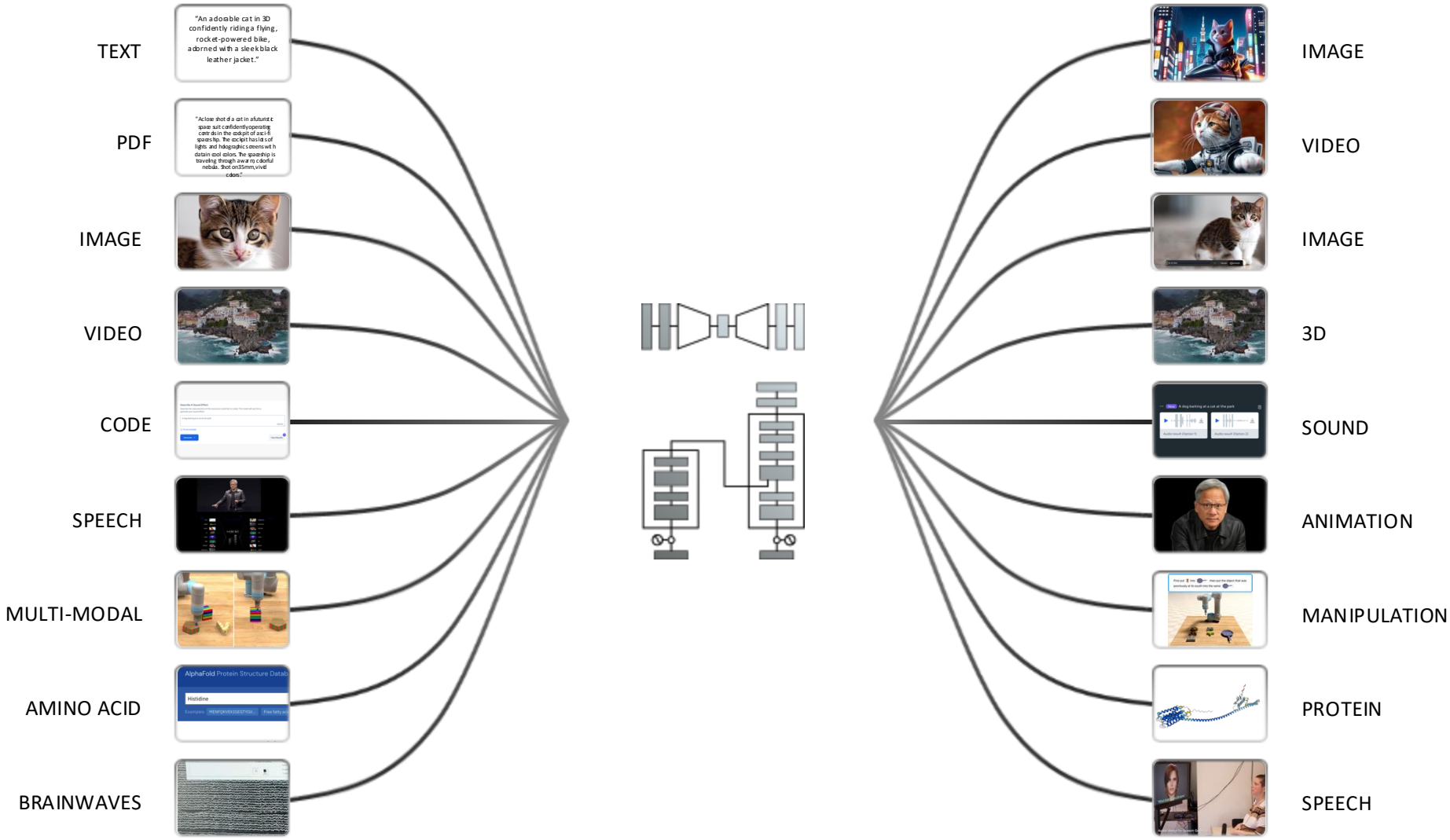
Accelerating Multimodal RAG Pipelines with NVIDIA and Open-Source Integration

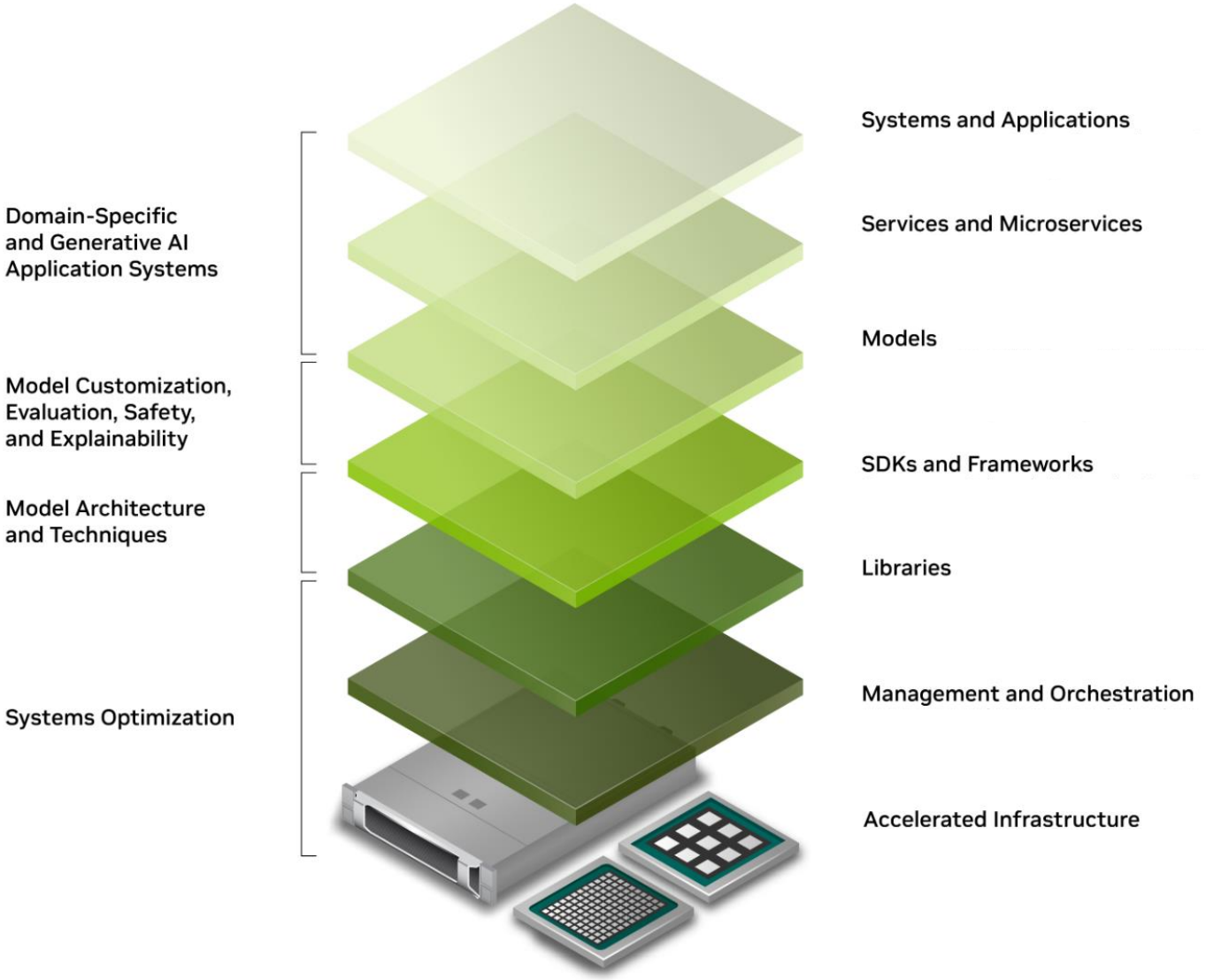
Jay Rodge, Developer Advocate - LLMs | AI, ML & Computer Vision Meetup

LLMs Can Learn and Understand Everything



NVIDIA Full Stack LLM Software Ecosystem

NVIDIA's deep, wide LLM ecosystem offers choices for developers across the entire stack.



NVIDIA Full Stack LLM Software Ecosystem

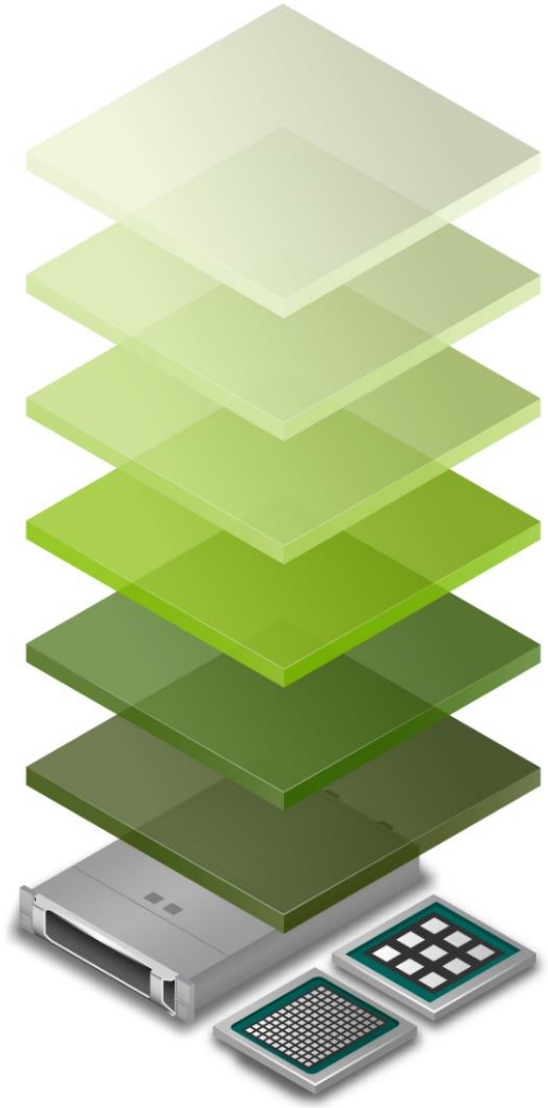
NVIDIA's deep, wide LLM ecosystem offers choices for developers across the entire stack.

