

KubeCon + CloudNativeCon Europe 2024

Deploying LLMs in CloudNative using LangChain

Ezequiel Lanza
AI Open Source Evangelist

Arun Gupta
VP/GM Open Ecosystem





Everybody wants
an **assistant**

Photo by [Ant Rozetsky](#) on [Unsplash](#)

Companies require a variety of solutions



Finance



Legal

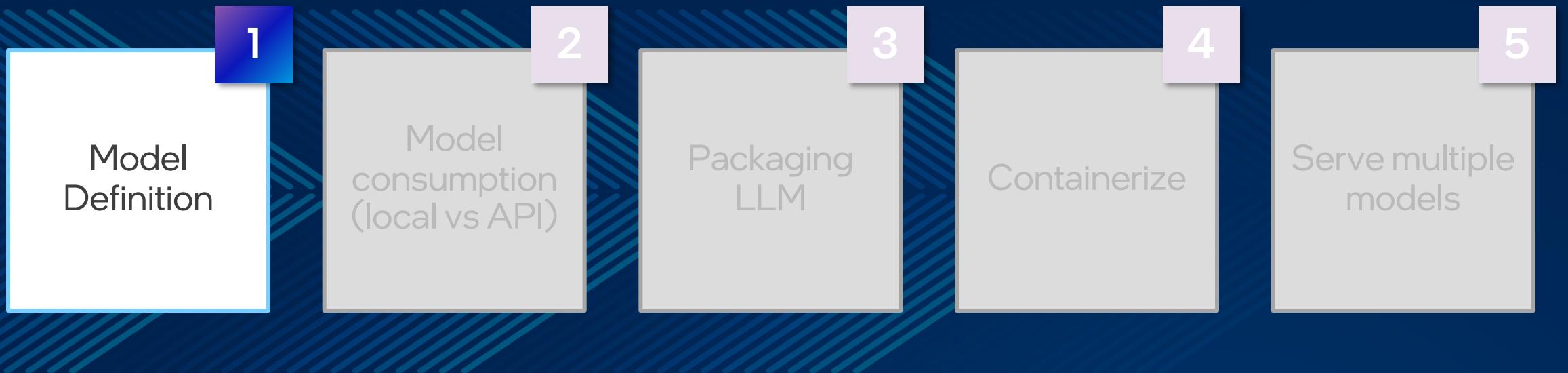


IT

...and others

One Way to Deploy your LLM in CloudNative







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)

Model definition: Considerations for picking a model

Hugging Face Leaderboard

The screenshot shows the Hugging Face Leaderboard interface. At the top, there are navigation links: LLM Benchmark, Metrics through time, About, and Submit here!. Below these are search and filter sections. The search bar contains the placeholder "Search for your model (separate multiple queries with `;` and press ENTER...)".

Model types: Pretrained (selected), fine-tuned on domain-specific datasets, chat models (RLHF, DPO, IFT, ...), base merges and moerges.

Precision: float16, bfloat16, 8bit, 4bit, GPTQ.

Model sizes (in billions of parameters): ~1.5, ~3, ~7, ~13, ~35, ~60, 70+.

Hide models: Private or deleted, Contains a merge/moerge, Flagged, MoE.

Table: Model, Average, ARC, Hellas.

| Model | Average | ARC | Hellas |
|---|---------|-------|--------|
| abacusai/Smaug-72B-v0.1 | 80.48 | 76.02 | 89.27 |
| ibivibiv/alpaca-dragon-72b-v1 | 79.3 | 73.89 | 88.16 |
| moreh/MoMo-72B-lora-1.8.7-DPO | 78.55 | 70.82 | 85.96 |
| cloudyu/TomGrc_FusionNet_34Bx2_MoE_v0.1_DPO_f16 | 77.91 | 74.06 | 86.74 |
| HanNayeoniee/LHK_DPO_v1 | 77.62 | 74.74 | 89.3 |
| cloudyu/TomGrc_FusionNet_34Bx2_MoE_v0.1_full_linear_DPO | 77.52 | 74.06 | 86.67 |

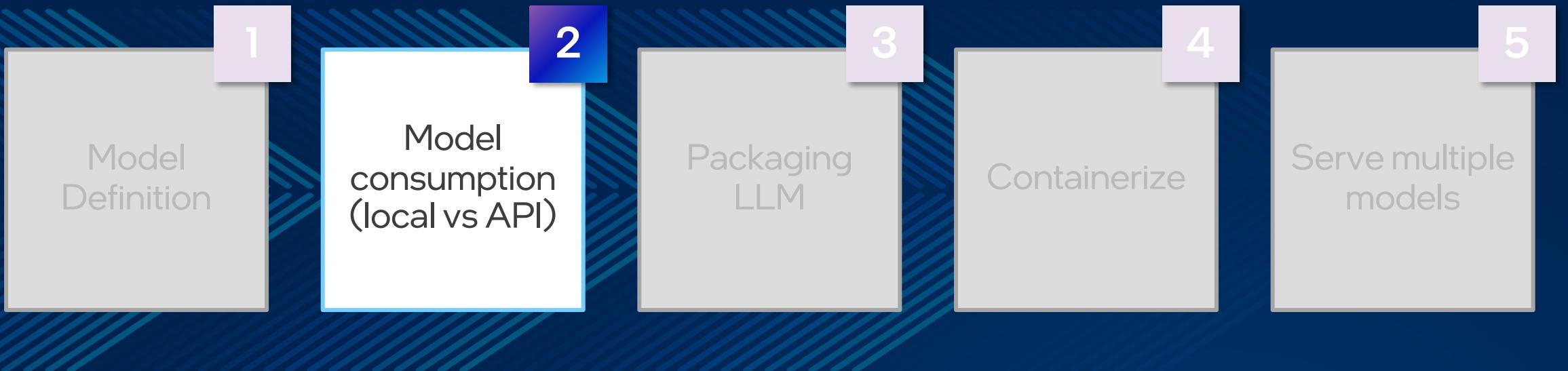
QR code at the bottom left.

