



**Connect**

# Agentic AI in Action

Red Hat & Intel Shaping the Future of Enterprise AI

Brussels

25 September 2025

Speaker  
Image

# Jurgen Eijmberts

Sales Strategy Specialist  
Intel



Speaker  
Image

# Hichem Kenniche

AI SSP, EMEA  
Red Hat



# Intel - RH Partnership

## Open-source software:

### Intel is committed

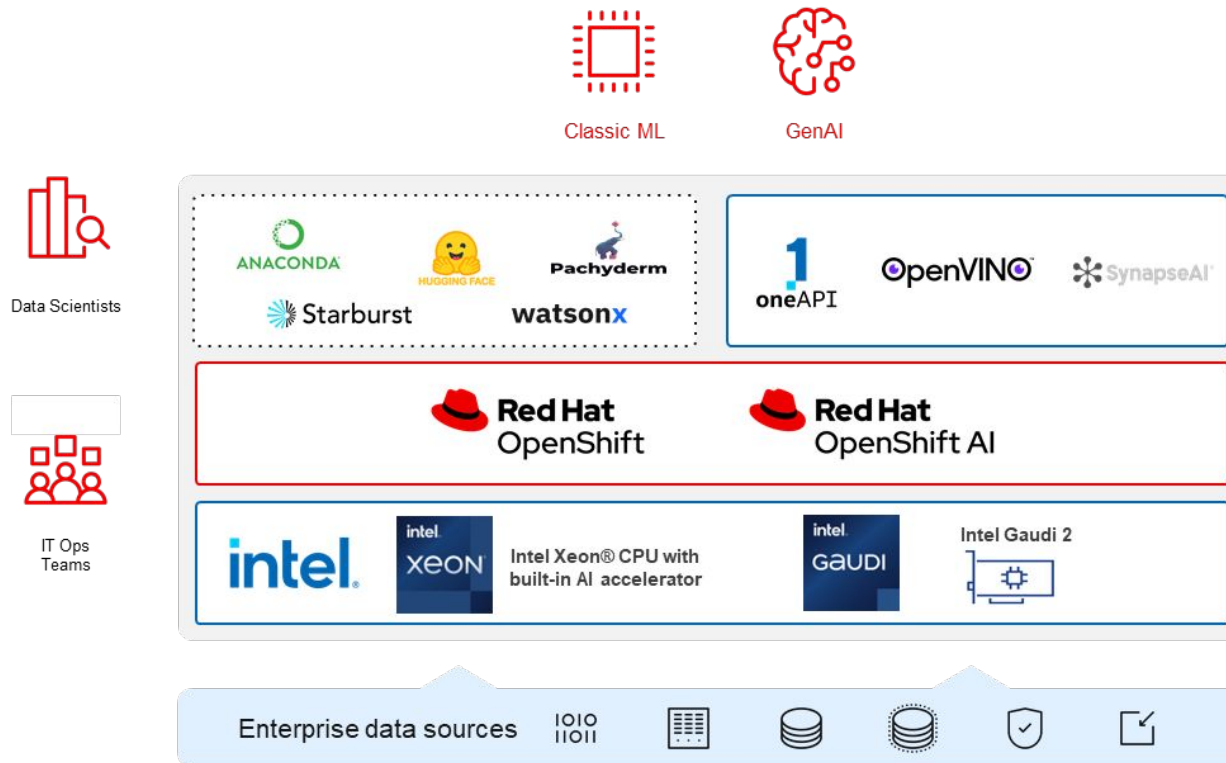
Intel® has a long history with Linux®, actively participating in open-source development and collaboration with the Linux community, to ensure hardware is well-supported and delivers optimal performance on Linux-based systems.

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



Intel is one of the key contributors to Linux  
Linux Kernel Corporate Contributor since 2007<sup>1</sup>

# 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

# Intel's AI Strategy and Capabilities

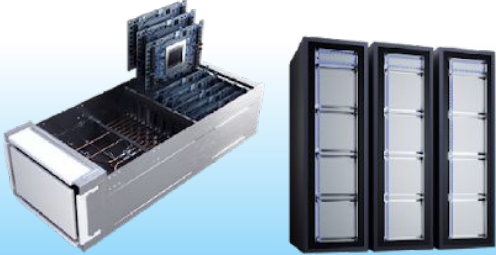
# Bringing AI Everywhere

## Intel's AI Strategy



AI PC Node  
AI Developer Productivity & Light Inference

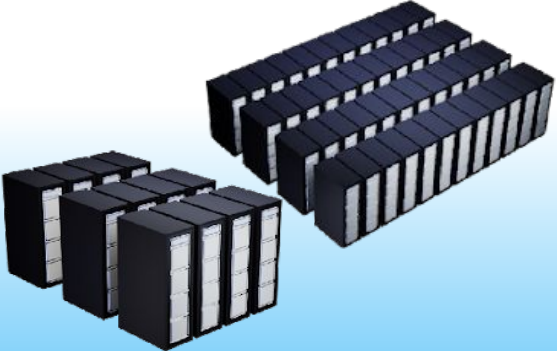
AI PC  
Broadest AI SW Ecosystem



Node  
Fine-tuning, Inference

Cluster  
Light Training, Tuning, Peak Inference

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



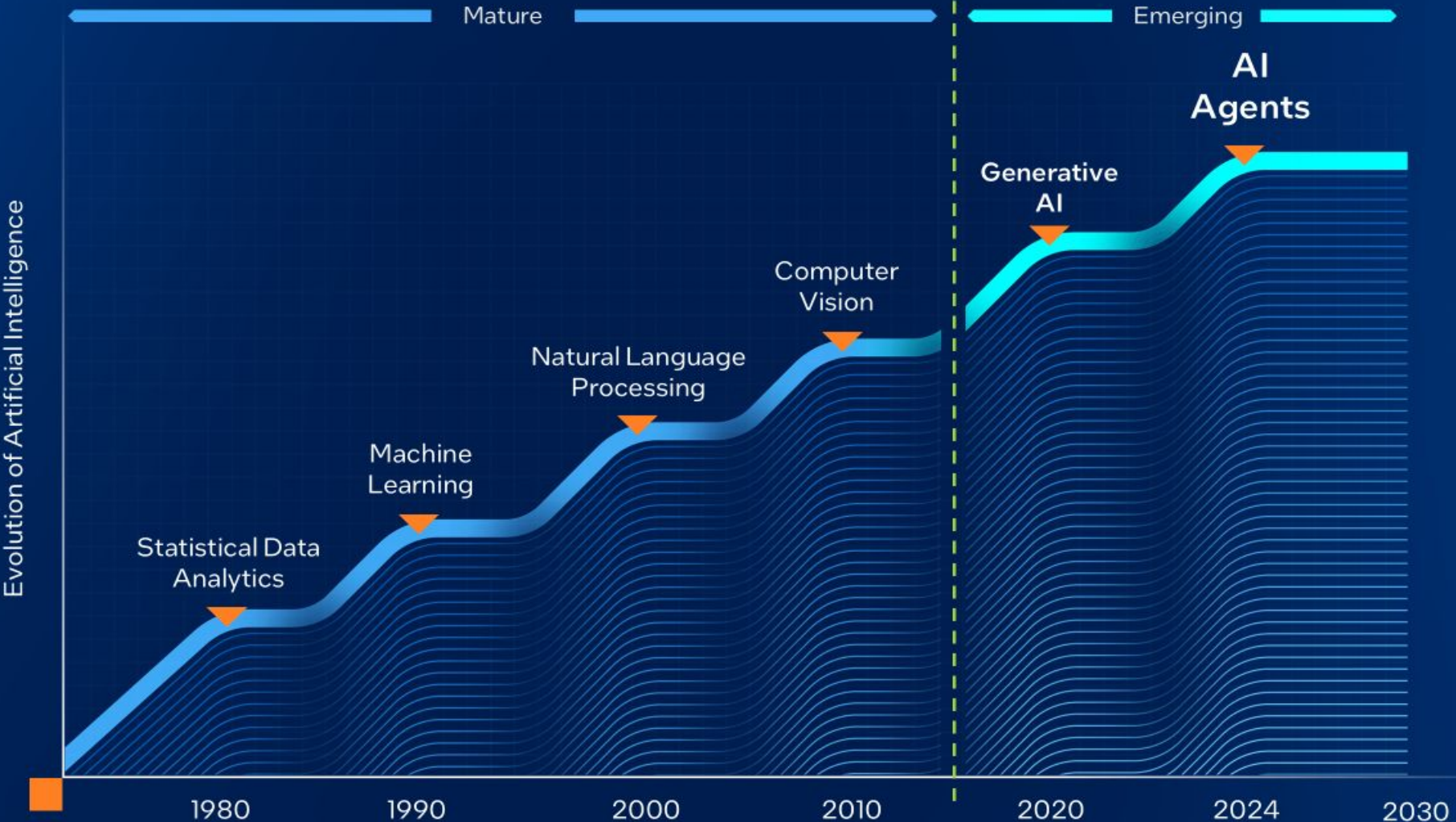
Super Cluster  
Training, Tuning, Peak Inference