Domain-Specific and Generative AI Application Systems

Model Customization, Evaluation, Safety, and Explainability

Model Architecture and Techniques

Systems Optimization



Systems and Applications
Ecosystem: LangChain, LlamaIndex, Weights & Biases
NVIDIA: AI Workbench, NeMo Guardrails

Services and Microservices
Ecosystem: AWS Bedrock, AzureML, Cohere,
Google Vertex AI, OpenAI APIs
NVIDIA: Avatar Cloud Engine (ACE), BioNeMo, NeMo, Picasso

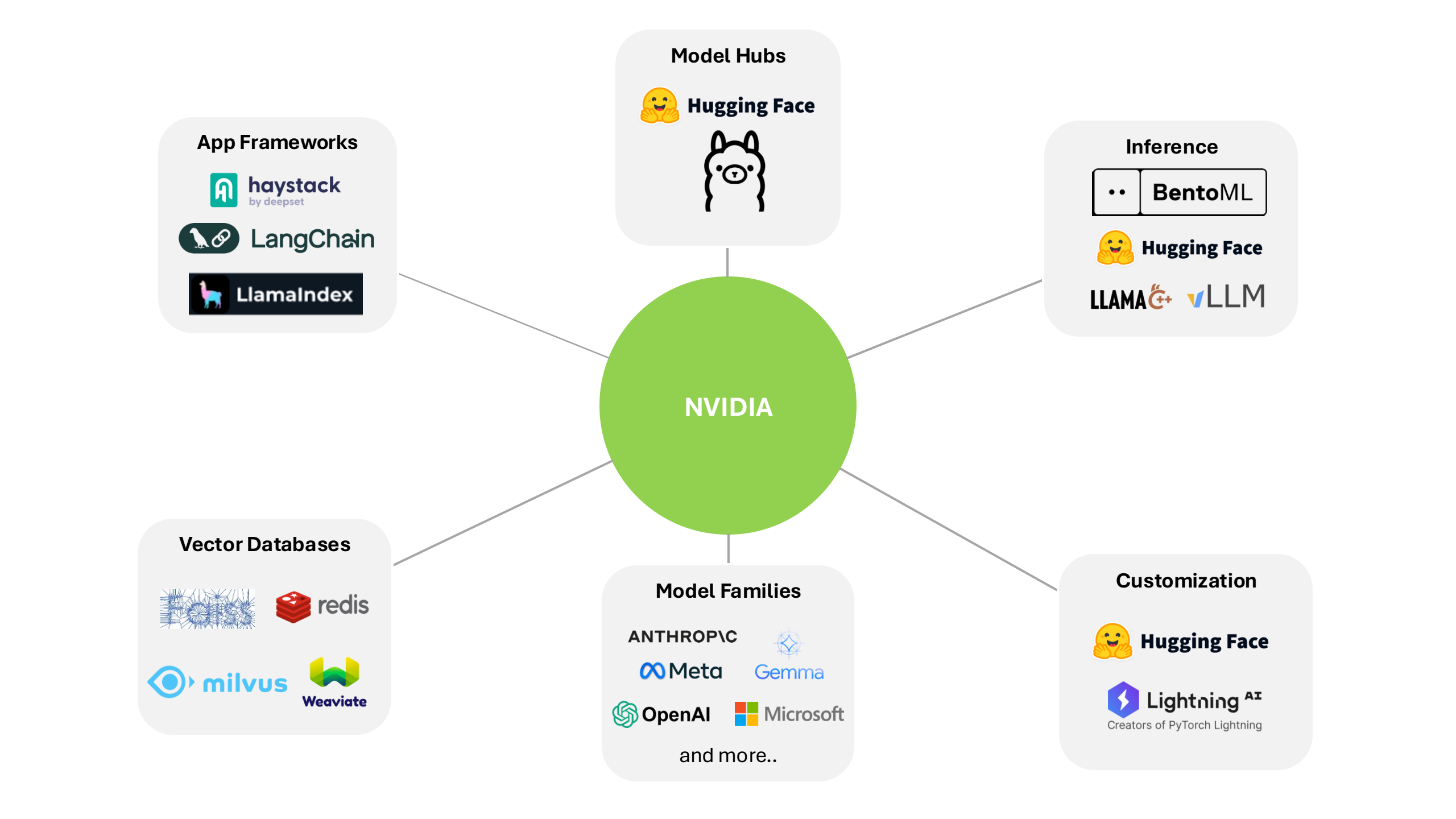
Models
Ecosystem: BLOOM, Llama, Mistral, MPT, OPT, Phi-2,
Getty Images AI Generator, Shutterstock 3D Generator
NVIDIA: BioMegatron, Edify, Nemotron

SDKs and Frameworks
Ecosystem: Colossal-AI, HuggingFace Transformers, PyTorch
NVIDIA: A2X, Megatron-LM, NeMo Framework, Riva, Picasso

Libraries
Ecosystem: XLA
NVIDIA: CUDA, CUTLASS, CV-CUDA, Megatron-Core, Megatron-LM,
NCCL, RAFT, Transformer Engine, TensorRT-LLM, Ray

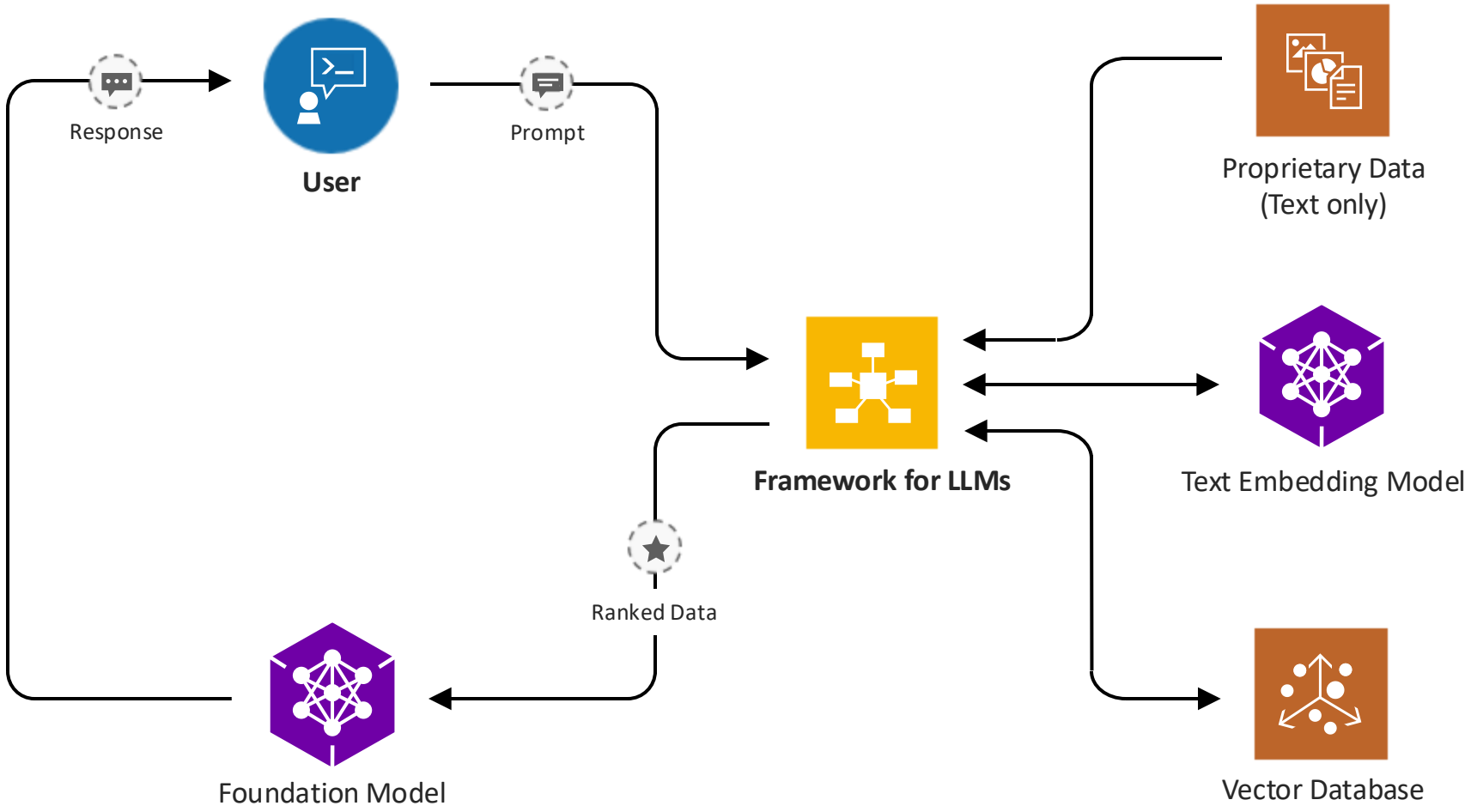
Management and Orchestration
Ecosystem: Kubernetes, Nephele, Slurm, VMware
NVIDIA: Base Command Platform

