Customizing Generative AI with Your Own AI Foundry [S61968]

Sophie Watson, Technical Marketing Engineer

Adriana Flores, Director NVIDIA AI Solution Architecture

Agenda

- Generative AI
- NVIDIA AI Foundry
- Domain Customization Journey
- AI Foundry Compute: DGX Cloud

Generative AI

Generative AI Journey

Driving Generative AI from Idea to Inference

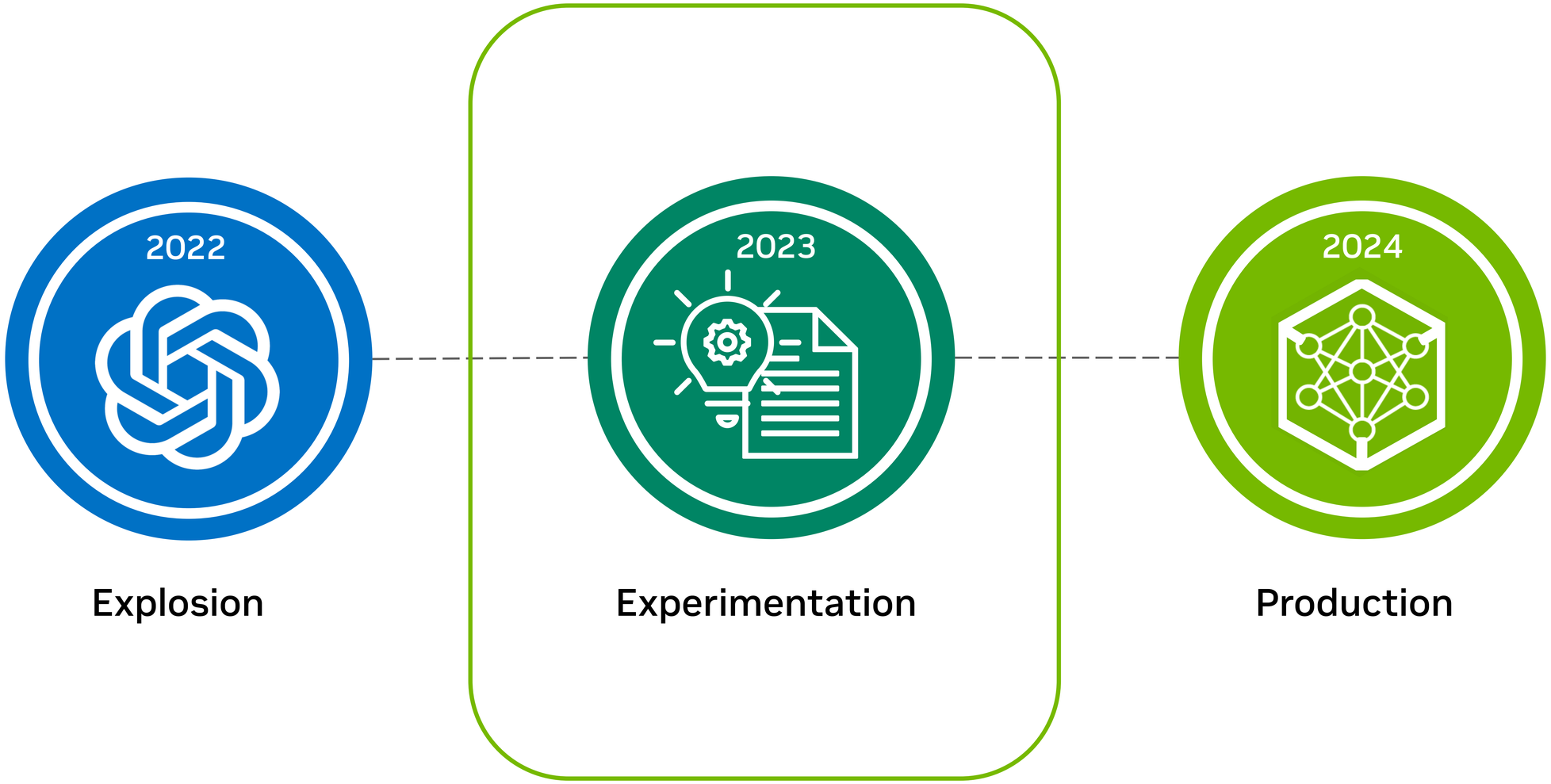
Explosion

Experimentation

Production

Generative AI Journey