Mega Cluster  
Large Scale Training & Inference

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

AI PC to Edge to Datacenter & Cloud

Efficient

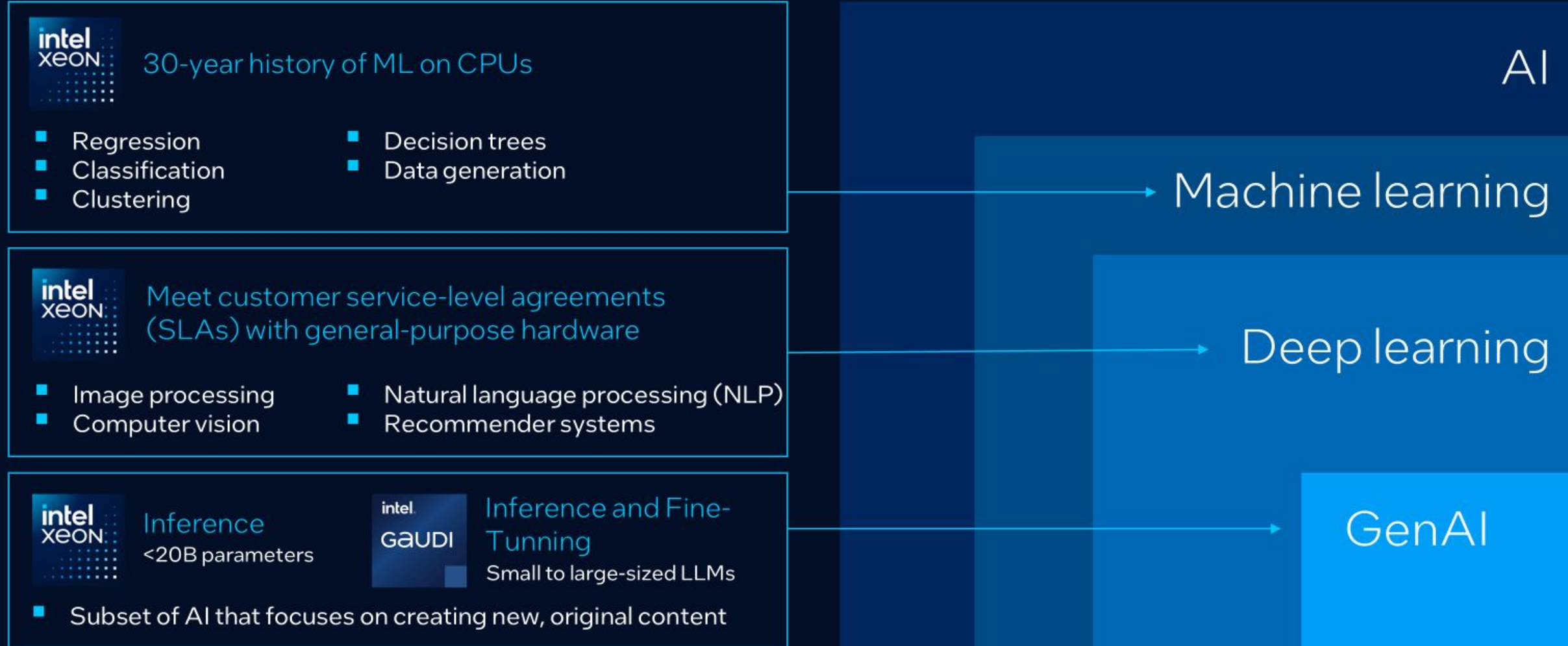
Performance per \$ & per W leadership

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)

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

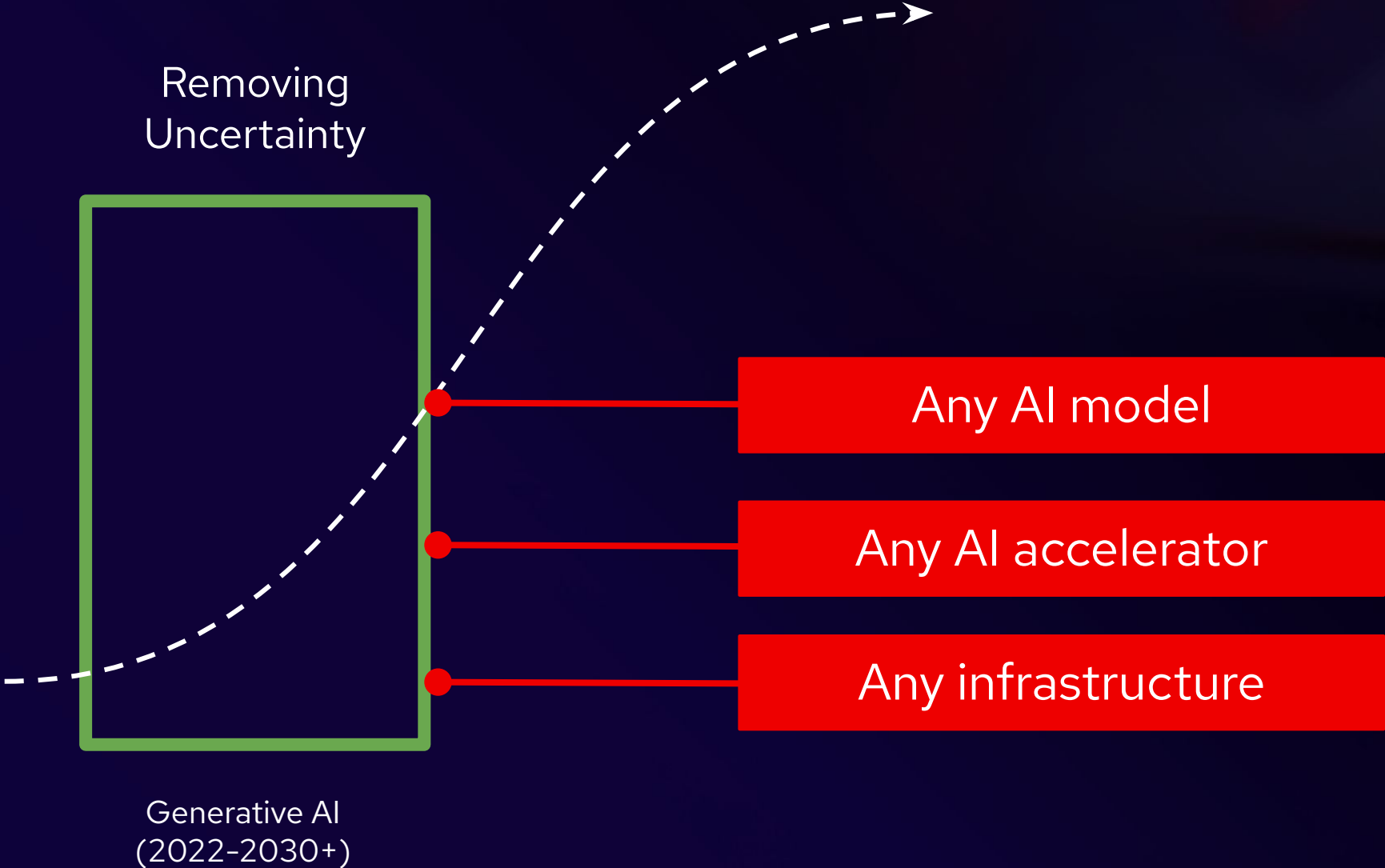


Classical Machine Learning

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 AI - Enabling AI Success





## 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**

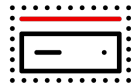
Trusted, Consistent and Comprehensive foundation