Accelerated Infrastructure
NVIDIA: DPUs, GPUs, InfiniBand

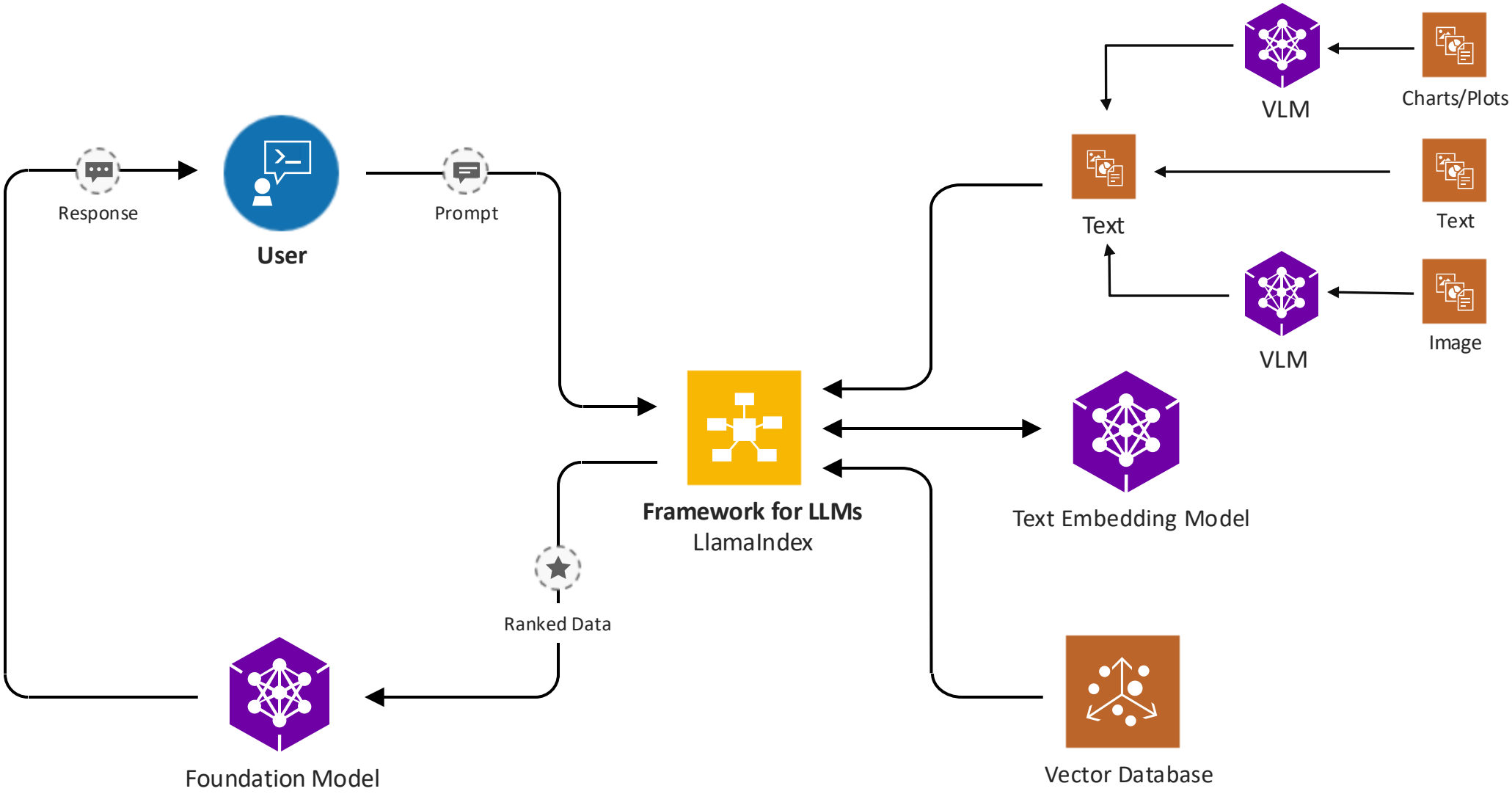

Model Hubs
Hugging Face
App Frameworks
haystack
by deepset
LangChain
LlamaIndex
Inference
BentoML
Hugging Face
LLAMA C++
vLLM
NVIDIA
Vector Databases
redis
milvus
Weaviate
Model Families
ANTHROP\C
Meta
Gemma
OpenAI
Microsoft
and more..
Customization
Hugging Face
Lightning AI
Creators of PyTorch Lightning

Retrieval Augmented Generation Workflow

Enable LLMs to provide up to date and domain specific answers

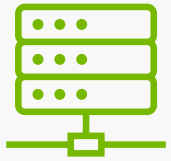


Multimodal RAG



CHALLENGES WITH MULTIMODAL RAG

Data Complexity



- Parsing diverse formats
- Handling unstructured data

Latency Concerns

- Slow processing times
- Delayed responses

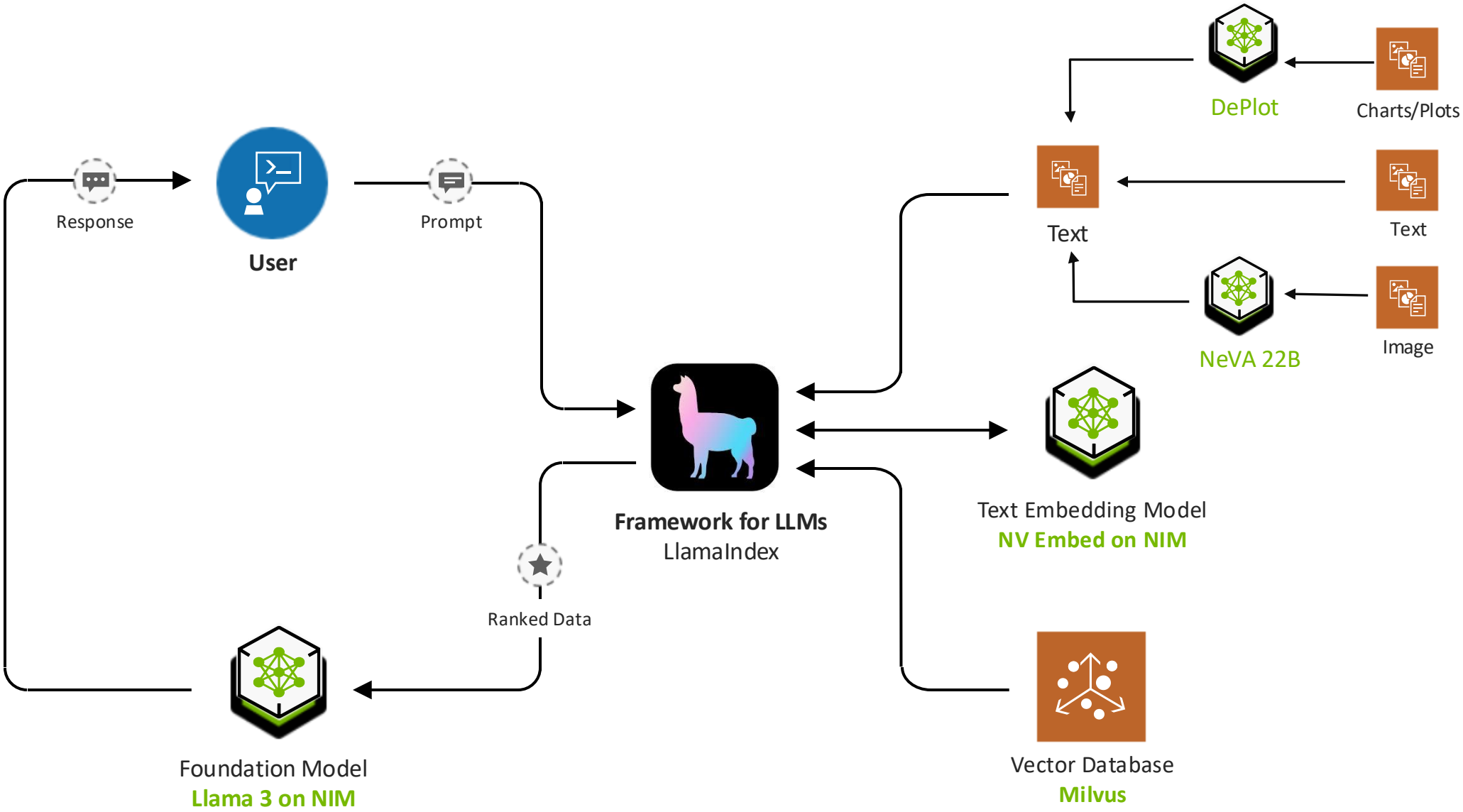
Compute Intensity

- High resource requirements
- Scaling for large-scale multimodal data processing

Integration Complexity

- Synthesizing information from multiple sources
- Complex system architecture

GPU-Accelerated Multimodal RAG



Deploy

Multimodal RAG

Choose input method:

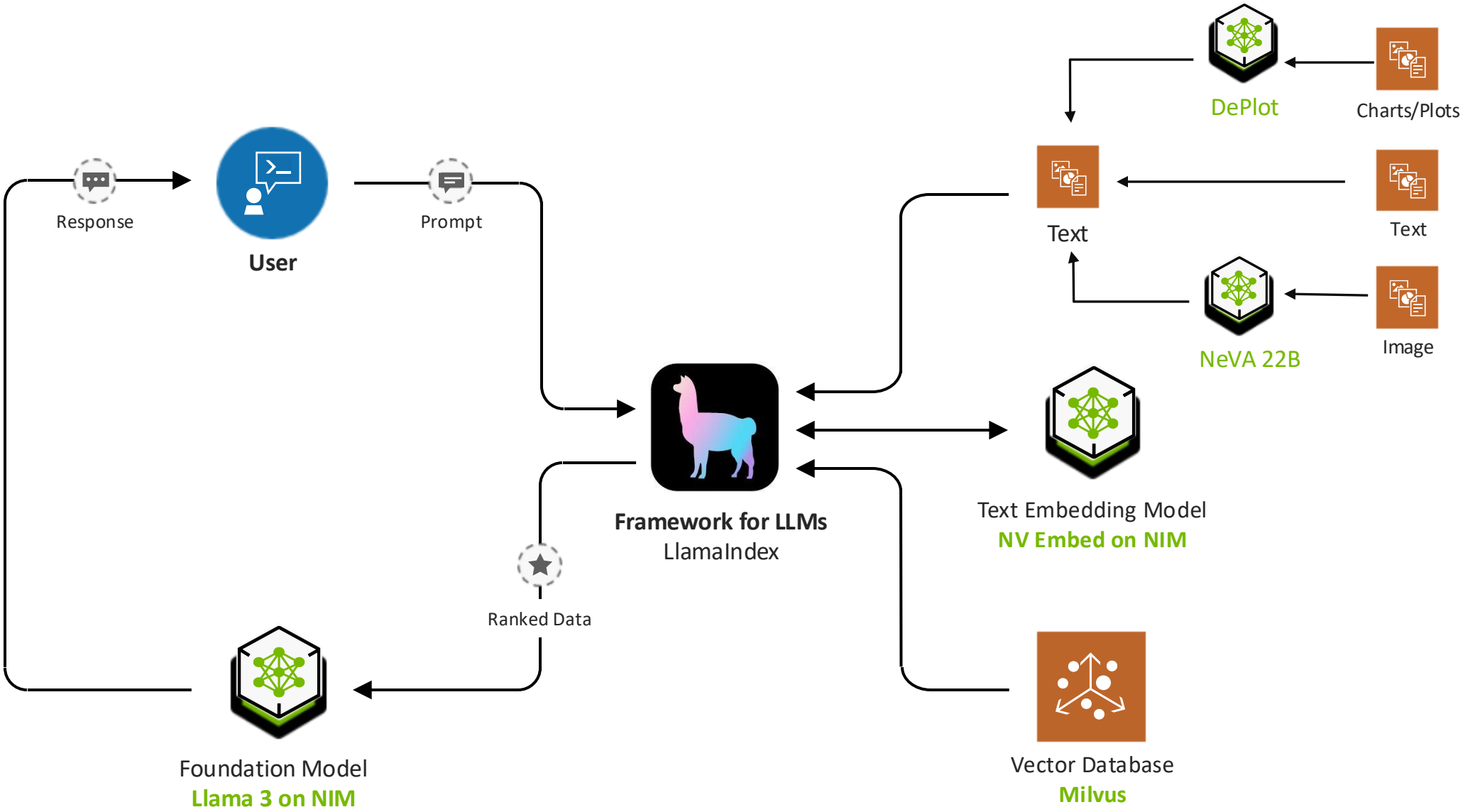
- (•) Upload Files
- ( ) Enter Directory Path

