KubeCon + CloudNativeCon Europe 2024

Deploying LLMs in CloudNative using LangChain

Ezequiel Lanza
AI Open Source Evangelist

Arun Gupta
VP/GM Open Ecosystem

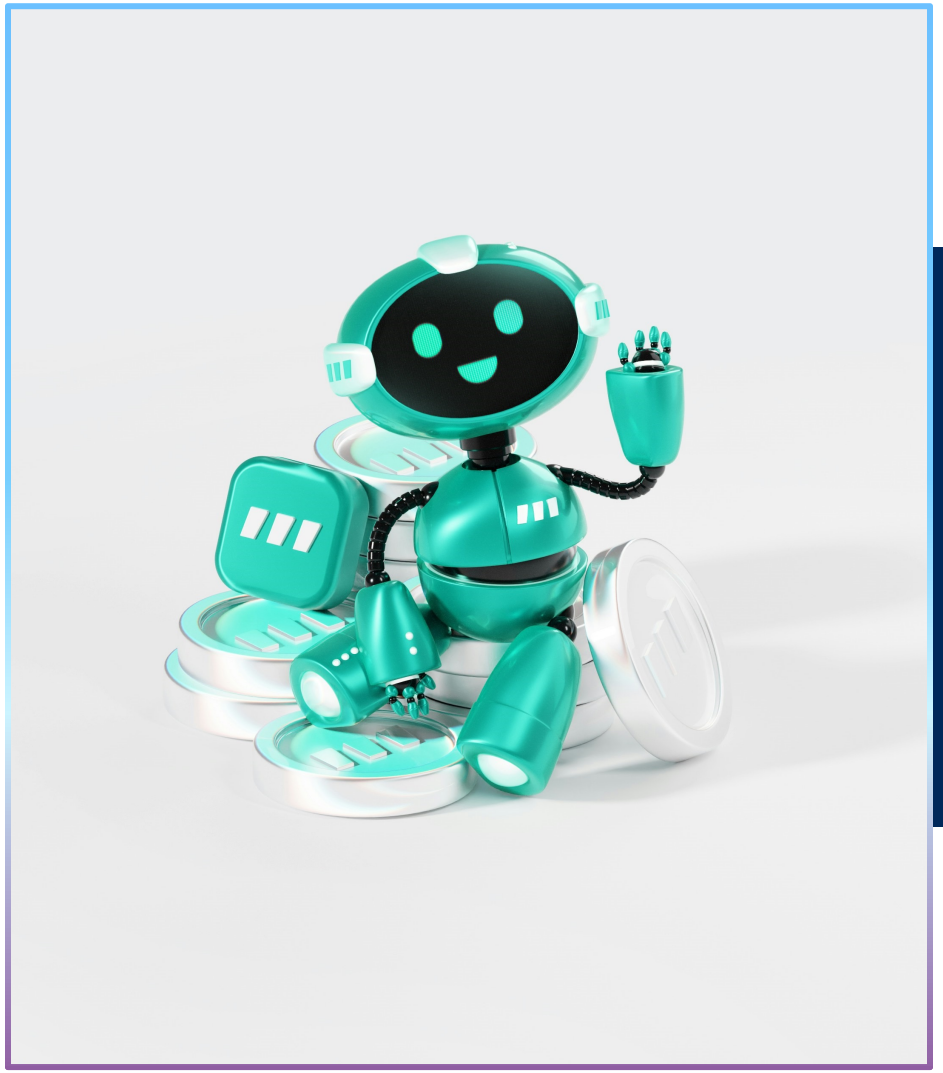


Photo by Ant Rozetsky on Unsplash

Everybody wants an **assistant**

Companies require a variety of solutions

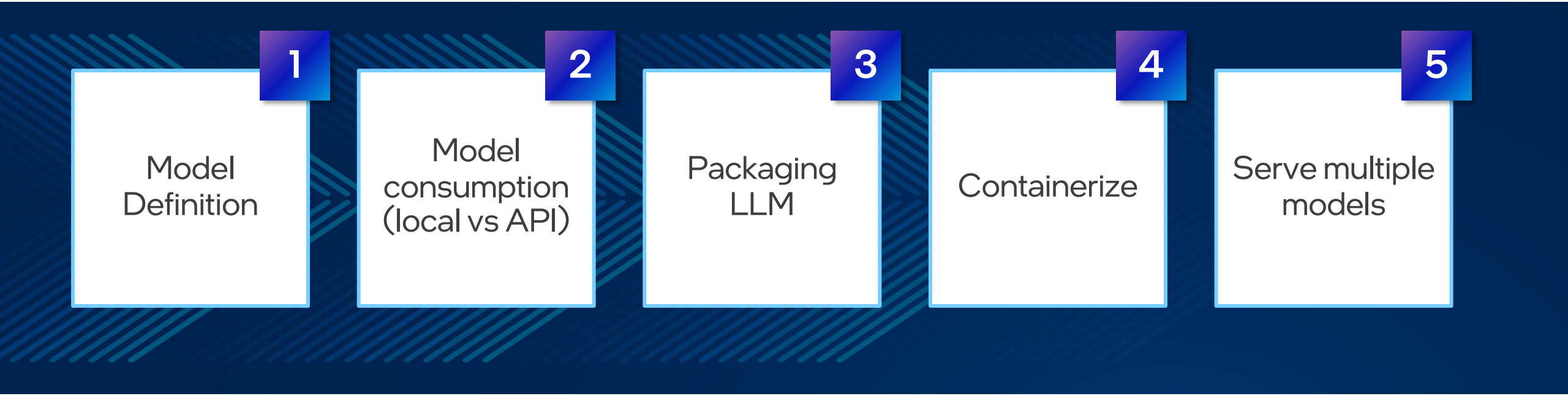
Finance

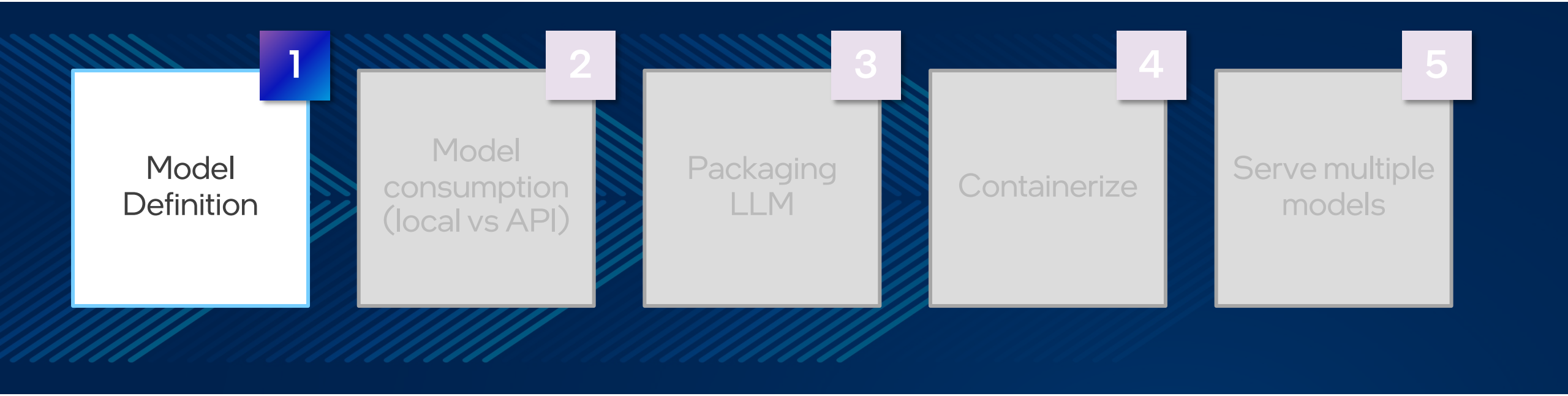
Legal

IT

...and others

One Way to Deploy your LLM in CloudNative




1
Model Definition
2
Model consumption (local vs API)
3
Packaging LLM
4
Containerize
5
Serve multiple models

1 Model definition

Define your problem

- Conversational chatbot
- Text summarization
- Classification
- Question Answering