Hardware Acceleration



Physical



Virtual



Private  
Cloud



Public  
Cloud



Edge

# Intel Gaudi AI Accelerators



# 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

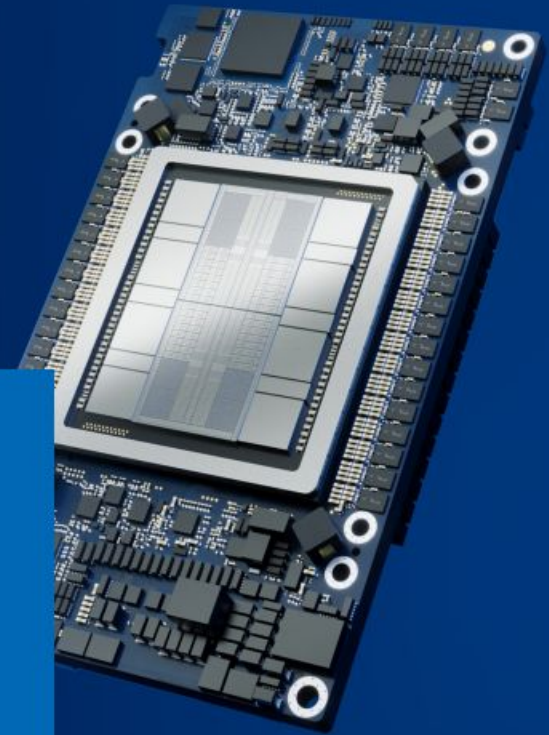
**92%**

### 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 NVIDIA H100 GPU and NVIDIA H200 GPU. Results may vary.

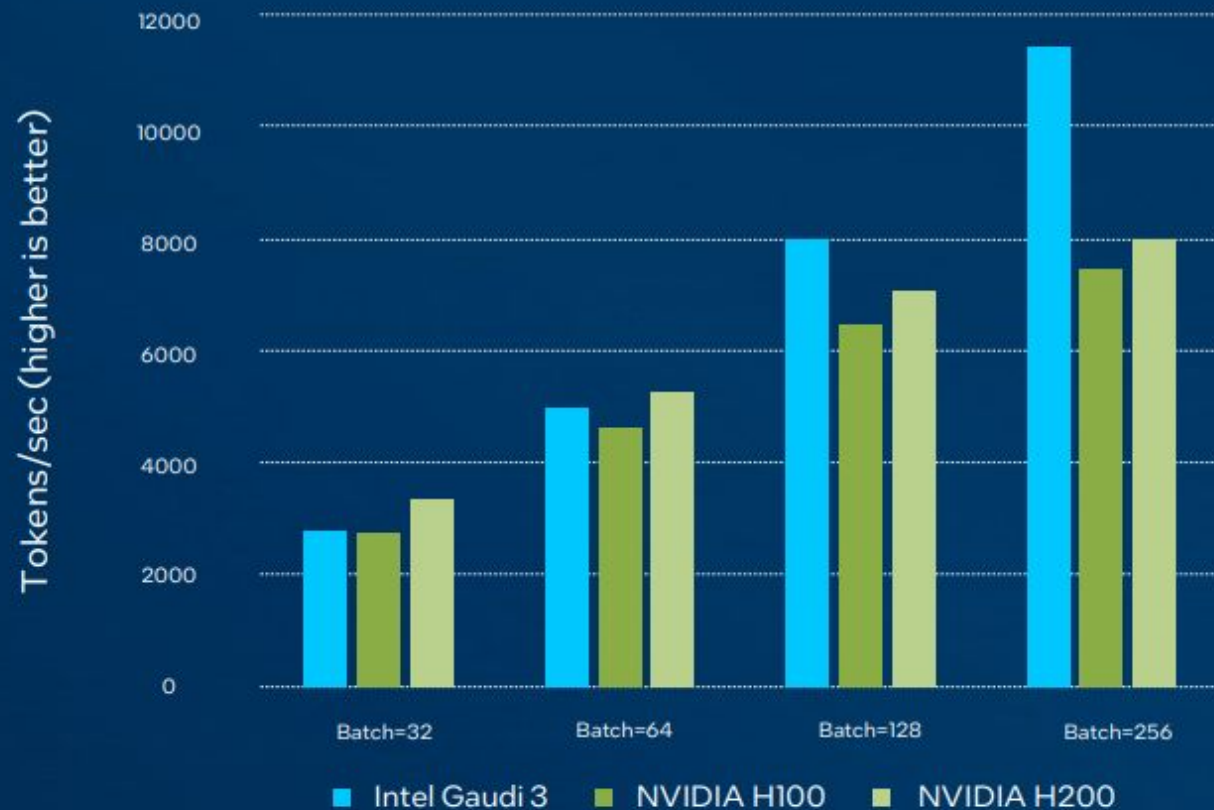
Refer to this link for the latest published Gaudi3 performance

<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.

Granite-3.1-8B-Instruct  
TP = 1, Input = 128, Output = 128 Inferencing Throughput Testing



Up to **36% higher**  
throughput than NVIDIA H200

Up to **200% higher**  
throughput than NVIDIA H100

For Large AI Workloads

\*Source: NVH200 comparisons based on [Signal65 Lab Insight: Intel Gaudi 3 Accelerates AI at Scale on IBM Cloud](#), April 2025.

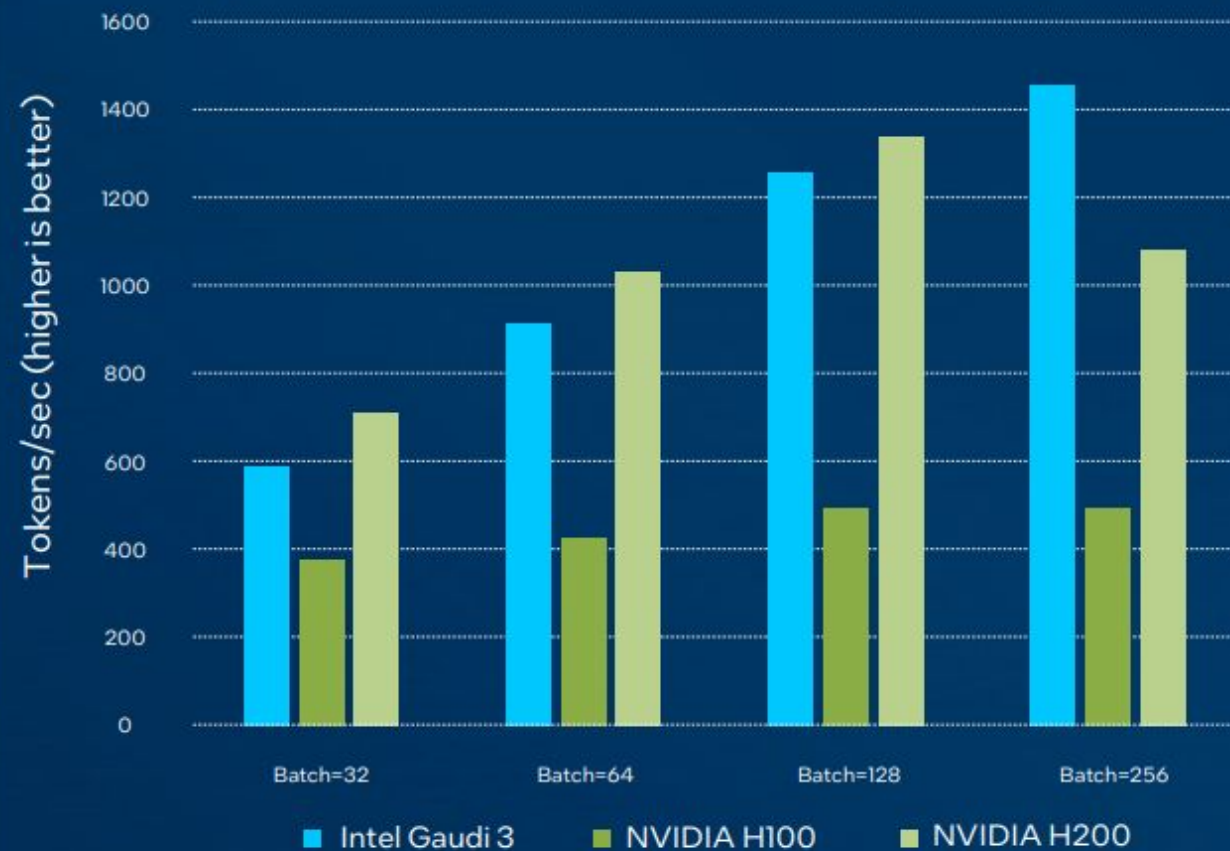
Reported numbers are inferencing results for Mixtral-8x7B-Instruct-v0.1 on Intel® Gaudi® 3 vs NVIDIA H200 GPU. Results may vary.

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.

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



# Intel Xeon Processors

intel  
xeon

# Intel<sup>®</sup> Xeon<sup>®</sup> 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



# 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	Intel Xeon 6 vs. 5th Gen Intel Xeon			5th Gen Intel Xeon vs. 3rd Gen Intel Xeon
<p>Llama2-7B</p> <p>Up to <b>1.38x</b> higher throughput</p> <p>with Intel Xeon 6980P vs. AMD EPYC 9965<sup>1</sup></p>	<p>GPTJ-6B</p> <p>Up to <b>2x</b> higher performance</p> <p>Intel Xeon 6980P vs. Intel Xeon 8592+<sup>2</sup></p>	<p>Llama-13B</p> <p>Up to <b>2x</b> higher performance</p> <p>Intel Xeon 6980P vs. Intel Xeon 8592+<sup>2</sup></p>	<p>Llama 2-7B</p> <p>Up to <b>2.3x</b> higher training performance</p> <p>Intel Xeon 6980P vs. Intel Xeon 8592+<sup>3</sup></p>	<p>Llama2-13B</p> <p>Up to <b>2.1x</b> real-time inference performance speedup</p> <p>5th Gen Intel Xeon vs. 3rd Gen Intel Xeon<sup>4</sup></p>

<sup>1,2,3,4</sup> See [9A231], [9A10], [9A6], and [A2] at [intel.com/processorclaims](https://www.intel.com/processorclaims). Intel® Xeon® 6 and 5th Generation Intel® Xeon® Scalable Processors. Results may vary.