Documentation/Community

Adoption
Support
Tutorials
Forums

Ethical considerations

Training Data
Bias



2

Model Consumption: Local vs External



Model Consumption: Considerations

Local Model



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

External Model

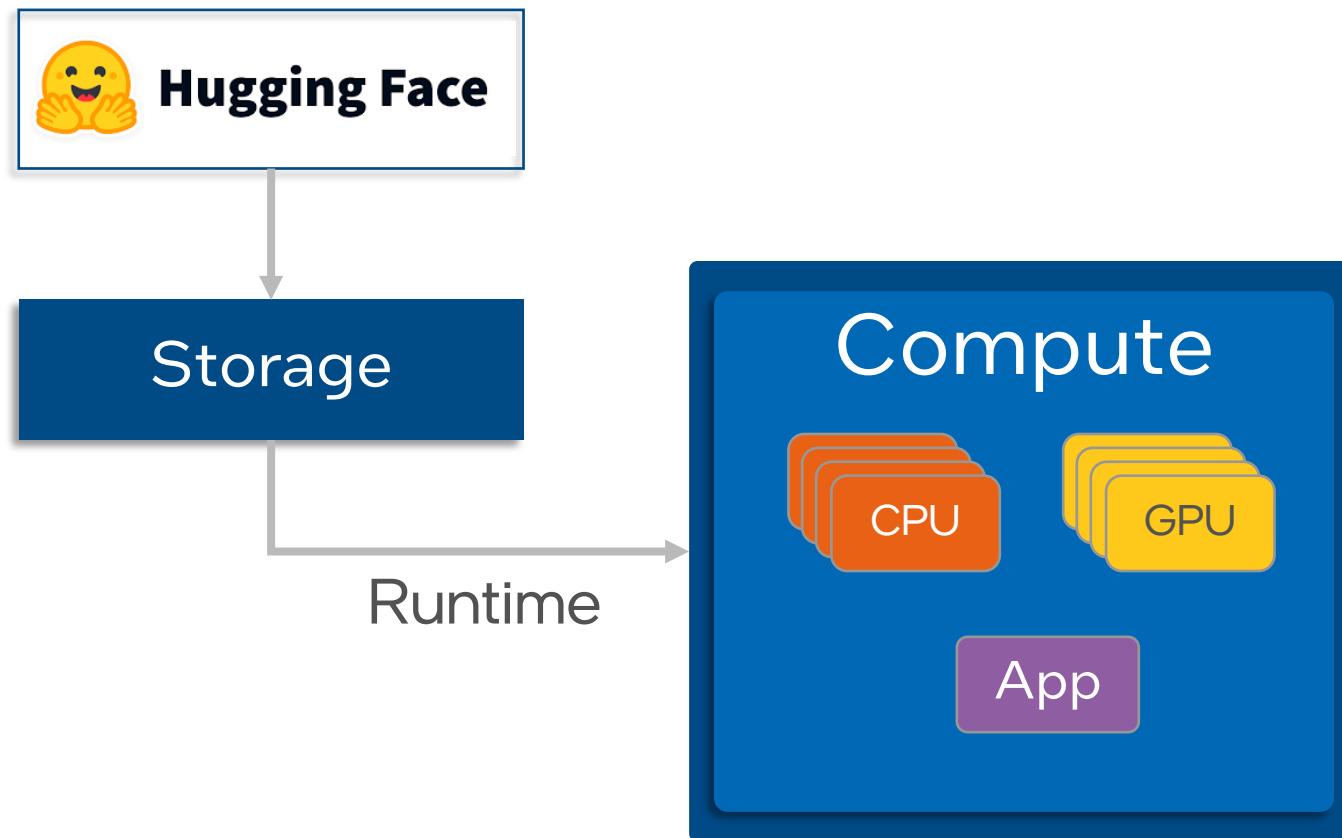


- Deliver quickly
- Scalability
- Less complex setup
- Elasticity
- High availability



2

Model Consumption: Local Inference



Considerations

Size (RAM)

