Red Hat Summit

Connect

Agentic AI in Action

Red Hat & Intel Shaping the Future of Enterprise AI

Copenhagen 9 October 2025

Red Hat | intel

Joachim Aertebjerg

Head of ISV Partnerships, EMEA
Intel

Kristoffer Nærland

EMEA Ecosystem Sales Specialist AI
Red Hat

Intel - RH Partnership

Open source software: Intel is committed

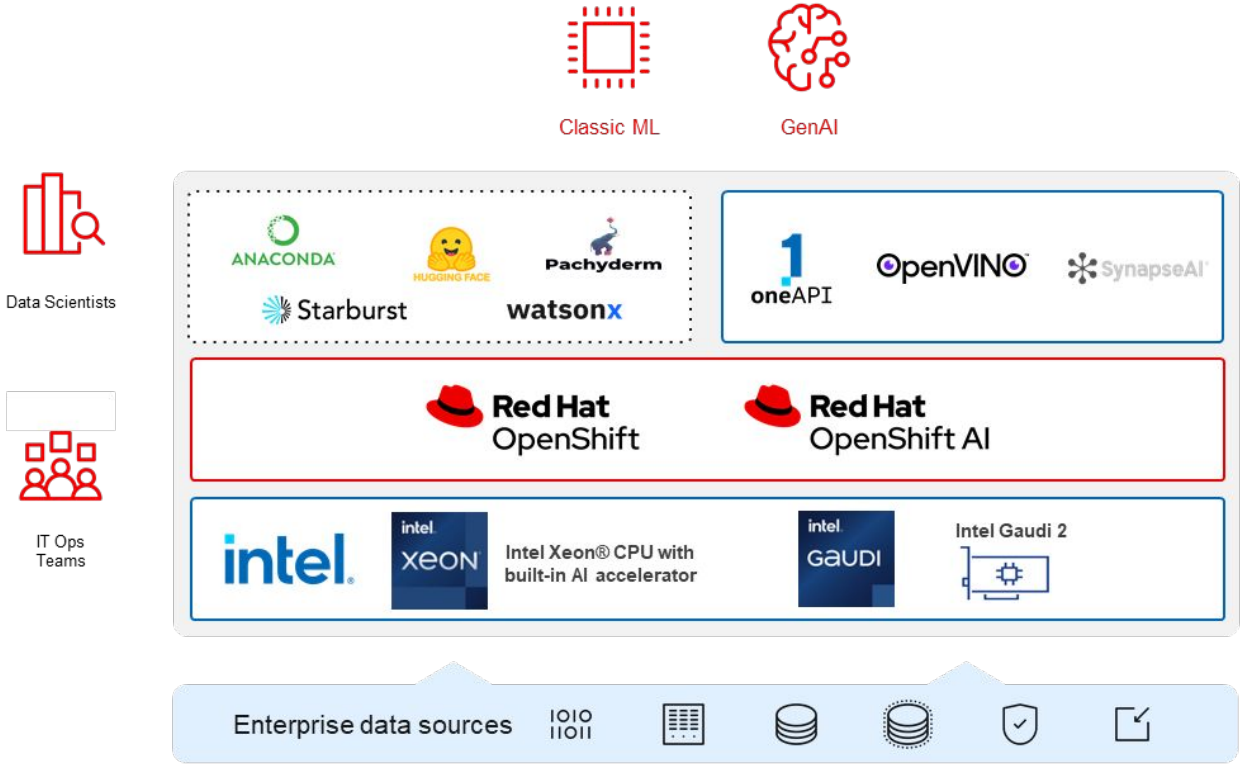
Intel® has a long history with Linux®, actively participating in open source development and collaboration with the Linux community, to ensure hardware is well-supported and delivers optimal performance on Linux-based systems.

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

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



- 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

Red Hat | intel.

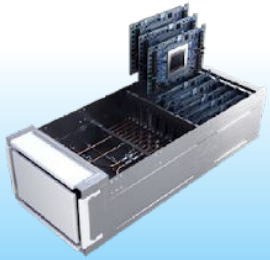
Intel's AI Strategy and Capabilities

Red Hat | intel.

Bringing AI Everywhere

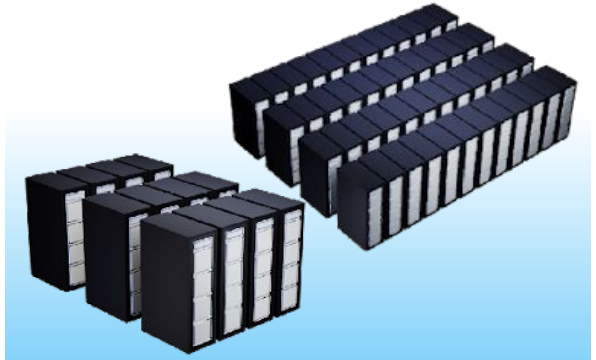
Intel's AI Strategy

AI PC Node
AI Developer Productivity & Light Inference



Node
Fine-tuning, Inference

Cluster
Light Training, Tuning, Peak Inference



Super Cluster
Training, Tuning, Peak Inference

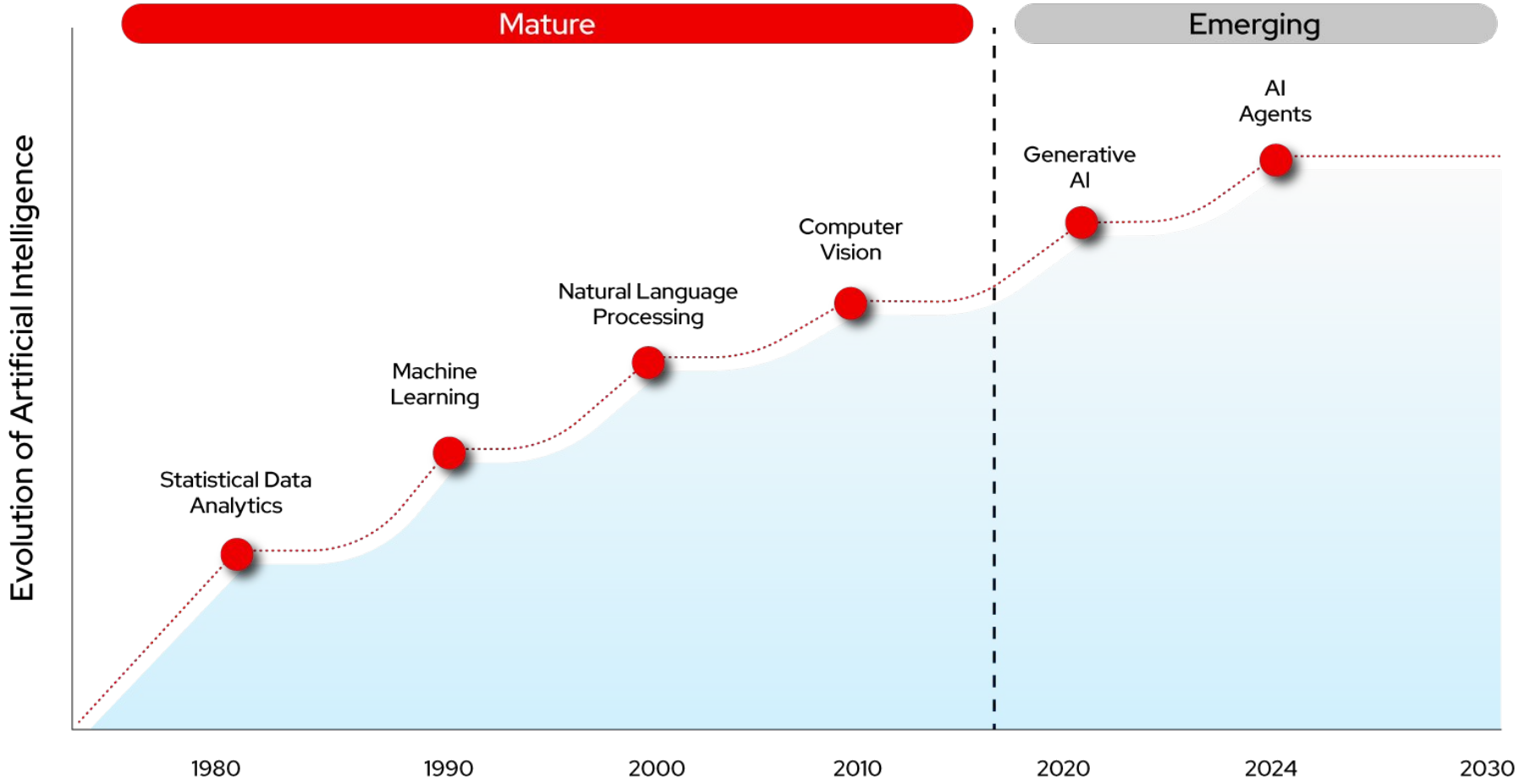
Mega Cluster
Large Scale Training & Inference

AI PC
Broadest AI SW Ecosystem

ENTERPRISE AI & EDGE AI
Open Standard, "Ready to Use"

DATA CENTER AI
AI Open, Scalable Systems & Reference Arch

Evolution of AI Applications in Enterprise Use Cases

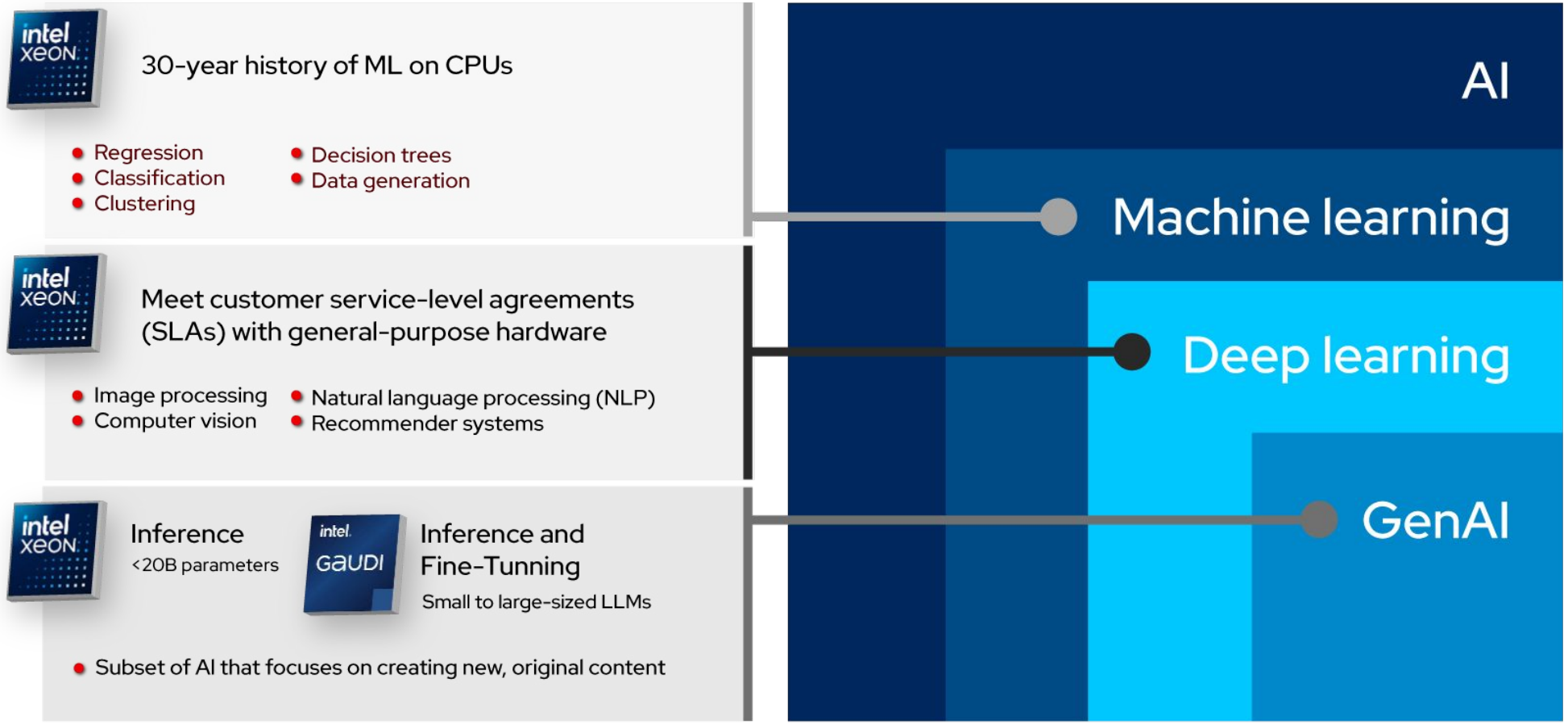


Intel's AI Strategy

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

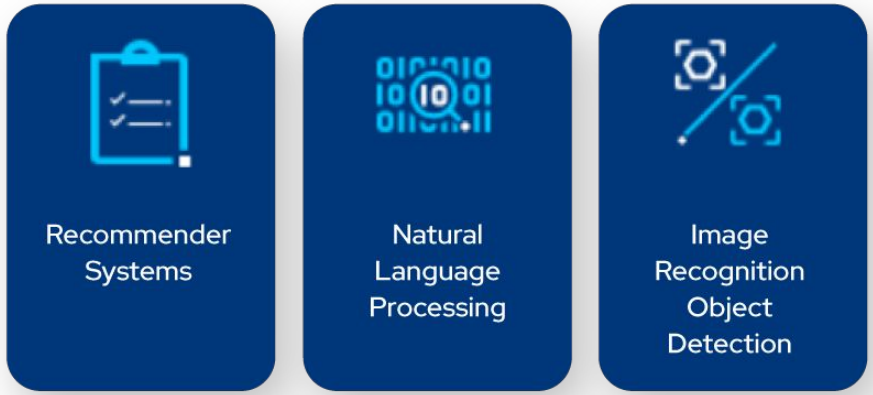
The AI Hierarchy: Mapping ML, DL, and GenAI with Intel

Discover how Intel® processors fuel AI workloads across inference, training, and next-generation GenAI applications



Intel® AMX Accelerates **DEEP LEARNING** Use Cases

Intel® Advanced Matrix Extensions (AMX)	Intel® Advanced Vector Extensions (AVX-512)
BF16, INT8, and FP16 precision	FP32 and FP64 precision

