

Enhancing Enterprise AI with RAG:

Boost your AI's intelligence by seamlessly merging real-time data with LLMs

Red Hat Summit Connect 2024

Milan, 19 November 2024



Red Hat

Codrin Bucur

Principal AI Specialist Solution Architect,
EMEA, Red Hat





Gianluca Cecchi

Technical Sales Specialist
SMG 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

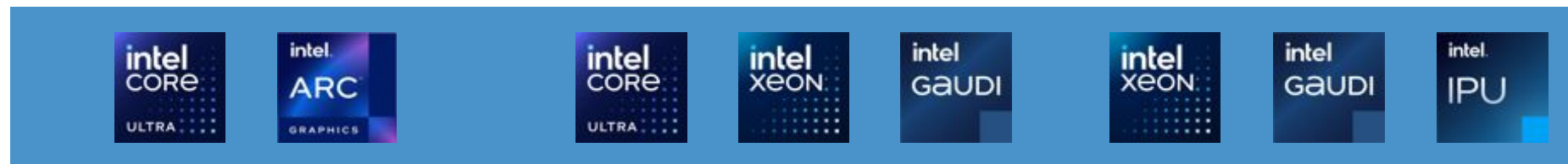
ENTERPRISE AI & EDGE AI
Open Standard, "Ready to Use"



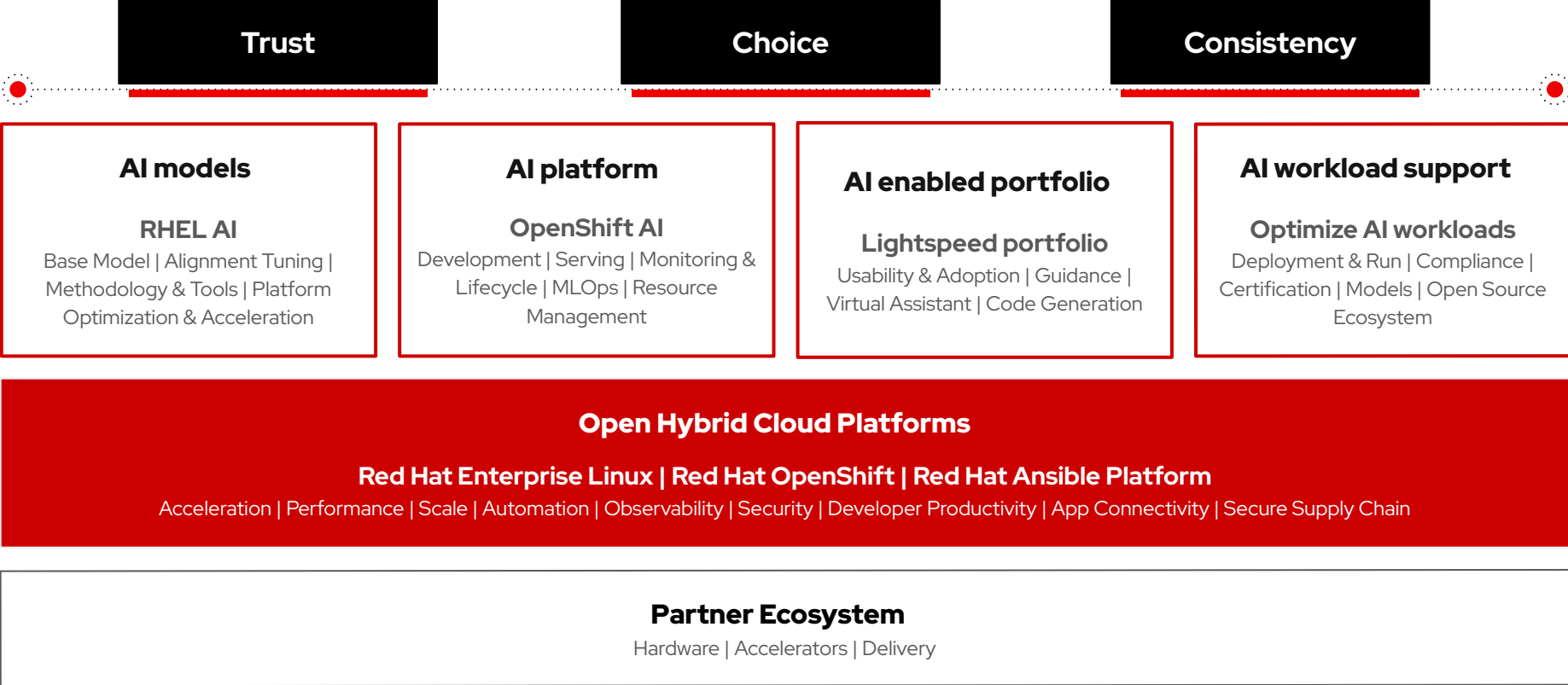
Super Cluster
Training, Tuning, Peak Inference

Mega Cluster
Large Scale Training & Inference

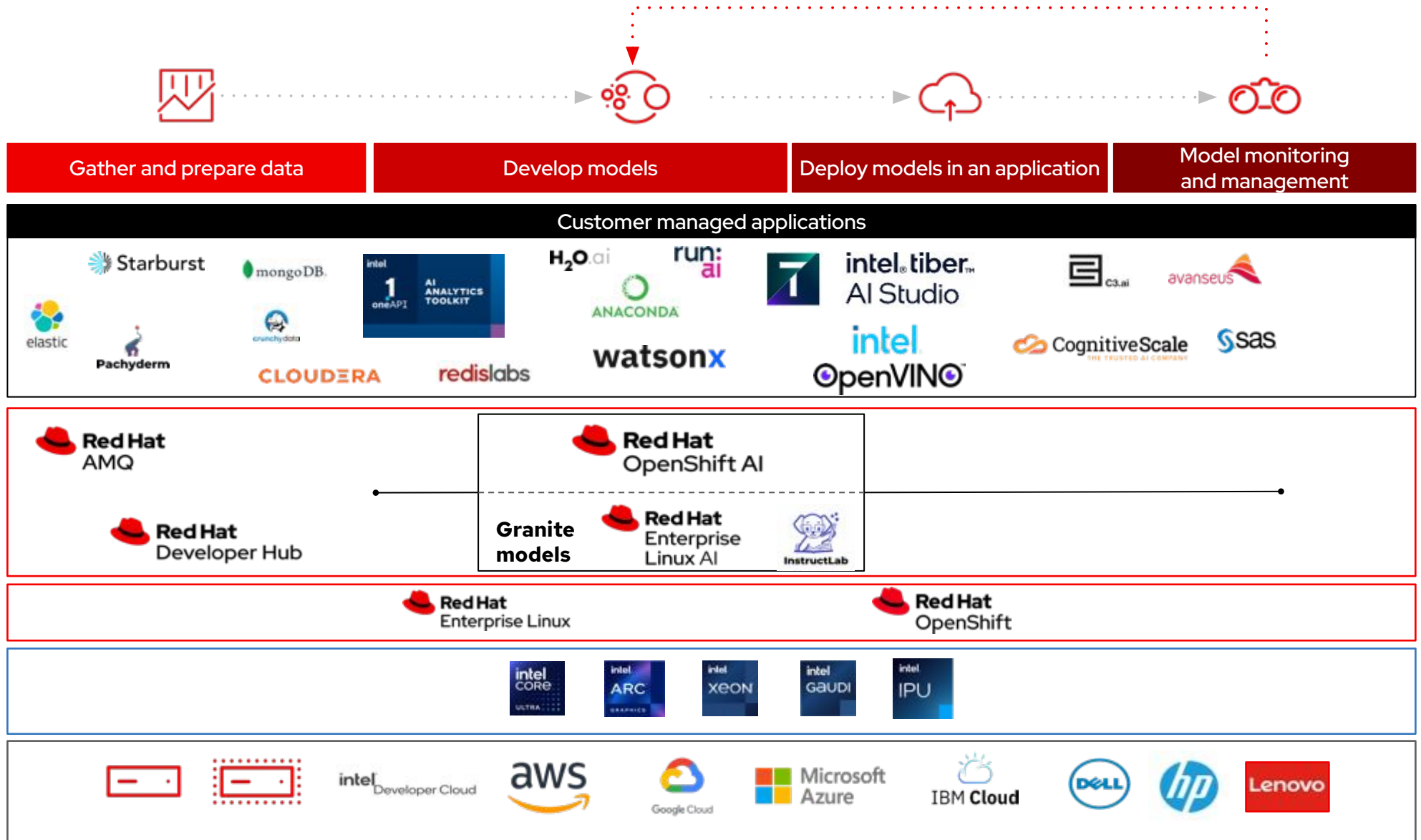
DATA CENTER AI
AI Open, Scalable Systems & Reference Arch