Drag and drop files here

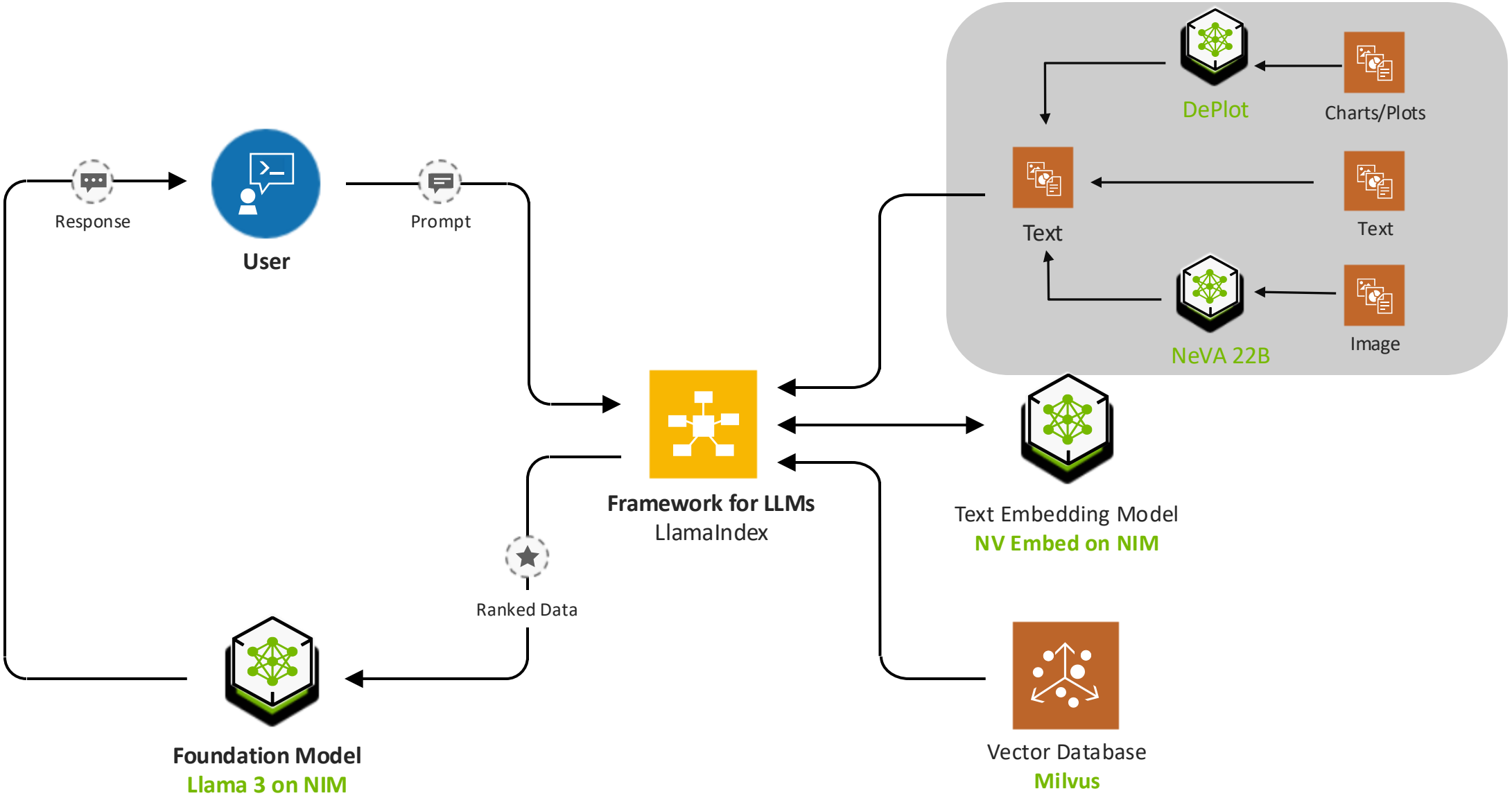
Drag and drop files here
Limit 200MB per file

Browse files

GPU-Accelerated Multimodal RAG



GPU-Accelerated Multimodal RAG



NVIDIA Inference Microservices

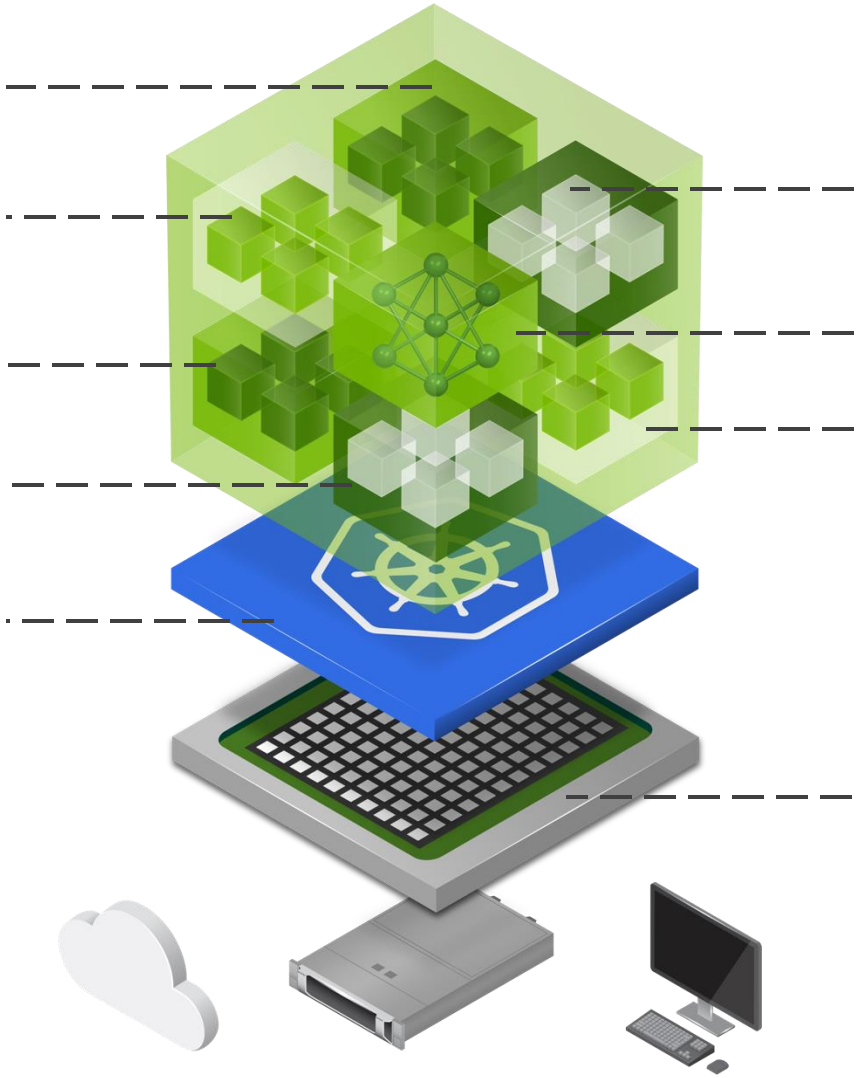
Standard APIs
Text, Speech, Image,
Video, 3D, Biology

NVIDIA Triton Inference Server
cuDF, CV-CUDA, DALI, NCCL,
Postprocessing Decoder

Cloud-Native Stack
GPU Operator, Network Operator

Enterprise Management
Health Check, Identity, Metrics,
Monitoring, Secrets Management

Kubernetes



NVIDIA TensorRT and TensorRT-LLM
cuBLAS, cuDNN, In-Flight Batching,
Memory Optimization, FP8 Quantization