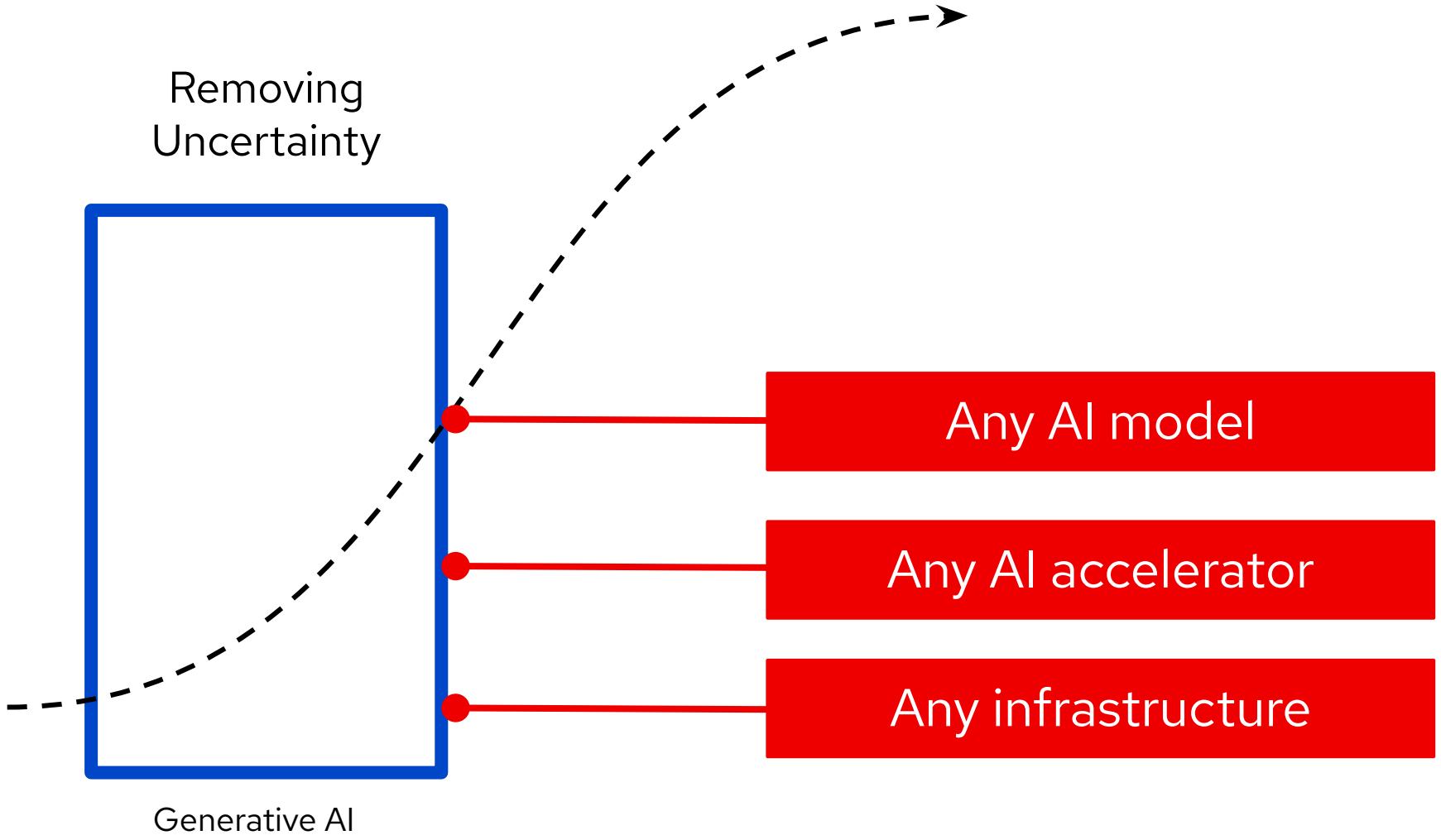


Many DL workloads are "mixed precision" and
5th Gen Xeon can seamlessly transition between AMX and AVX-512 as needed

Red Hat's AI Strategy and Capabilities

Red Hat | intel

Red Hat AI - Enabling AI Success



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**

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

Accelerate Agentic AI delivery and stay at the forefront of innovation

Red Hat AI

Red Hat AI
Inference Server

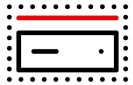
Red Hat
Enterprise Linux AI

Red Hat
OpenShift AI

Trusted, Consistent and Comprehensive foundation

NVIDIA AMD intel Hardware Acceleration Google aws IBM

Physical



Virtual

Private Cloud

Public Cloud

Edge

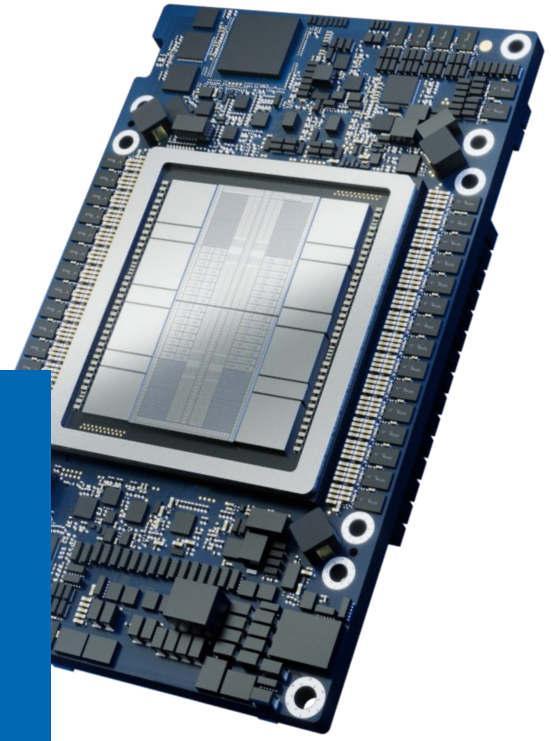
* NVIDIA, AMD, Intel, Google TPU supported in Red Hat AI. AWS Inferentia/Neuron IBM AIU are on our roadmap

Intel Gaudi AI Accelerators

Red Hat | intel.

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.

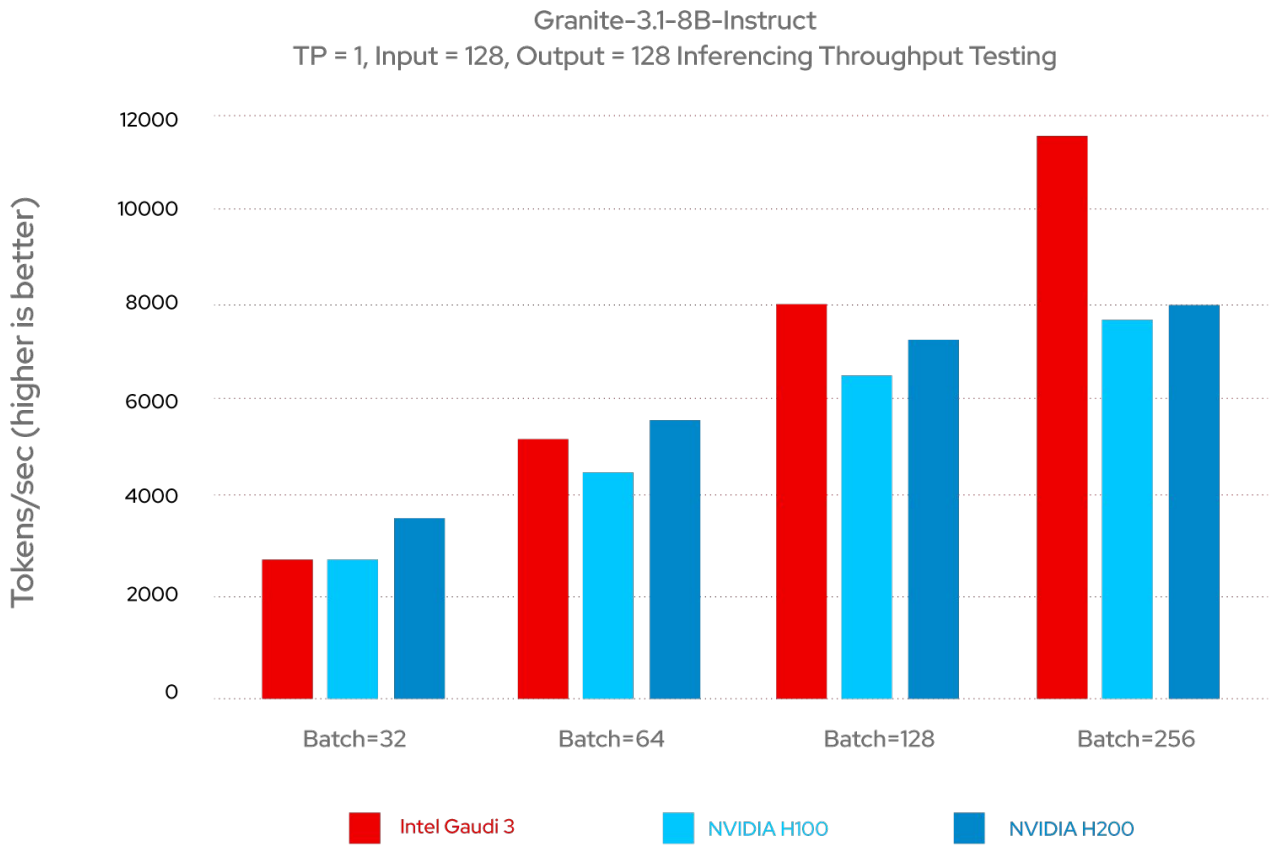
Up to **43% higher**

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.

Red Hat | intel.

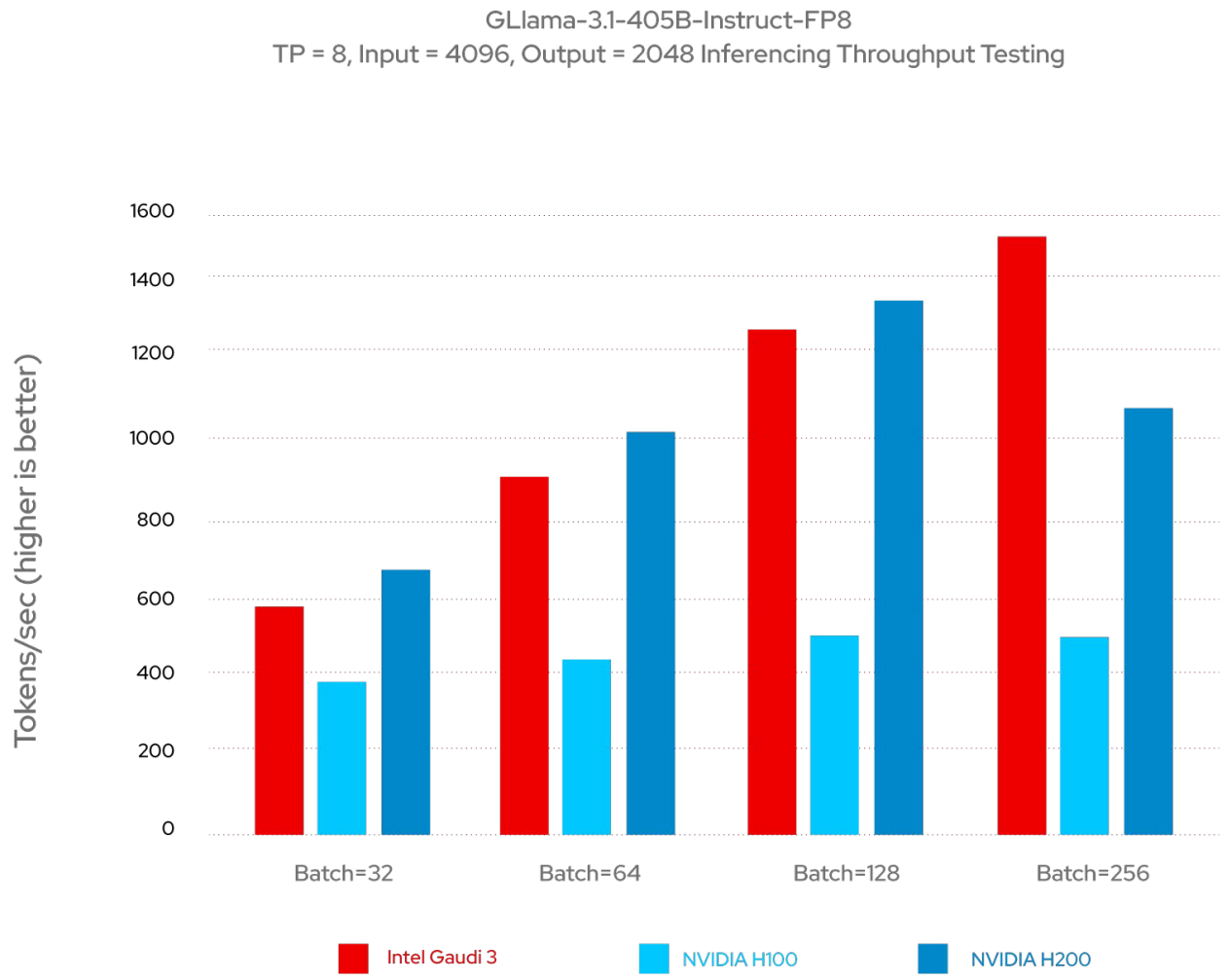
Up to **36% higher**

throughput than NVIDIA H200

Up to **200% higher**

throughput than NVIDIA H100

For Large AI Workloads



*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.

Red Hat | intel.

Intel Xeon Processors

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

Red Hat | intel.

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 Xeon4

Intel Confidential Computing

Red Hat | intel.

Confidential Computing and Post Quantum Crypto for Information & Data Security

Intel® Software Guard Extensions (Intel® SGX)

Smallest Trust Boundary - Confidential data access is restricted to attested application code

Intel® Trust Domain Extensions (Intel® TDX)

Virtual machine isolation from cloud stack, admins, and other tenants

Post Quantum

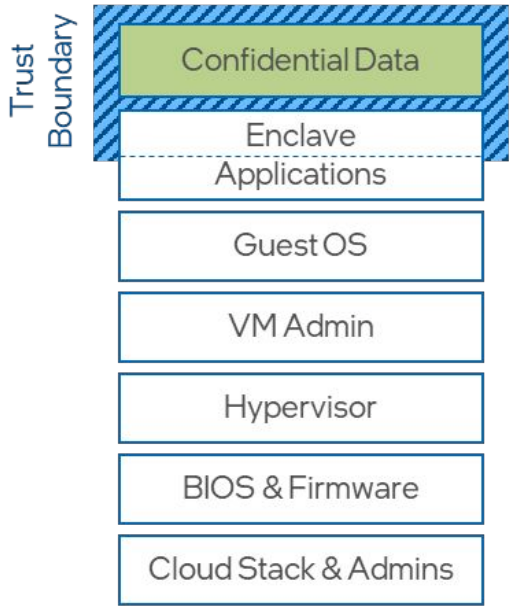
Intel adds Quantum attack protection while providing 1.89 Tb IPsec throughput.

Performant Post-Quantum Cryptography (PQC) leveraging the Intel NetSec Accelerator and Arqit SKA-Platform™ for PQC.