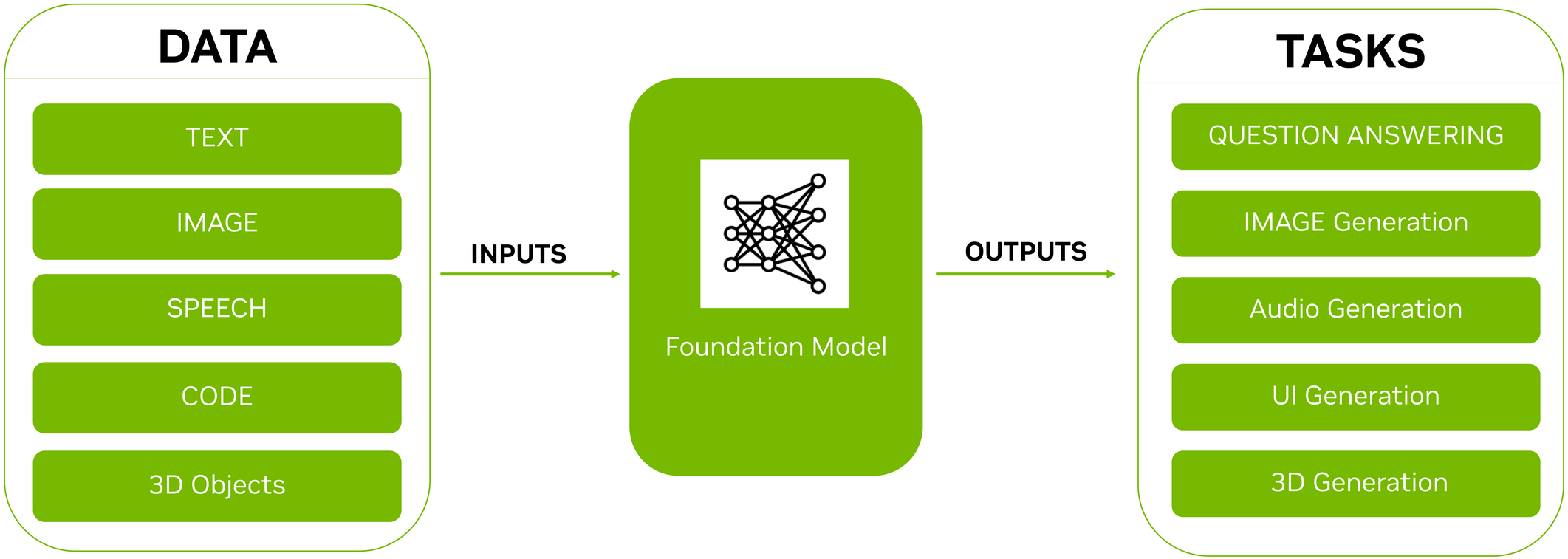
Driving Generative AI from Idea to Inference



Generative AI Journey

Driving Generative AI from Idea to Inference

How Does Generative AI Work?



Target Use Cases for Generative AI

AI assistants are driving the explosion of POCs

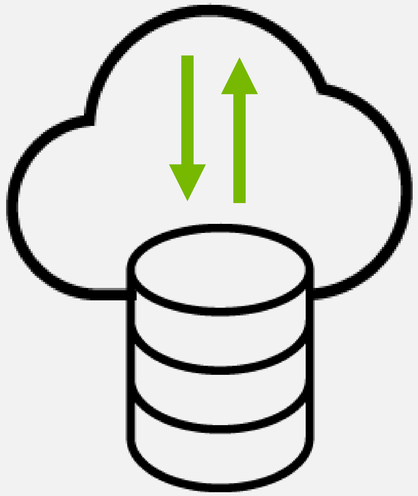
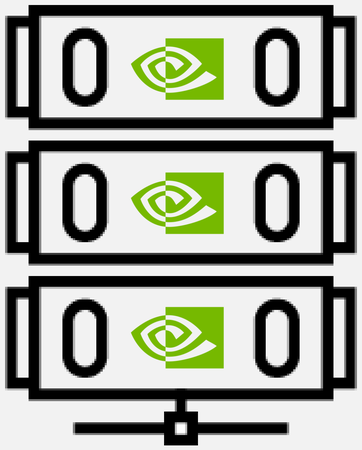
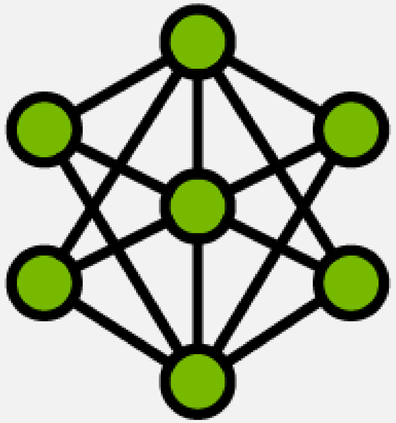
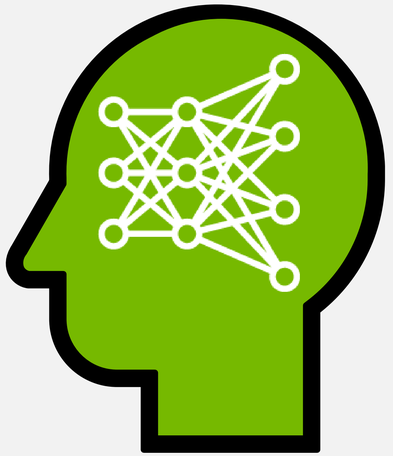
Intelligent Chatbot