Optimized Model
Single GPU, Multi-GPU, Multi-Node

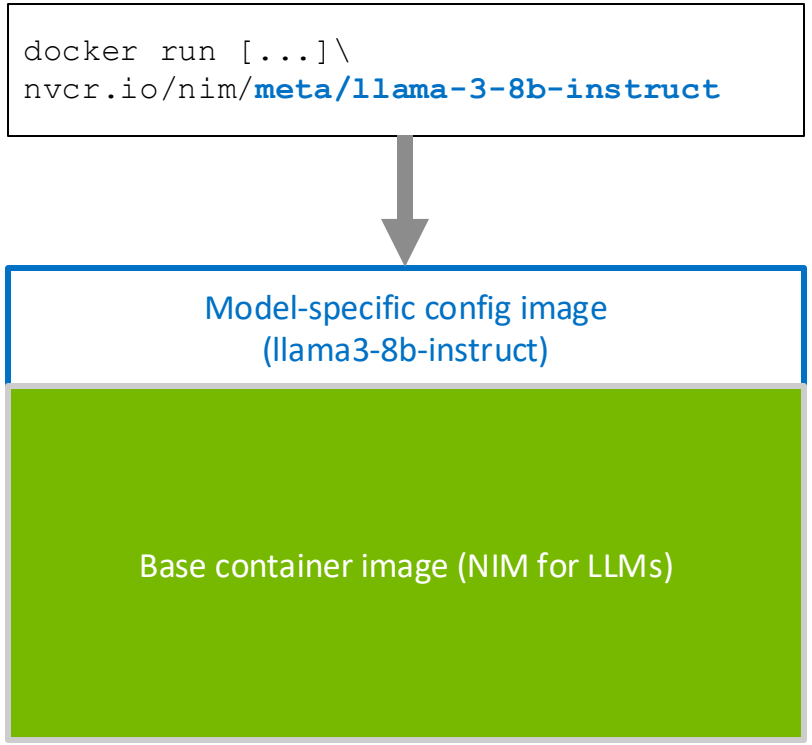
Customization Cache
P-Tuning, LoRA, Model Weights

CUDA

NVIDIA NIM for LLM

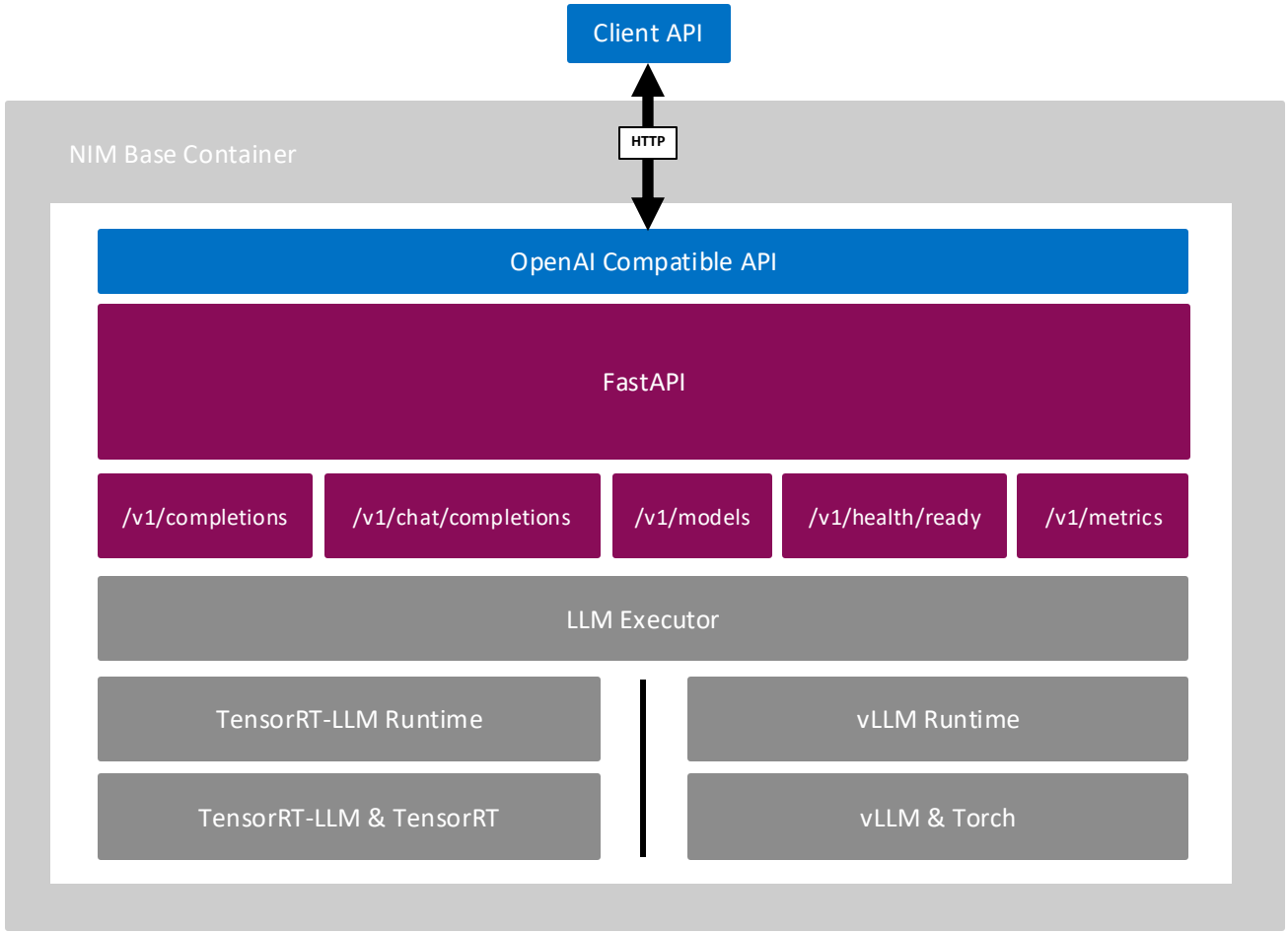
Deploy an Optimized NIM with a Single Command

- NVIDIA NIM for LLM is a container specific to each model
- When a model container is launched, it:
  - Detects the hardware it is running on
  - Mounts the cache for model and asset data
  - Selects the most optimal model given the hardware
  - Downloads the optimized model file from NGC
  - Loads the model file and starts serving

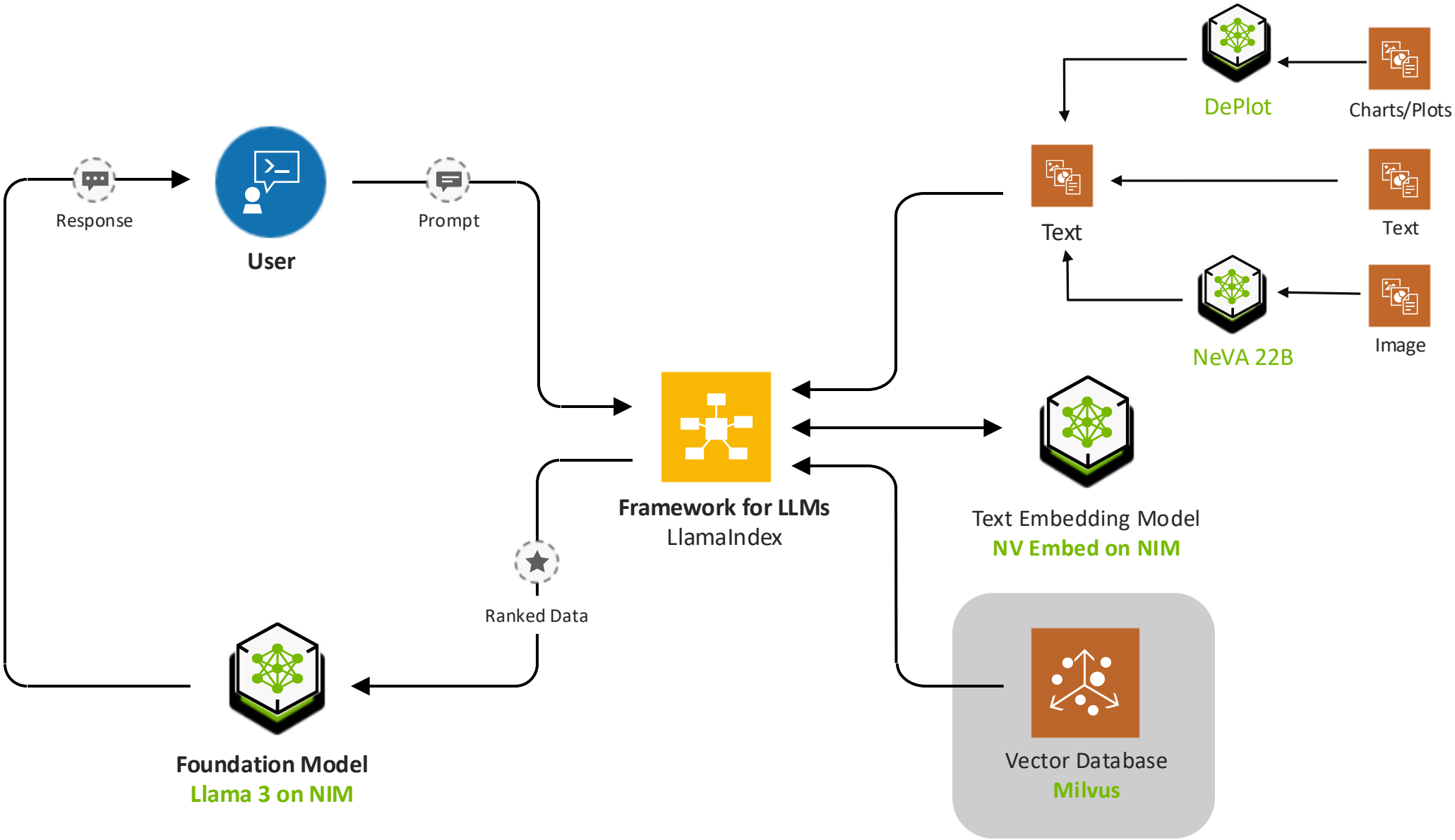


NVIDIA NIM for LLM Architecture

- HTTP REST API conforms to OpenAI specification for easy developer integration
- Liveness, health check and metrics endpoints for monitoring and enterprise management
- NVIDIA NIM includes multiple LLM runtimes
  - TensorRT-LLM and vLLM
  - Runtime is selected based on detected hardware and available optimized engines, with preference given to optimized engines