1 - Data Sharing is a Key Digital Transformation Capability (gartner.com) / 2 - Privilege elevation exploits used in over 50% of insider attacks (bleepingcomputer.com) / 3 - Identifying global privacy laws, relevant DPAs | IAPP.

App Isolation

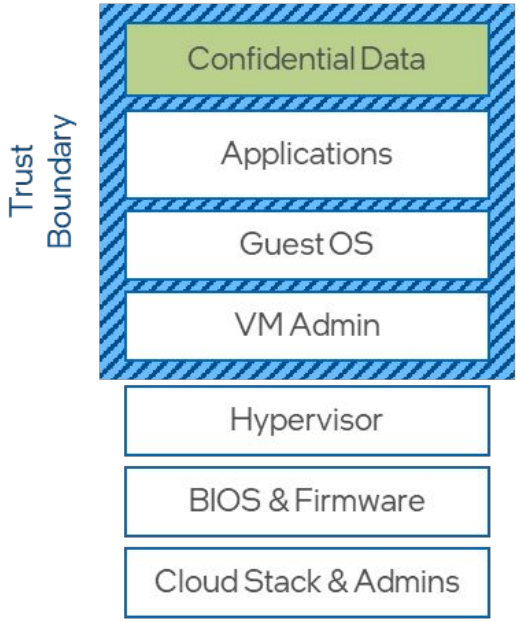
Intel® SGX



Smallest trust boundary for greatest data protection & code integrity

VM Isolation

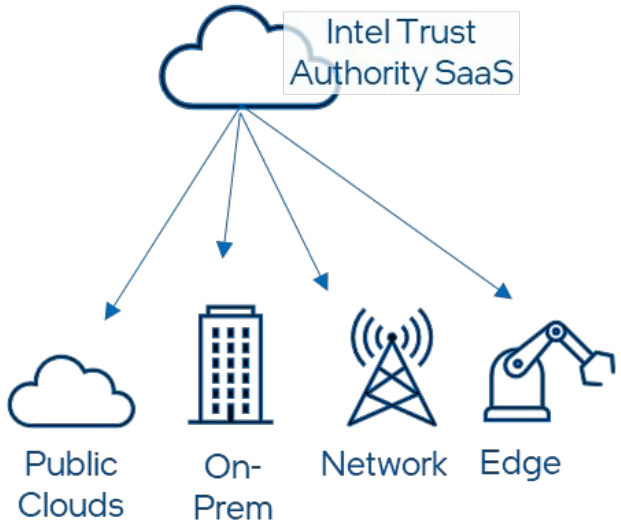
Intel® TDX



Most straightforward path to greater security for legacy apps

Trust Services

Intel® Tiber™ Trust Authority



Uniform, independent attestation of trustworthy environments

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

Red Hat | intel.

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.

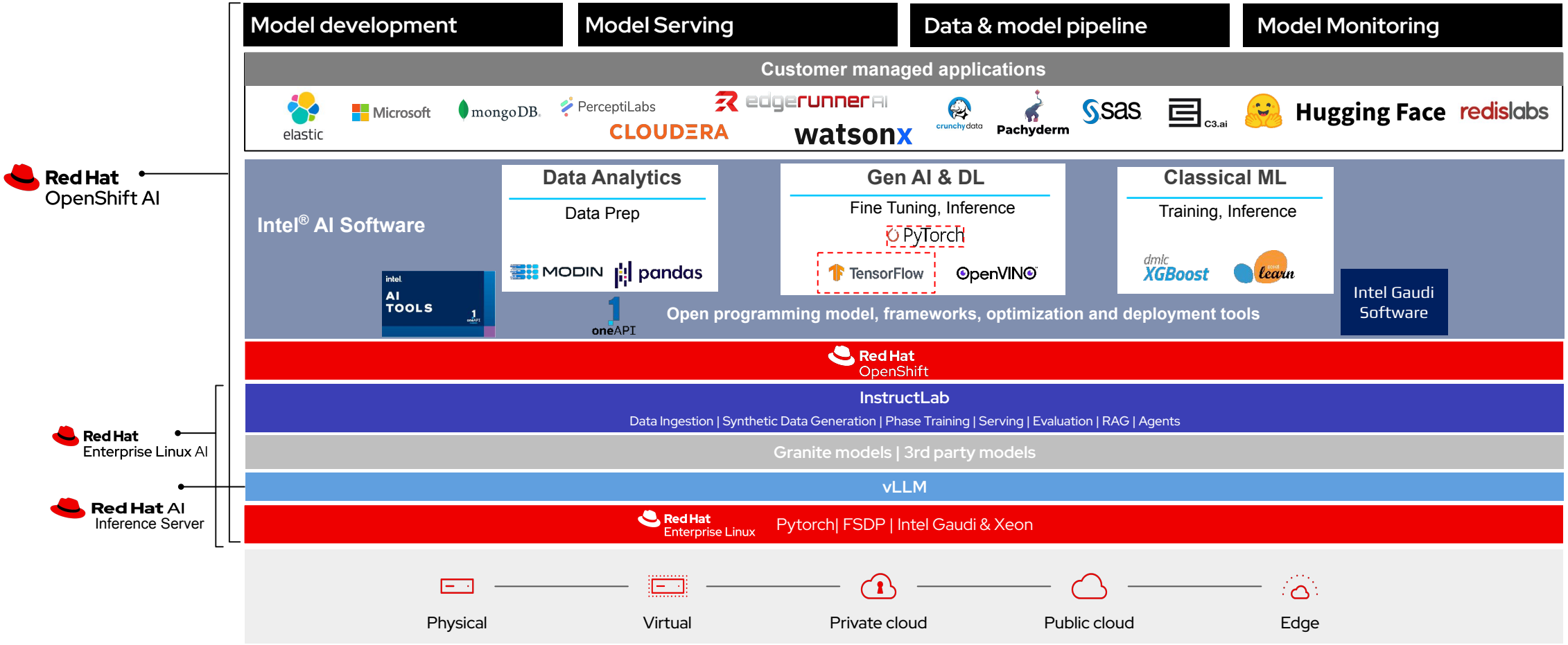
Red Hat | intel.

Intel AI Software

Red Hat | intel.

Red Hat AI with Intel AI platform

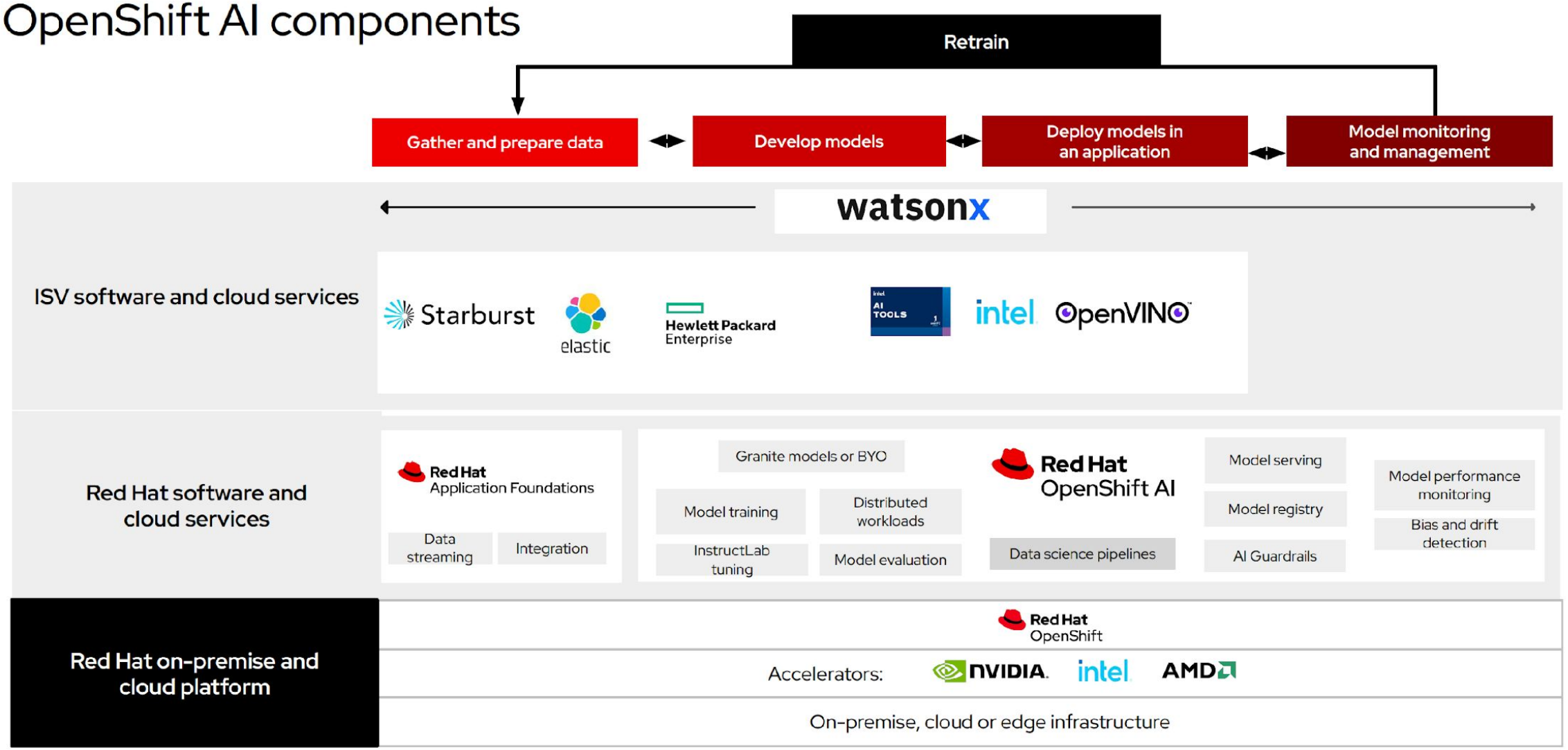
Generative AI and MLOps capabilities for building flexible, trusted AI solutions at scale



Red Hat AI Platform

Red Hat | intel

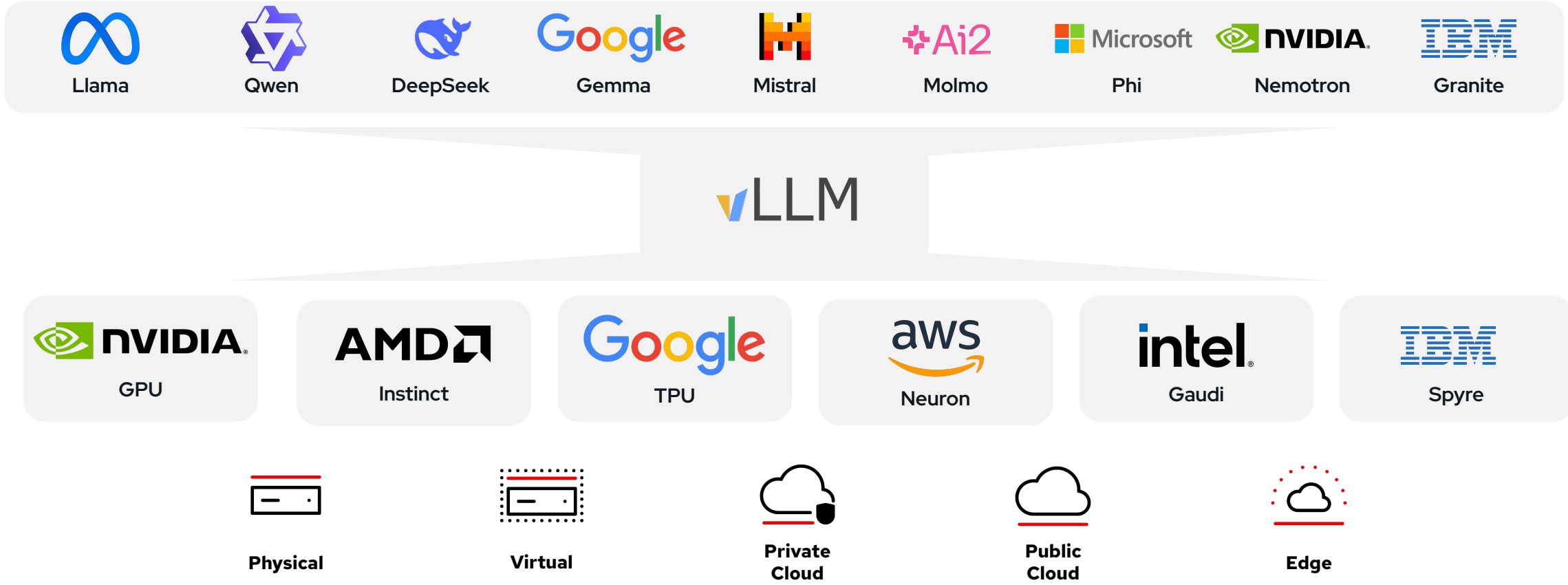
OpenShift AI components



Red Hat

Red Hat AI the inference engine for the hybrid cloud

vLLM supports the key models on the key hardware accelerators



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

Intro to Agentic AI

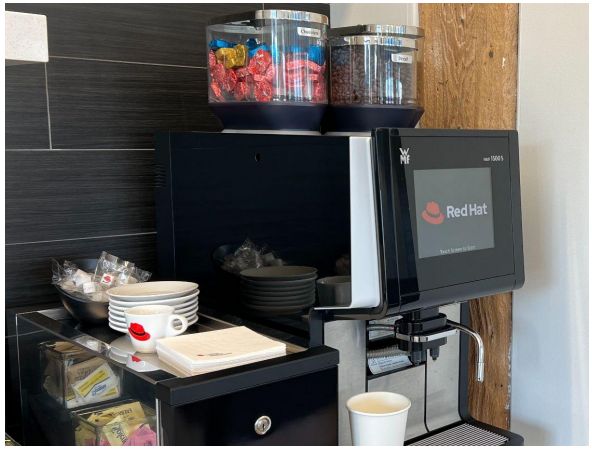
... Imagine you're the CEO and need to provide coffee to your employees

Expense Starbucks for all

- Easy, quick, consistent
- Linear cost & complexity

Sort of like...

- OpenAI, Gemini, etc.



