



Connect

Agentic AI in Action

Red Hat & Intel Shaping the Future of Enterprise AI

Utrecht

15 October 2025



Jurgén Eijmberts

Technical Sales Specialist
Intel Corporation



Kyra Goud

OpenShift SSA, Western Europe
Red Hat

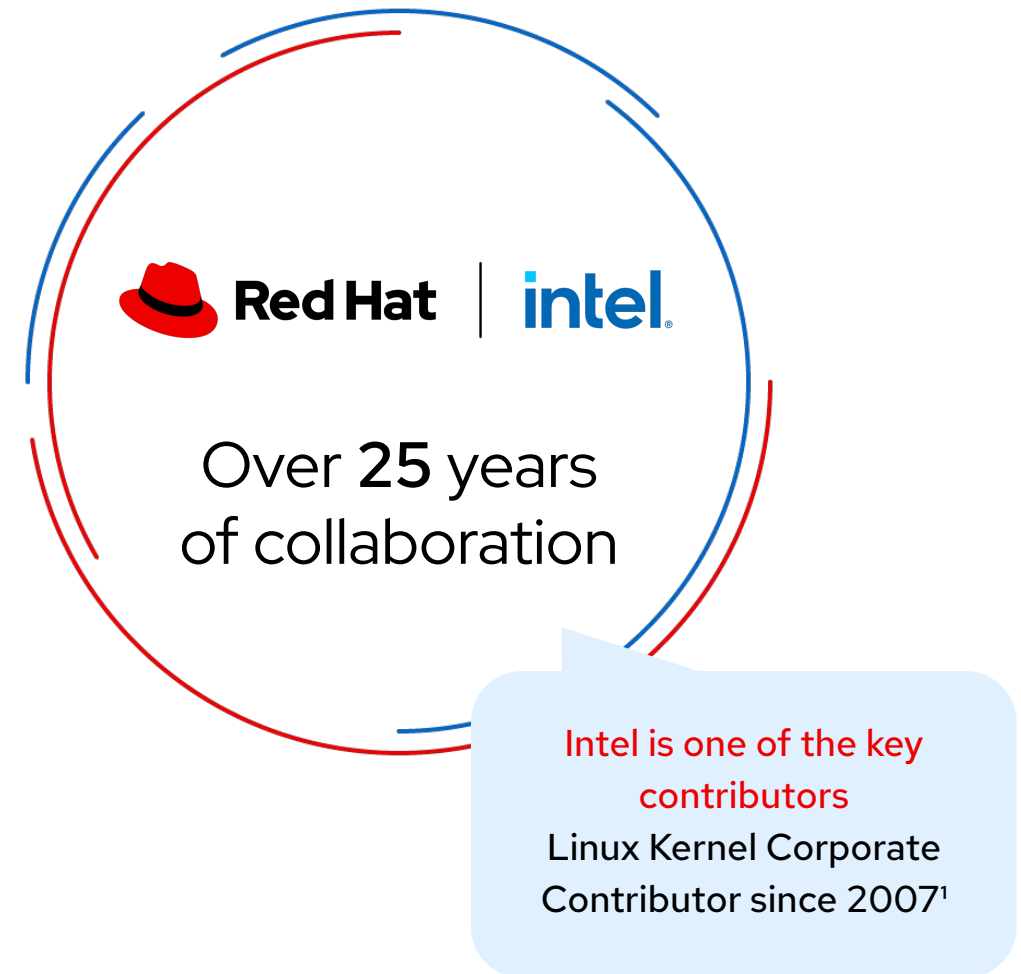


Intel – RH Partnership

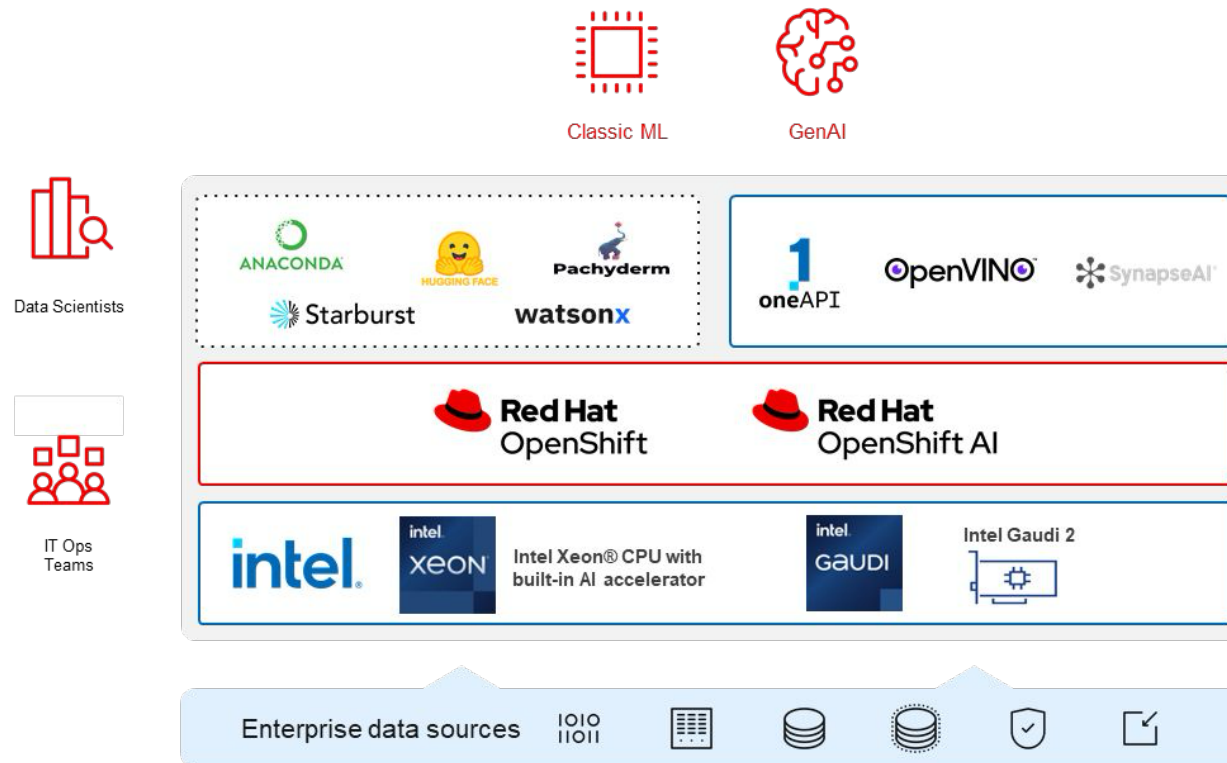
Open source software: Intel is committed

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.



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

Broadest AI SW Ecosystem



ENTERPRISE AI & EDGE AI

Open Standard, "Ready to Use"

