Red Hat Summit

Connect

RAG mit dem Llama Stack leicht gemacht

Stefan Bergstein
Senior Principal
Chief Architect

Rene Schramowski
OpenShift Sales Specialist

Tala Ismail
Account Solution Architect

Hakki Kayali
Associate Account
Solution Architect

Today's discussion

- Introductions
- Red Hat OpenShift AI - quick introduction
- Introduction to Llama Stack & RAG
- Hands-On Workshop
- Wrap-up

Stefan Bergstein

Chief Architect

Red Hat

René Schramowski

Platform Specialist
Red Hat

Tala Ismail

Account Solution Architect
Red Hat

Hakki Kayali

Account Solution Architect
Red Hat

Red Hat OpenShift AI – quick introduction

Red Hat's AI/ML engineering is 100% open source

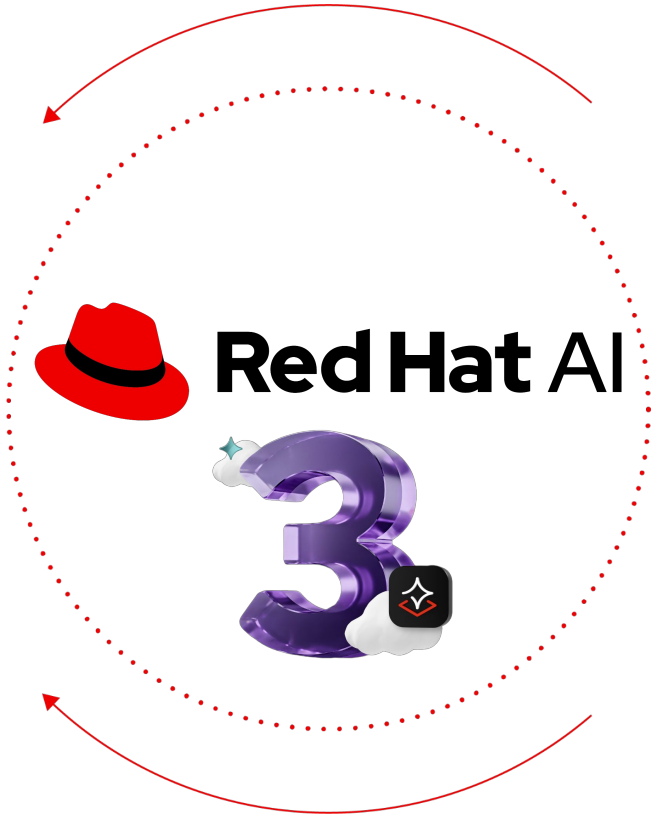
Red Hat

Flexible and Efficient Inference

- GA distributed inference (llm-d)
- New validated and optimized models
- vLLM enhancements
- LLM Compressor GA

Connecting Models to Data

- Modular and extensible approach for: data ingestion, synthetic data generation, tuning, evaluations.
- RAG enhancements & partner integrations
- Feature Store GA



Agentic AI

- AI experiences: AI hub and gen AI studio
- Model Context Protocol support & MCP Server access in gen AI studio
- Llama Stack API integration

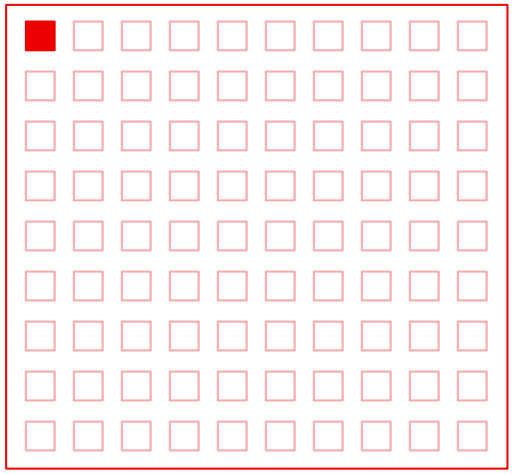
AI Platform

- Model catalog and registry GA
- Model as a Service provider enhancements and API Mgt integration
- GPU as a Service enhancements