Machine for everyone

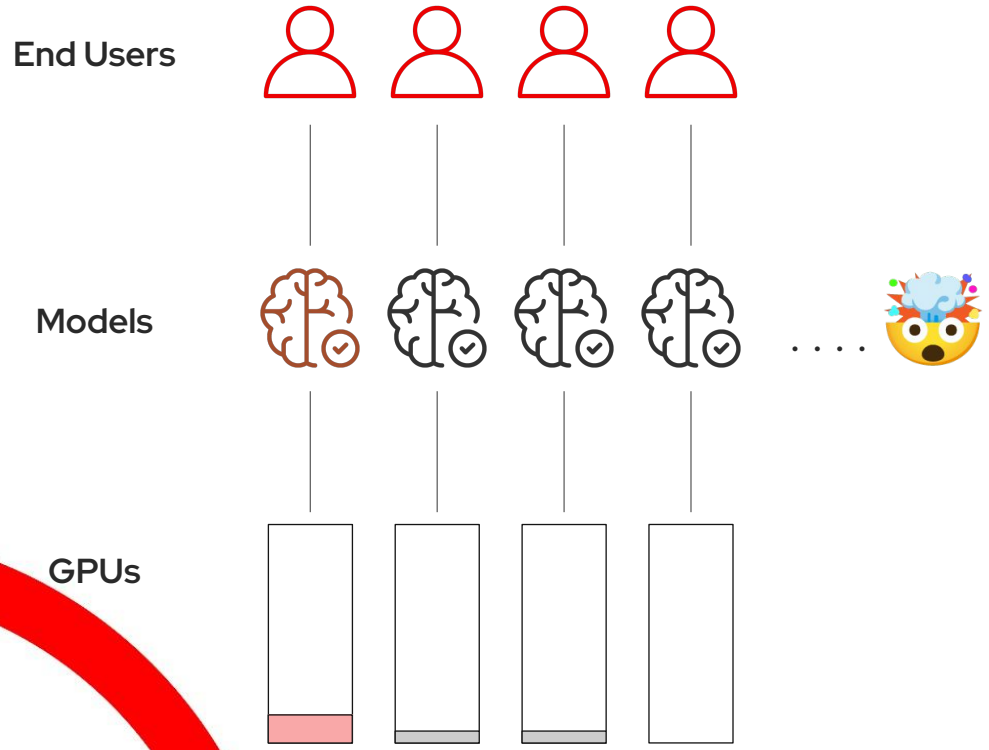
- Easy, quick, consistent
- Economy of scale
- Models-as-a-Service

Full time Professional Barista

- Great for short time periods
- Expenses high but limited
- AI Infrastructure (aka GPUs) as-a-Service

CPUs & **especially GPUs**

Infrastructure as a Service can be costly



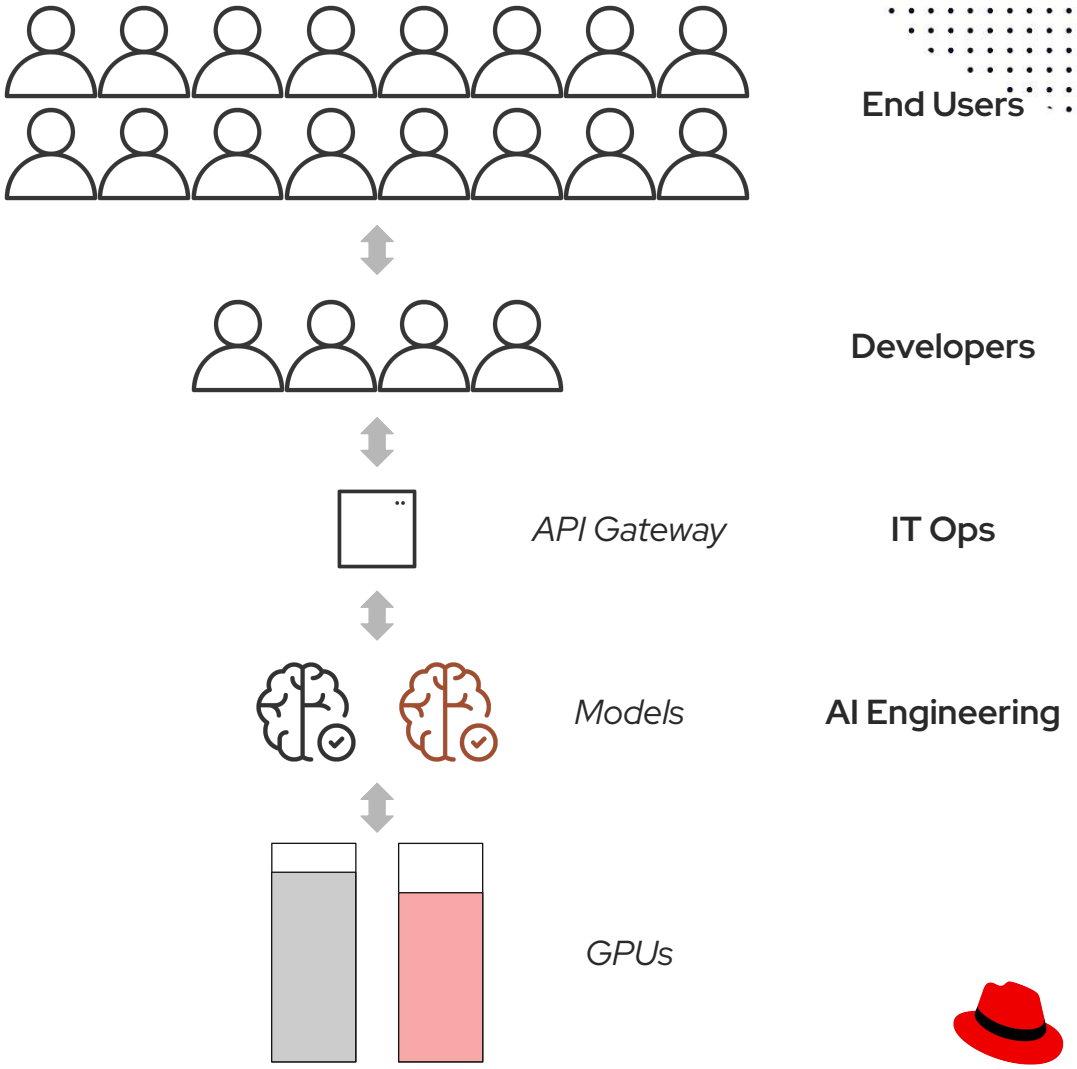
Self-Service is good for plentiful resources & small teams

- Throwing GPUs at the problem is risky
- Few people know how to use them correctly
- Leads to duplication and underutilization
- Leads to high costs
- Most people want an LLM endpoint, not a GPU

Models as a Service (MaaS)

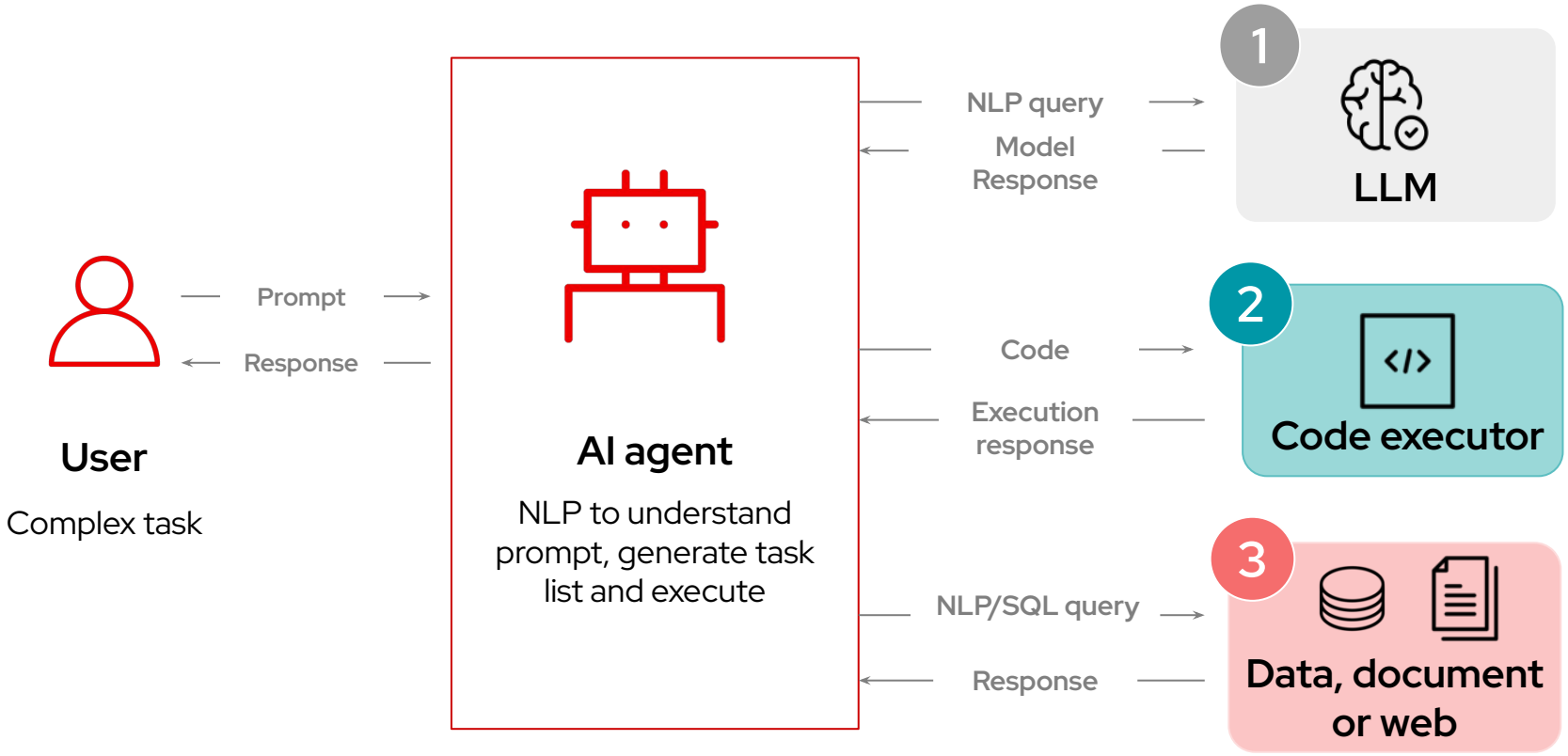
Offering AI **models as *the* service** to a larger audience

- IT serves common models centrally
  - Generative AI focus, applicable to any model
  - Centralized pool of hardware
  - Platform Engineering for AI
  - AI management (versioning, regression testing, etc)
- Models available through API Gateway
- 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



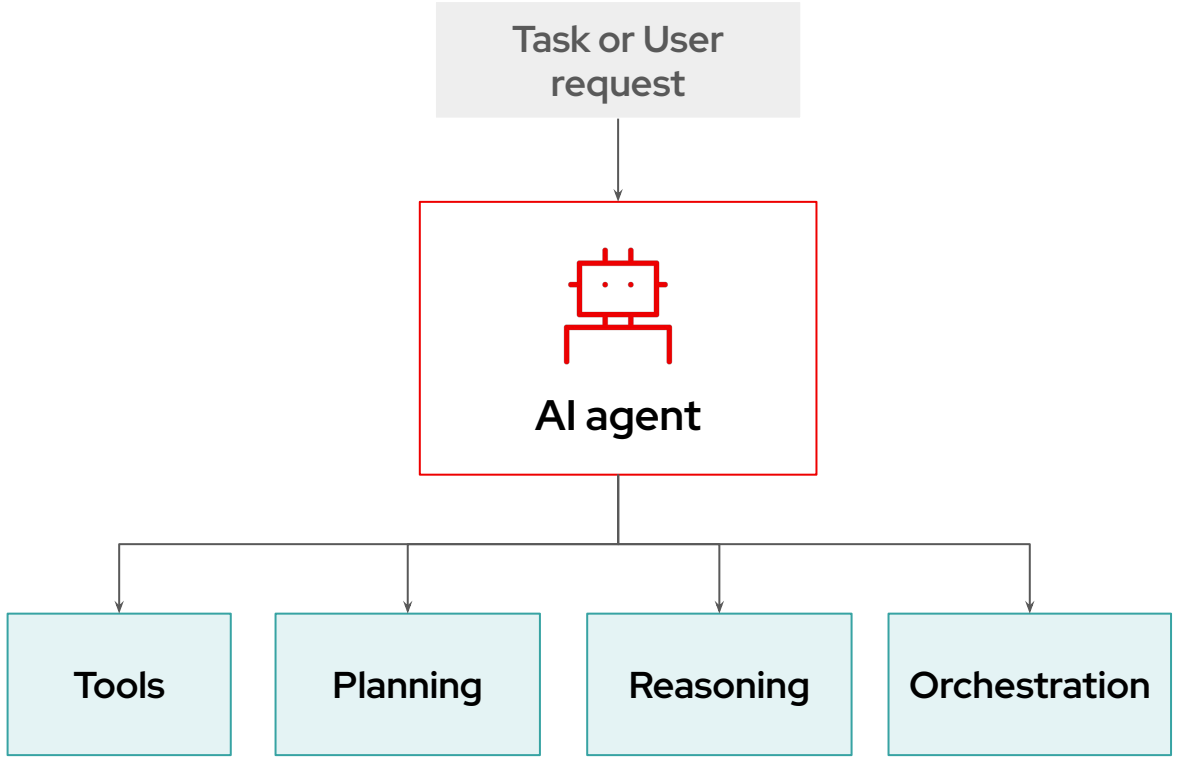
AI agents integrate models, functions & tools

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



Red Hat

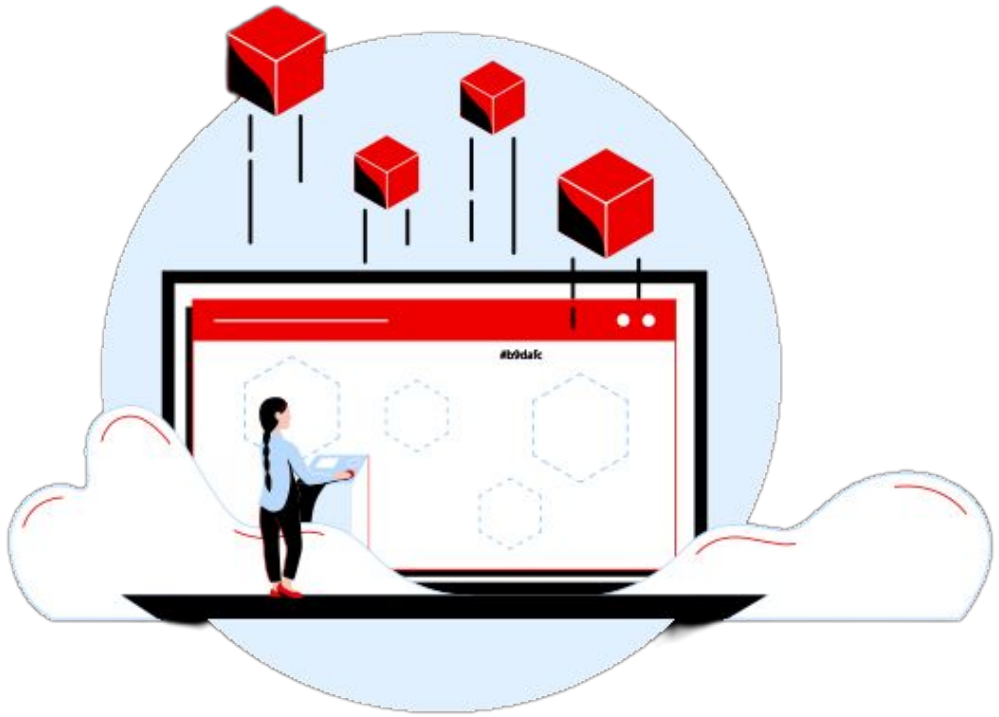
The components of an AI Agent system



- **Tool Utilization:** Leverages external tools to gather data and perform tasks.
- **Planning and Execution:** Develops and executes multistep plans to achieve goals autonomously.
- **Reasoning:** Applies logic and contextual understanding to make informed decisions.
- **Orchestration:** Coordinates actions, tools, and agents to dynamically adjust and complete tasks.
- **Communication protocols:** enables the connections between the components.