Red Hat's AI Strategy



Intel Enterprise AI with Red Hat® OpenShift® AI



ISV software and services including INTEL

Red Hat Software and cloud services Hybrid, multi-cloud platform

Trusted, cloud-ready platform

Intel® Core™, Intel® Arc™, Intel® Xeon®, Intel® Gaudi®, Intel® IPU

Deploy anywhere

OPEA – Open Platform for Enterprise AI

OPEA - Open Platform for Enterprise AI

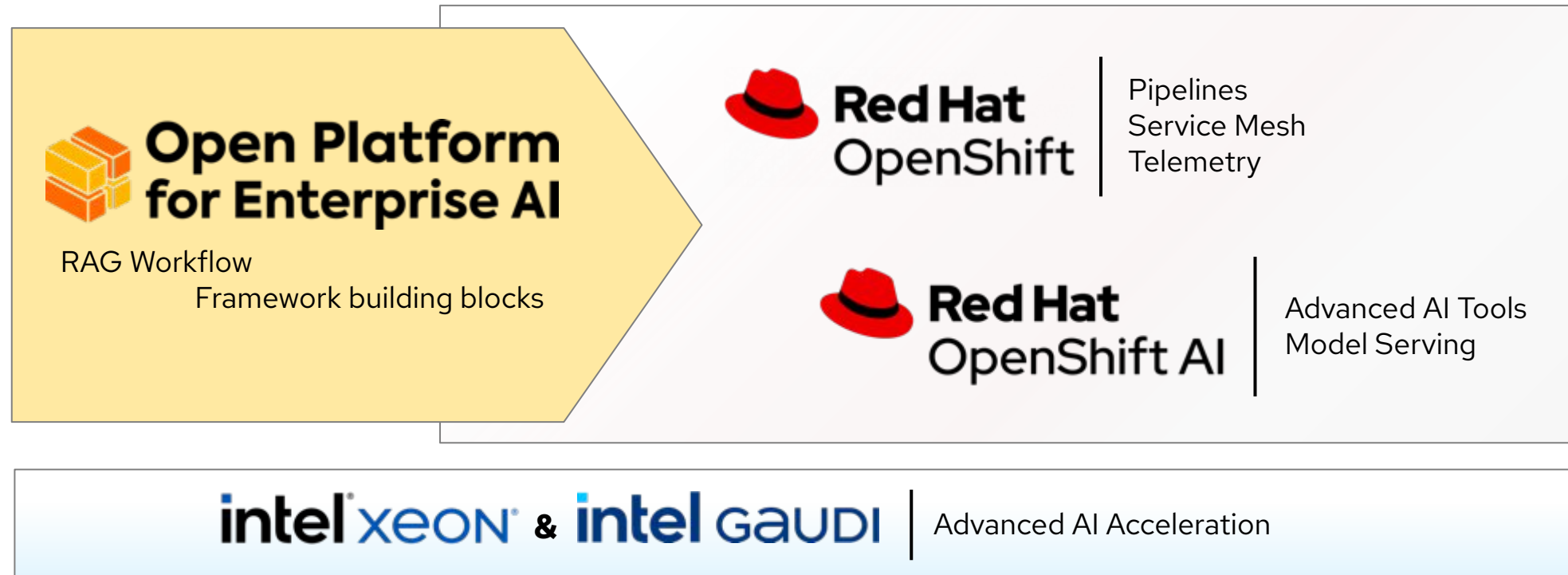
By The Linux Foundation

- ▶ 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**¹
- Projected **50% faster inferencing**²
- Projected **40% better power efficiency**³



Freedom to Scale without Lock-in

- Open standard ethernet networking vs proprietary InfiniBand
- 24x200 GbE ports of industry-standard RoCE on every Gaudi®³
- 33% more I/O peak throughput vs H100 for massive scale-up within the server⁴



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

*Public benchmarks on Gaudi 2 and Gaudi 3 available at: <https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html>

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

² 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.

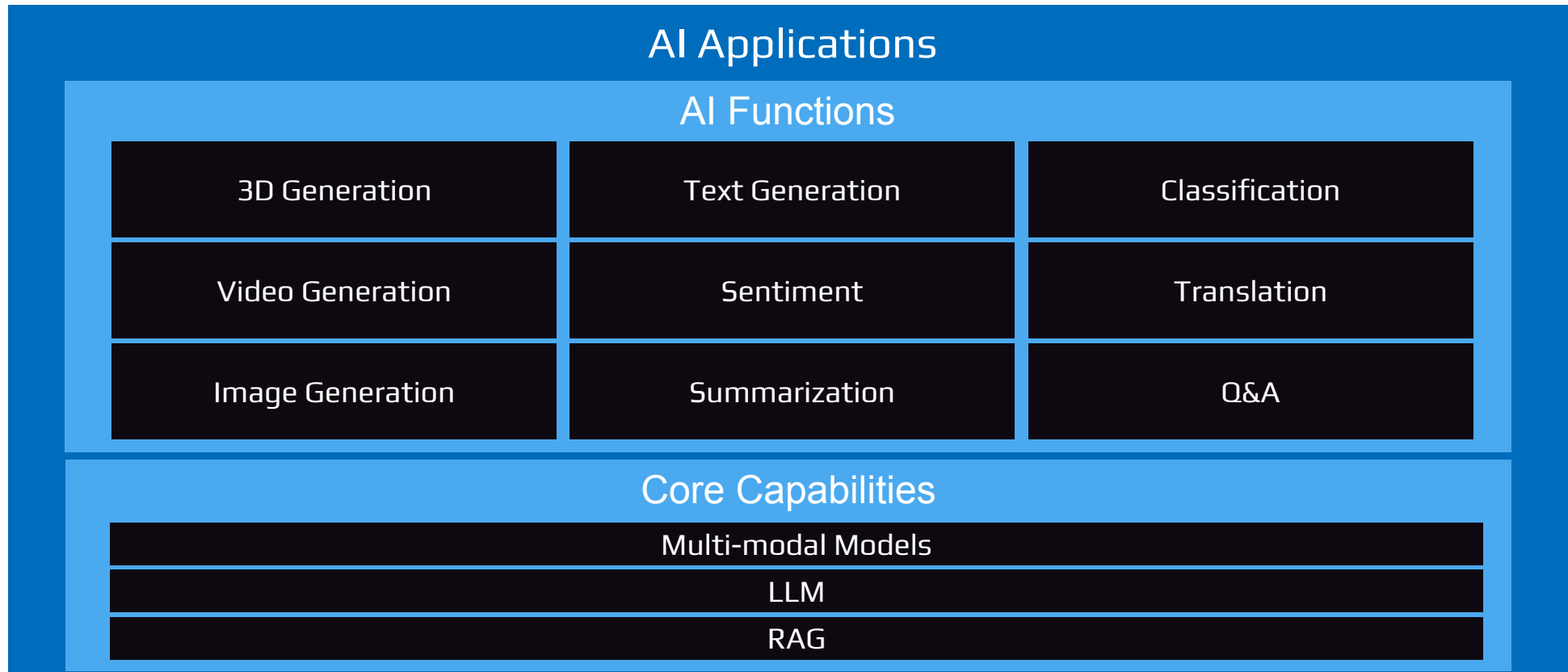
³ Source: NV comparison based on <https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8> , Mar 28th, 2024. Reported numbers are per GPU.

¹⁻³ 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.

⁴ 900 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.

intel | **IBM**

*

A photograph of the Intel Gaudi 3 AI Accelerator card, showing the central chip with the 'intel GAUDI' logo and various connectors on the blue PCB.