Knowledge Base Copilot

Code Generation

Customizing a Generative AI model

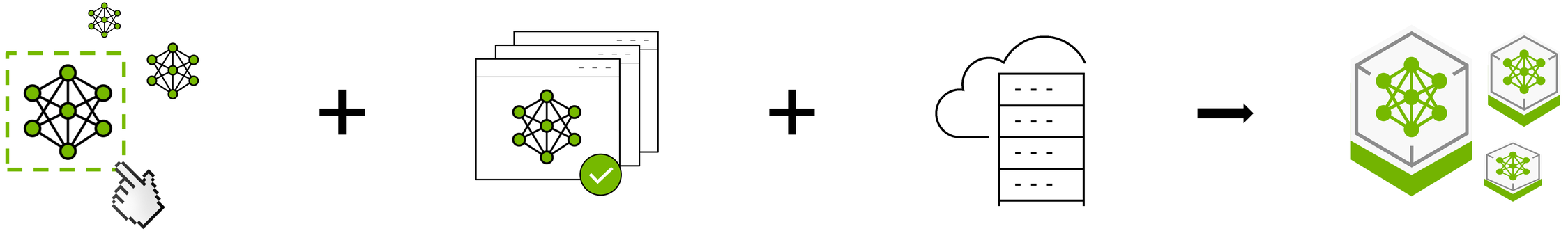
What do you need?

Training Data	Accelerated Computing	Training and Inference Tools	AI Expertise
			

Building an Enterprise AI Foundry

What is an AI Foundry?

An AI foundry *is a service that consists of:*



Leading state-of-the-art, pre-trained models

Tools to easily customize models using proprietary data

Cloud-native, accelerated infrastructure

Your customized enterprise model

Why do AI Foundry?

Building an Enterprise AI Foundry

Protect Company IP	Align to Your Brand	Secure Data Access	Maintain Control Anywhere	Production Ready
				
Keep intellectual property and confidential information secure and reduce risk of exposure	Tailored to AI applications to your company's unique standards and ethics	Protect against data breaches and accidental leakage of private information	Self hosted AI development and deployment environments	Enterprise grade platform include CVE patching, feature branches, and support