Single platform to run any model, on any accelerator, on any cloud

Enterprises need models aligned to their private data

LLMs are trained with a range of public data, not enterprise-relevant data

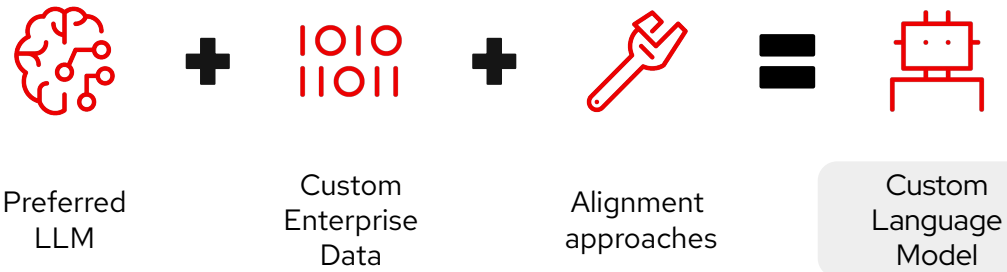


Less than 1% of all enterprise data is represented in foundation models

Enterprise organizations need to

1. Start from a trusted base model
2. Create a new representation of their data
3. Deploy, scale, and create value with their AI

Customize your preferred model using enterprise data to build an efficient, cost-effective solution.



Red Hat AI provides:

- ✓ Validated and optimized models ready-to-use
- ✓ Data ingestion capabilities
- ✓ Synthetic data generation pipelines
- ✓ Multiple alignment techniques

Red Hat

Red Hat AI repository on Hugging Face

Collection of third-party models

Llama

Qwen

Gemma

Mistral, Voxtral

DeepSeek

Microsoft Phi

Molmo

Granite

NVIDIA Nemotron

OpenAI GPT-oss

KIMI K2

SMOLI M3 3B

Choice of Models

- Transformers (Dense, MOE), Multi-modal LLMs, Embeddings Models, Hybrid / Novel Attention, Vision
- Hugging Face compatible (safe tensors), OCI-compatible containers

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

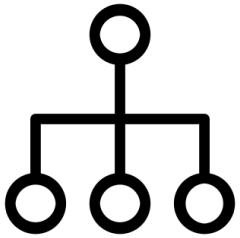
What's next: multiple approaches?

Build customized AI solutions that address domain specific business cases

Coming soon

Prompt design

Prompt tuning and engineering



Design and engineer the prompts to enhance GenAI model responses and achieve more specific and accurate outcomes.

Enhanced

RAG

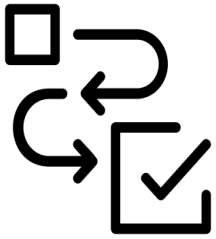
Retrieval Augmented Generation

Enhance Gen AI model generated text by retrieving relevant information from external sources, improving accuracy and depth of model's responses.

Enhanced

Fine tuning

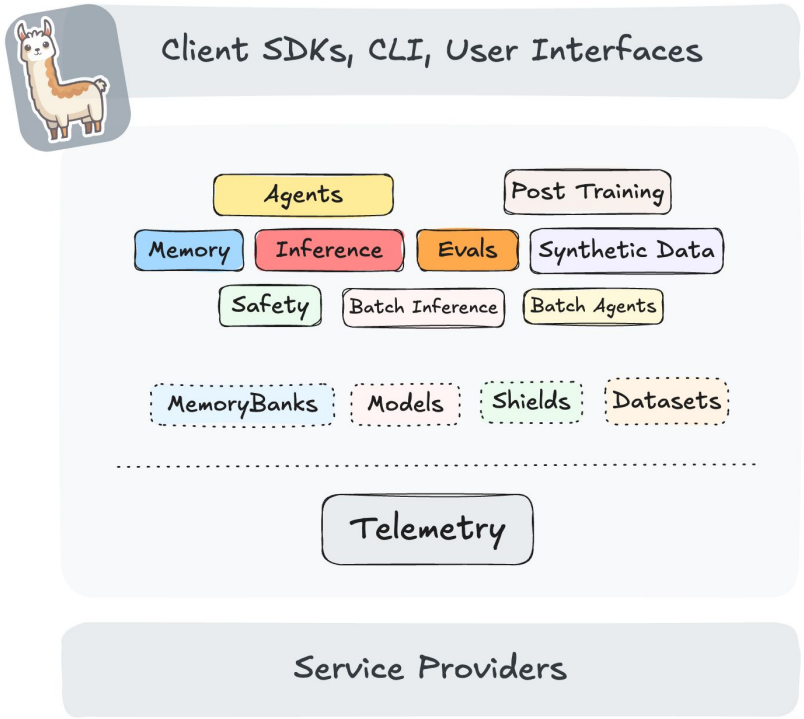
InstructLab, OSFT, LoRA and QLoRA



Customize a base model for specific tasks or private data, using a range of approaches—from full fine-tuning to parameter-efficient methods—to balance performance and efficiency.