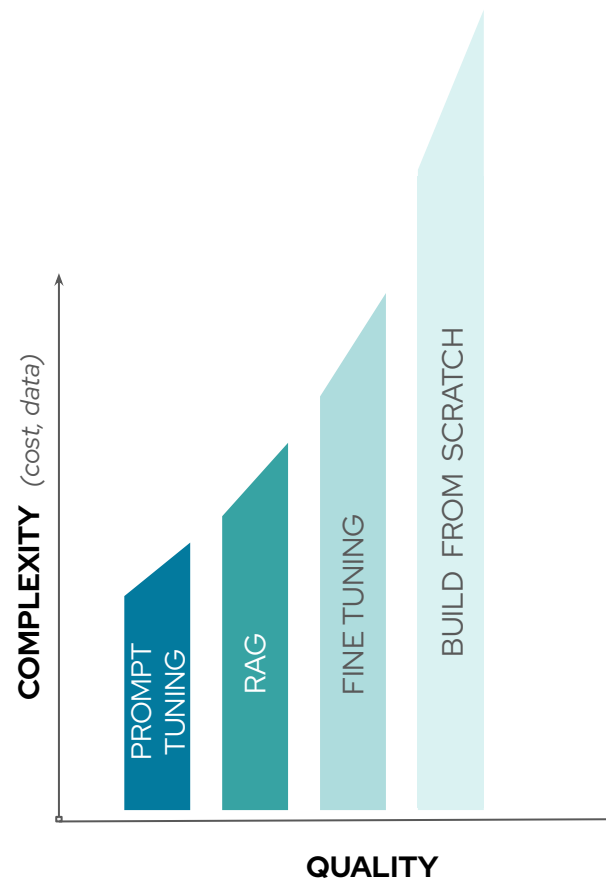
*See more at:
<https://newsroom.ibm.com/blog-intel-and-ibm-collaborate-to-provide-better-cost-performance-for-ai-innovation>
and
<https://www.intel.com/content/www/us/en/newsroom/news/intel-ibm-deliver-enterprise-ai-in-the-cloud.html>

Retrieval Augmented Generation (RAG) Explained

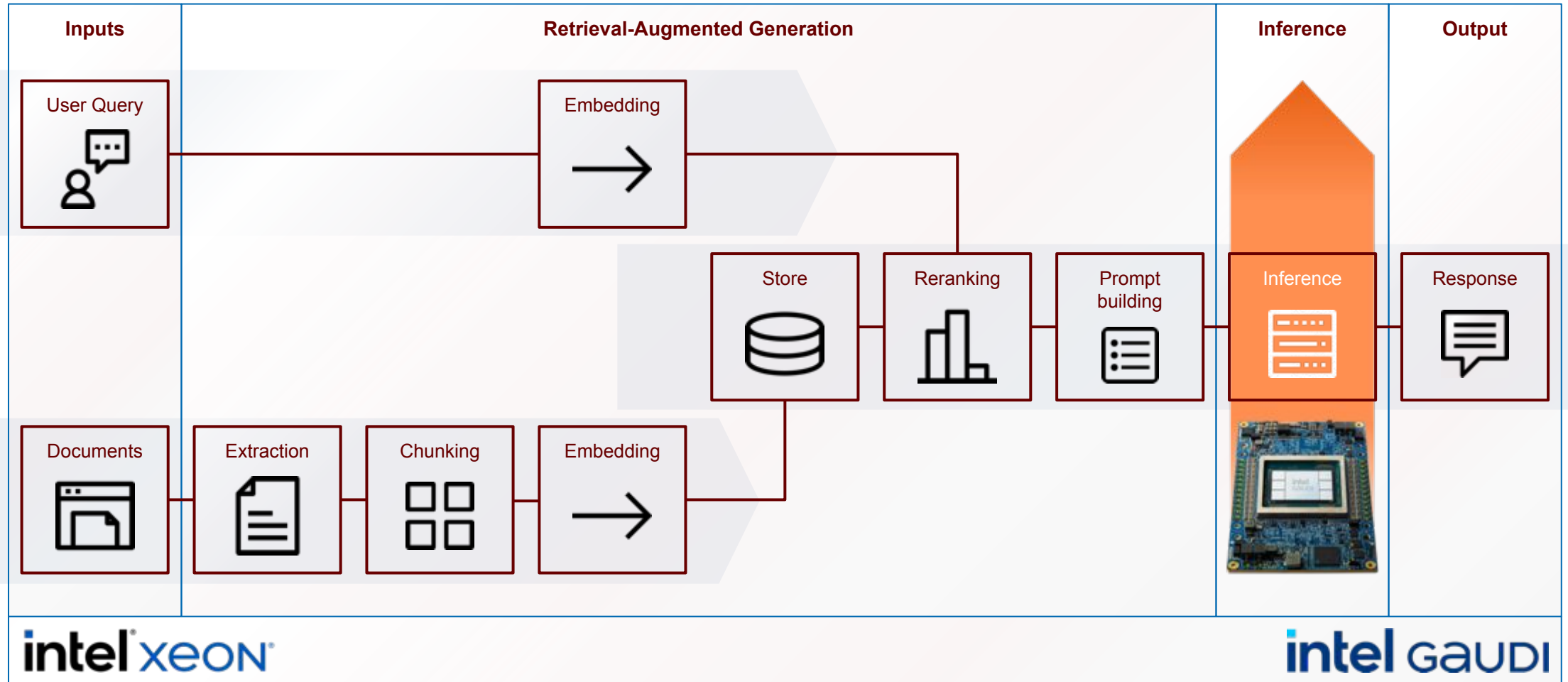
The balancing act of using foundation models

Foundation models will still need more work to be useful

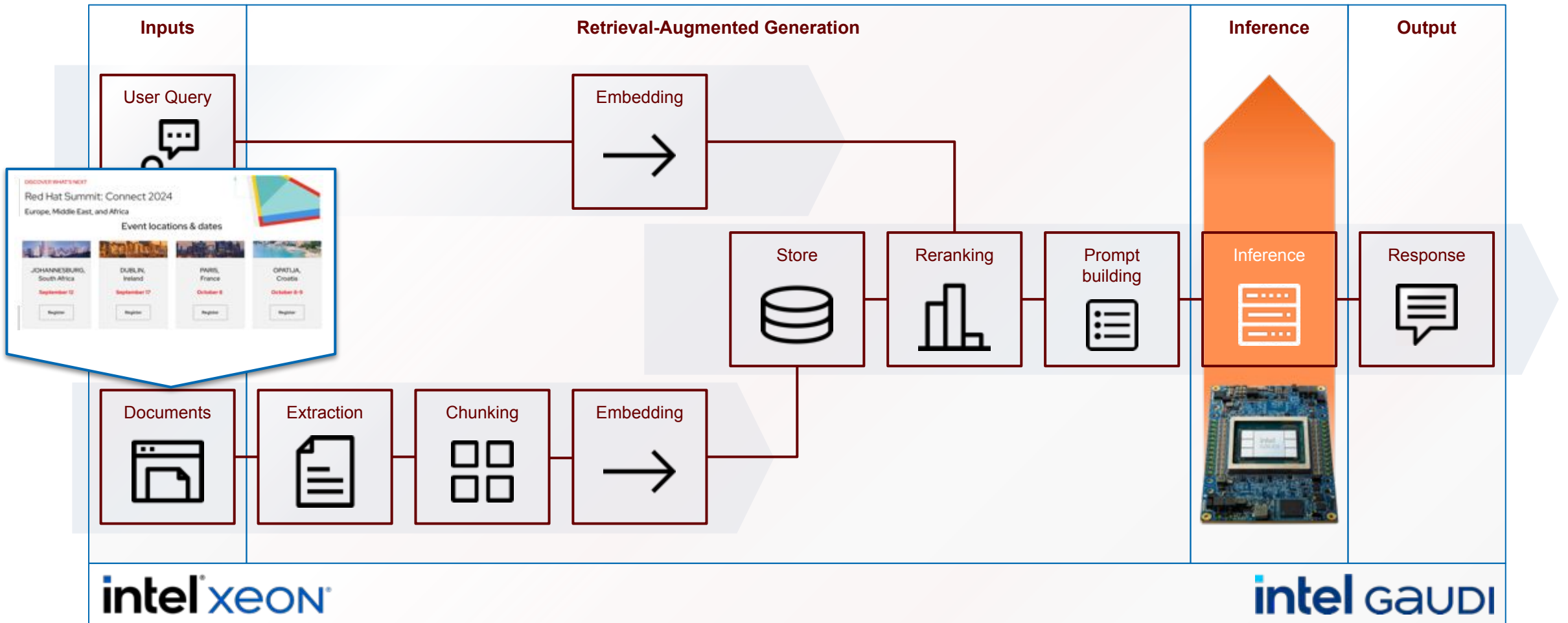
- ▶ Prompt tuning
- ▶ Retrieval-Augmented Generation (RAG)
- ▶ Fine tuning foundation models
- ▶ Training a Foundation Model from scratch