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

Access

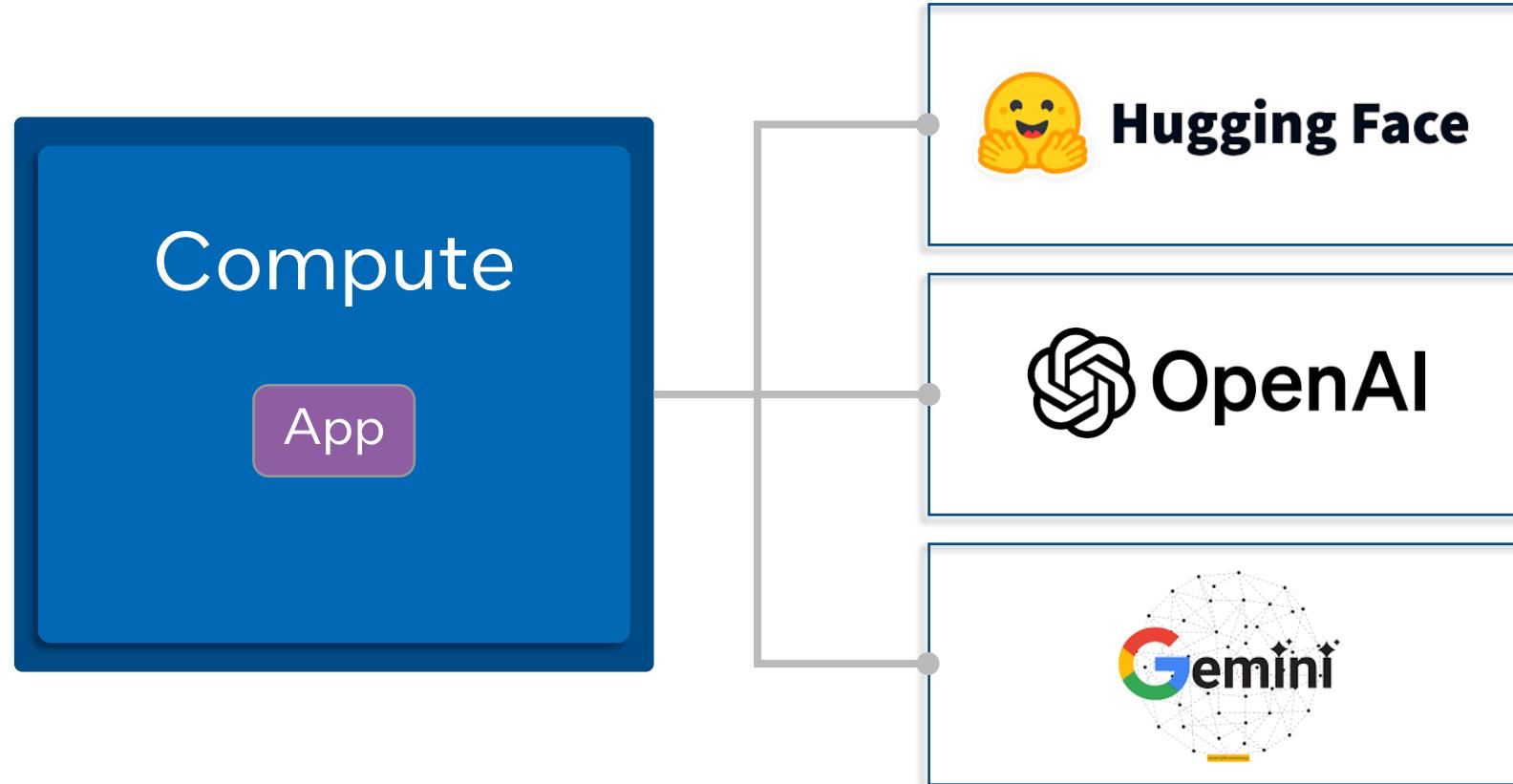
- Local device
- Cloud
- File Server





2

Model Consumption: External Inference



OpenAI & other LLM API Pricing Calculator

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

Input Tokens Output Tokens API Calls

100000 300000 8000

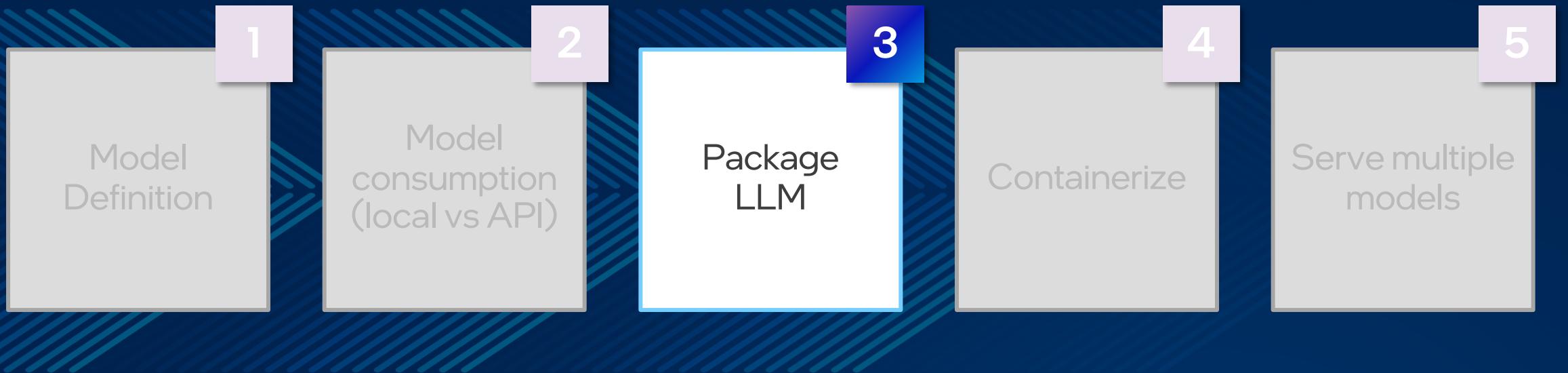
Calculate by

Tokens Words Characters

Pricing Calculations

The following pricing calculations are based on the input tokens, output tokens, and API calls you have entered above.

| Provider | Model | Context | Input/1k Tokens | Output/1k Tokens | Per Call | Total |
|------------------------|----------------|---------|-----------------|------------------|-----------|-------------|
| Chat/Completion Models | | | | | | |
| OpenAI / Azure | GPT-3.5 Turbo | 16K | \$0.0005 | \$0.0015 | \$0.5000 | \$4000.00 |
| OpenAI / Azure | GPT-4 Turbo | 128K | \$0.01 | \$0.03 | \$10.0000 | \$80000.00 |
| OpenAI / Azure | GPT-4 | 8K | \$0.03 | \$0.06 | \$21.0000 | \$168000.00 |
| Anthropic | Claude Instant | 100K | \$0.0008 | \$0.0024 | \$0.8000 | \$6400.00 |
| Anthropic | Claude 2.1 | 200K | \$0.008 | \$0.024 | \$8.0000 | \$64000.00 |
| Meta (via Anyscale) | Llama 2 70b | 4K | \$0.001 | \$0.001 | \$0.4000 | \$3200.00 |
| Google | Gemini 1.0 Pro | 32K | \$0.0005 | \$0.0015 | \$0.5000 | \$4000.00 |



3

Challenges and Requirements of Serving Multiple LLMs



Business needs
different types of LLMs



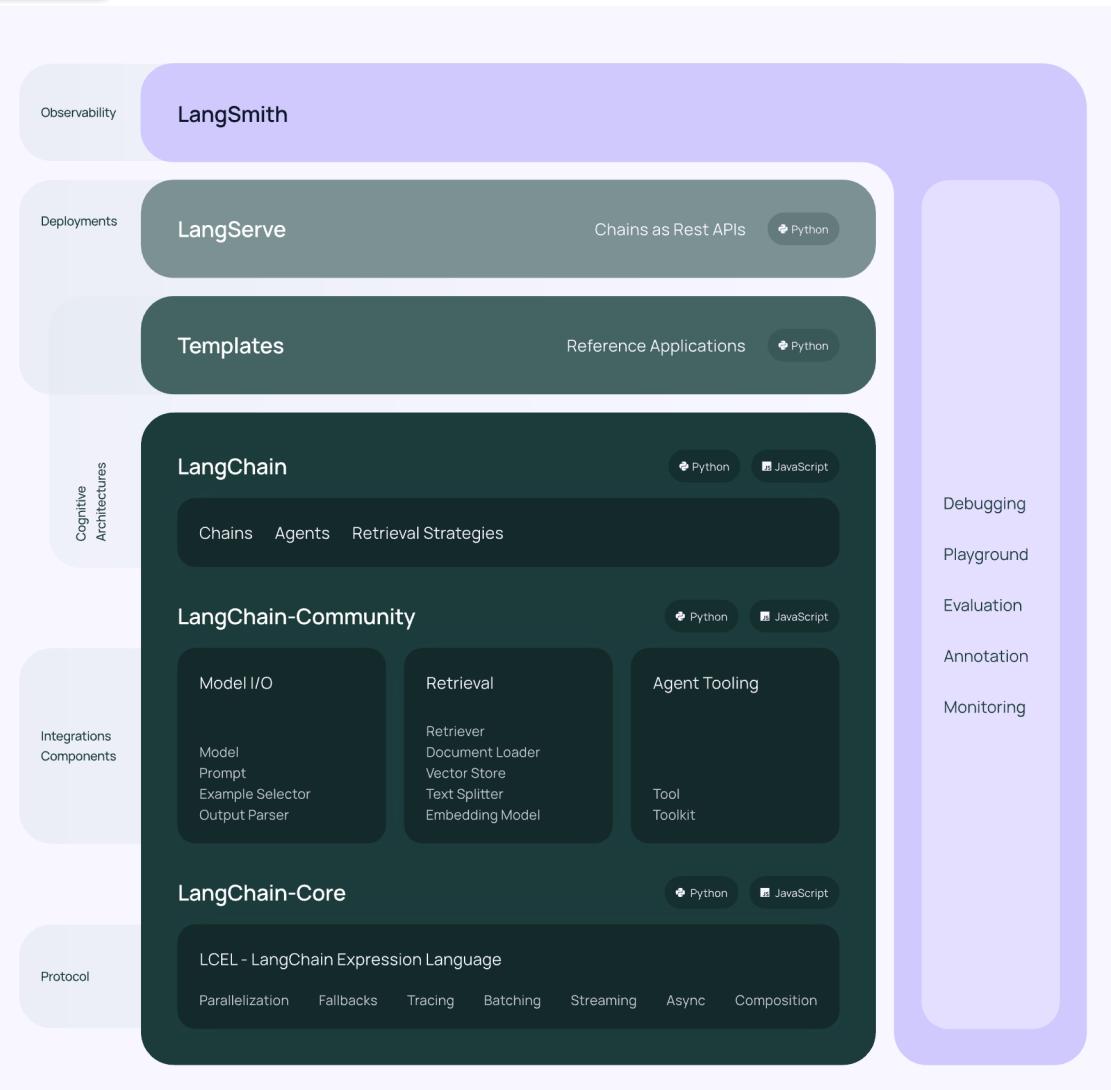
Each model has
different
compute/storage
requirements



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

"Tell me about K8s"



Prompt Template

`chain = prompt | pipeline`



Chain

`result = chain.invoke({"query": question})`

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.

```
# Create the prompt
```

```
custom_prompt_template="""
You are a very smart and educated
assistant to guide the user to understand the concepts. Please
Explain the answer
```

Model

```
If you don't know the answer, just say that you don't know, don't try
to make up an answer."temperature": 0.01})
```