Pick your model/strategy

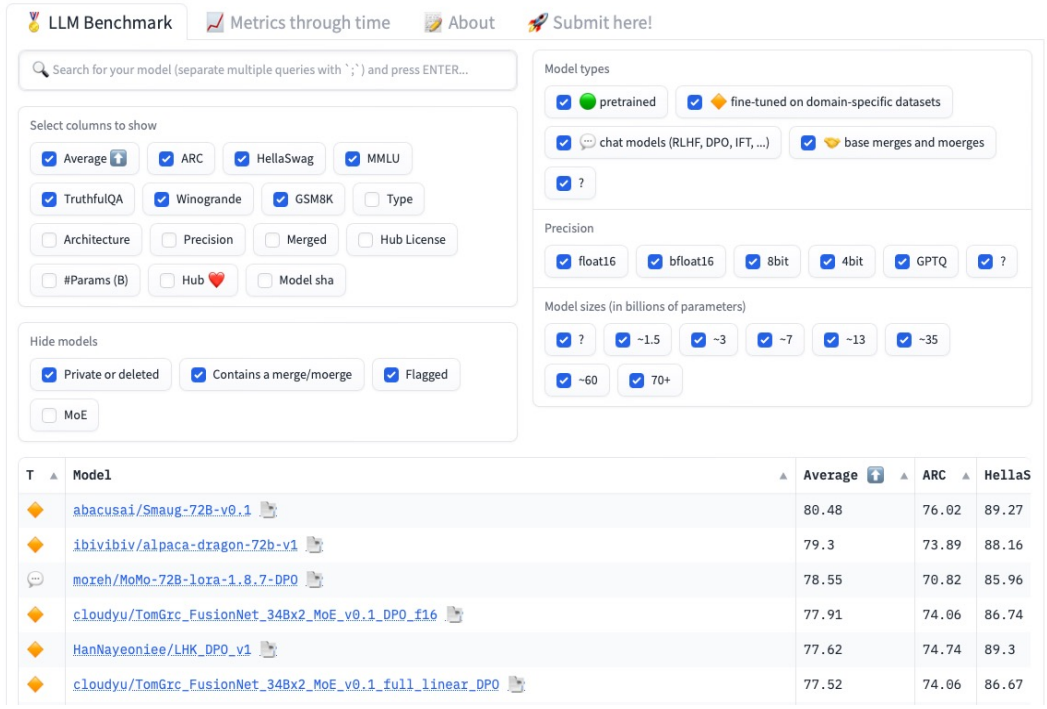
- Foundation Model (General knowledge)- 70b/7b
- Fine-tune a model (Context knowledge) Own Data
- Retrieval Augmented Generation (RAG)

Find tools

- HuggingFace
- LangChain
- Additional tools based on the model decided (Milvus/Chroma/FAISS for RAG, or tuning tools if a model will be fine-tuned)

1 Model definition: Considerations for picking a model

Hugging Face Leaderboard



Documentation/Community

Adoption

Support

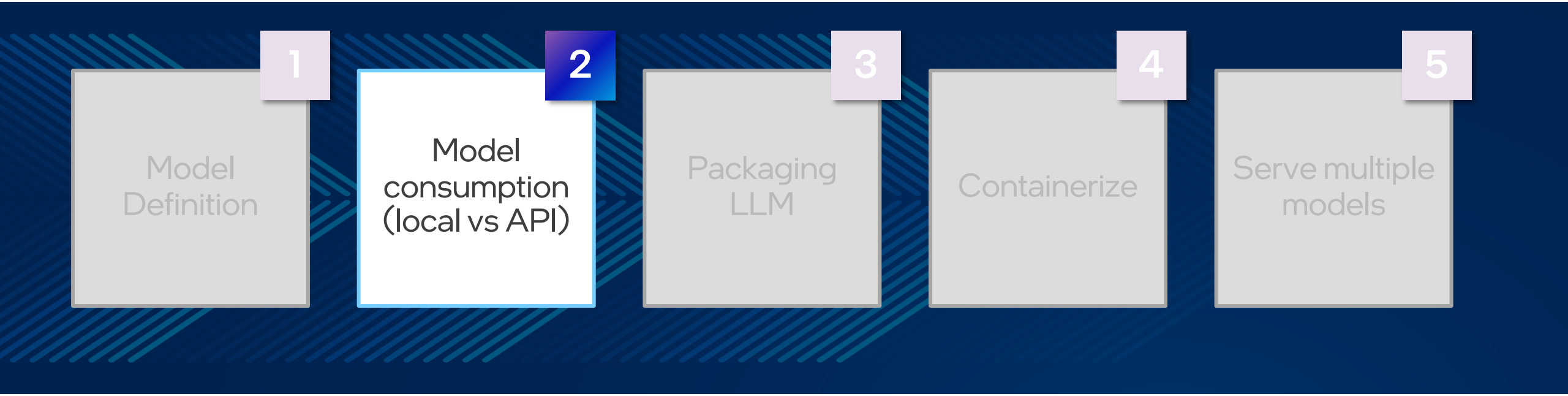
Tutorials

Forums

Ethical considerations

Training Data

Bias


1
Model Definition
2
Model consumption (local vs API)
3
Packaging LLM
4
Containerize
5
Serve multiple models

2 Model Consumption : Local vs External

2 Model Consumption : Considerations

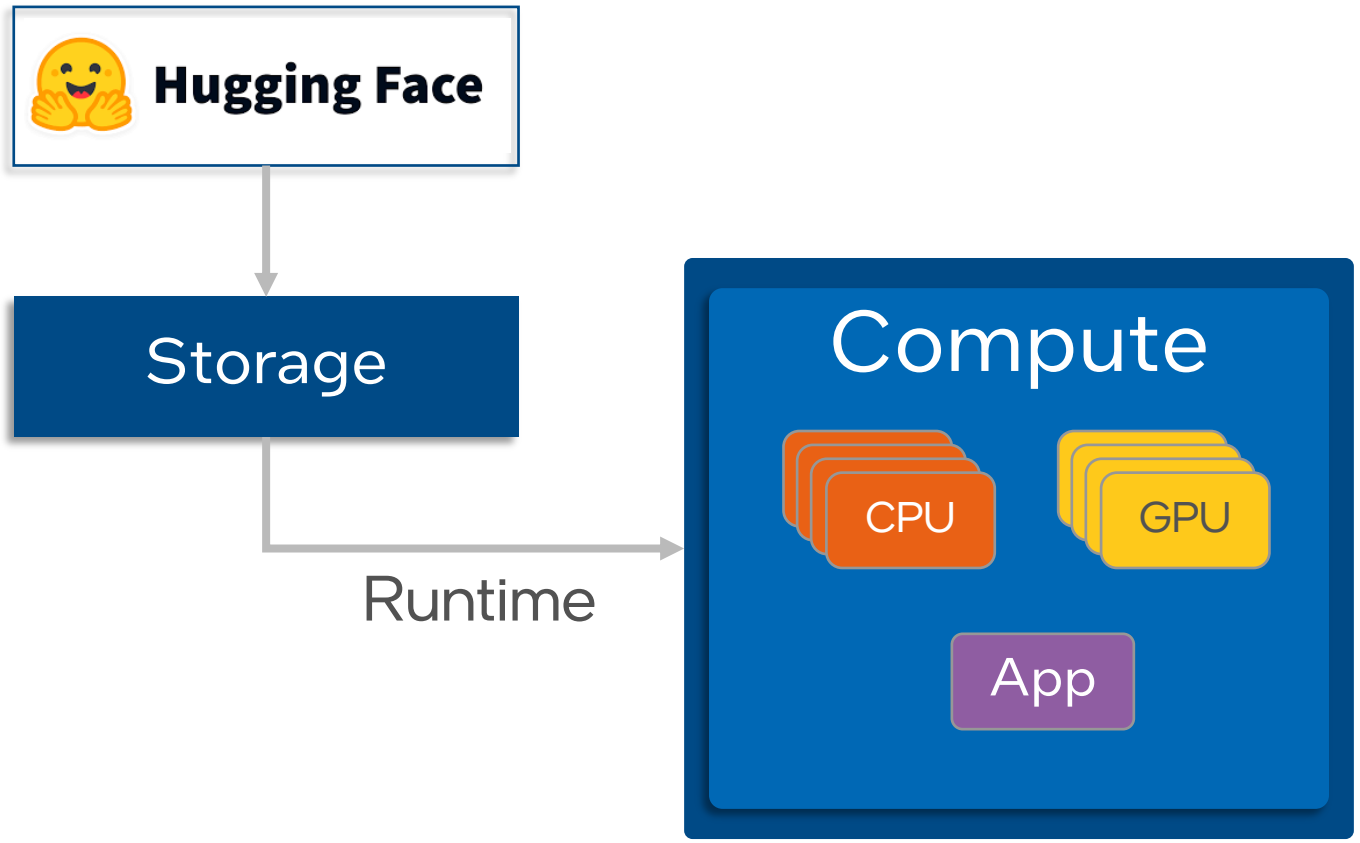
Local Model

- Data privacy
- Offline usage
- Cost
- Censorship
- Better customization

External Model

- Deliver quickly
- Scalability
- Less complex setup
- Elasticity
- High availability