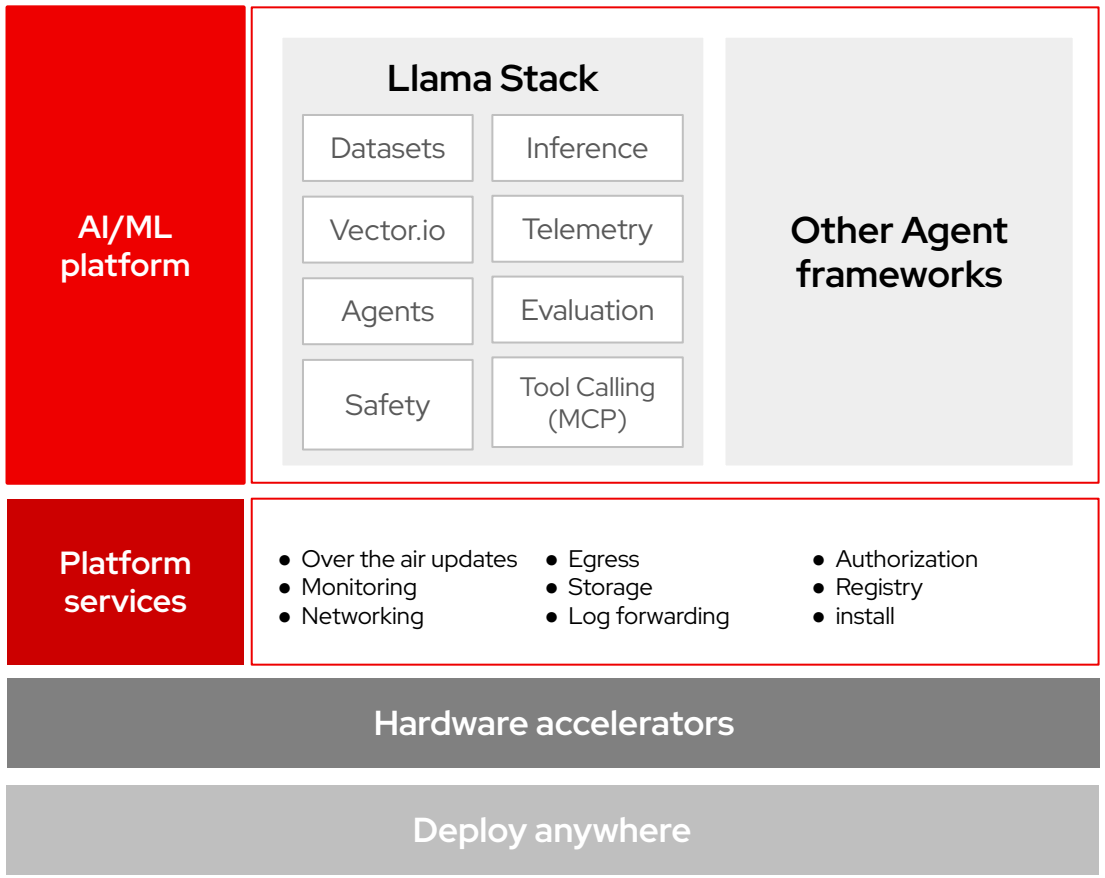
Red Hat AI provides an agile, stable foundation to accelerate the development and deployment of AI agentic workflows.

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



Red Hat

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.

Red Hat

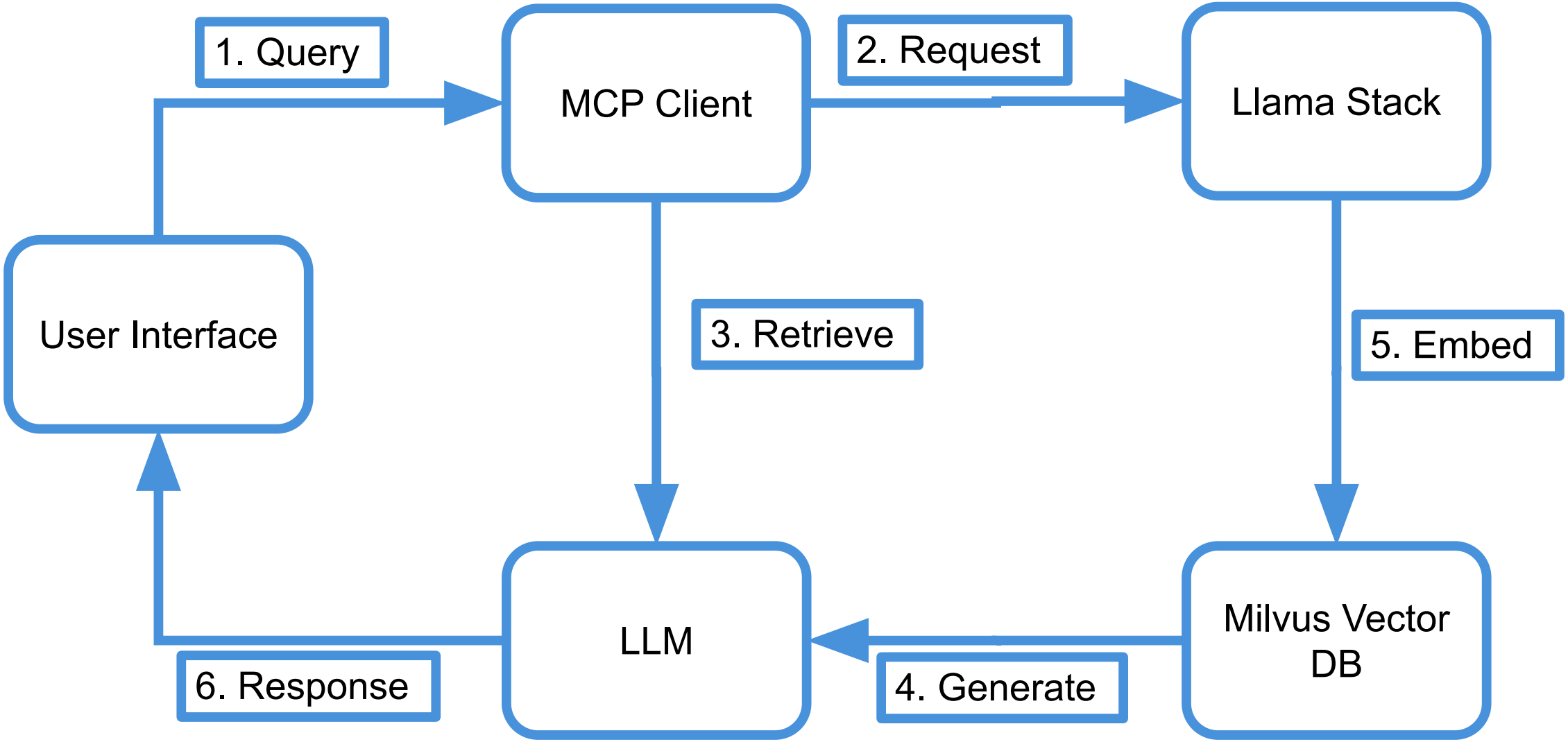
RAG vs RAG + MCP

	RAG	RAG + MCP
Context	Token-limited	Session memory, multi-step
Integration	Custom APIs	Standardized plug-in tools
Agents	Hard to scale	Multi-agent coordination
Deployment	Manual, standalone	Cloud-ready, replicable
Monitoring	Isolated errors	Centralized logging
Resources	Ad hoc calls	Batching, caching

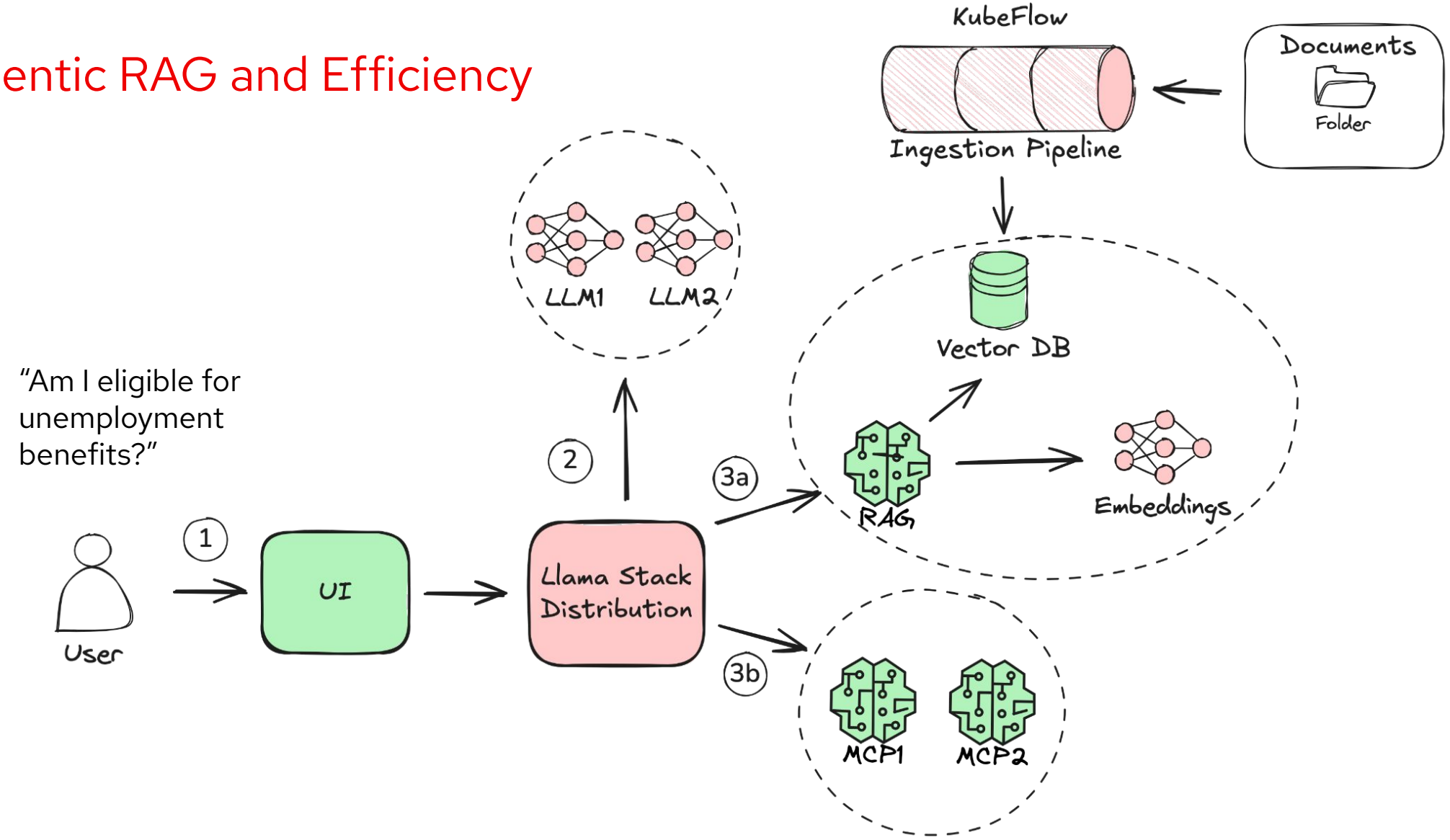
Agentic AI Demo

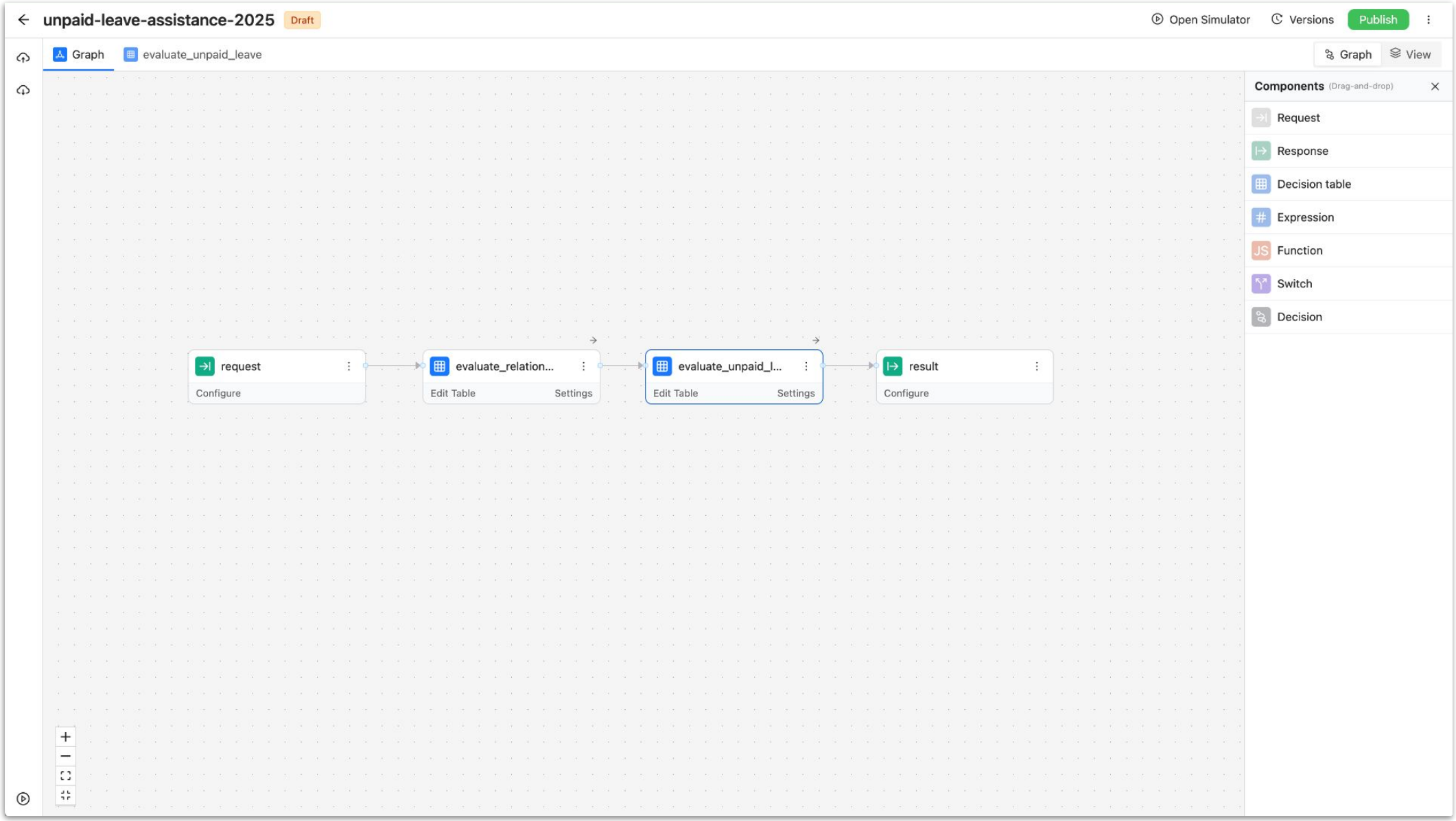
Red Hat | intel.

