
KubeCon
CloudNativeCon
Europe 2025
LONDON BUS

KubeCon | CloudNativeCon
Europe 2025

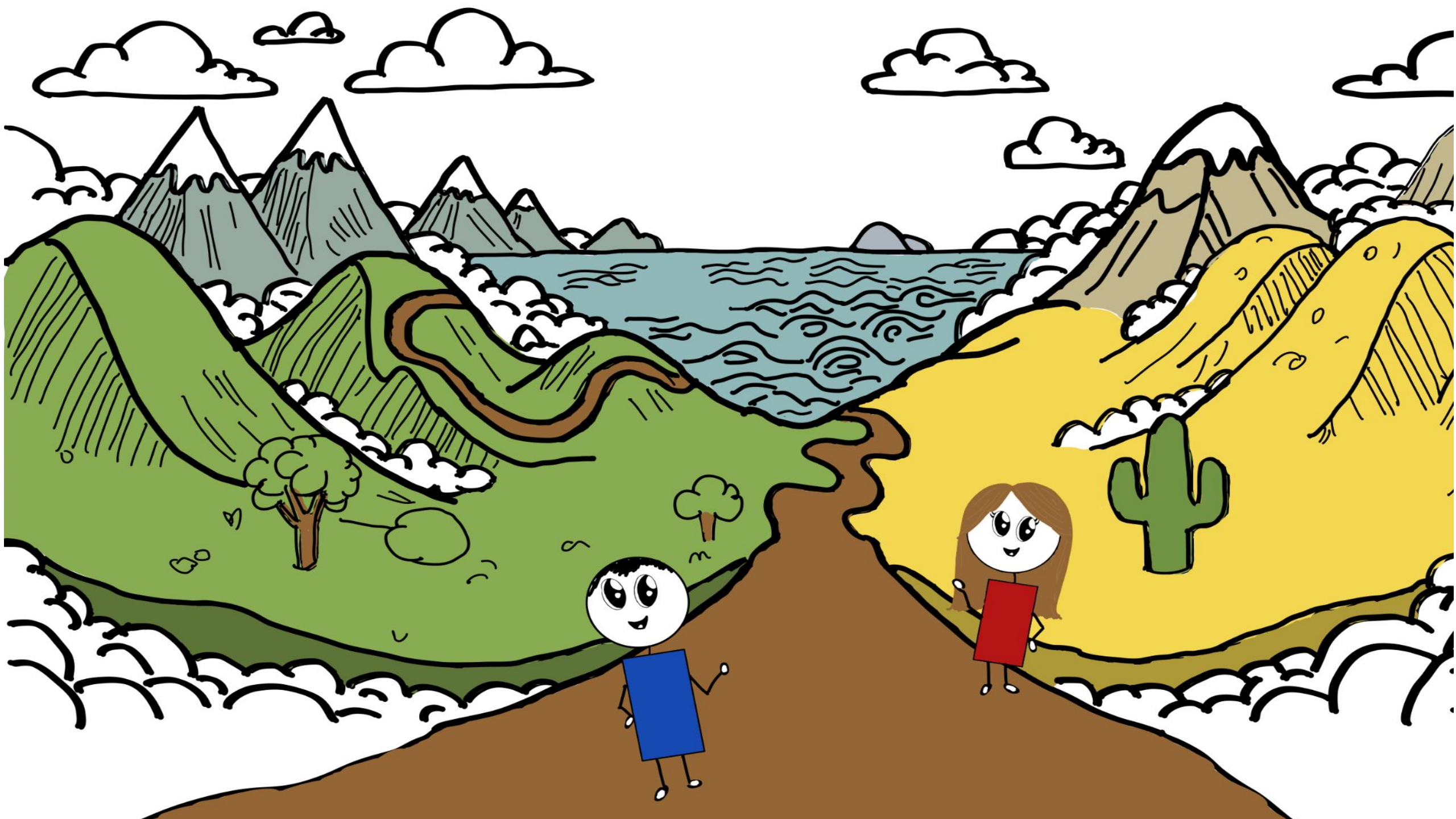
# *The Explorer's Guide To Cloud Native GenAI Platform Engineering*

Alexa Griffith, Bloomberg
Max Körbächer, Liquid Reply

**Alexa Griffith**
Senior Software Engineer
Bloomberg

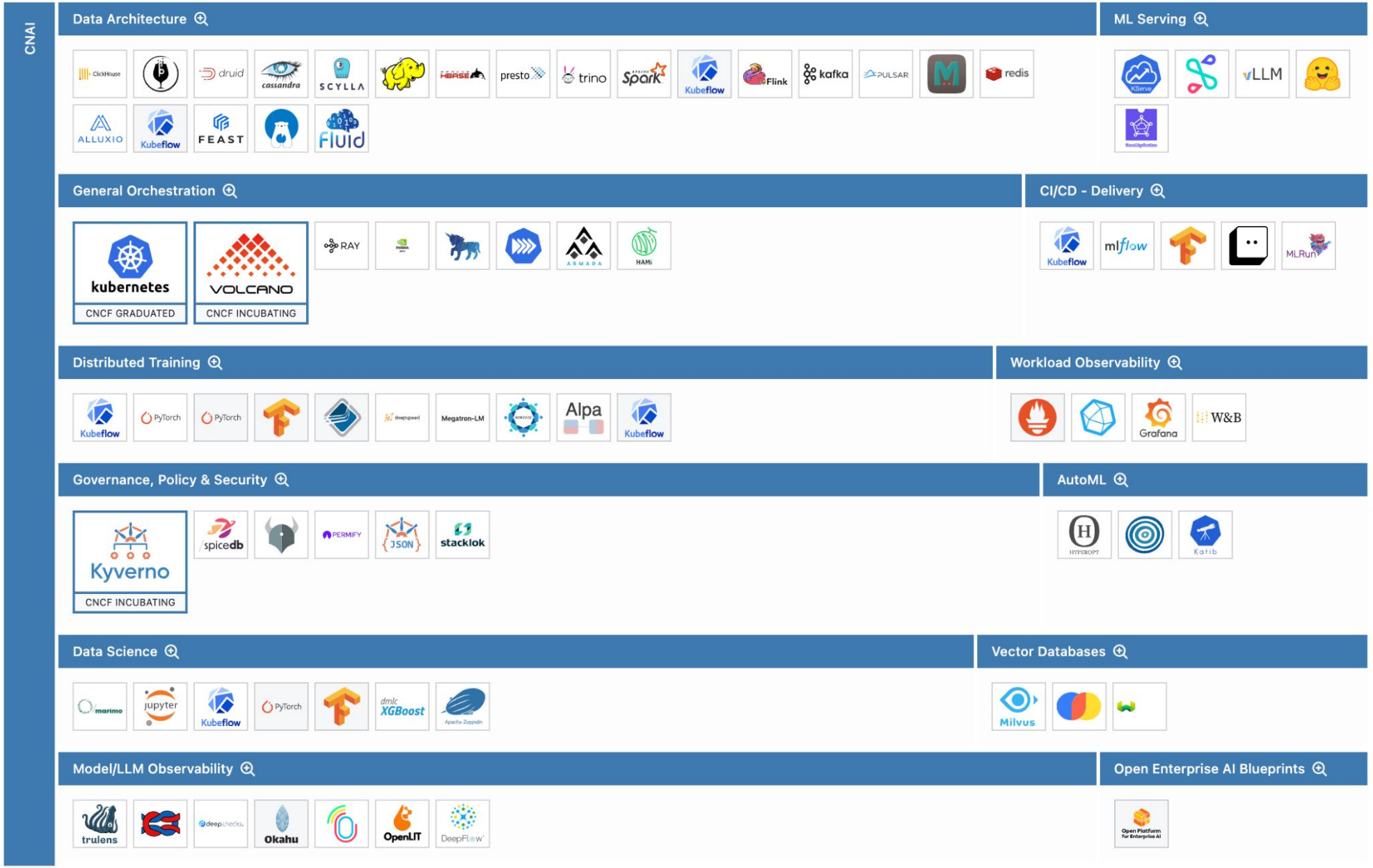
**Max Körbächer**
Cloud Native Advisor & Managing Director
Liquid Reply

KubeCon | CloudNativeCon
Europe 2025



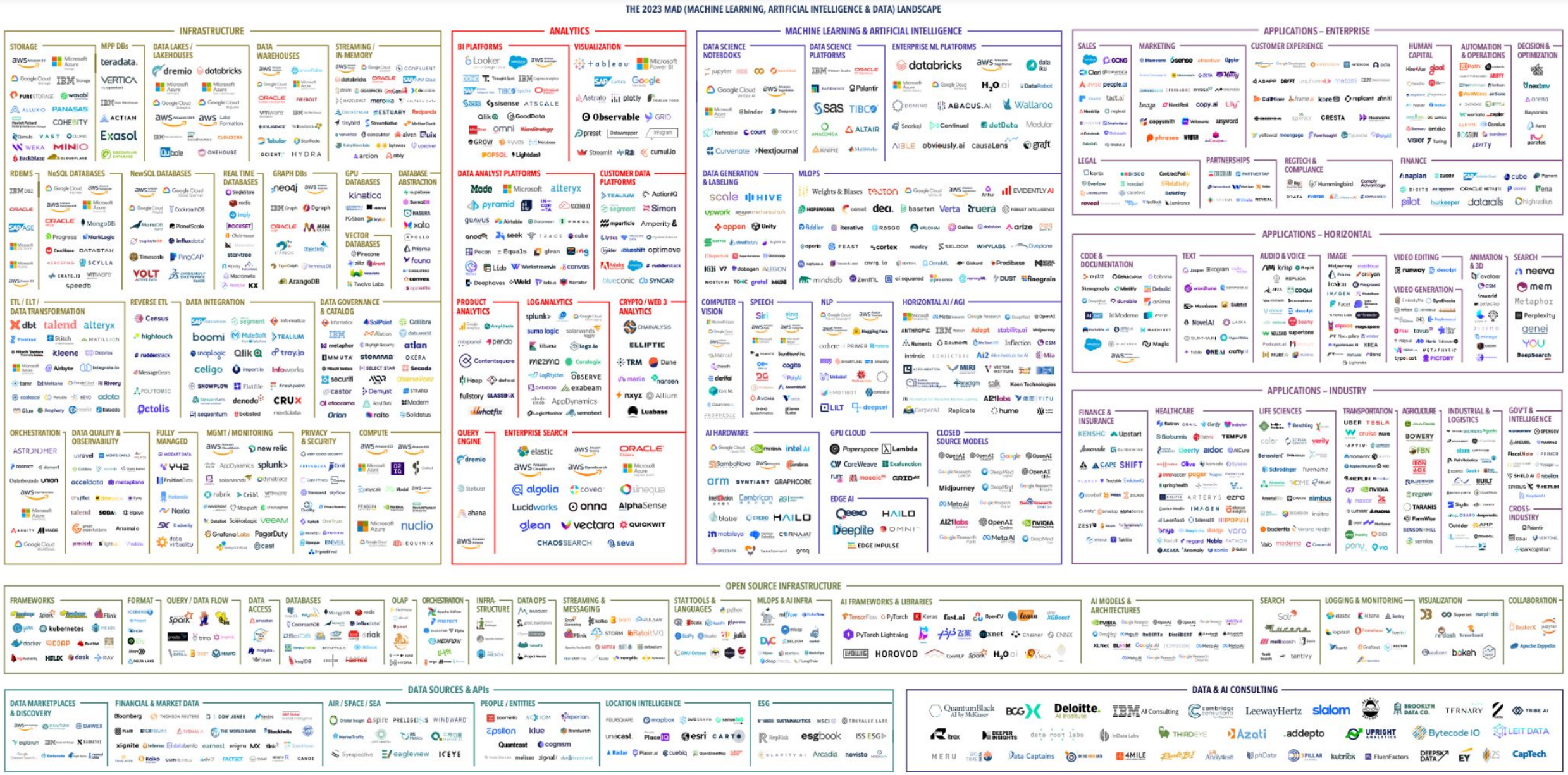
# AI Happens alongside the Cloud Native world

KubeCon CloudNativeCon Europe 2025



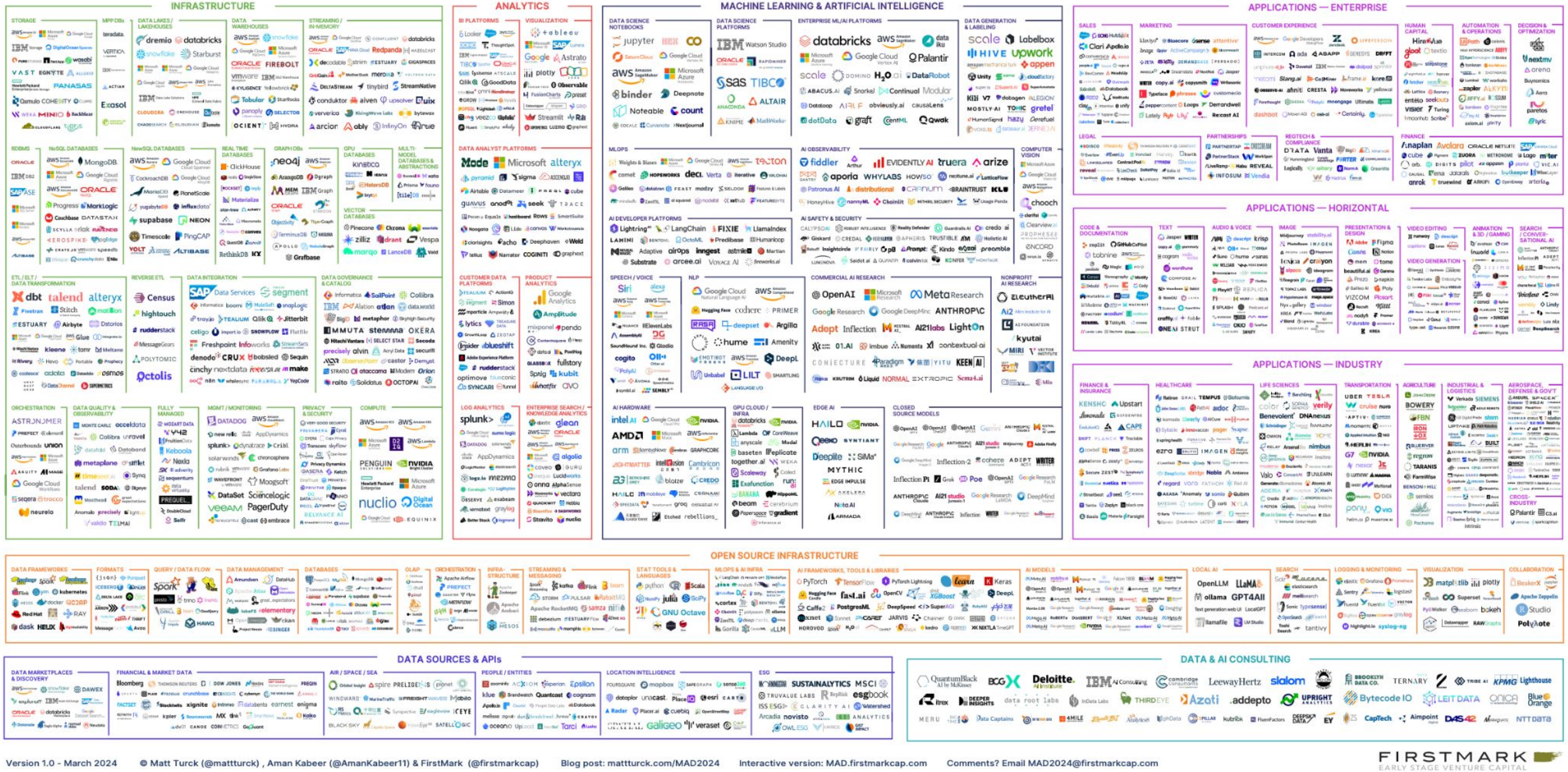
# The 2023 MAD Landscape

KubeCon CloudNativeCon Europe 2025



# The 2024 MAD Landscape

KubeCon CloudNativeCon Europe 2025



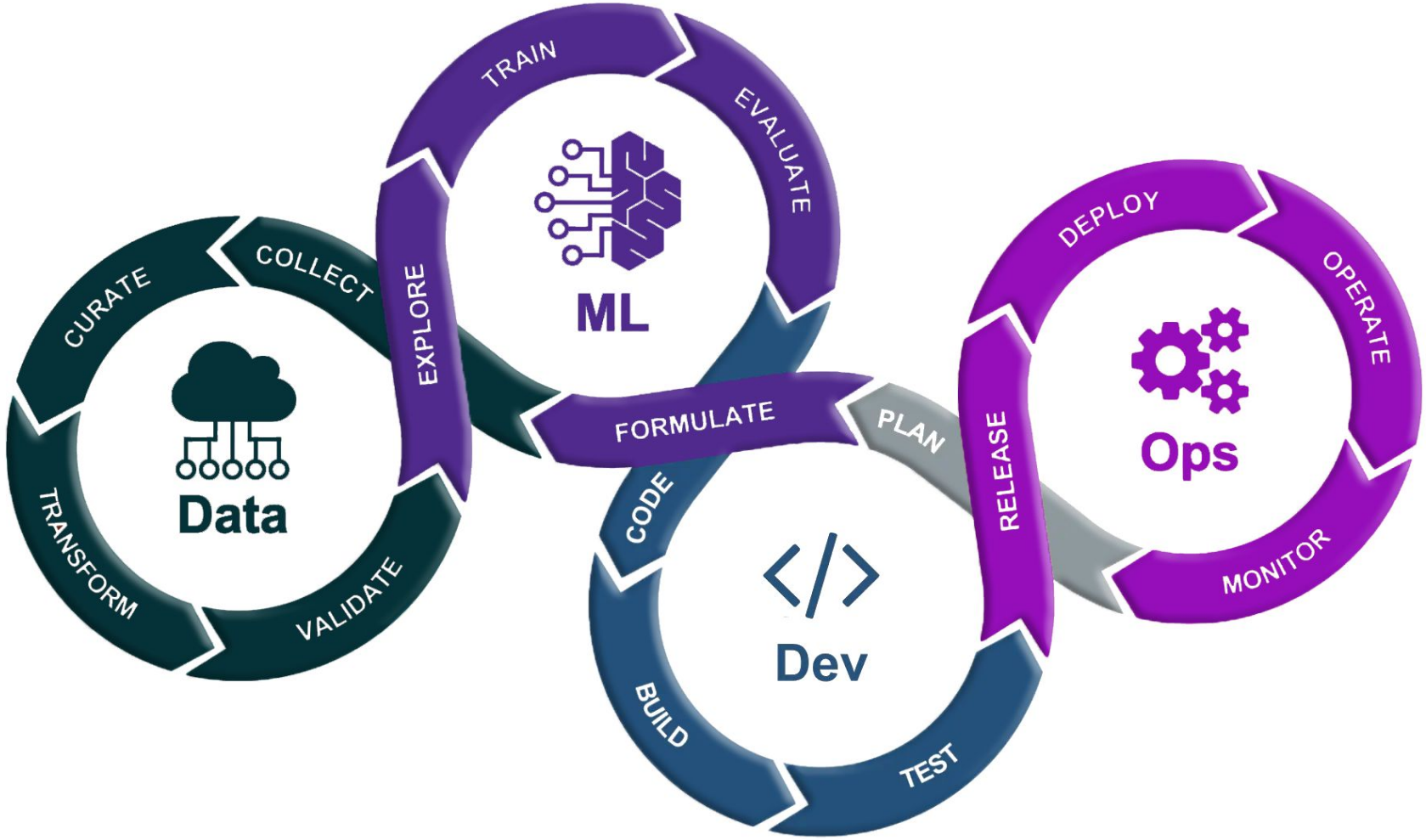

Version 1.0 - March 2024  Blog post: mattturck.com/MAD2024 Interactive version: MAD.firstmarkcap.com Comments? Email MAD2024@firstmarkcap.com

FIRSTMARK EARLY STAGE VENTURE CAPITAL

# Kubernetes is a platform to build platforms

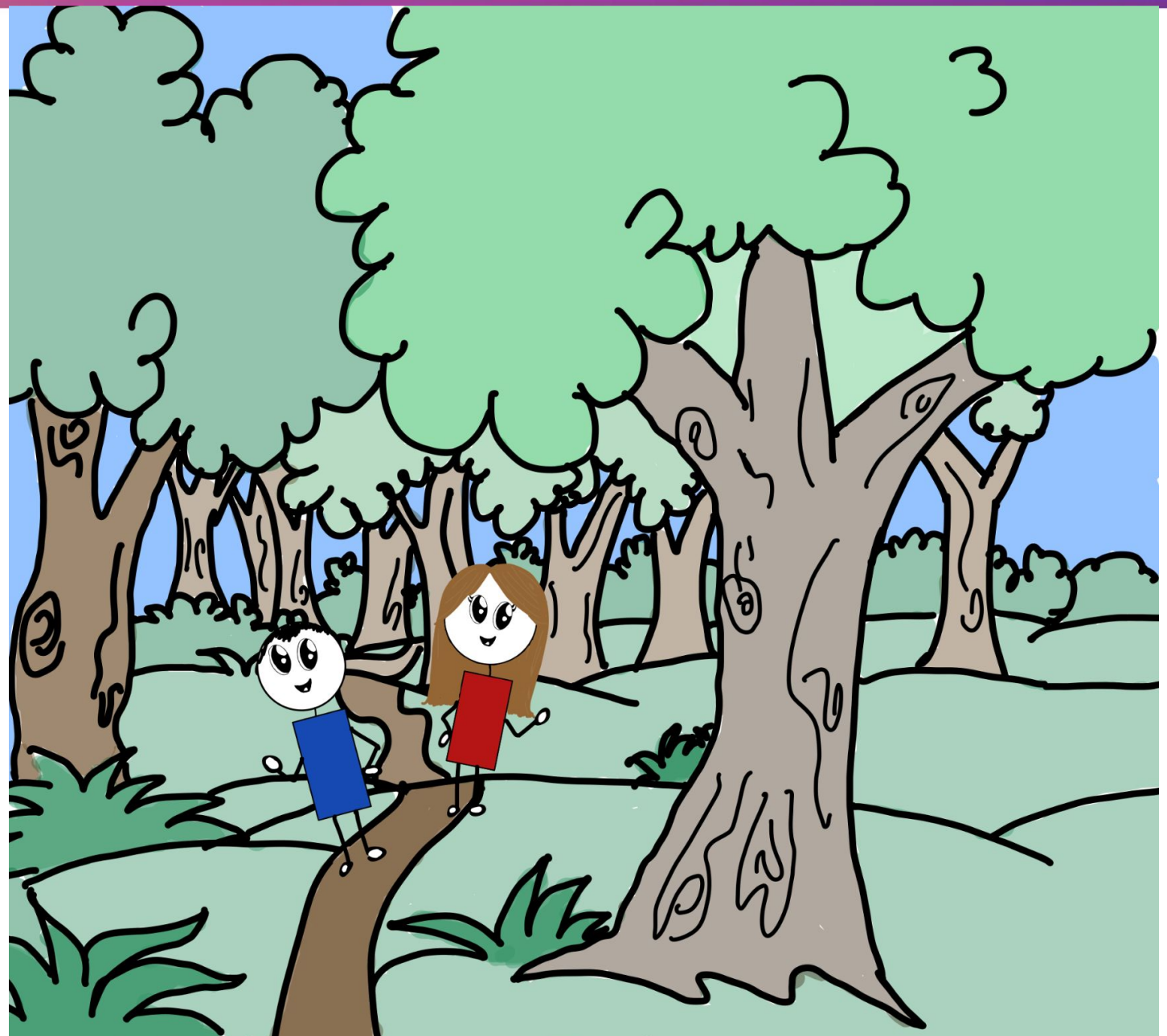
2020 - Infrastructure Abstraction & Automation
2022 - Going to Edge, Telco & Fleets
2023 - Internal Development Platforms (IDP)

**AI/ML/LLM in Cloud Native**

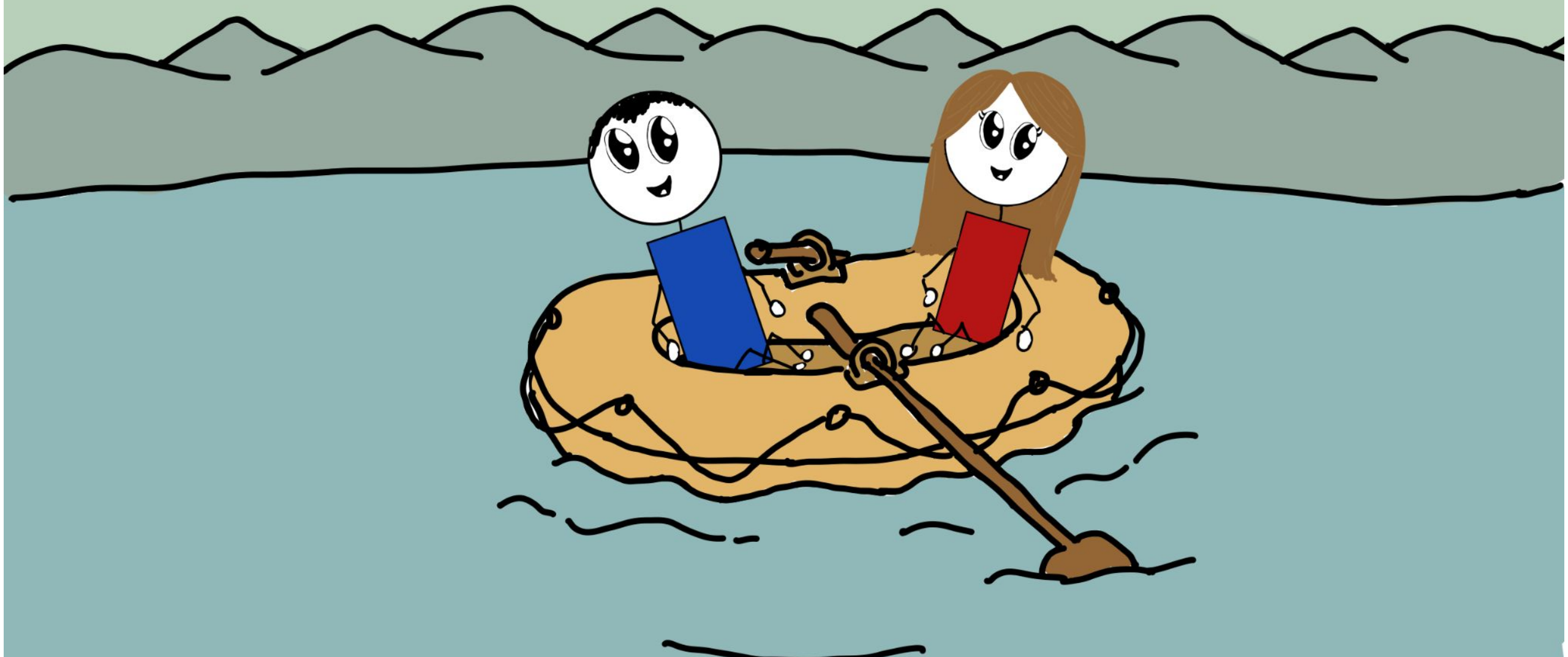



Source: SCGupta

# What are my options to run GenAI workloads and enable others to do so too?

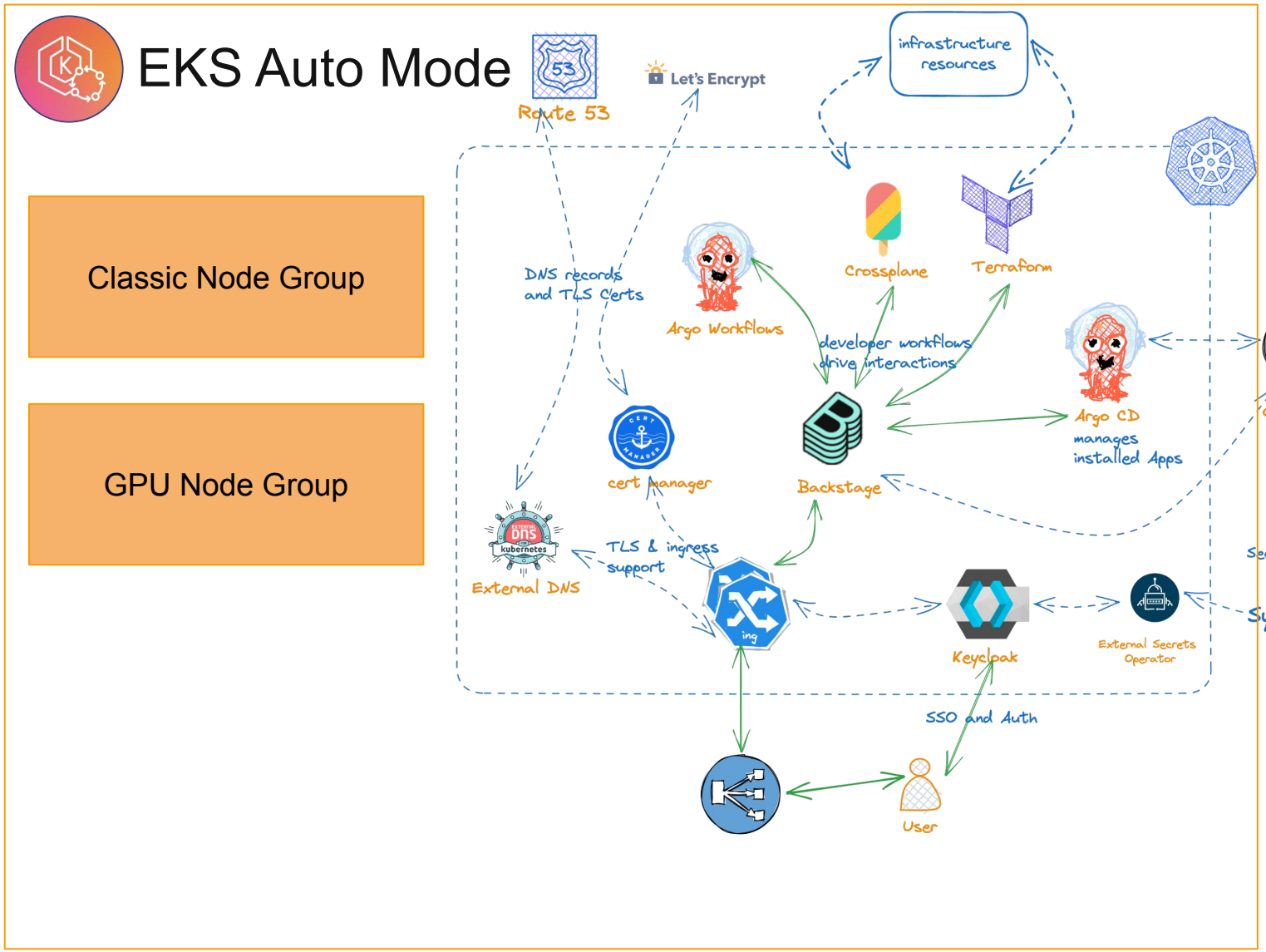


# Stage 1: TVP

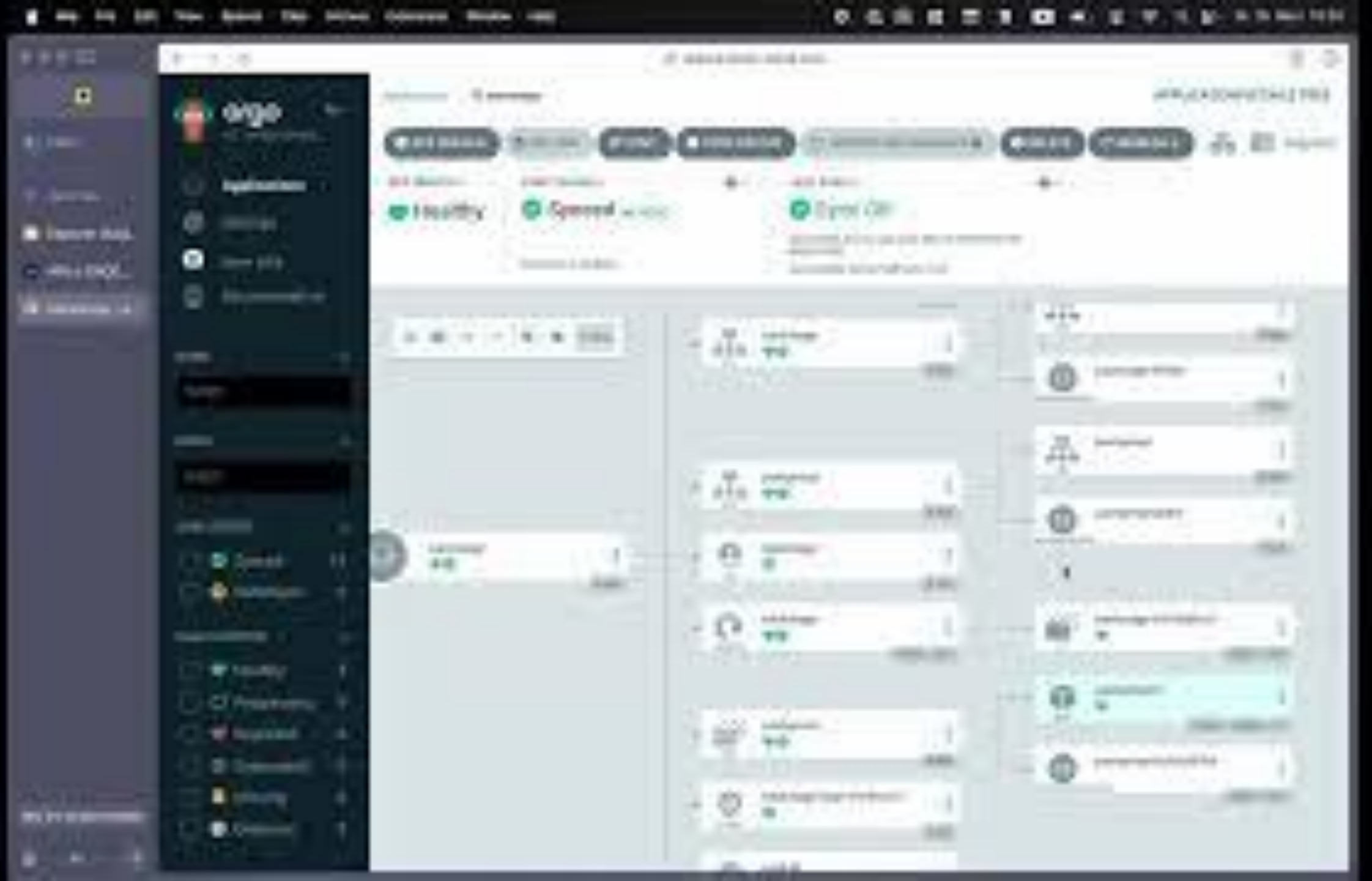


# “Minimal IDP”

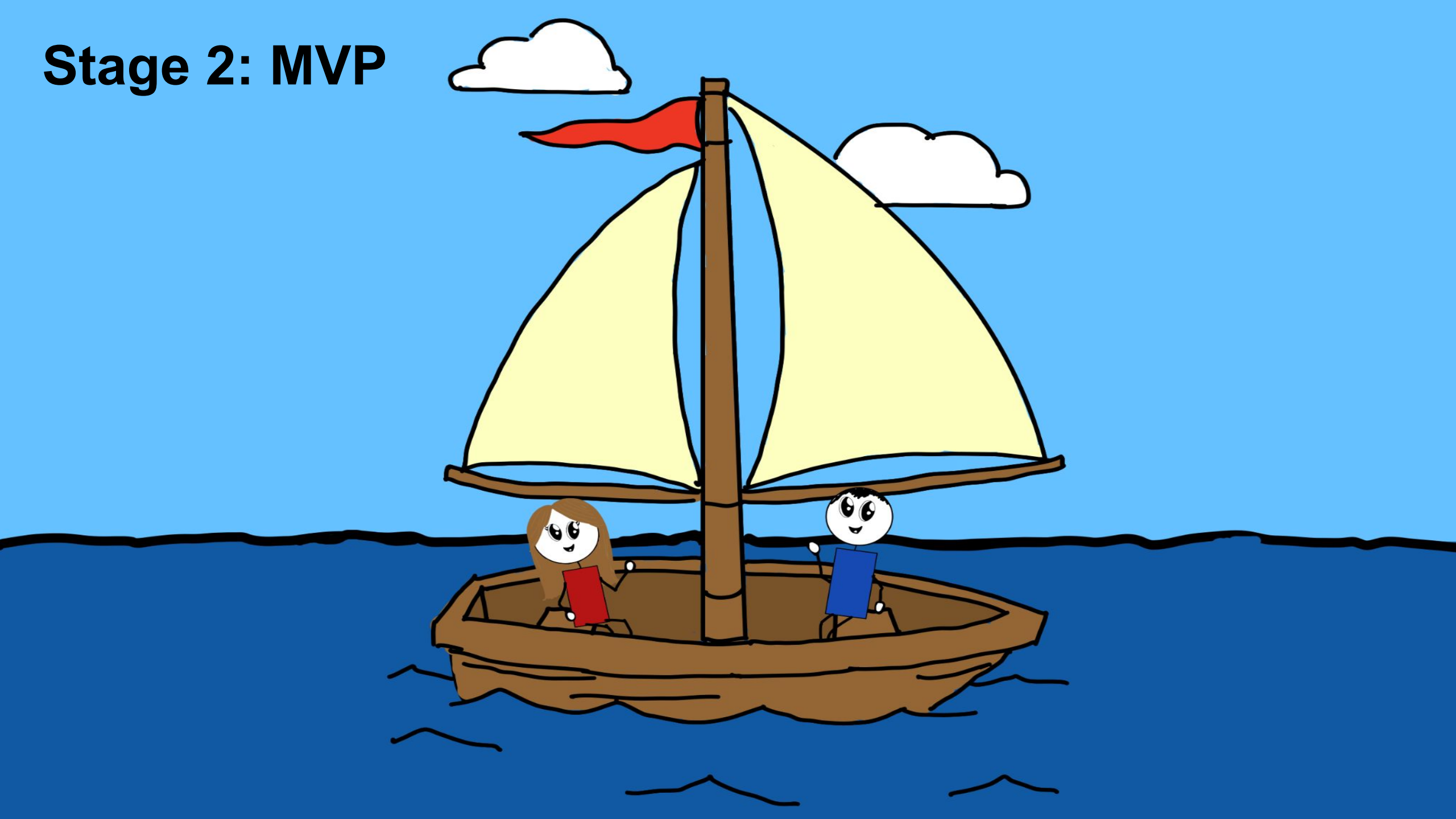





CNOE


argo
Healthy
Synced

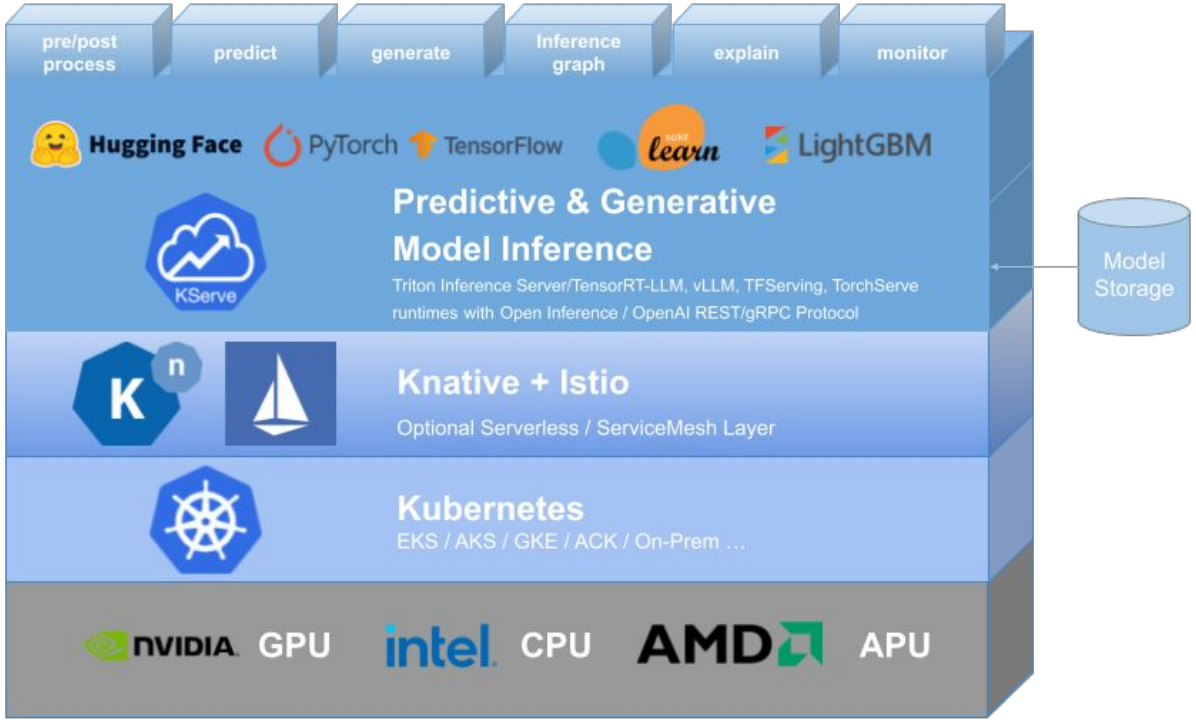
# Stage 2: MVP



# KServe: Powering the AI Expedition

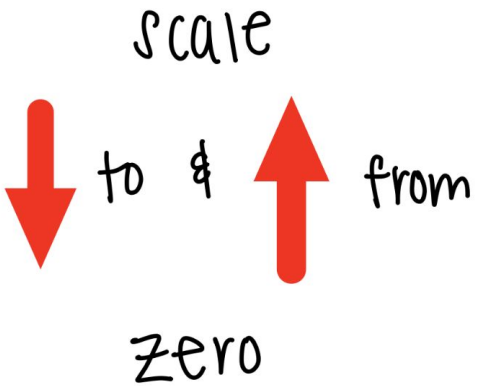


- Simplifies scalable model serving on Kubernetes
- Supports multiple model runtimes out of the box, including LLMs

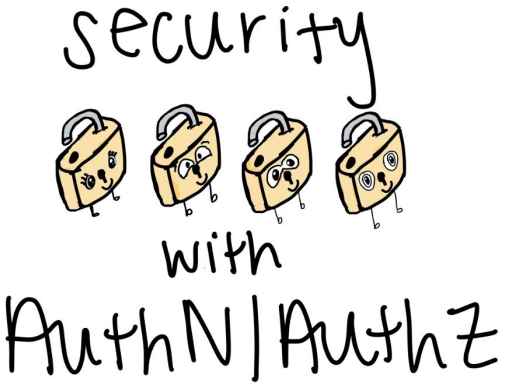


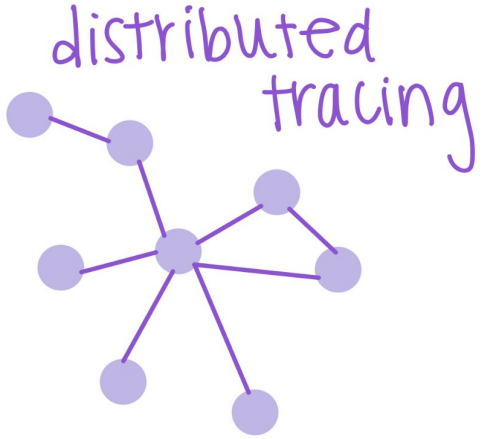
# KServe: Powering the AI Expedition

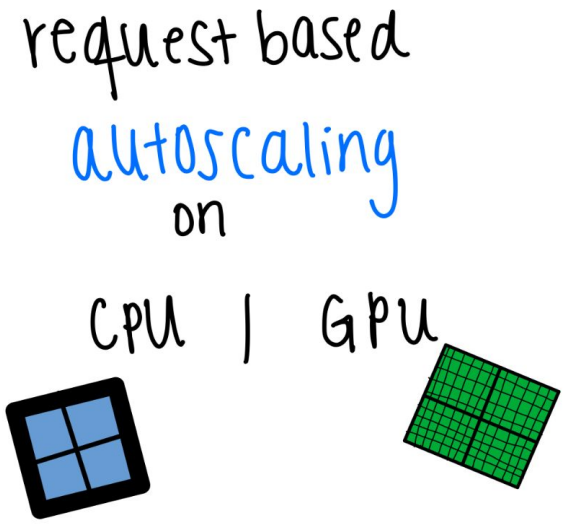


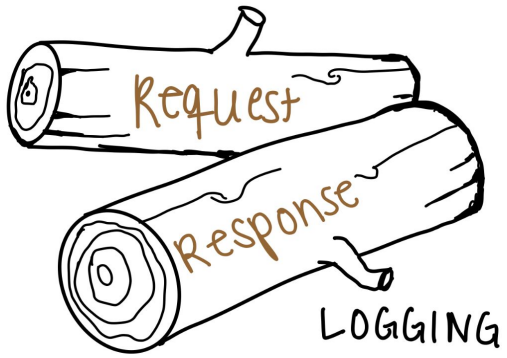




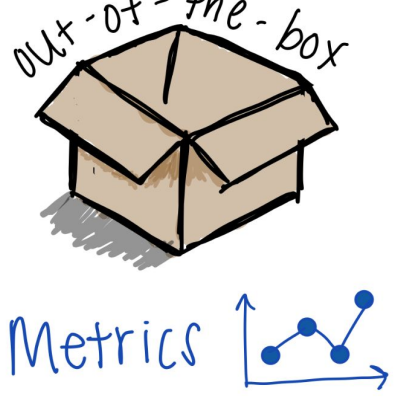


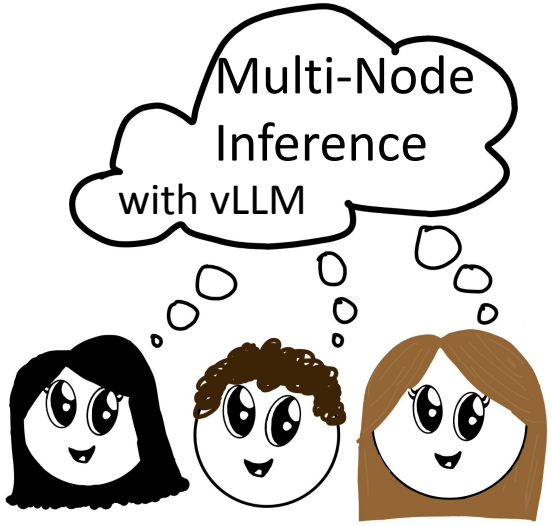




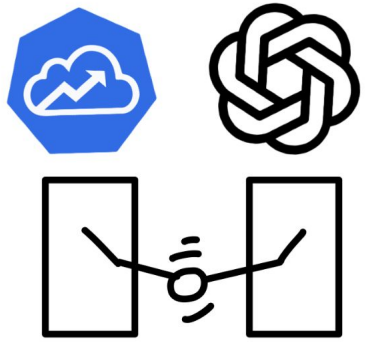
# Building the AI Vessel: KServe for GenAI

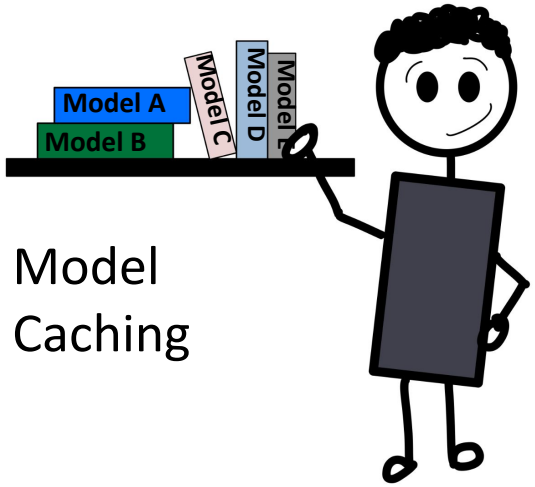


LLM
Metric-based
Autoscaling

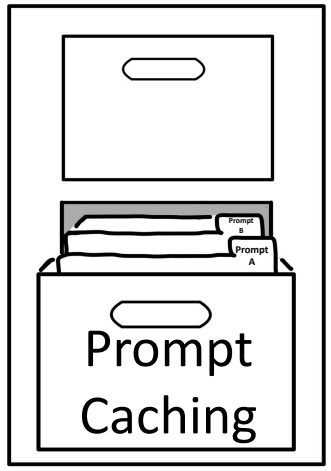



OpenAI Protocol Support

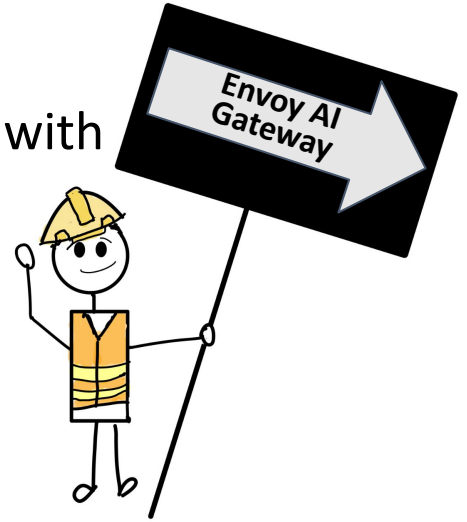





Model
Caching

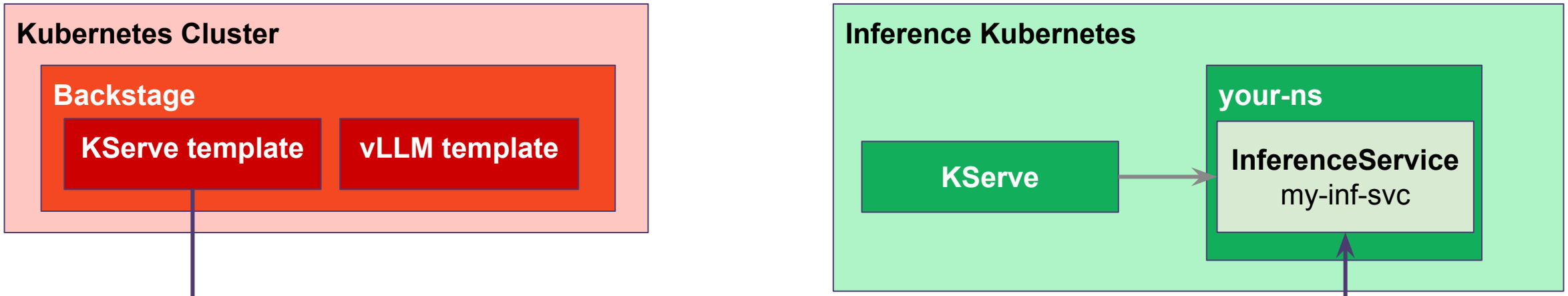



Traffic
Management with



# Utilize Backstage for Inferencing & More

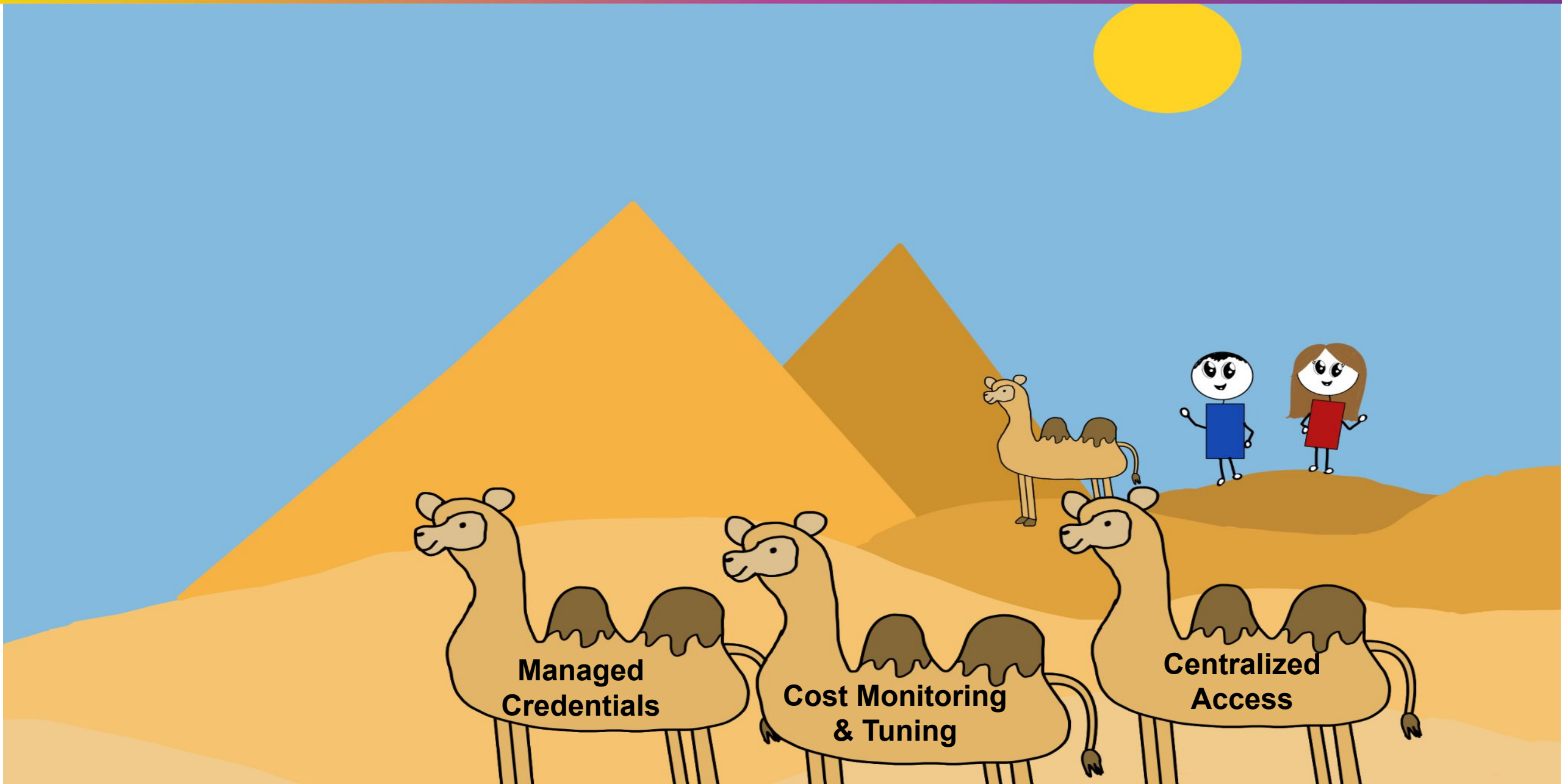




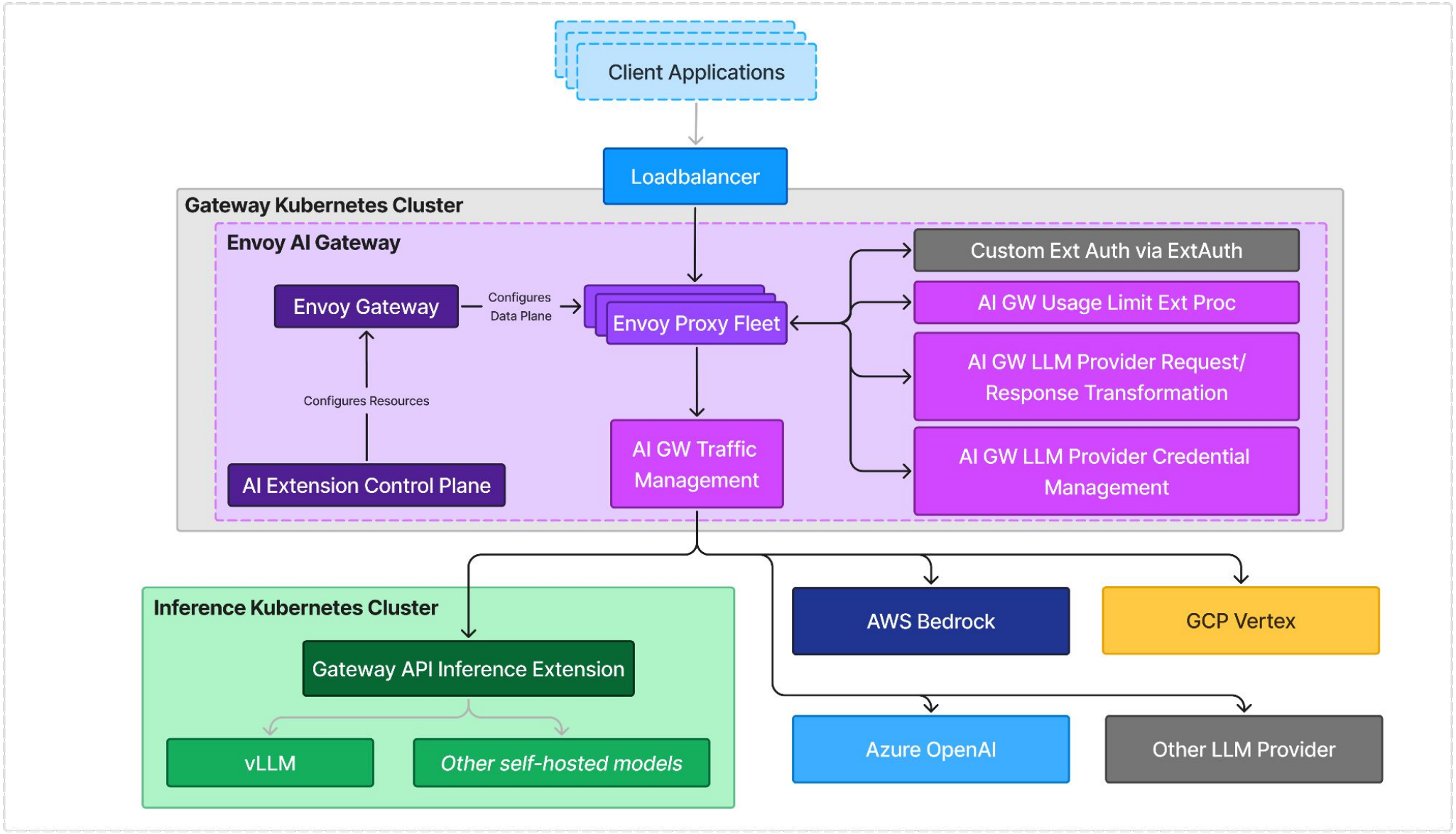

```yaml
apiVersion: serving.kserve.io/v1beta1
kind: InferenceService
metadata:
  name: ${{values.name | dump}}
  annotations:
    backstage.io/kubernetes-namespace: "${{values.namespace}}"
spec:
  predictor:
    model:
      modelFormat:
        name: ${{values.name | dump}}
      storageUri: ${{values.url}}
```

# Navigating Access with Envoy AI Gateway

KubeCon CloudNativeCon Europe 2025

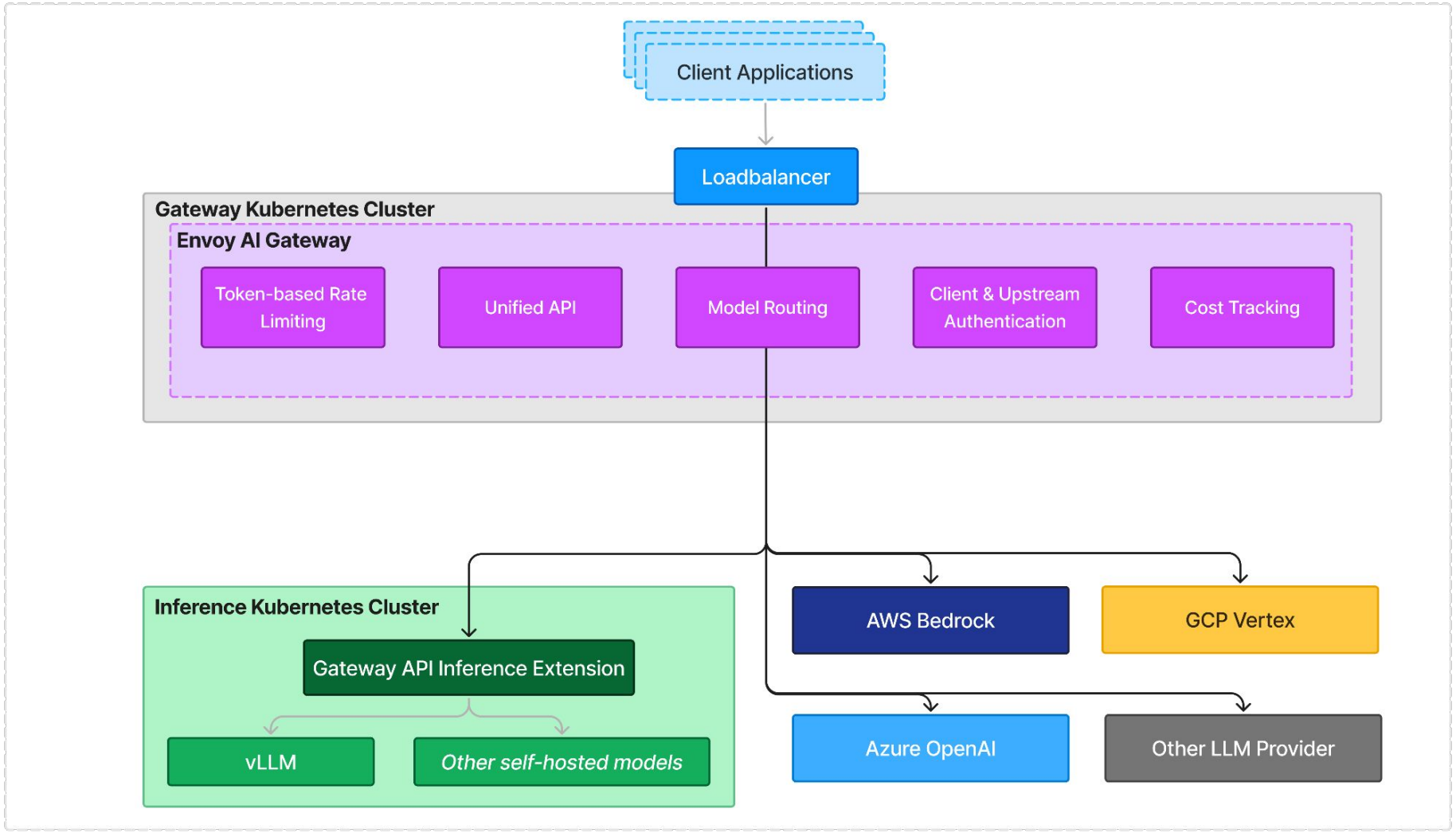


# Through the Envoy AI Gateway



# Through the Envoy AI Gateway

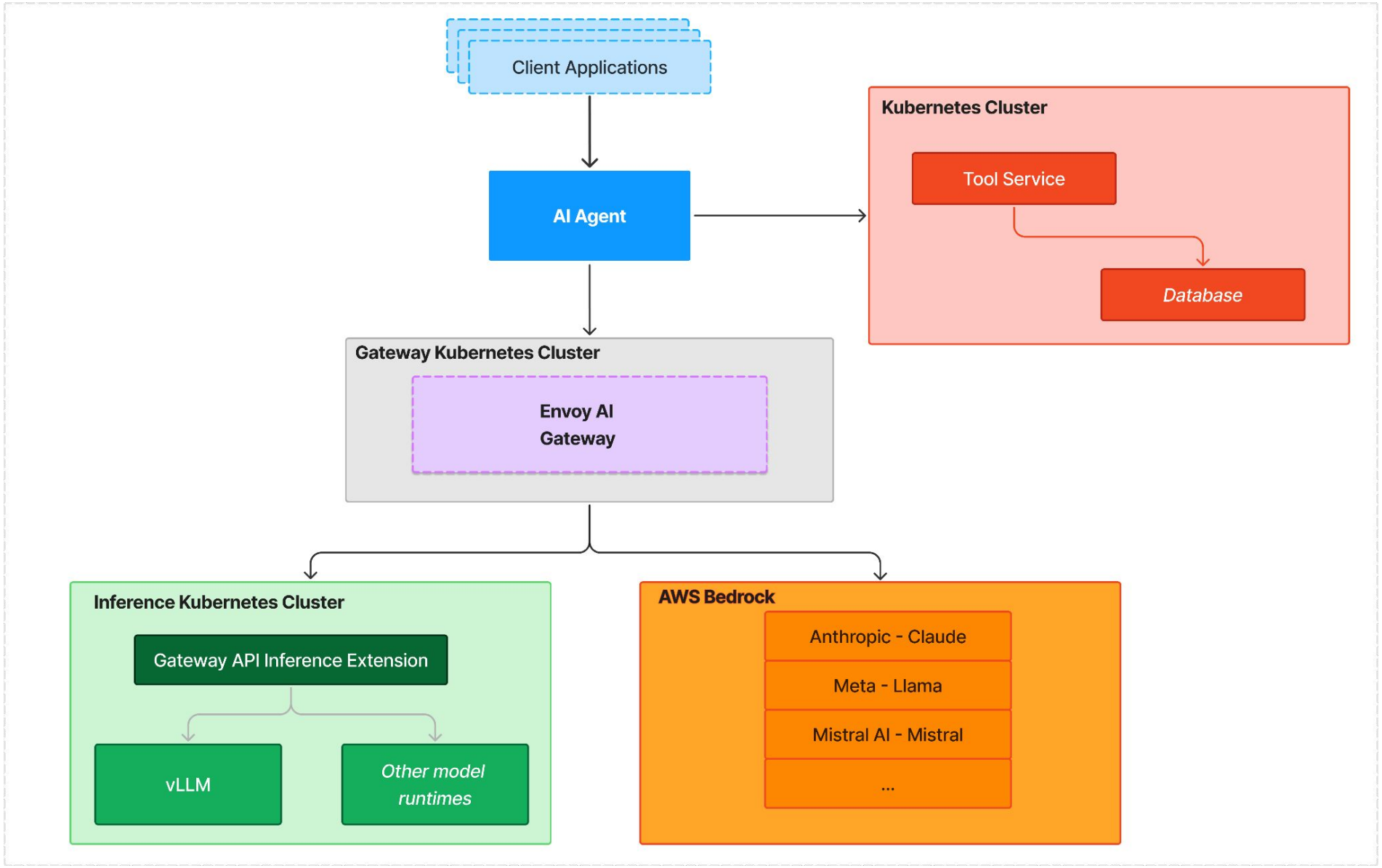





[KServe Envoy AI Gateway Integration]

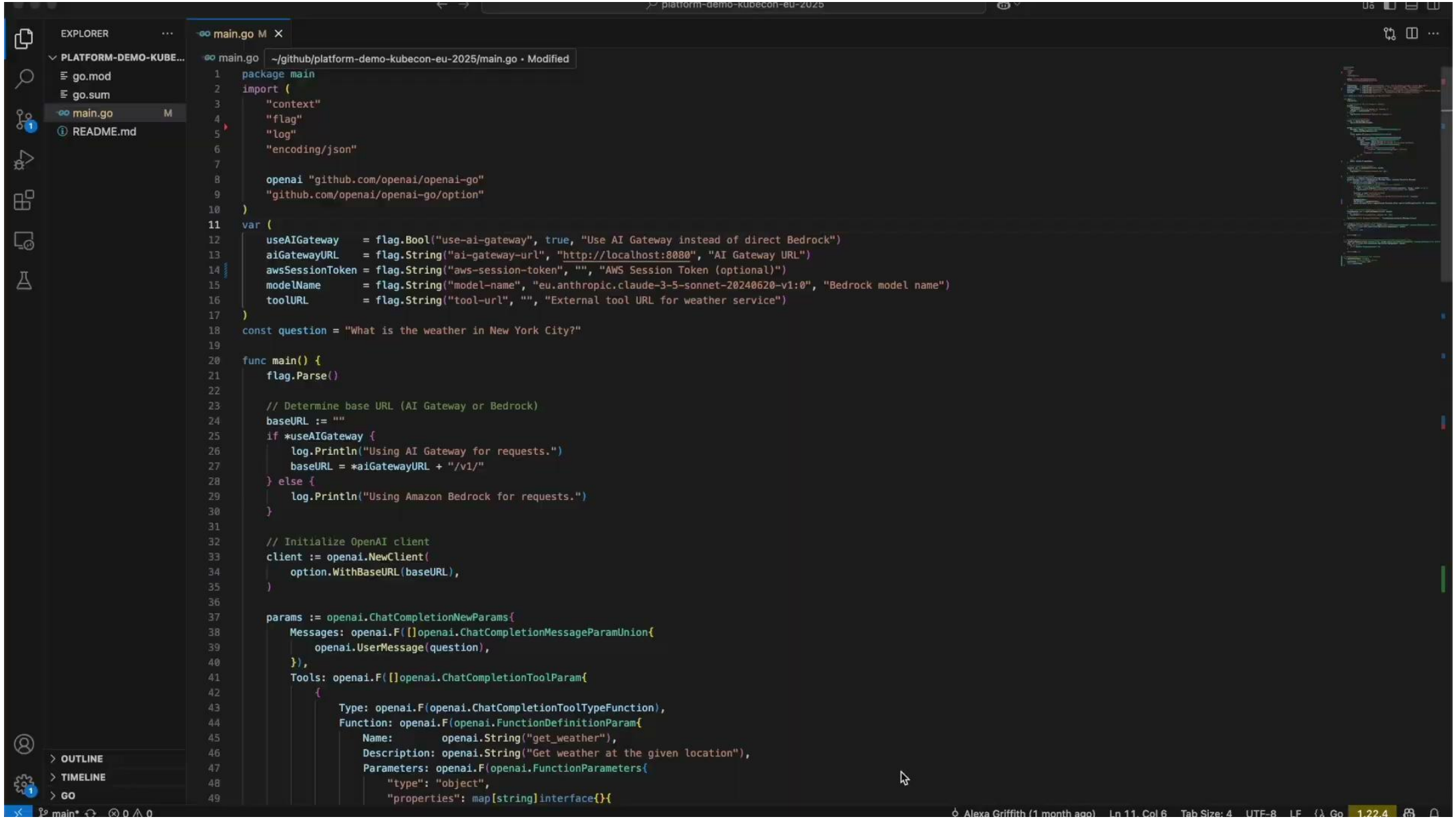
# Map of the Route Ahead





# AI Gateway Demo

KubeCon CloudNativeCon Europe 2025



# Charting the Course for Managed LLMs



**AI without the Overhead**

**Optimized & Up-to-Date Models**

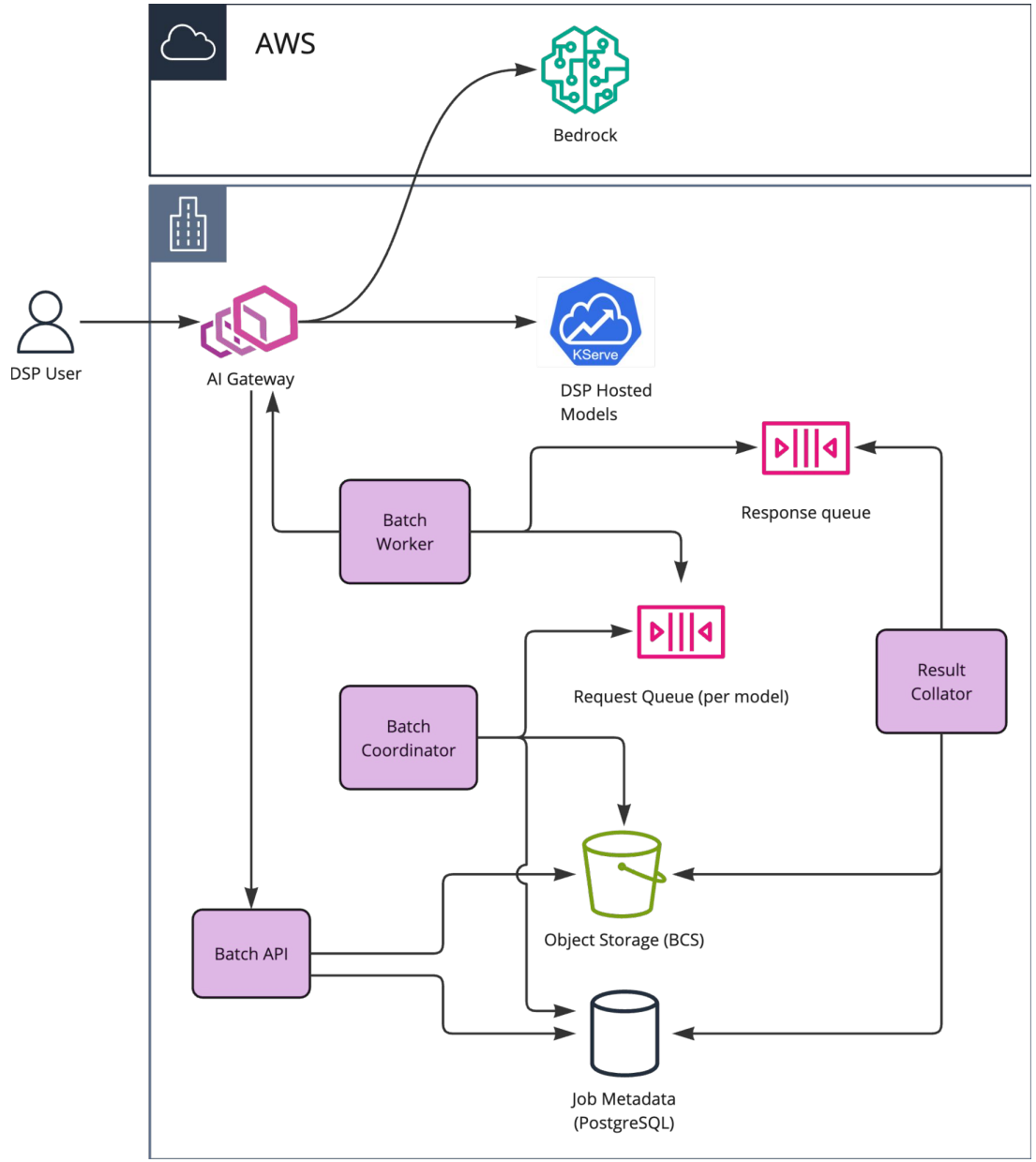
**Unified Access**

**Cloud-Native Scalability**

# Stage 3: Advanced Features

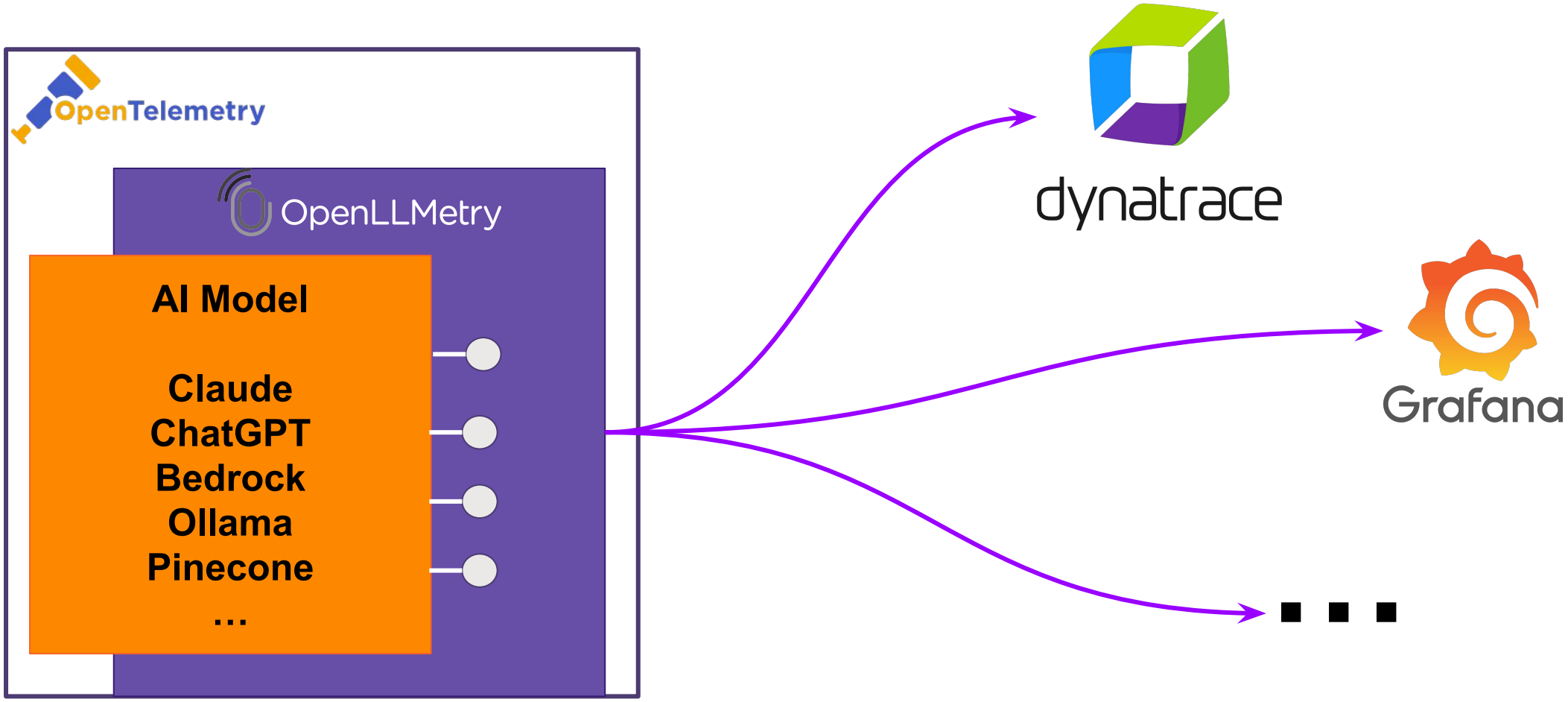
# Expedition Mode: Scaling AI with Batch Inference





# LLM Observability Observatory





## Service Health & Performance

Open Problems

2

# of Total Requests

796.00

≈ 150%

Cost

840.88US$

≈ 60,83%

Service Health

API Request Duration

1.07s

≈ 4,82%

PII Request Duration

4.49s

≈ 61,05%

Guard AI Processed

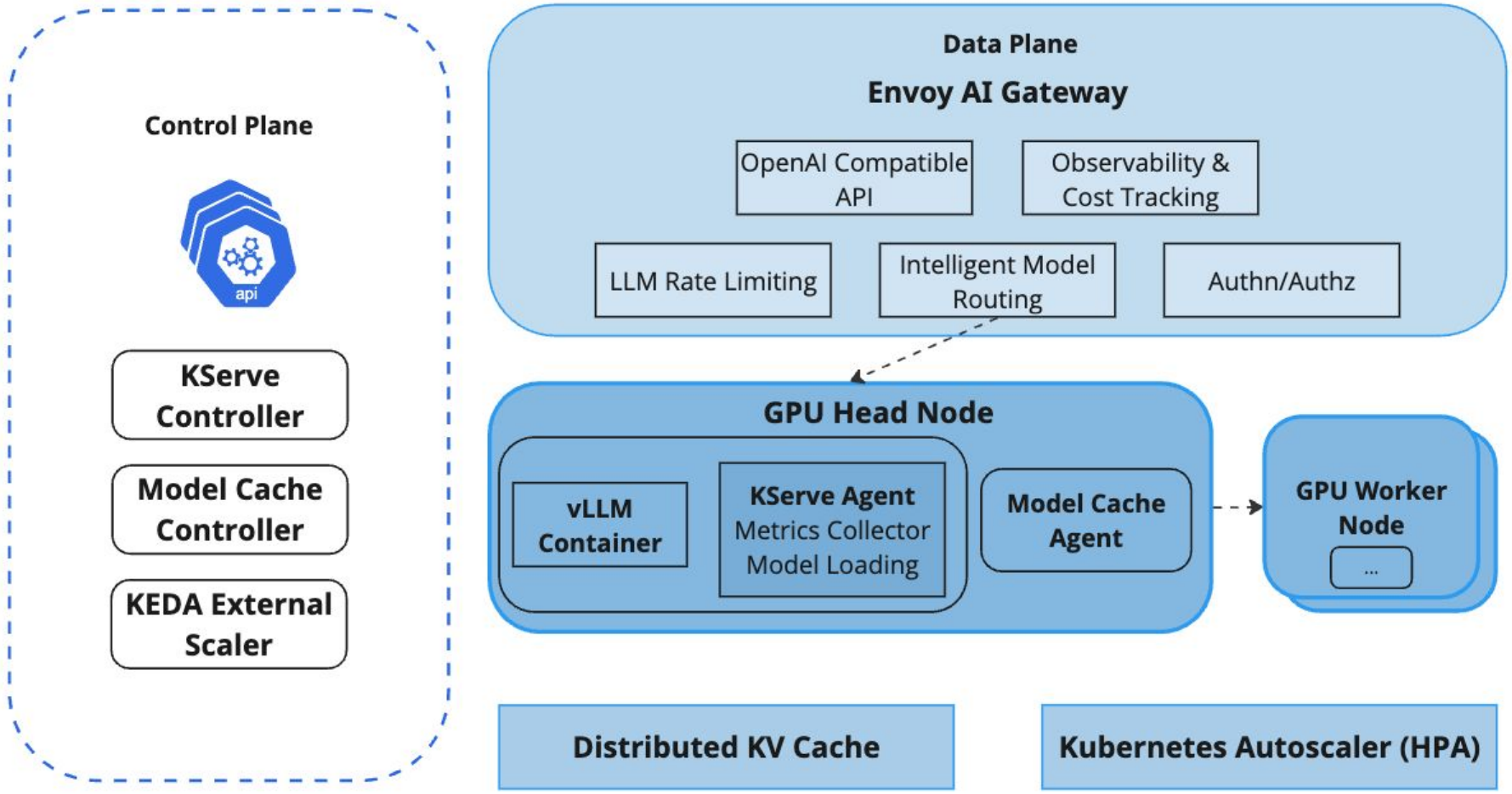
## Service Quality & Guardrails

| Guardrail Executions | Toxicity | PII Leaks | Denied Topics | Grounding |
|---|---|---|---|---|
| 7.55% | 14.23% | 14.00% | 14.23% | 0.74 |