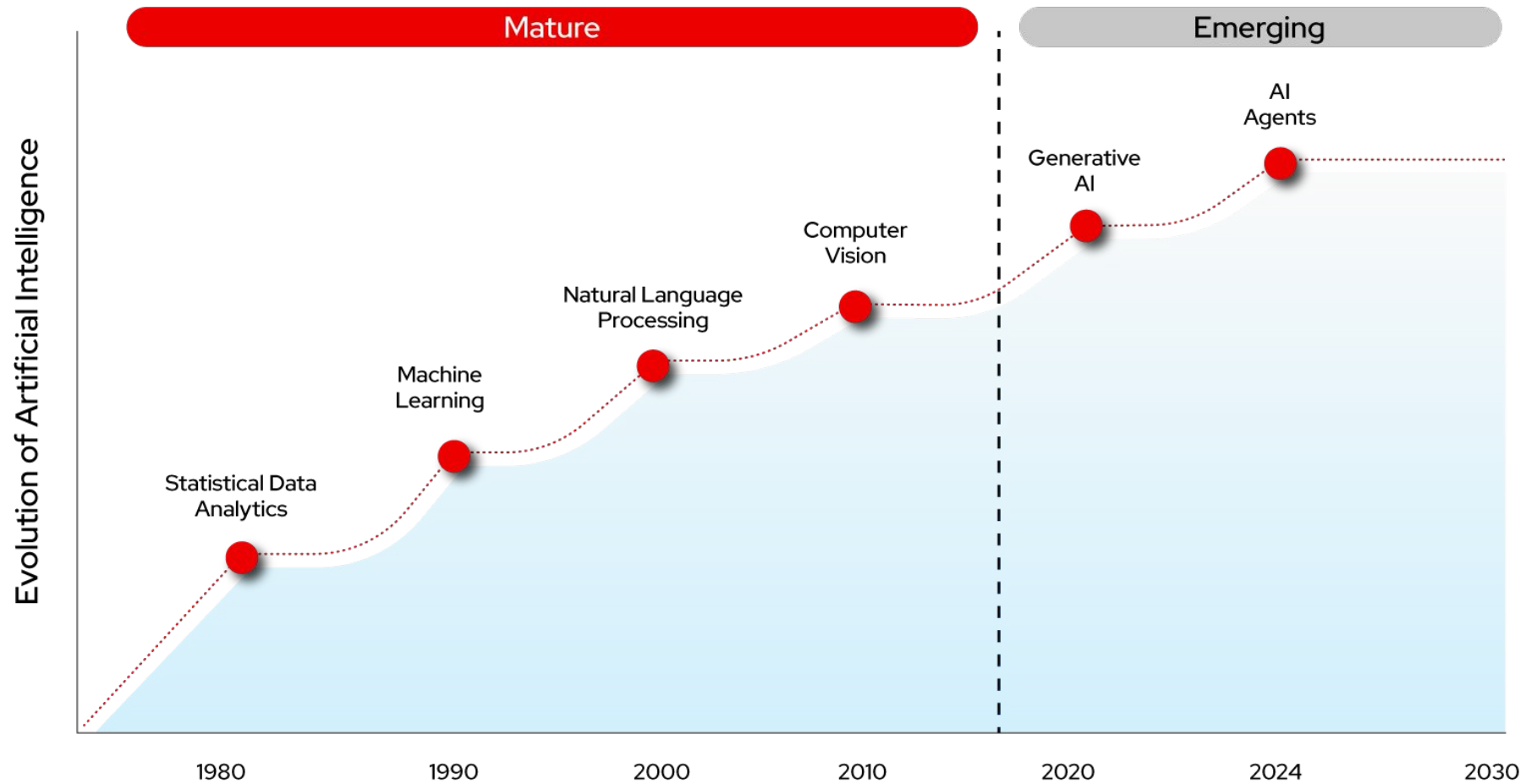


DATA CENTER AI

AI Open, Scalable Systems & Reference Arch



Evolution of AI Applications in Enterprise Use Cases

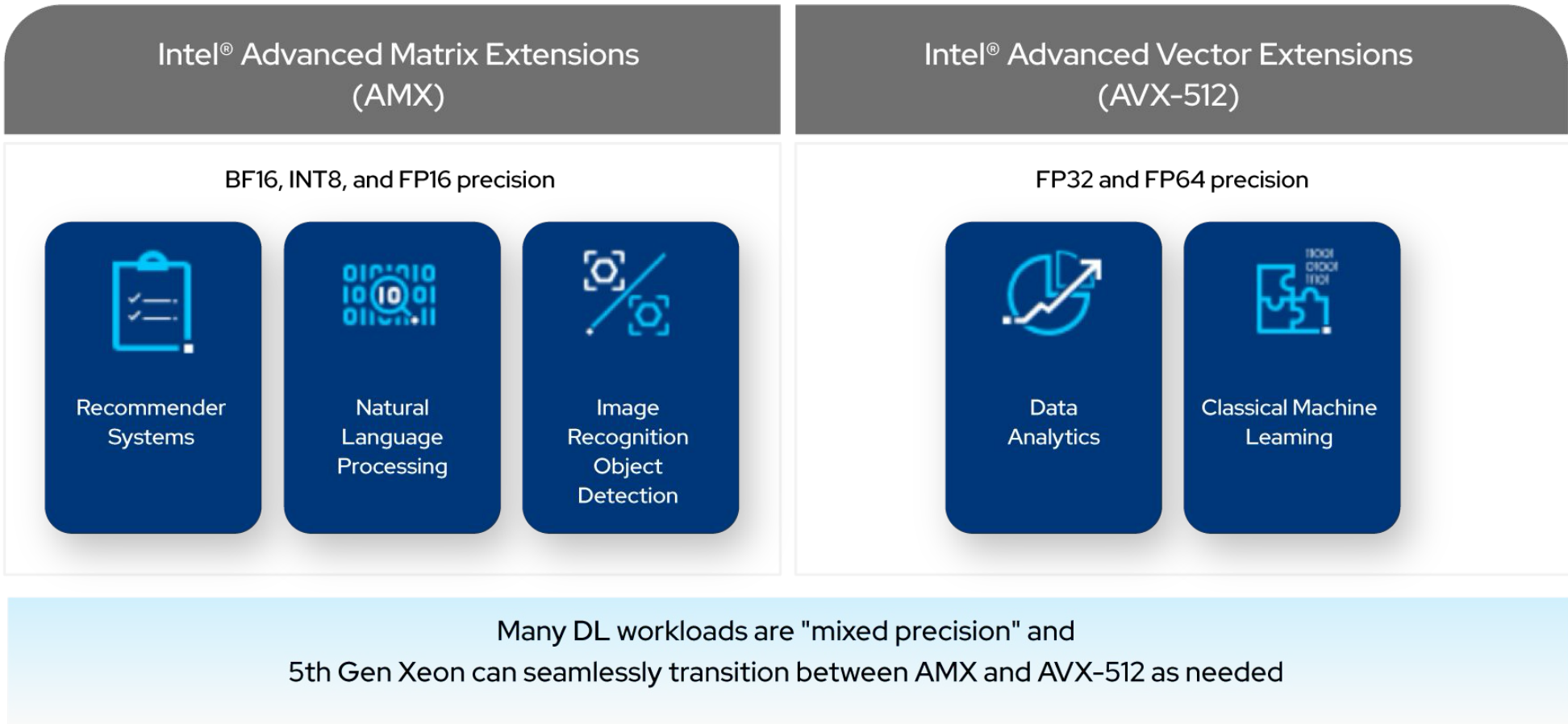


Intel's AI Strategy



- 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

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



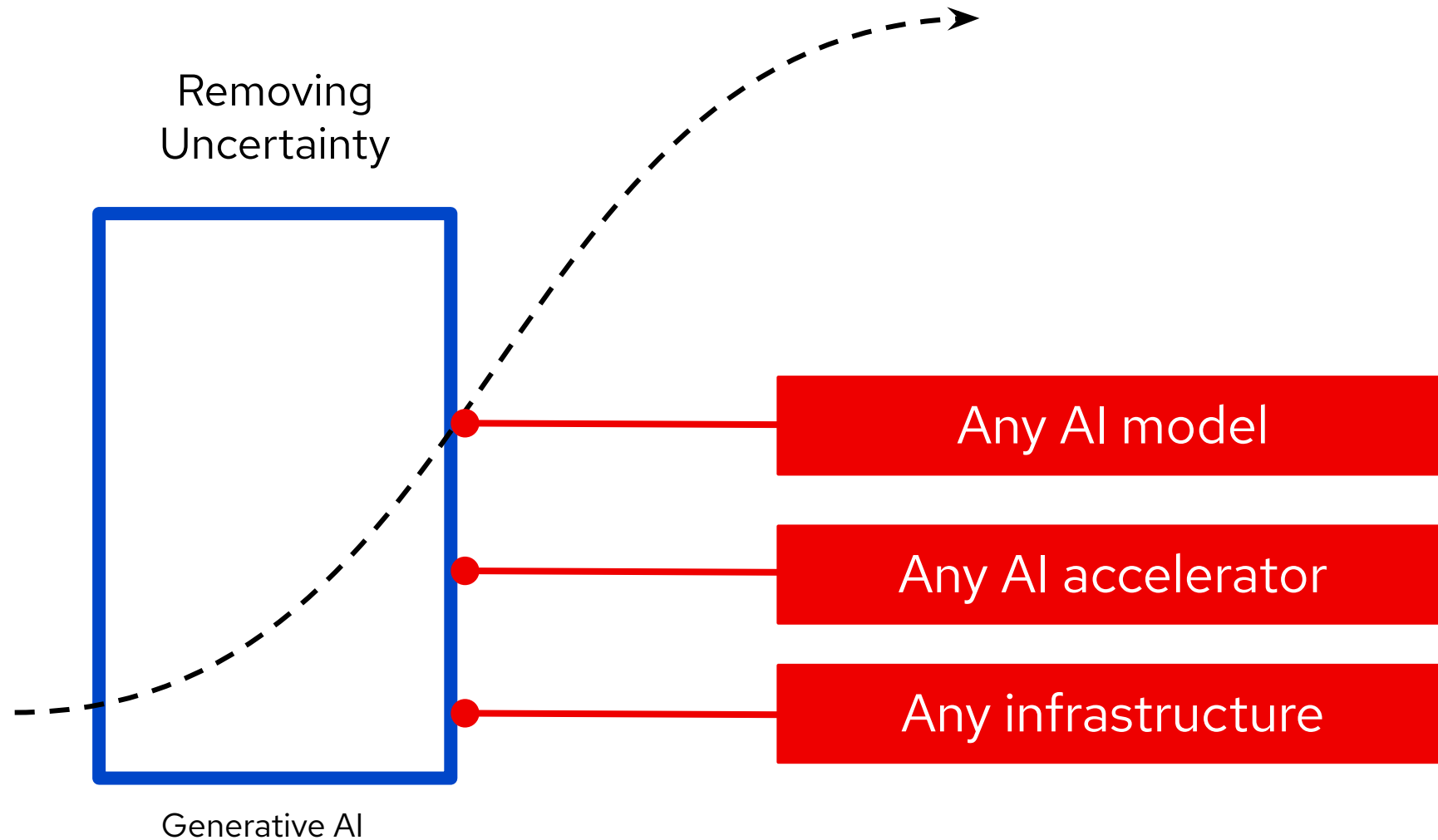
AI Gold Deck

Public

intel ai

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





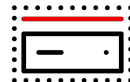
