CLOUD NATIVE & KUBERNETES

AI DAY

NORTH AMERICA

CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA

Advancing Cloud Native AI Innovation Through Open Collaboration

Yuan Tang

Principal Software Engineer

Open Source Leadership

Red Hat

Kubeflow

KServe

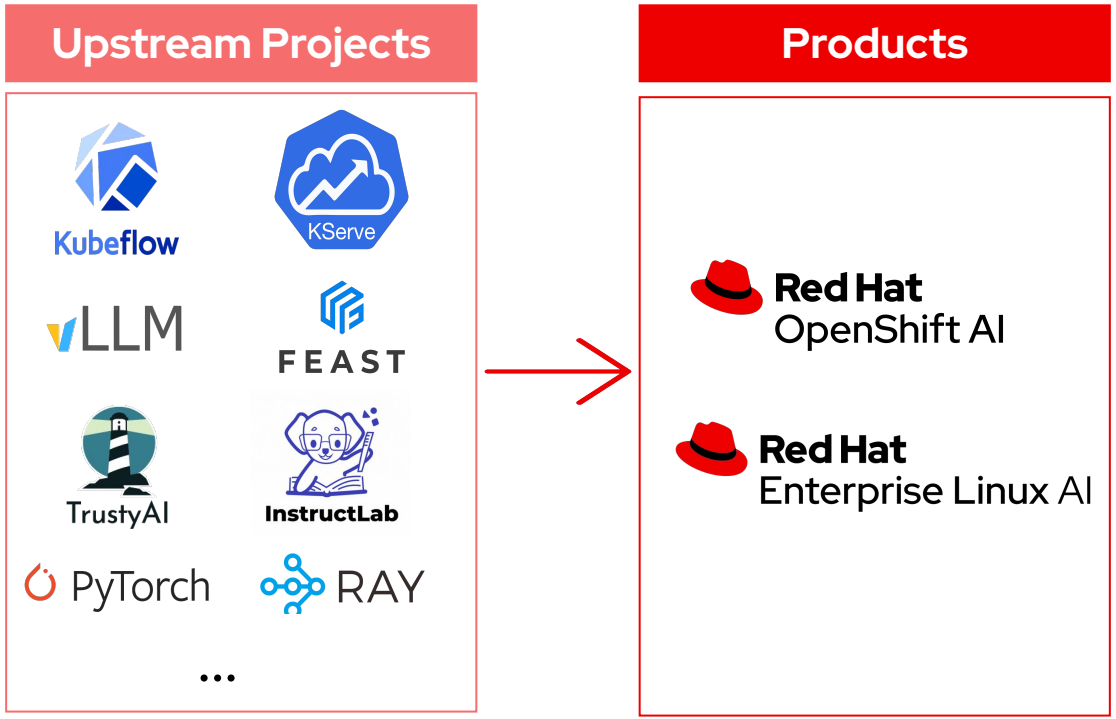