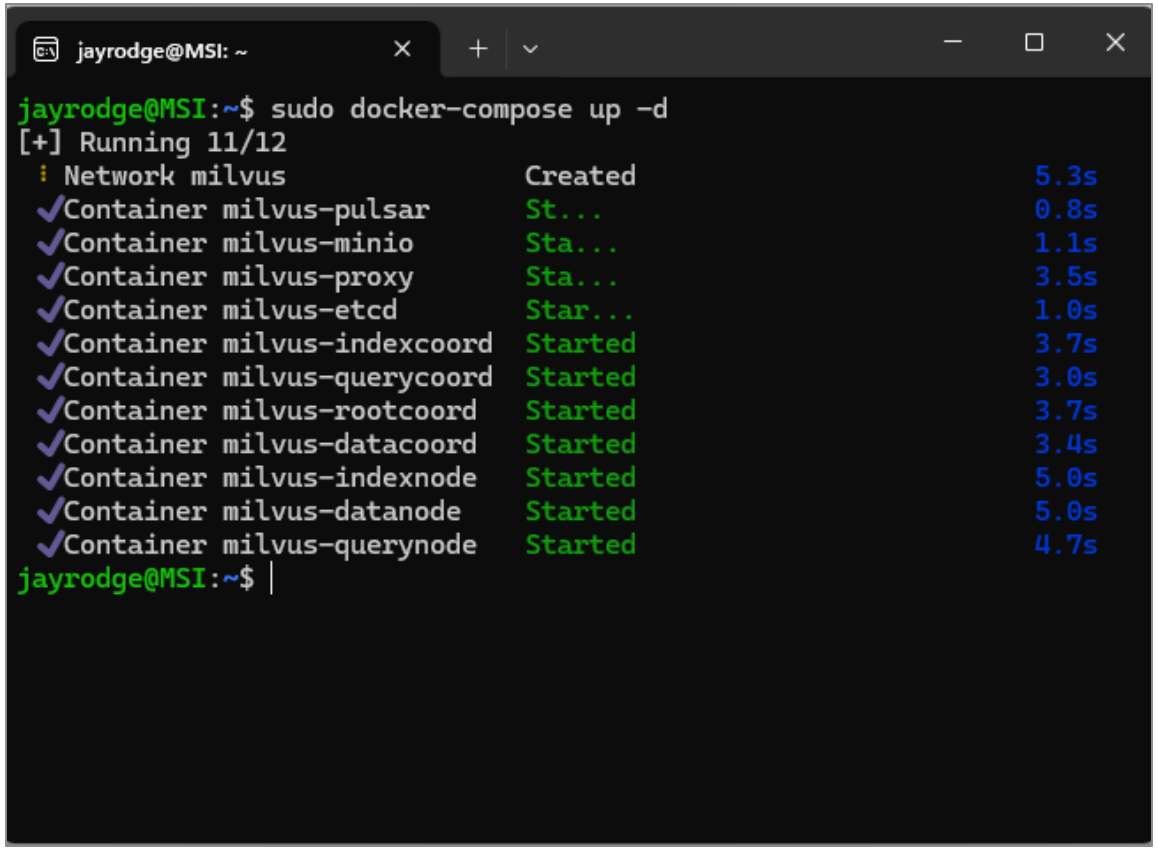


GPU-Accelerated Multimodal RAG



Milvus: Cloud-Native Vector Database

Getting Started



Starting Milvus Server

Step 1

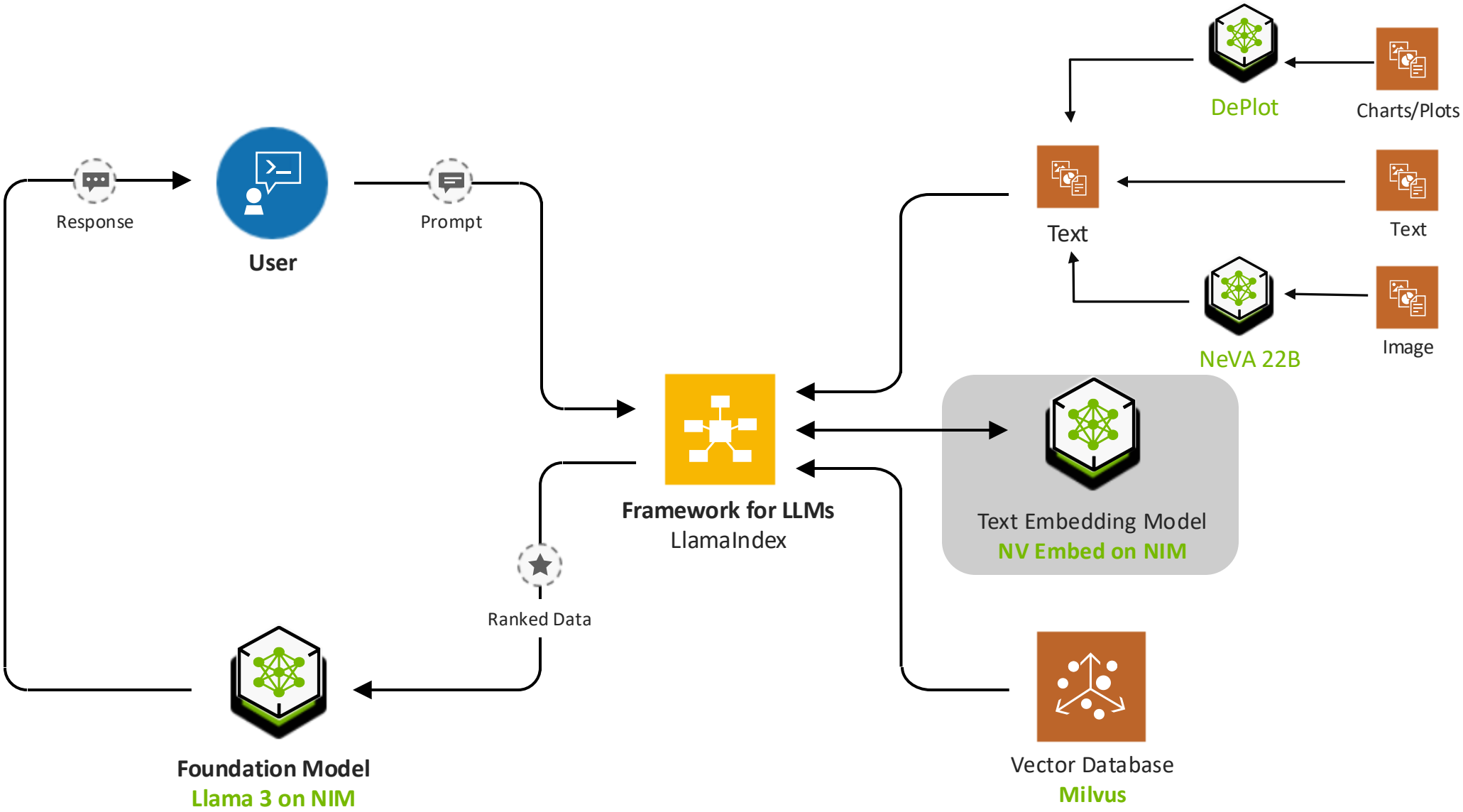
```python
from llama_index.vector_stores.milvus
import MilvusVectorStore


vector_store = MilvusVectorStore(
    host = "127.0.0.1",
    port = 19530,
    dim = 384
)
```