# Intel Confidential Computing

# 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

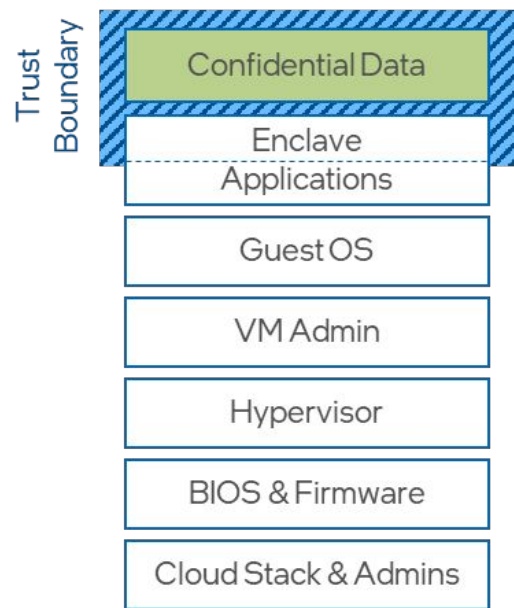
Intel adds Quantum attack protection while providing 1.89 Tb IPsec throughput.

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



# App Isolation

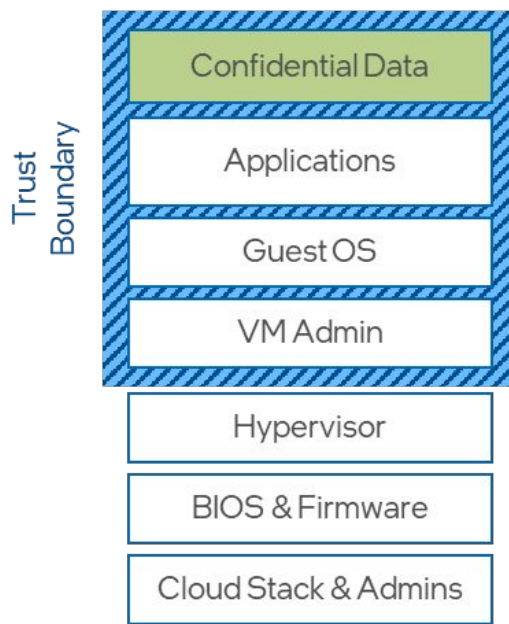
*Intel® SGX*



Smallest trust boundary for greatest data protection & code integrity

# VM Isolation

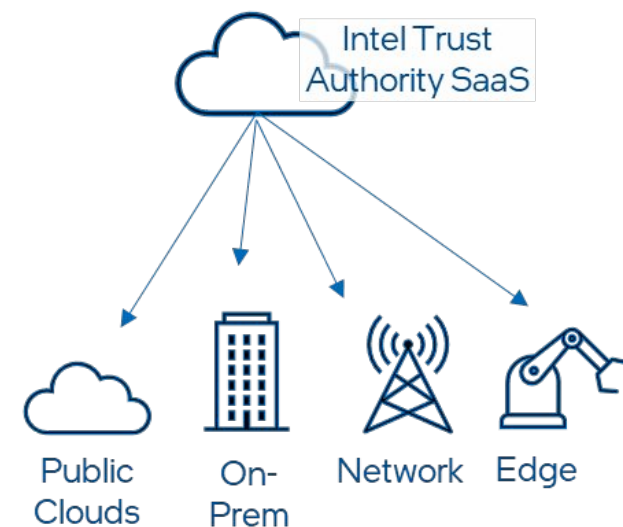
*Intel® TDX*



Most straightforward path to greater security for legacy apps

# Trust Services

*Intel® Tiber™ Trust Authority*



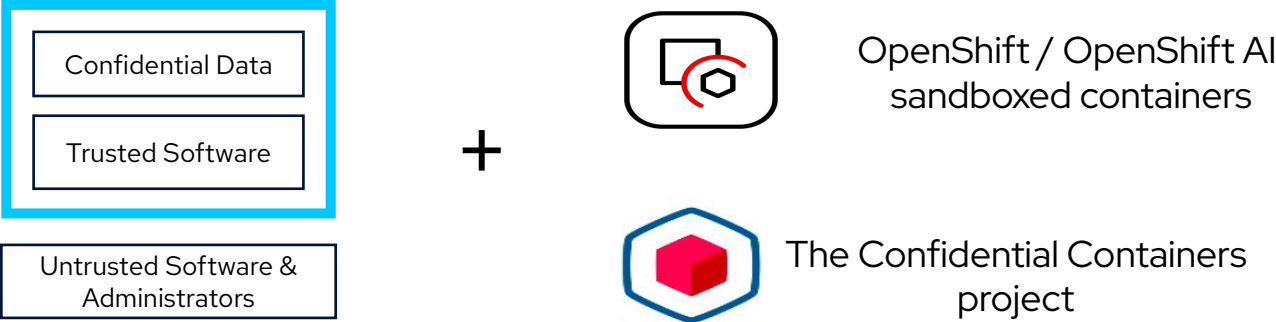
Uniform, independent attestation of trustworthy environments

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

# 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)

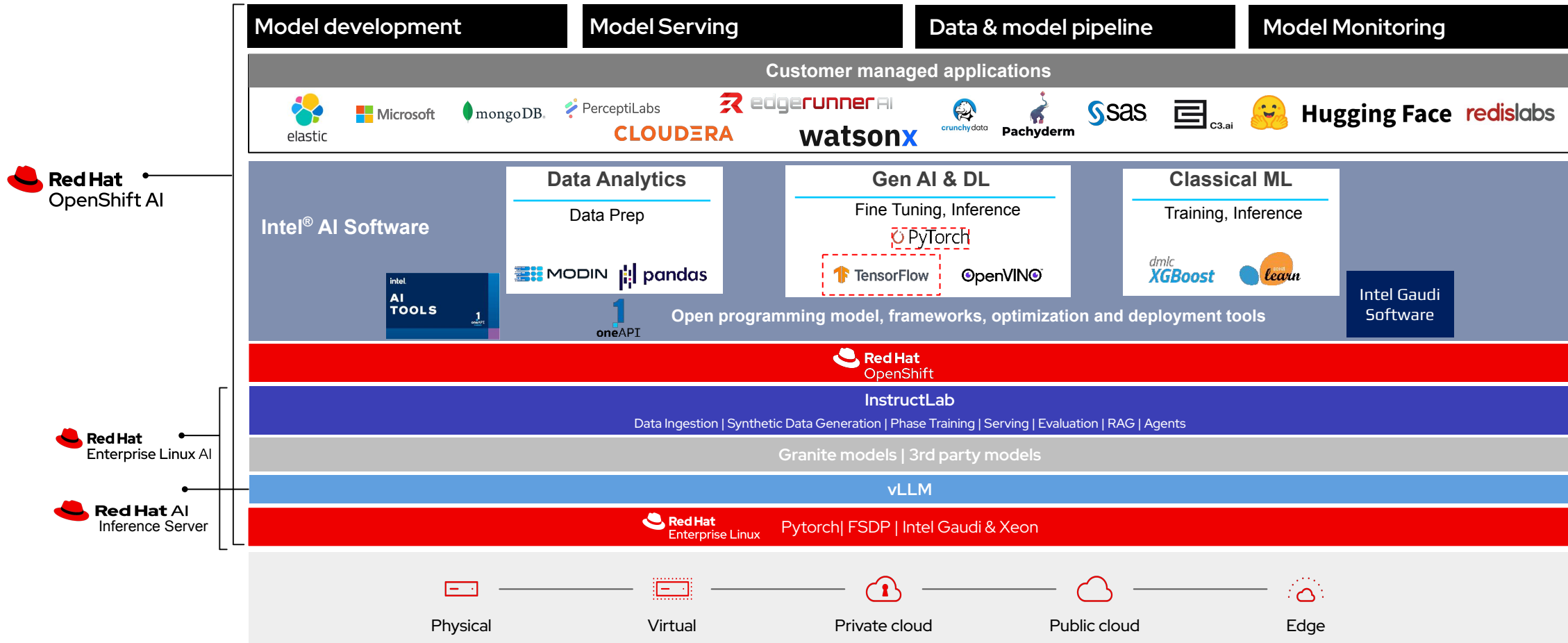


Confidential Computing is about **protecting data in-use**.  
You do not **have to trust** the system admins of the providers any longer.