Question: {question}



Pipeline

Only return the helpful answer below and nothing else. Give an answer
in 1000 characteres at maximum please

Helpful answer:

`question = "Tell me about K8s"`

`result = chain.invoke({"query": question})`

LLM Logic: Local_Optimized

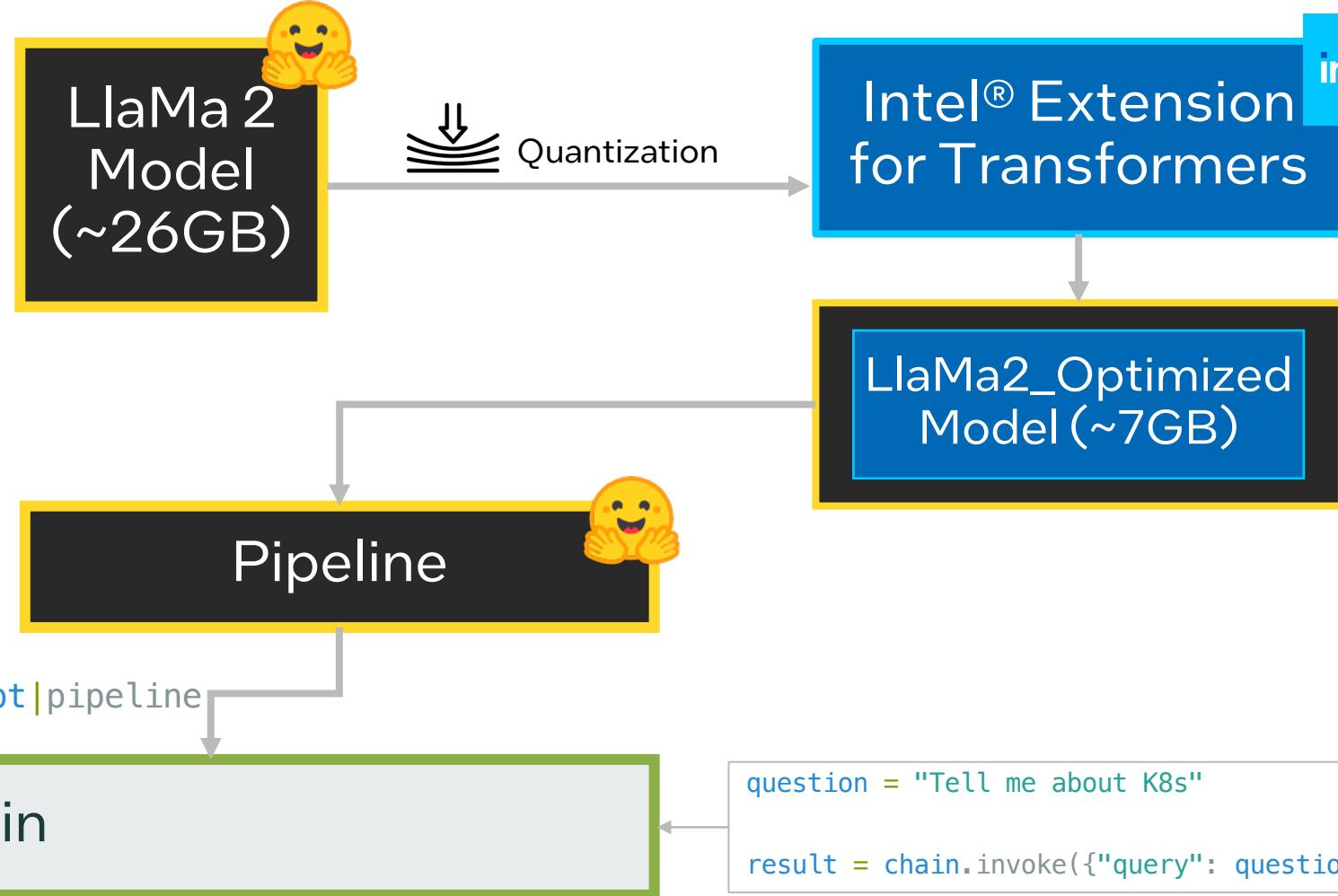
"Tell me about K8s"