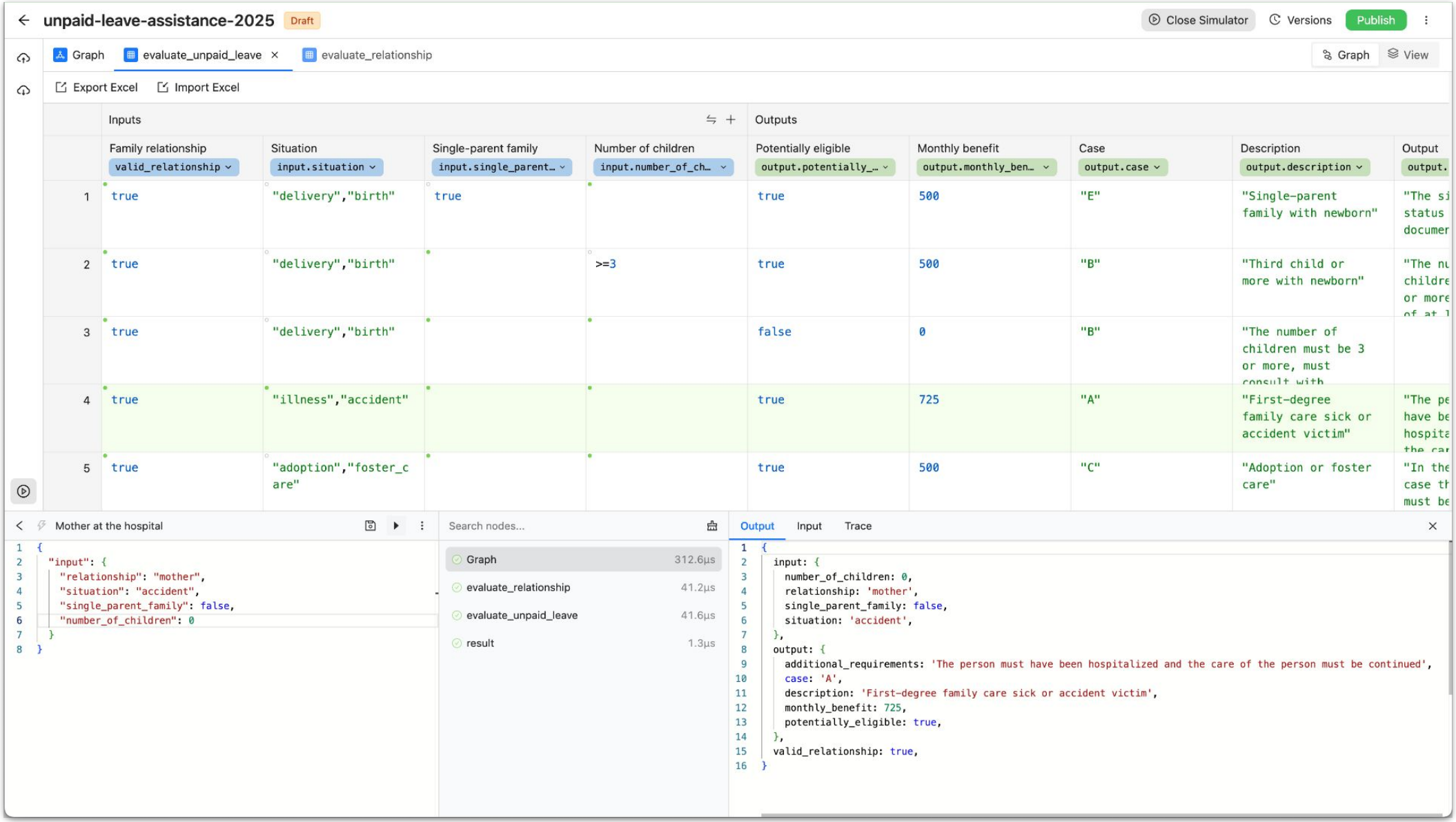
Agentic AI Demo Architecture

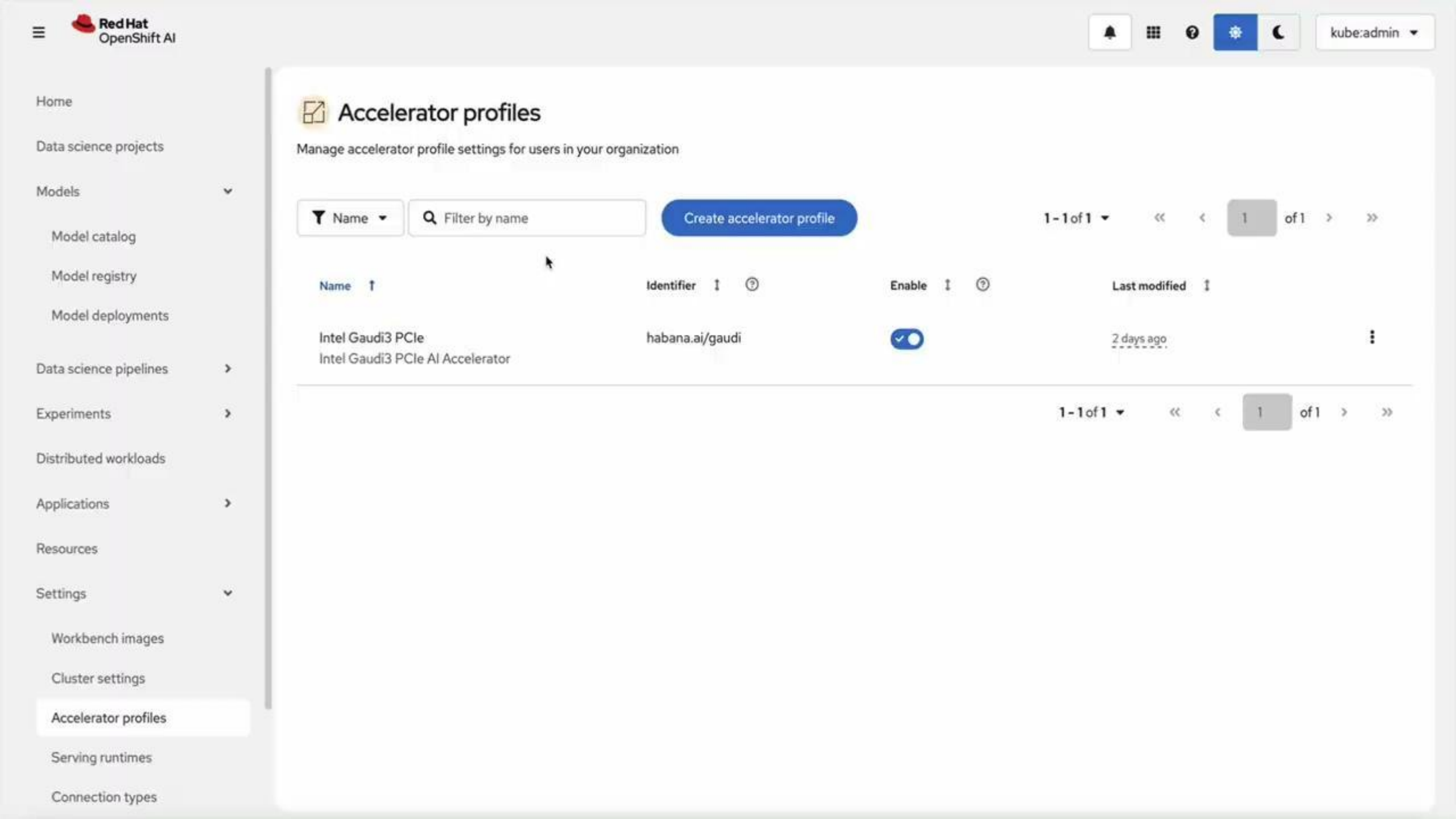


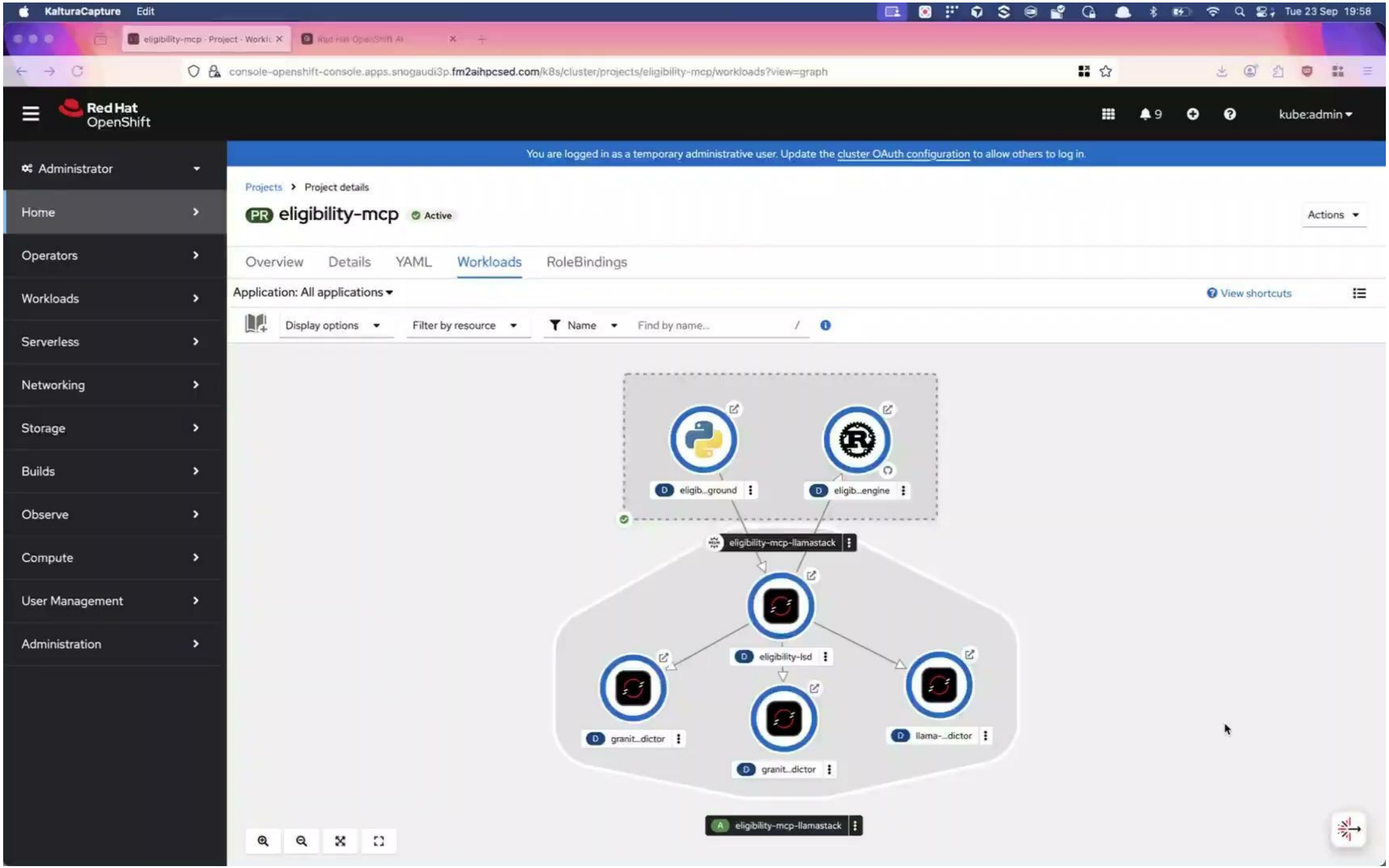
Agentic RAG and Efficiency

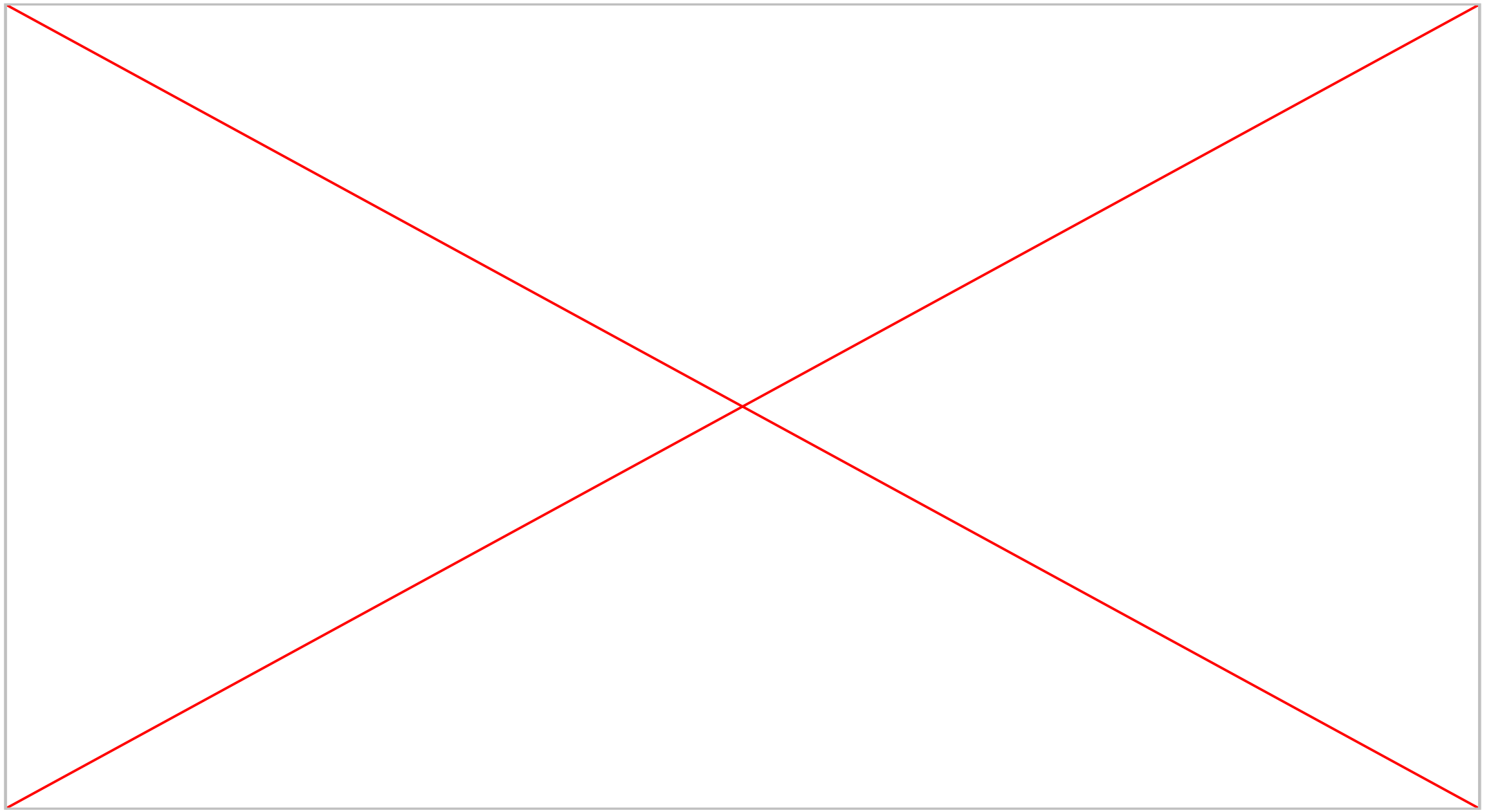







Red Hat OpenShift AI
kube:admin
Home
Data science projects
Models
Model catalog
Model registry
Model deployments
Data science pipelines
Experiments
Distributed workloads
Applications
Resources
Settings
Workbench images
Cluster settings
Accelerator profiles
Serving runtimes
Connection types
Accelerator profiles
Manage accelerator profile settings for users in your organization
Name
Filter by name
Create accelerator profile
1 - 1 of 1
1
of 1
Name
Identifier
Enable
Last modified
Intel Gaudi3 PCIe
Intel Gaudi3 PCIe AI Accelerator
habana.ai/gaudi
2 days ago
1 - 1 of 1
1
of 1


KalturaCapture Edit
Tue 23 Sep 19:58