NVIDIA AI Enterprise

NVIDIA AI Foundry

NVIDIA AI Foundry

Foundry of Domain Specific Generative AI models

Proprietary Data

Compute
OnPrem & Cloud
NVIDIA DGX

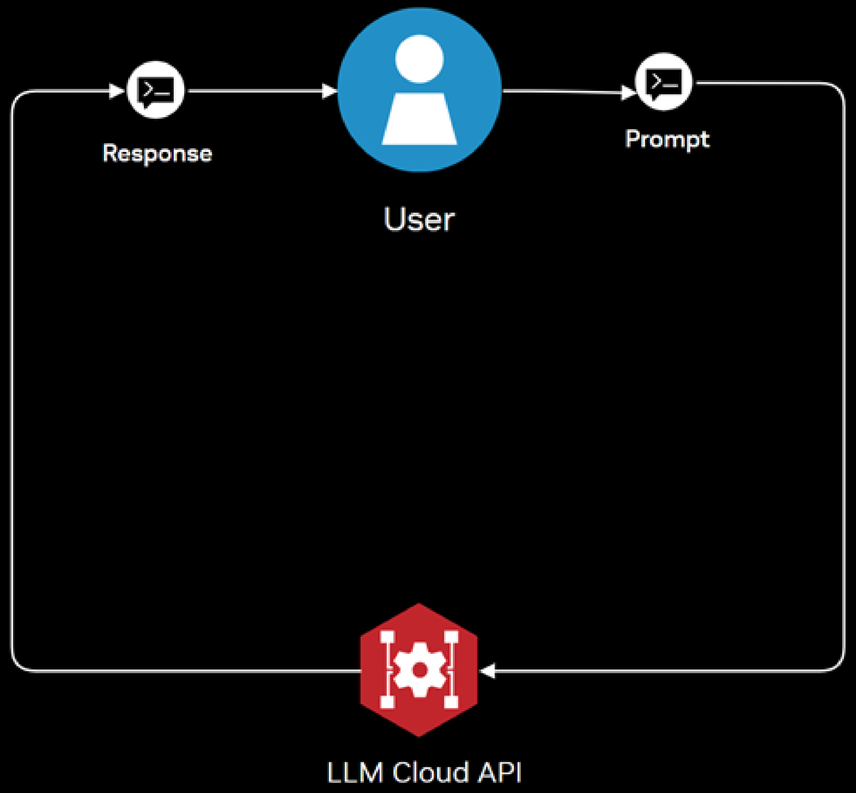
Process & Tools
ai.nvidia.com
NVIDIA NIM
NVIDIA NeMo