# Intel AI Software

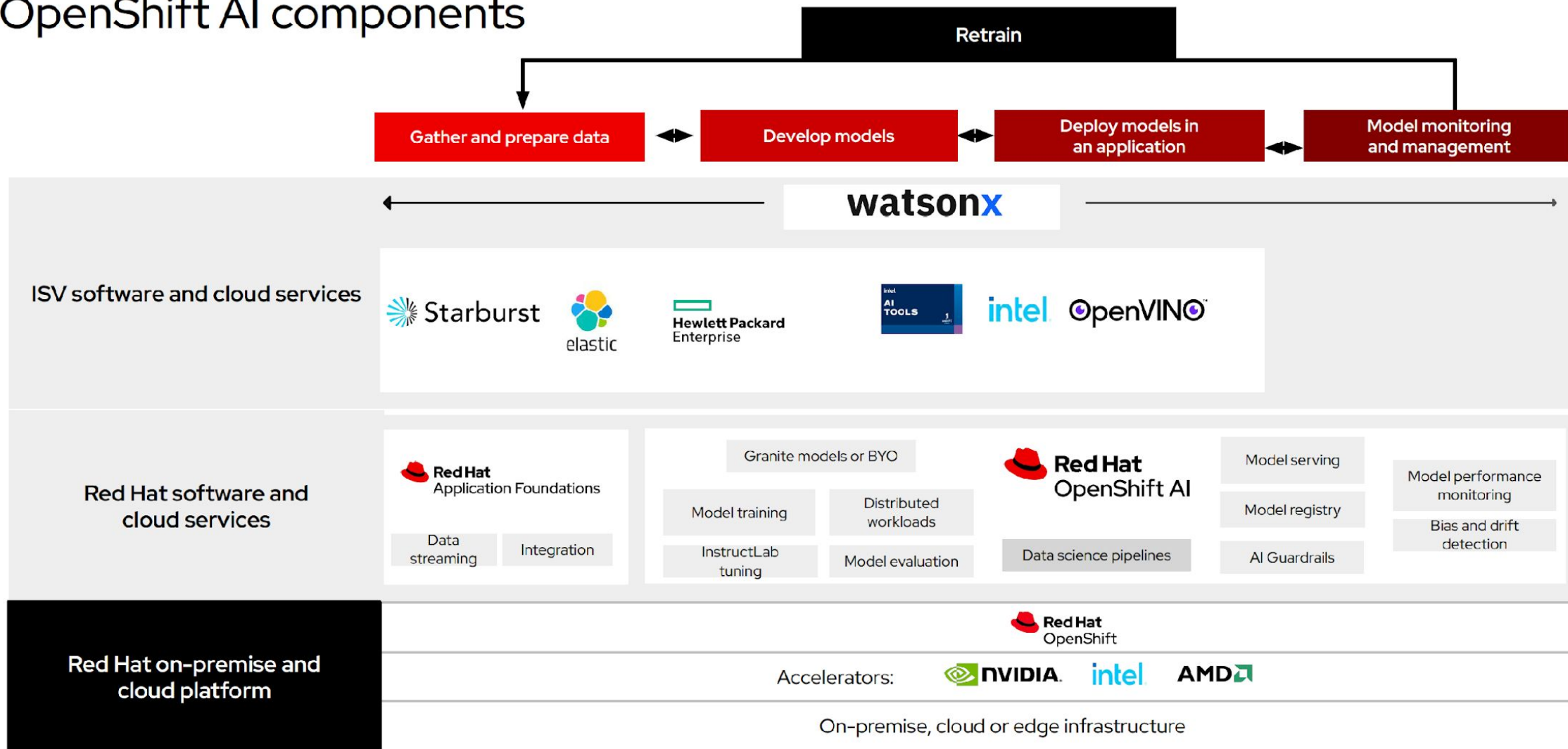
# Red Hat AI with Intel AI platform

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



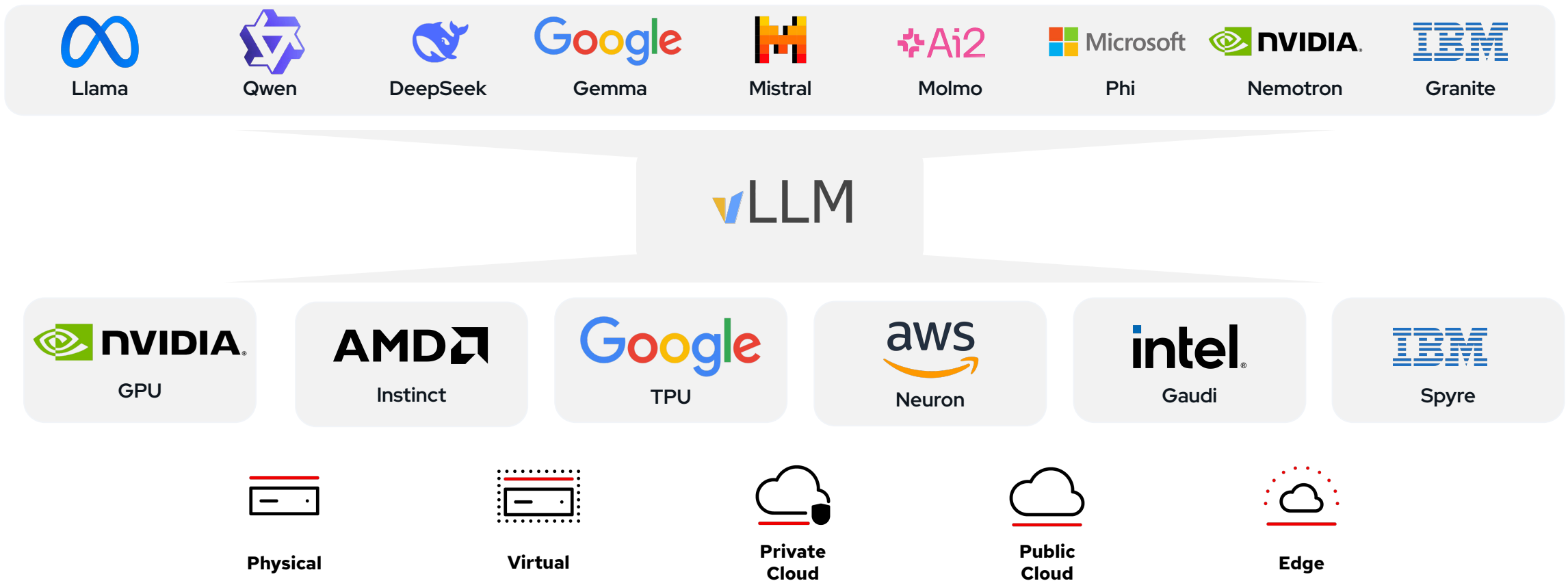
# Red Hat AI Platform

# OpenShift AI components



# 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



Microsoft

Phi



Molmo

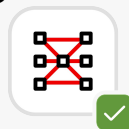


Granite



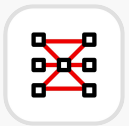
Nemotron

## Validated models



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

## Optimized models



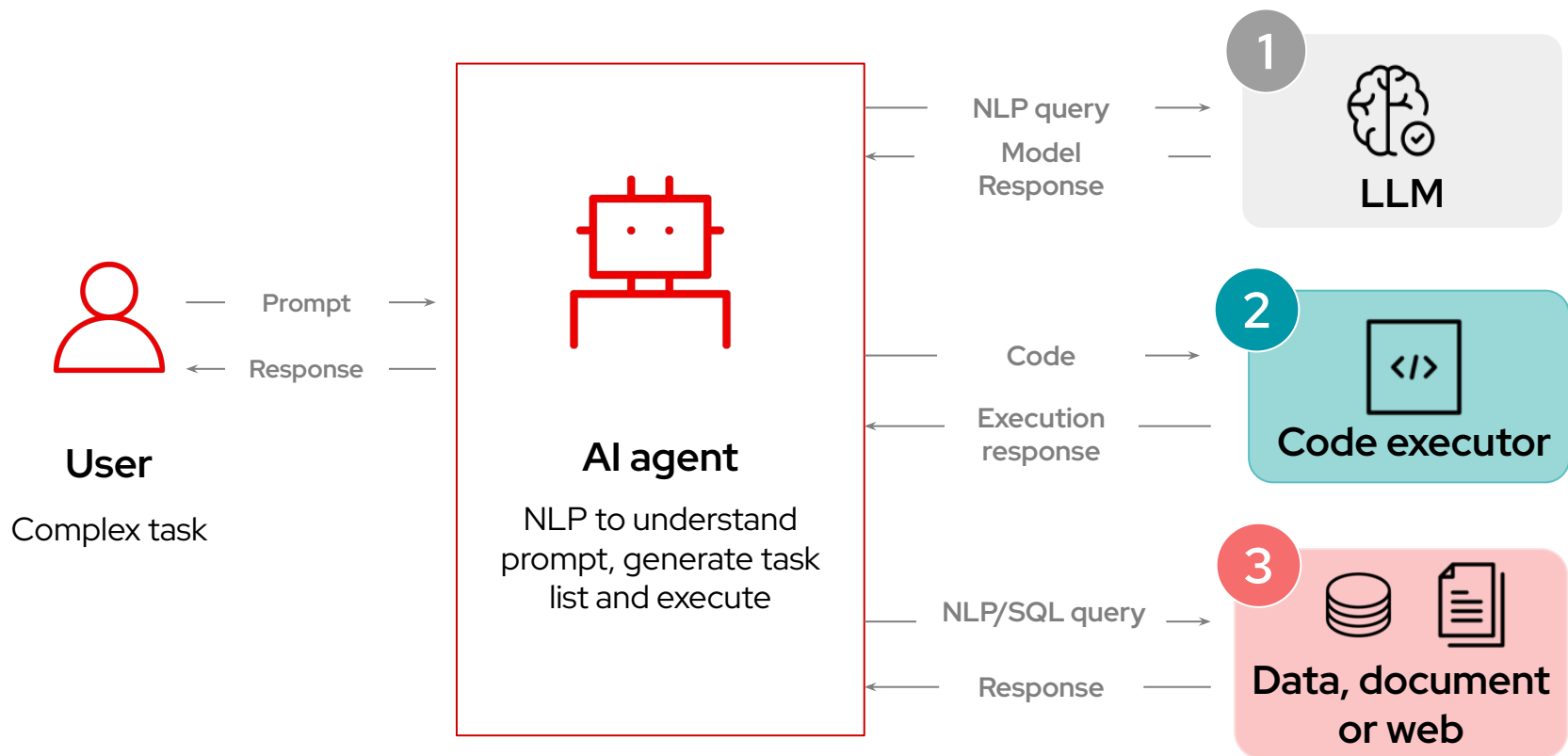
- ▶ Compressed for speed and efficiency
- ▶ Designed to run faster, use fewer resources, maintain accuracy
- ▶ Done using LLM Compressor with latest algorithms