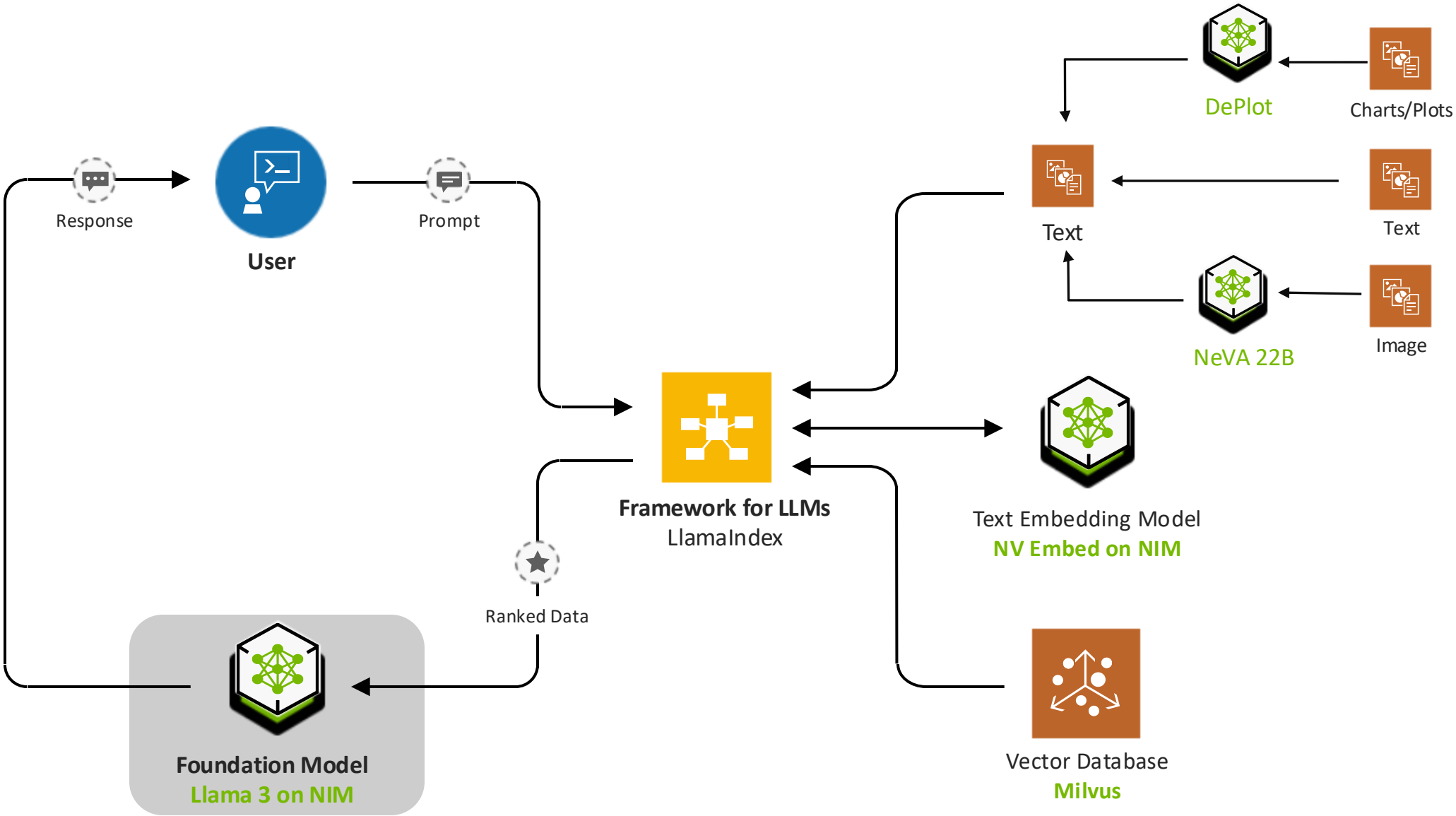
Querying through client

Step 2

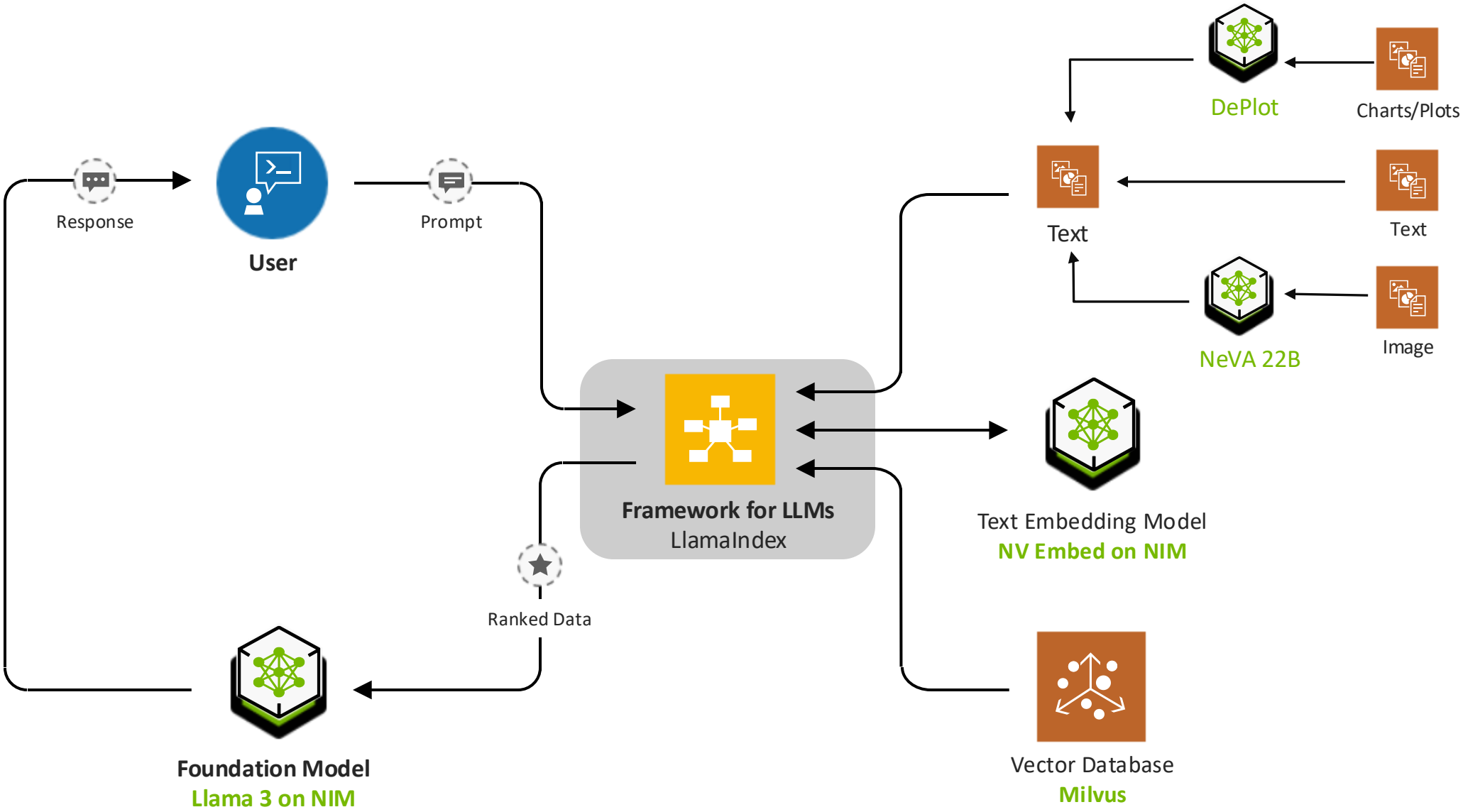
GPU-Accelerated Multimodal RAG



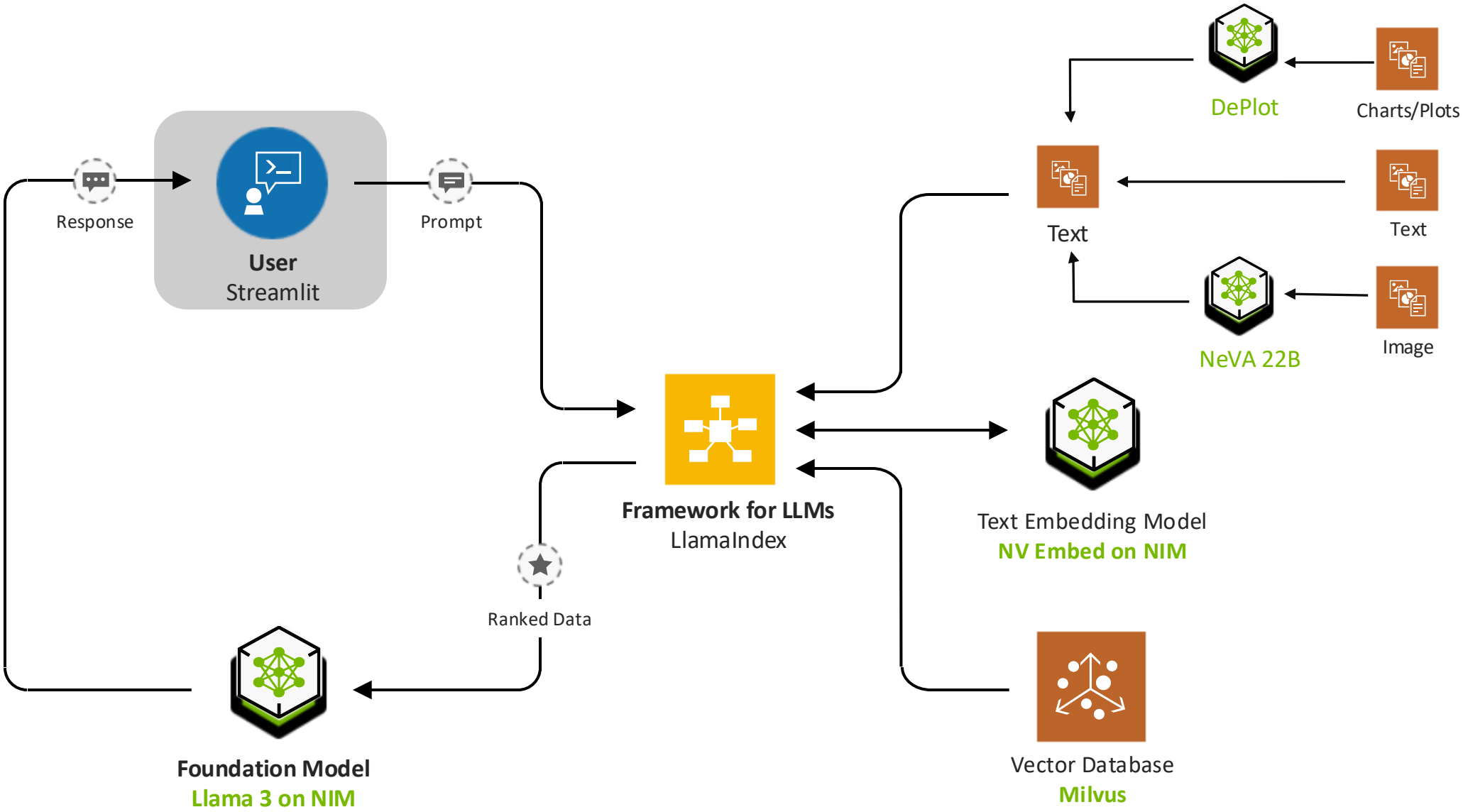
GPU-Accelerated Multimodal RAG



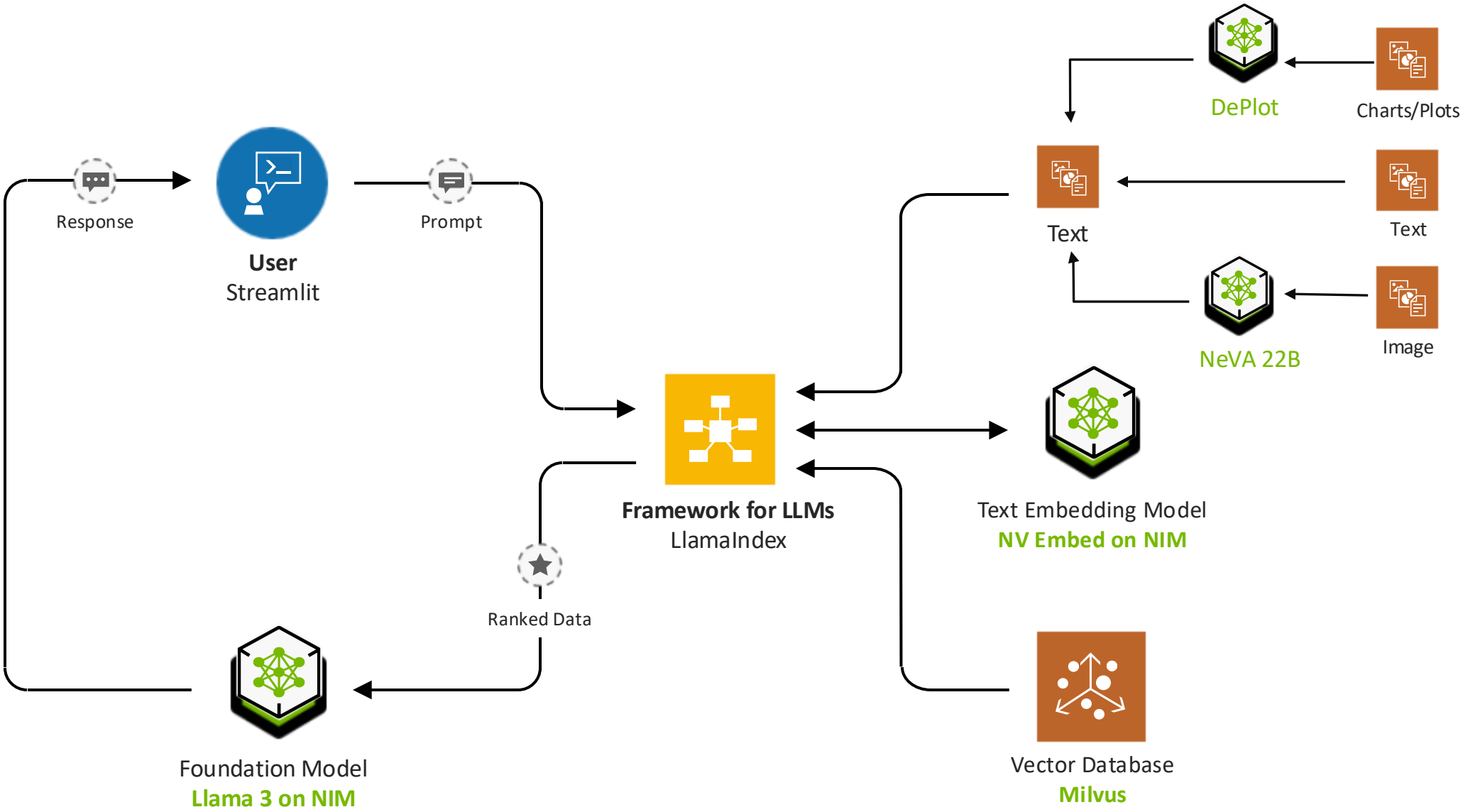
GPU-Accelerated Multimodal RAG



GPU-Accelerated Multimodal RAG



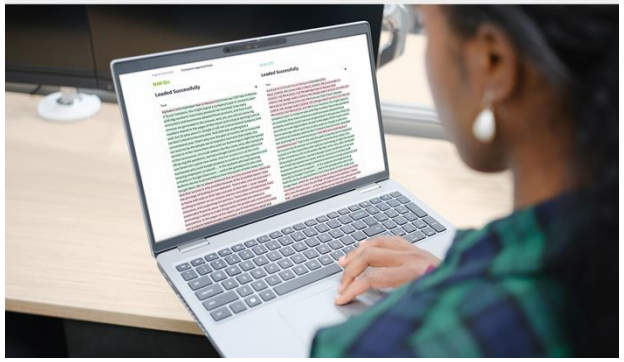
GPU-Accelerated Multimodal RAG



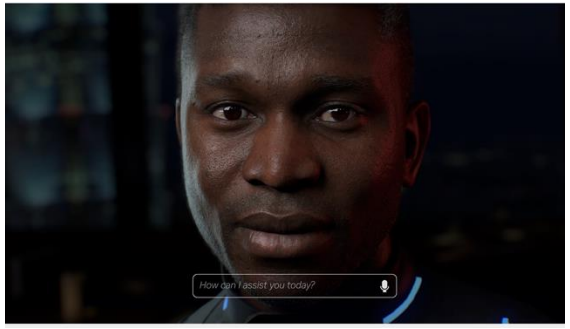
NVIDIA NIM Agent Blueprints

Available on build.nvidia.com

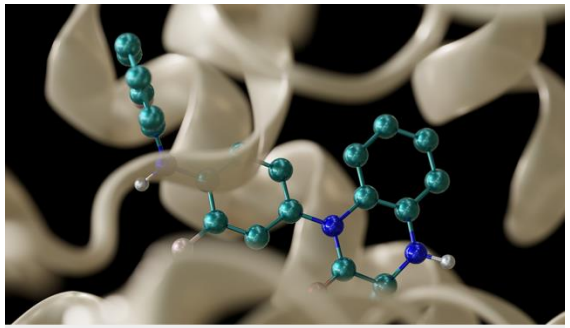
Multimodal PDF Data Extraction for Enterprise RAG



Digital Humans for Customer Service



Generative Virtual Screening for Drug Discovery



• • •

monthly release

NVIDIA NIM Agent Blueprint

Example Application

Interactive experience that can be easily replicated