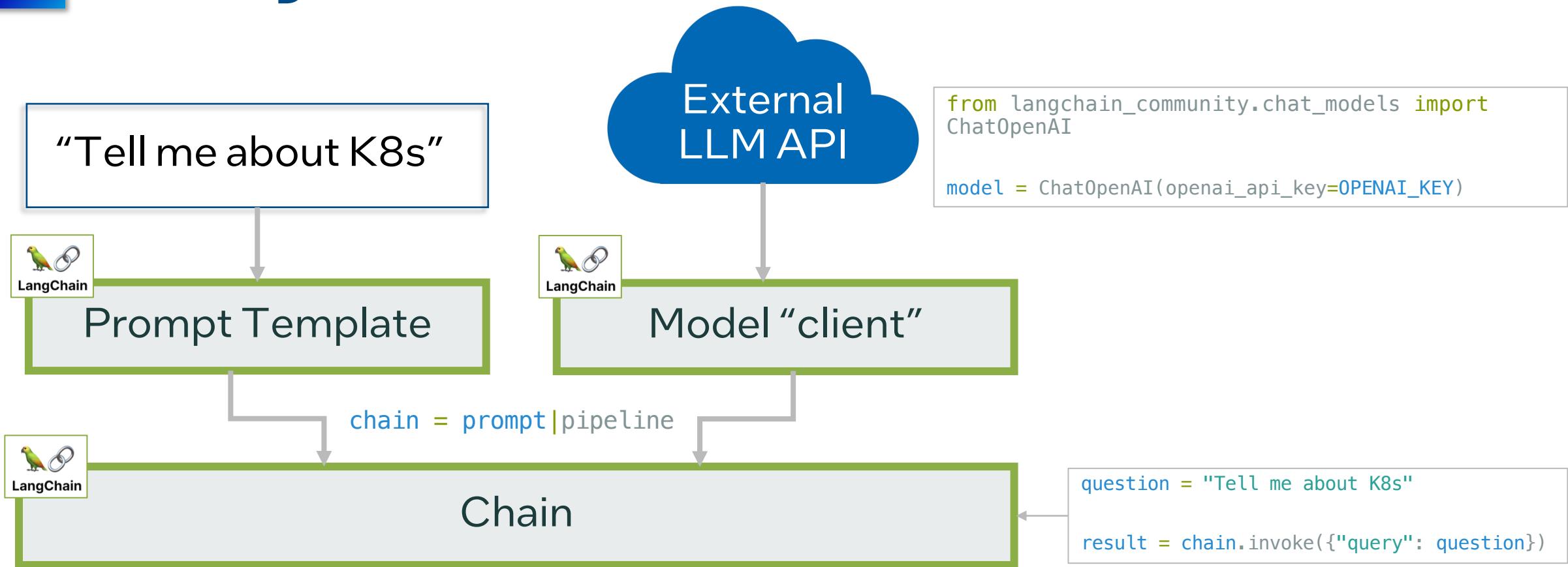
Prompt Template



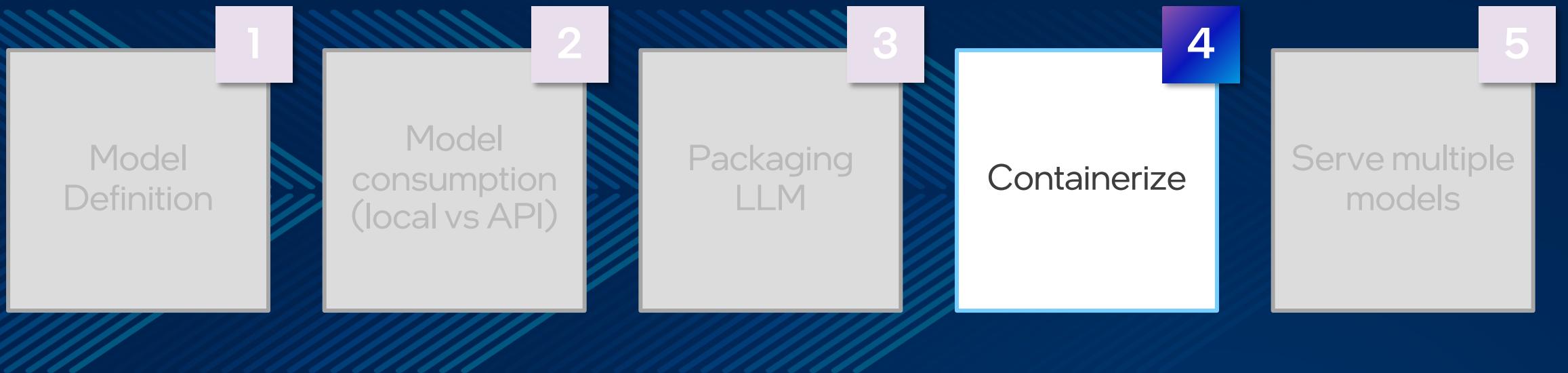
Chain



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.



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.



Cloud Native is Platform of Choice for deploying LLMs

Scalability

Right-scale LLM
deployments

Resource Management

CPU/memory
limits, resource
quotas, priority
classes

Portability

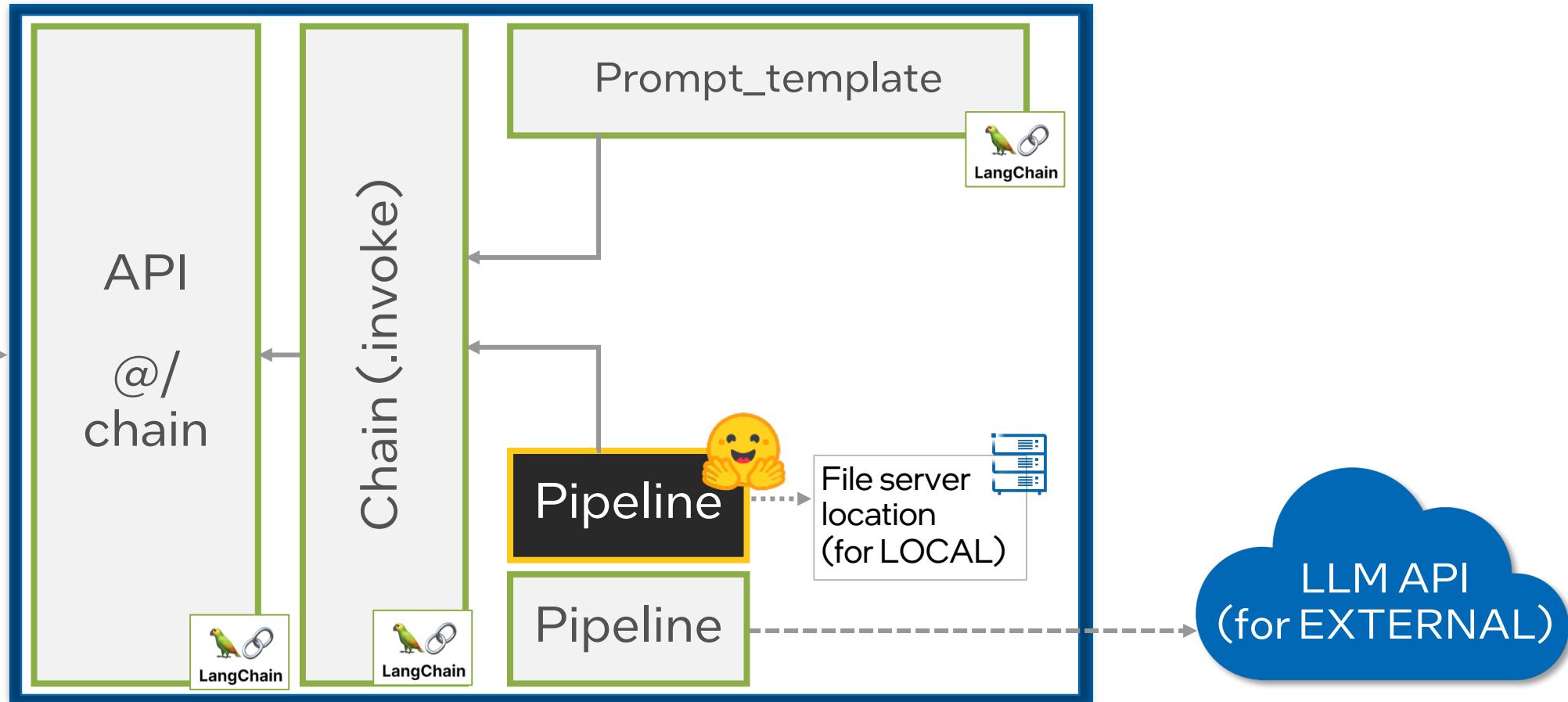
On-prem/
cloud/edge or
small/large node
clusters