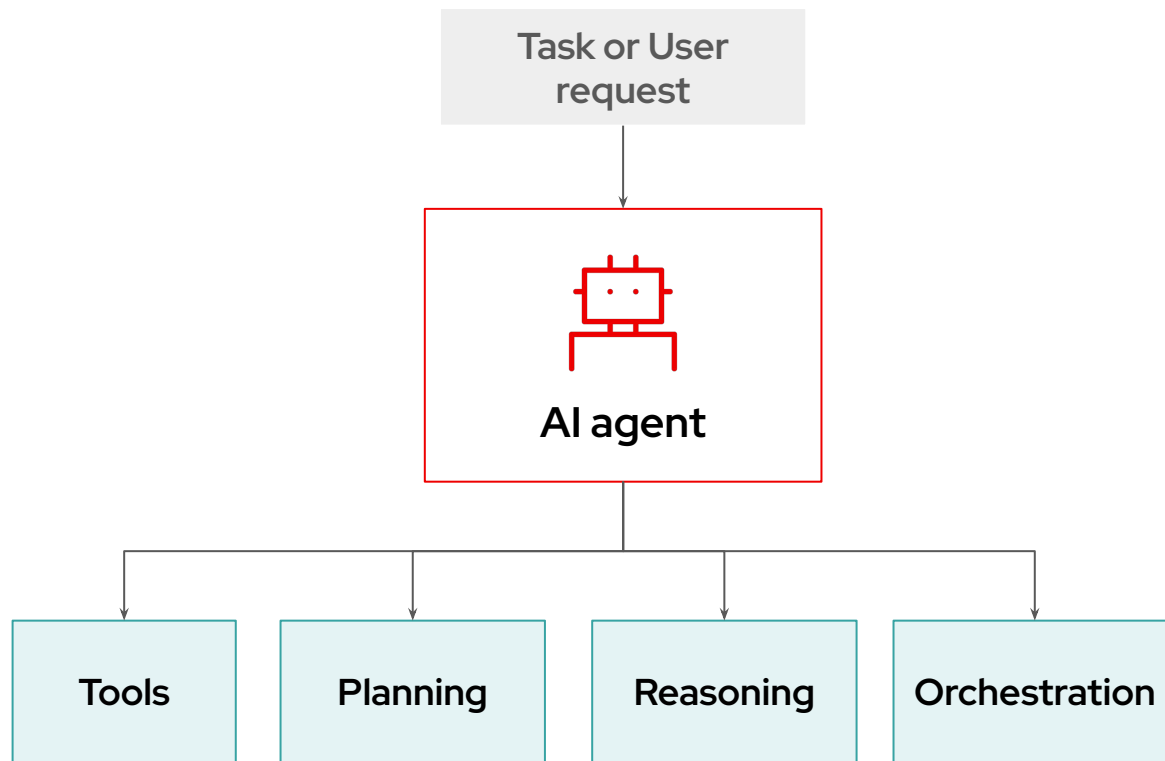
# Intro to Agentic AI

# AI agents integrate models, functions & tools

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



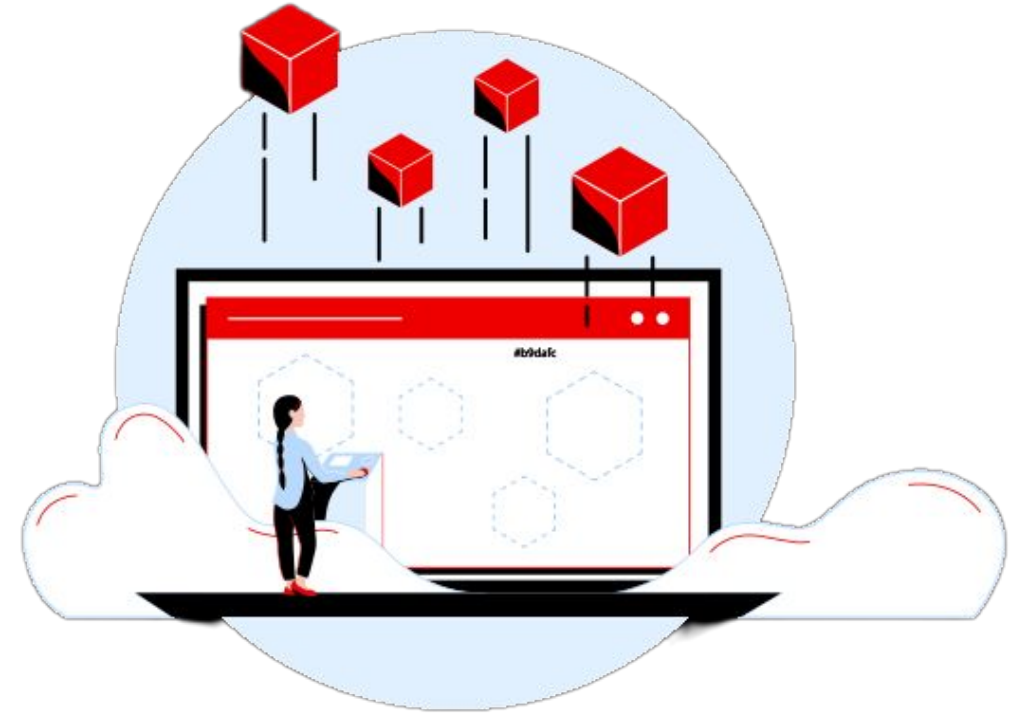
# The components of an AI Agent system



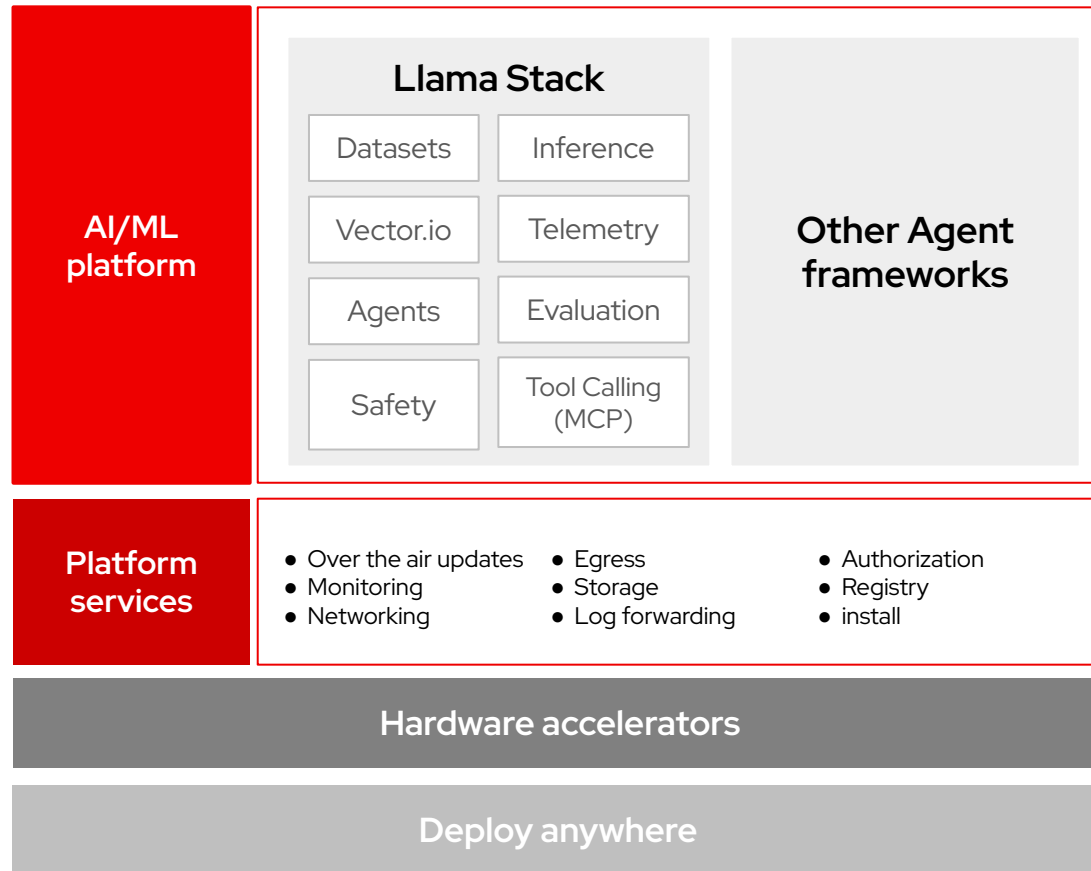
- ▶ **Tool Utilization:** Leverages external tools to gather data and perform tasks.
- ▶ **Planning and Execution:** Develops and executes multistep plans to achieve goals autonomously.
- ▶ **Reasoning:** Applies logic and contextual understanding to make informed decisions.
- ▶ **Orchestration:** Coordinates 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.