Trusted, Consistent and Comprehensive foundation



Hardware Acceleration



Physical



Virtual



Private
Cloud



Public
Cloud

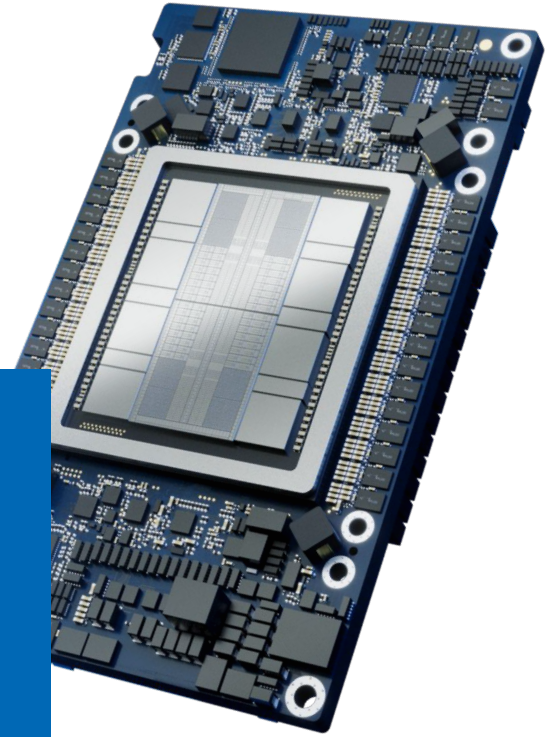


Edge

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)

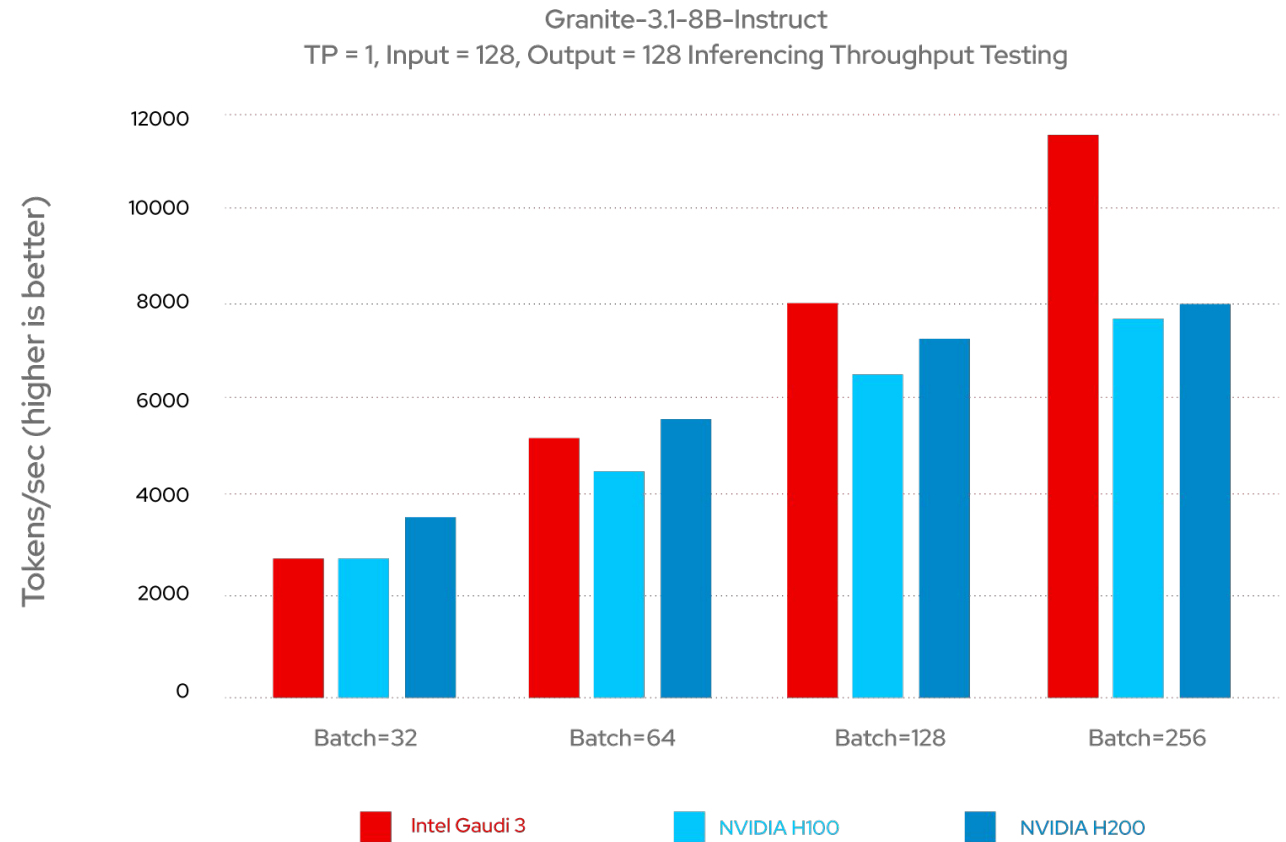
on Llama-3.1-405B-Instruct-FP8
vs. leading GPU competitor
with large context sizes



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. 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.