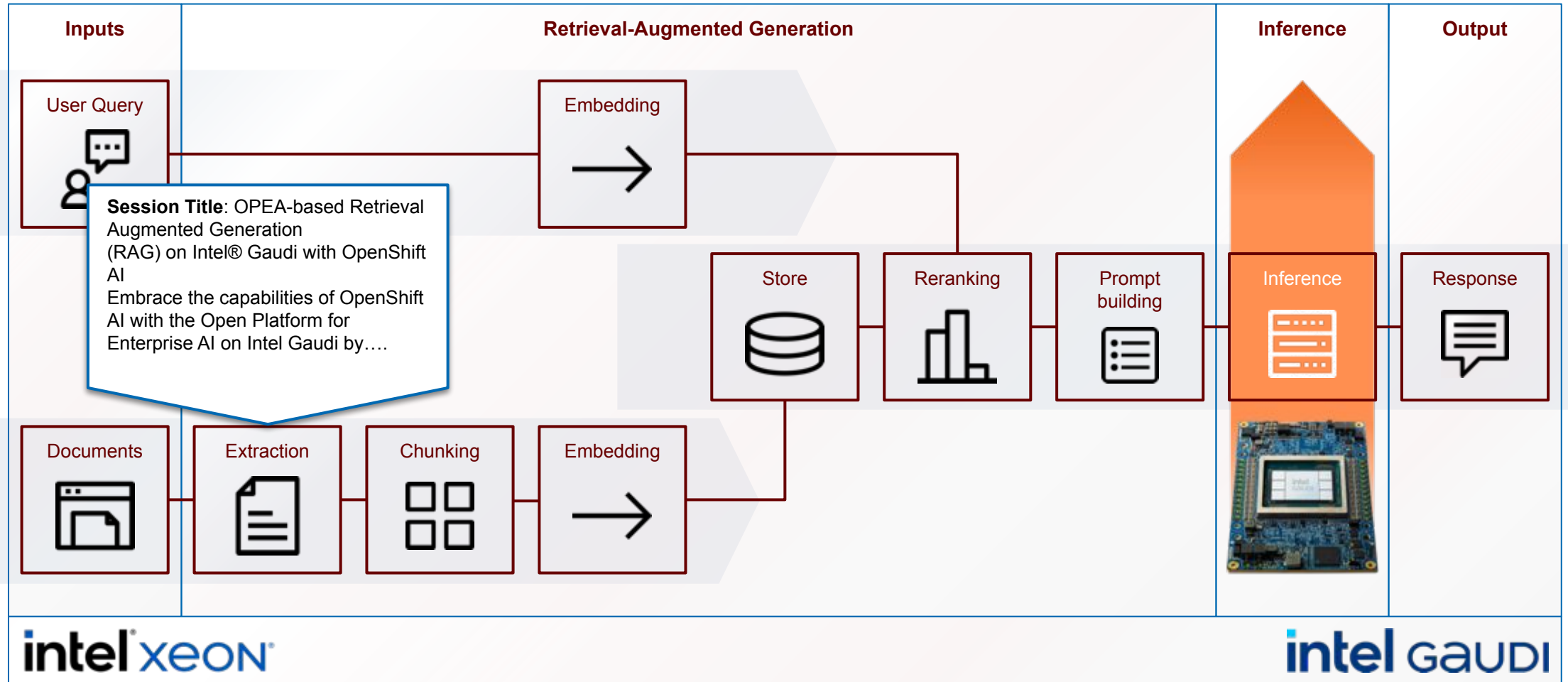
Retrieval Augmented Generation (RAG)



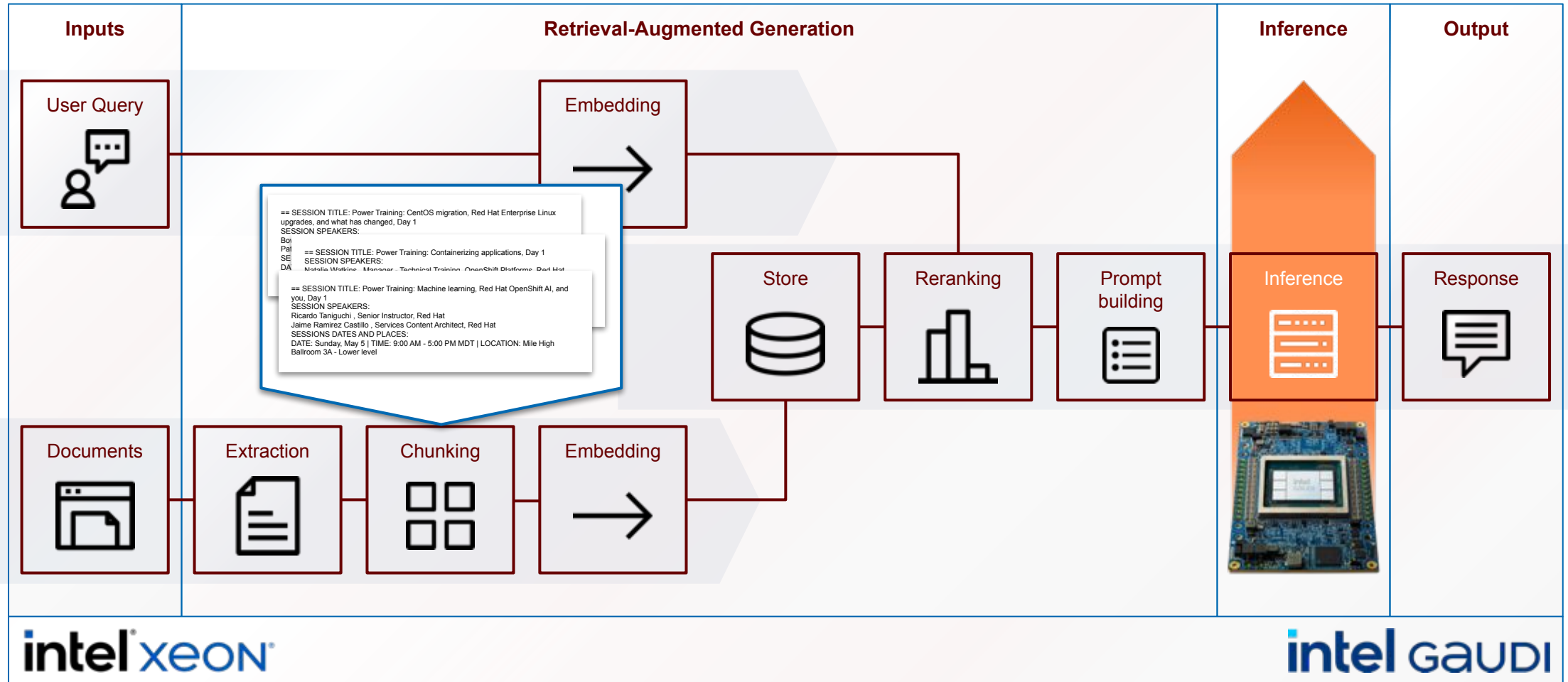
Retrieval Augmented Generation (RAG)