Sample Data

Public data for workflow testing

Reference Code

Leverage proven pre-trained models

Architecture

Reference architecture including API definitions, NIM, and more

Customization Tools

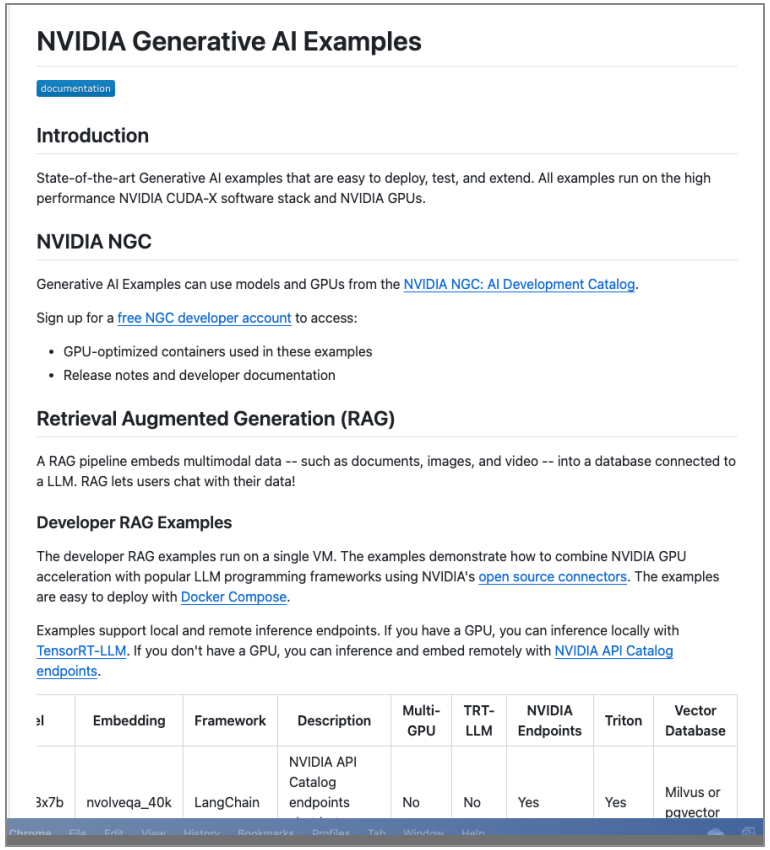
Customize and evaluate models

Orchestration Tools

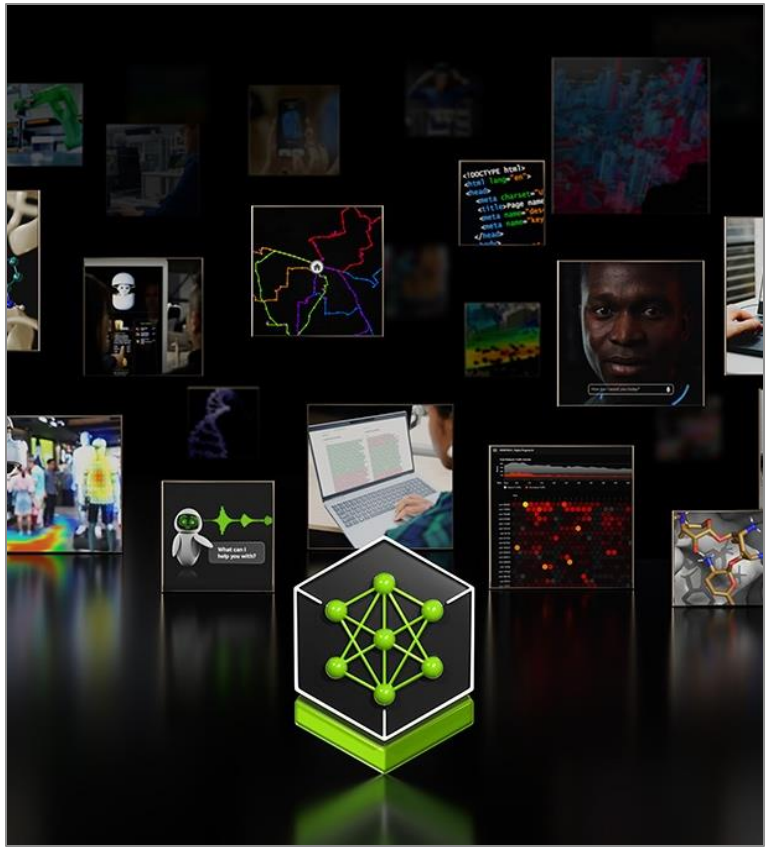
Deploy and manage workflow microservices

GET STARTED

Multimodal RAG project



GenAI Examples



NIM Agent Blueprints

NVIDIA and LlamaIndex Developer Contest

Enter for a chance to win cash prizes, an NVIDIA GeForce RTX™ 4080 SUPER GPU, and more

Register Now

Thank You

NVIDIA Developer Program

Your guide to the latest AI advancements powered by NVIDIA GPUs.

Get updates directly in your inbox by subscribing to newsletter at: https://developer.nvidia.com/developer-program


NVIDIA