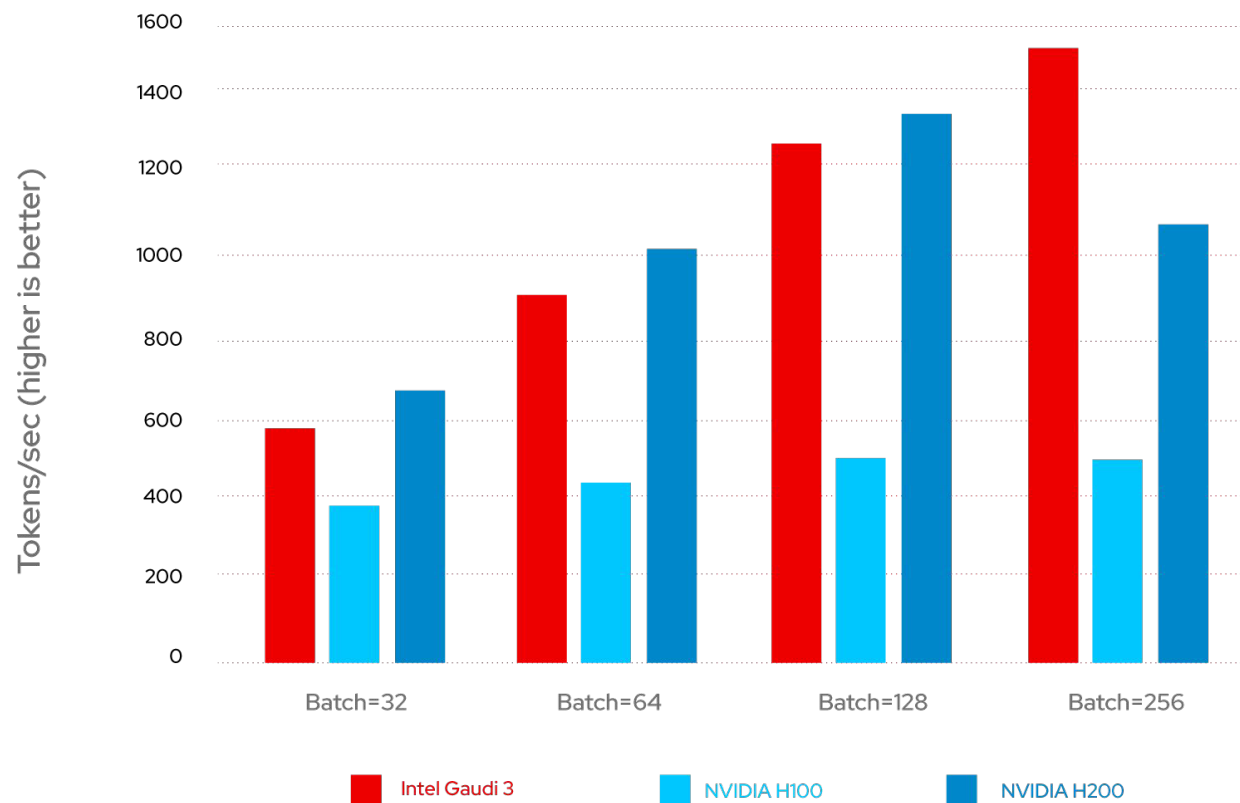
Results may vary.

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 NVIDIA 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 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](https://www.intel.com/processorclaims): Intel Xeon 6. Results may vary

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

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

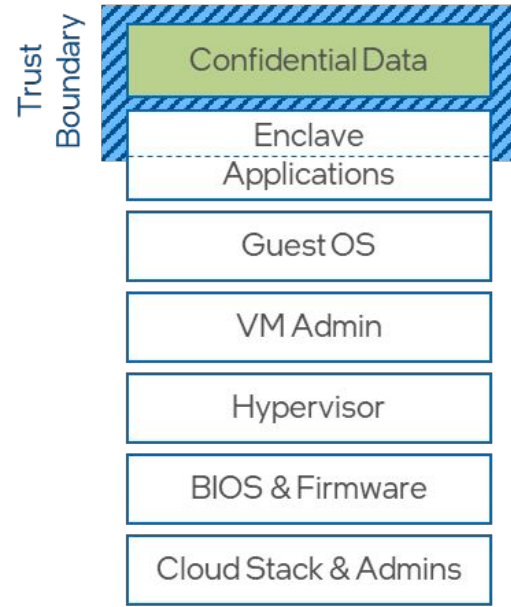
Post Quantum

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

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

App Isolation

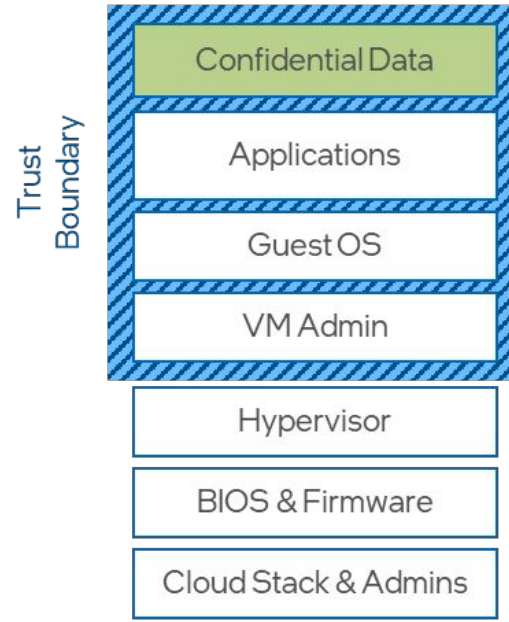
Intel® SGX



Smallest trust boundary for greatest data protection & code integrity

VM Isolation

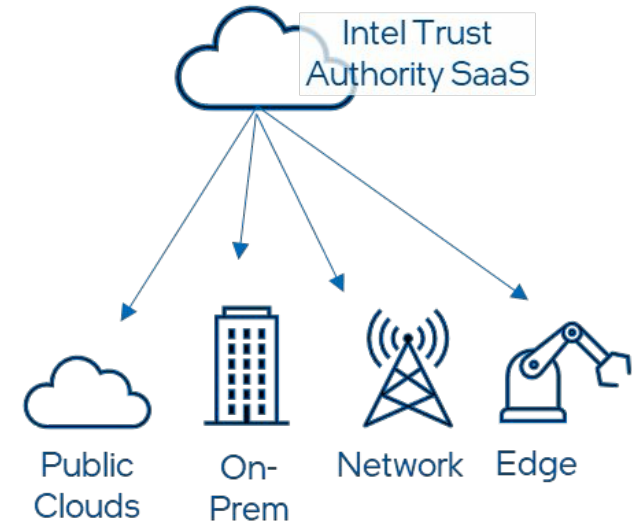
Intel® TDX



Most straightforward path to greater security for legacy apps

Trust Services

Intel® Tiber™ Trust Authority



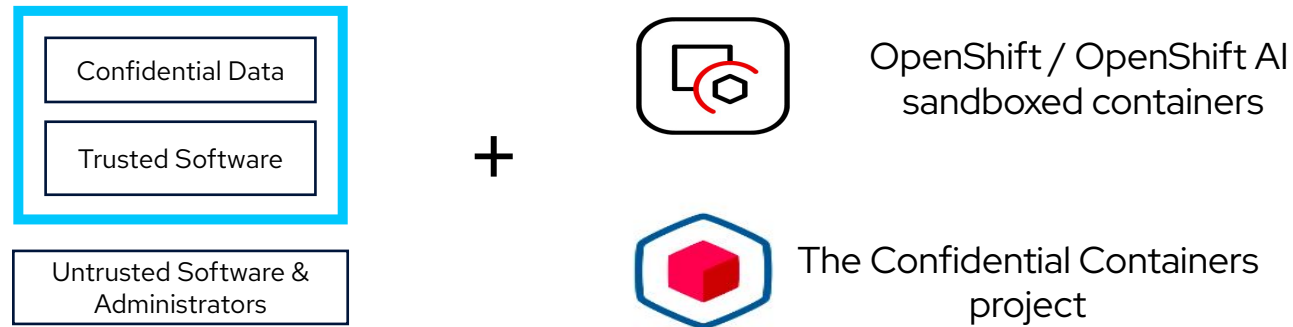
Uniform, independent attestation of trustworthy environments