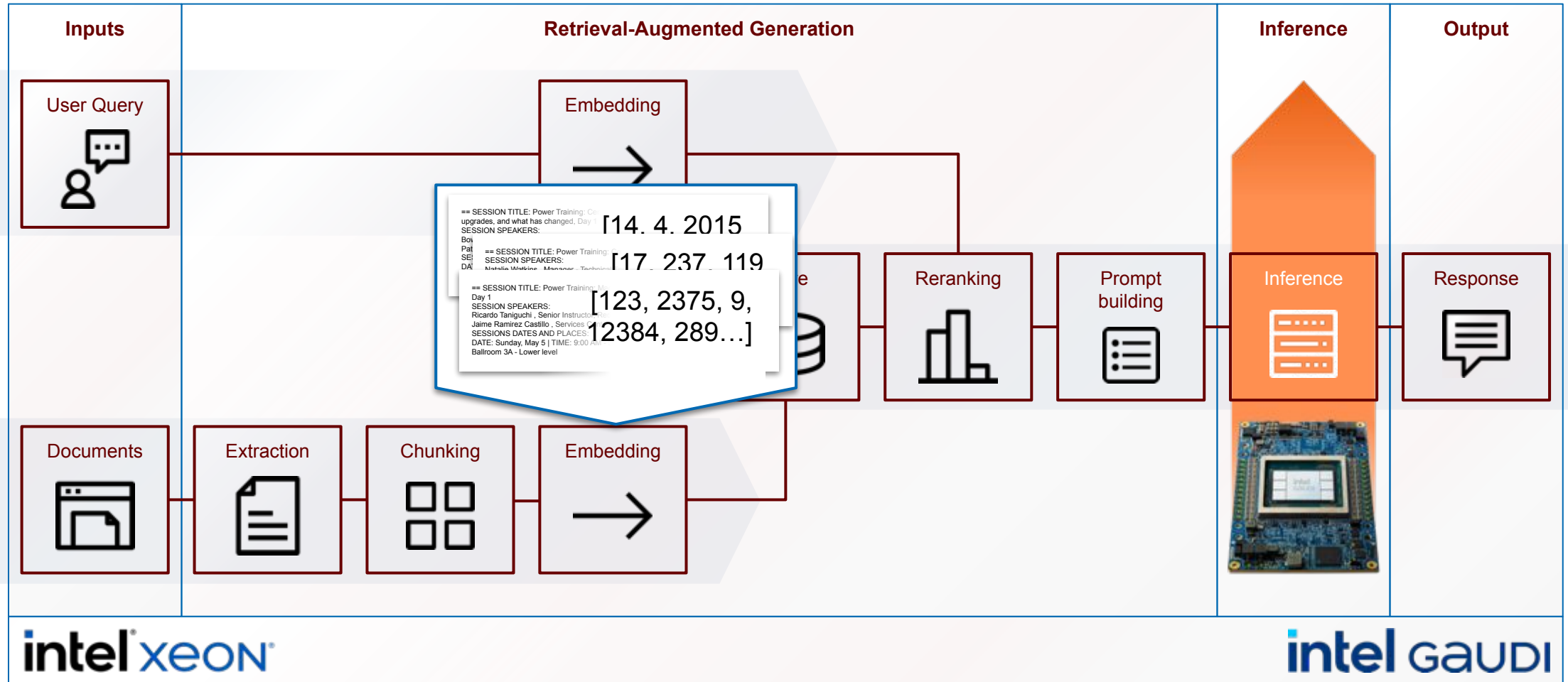
Retrieval Augmented Generation (RAG)



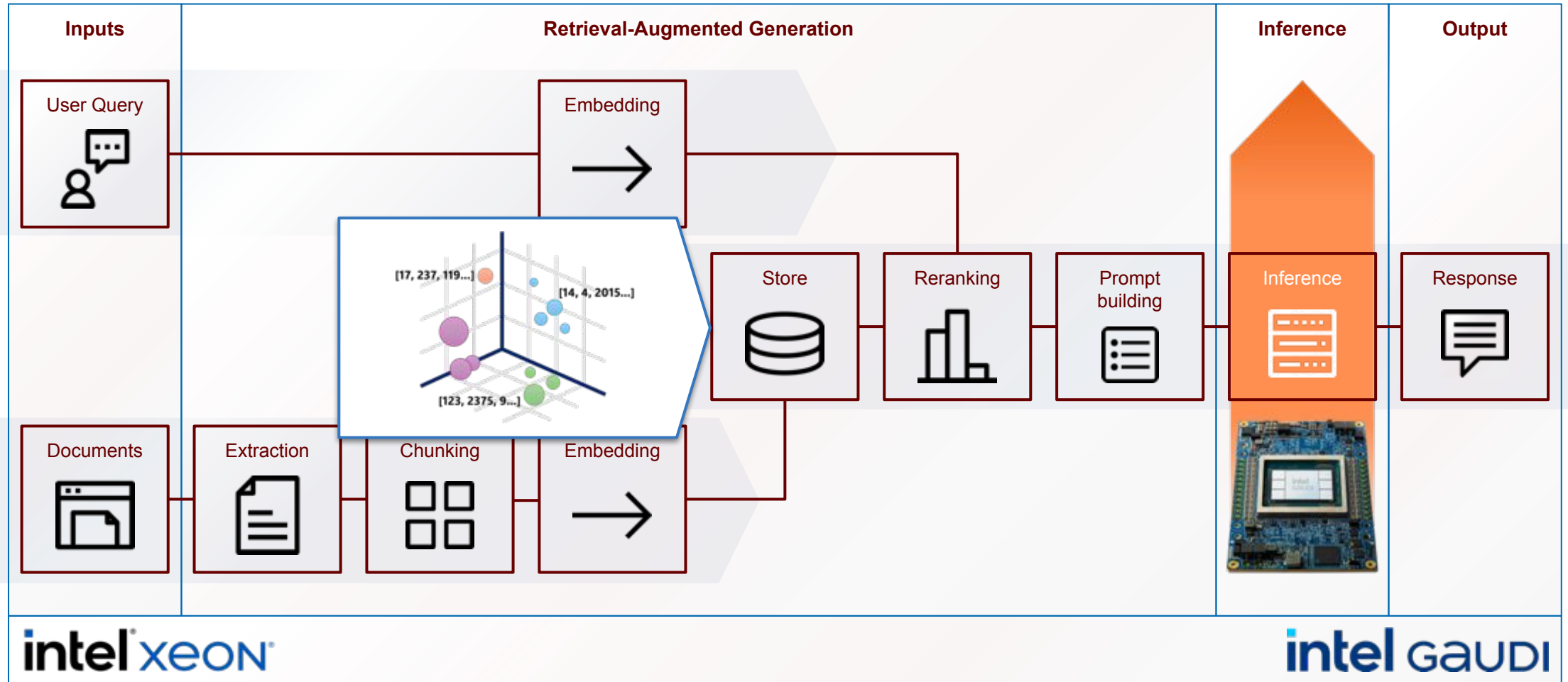
Retrieval Augmented Generation (RAG)