Red Hat

Introduction to Llama Stack & RAG



Llama Stack is an open source AI control plane

It gives enterprises a consistent foundation to build and run autonomous AI systems

Llama Stack Overview

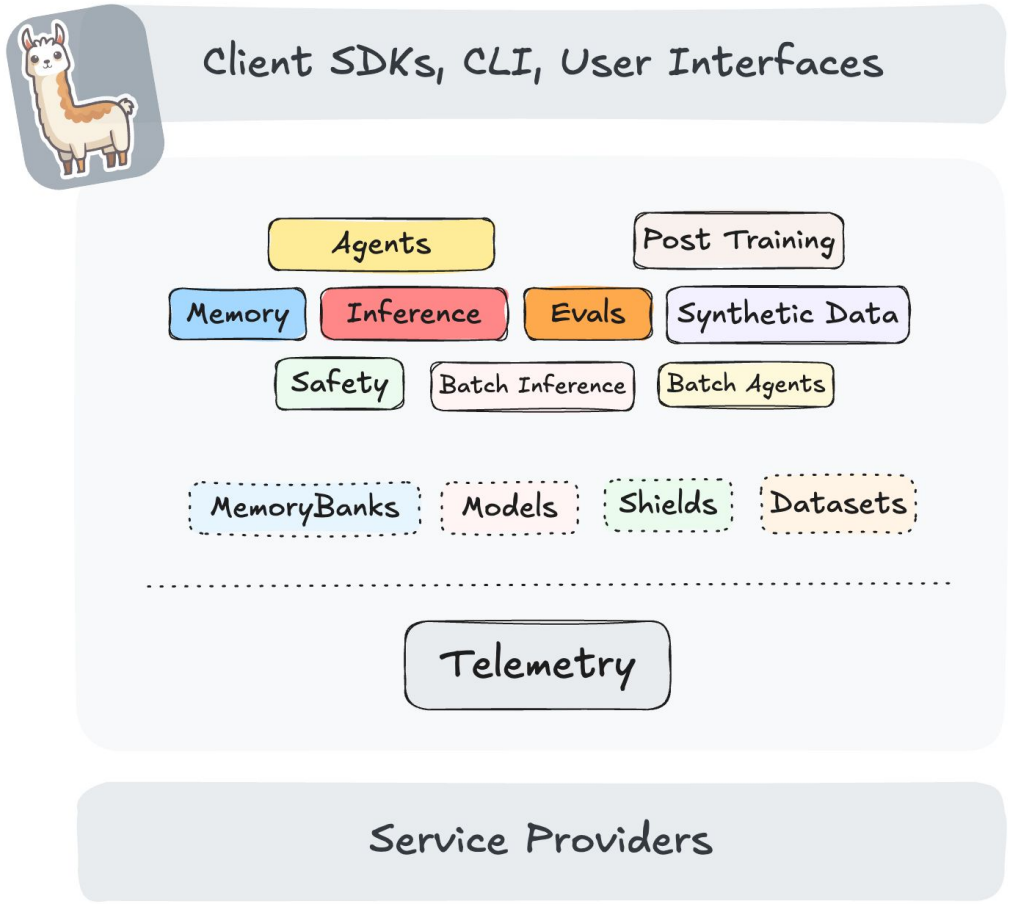
Out of the box, Llama-Stack provides:

Unified APIs: A single interface for inference, RAG (retrieval-augmented generation), agent orchestration, tool calling, safety guardrails, memory management, evaluation, and telemetry.

Plug-and-Play Distributions: Pre-packaged distributions that let you start locally (e.g., using Ollama) and later move to production (e.g., cloud or on-prem deployments) without changing your code.