WG Serving

argo

CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA

Our Commitment to Open Source

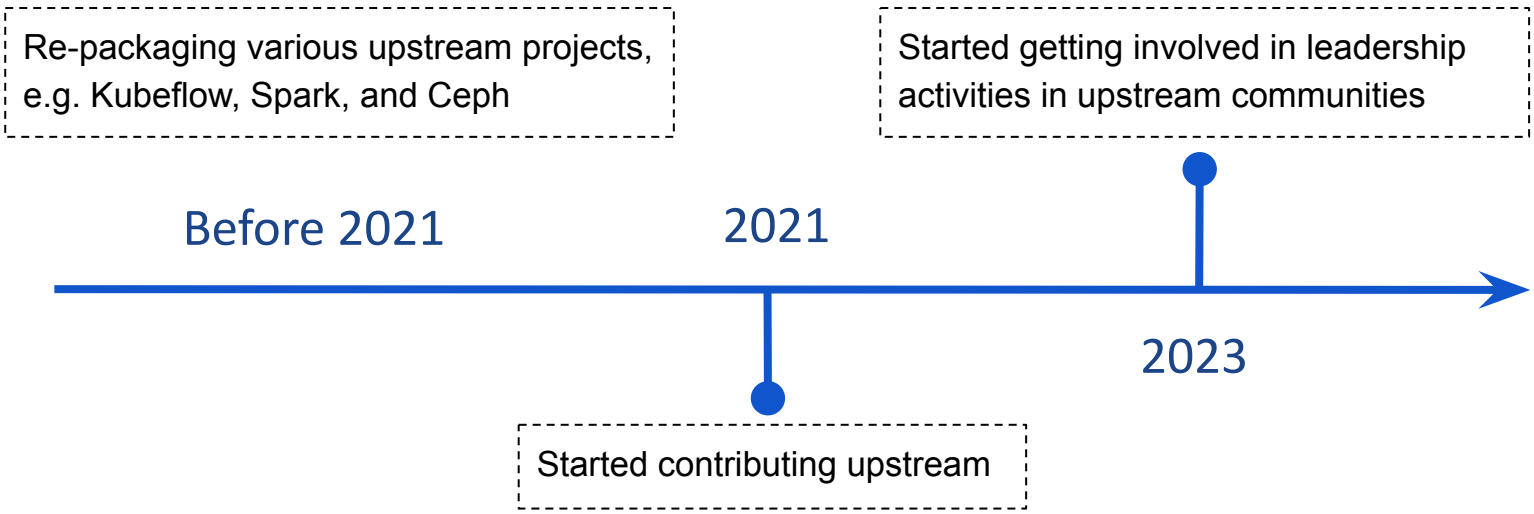
Red Hat AI = 100% Open Source



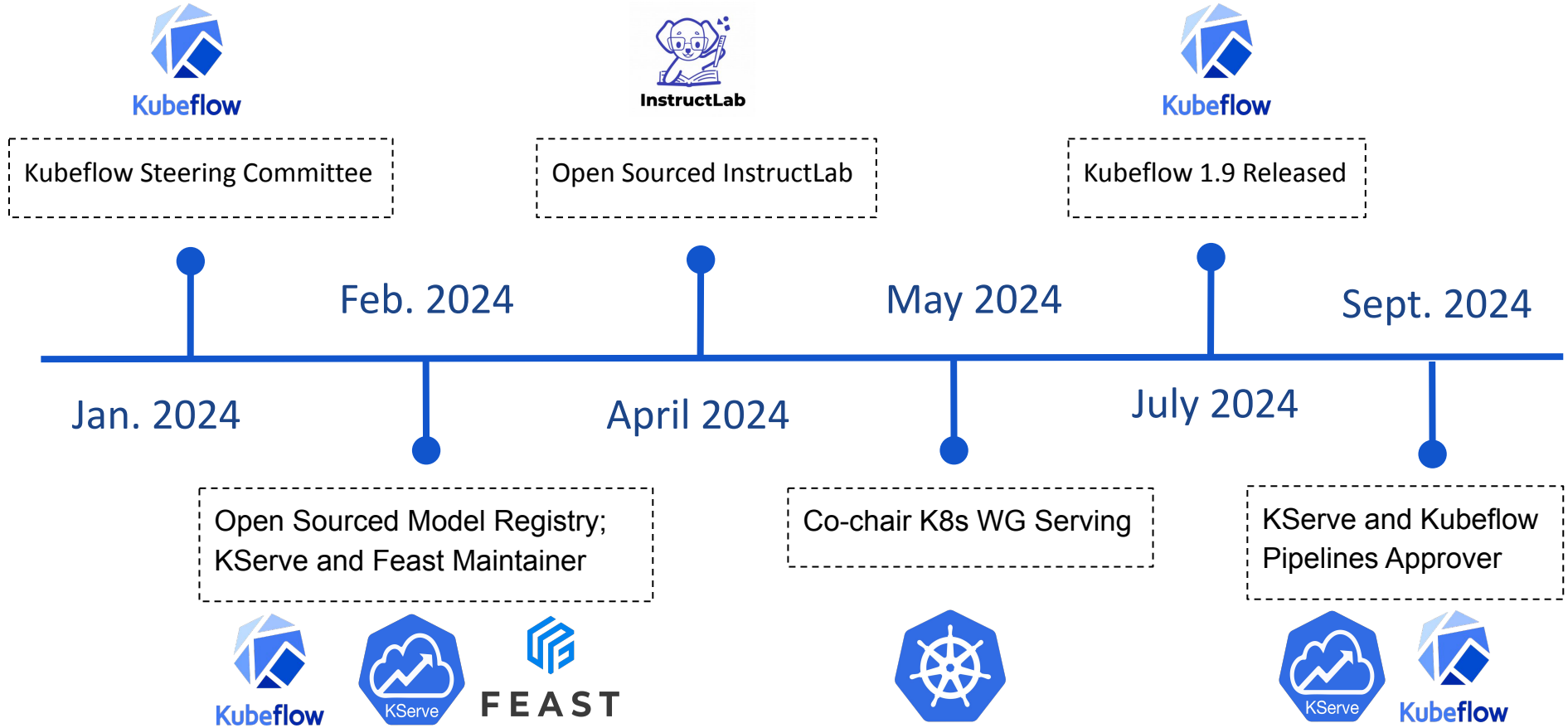
CLOUD NATIVE & KUBERNETES
AI DAY
NORTH AMERICA

Red Hat Open Source AI Journey

Red Hat Open Source AI Journey



Red Hat Open Source AI Journey



Open Source Contributions

Workflow

Kubeflow

argo

Elyra

Data

FEAST

InstructLab

Training & Tuning

Kubeflow

RAY

Kueue

InstructLab

CodeFlare

Evaluation

TrustyAI

InstructLab

Serving

KServe

Kubeflow

TrustyAI

kubernetes

InstructLab

vLLM

K n

Dashboard & IDE

Kubeflow

Elyra

CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA

Contribution Highlights

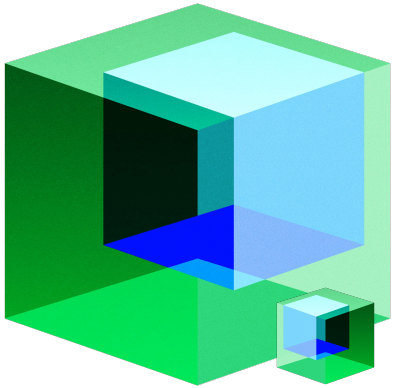
Contribution Highlights: InstructLab

InstructLab is **an open source project** for enhancing large language models **(LLMs)** used in generative AI applications.

Accessible ◆ Cost-Efficient ◆ Open Source

Contribution Highlights: Granite Models

New Granite 3.0 Language Models under Apache 2.0



New dense architecture

State-of-the-art training and data recipes

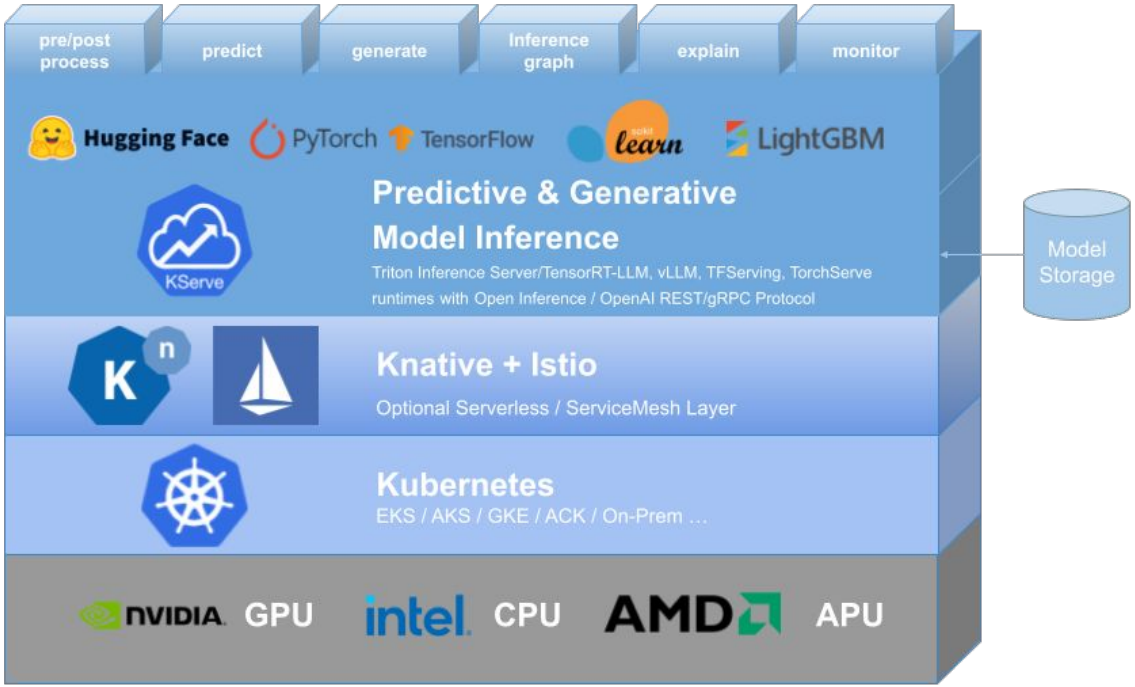
12T+ tokens training data

Designed for enterprise tasks:

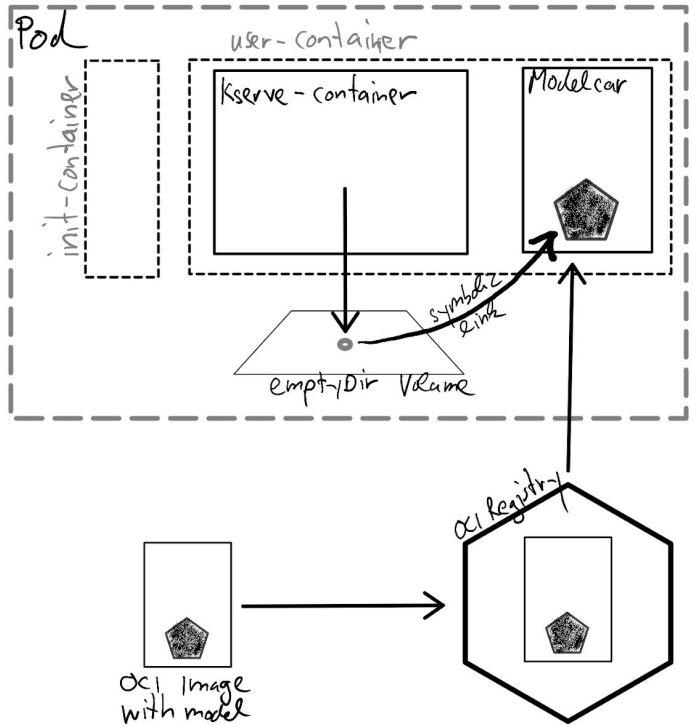
→ Language (RAG, summarization, entity extraction, classification, etc.)
→ Code (generation, translation, bug fixing)
→ Agents (tool use, advanced reasoning)
→ Multilingual support (en, de, es, fr, ja pt, ar, cs, it, ko, nl, zh)