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

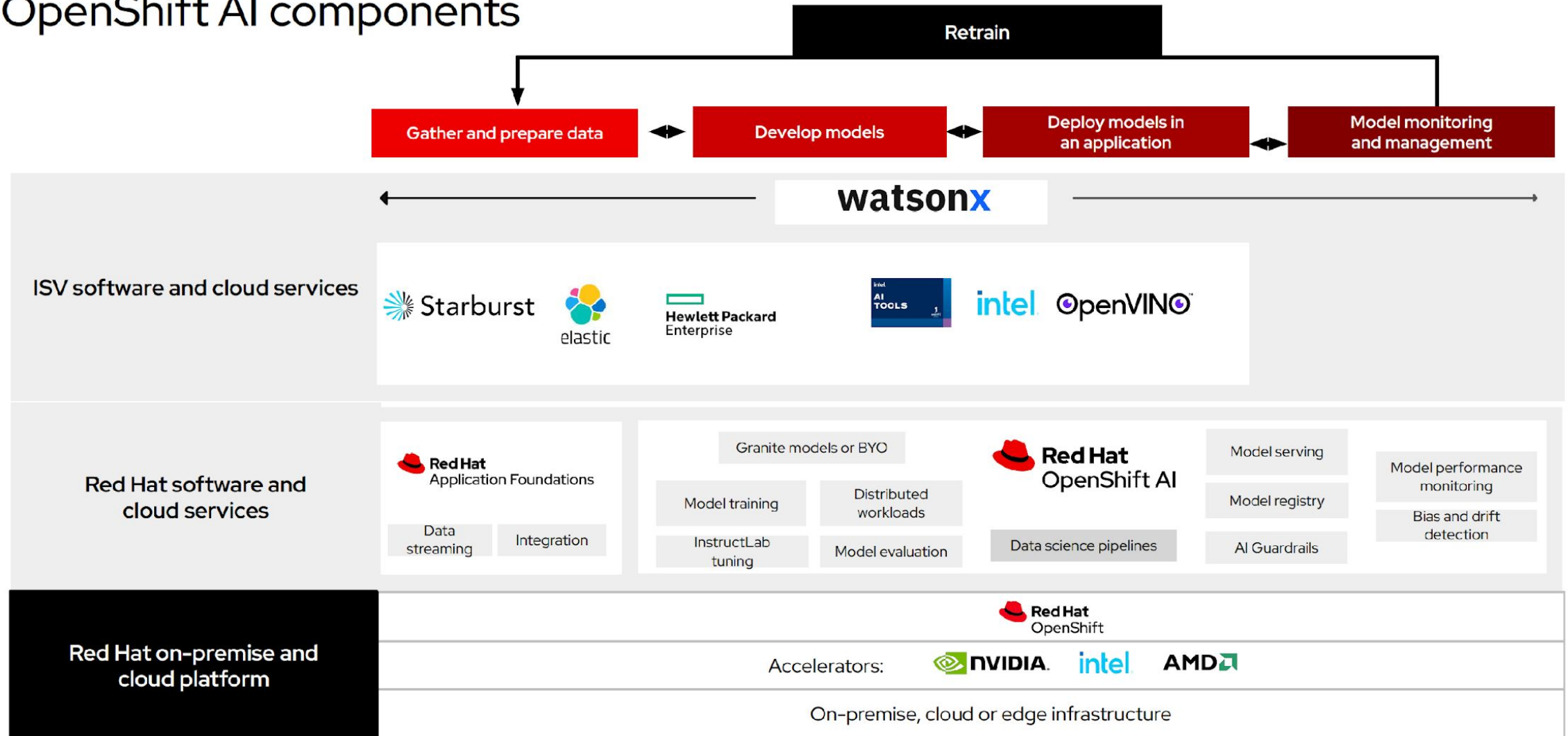
Hardware-Based Protection of Data In-Use
With Intel Trusted Domain Extensions (TDX)