≈ 3,46%

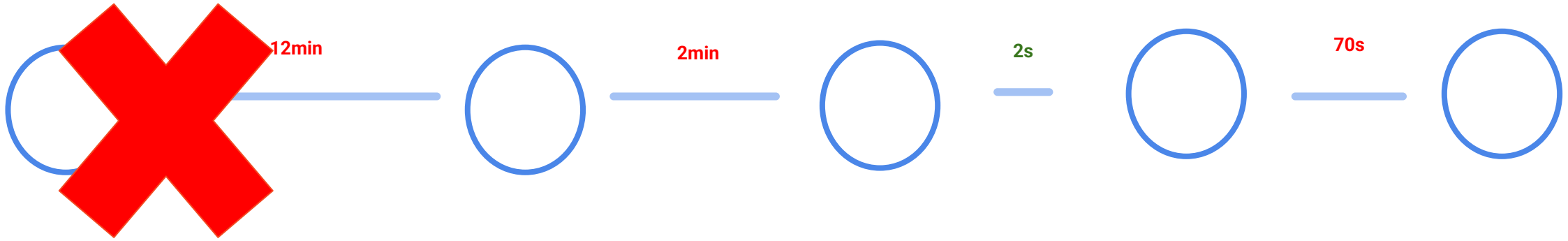
# Upgrades for the Expedition





# Model Caching






**Init Container**

Download models from Model Storage to local volume

**Image Pull (first time)**

Pull serving runtime image (vLLM) from remote artifactory registry

**Create model server container**

Create and start the model server container

**Load model**

Load model from mounted volume into the model server GPU memory

**Ready**

Ready for inference

# Prompt Caching



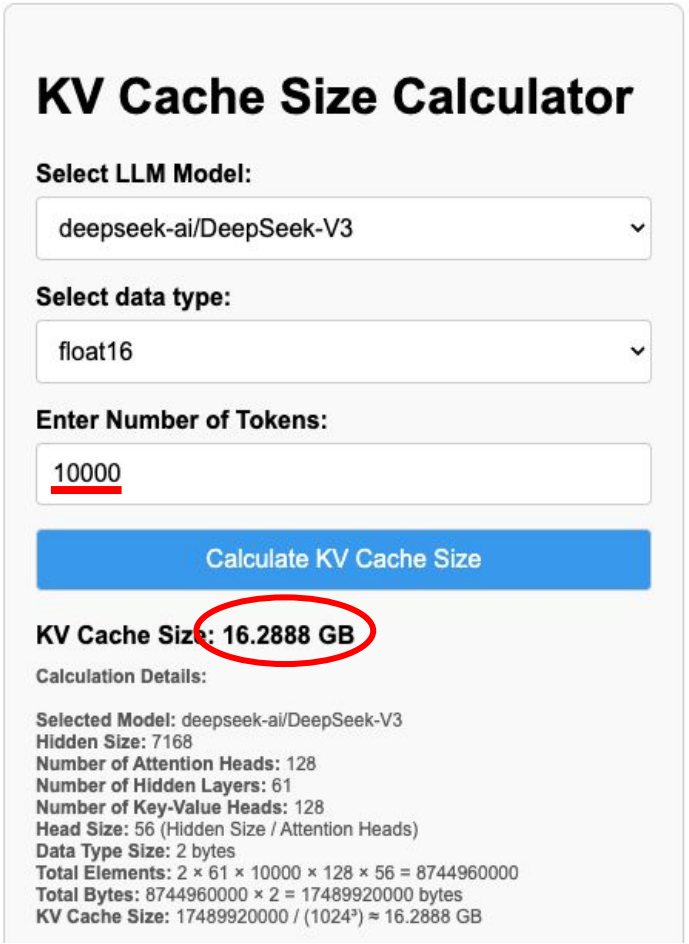
🧠 **KVCache**

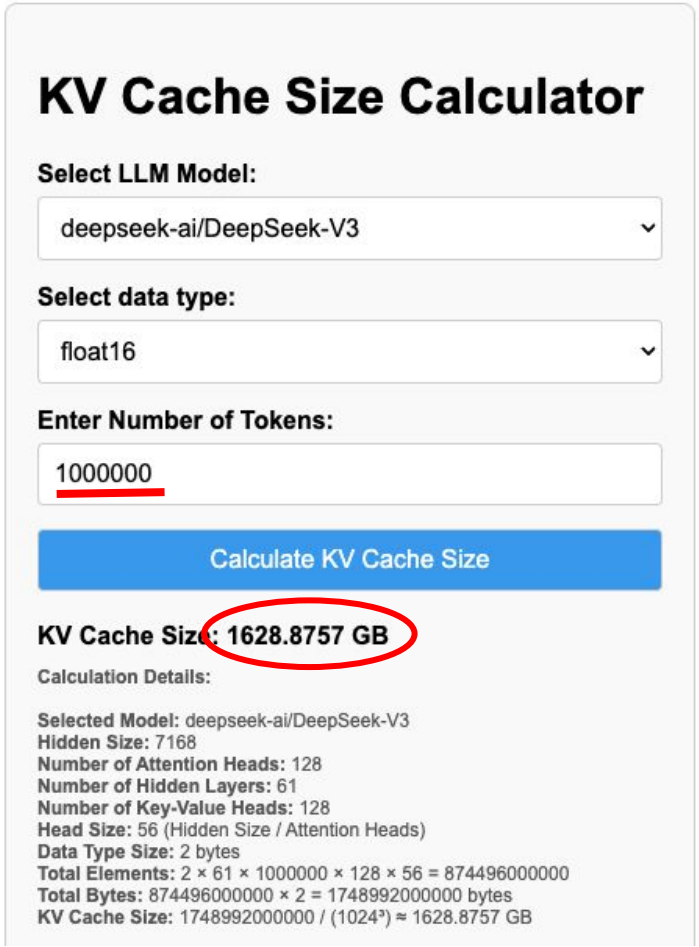
Grows too fast! 📈

💡 **LMCache**

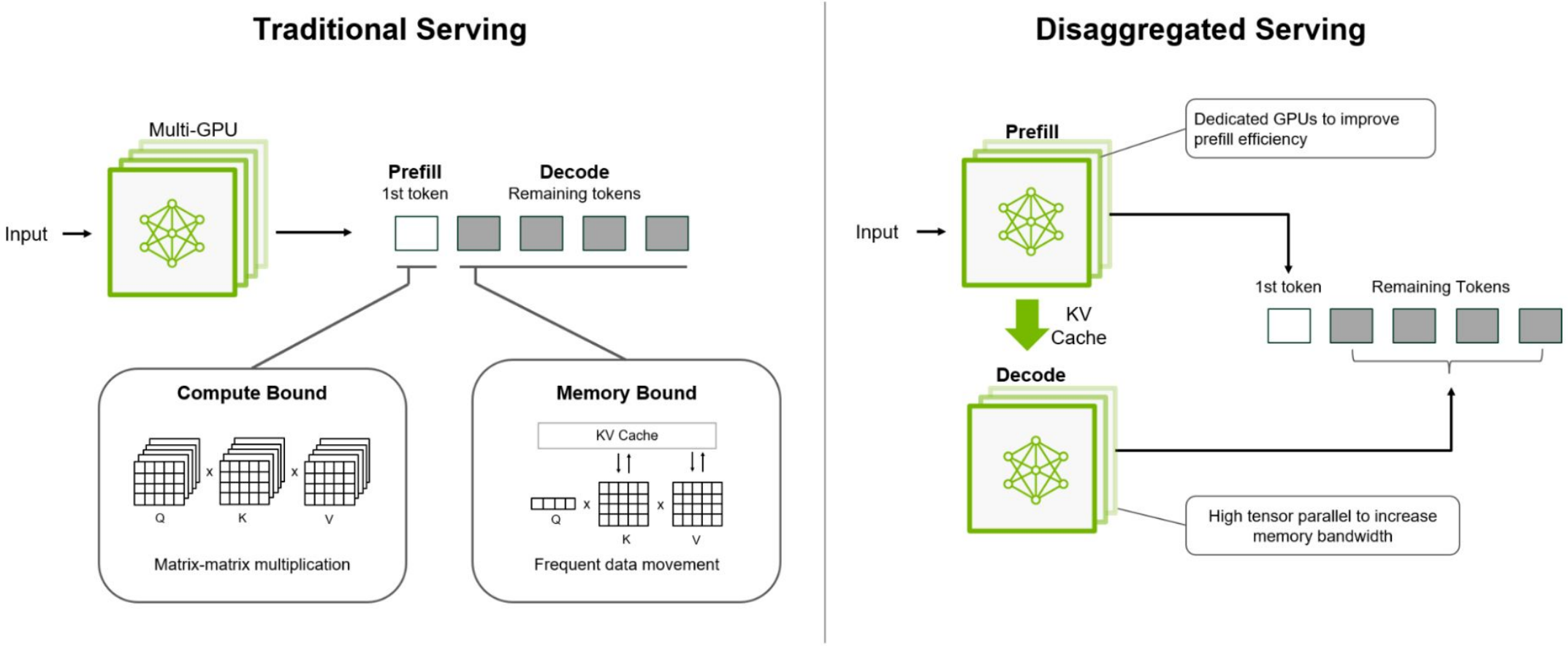
Reuses common prefixes.

✅ Faster inference!



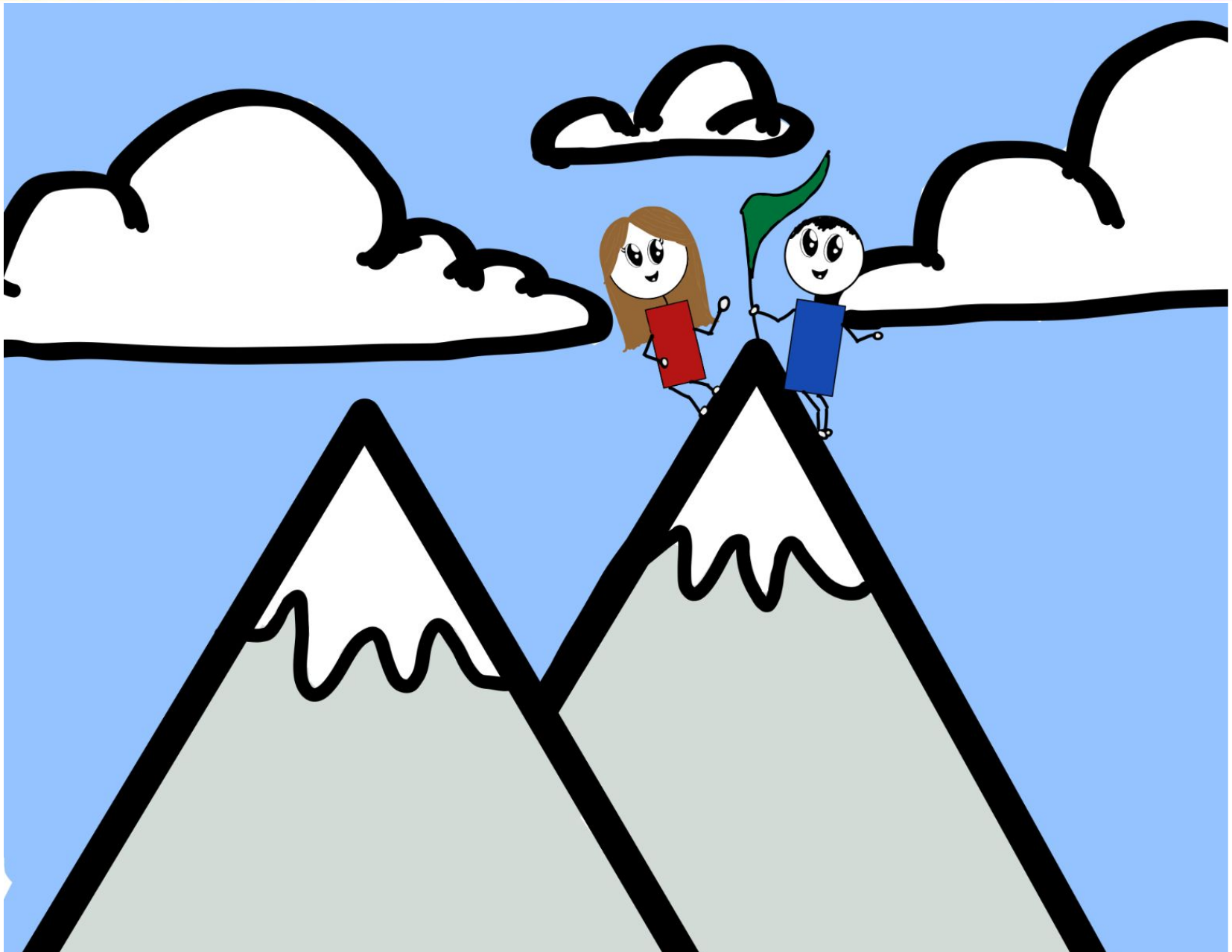
# Disaggregated Serving




https://developer.nvidia.com/blog/introducing-nvidia-dynamo-a-low-latency-distributed-inference-framework-for-scaling-reasoning-ai-models/

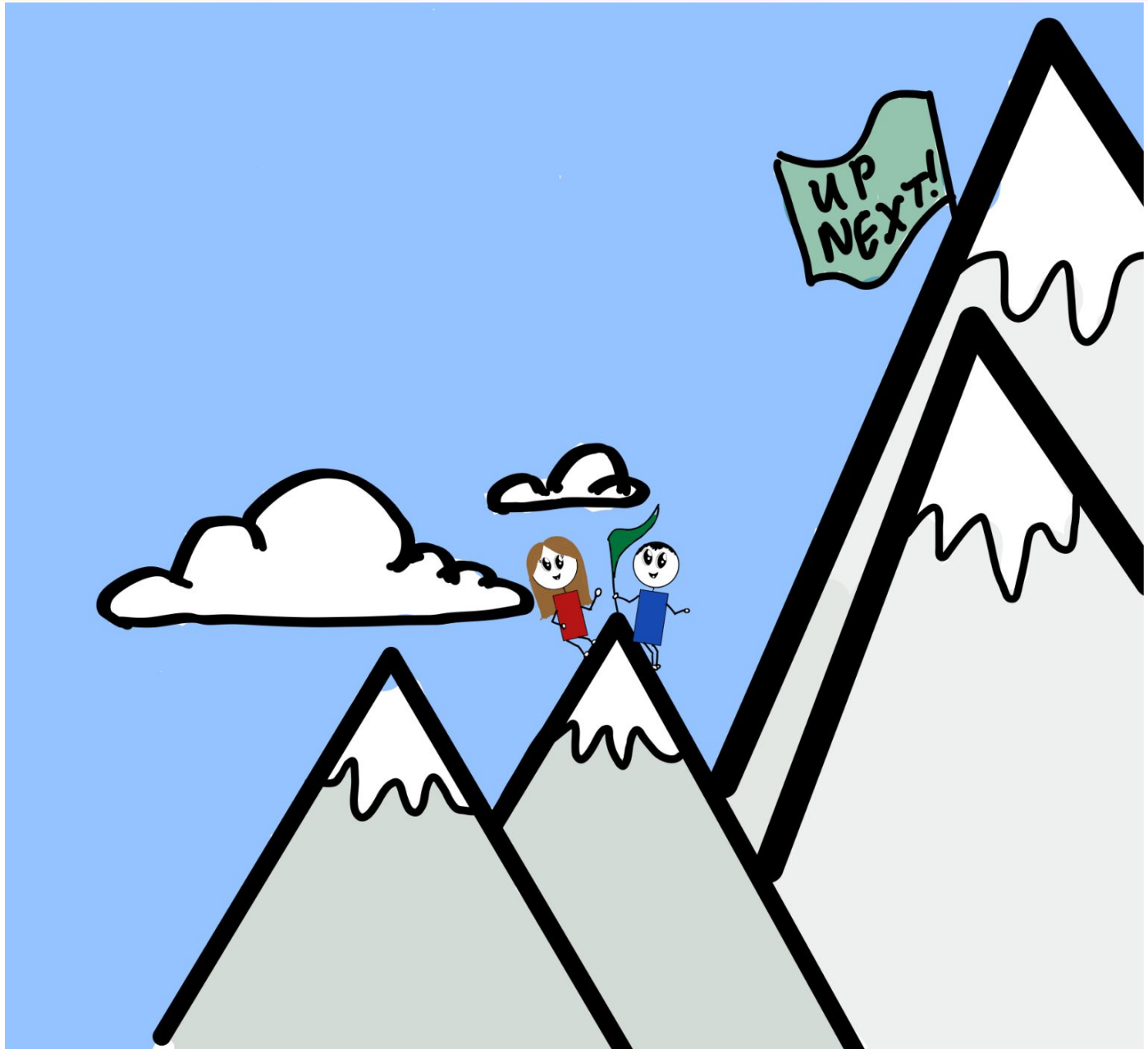
# End of the Journey





# End of the Journey (or is it?)





# We are hiring: bloomberg.com/engineering

- Gen AI Platform Team Lead - AI Engineering (London)
- Senior Software Engineer - Data Technologies Compute Platform (London)
- Senior Software Engineer - AI Hardware (NYC)
- Senior ML Ops Engineer - AI Engineering (NYC)
- Senior Software Engineer - Bare Metal as a Service (NYC)
- Senior Software Engineer - Public Cloud IAM and Org Management (NYC)
- Senior Software Engineer - Public Cloud Pipelines (NYC)
- Senior Software Engineer - Public Cloud Visibility (NYC)
- Senior Java Engineer - Search Infrastructure (NYC)
- Technical Product Manager, GenAI Developer Platform - CTO Office (NYC)
- Technical Product Manager, LLM Platforms - CTO Office (NYC)

Bloomberg

**Contact Our Recruiters**

**TechAtBloomberg.com**

We are hiring: https://join.com/companies/liquidreply

- Business Unit Manager - Sovereign Cloud - Germany
- Developer Relations Engineer - Platform Engineering, Open Source & Sovereign Cloud - Germany
- Platform Engineer - Germany
- Cloud Engineer - Germany

REPLY LIQUID

Grab your free copy with signing of
**“Platform Engineer for Architects - Crafting moden platforms as a product”**


- Today - Booth S240 @Dynatrace during booth crawl
- Tomorrow - Booth S641@Syntasso during lunch break