○ Agent-based

Uses an Agent (Regular or ReAct) with tools.

Select Document Collections to use in RAG queries

Choose an option

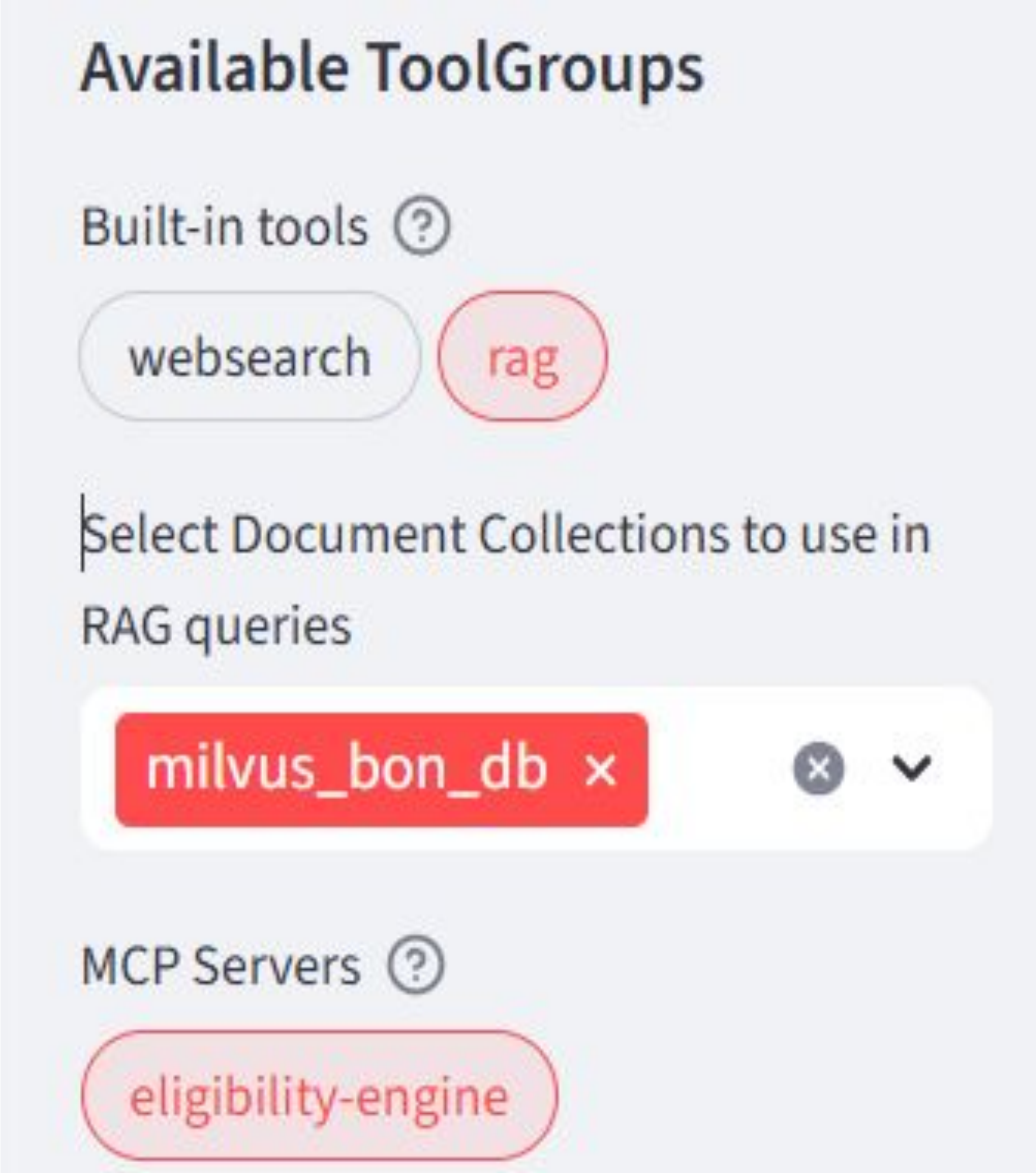
Sampling Parameters

Temperature

0.10

Chat

How can I help you?