2 Model Consumption: Local Inference



Considerations

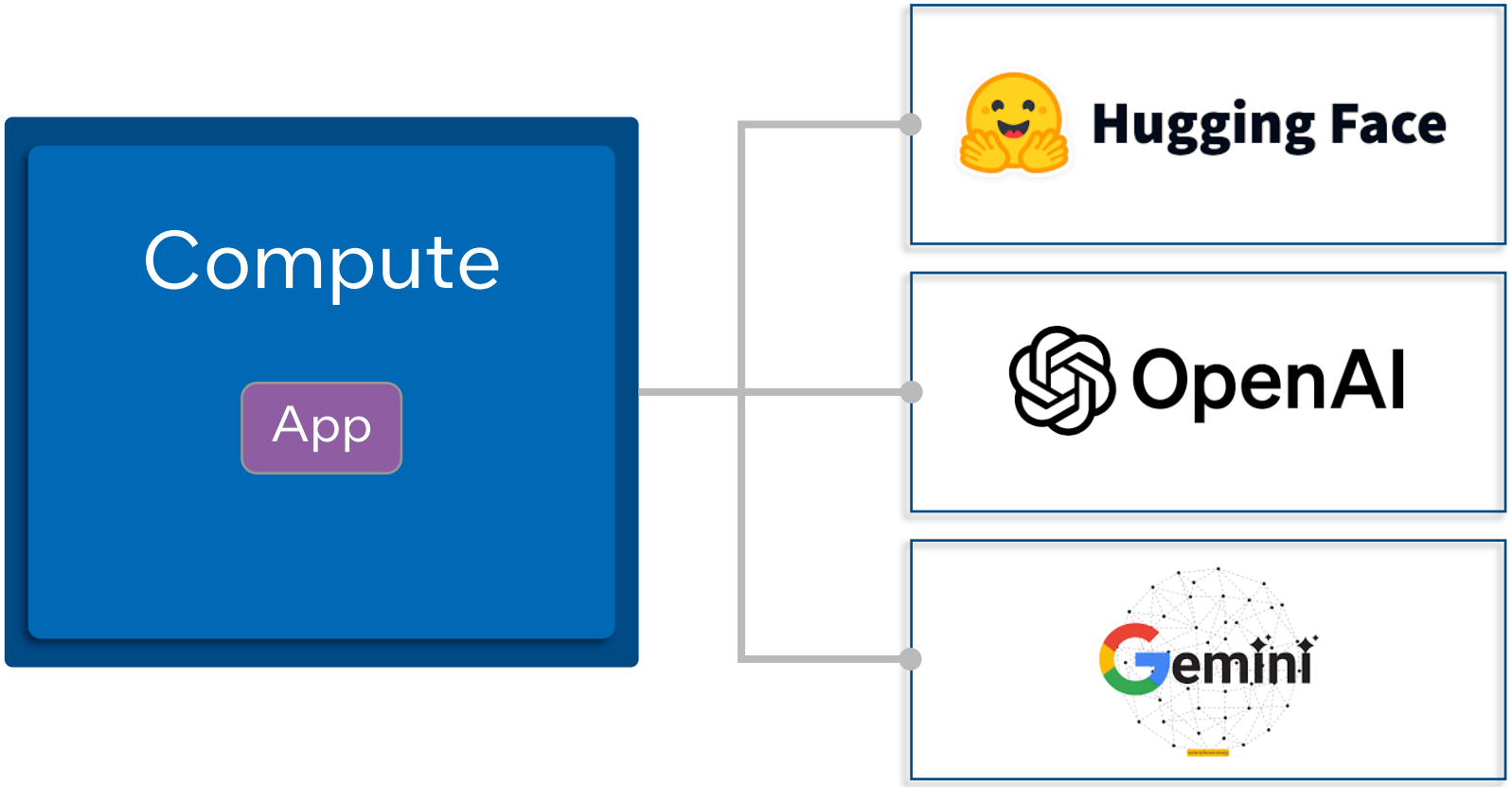
Size (RAM)