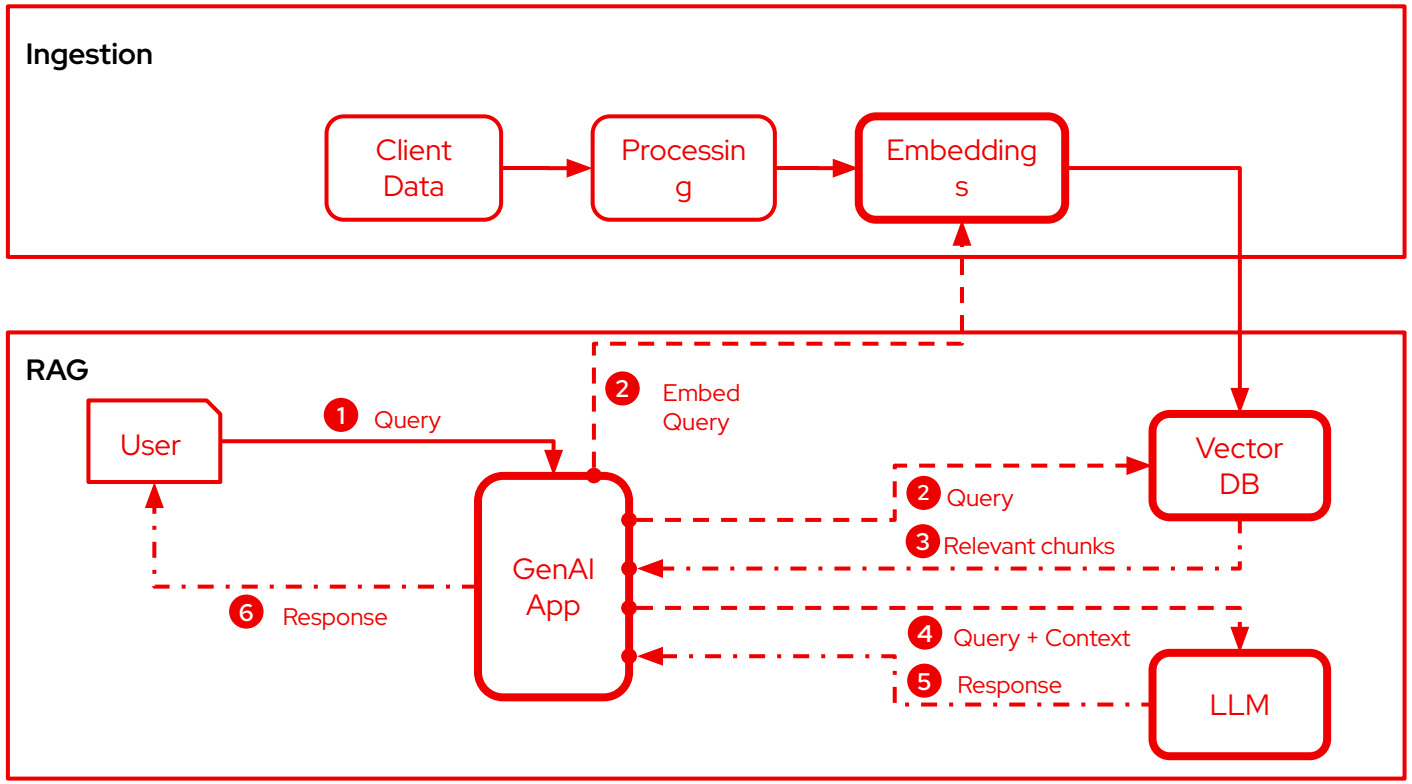
Extensibility: A plugin architecture that supports multiple providers (Meta Reference, Together, Fireworks, etc.), making it flexible for various deployment scenarios.

Built-in Safety & Monitoring: Integrated safety features (like Llama Guard) and observability tools to monitor model performance and usage.