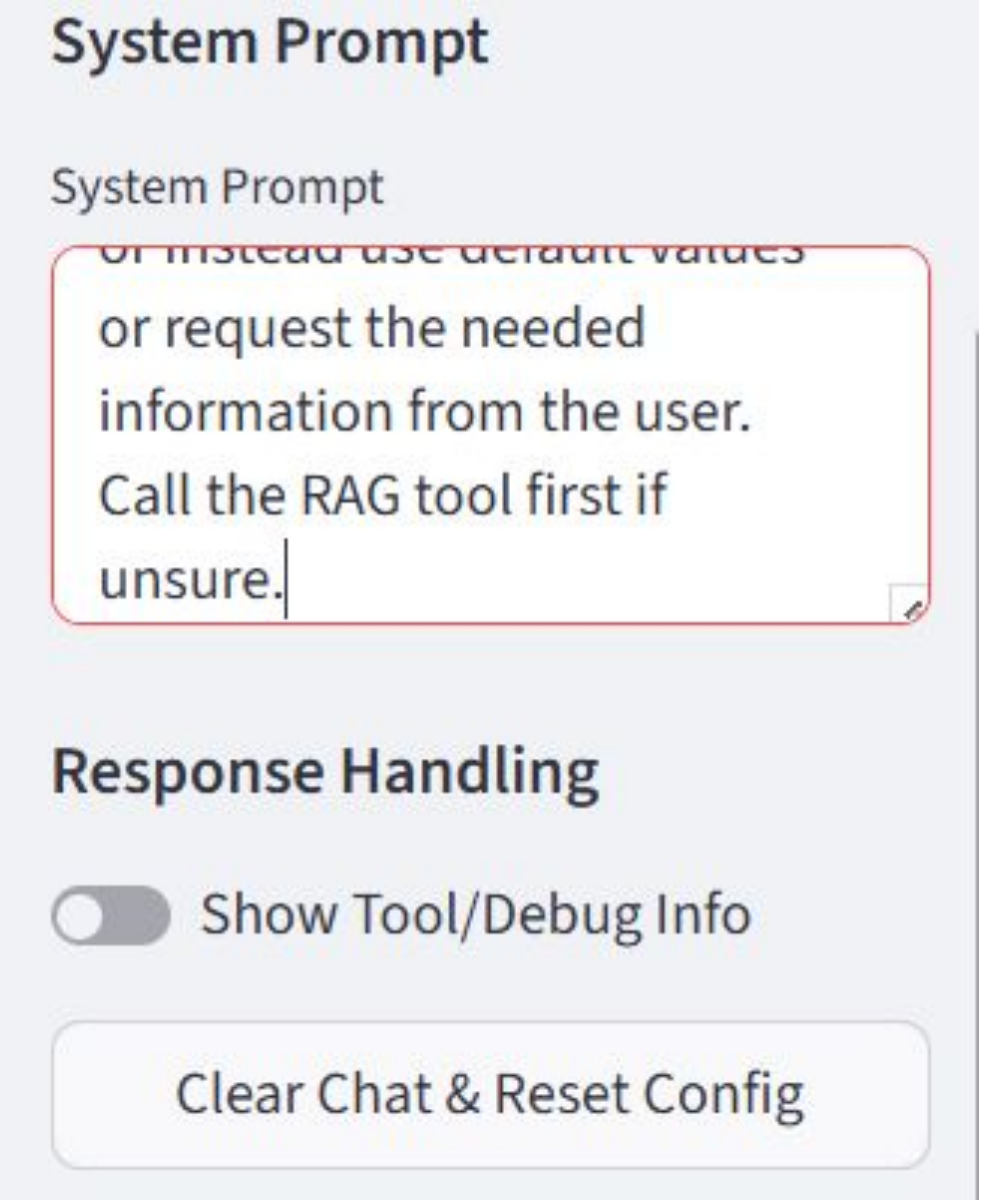


Available ToolGroups

Built-in tools: websearch, **rag**

Select Document Collections to use in RAG queries: milvus_bon_db

MCP Servers: **eligibility-engine**



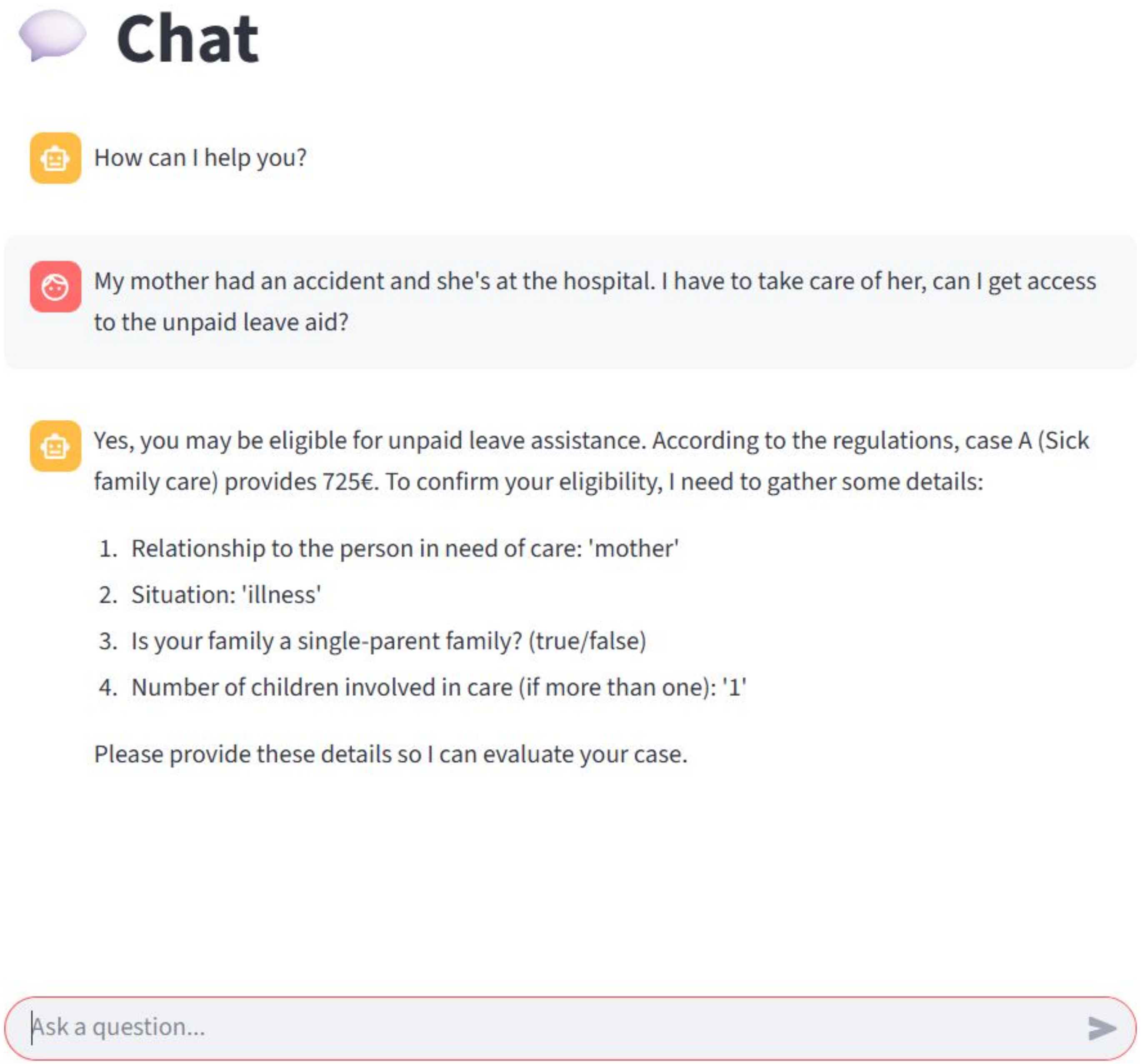
System Prompt

System Prompt:

or request the needed information from the user. Call the RAG tool first if unsure.

Response Handling

○ Show Tool/Debug Info



Chat

How can I help you?

My mother had an accident and she's at the hospital. I have to take care of her, can I get access to the unpaid leave aid?

Yes, you may be eligible for unpaid leave assistance. According to the regulations, case A (Sick family care) provides 725€. To confirm your eligibility, I need to gather some details:

1. Relationship to the person in need of care: 'mother'
2. Situation: 'illness'
3. Is your family a single-parent family? (true/false)
4. Number of children involved in care (if more than one): '1'

Please provide these details so I can evaluate your case.

Q & A

Apply for a **free** Gaudi 3 Proof of Concept in **30 seconds**

Choose your GenAI or Virtualization PoC:

- Building Inference, RAG, AgenticAI, Model-as-a-Service, and other AI Use Cases with Intel Gaudi and Xeon
- Optimize fine tuning with intel Gaudi

Why work with Intel + Red Hat?

- Benefit from access to free highly qualified experts from Red Hat and Intel and free access to the latest hardware to build your AI use case / application.

If selected, a Intel / Red Hat representative will contact you via email.

Red Hat Summit

Connect

Thank you

in linkedin.com/company/red-hat

f facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat