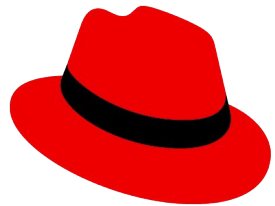


# Enhancing Enterprise AI with RAG: Boost your AI's intelligence by seamlessly merging real-time data with LLMs

**Red Hat Summit Connect 2024 Zurich**

*Zurich, 15 January 2025*

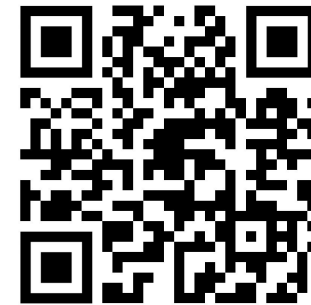


**Red Hat**

# Codrin Bucur

Principal AI Specialist Solution Architect,  
EMEA

Red Hat





# Hind Azegrouz

AI Inference Lead, EMEA

Intel



Over **25** Years of Collaboration



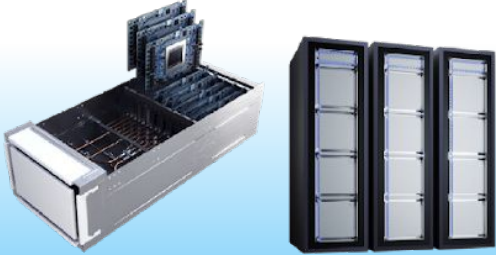
# Bringing AI Everywhere

## Intel's AI Strategy



AI PC Node  
AI Developer Productivity & Light Inference

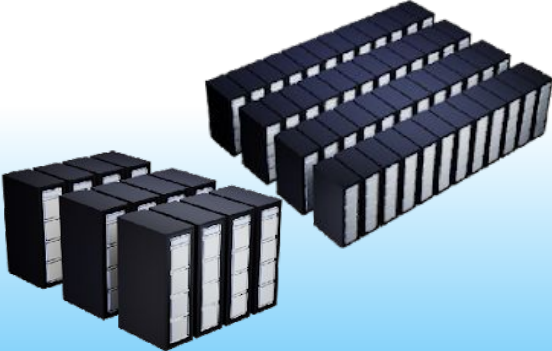
AI PC  
Broadest AI SW Ecosystem



Node  
Fine-tuning, Inference

Cluster  
Light Training, Tuning, Peak Inference

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



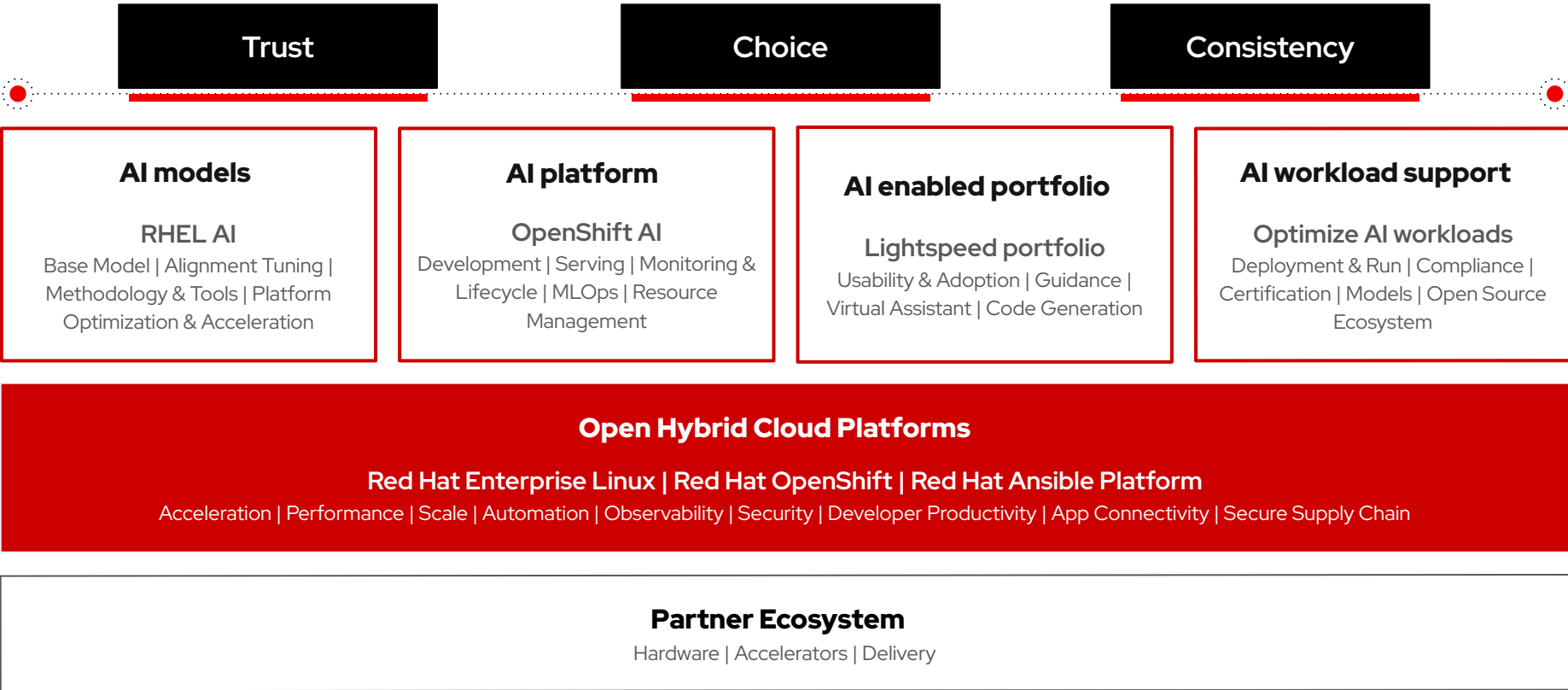
Super Cluster  
Training, Tuning, Peak Inference

Mega Cluster  
Large Scale Training & Inference

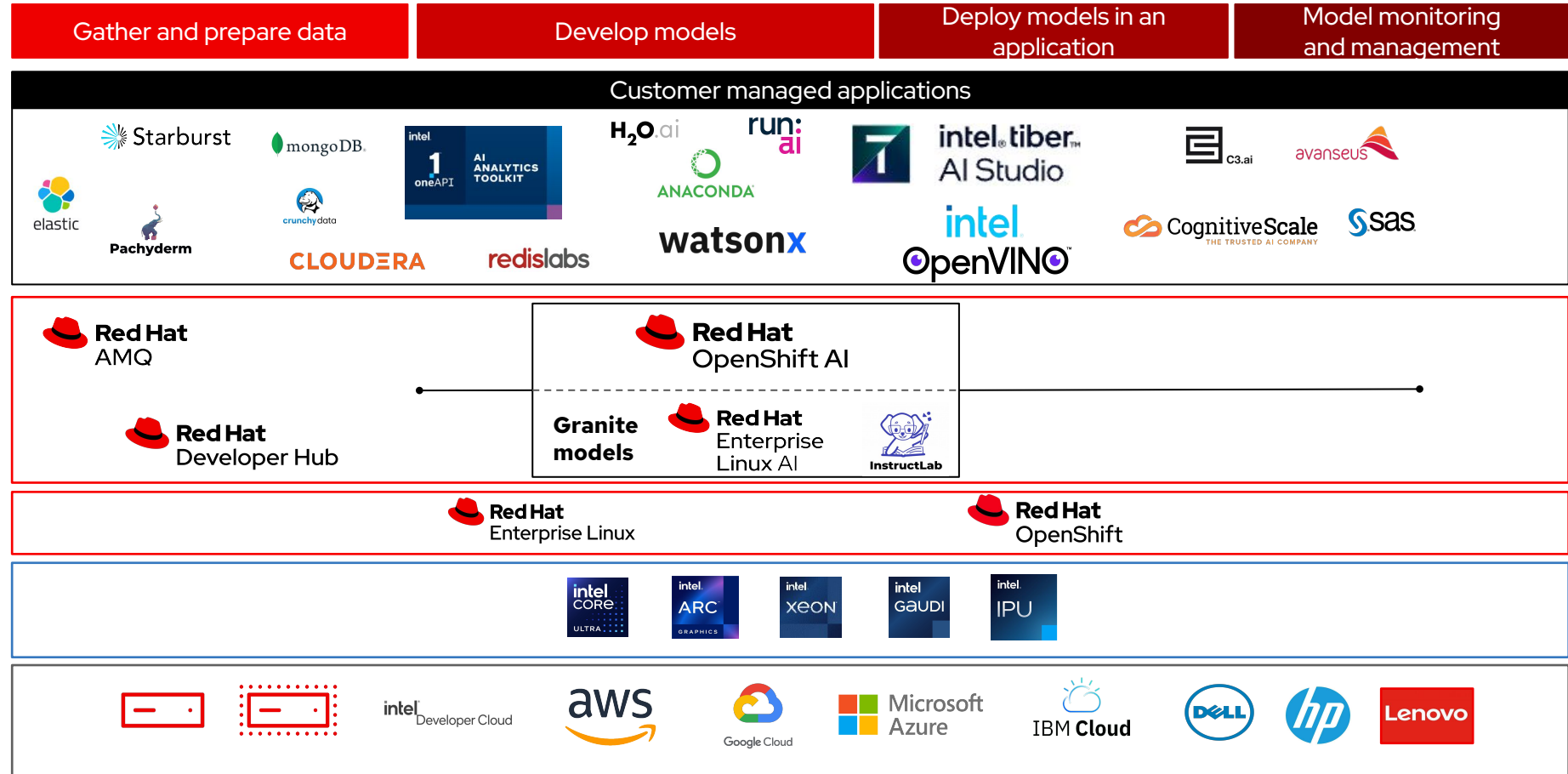
DATA CENTER AI  
AI Open, Scalable Systems & Reference Arch



# Red Hat's AI Strategy



# Intel Enterprise AI with Red Hat® OpenShift® AI



# OPEA – Open Platform for Enterprise AI



# OPEA - Open Platform for Enterprise AI

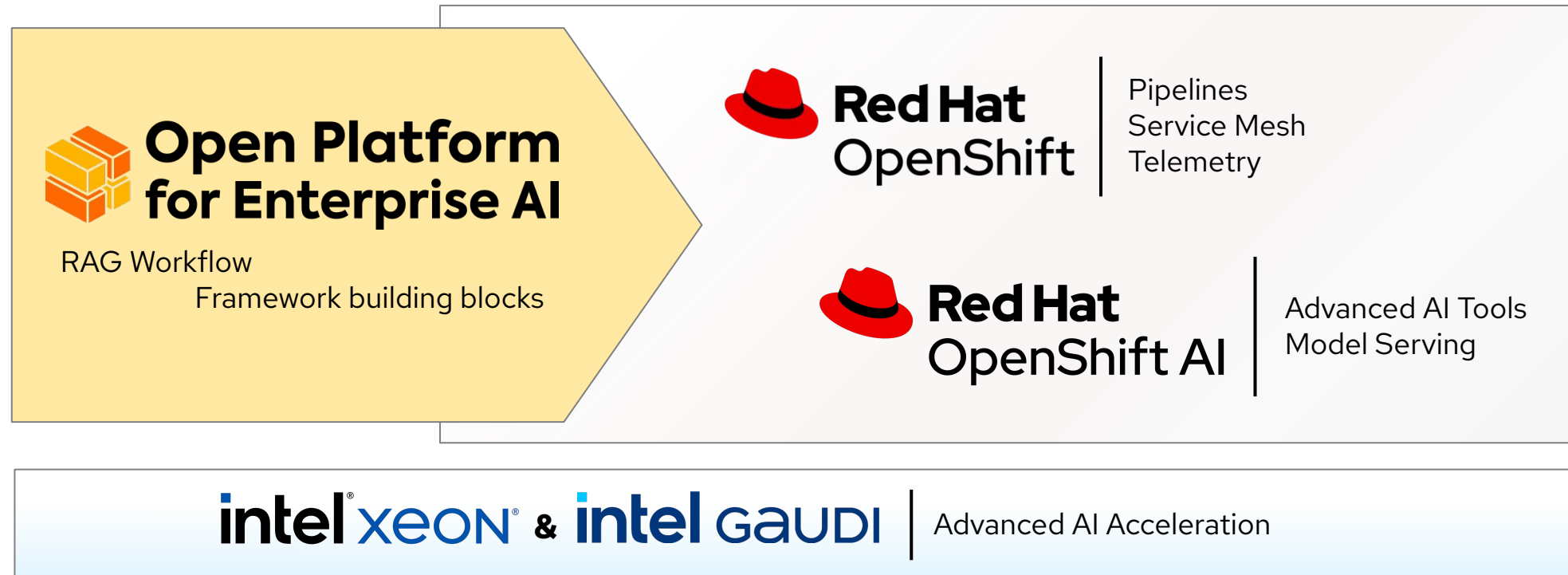
By The Linux Foundation

- ▶ Ecosystem orchestration framework for GenAI
- ▶ OPEA.dev
- ▶ GitHub: <https://github.com/opea-project>
- ▶ Contributors:



# OPEA with OpenShift AI

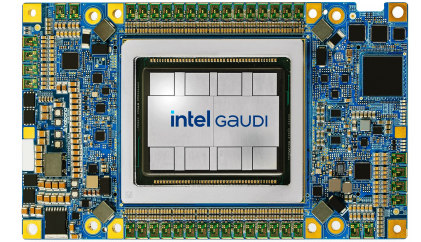
OpenShift AI makes OPEA more enterprise ready



# Intel Gaudi AI Accelerators

# Introducing the Intel® Gaudi® 3 Accelerator

Breaking benchmarks, not budgets



## Competitive Gen AI Performance over H100

- Projected **50% faster time to train**<sup>1</sup>
- Projected **50% faster inferencing**<sup>2</sup>
- Projected **40% better power efficiency**<sup>3</sup>



## Freedom to Scale without Lock-in

- Open standard ethernet networking vs proprietary InfiniBand
- 24x200 GbE ports of industry-standard RoCE on every Gaudi®<sup>3</sup>
- 33% more I/O peak throughput vs H100 for massive scale-up within the server<sup>4</sup>



## Open Development on GenAI platforms

- Integrated open-source PyTorch framework with optimized model library on Hugging Face
- Migrate models on open software from H100 with as few as 3 lines of code

<sup>1</sup> NV H100 comparison based on : <https://developer.nvidia.com/deep-learning-performance-training-inference/training>, Mar 28th 2024 -> "Large Language Model" tab.

<sup>2</sup> Source: NV H100 comparison based on <https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8> , Mar 28th, 2024. Reported numbers are per GPU.

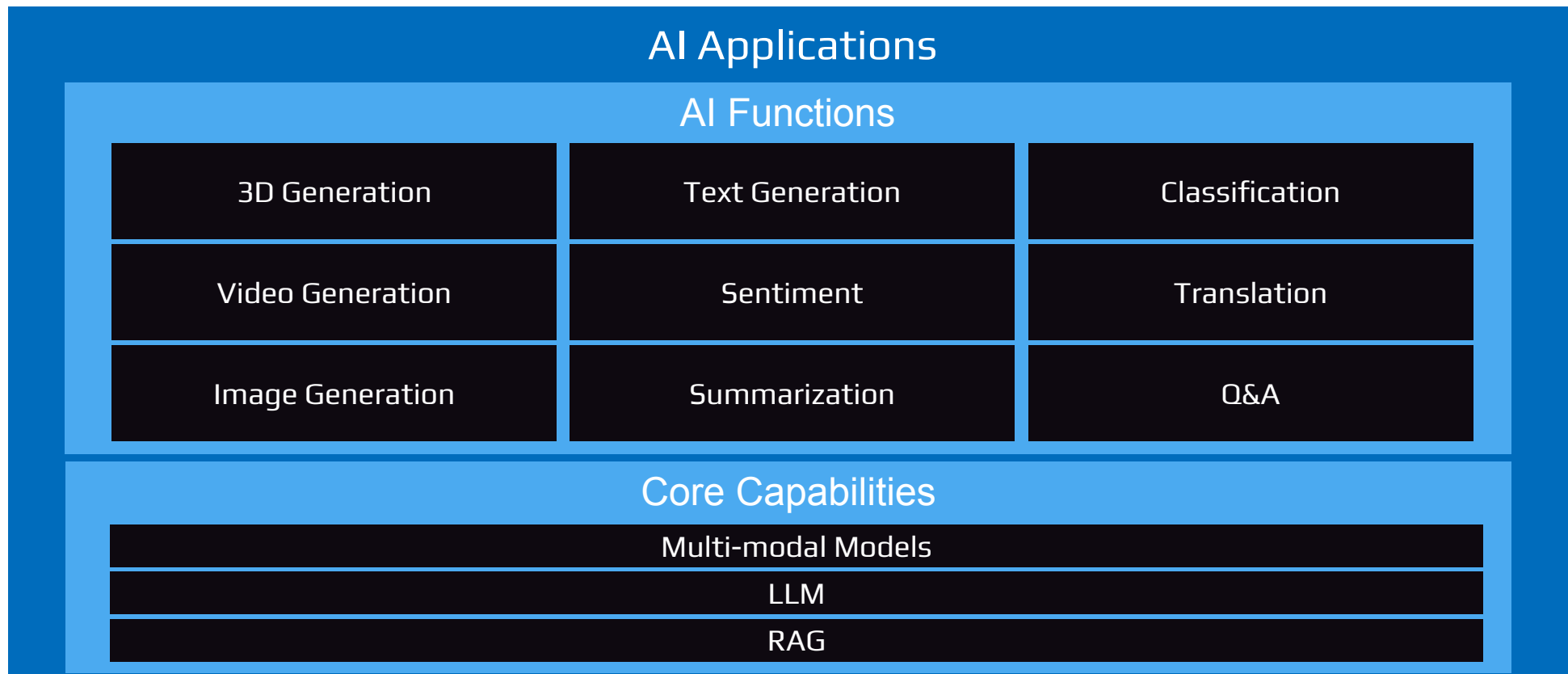
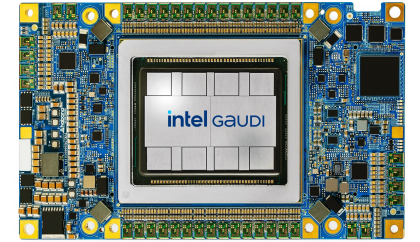
<sup>3</sup> Source: NV comparison based on <https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8> , Mar 28th, 2024. Reported numbers are per GPU.

<sup>1-3</sup> Vs Intel® Gaudi® 3 projections for LLAMA2-7B, LLAMA2-70B & Falcon 180B Power efficiency for both Nvidia and Gaudi3 based on internal estimates. Results may vary.

<sup>4</sup> 900 GB/s NVLink connectivity on H100 vs. 1200 GB/s on Gaudi 3

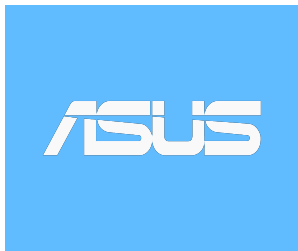
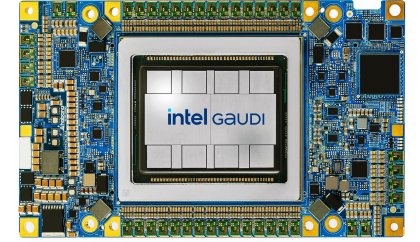
# Intel Gaudi AI Accelerators

Broad Application Support with Focus on Multi-Modal, LLM and RAG



# Intel® Gaudi® 3 AI Accelerator

## Launch Partners



IBM and Intel announce a global collaboration to integrate Intel® Gaudi® 3 accelerators with watsonx on IBM Cloud.



\*See more at:  
<https://newsroom.ibm.com/blog-intel-and-ibm-collaborate-to-provide-better-cost-performance-for-ai-innovation>  
and  
<https://www.intel.com/content/www/us/en/newsroom/news/intel-ibm-deliver-enterprise-ai-in-the-cloud.html>

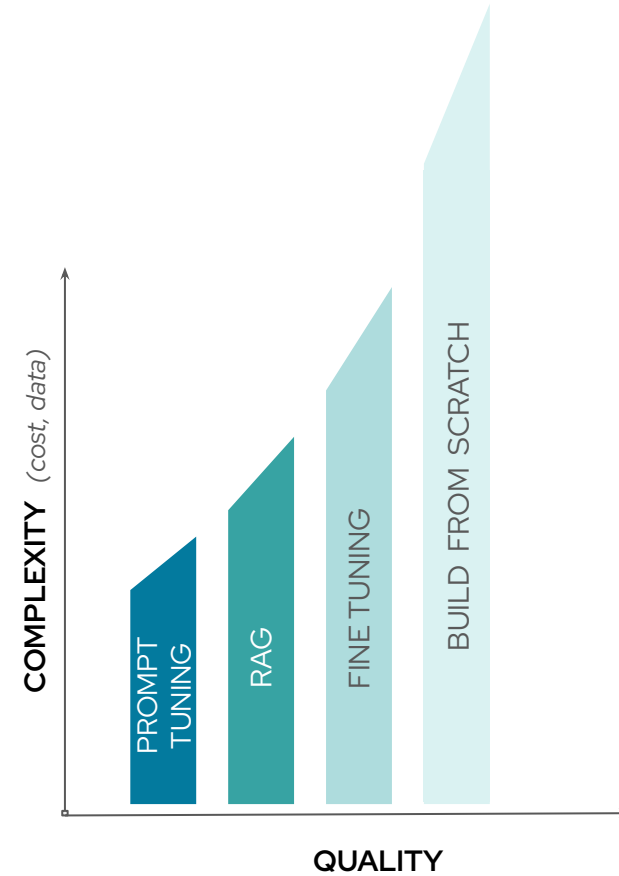


# Retrieval Augmented Generation (RAG) Explained

# The balancing act of using foundation models

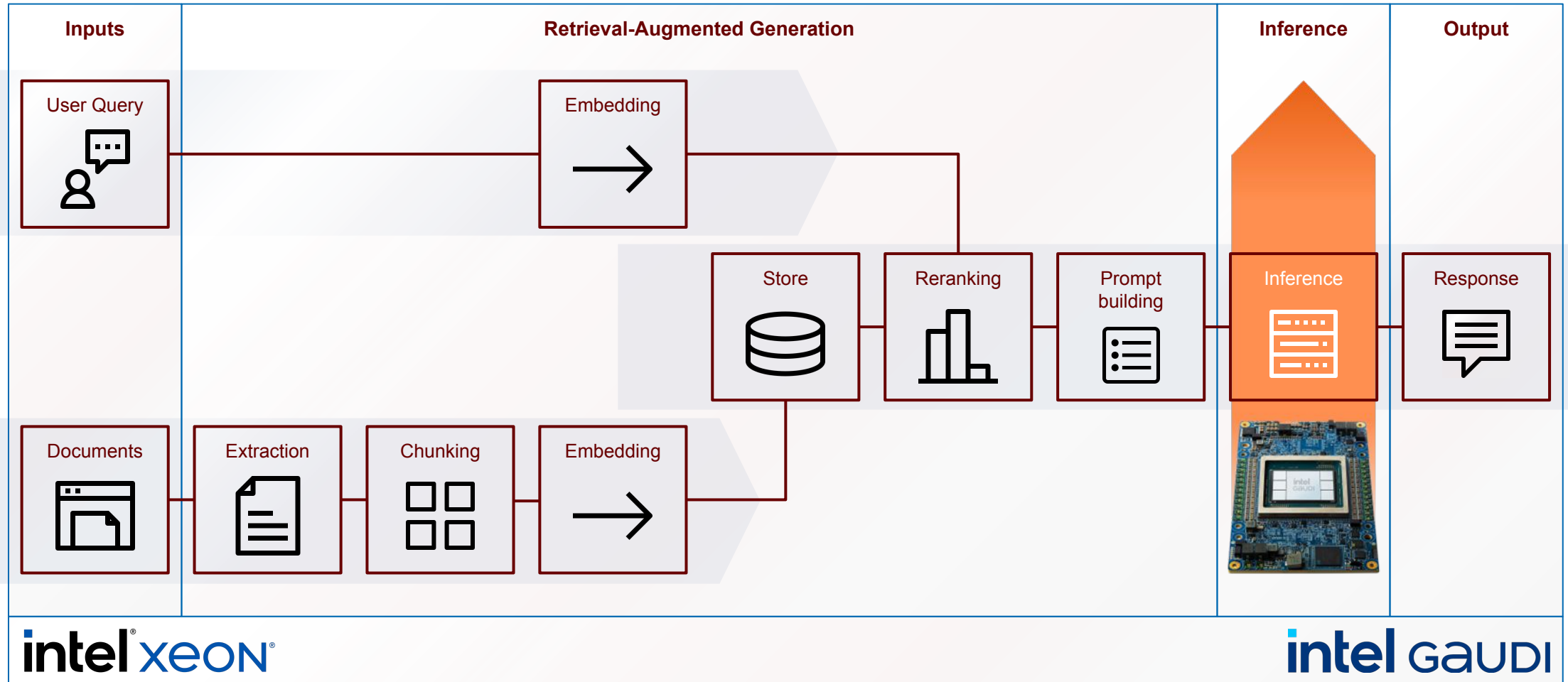
Foundation models will still need more work to be useful

- ▶ Prompt tuning
- ▶ Retrieval-Augmented Generation (RAG)
- ▶ Fine tuning foundation models
- ▶ Training a Foundation Model from scratch

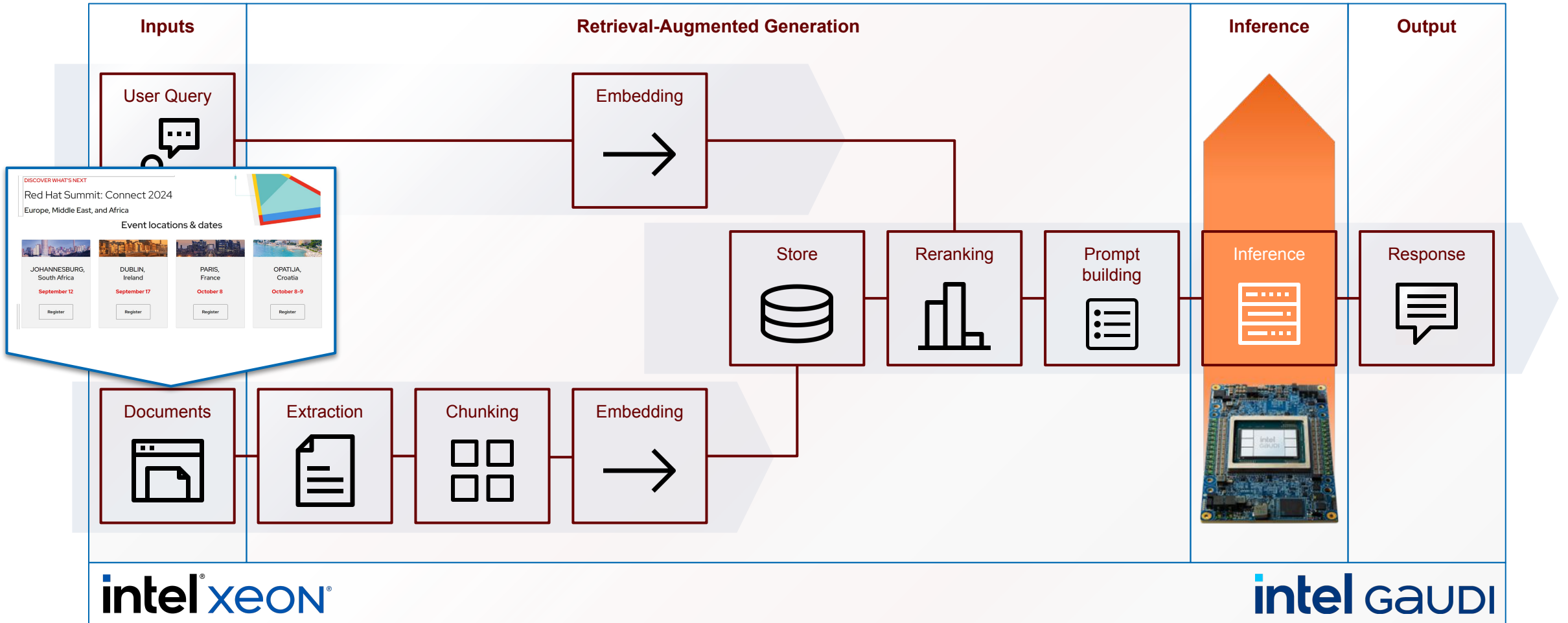




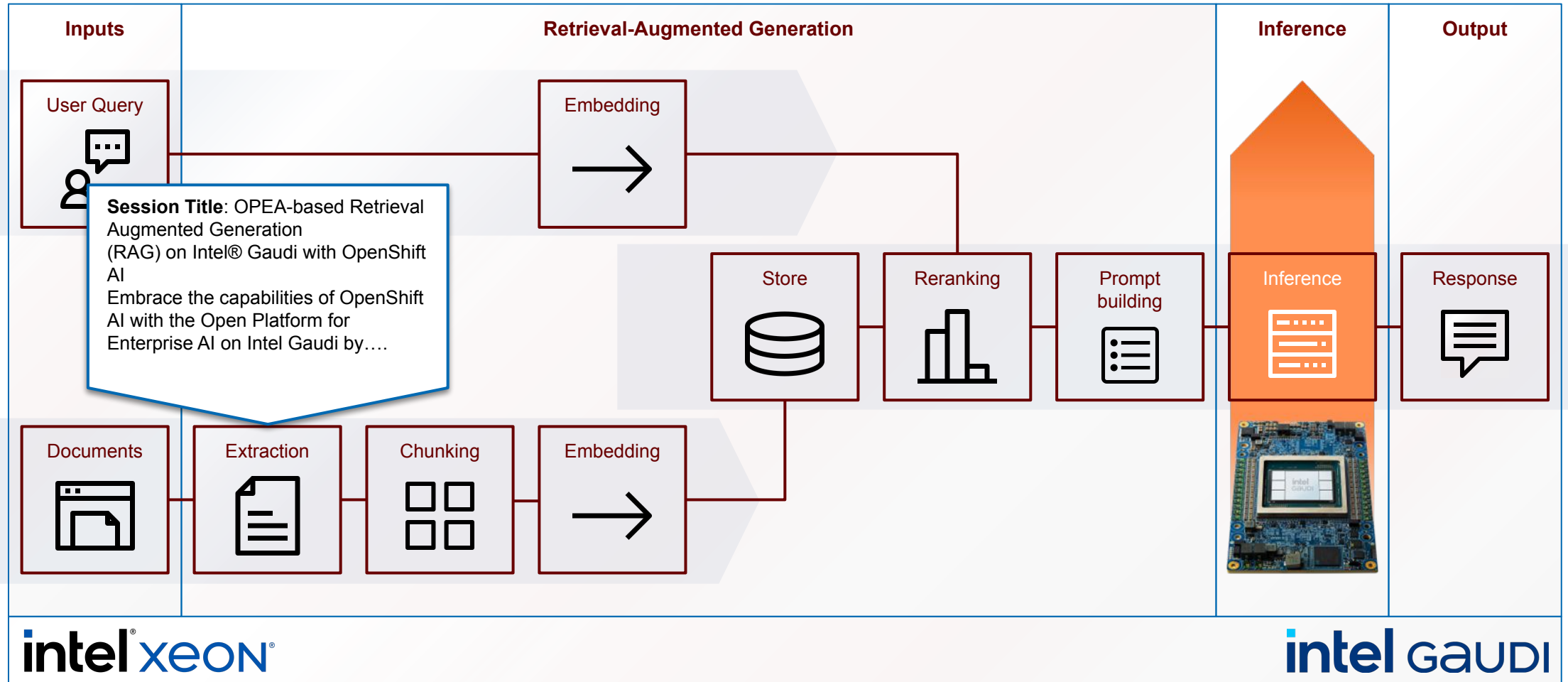
# Retrieval Augmented Generation (RAG)



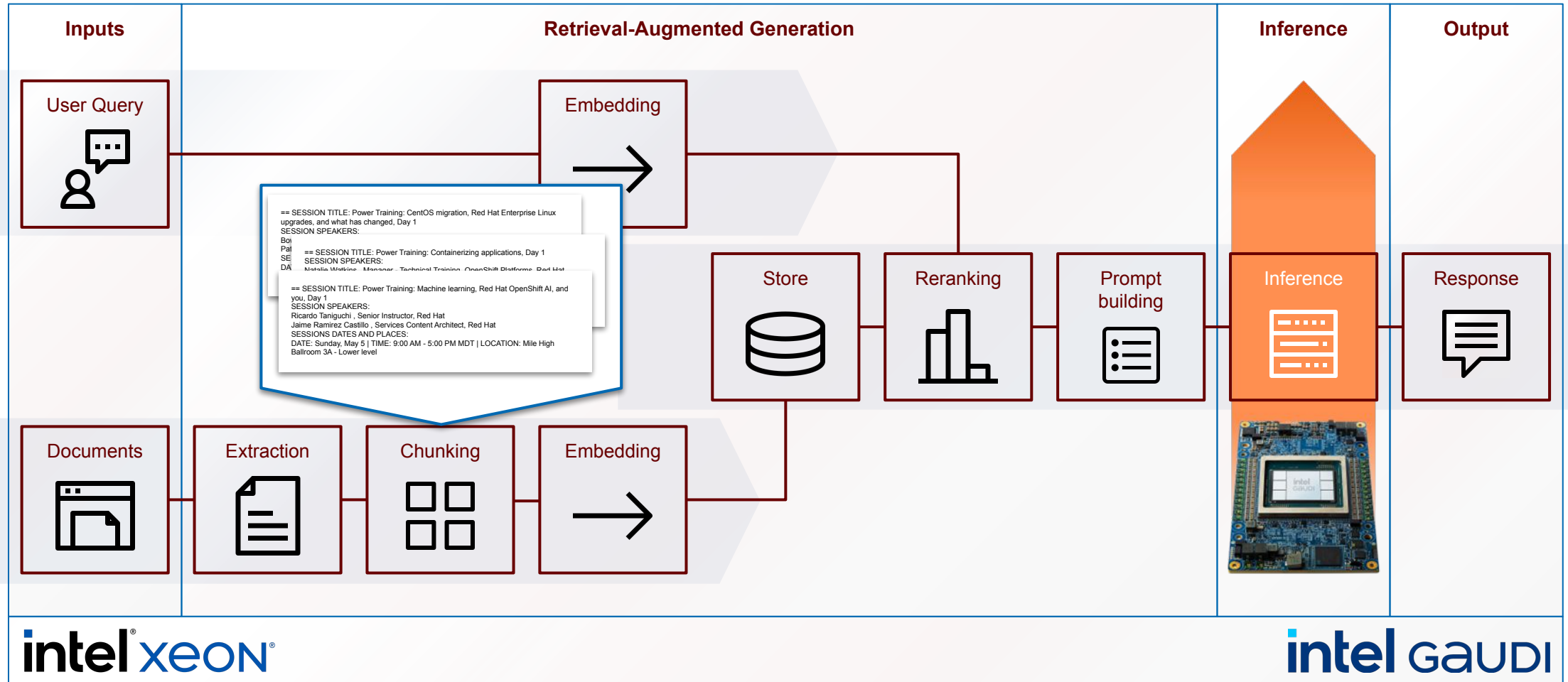
# Retrieval Augmented Generation (RAG)



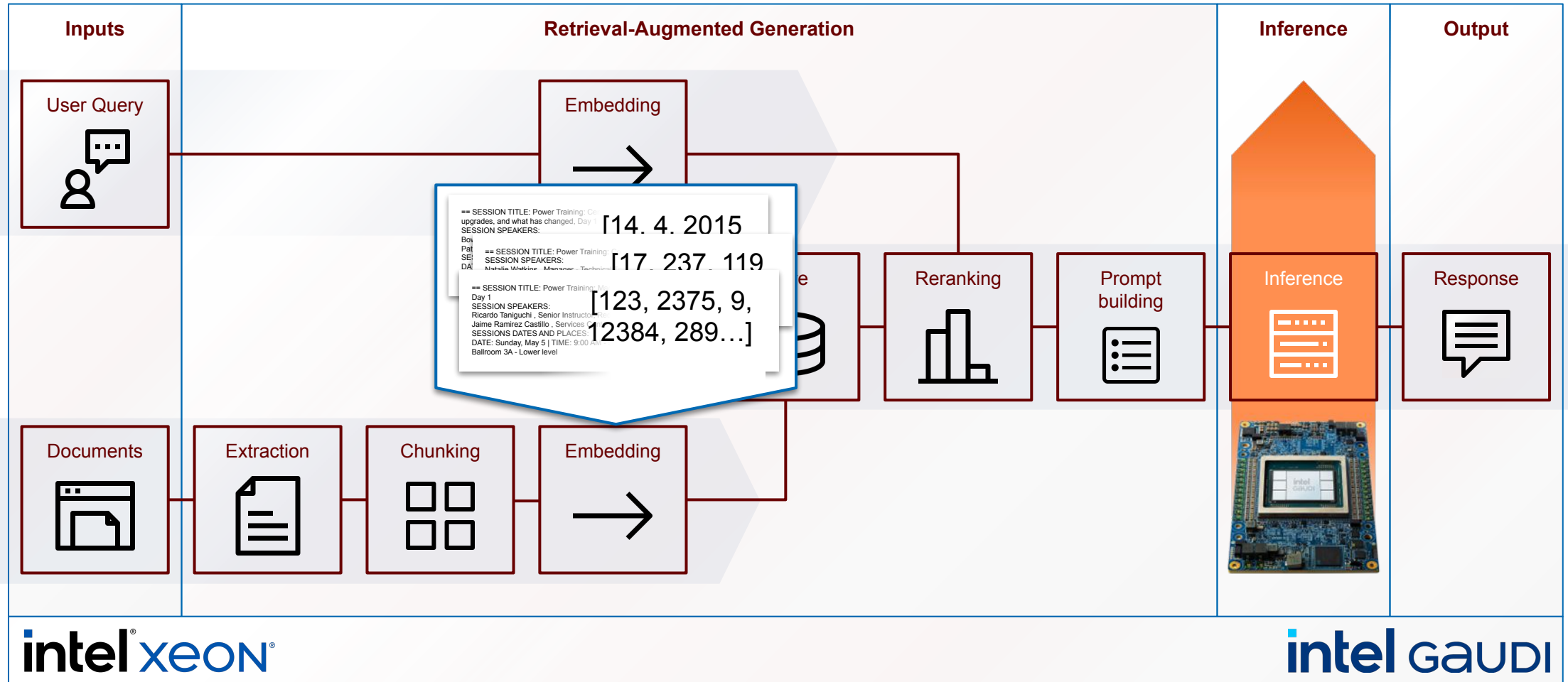
# Retrieval Augmented Generation (RAG)



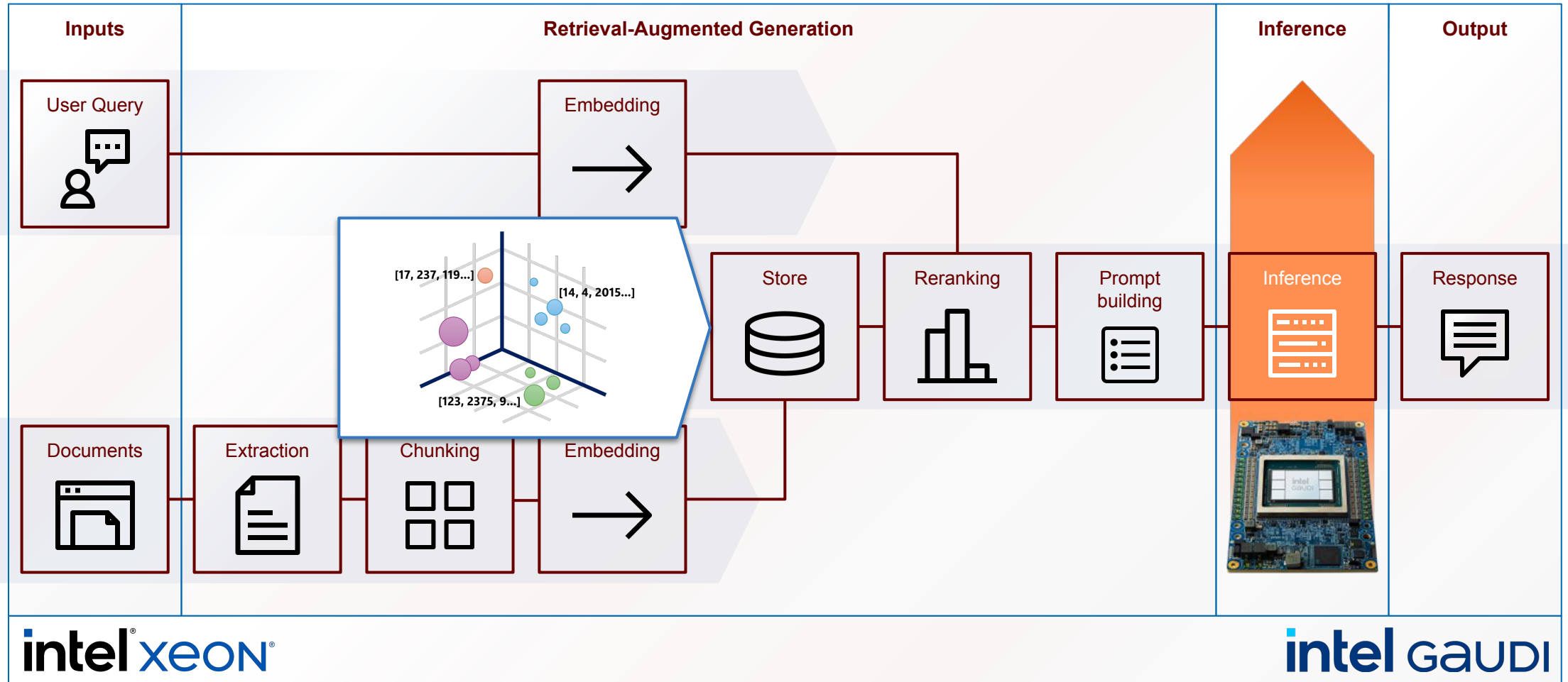
# Retrieval Augmented Generation (RAG)



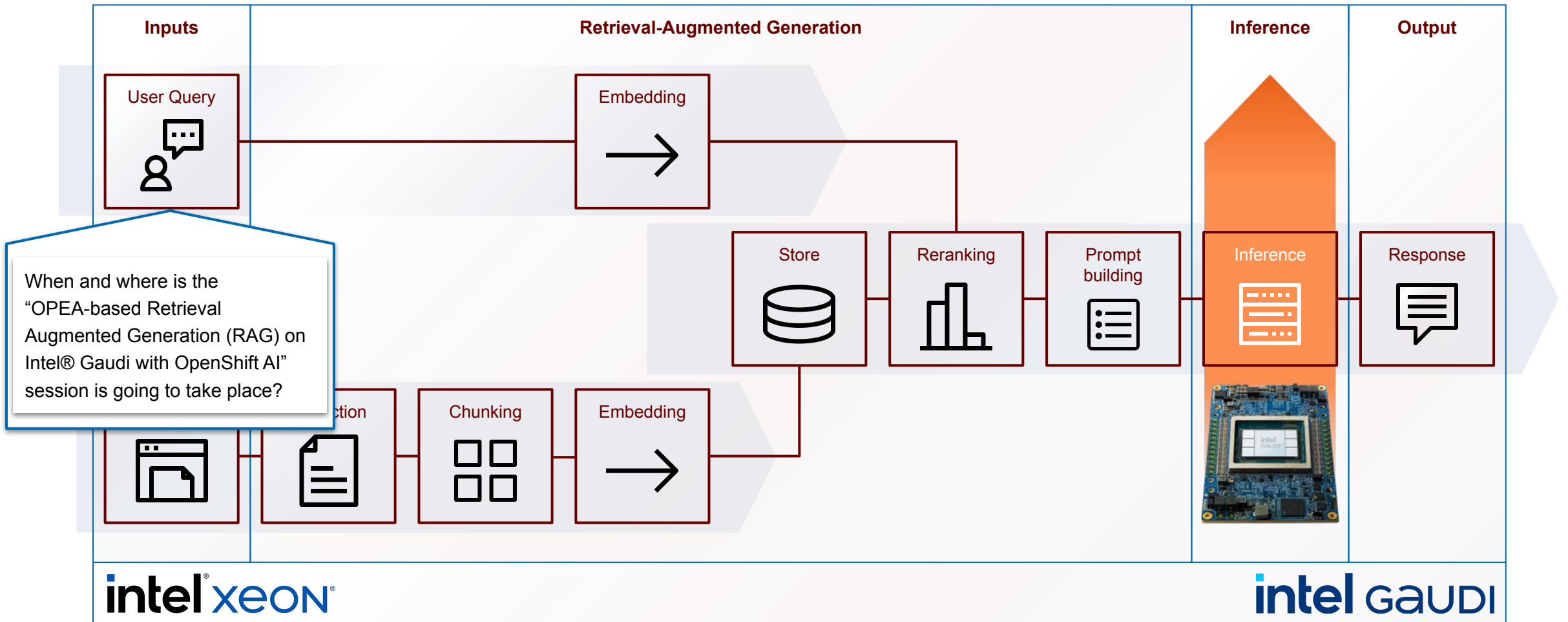
# Retrieval Augmented Generation (RAG)



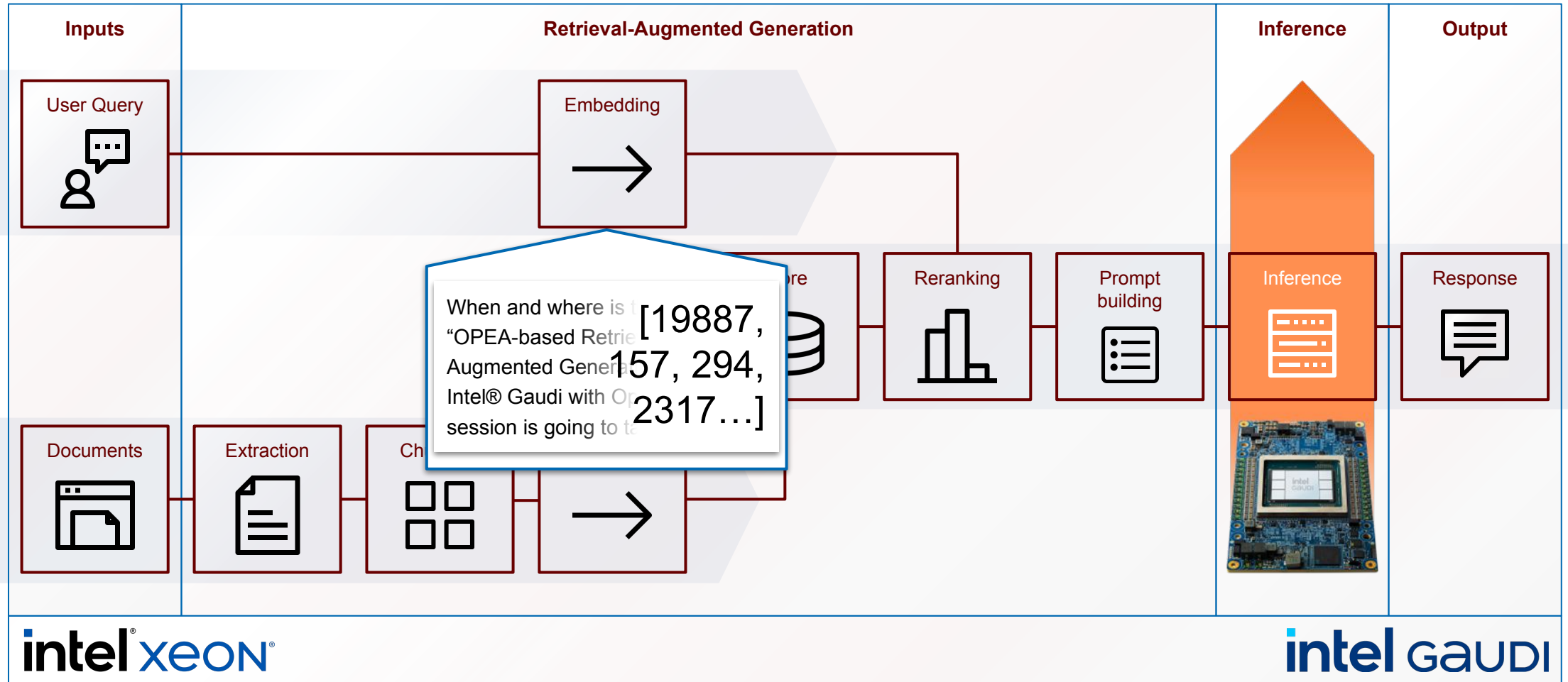
# Retrieval Augmented Generation (RAG)



# Retrieval Augmented Generation (RAG)

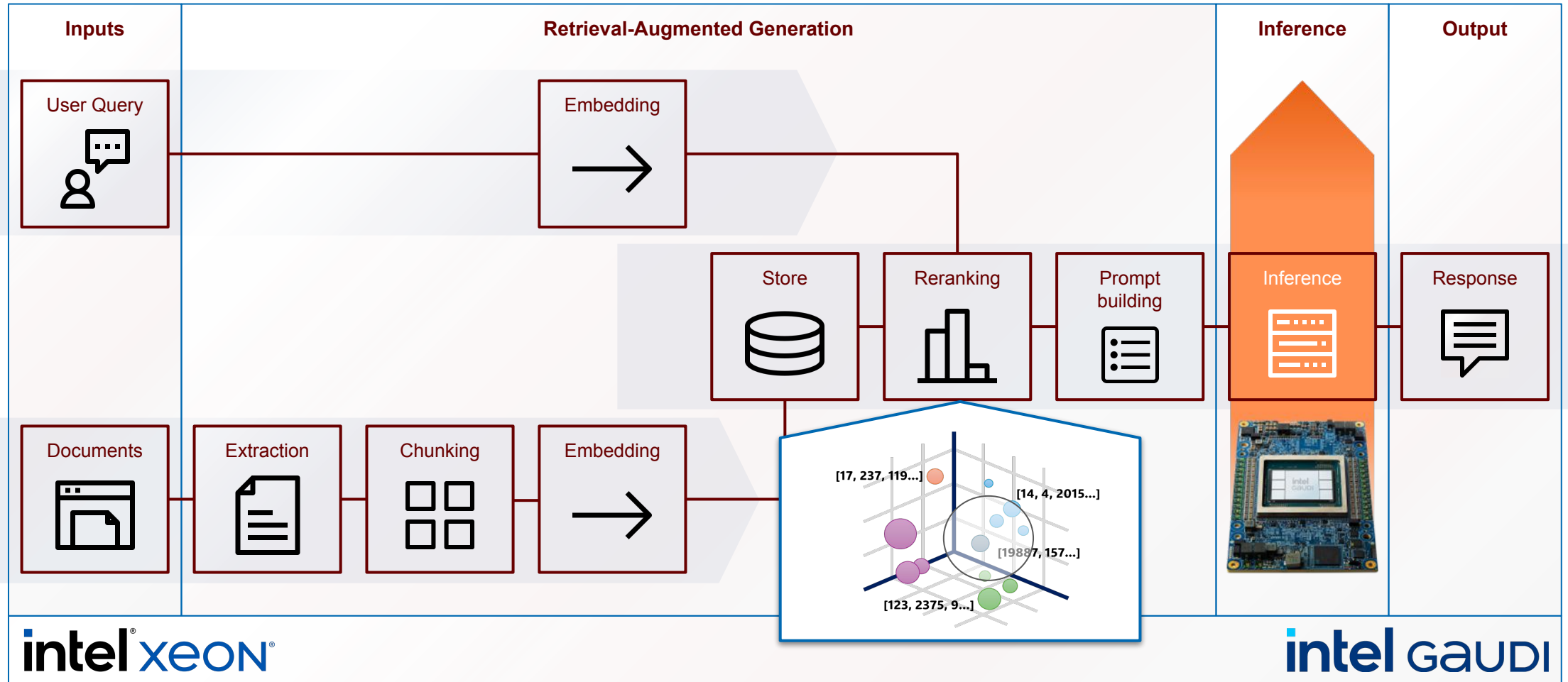


# Retrieval Augmented Generation (RAG)

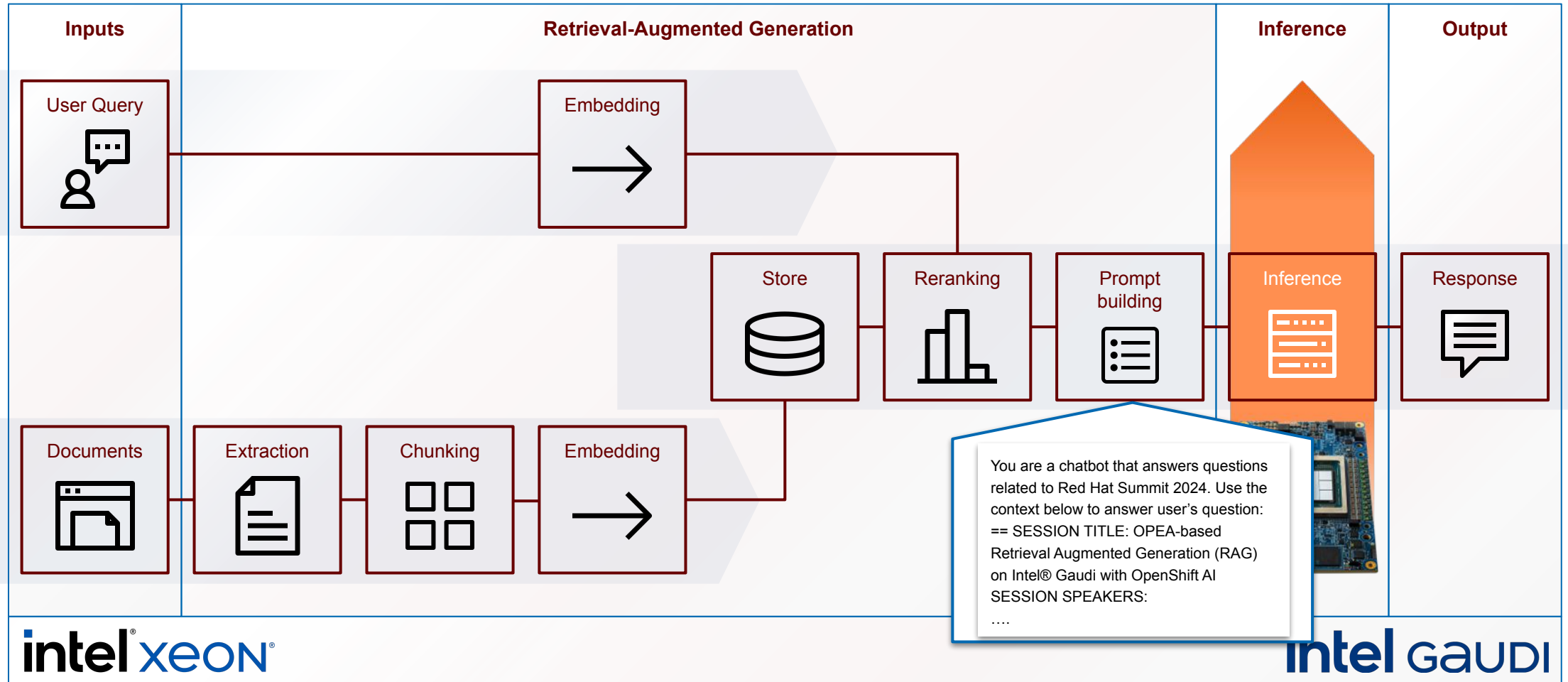




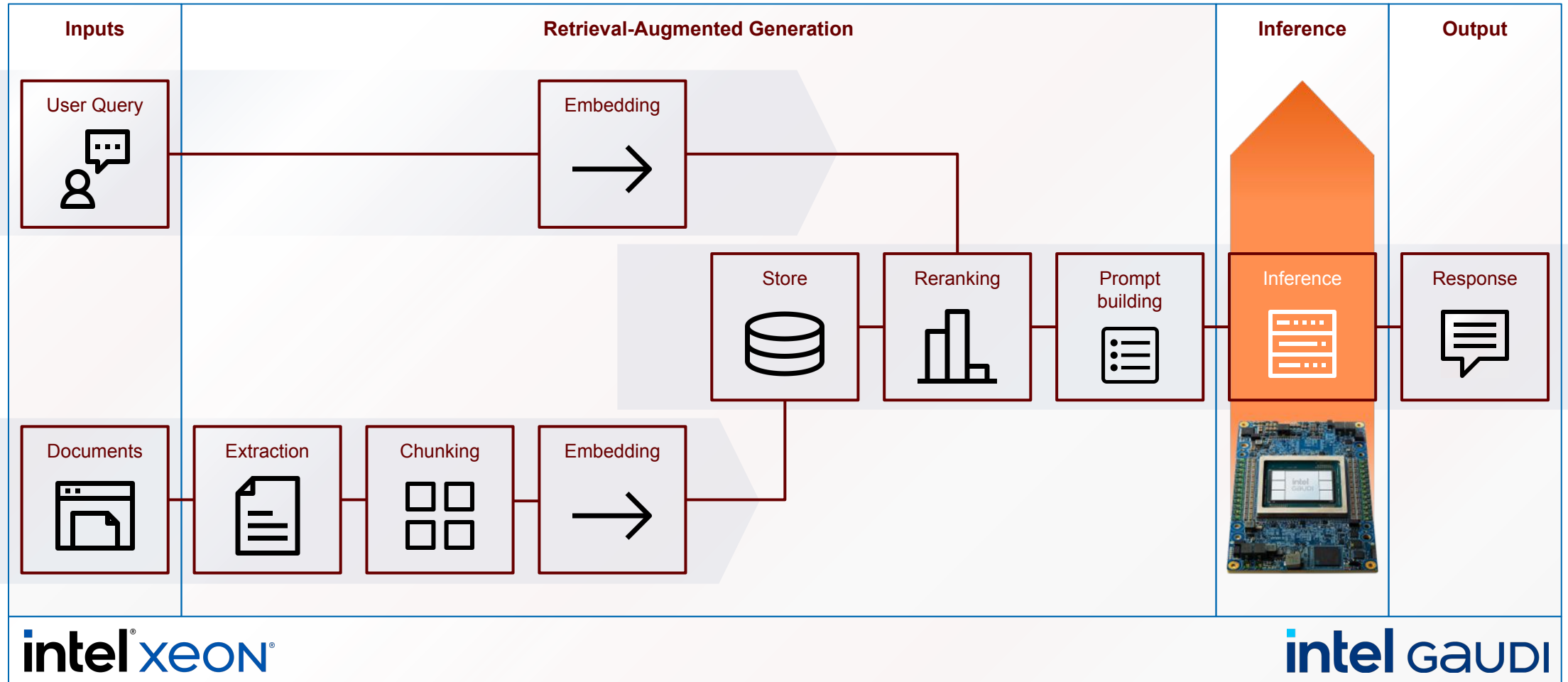
# Retrieval Augmented Generation (RAG)



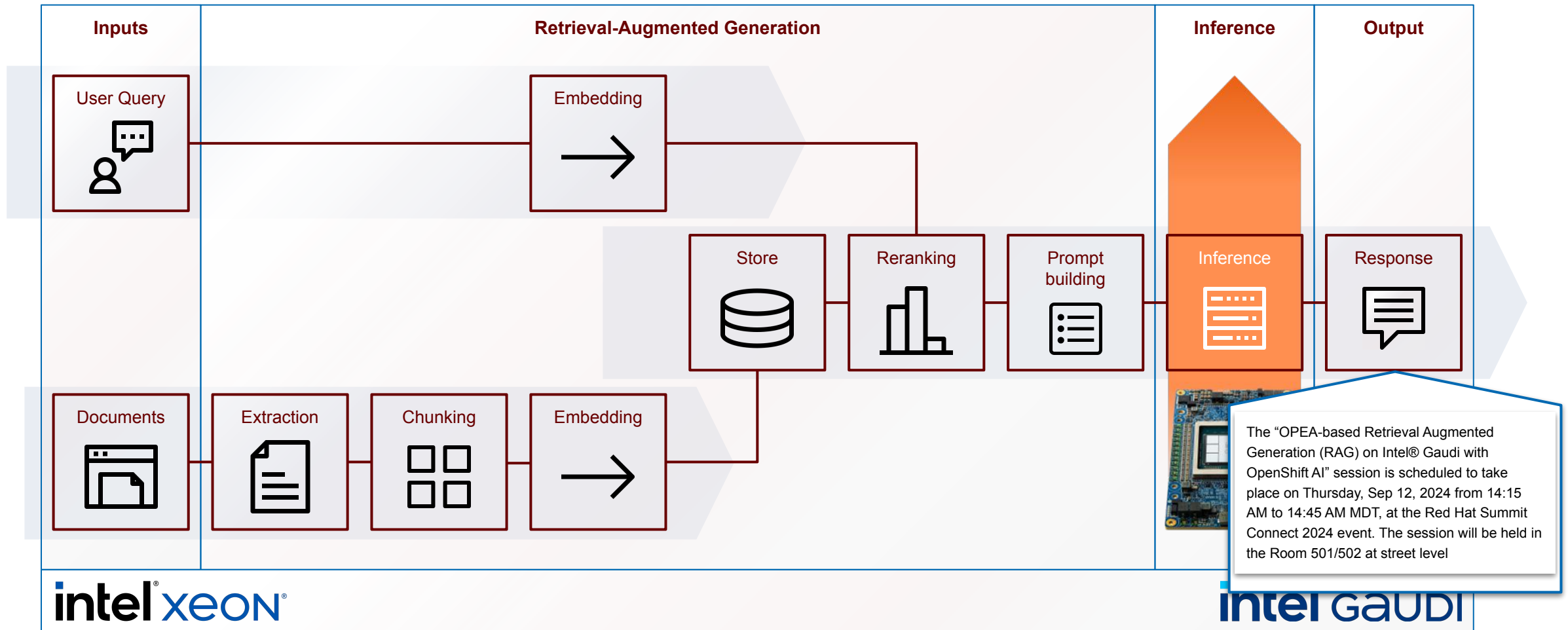
# Retrieval Augmented Generation (RAG)



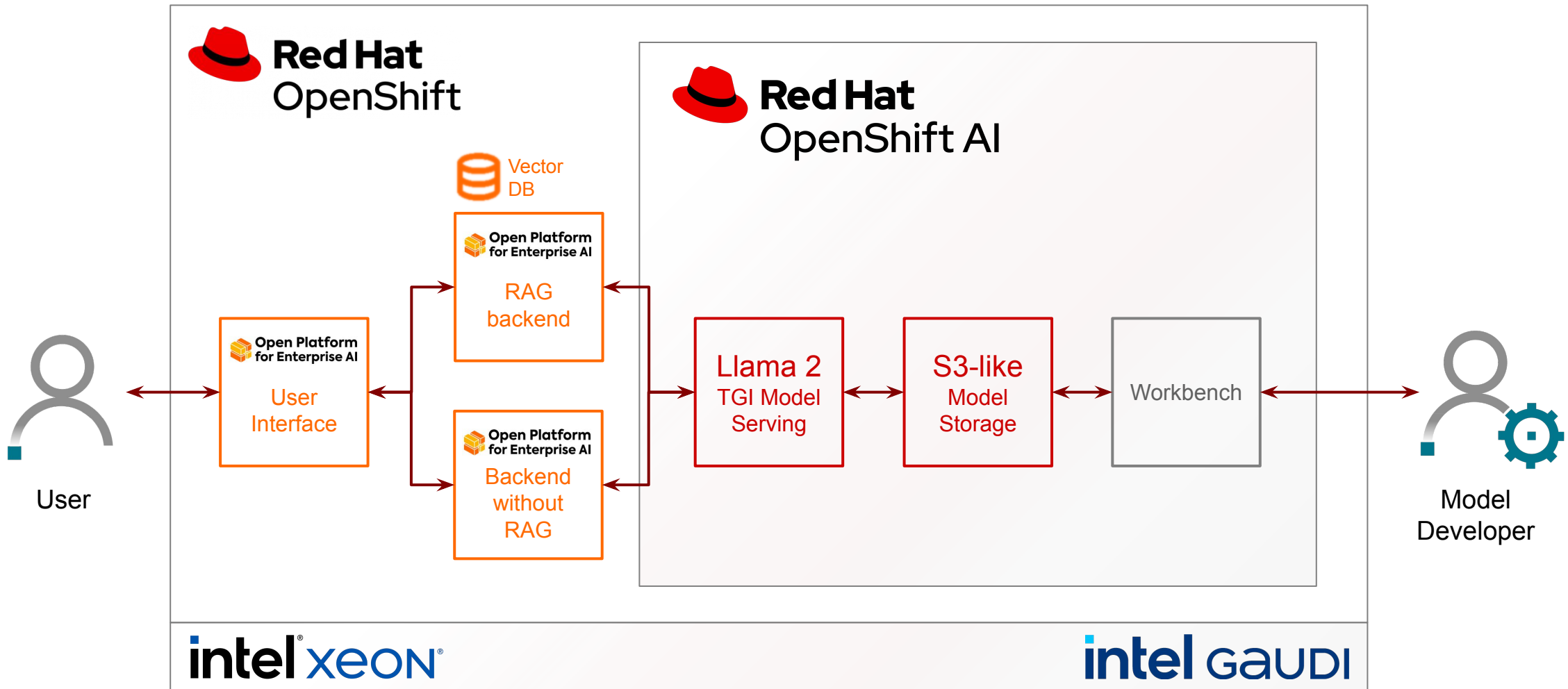
# Retrieval Augmented Generation (RAG)



# Retrieval Augmented Generation (RAG)



# Retrieval Augmented Generation (RAG) Chatbot Demo



- Administrator ▾
- Home >
- Operators ▾
  - OperatorHub
  - Installed 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.

Project: All Projects ▾

## Installed Operators

Installed Operators are represented by ClusterServiceVersions within this Namespace. For more information, see the [Understanding Operators documentation](#). Or create an Operator and ClusterServiceVersion using the [Operator SDK](#).

Name ▾

Name	Namespace	Managed Namespaces	Status	Last updated	Provided APIs
 <b>Intel® Gaudi AI SW Tools Operator</b> 0.0.1 provided by Intel	NS openshift-operators	All Namespaces	<span style="color: green;">✔</span> Succeeded Up to date	 Oct 1, 2024, 10:45 AM	GaudiAIToolsContainer
 <b>Intel Gaudi AI accelerator</b> 1.17.0-495 provided by Habana Labs Ltd.	NS habana-ai-operator	NS habana-ai-operator	<span style="color: green;">✔</span> Succeeded Up to date	 Sep 27, 2024, 1:10 PM	Device Config
 <b>Kernel Module Management</b> 2.1.1 provided by Red Hat	NS openshift-kmm	All Namespaces	<span style="color: green;">✔</span> Succeeded	 Oct 3, 2024, 8:17 PM	PreflightValidation PreflightValidationOCP Module NodeModulesConfig
 <b>Node Feature Discovery Operator</b> 4.16.0-202409202304 provided by Red Hat	NS openshift-nfd	NS openshift-nfd	<span style="color: green;">✔</span> Succeeded Up to date	 Oct 1, 2024, 11:07 PM	NodeFeatureDiscovery NodeFeatureRule NodeFeature
 <b>Package Server</b> 0.0.1-snapshot provided by Red Hat	NS openshift-operator-lifecycle-manager	NS openshift-operator-lifecycle-manager	<span style="color: green;">✔</span> Succeeded	 Sep 12, 2024, 3:46 PM	PackageManifest
 <b>Red Hat OpenShift AI</b> 2.13.0 provided by Red Hat, Inc.	NS redhat-ods-operator	All Namespaces	<span style="color: green;">✔</span> Succeeded Up to date	 Sep 15, 2024, 11:59 PM	Data Science Cluster DSC Initialization FeatureTracker
 <b>Red Hat OpenShift Serverless</b> 1.33.2 provided by Red Hat	NS openshift-serverless	All Namespaces	<span style="color: green;">✔</span> Succeeded Up to date	 Sep 27, 2024, 1:10 PM	Knative Serving Knative Eventing Knative Kafka
 <b>Red Hat OpenShift Service Mesh</b> 2.6.2-0 provided by Red Hat, Inc.	NS openshift-operators	All Namespaces	<span style="color: green;">✔</span> Succeeded Up to date	 Oct 7, 2024, 3:05 AM	Istio Service Mesh Control Plane Istio Service Mesh Member Istio Service Mesh Member Roll

- Applications >
- Data Science Projects
- Data Science Pipelines
- Model Serving
- Resources
- Settings ▾
  - Notebook images
  - Cluster settings
  - Accelerator profiles
  - Serving runtimes
  - User management

## 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
Text Generation Inference on Habana Gaudi ?	<input checked="" type="checkbox"/>	Single-model	REST
Caikit TGIS ServingRuntime for KServe ? Pre-installed	<input checked="" type="checkbox"/>	Single-model	REST
OpenVINO Model Server ? Pre-installed	<input checked="" type="checkbox"/>	Single-model	REST
OpenVINO Model Server ? Pre-installed			
TGIS Standalone ServingRuntime for KServe ? Pre-installed			

To accelerate your OpenShift AI model with Intel® Gaudi® 2, you need a suitable Serving runtime



Home

Applications

Data Science Projects

Data Science Pipelines

Experiments

Distributed Workload Metrics

Model Serving

Resources

Settings

Data Science Projects > gaudi-llama

## gaudi-llama

Overview Workbenches Pipelines **Models** Cluster storage Data connections Permissions

### Models and model servers

Deploy model

Single-model serving enabled

Model name	Serving runtime	Inference endpoint	API protocol	Status
gaudi-llama3	tgi-gaudi-llama3	Internal Service	REST	✓
Framework	llm			
Model server replicas	1			
Model server size	Custom			
	16 CPUs, 128Gi Memory requested			
	16 CPUs, 128Gi Memory limit			
Accelerator	gaudi			
Number of accelerators	4			

You are logged in as a temporary administrative user. Update the [cluster OAuth configuration](#) to allow others to log in.

- Administrator
- Home
- Operators
- Workloads
  - Pods**
  - Deployments
  - DeploymentConfigs
  - StatefulSets
  - Secrets
  - ConfigMaps
  - CronJobs
  - Jobs
  - DaemonSets
  - ReplicaSets
  - ReplicationControllers
  - HorizontalPodAutoscalers
  - PodDisruptionBudgets
- Serverless
- Networking
- Storage
- Builds
- Observe
- Compute

Project: gaudi-demo

### Pods

Create Pod

Filter Name Search by name...

Name ↑	Status ↓	Ready ↓	Restarts ↓	Owner ↓	Memory ↓	CPU ↓	Created ↓
chatqna-non-rag-redis-75bc5549fc-429wm	Running	1/1	0	RS chatqna-non-rag-redis-75bc5549fc	326.0 MiB	0.001 cores	Oct 4, 2024, 10:01 AM
chatqna-rag-redis-5b5c79d846-6fksk	Running	1/1	0	RS chatqna-rag-redis-5b5c79d846	840.8 MiB	0.001 cores	Oct 4, 2024, 10:04 AM
redis-vector-db-6d69cc9495-hvxx8	Running	1/1	0	RS redis-vector-db-6d69cc9495	140.0 MiB	0.001 cores	Oct 4, 2024, 10:04 AM
ui-demo-6f688c8486-hlgk	Running	1/1	0	RS ui-demo-6f688c8486	304.7 MiB	0.001 cores	Oct 4, 2024, 10:16 AM

You are logged in as a temporary administrative user. Update the [cluster OAuth configuration](#) to allow others to log in.

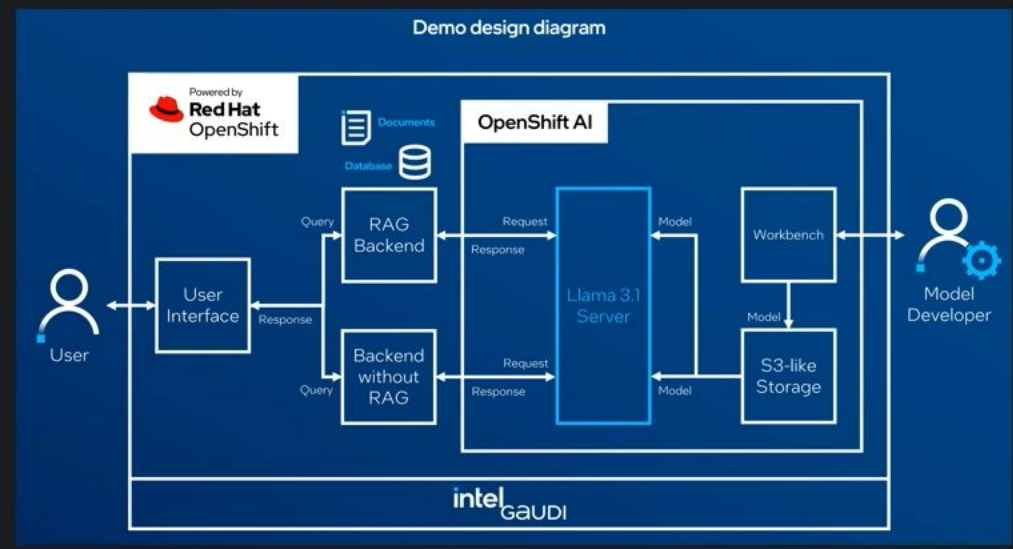
Project: gaudi-demo

### Pods

Create Pod

Filter Name Search by name...

Name ↑	Status ↓	Ready ↓	Restarts ↓	Owner ↓	Memory ↓	CPU ↓	Created ↓
chatqna-non-rag-redis-75bc5549fc-429wm	Running	1/1	0	RS chatqna-non-rag-redis-75bc5549fc	326.0 MiB	0.001 cores	Oct 4, 2024, 10:01 AM
chatqna-rag-redis-5b5c79d846-6fksk	Running	1/1	0	RS chatqna-rag-redis-5b5c79d846	840.8 MiB	0.001 cores	Oct 4, 2024, 10:04 AM
redis-vector-db-6d69cc9495-hvxx8	Running	1/1	0	RS redis-vector-db-6d69cc9495	140.0 MiB	0.001 cores	Oct 4, 2024, 10:04 AM
ui-demo-6f688c8486-hlghk	Running	1/1	0	RS ui-demo-6f688c8486	304.7 MiB	0.001 cores	Oct 4, 2024, 10:16 AM



There are four more applications deployed for the demo purposes.

- Administrator
- Home
- Operators
- Workloads
  - Pods
  - Deployments
  - DeploymentConfigs
  - StatefulSets
  - Secrets
  - ConfigMaps
  - CronJobs
  - Jobs
  - DaemonSets
  - ReplicaSets
  - ReplicationControllers
  - HorizontalPodAutoscalers
  - PodDisruptionBudgets
- Serverless
- Networking
- Storage
- Builds
- Observe
- Compute

You are logged in as a temporary administrative user. Update the [cluster OAuth configuration](#) to allow others to log in.

Project: gaudi-demo

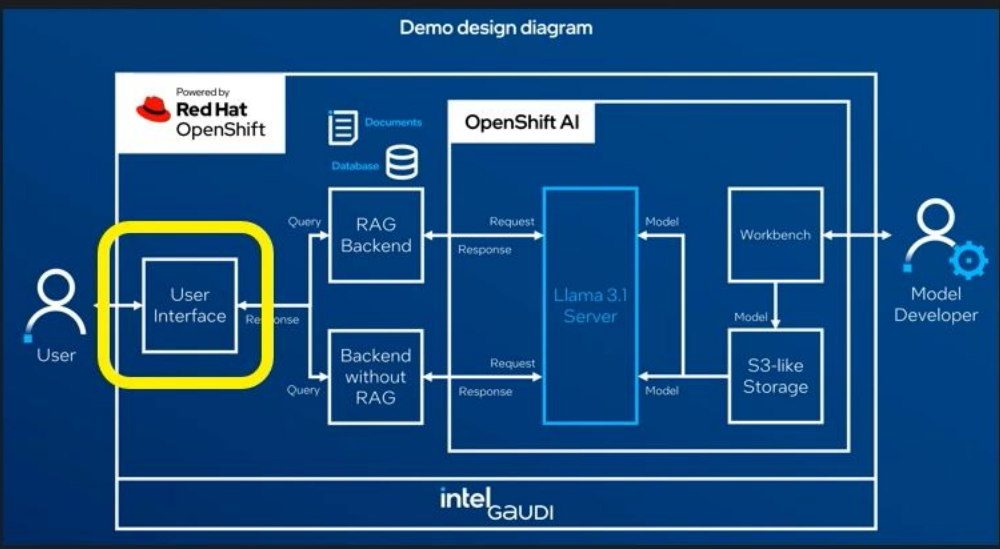
### Pods

Create Pod

Filter Name Search by name...

Name ↑	Status ↓	Ready ↓	Restarts ↓	Owner ↓	Memory ↓	CPU ↓	Created ↓
chatqna-non-rag-redis-75bc5549fc-429wm	Running	1/1	0	RS chatqna-non-rag-redis-75bc5549fc	326.0 MiB	0.001 cores	Oct 4, 2024, 10:01 AM
chatqna-rag-redis-5b5c79d846-6fksk	Running	1/1	0	RS chatqna-rag-redis-5b5c79d846	840.8 MiB	0.001 cores	Oct 4, 2024, 10:04 AM
redis-vector-db-6d69cc9495-hvxx8	Running	1/1	0	RS redis-vector-db-6d69cc9495	140.0 MiB	0.001 cores	Oct 4, 2024, 10:04 AM
ui-demo-6f688c8486-hlghk	Running	1/1	0	RS ui-demo-6f688c8486	304.7 MiB	0.001 cores	Oct 4, 2024, 10:16 AM

- Administrator
- Home
- Operators
- Workloads
  - Pods**
  - Deployments
  - DeploymentConfigs
  - StatefulSets
  - Secrets
  - ConfigMaps
  - CronJobs
  - Jobs
  - DaemonSets
  - ReplicaSets
  - ReplicationControllers
  - HorizontalPodAutoscalers
  - PodDisruptionBudgets
- Serverless
- Networking
- Storage
- Builds
- Observe
- Compute



The UI to interact with the model.

You are logged in as a temporary administrative user. Update the [cluster OAuth configuration](#) to allow others to log in.

Project: gaudi-demo

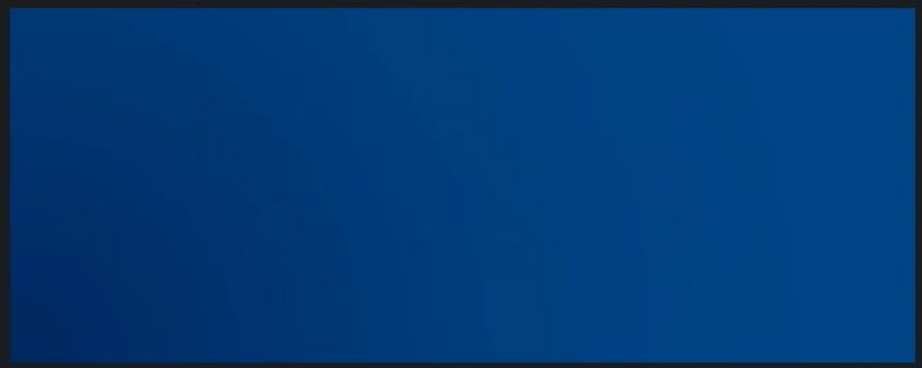
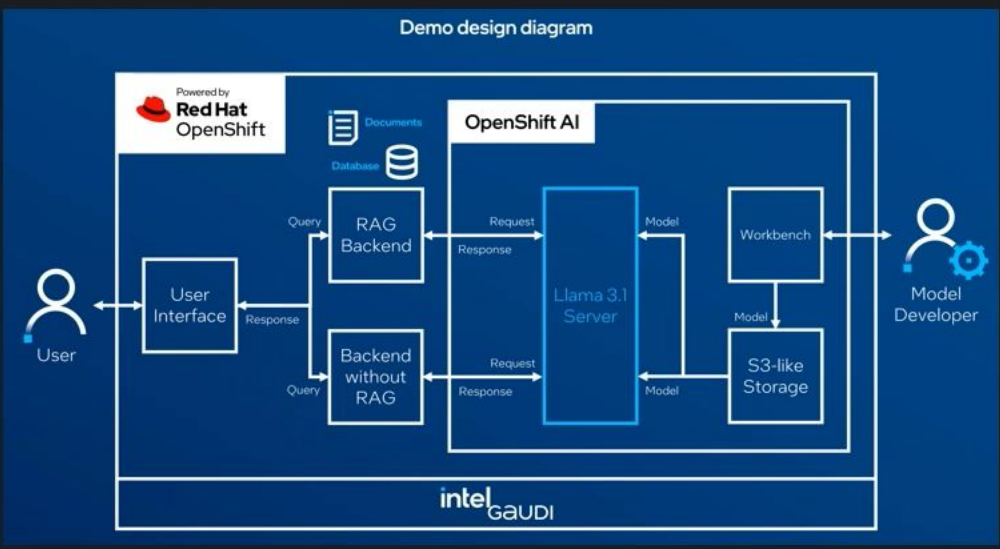
### Pods

Create Pod

Filter Name Search by name...

Name ↑	Status ↓	Ready ↓	Restarts ↓	Owner ↓	Memory ↓	CPU ↓	Created ↓
chatqna-non-rag-redis-75bc5549fc-429wm	Running	1/1	0	RS chatqna-non-rag-redis-75bc5549fc	326.0 MiB	0.001 cores	Oct 4, 2024, 10:01 AM
chatqna-rag-redis-5b5c79d846-6fksk	Running	1/1	0	RS chatqna-rag-redis-5b5c79d846	840.8 MiB	0.001 cores	Oct 4, 2024, 10:04 AM
redis-vector-db-6d69cc9495-hvxx8	Running	1/1	0	RS redis-vector-db-6d69cc9495	140.0 MiB	0.001 cores	Oct 4, 2024, 10:04 AM
ui-demo-6f688c8486-hlghk	Running	1/1	0	RS ui-demo-6f688c8486	304.7 MiB	0.001 cores	Oct 4, 2024, 10:16 AM

- Administrator
- Home
- Operators
- Workloads
  - Pods**
  - Deployments
  - DeploymentConfigs
  - StatefulSets
  - Secrets
  - ConfigMaps
  - CronJobs
  - Jobs
  - DaemonSets
  - ReplicaSets
  - ReplicationControllers
  - HorizontalPodAutoscalers
  - PodDisruptionBudgets
- Serverless
- Networking
- Storage
- Builds
- Observe
- Compute



What's on your mind?



Stock Llama2 70b

Stock Llama2 70b with RAG





### Stock Llama2 70b

Hi!

Hello! What can I help you with today?




### Stock Llama2 70b with RAG

Hi!

Hi! Welcome to Red Hat Summit 2024! We're excited to have you join us for this year's event. Can we help you with something specific? Are you looking for information on a particular topic or session? Or perhaps you'd like to know more about our keynote speakers or networking opportunities? Let us know, and we'll do our best to assist you. See you soon!

#### Sources

 Red Hat Summit 2024 s...



# Summary



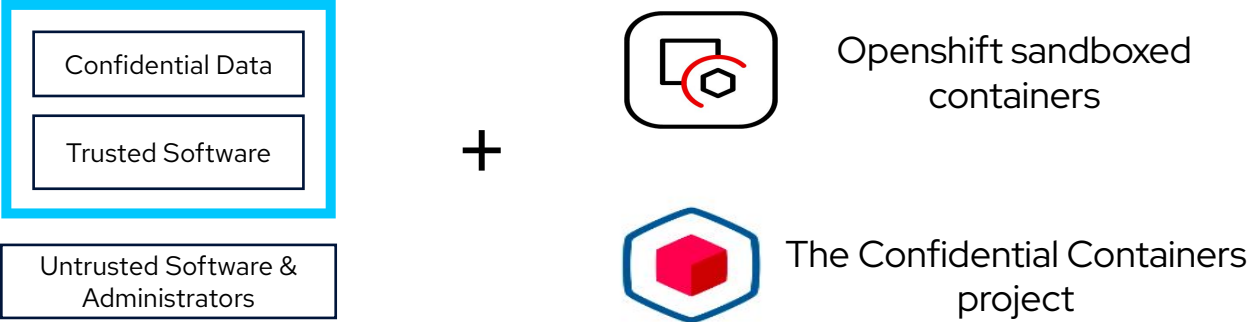
# Key Takeaways

- ▶ RAG enhances AI development
- ▶ OPEA simplifies AI deployment
- ▶ OpenShift AI integrates into DevOps workflow
- ▶ Intel Gaudi 3 accelerates AI training and inference

# 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.

# Confidential AI Helps Protect Data & Models In-Use

Utilizing Confidential Computing for Containers with Intel TDX

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

Come visit the Intel and Red Hat booth on the showfloor to learn more about Confidential Computing

Confidential Data  
Trusted Software  
Untrusted Software & Administrators

OpenShift Sinoex Containers  
The confidential containers project



Learn more!



Learn more!

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

# Q&A

Red Hat  
**Summit**

**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)

## CODRIN BUCUR

Principal AI Specialist Solution Architect  
Red Hat EMEA



**Bio:** As an Principal AI Specialist Solution Architect, Codrin is supporting Red Hat customers and partners in EMEA with their data science, AI/ML and MLOps needs and best practices. Previously, as Architect and TSM in Red Hat Consulting Alps for 7+ years, Codrin has supported customers with their adoption of Red Hat container platform, integration and middleware technologies.

**Contact:** [cbucur@redhat.com](mailto:cbucur@redhat.com)

<https://www.linkedin.com/in/codrin>



**Hind Azegrouz, PhD**  
**AI Inference Lead, EMEA**  
**Intel**



**Bio:**Hind Azegrouz, PhD is EMEA Lead for AI Inference at INTEL. Previous to her current role Hind Azegrouz was Data and AI architect at Repsol, Advanced analytics manager at Avanade, Advanced research fellow in Massachusetts Institute of Technology, research scientist at the Spanish National Center for Cardiovascular research. Hind Azegrouz pursued her PhD studies in the Edinburgh joint research institute (led by university of Edinburgh and heriot watt university) with focus on computer vision applications, she is also an electronics engineer from ENSEIB (Bordeaux, France). Hind Azegrouz is also assistant professor at IE business school where she teaches Compute

Contact: [Hind.Azegrouz@intel.com](mailto:Hind.Azegrouz@intel.com)  
<https://www.linkedin.com/in/hindazegrouz/>