# A modular approach to building AI agents



## Red Hat AI allows to:

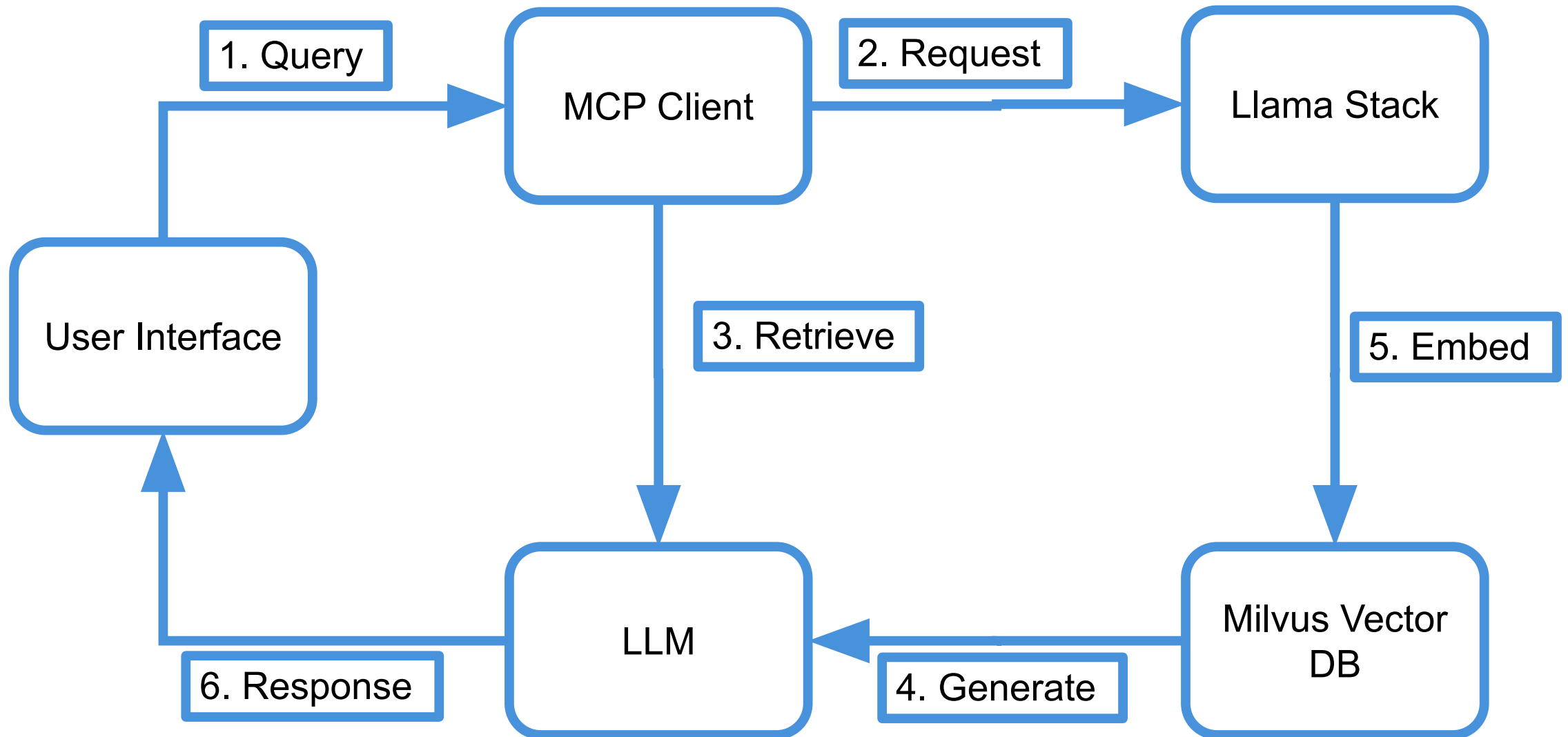
- ▶ Build agents using **Llama Stack's native capabilities and implementations.**
- ▶ **Bring compatible Llama Stack implementations** to OpenShift AI.
- ▶ **Use your own agent framework** and selectively incorporate Llama Stack APIs.
- ▶ **Build with Core Primitives** and manage your own agent framework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

# RAG vs RAG + MCP

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

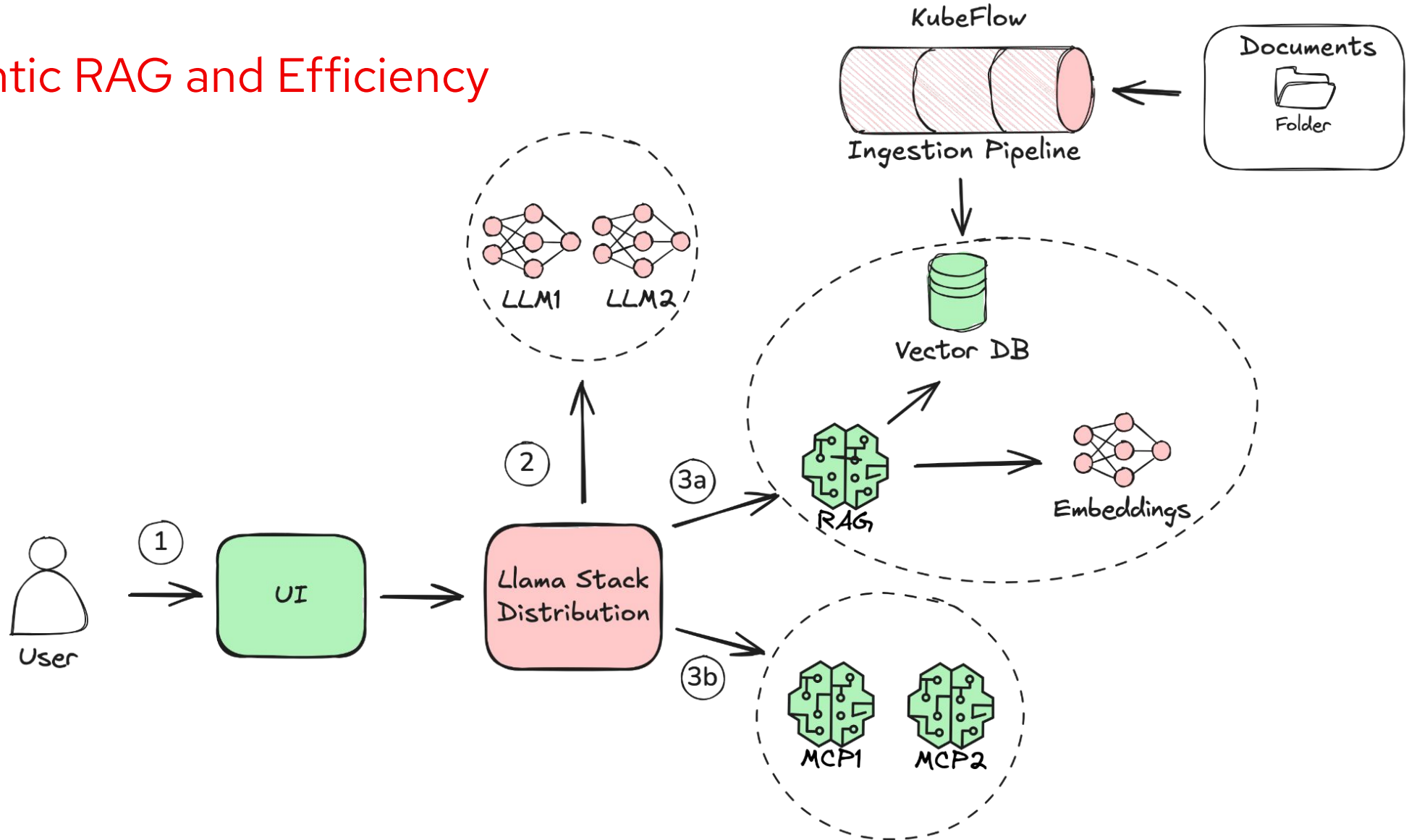
# Agentic AI Demo

# Agentic AI Demo Architecture





# Agentic RAG and Efficiency



← unpaid-leave-assistance-2025 Draft Open Simulator Versions Publish

Graph evaluate\_unpaid\_leave Graph View

**Components** (Drag-and-drop) ×

- Request
- Response
- Decision table
- Expression
- Function
- Switch
- Decision

```
graph LR; request[request] --> evaluate_relation[evaluate_relation...]; evaluate_relation --> evaluate_unpaid_l[evaluate_unpaid_l...]; evaluate_unpaid_l --> result[result];
```

request :  
Configure

evaluate\_relation... :  
Edit Table Settings

evaluate\_unpaid\_l... :  
Edit Table Settings

result :  
Configure

+  
-  
↺  
↻  
▶

← unpaid-leave-assistance-2025 Draft Close Simulator Versions Publish

Graph evaluate\_unpaid\_leave × evaluate\_relationship Graph View

Export Excel Import Excel

	Inputs				Outputs				
	Family relationship	Situation	Single-parent family	Number of children	Potentially eligible	Monthly benefit	Case	Description	Output
	valid_relationship	input.situation	input.single_parent...	input.number_of_ch...	output.potentially_...	output.monthly_ben...	output.case	output.description	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 1"
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 Search nodes... Output Input Trace

```

1 {
2   "input": {
3     "relationship": "mother",
4     "situation": "accident",
5     "single_parent_family": false,
6     "number_of_children": 0
7   }
8 }