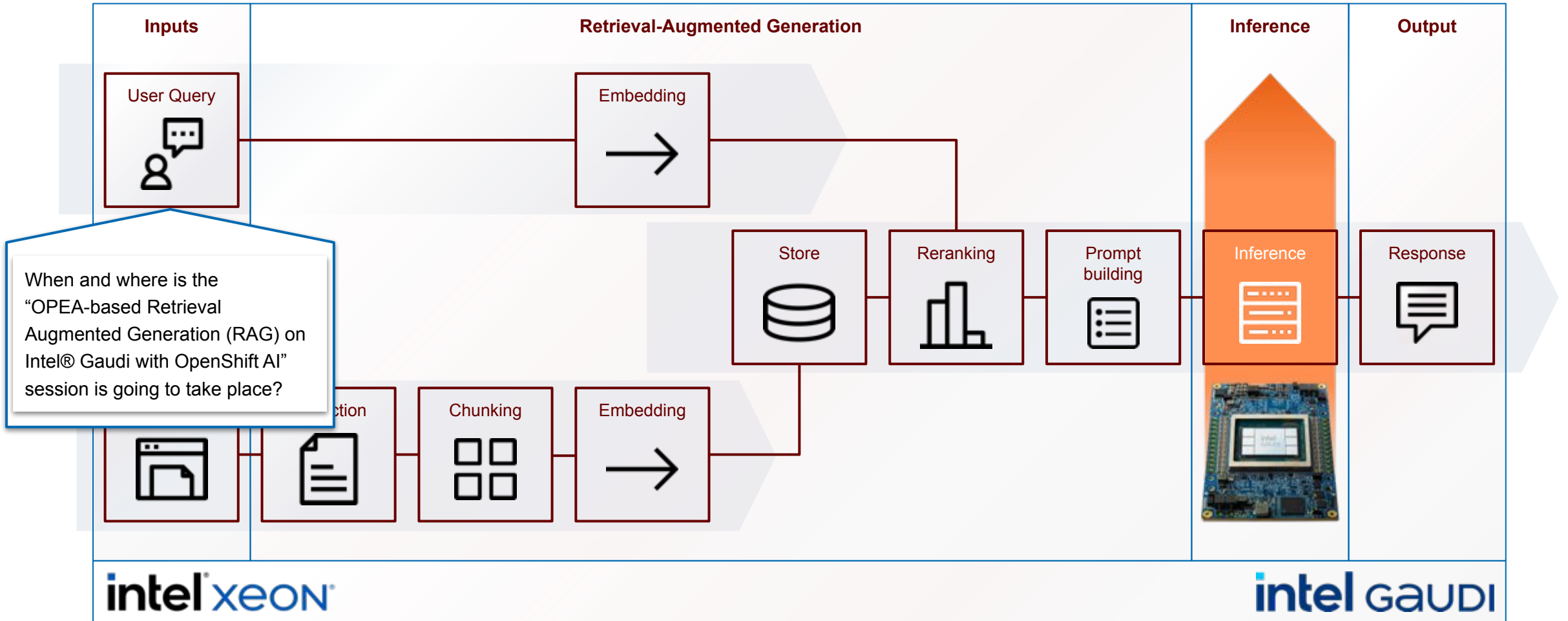
Retrieval Augmented Generation (RAG)



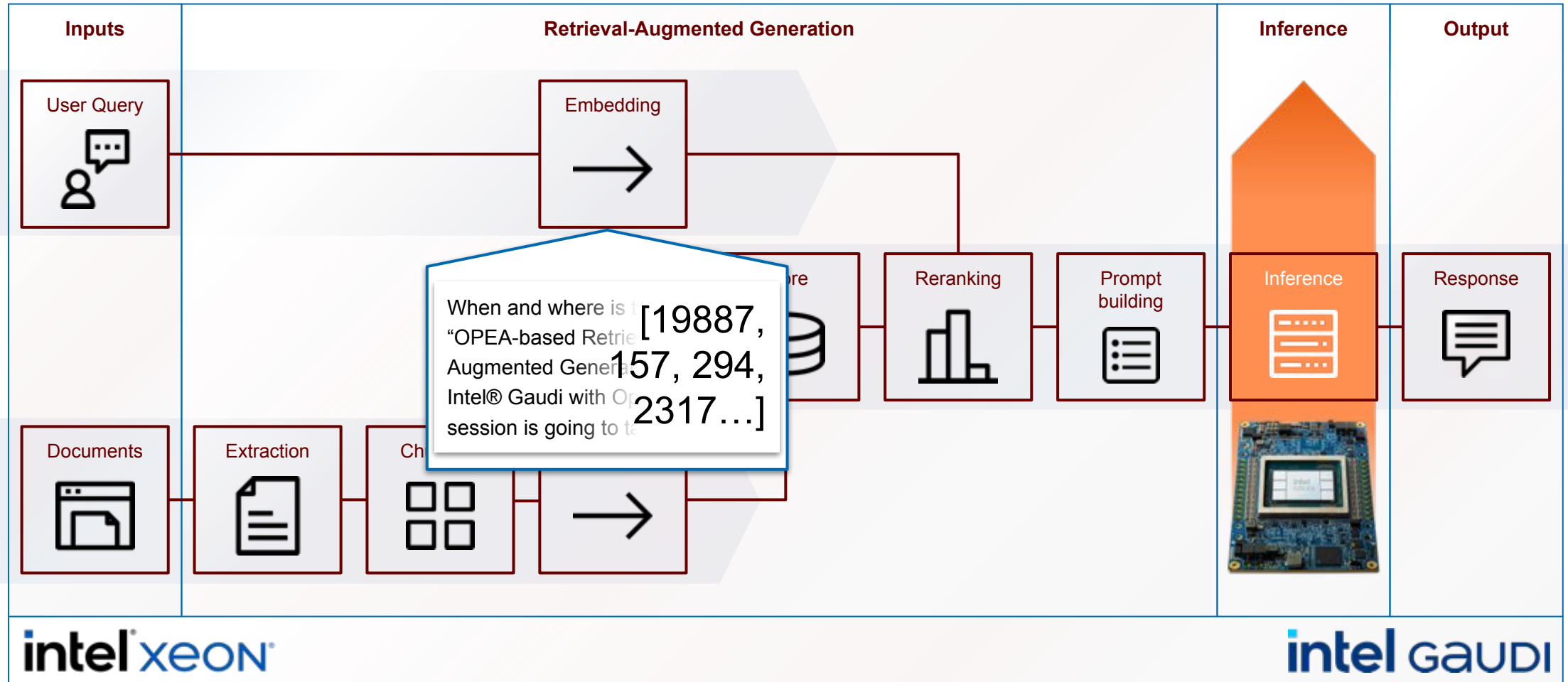
Retrieval Augmented Generation (RAG)



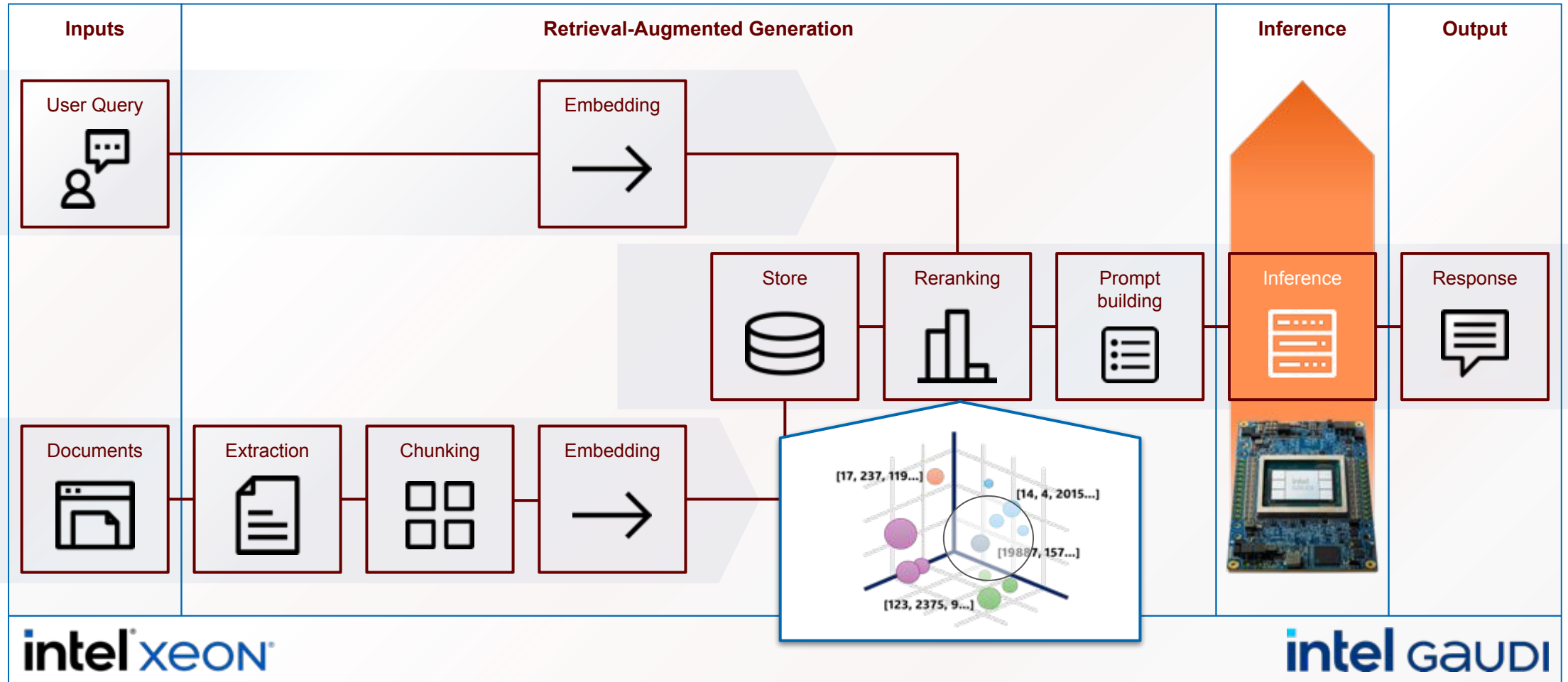
Retrieval Augmented Generation (RAG)