Observability

Long-running
processes

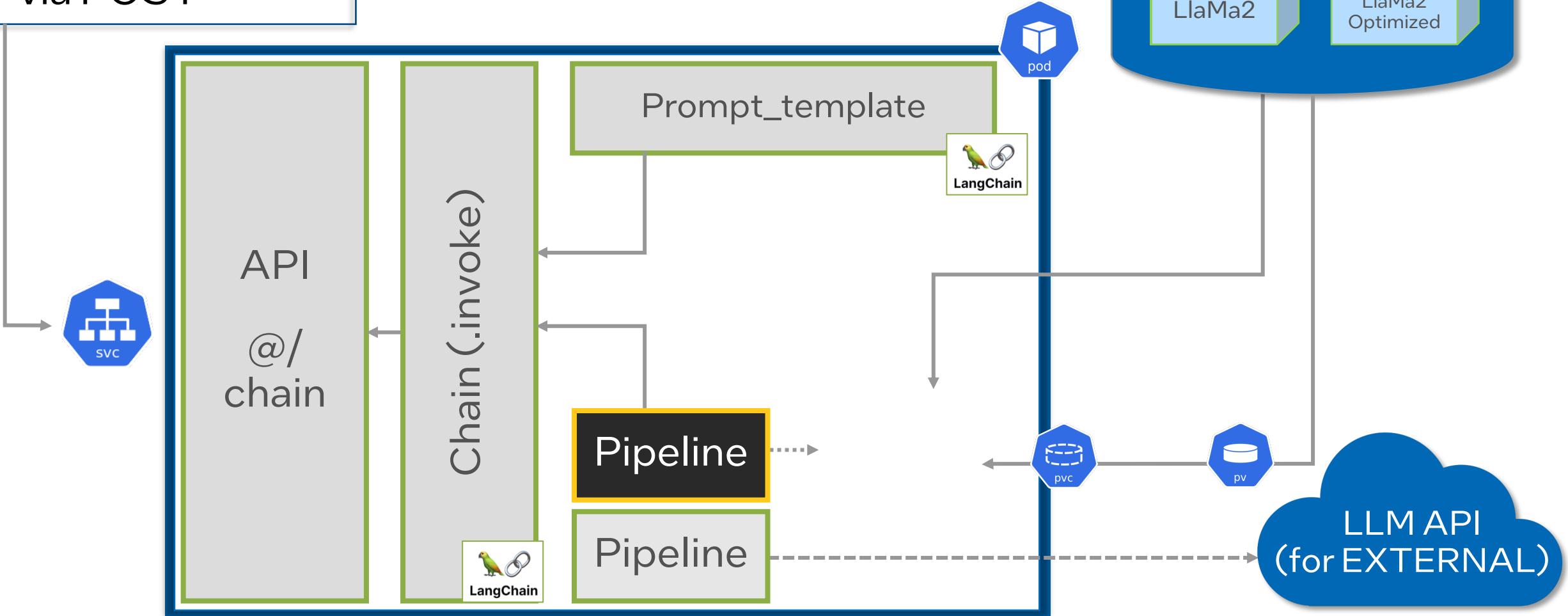


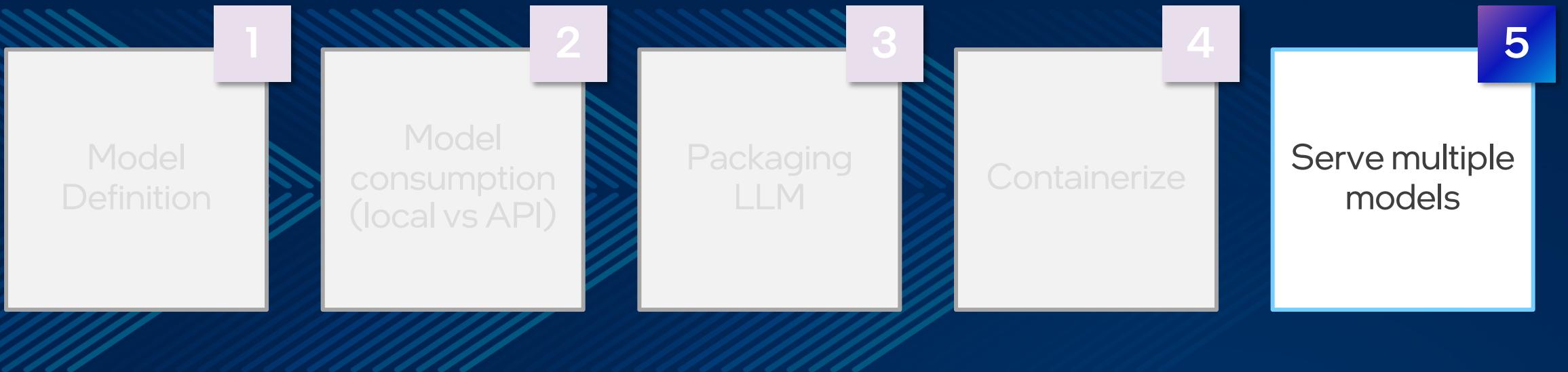
“Query”: Question
–via POST



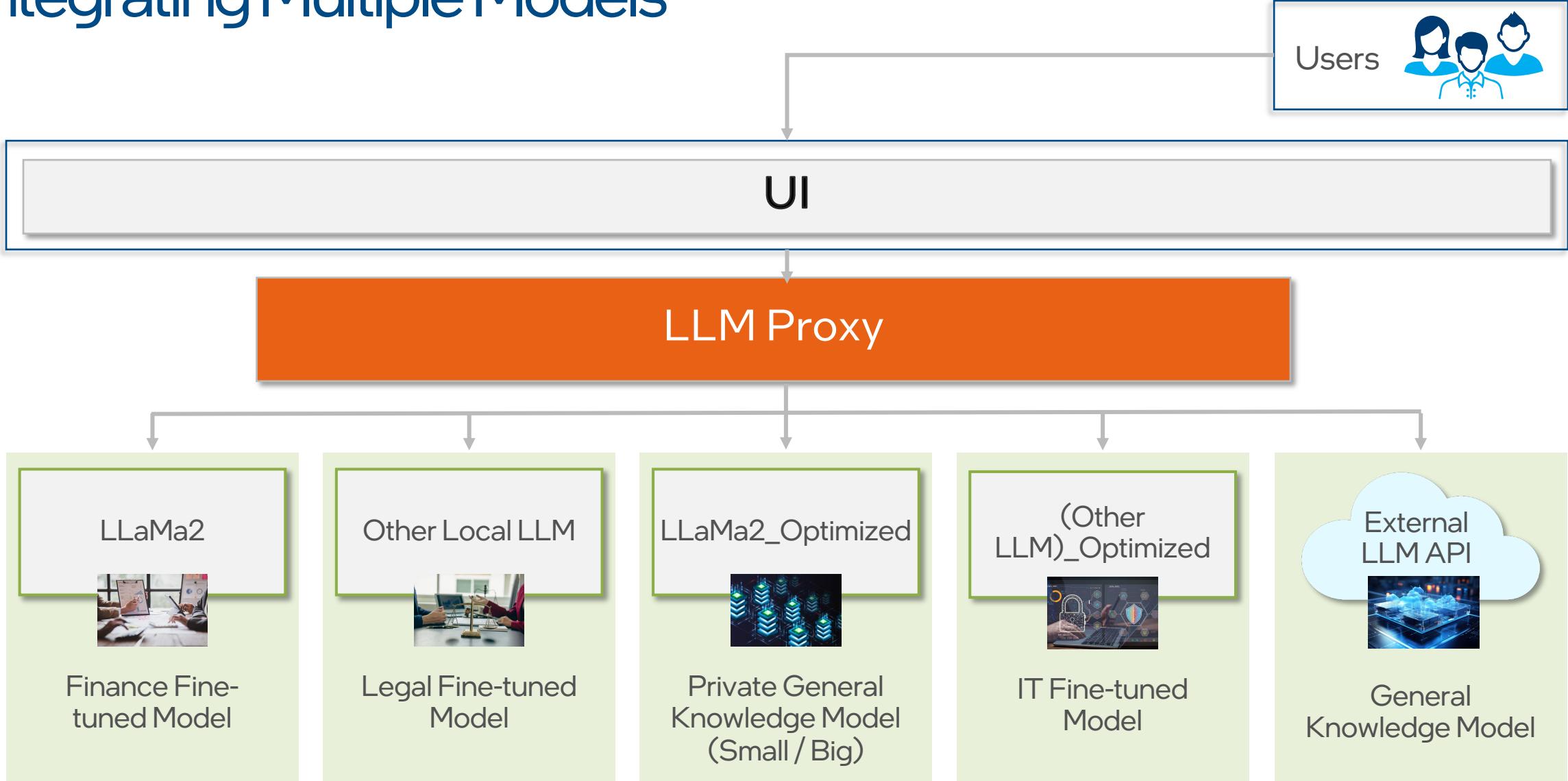
4 Inside the Container

“Query”: Question