- 7B: ~ 26GB
- 70B : ~ 120GB
- 7B (optimized): ~ 7GB
- 70B (optimized): ~ 36GB

Access

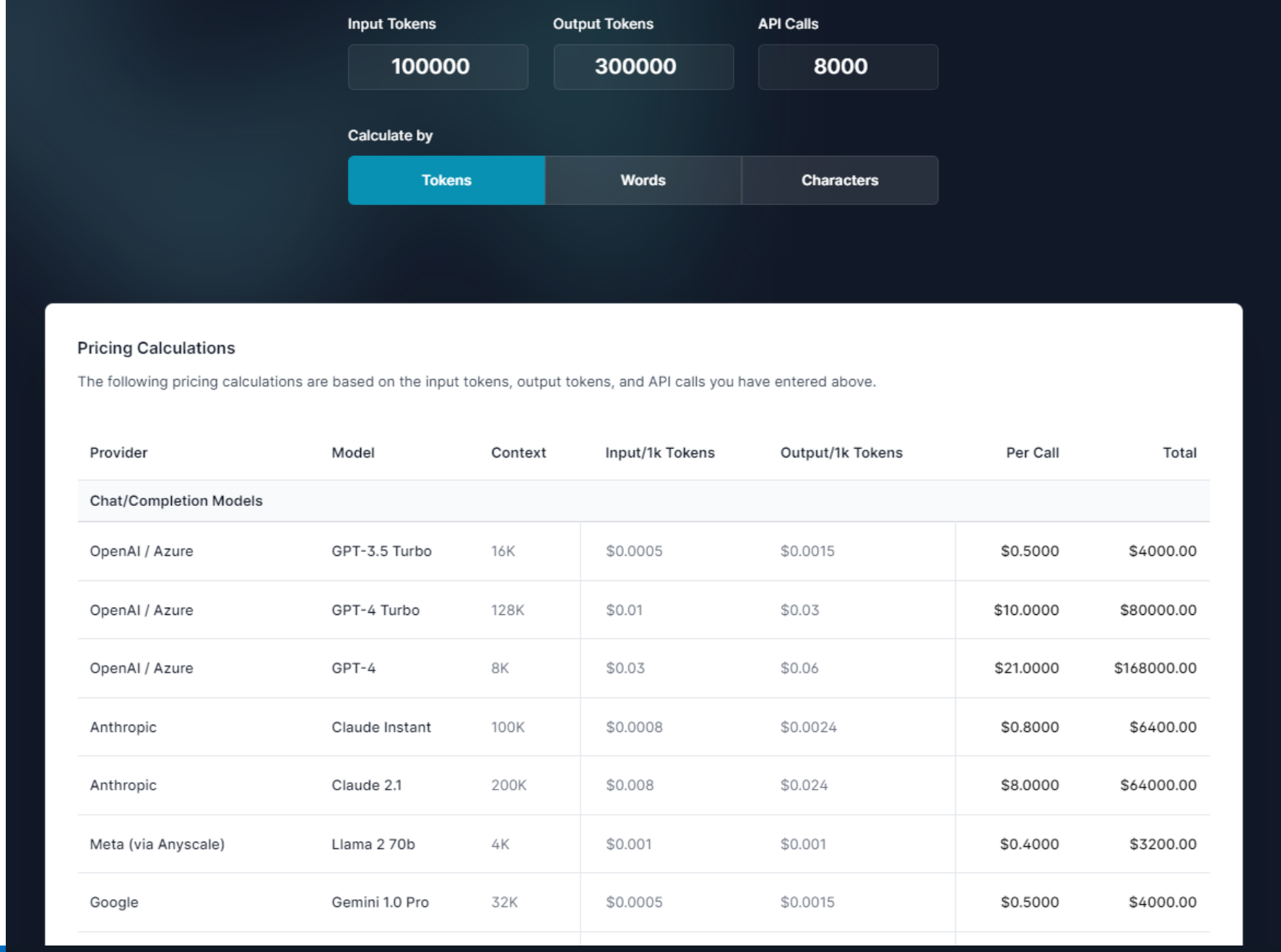
- Local device
- Cloud
- File Server

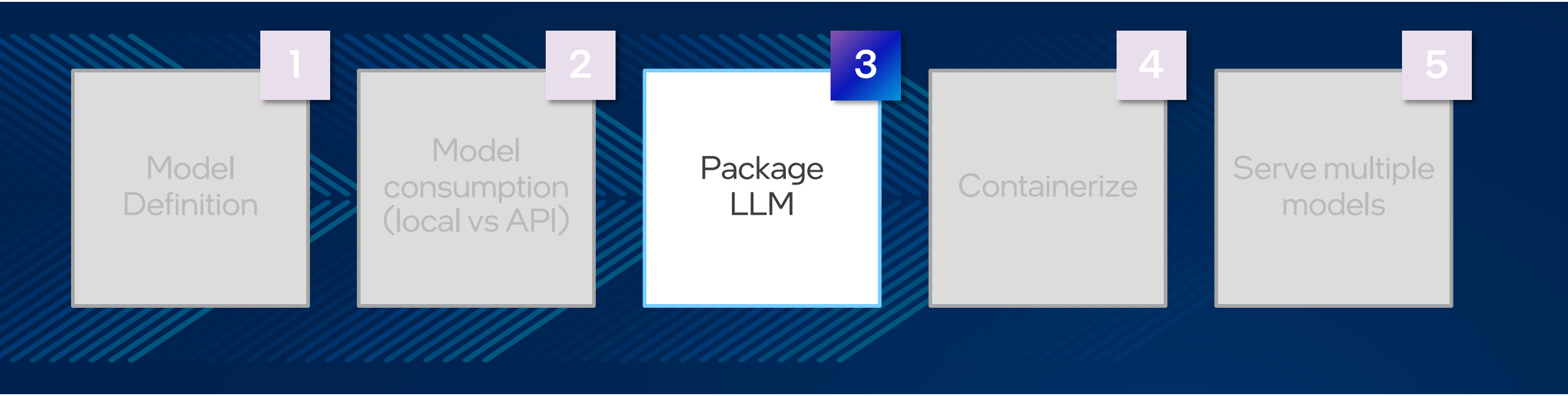
2 Model Consumption: External Inference



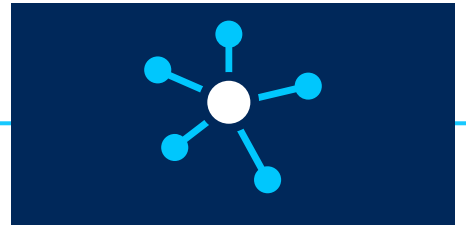
OpenAI & other LLM API Pricing Calculator