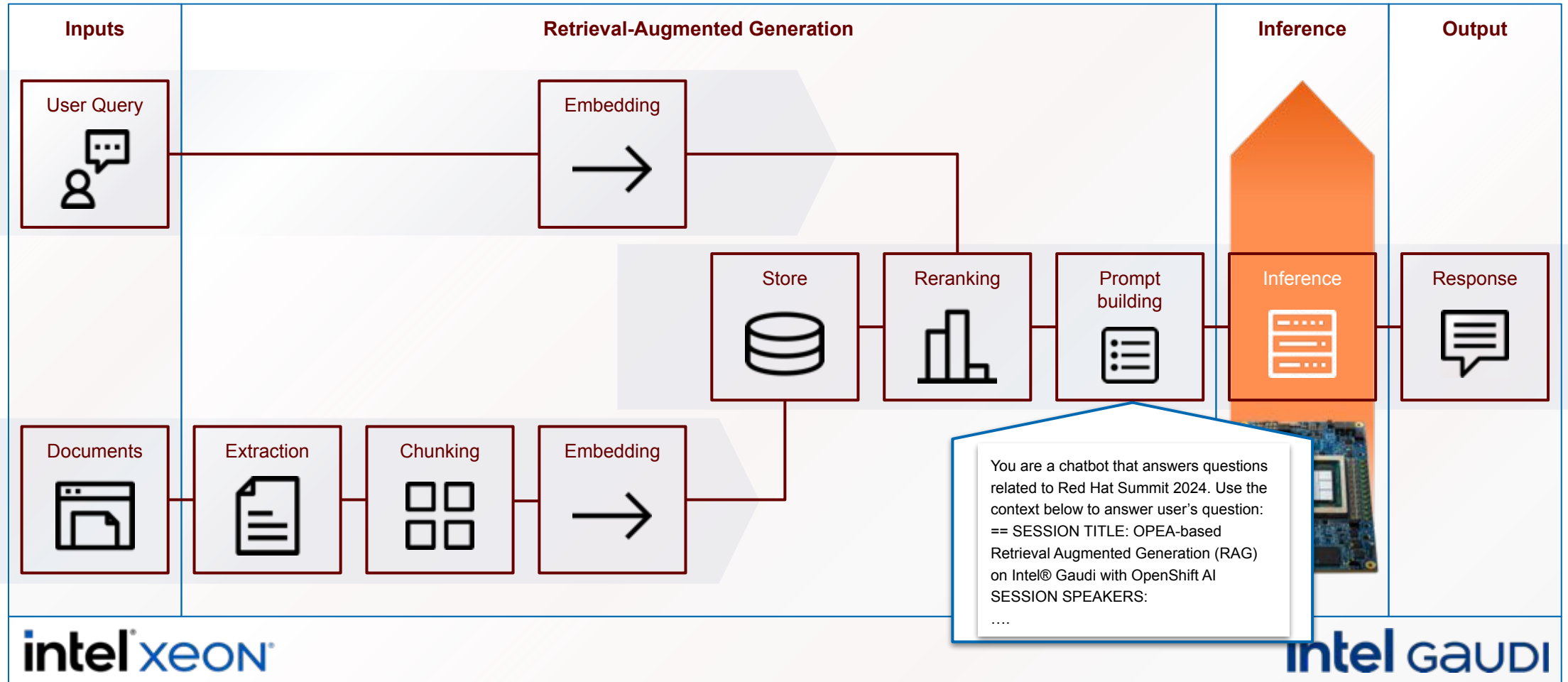
Retrieval Augmented Generation (RAG)



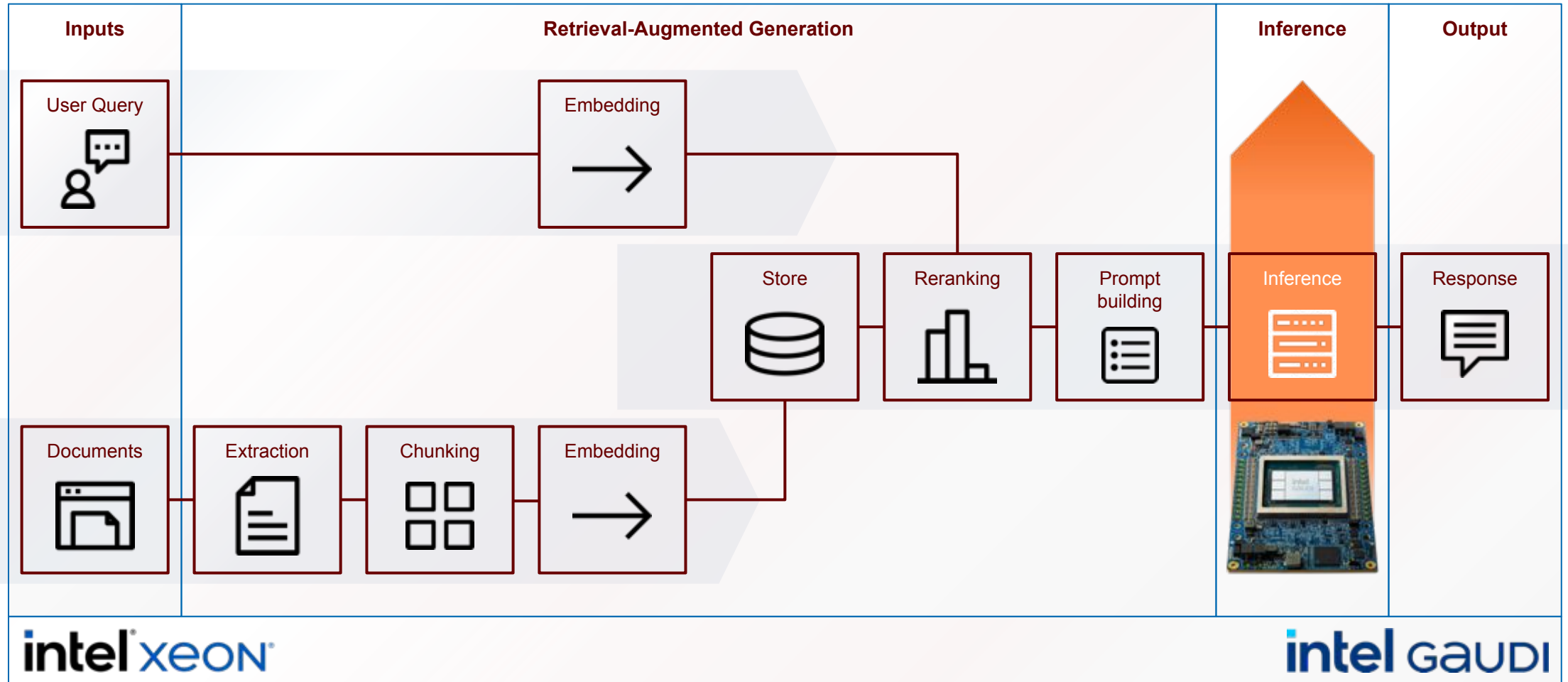
Retrieval Augmented Generation (RAG)