Expertise
Methodology to Build Domain Adapted Models

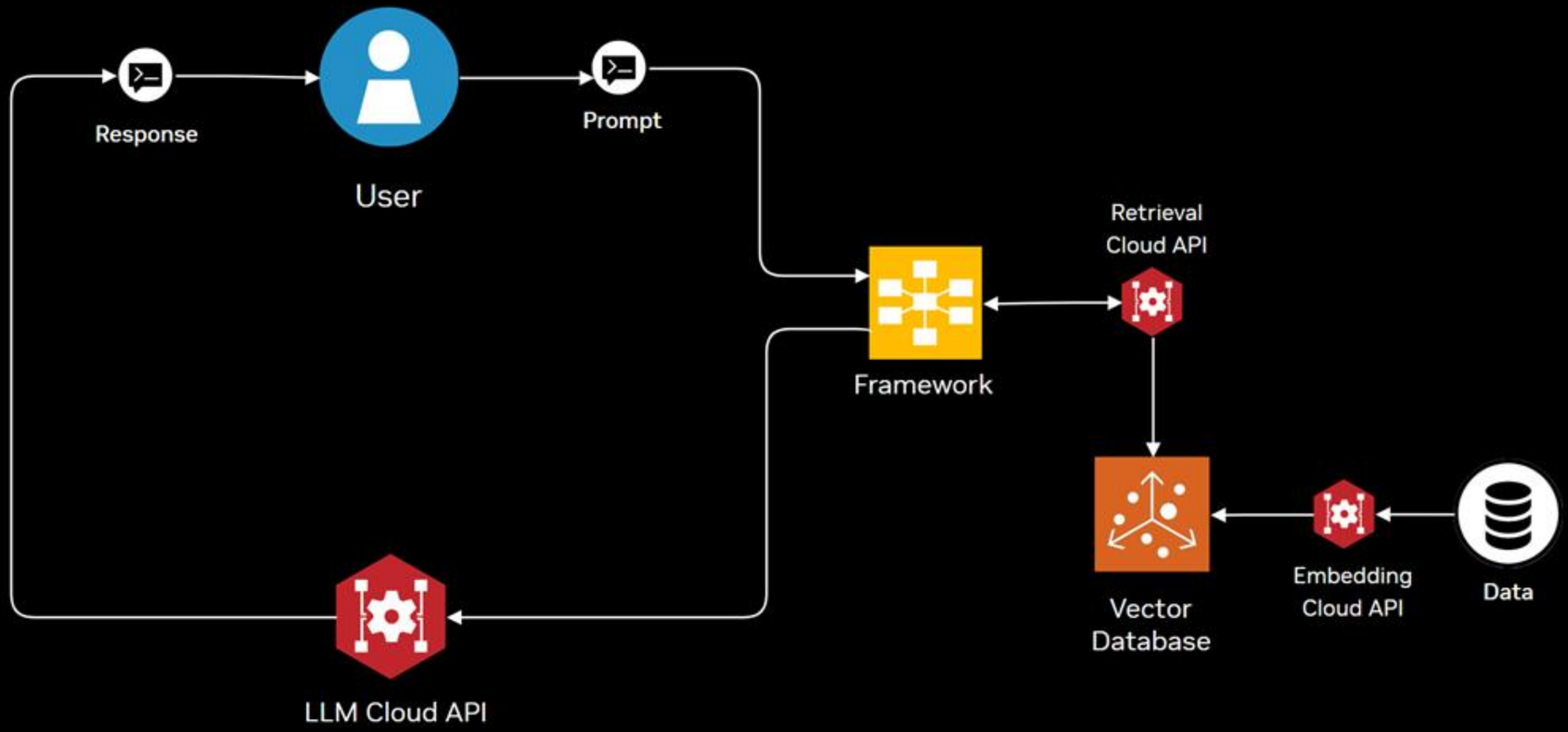
Domain Specific Use Case

Domain Customization Journey

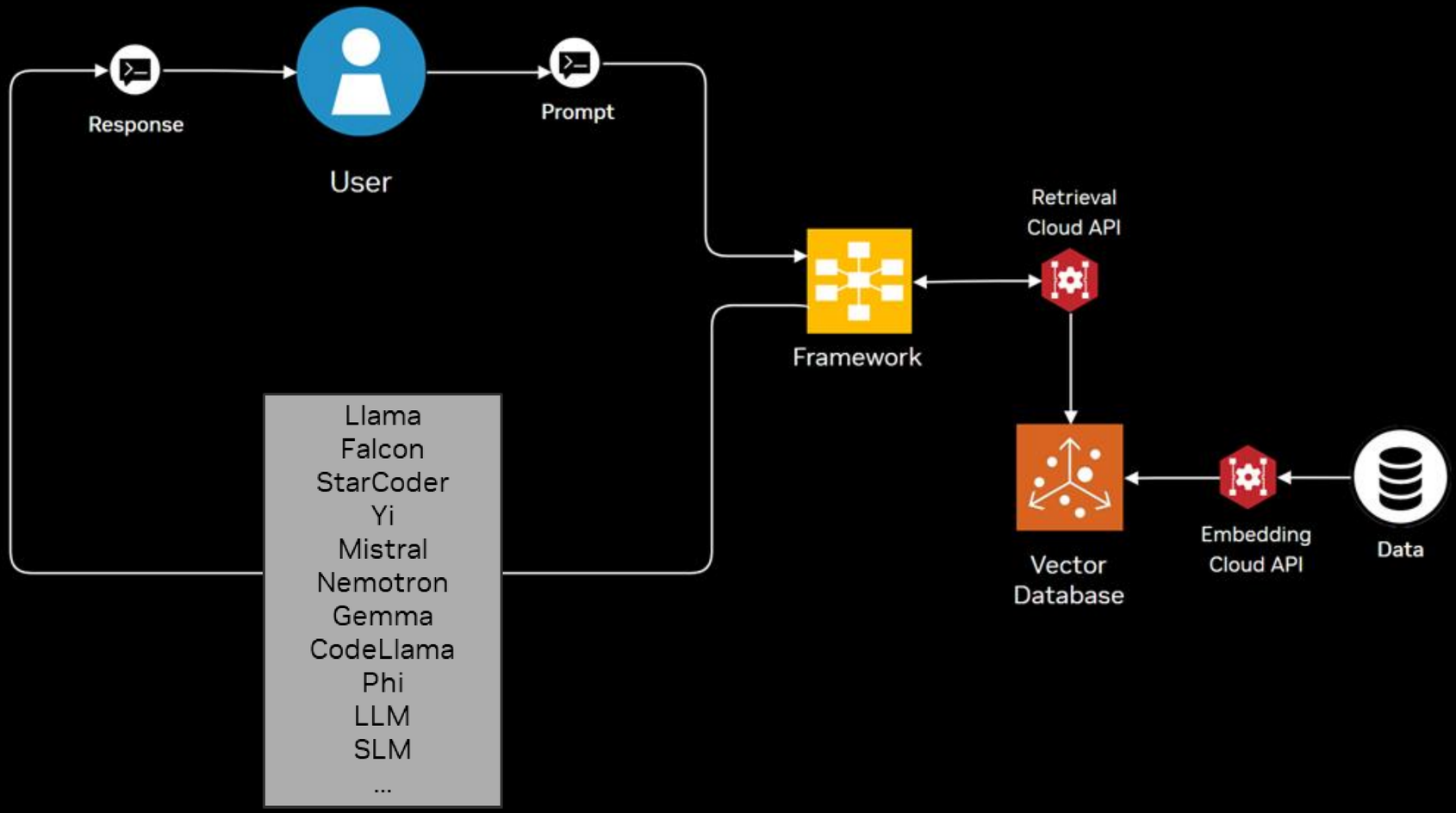
2023 was All About Exploration of GenAI LLMs