console-openshift-console.apps.snogaudi3p.fm2aihpcsed.com/k8s/cluster/projects/eligibility-mcp/workloads?view=graph
Red Hat OpenShift
kube:admin
Administrator
Home
Operators
Workloads
Serverless
Networking
Storage
Builds
Observe
Compute
User Management
Administration
You are logged in as a temporary administrative user. Update the cluster OAuth configuration to allow others to log in.
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...
eligib...ground
eligib...engine
eligibility-mcp-llamastack
eligibility-lsd
granit...dictor
granit...dictor
llama-...dictor
eligibility-mcp-llamastack

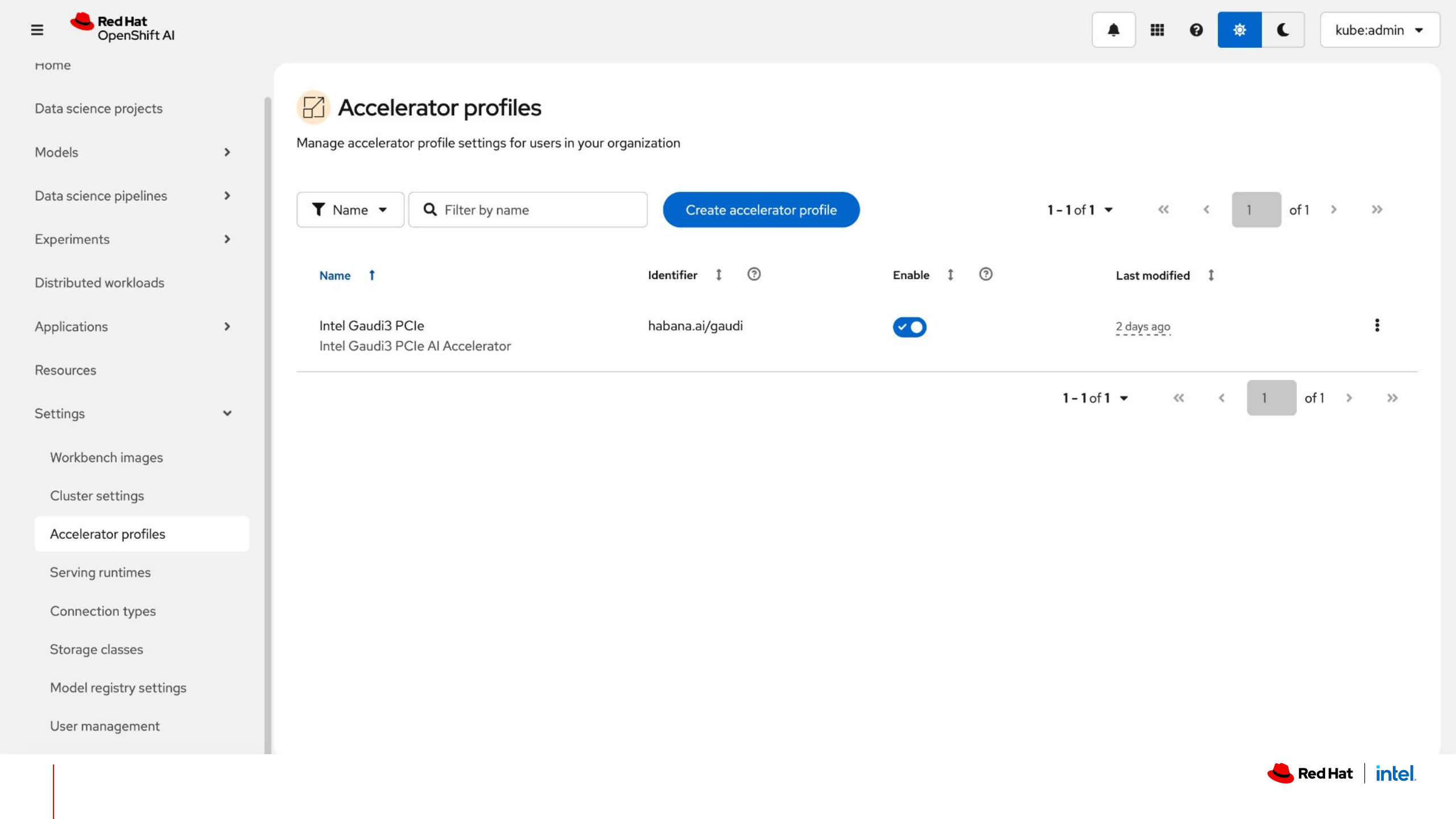


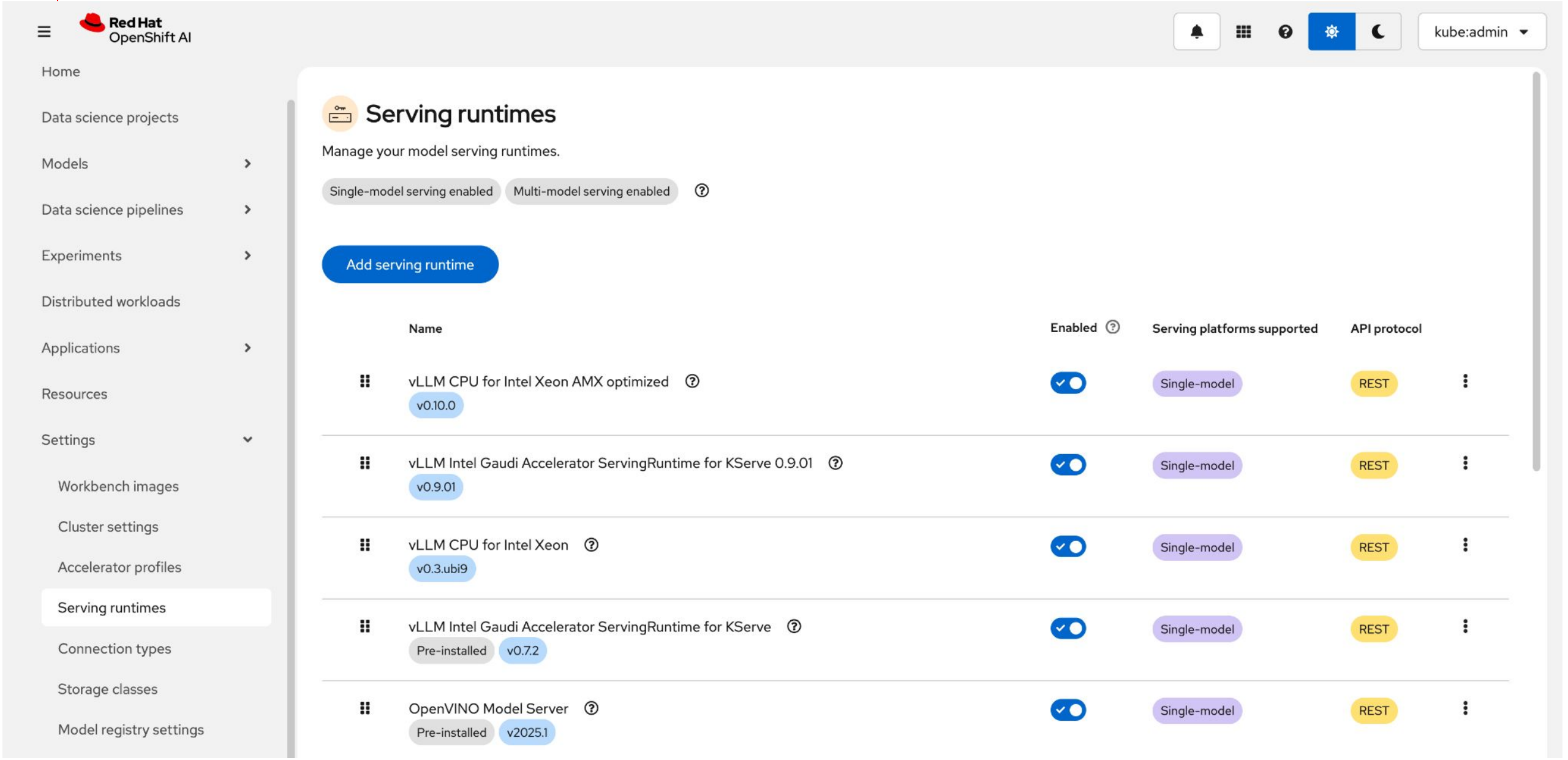
Demo – OpenShift AI Model Serving with Gaudi and Xeon

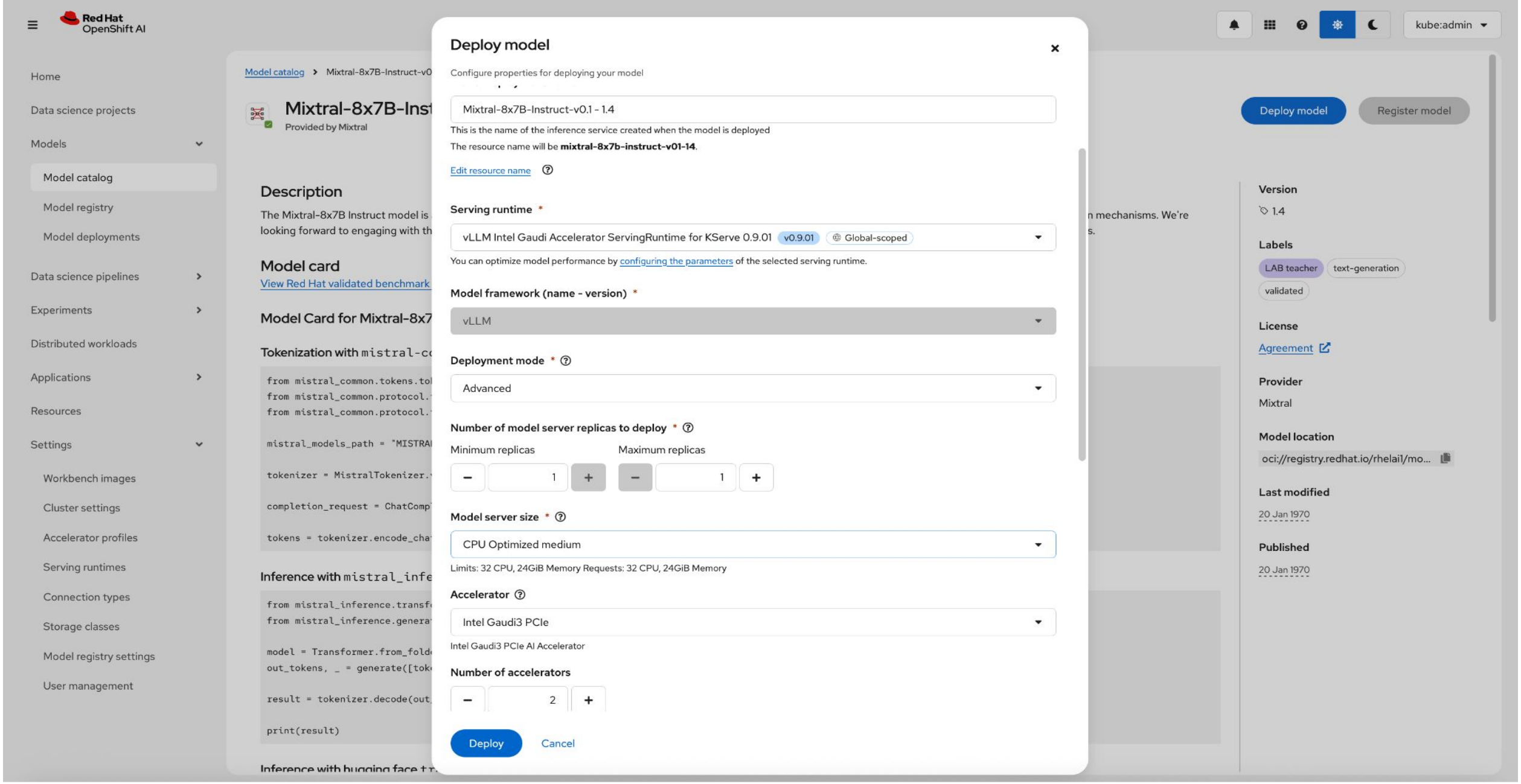
Demo Screenshots for Distributing the Deck (Unskip screenshot slides when generating PDF)

Agentic AI Demo
- Admin Video Slides

Red Hat | intel.