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

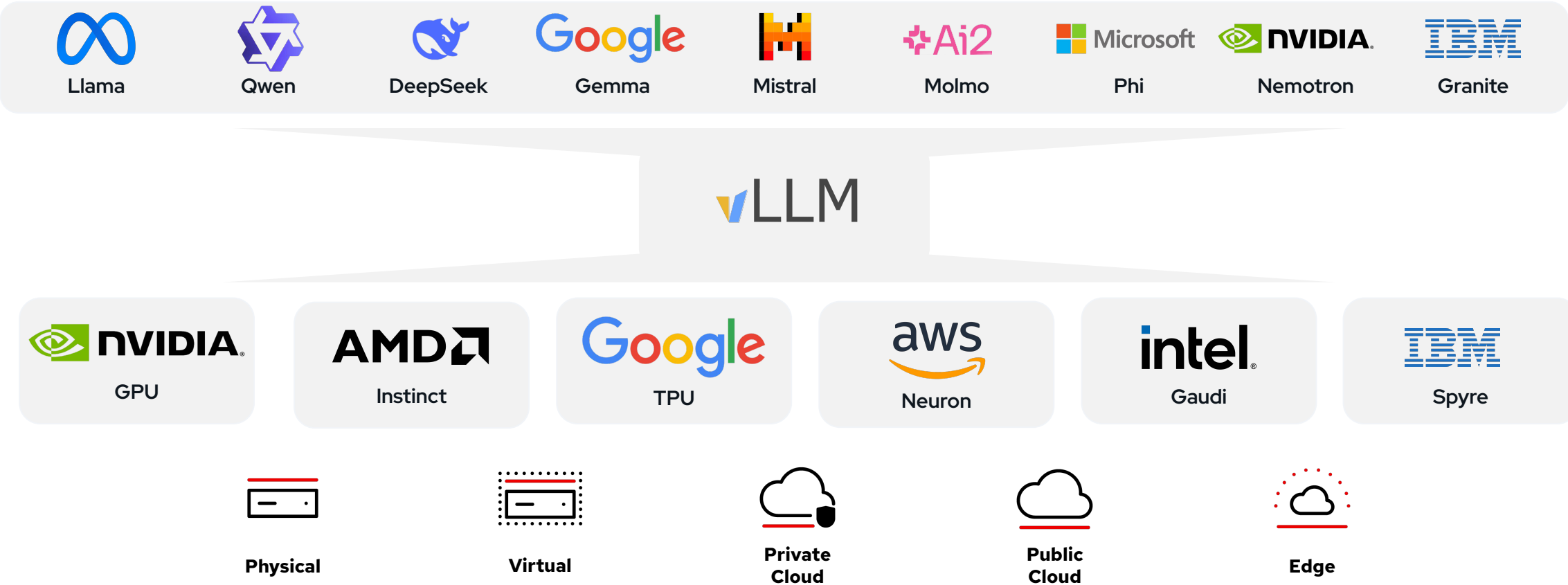
Red Hat AI Platform

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



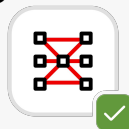
Granite



NVIDIA

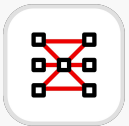
Nemotron

Validated models



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

Optimized models

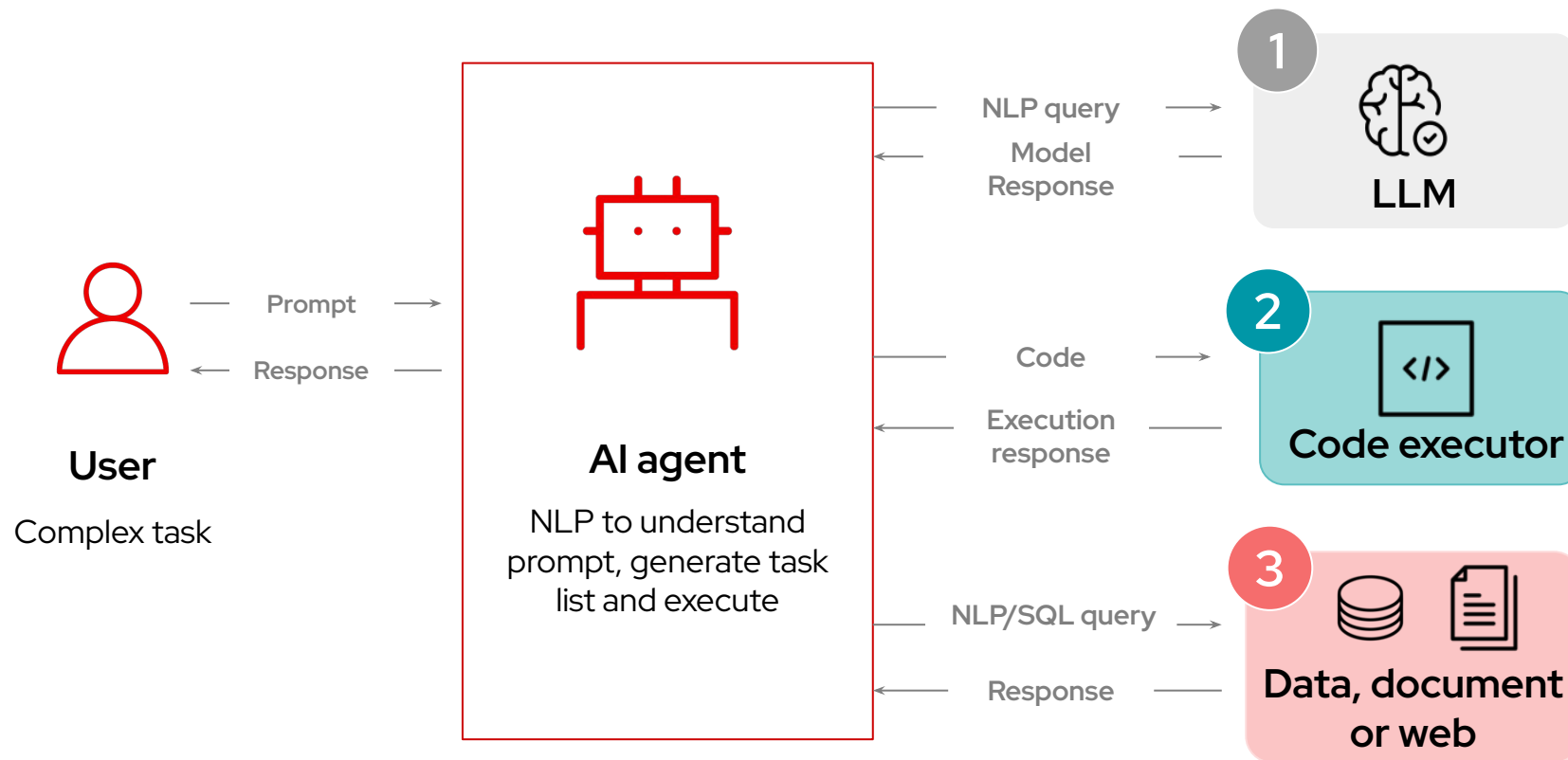


- ▶ Compressed for speed and efficiency
- ▶ Designed to run faster, use fewer resources, maintain accuracy
- ▶ Done using LLM Compressor with latest algorithms