Contribution Highlights: Faster Model Serving

KServe



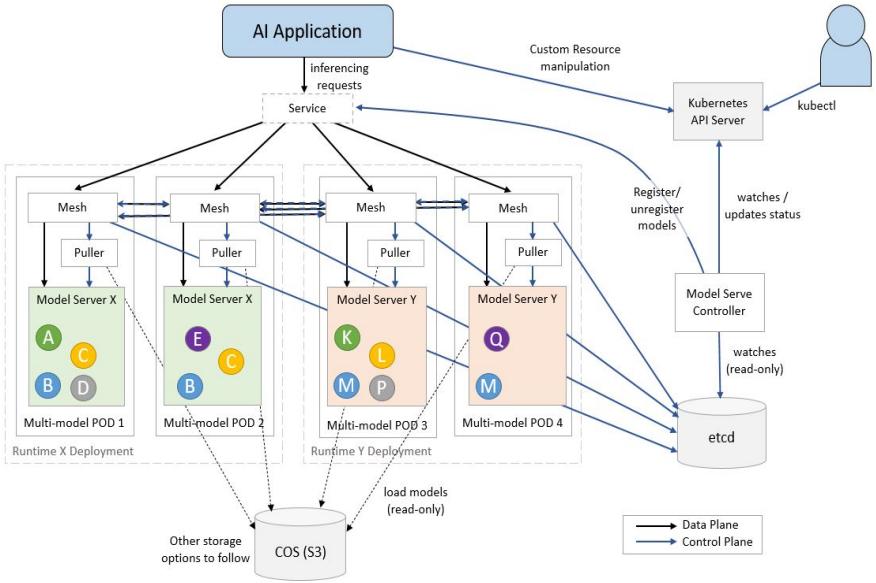
KServe ModelCar



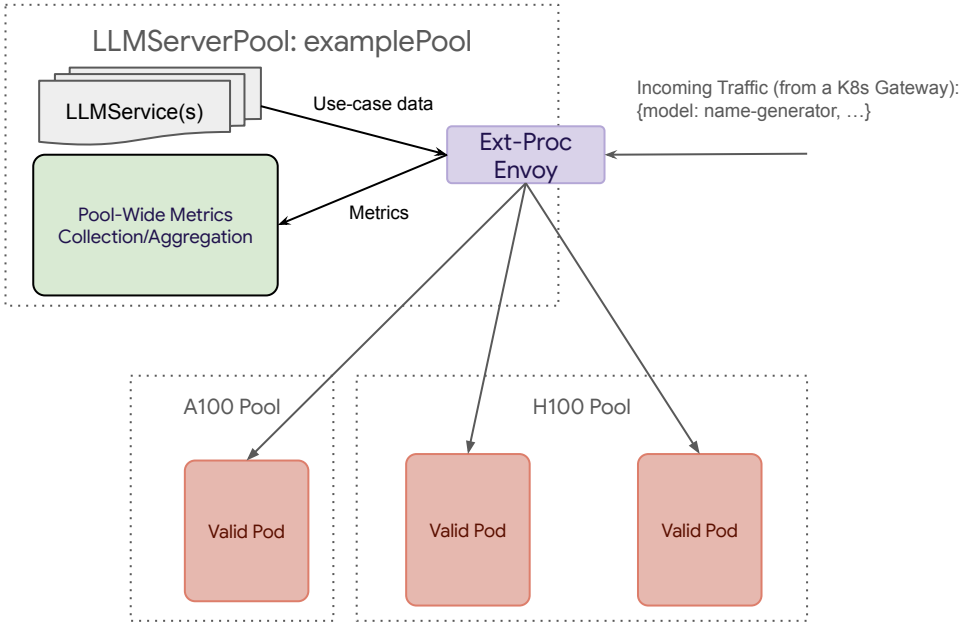
Contribution Highlights: Faster Model Serving

CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA

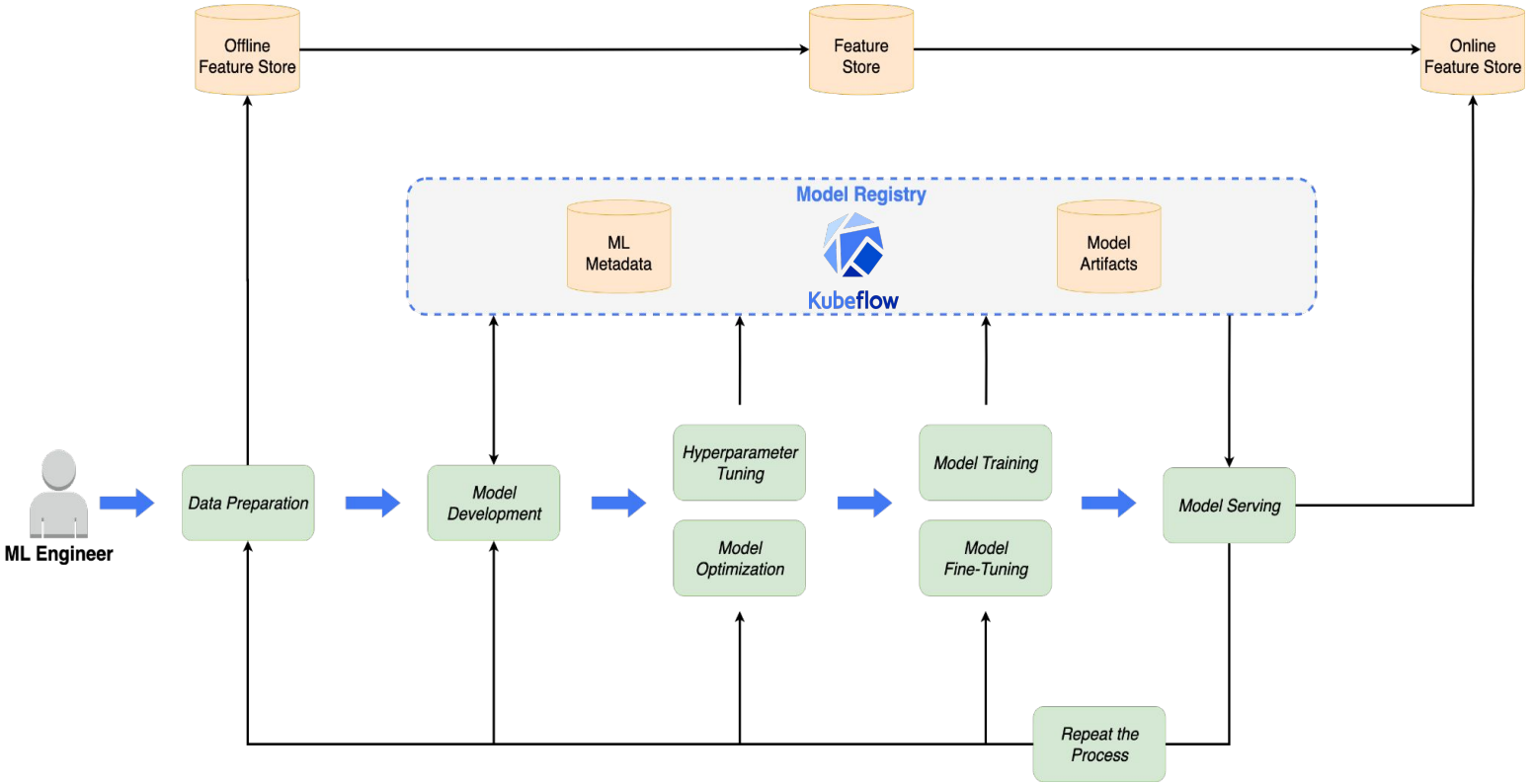
KServe ModelMesh



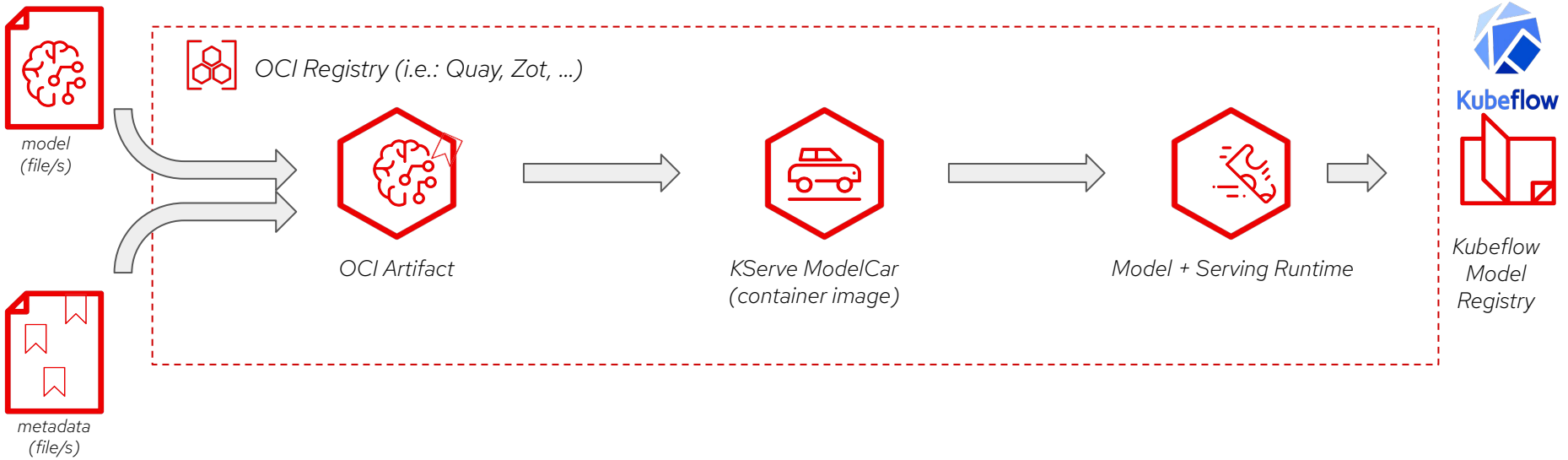
LLM Instance Gateway



Contribution Highlights: Better Model Management



Contribution Highlights: Better Model Management



Contribution Highlights: Safer Model Usage

TrustyAI

Core Features

TrustyAI REST Service

Compute bias and drift metrics

LLM Evaluation

Benchmark LLM capabilities over a variety of tasks

LLM Guardrails

Moderate interaction pathways between users and LLMs

CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA

Our Collaborative Approach

Our Collaborative Approach

- **Mentoring** next generation of contributors
  - 8 successful Kubeflow GSoC projects.
- **Collaborating** with project maintainers, contributors, and corporate partners
  - KServe
    - 15 maintainers from 6 companies and 250+ contributors
    - Engaging the KServe Community, The Impact of Integrating a Solutions with Standardized CNCF Projects | **Thursday** 5:25pm - 6:00pm
  - K8s WG Serving
    - 250+ community members
    - WG Serving: Accelerating AI/ML Inference Workloads on Kubernetes | **Friday** 11:55am - 12:30pm
- **Listening** to feedback from end users and adopters
  - Community surveys and interactions with project end-users
  - Kubeflow and KServe end users will be featured at KubeCon keynotes

Join Our Collaboration!

CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA

Email: terrytangyuan@gmail.com

Social handle: @TerryTangYuan

terrytangyuan.xyz