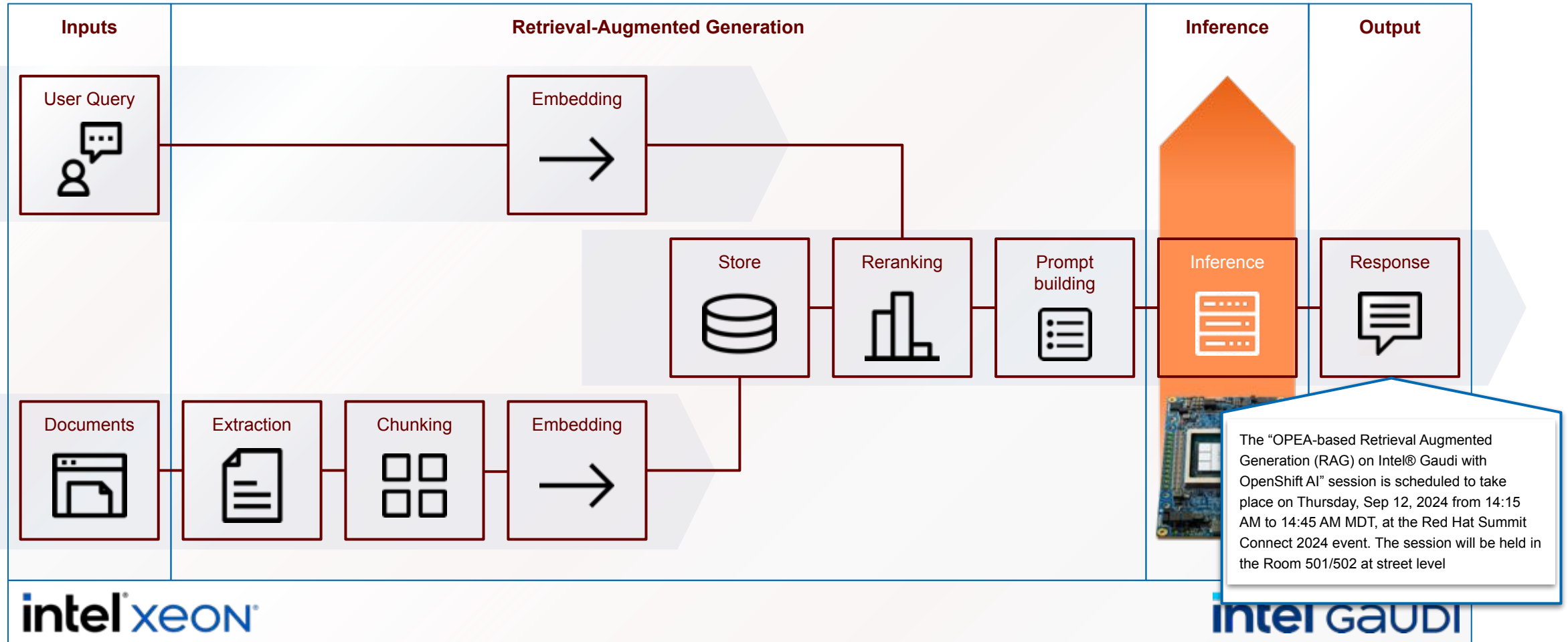
Retrieval Augmented Generation (RAG)



Retrieval Augmented Generation (RAG)

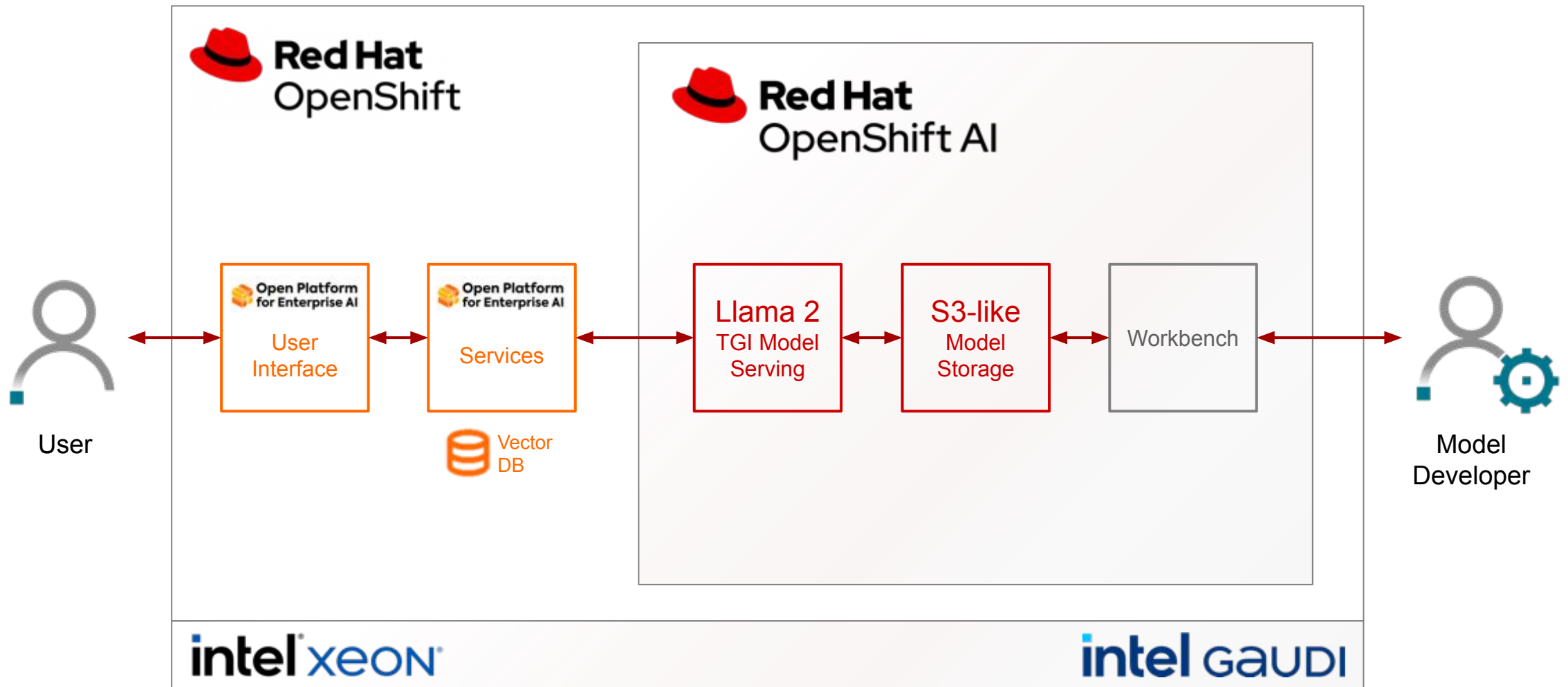


Retrieval Augmented Generation (RAG)



Retrieval Augmented Generation (RAG) Chatbot Demo

Attenzione: video originale non disponibile nella versione .pdf



Summary

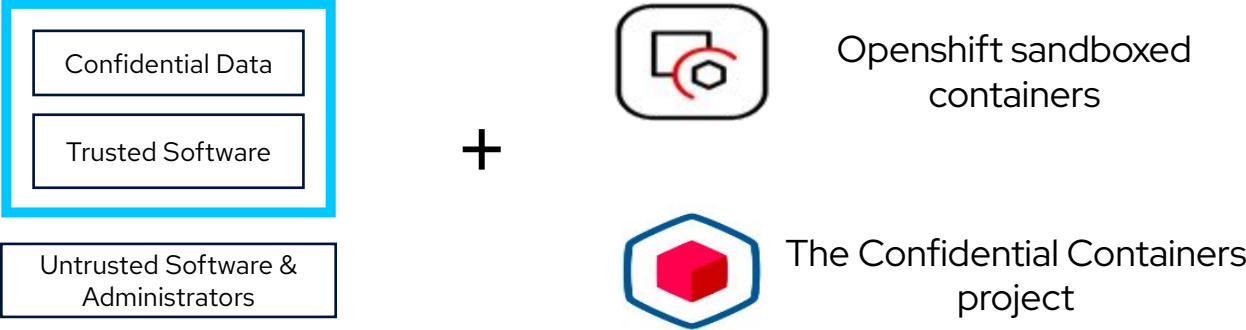
Key Takeaways

- ▶ 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.

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

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

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