https://docsbot.ai/tools/gpt-openai-api-pricing-calculator




1
Model Definition
2
Model consumption (local vs API)
3
Package LLM
4
Containerize
5
Serve multiple models

3 Challenges and Requirements of Serving Multiple LLMs



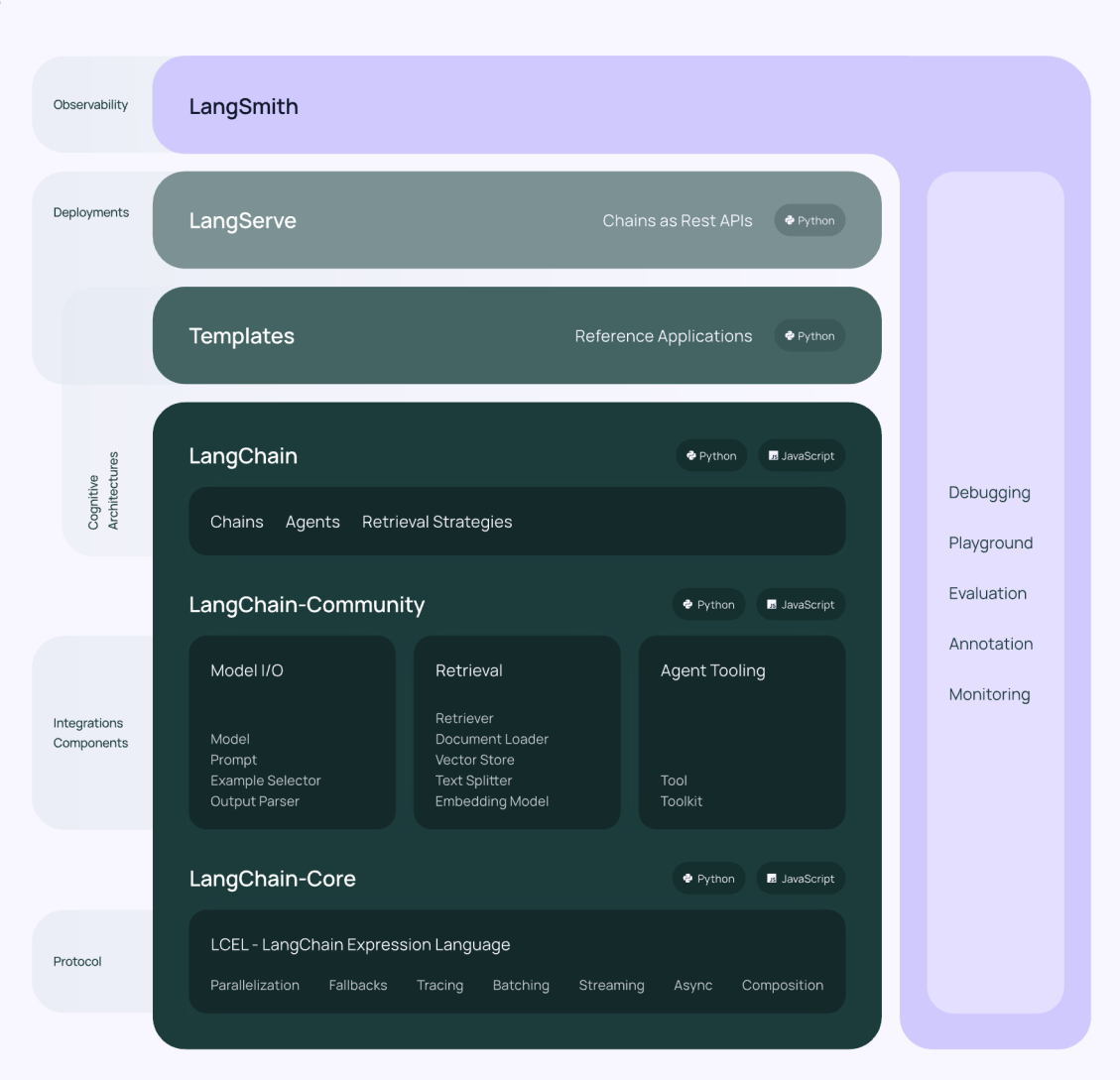
Business needs different types of LLMs

Each model has different compute/storage requirements

Each model has a different way to interact

Business needs a unified way to interact with models

3 Exposing LLMs: LangChain

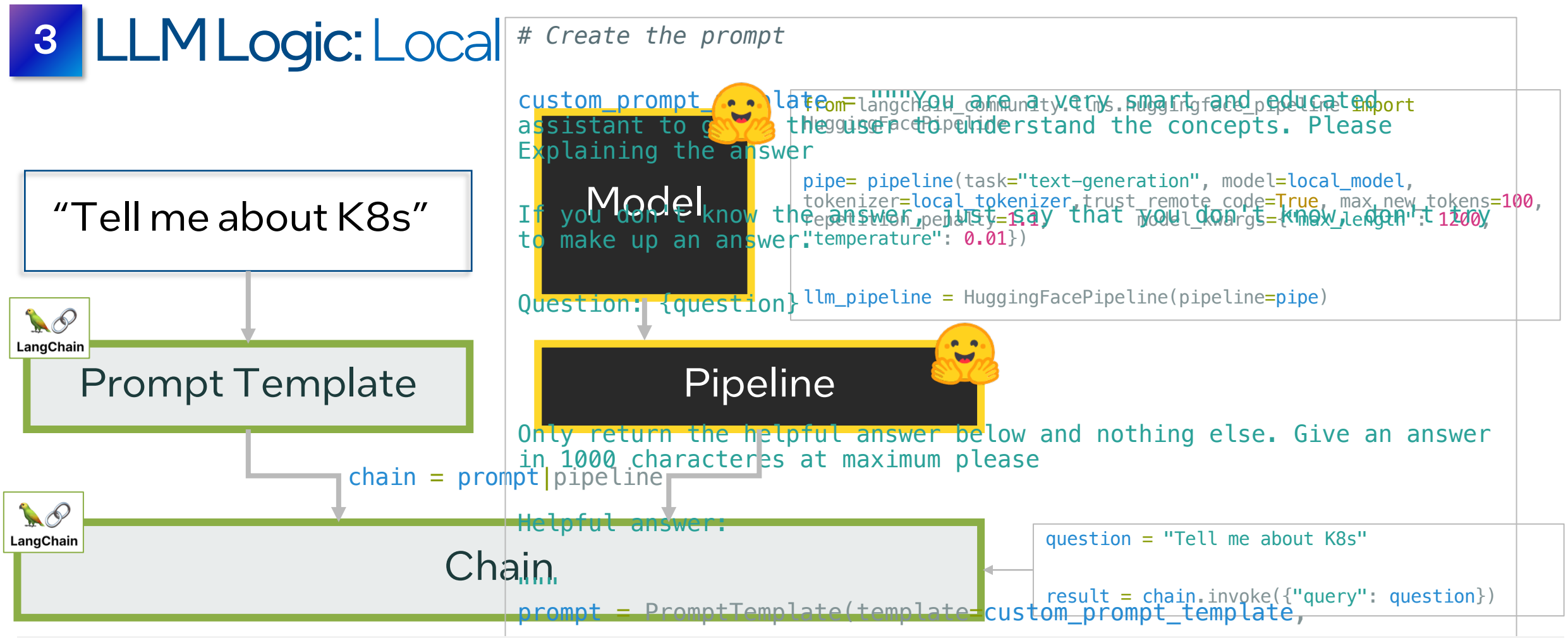


LangChain: A framework for building apps powered by LLMs

- Python and JS/Typescript library
- Native support for 80+ LLMs, open source models supported by templates
- Supports RAG pipelines, 75+ vector stores
- LangServe: Deploy LangChain chains as REST API
- LangSmith: Developer platform

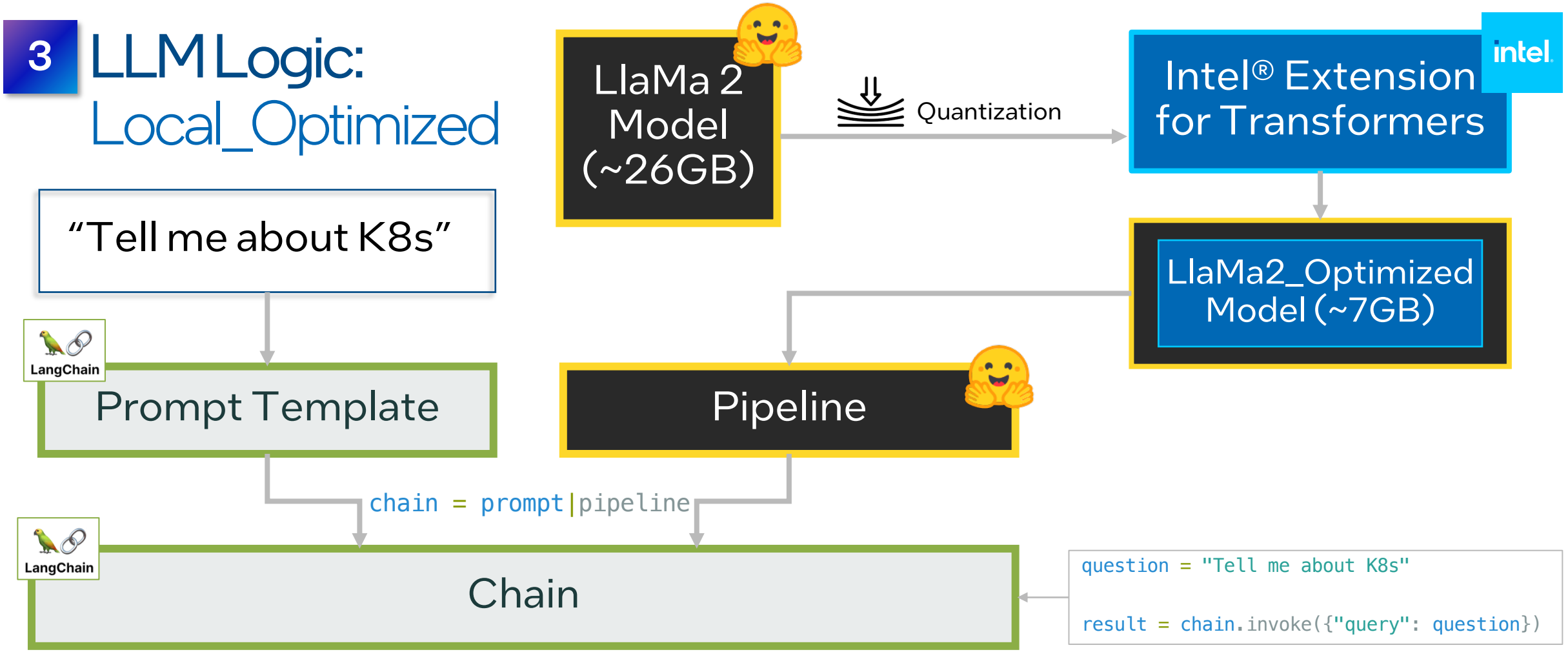
Image from https://github.com/langchain-ai/langchain/blob/master/docs/static/svg/langchain_stack.svg

3 LLM Logic: Local



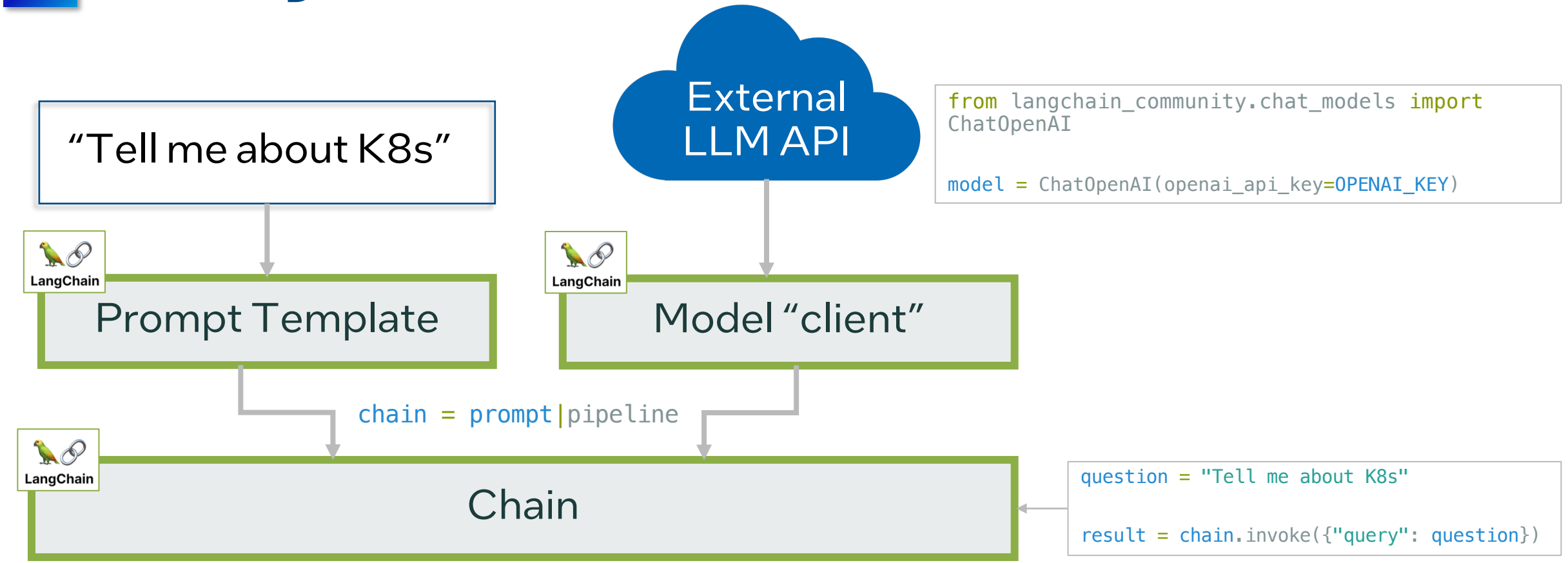
K8s, short for Kubernetes, is an open-source platform designed to automate deploying, scaling, and managing containerized applications. It allows users to easily manage multiple containers across clusters of hosts. Kubernetes provides features such as load balancing, self-healing, storage orchestration, and automated rollouts and rollbacks. It has become a popular tool for managing containerized applications in production environments due to its flexibility, scalability, and robustness. Overall, Kubernetes simplifies the process of managing containers and helps organizations efficiently deploy and scale their applications.