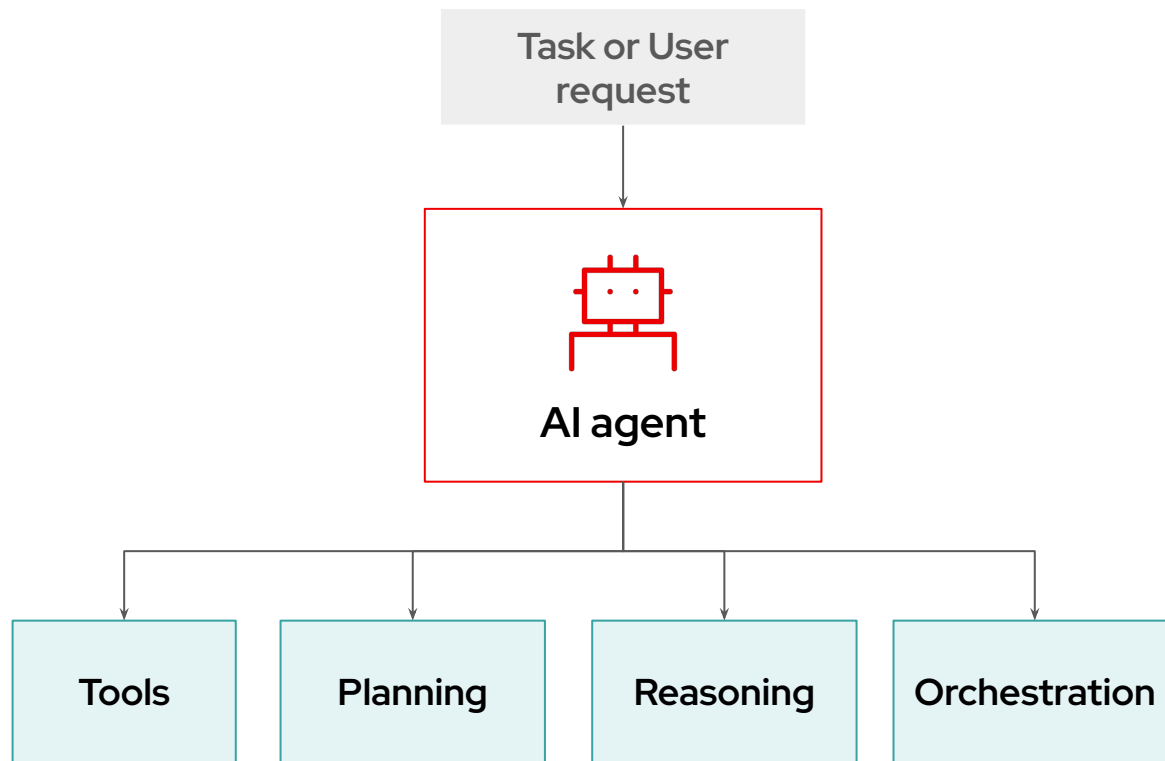
Intro to Agentic AI

AI agents integrate models, functions & tools

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



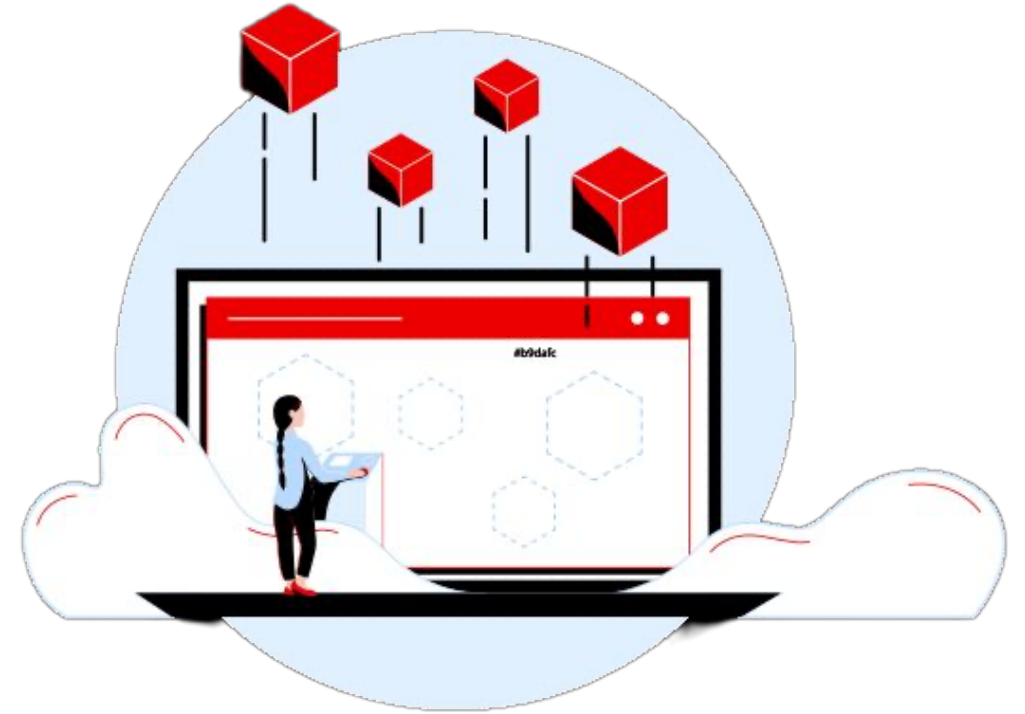
The components of an AI Agent system



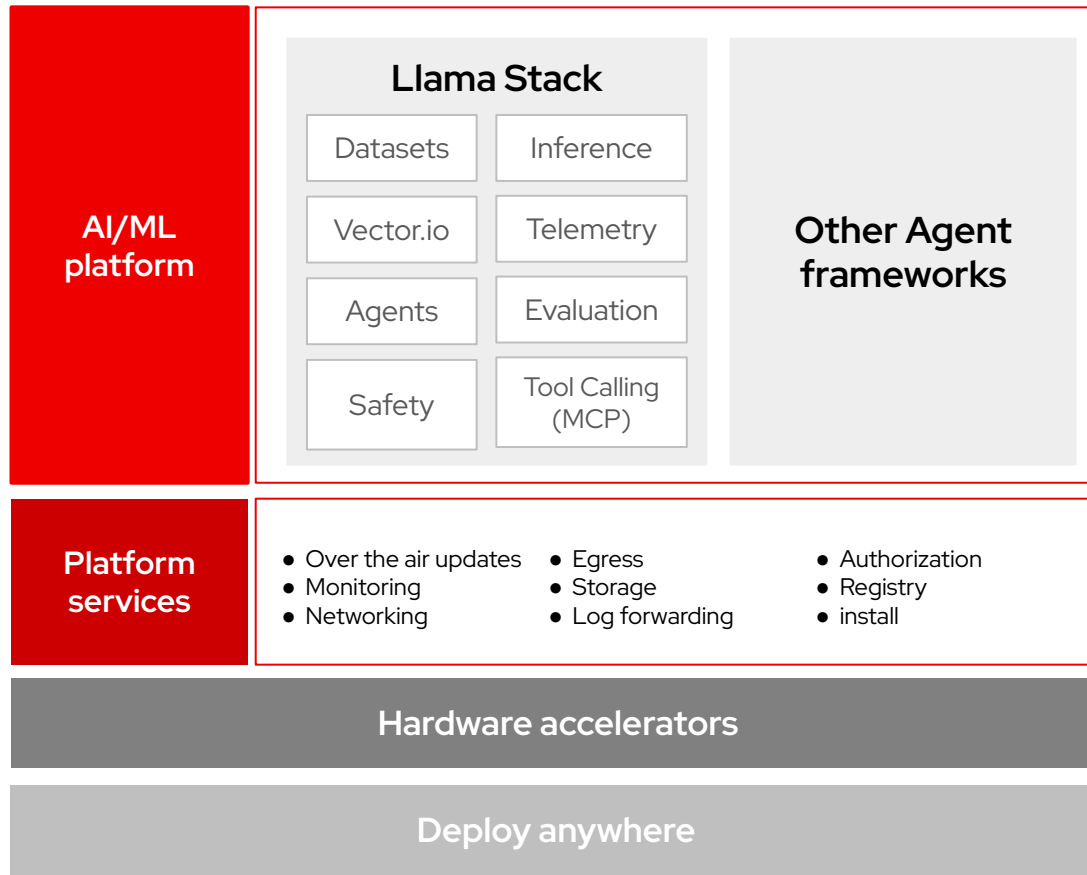
- ▶ **Tool Utilization:** Leverages external tools to gather data and perform tasks.
- ▶ **Planning and Execution:** Develops and executes multistep plans to achieve goals autonomously.
- ▶ **Reasoning:** Applies logic and contextual understanding to make informed decisions.
- ▶ **Orchestration:** Coordinates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

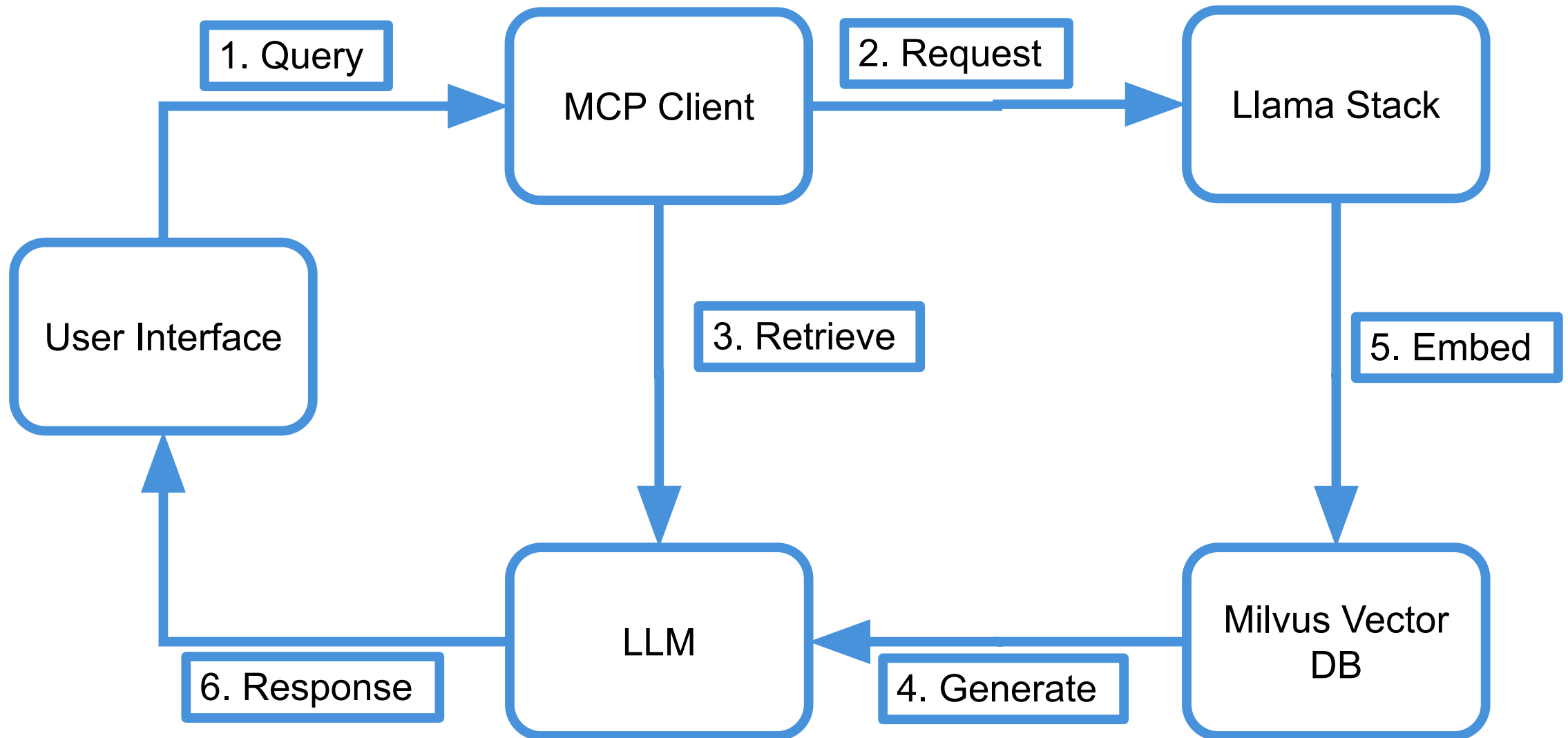


Red Hat AI allows to:

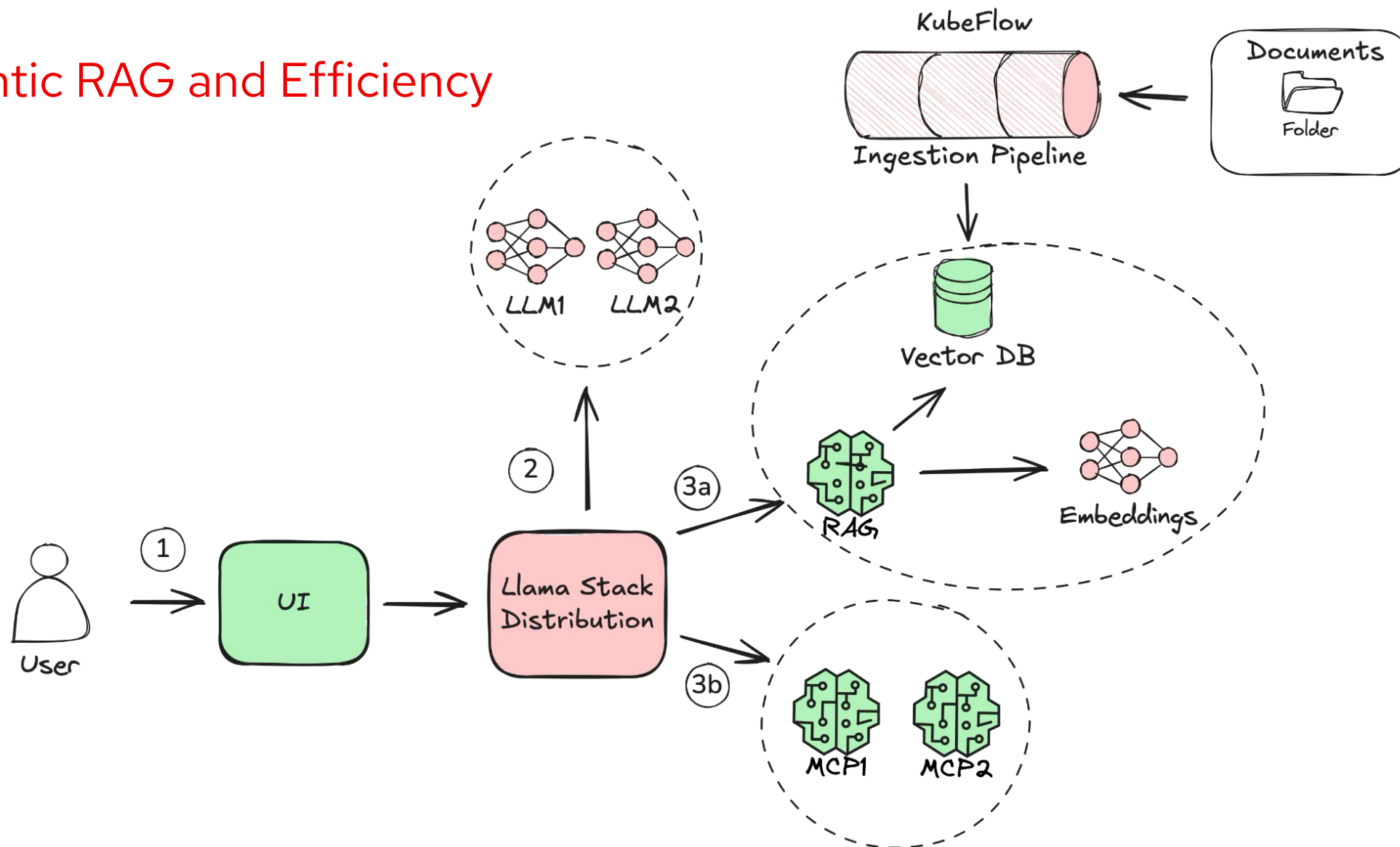
- ▶ Build agents using **Llama Stack's native capabilities and implementations**.
- ▶ **Bring compatible Llama Stack implementations** to OpenShift AI.
- ▶ **Use your own agent framework** and selectively incorporate Llama Stack APIs.
- ▶ **Build with Core Primitives** and manage your own agent framework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

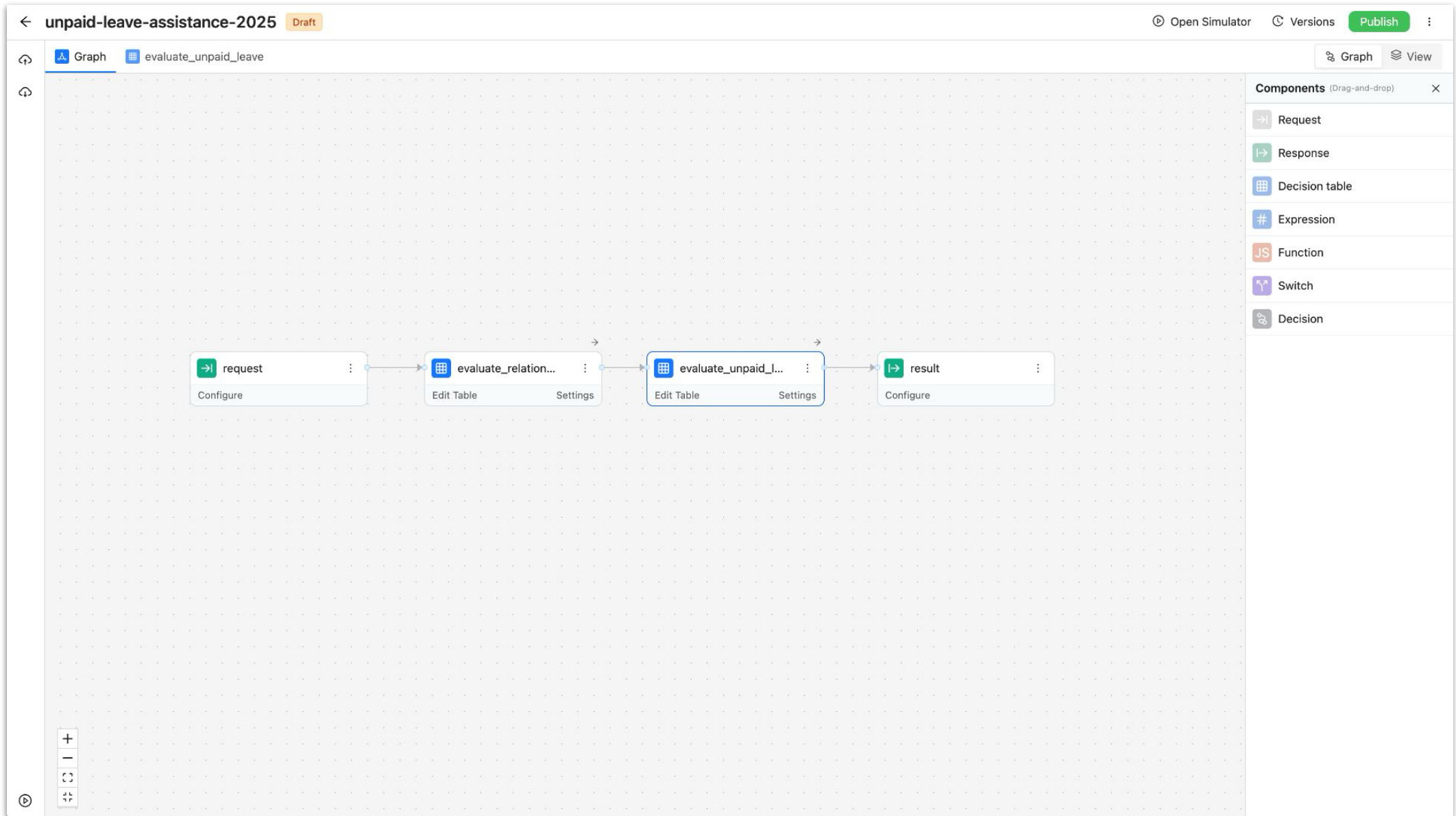
Agentic AI Demo

Agentic AI Demo Architecture



Agentic RAG and Efficiency





unpaid-leave-assistance-2025

Draft

Close Simulator

Versions

Publish

Graph

evaluate_unpaid_leave

evaluate_relationship

Export Excel

Import Excel

	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
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_c are"			true	500	"C"	"Adoption or foster care"	"In the case th must be"

Mother at the hospital

Search nodes...

Graph

evaluate_relationship

evaluate_unpaid_leave

result

Output

Input

Trace

```

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

```

```

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 }

```

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



facebook.com/redhatinc



youtube.com/user/RedHatVideos



twitter.com/RedHat