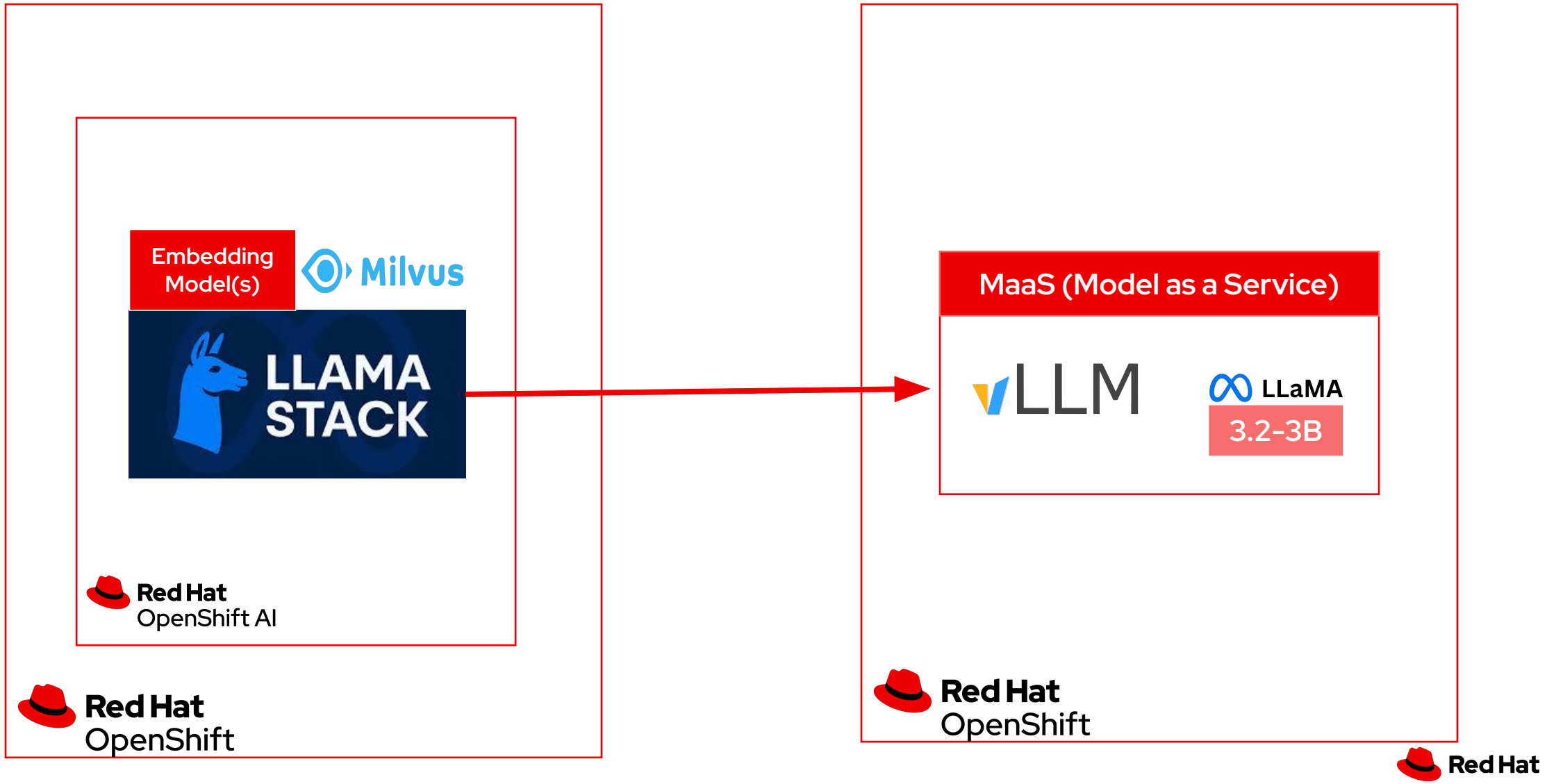


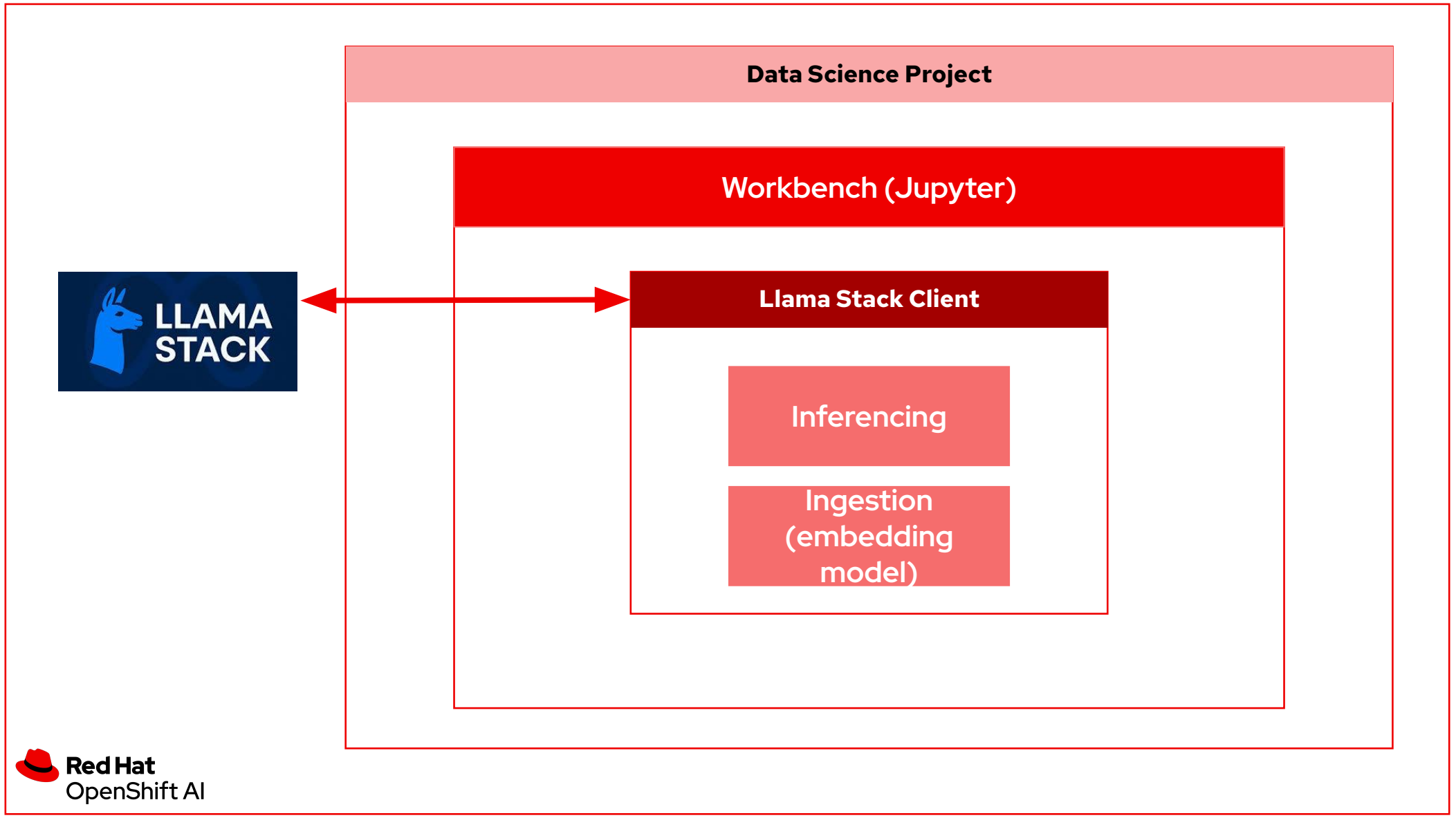
How does RAG (Retrieval Augmented Generation) work?

- RAG **improves LLM answers** by adding external knowledge
- It retrieves relevant info from a data source and provides it to the model
- Helps the LLM give **accurate, up-to-date responses**



Workshop Time!


Embedding Model(s)
Milvus
LLAMA STACK
Red Hat OpenShift AI
Red Hat OpenShift
MaaS (Model as a Service)
vLLM
LLaMA
3.2-3B
Red Hat OpenShift
Red Hat


Data Science Project
Workbench (Jupyter)
LLAMA STACK
Llama Stack Client
Inferencing
Ingestion (embedding model)
Red Hat OpenShift AI

Red Hat

Call to Action

Lab Link:

red.ht/raglab

Repo Link:

red.ht/raglab-repo

Red Hat Summit

Connect

Thank you


linkedin.com/company/red-hat


facebook.com/redhatinc


youtube.com/user/RedHatVideos


twitter.com/RedHat