3 LLM Logic: Local_Optimized

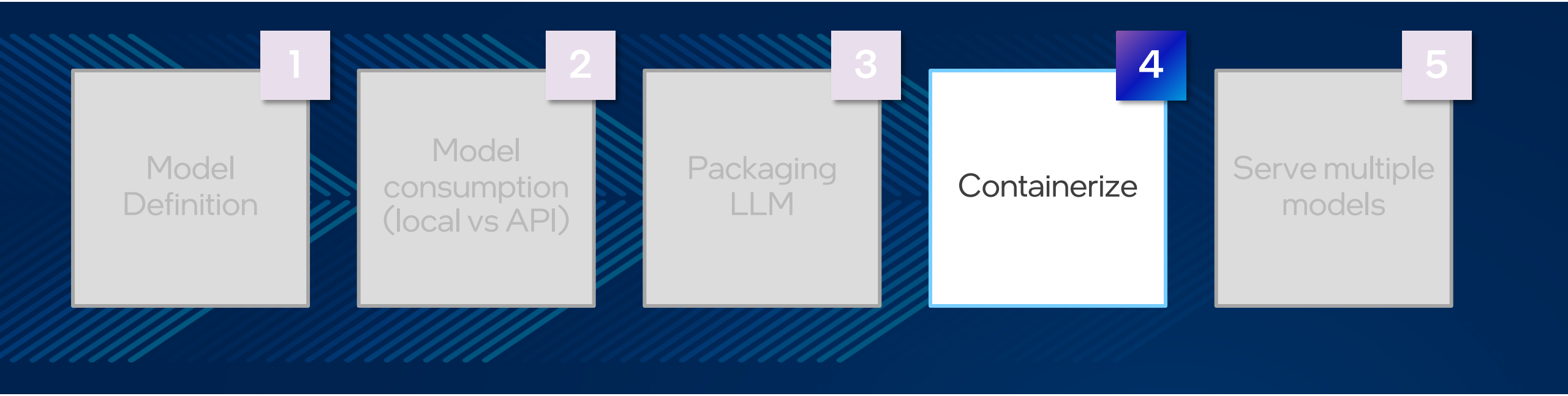


K8s, short for Kubernetes, is an open-source platform designed to automate deploying, scaling, and managing containerized applications. It allows users to easily manage multiple containers across clusters of hosts. Kubernetes provides features such as load balancing, self-healing, storage orchestration, and automated rollouts and rollbacks. It has become a popular tool for managing containerized applications in production environments due to its flexibility, scalability, and robustness. Overall, Kubernetes simplifies the process of managing containers and helps organizations efficiently deploy and scale their applications.

3 LLM Logic: External



K8s, short for Kubernetes, is an open-source platform designed to automate deploying, scaling, and managing containerized applications. It allows users to easily manage multiple containers across clusters of hosts. Kubernetes provides features such as load balancing, self-healing, storage orchestration, and automated rollouts and rollbacks. It has become a popular tool for managing containerized applications in production environments due to its flexibility, scalability, and robustness. Overall, Kubernetes simplifies the process of managing containers and helps organizations efficiently deploy and scale their applications.