Red Hat OpenShift AI
kube:admin
Home
Data science projects
Models
Data science pipelines
Experiments
Distributed workloads
Applications
Resources
Settings
Workbench images
Cluster settings
Accelerator profiles
Serving runtimes
Connection types
Storage classes
Model registry settings
User management
Accelerator profiles
Manage accelerator profile settings for users in your organization
Name
Filter by name
Create accelerator profile
1 - 1 of 1
1
of 1
Name
Identifier
Enable
Last modified
Intel Gaudi3 PCIe
Intel Gaudi3 PCIe AI Accelerator
habana.ai/gaudi
2 days ago
1 - 1 of 1
1
of 1
Red Hat
intel.


Red Hat OpenShift AI
kube:admin
Home
Data science projects
Models
Data science pipelines
Experiments
Distributed workloads
Applications
Resources
Settings
Workbench images
Cluster settings
Accelerator profiles
Serving runtimes
Connection types
Storage classes
Model registry settings
Serving runtimes
Manage your model serving runtimes.
Single-model serving enabled
Multi-model serving enabled
Add serving runtime
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


Red Hat OpenShift AI
kube:admin
Home
Data science projects
Models
Model catalog
Model registry
Model deployments
Data science pipelines
Experiments
Distributed workloads
Applications
Resources
Settings
Workbench images
Cluster settings
Accelerator profiles
Serving runtimes
Connection types
Storage classes
Model registry settings
User management
Model catalog > Mixtral-8x7B-Instruct-v0
Mixtral-8x7B-Inst
Provided by Mixtral
Deploy model
Register model
Description
The Mixtral-8x7B Instruct model is
looking forward to engaging with th
Model card
View Red Hat validated benchmark
Model Card for Mixtral-8x7
Tokenization with mistral-c
from mistral_common.tokens.to
from mistral_common.protocol.
from mistral_common.protocol.
mistral_models_path = "MISTRA
tokenizer = MistralTokenizer.
completion_request = ChatComp
tokens = tokenizer.encode_cha
Inference with mistral_infe
from mistral_inference.transf
from mistral_inference.genera
model = Transformer.from_fold
out_tokens, _ = generate([tok
result = tokenizer.decode(out
print(result)
Inference with hugging face t
n mechanisms. We're
s.
Version
1.4
Labels
LAB teacher
text-generation
validated
License
Agreement
Provider
Mixtral
Model location
oci://registry.redhat.io/rhelai1/mo...
Last modified
20 Jan 1970
Published
20 Jan 1970
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

Overview | Workbenches | Pipelines | Models | Cluster storage | Connections | Permissions | Settings

Models and model servers

Deploy model | Single-model serving enabled

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	✓

Framework vLLM

Model server replicas 1

Model server size Custom
64 CPUs, 24GiB Memory requested
64 CPUs, 24GiB Memory limit

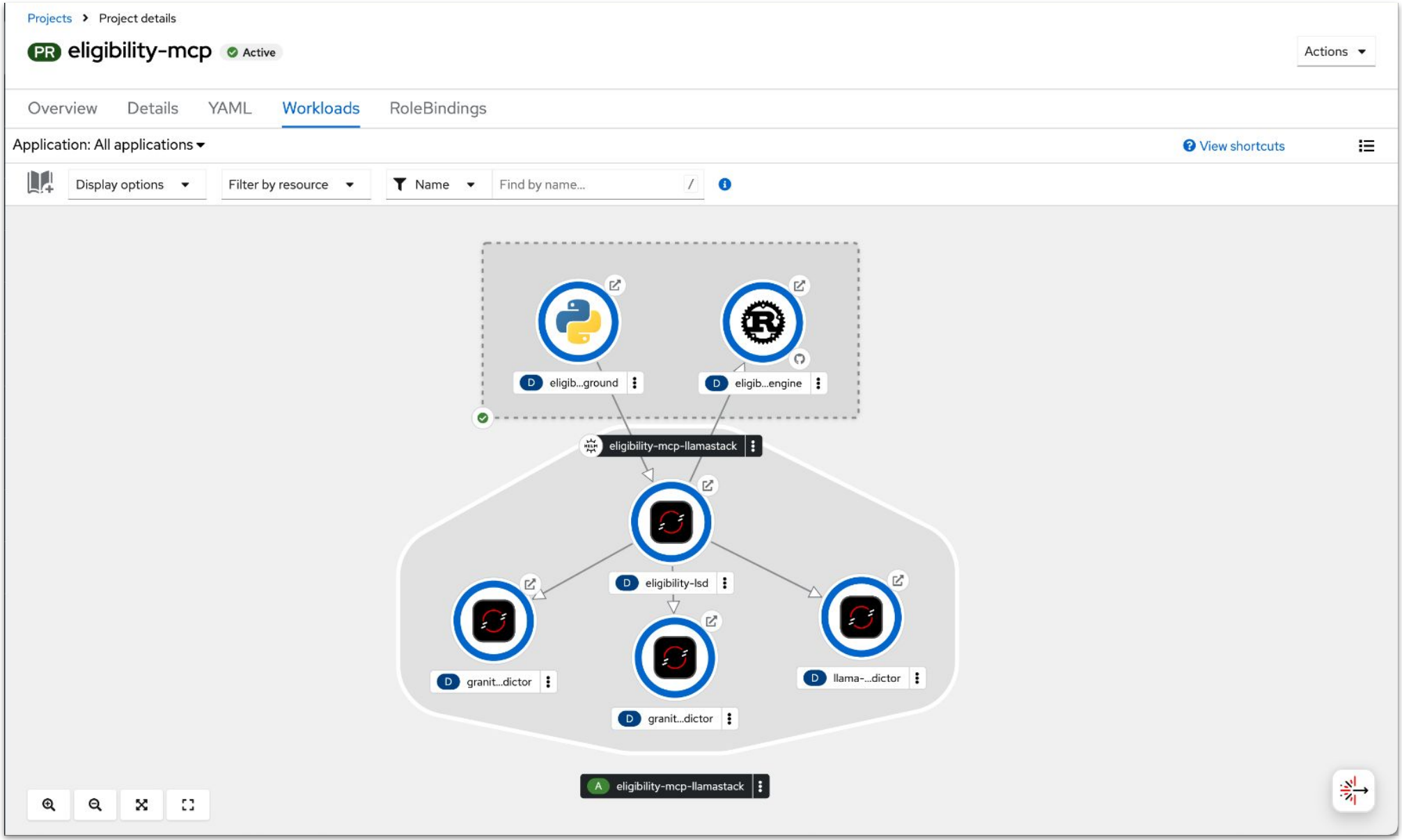
Accelerator No accelerator selected

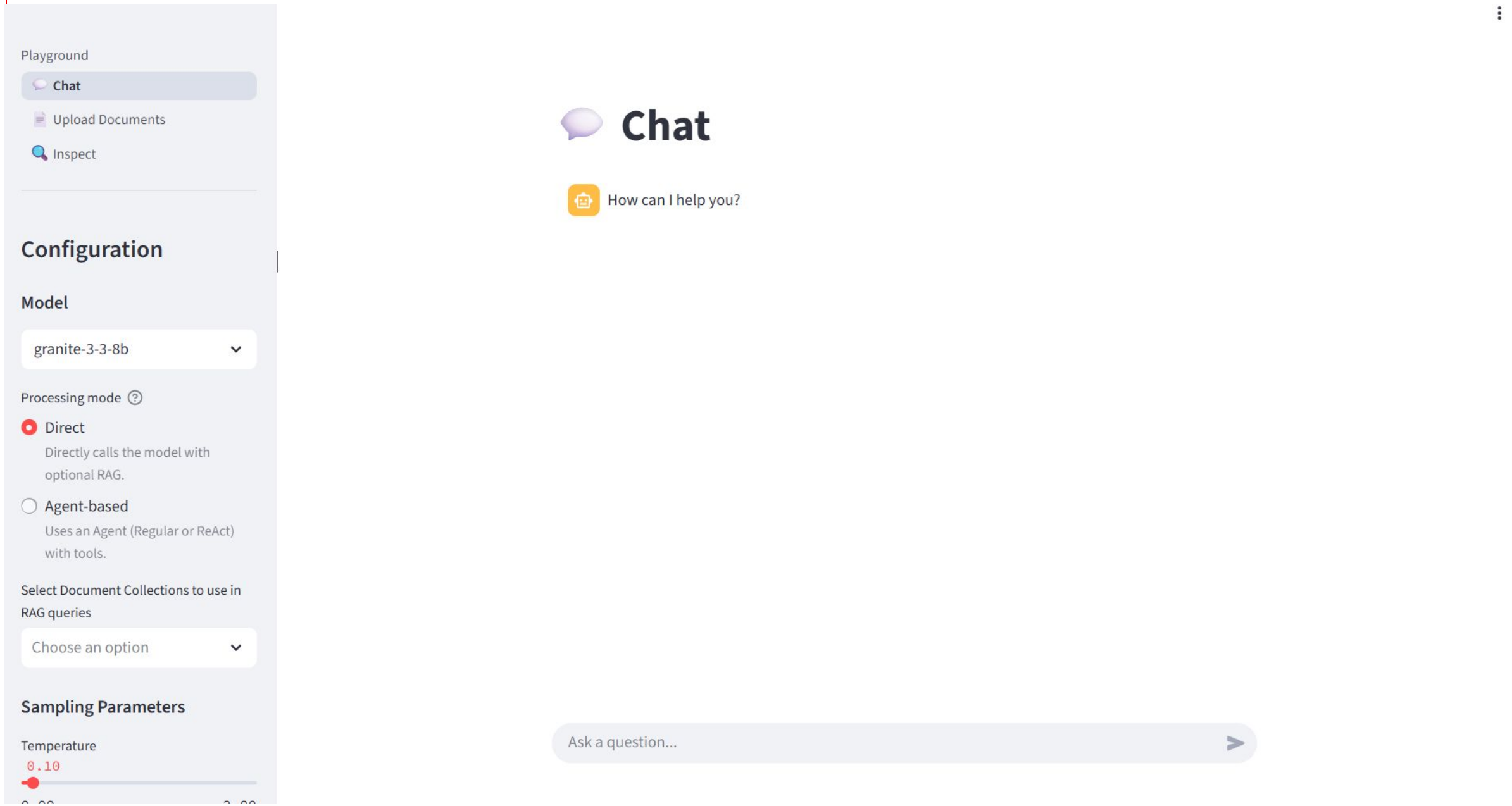
Token authentication ⚠ Tokens disabled

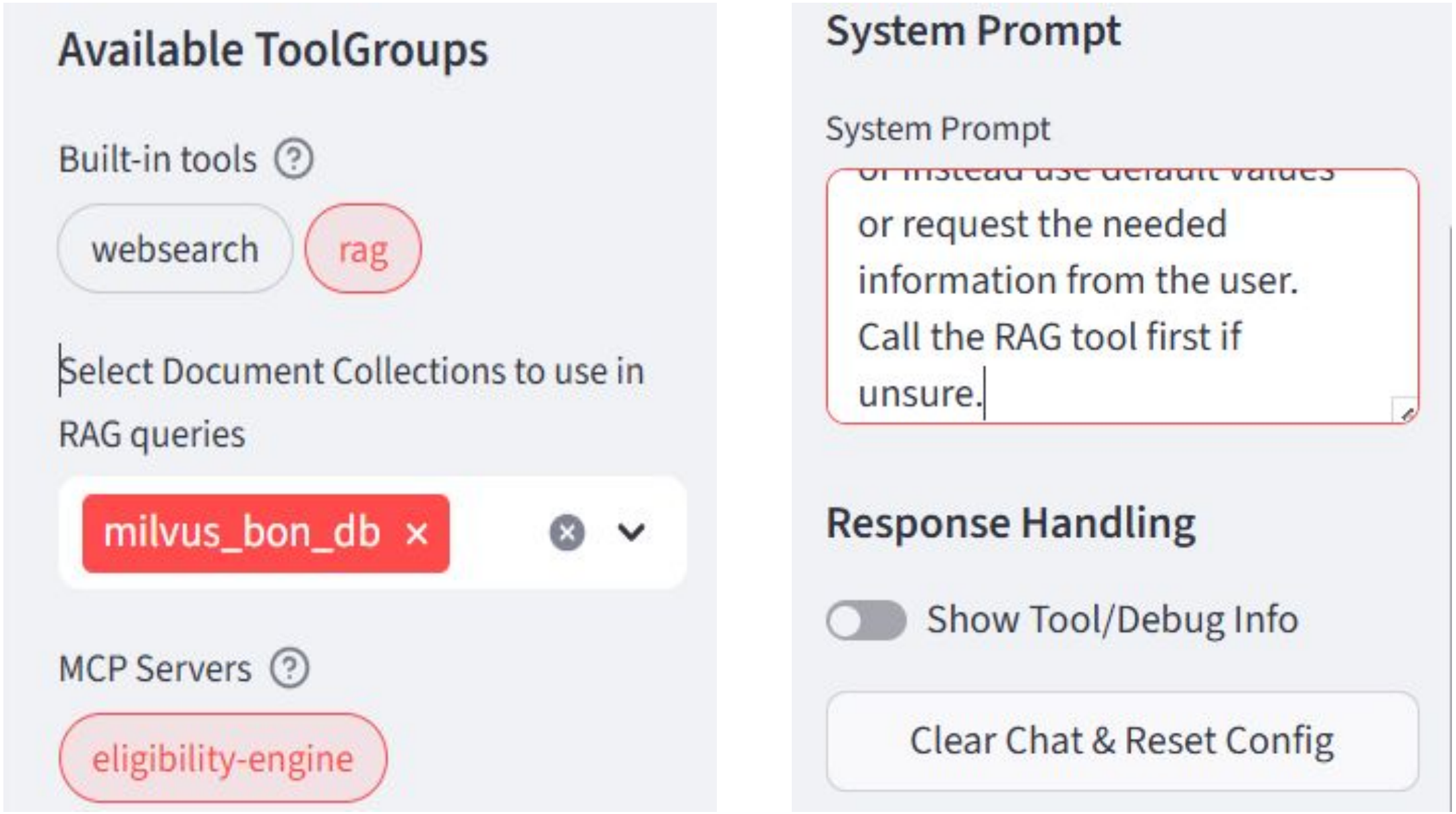
Model deployment name	Serving runtime	Inference endpoint	API protocol	Status
Granite 3.3 8B	vLLM Intel Gaudi Accelerator ServingRuntime for KServe 0.9.01	Internal endpoint details	REST	✓

Agentic AI Demo - Agentic AI Application Video Slides

Red Hat | intel.







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.

Ask a question...

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 finetuning 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.

Come visit the Intel and Red Hat booths to learn more!

Red Hat Summit

Connect

Thank you


linkedin.com/company/red-hat


facebook.com/redhatinc


youtube.com/user/RedHatVideos


twitter.com/RedHat