–via POST



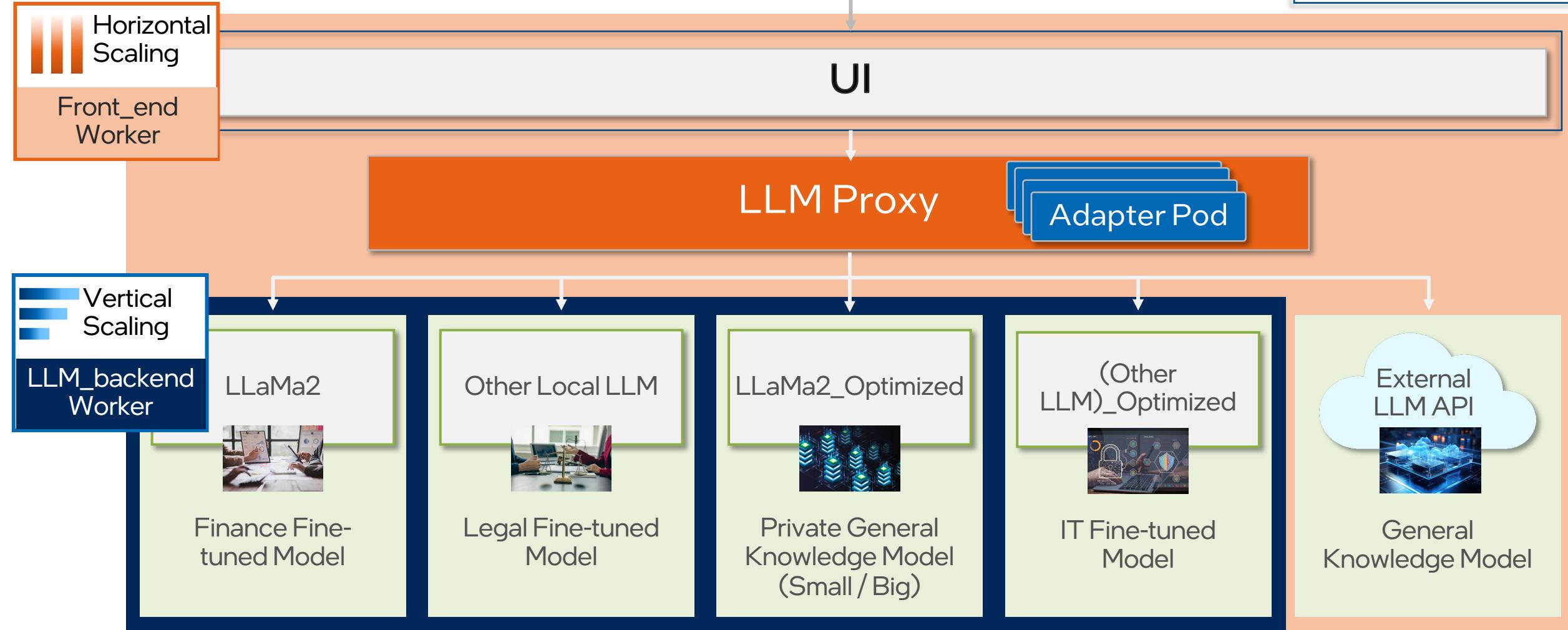


5 Integrating Multiple Models

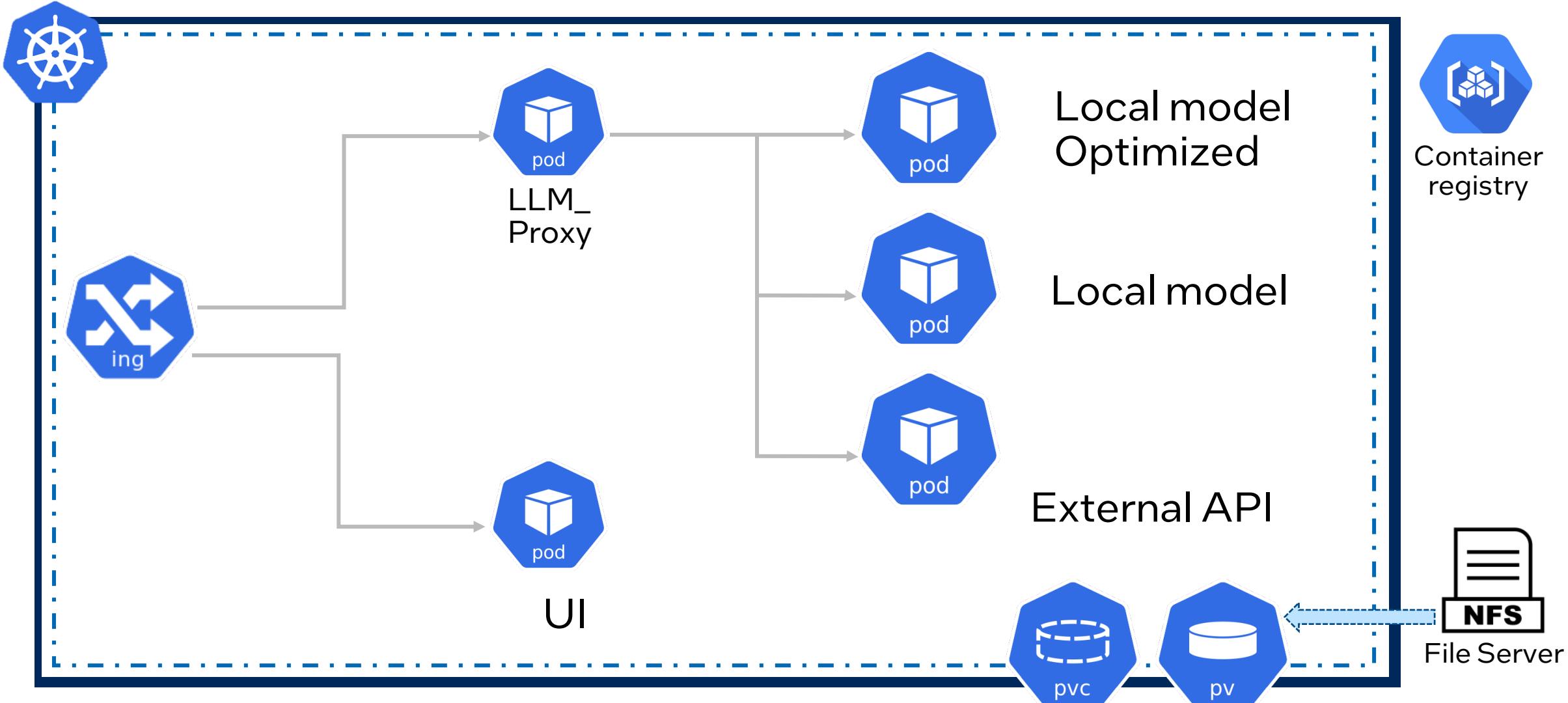


5 Integrating Multiple Models

Multiple pods



Integrating Multiple Models: Multiple pods



RECAP:

Logic Architecture

Pre

Local models are downloaded at each container launch

1

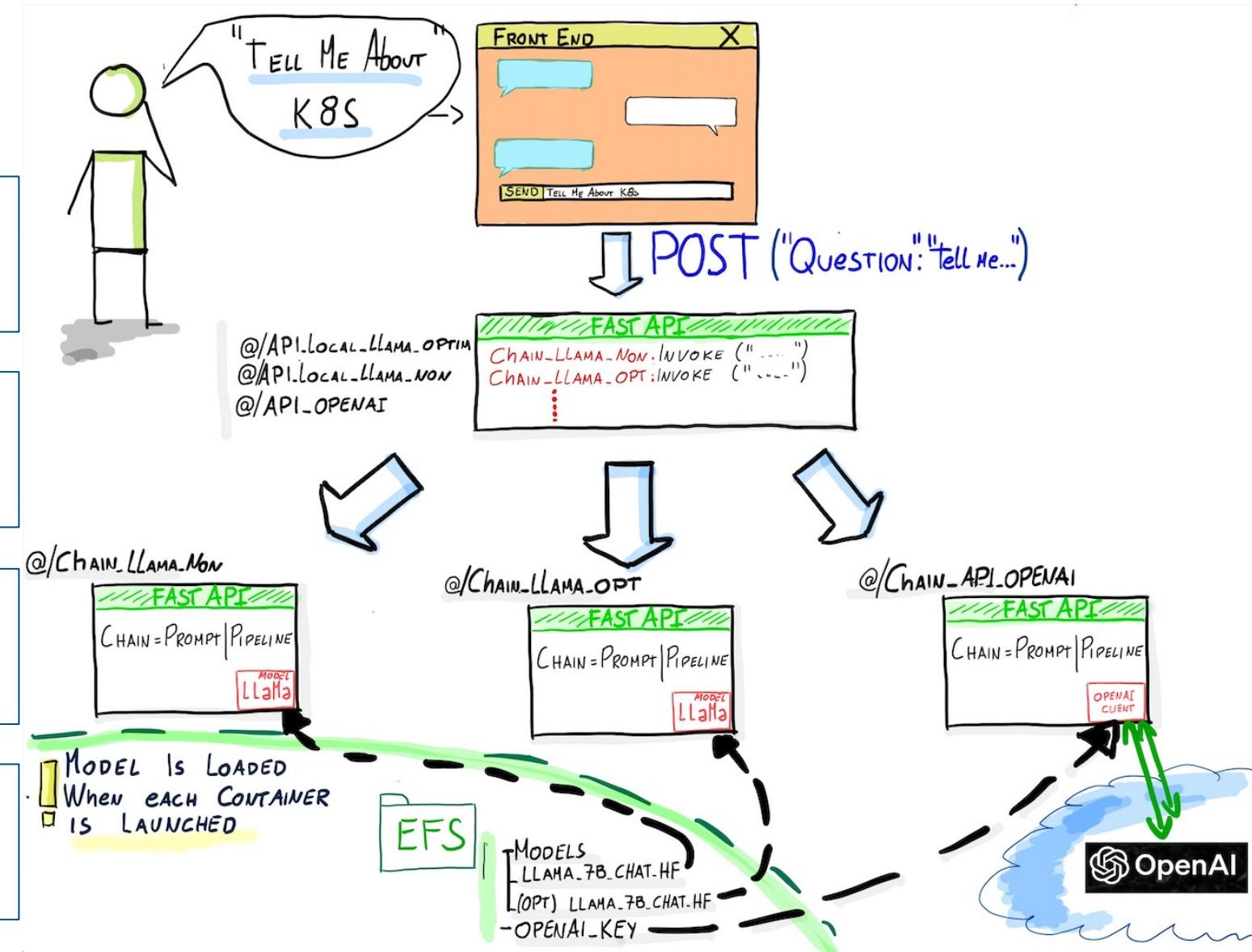
UI(Front_end) sends the message to the LLM_proxy

2

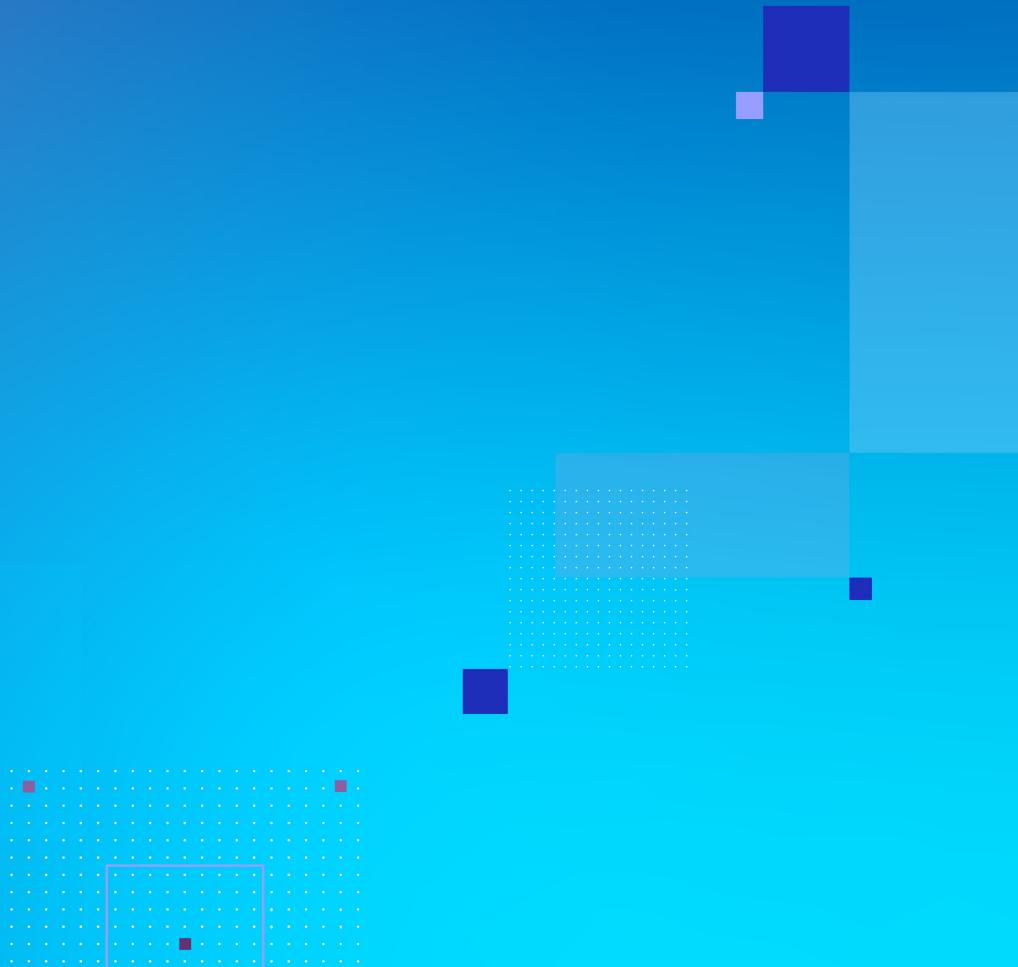
LLM_proxy sends the message to the selected LLM (.INVOKE)

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.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