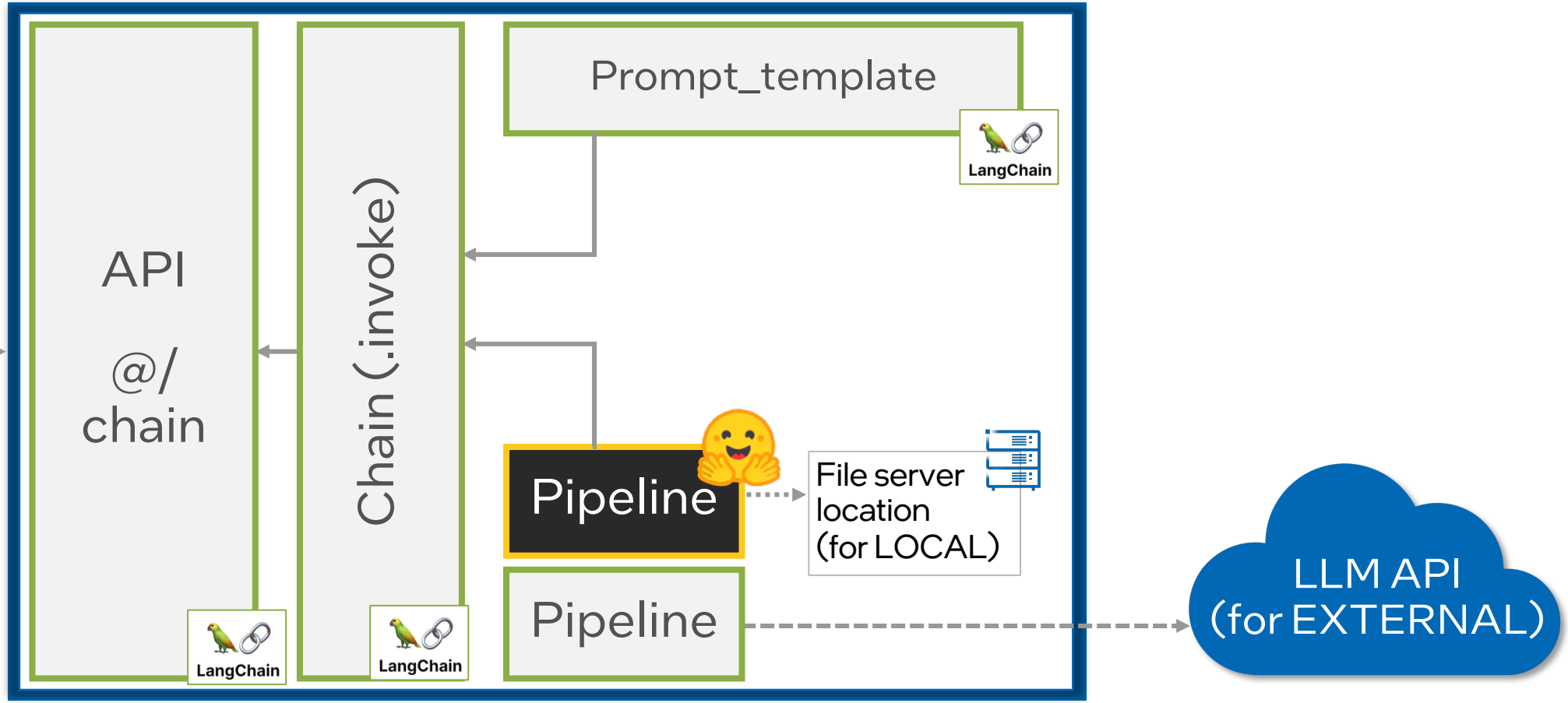

1
Model Definition
2
Model consumption (local vs API)
3
Packaging LLM
4
Containerize
5
Serve multiple models

Cloud Native is Platform of Choice for deploying LLMs

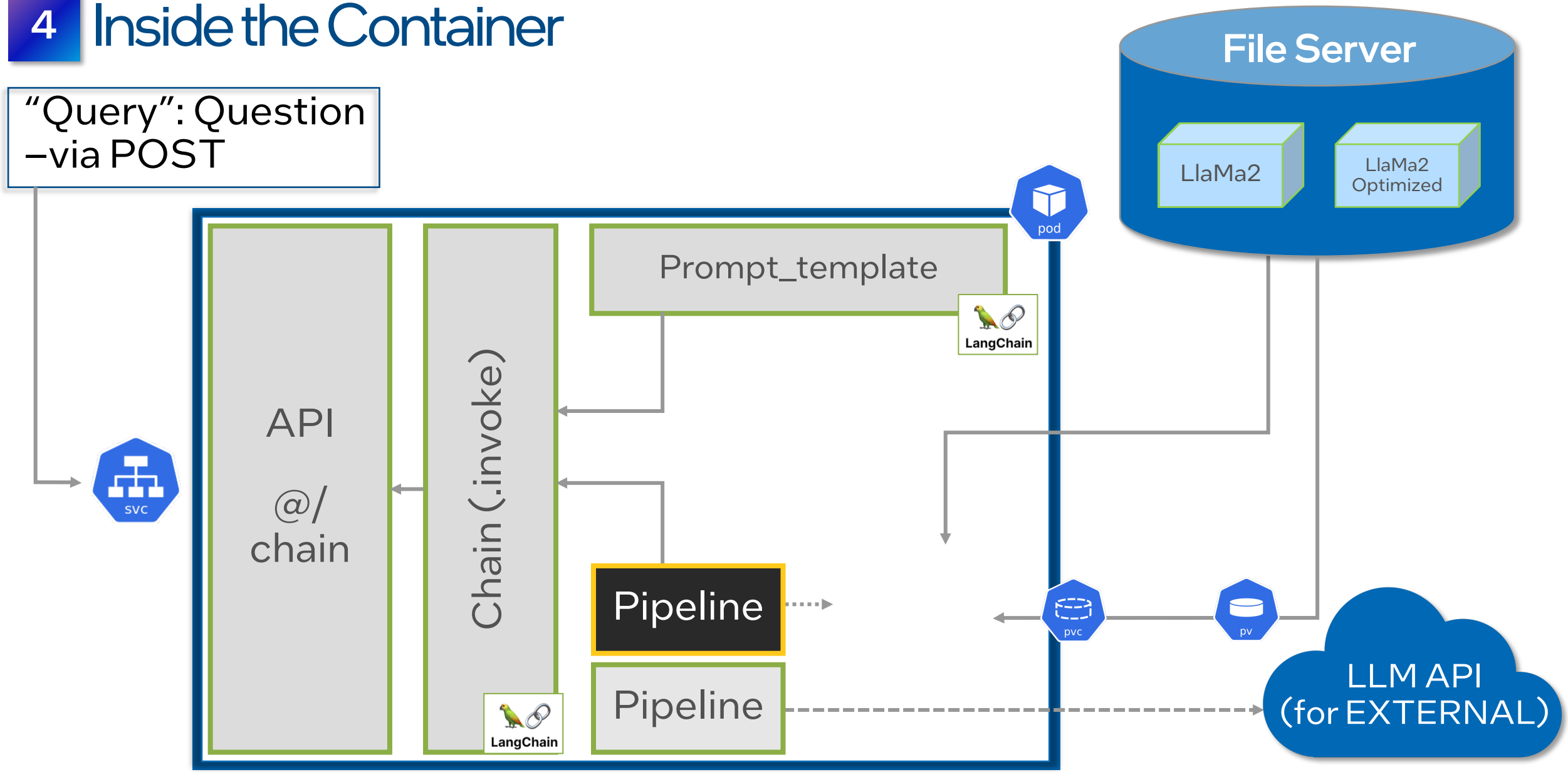
Scalability	Resource Management	Portability	Observability
Right-scale LLM deployments	CPU/memory limits, resource quotas, priority classes	On-prem/ cloud/edge or small/large node clusters	Long-running processes

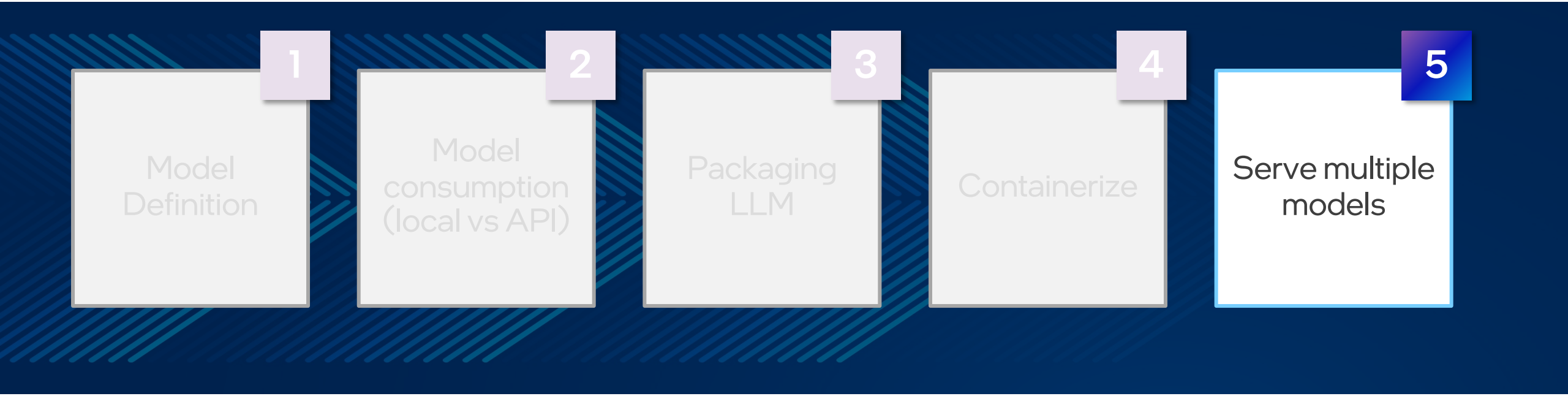
4 Inside the Container

"Query": Question
–via POST

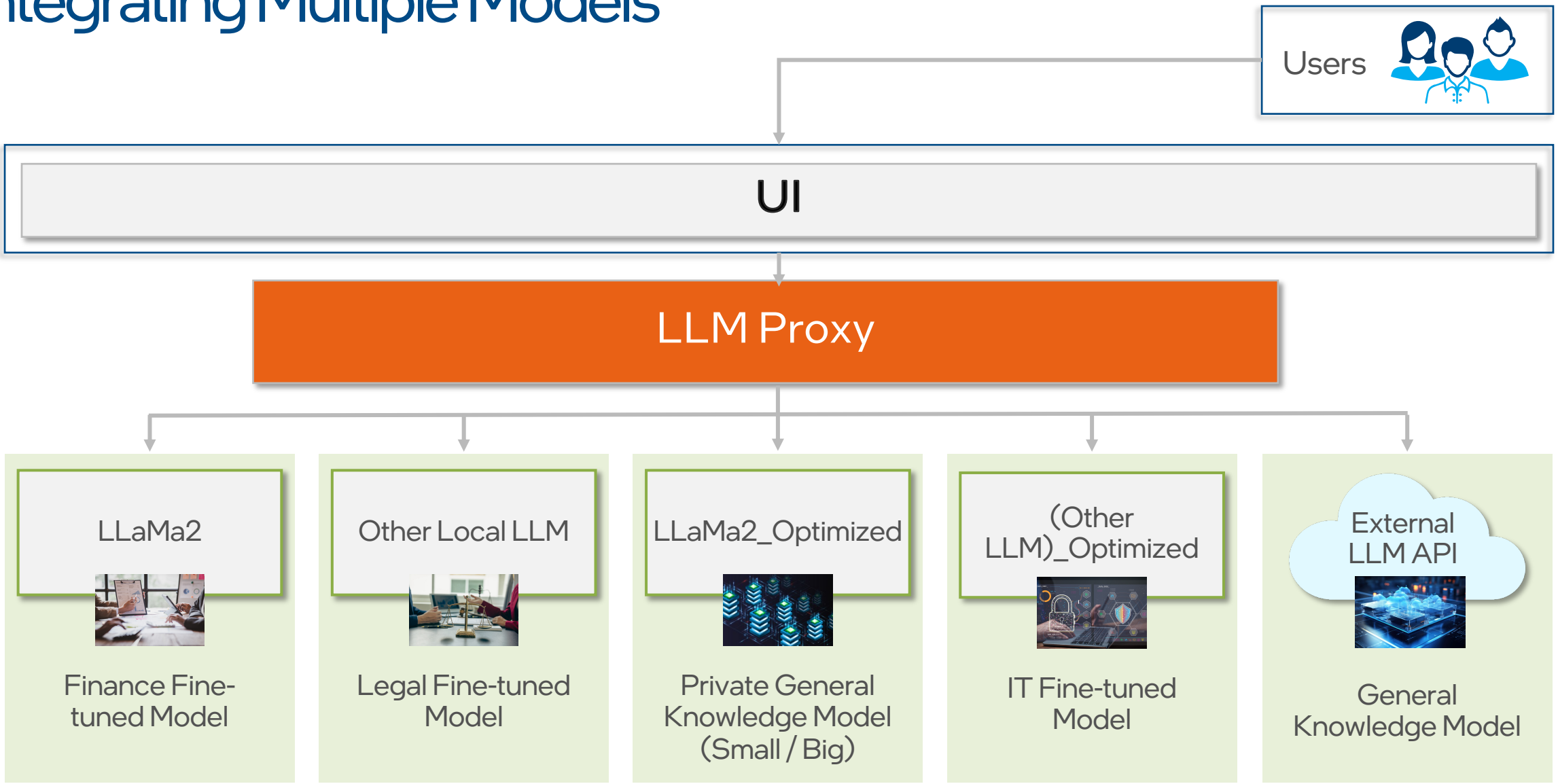


4 Inside the Container




1
Model Definition
2
Model consumption (local vs API)
3
Packaging LLM
4
Containerize
5
Serve multiple models

5 Integrating Multiple Models



5 Integrating Multiple Models

Multiple pods

Users

Horizontal Scaling

Front_end Worker

UI

LLM Proxy

Adapter Pod

Vertical Scaling

LLM_backend Worker

LLaMa2

Finance Fine-tuned Model

Other Local LLM

Legal Fine-tuned Model

LLaMa2_Optimized

Private General Knowledge Model (Small / Big)