```

Graph 312.6µs  
 evaluate\_relationship 41.2µs  
 evaluate\_unpaid\_leave 41.6µs  
 result 1.3µs

```

1 {
2   input: {
3     number_of_children: 0,
4     relationship: 'mother',
5     single_parent_family: false,
6     situation: 'accident',
7   },
8   output: {
9     additional_requirements: 'The person must have been hospitalized and the care of the person must be continued',
10    case: 'A',
11    description: 'First-degree family care sick or accident victim',
12    monthly_benefit: 725,
13    potentially_eligible: true,
14  },
15  valid_relationship: true,
16 }

```

# 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

- 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

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	<input checked="" type="checkbox"/>	2 days ago	⋮

1-1 of 1
<< < 1 of 1 > >>

- 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	<input checked="" type="checkbox"/>	Single-model	REST
■■■ vLLM Intel Gaudi Accelerator ServingRuntime for KServe 0.9.01 ? v0.9.01	<input checked="" type="checkbox"/>	Single-model	REST
■■■ vLLM CPU for Intel Xeon ? v0.3.ubi9	<input checked="" type="checkbox"/>	Single-model	REST
■■■ vLLM Intel Gaudi Accelerator ServingRuntime for KServe ? Pre-installed v0.7.2	<input checked="" type="checkbox"/>	Single-model	REST
■■■ OpenVINO Model Server ? Pre-installed v2025.1	<input checked="" type="checkbox"/>	Single-model	REST



- 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-v0.1

**Mixtral-8x7B-Instruct-v0.1**  
Provided by Mixtral

### Description

The Mixtral-8x7B Instruct model is looking forward to engaging with th

### Model card

[View Red Hat validated benchmark](#)

### Model Card for Mixtral-8x7B-Instruct-v0.1

#### Tokenization with mistral-common

```

from mistral_common.tokens.tokenizers.mistral import MistralTokenizer
from mistral_common.protocol.inference.common import CompletionRequest
from mistral_common.protocol.inference.openai_compat import OpenAICompatRequest

mistral_models_path = "MISTRAL_MODELS_PATH"

tokenizer = MistralTokenizer.from_pretrained(mistral_models_path)

completion_request = ChatCompletionRequest(
    model="mistral-large-latest",
    messages=[{"role": "user", "content": "Hello!"}],
)

openai_compat_request = OpenAICompatRequest.from_completion_request(completion_request)

tokens = tokenizer.encode_chat_completion(openai_compat_request)
  
```

#### Inference with mistral-inference

```

from mistral_inference.transformers import Transformer
from mistral_inference.generator import MistralGenerator

model = Transformer.from_folder(mistral_models_path)

generator = MistralGenerator(model)

out_tokens, _ = generate([tokens])

result = tokenizer.decode(out_tokens)

print(result)
  
```

Inference with hugging face tr

## 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      Maximum replicas

- 1 +      - 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

Deploy model Register model

### Version

1.4

### Labels

- LAB teacher
- text-generation
- validated

### License

[Agreement](#)

### Provider

Mixtral

### Model location

oci://registry.redhat.io/rhelai/mo...

### Last modified

20 Jan 1970

### Published

20 Jan 1970



## Models and model servers ?

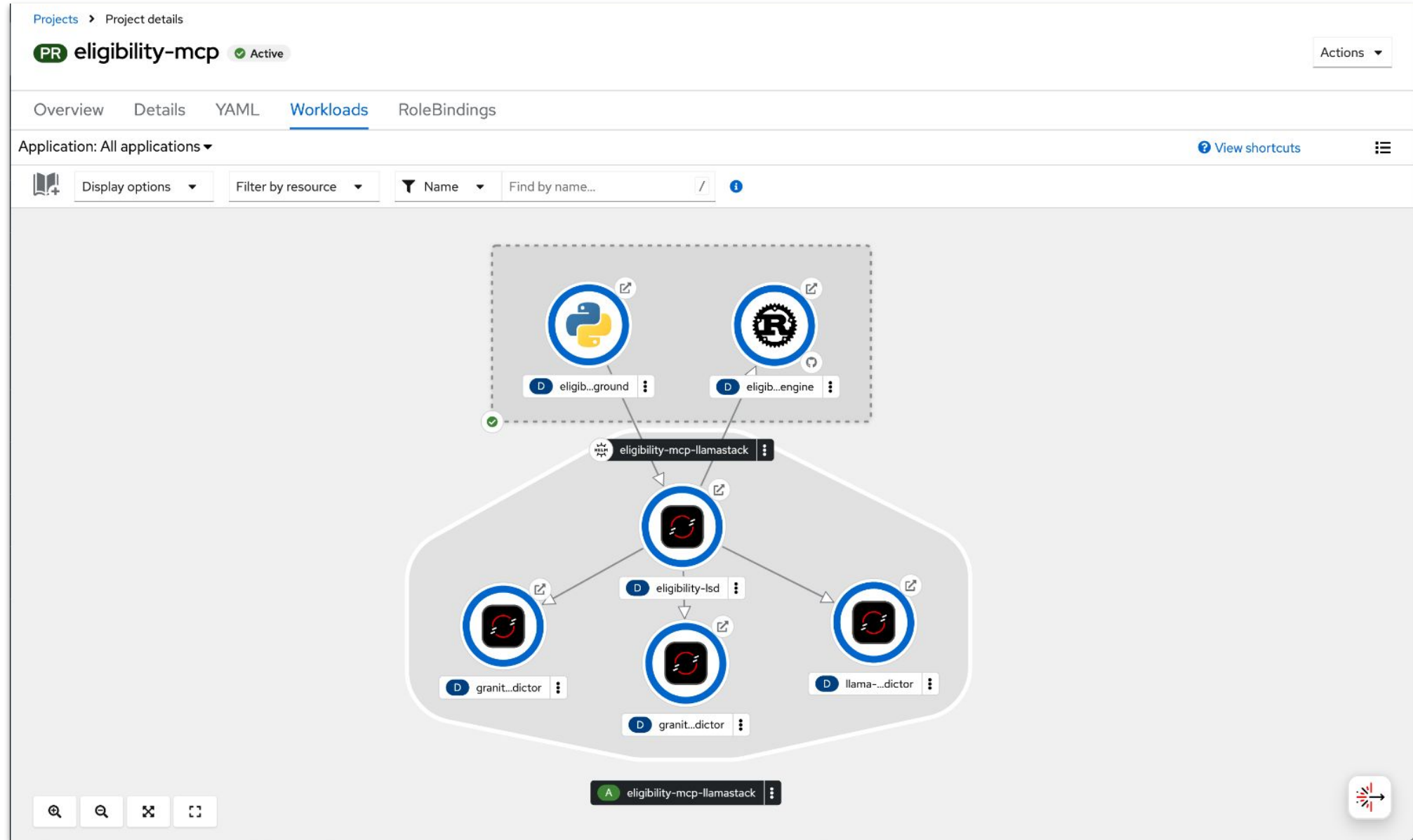
Deploy model

Single-model serving enabled

Model deployment name <span>↑</span>	Serving runtime	Inference endpoint	API protocol	Status
<span>▼</span> <a href="#">Granite 3.3 2B</a> <span>?</span>	vLLM CPU for Intel Xeon AMX optimized	<a href="#">Internal endpoint details</a>	REST	<span>✓</span> <span>⋮</span>
<b>Framework</b>	vLLM			
<b>Model server replicas</b>	1			
<b>Model server size</b>	Custom 64 CPUs, 24GiB Memory requested 64 CPUs, 24GiB Memory limit			
<b>Accelerator</b>	No accelerator selected			
<b>Token authentication</b>	<span>⚠</span> Tokens disabled			
<span>▶</span> <a href="#">Granite 3.3 8B</a> <span>?</span>	vLLM Intel Gaudi Accelerator ServingRuntime for KServe 0.9.01	<a href="#">Internal endpoint details</a>	REST	<span>✓</span> <span>⋮</span>

# Agentic AI Demo

- Agentic AI
- Application Video
- Slides



Playground

Chat

Upload Documents

Inspect

## 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.

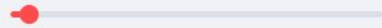
Select Document Collections to use in RAG queries

Choose an option


### Sampling Parameters

Temperature

0.10



 Chat

 How can I help you?

Ask a question...



## Available ToolGroups

Built-in tools ?

websearch

rag

Select Document Collections to use in RAG queries

milvus\_bon\_db x



MCP Servers ?

eligibility-engine

## System Prompt

System Prompt

or instead use default values  
or request the needed  
information from the user.  
Call the RAG tool first if  
unsure.

## Response Handling

Show Tool/Debug Info

Clear Chat & Reset Config

# 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!



Connect

# Thank you



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



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



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



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