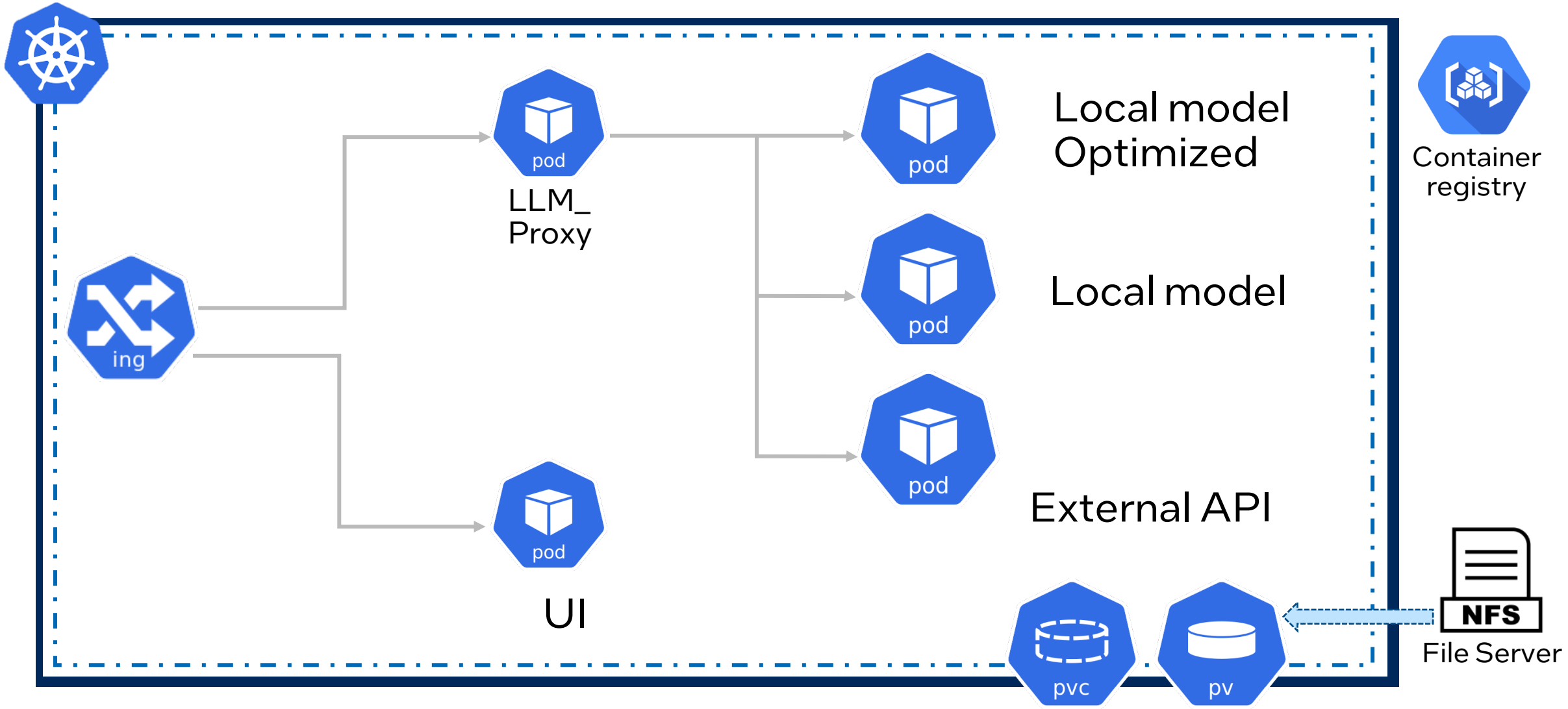
(Other LLM)_Optimized

IT Fine-tuned Model

External LLM API

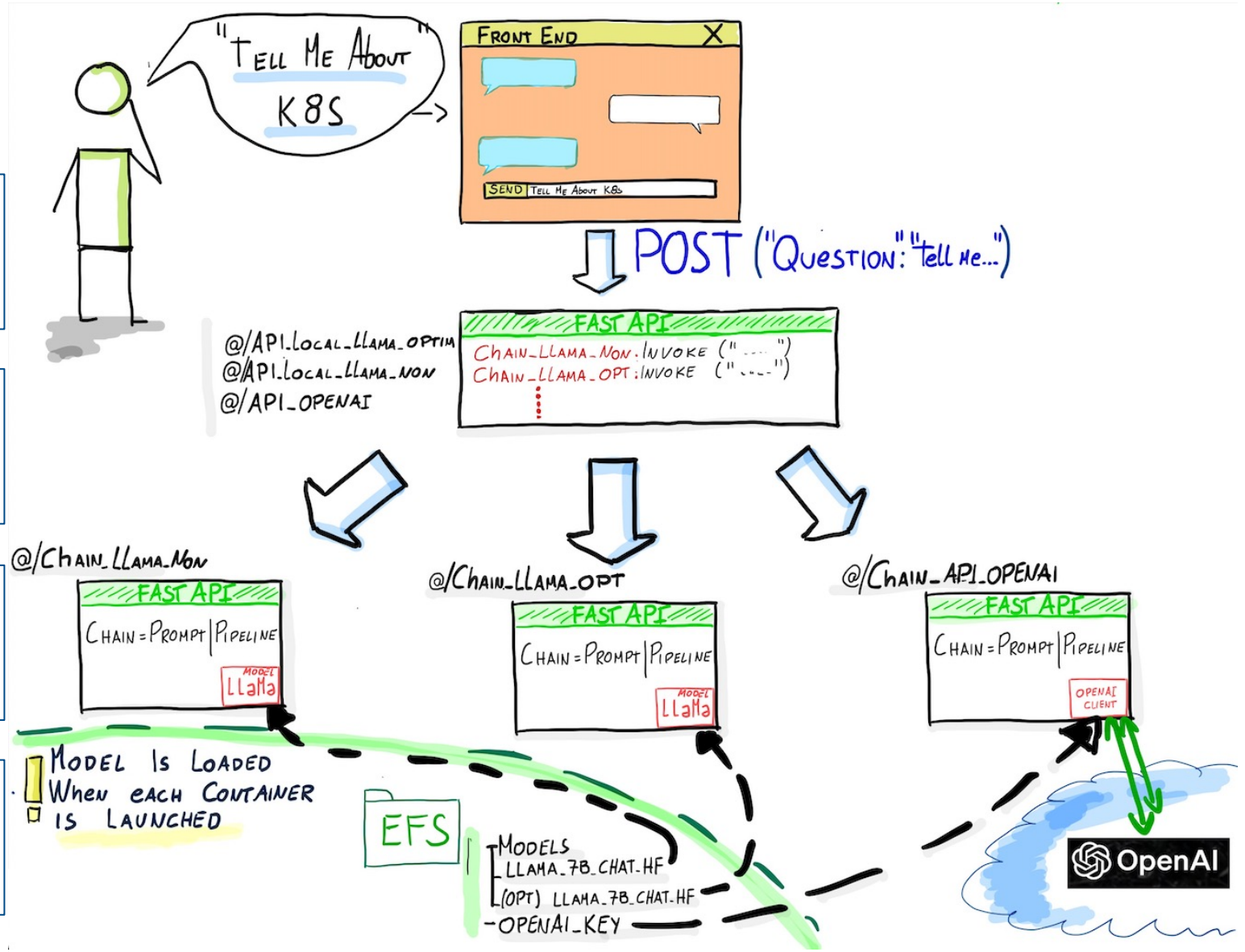
General Knowledge Model

5 Integrating Multiple Models: Multiple pods



RECAP:
Logic Architecture

Pre	Local models are downloaded at each container launch
1	UI (Front_end) sends the message to the LLM_proxy
2	LLM_proxy sends the message to the selected LLM (.INVOKE)
3	After processed the answer is sent back to the UI.



DEMO

Conclusions

- Choice of model depends on your business use case
- LangChain makes it simple to use multiple models in the same environment
- Cloud Native is the Platform of Choice for deploying LLMs
- Optimizations play a key role in enabling local deployments

Call to action

DEMO
Git clone / contribute

Open.intel
Website

Open.intel
Podcast

Notices and disclaimers

Intel is committed to respecting human rights and avoiding complicity in human rights abuses. See Intel's Global Human Rights Principles. Intel® products and software are intended only to be used in applications that do not cause or contribute to a violation of an internationally recognized human right.

Intel® technologies may require enabled hardware, software, or service activation. No product or component can be absolutely secure. Your costs and results may vary.