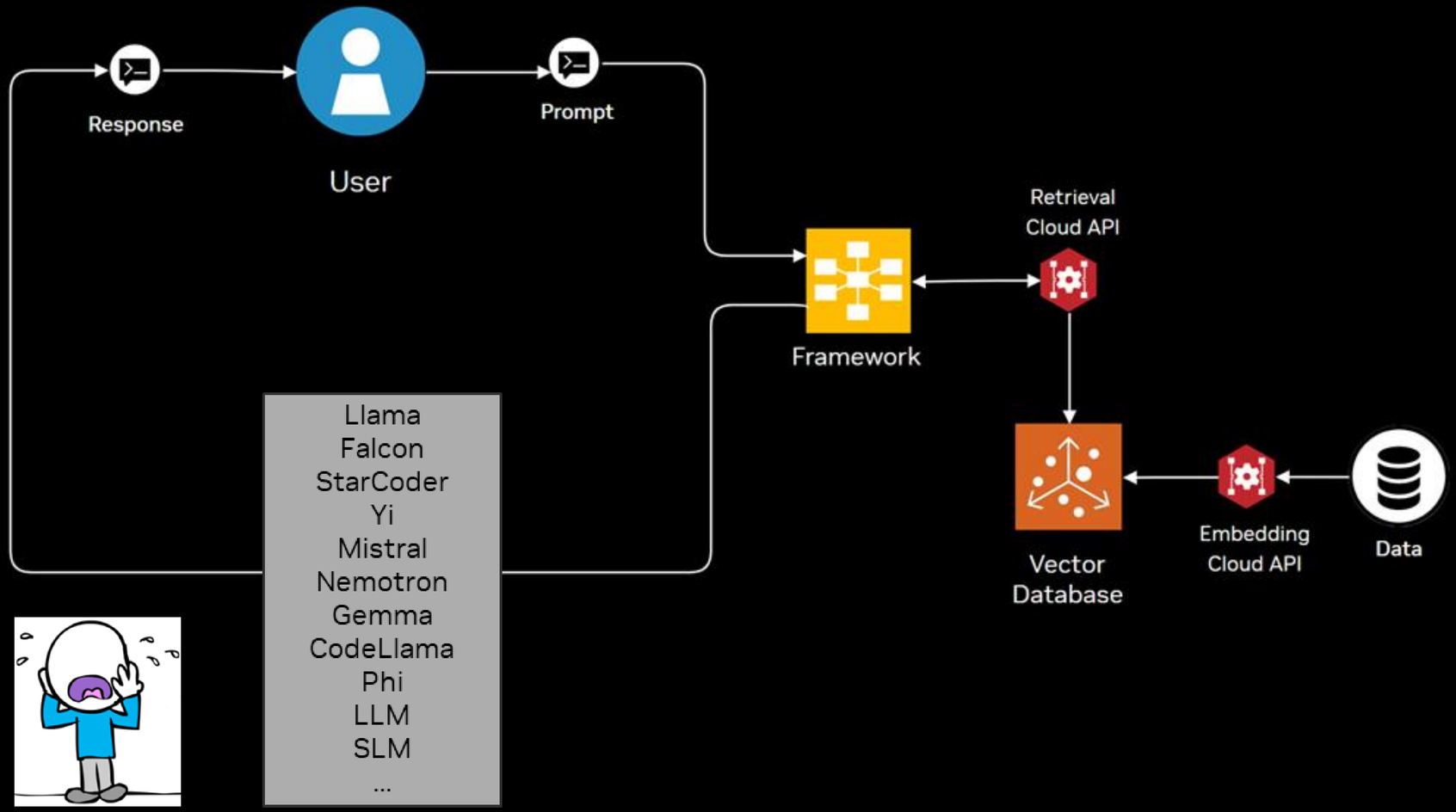
Then we started adding RAGs



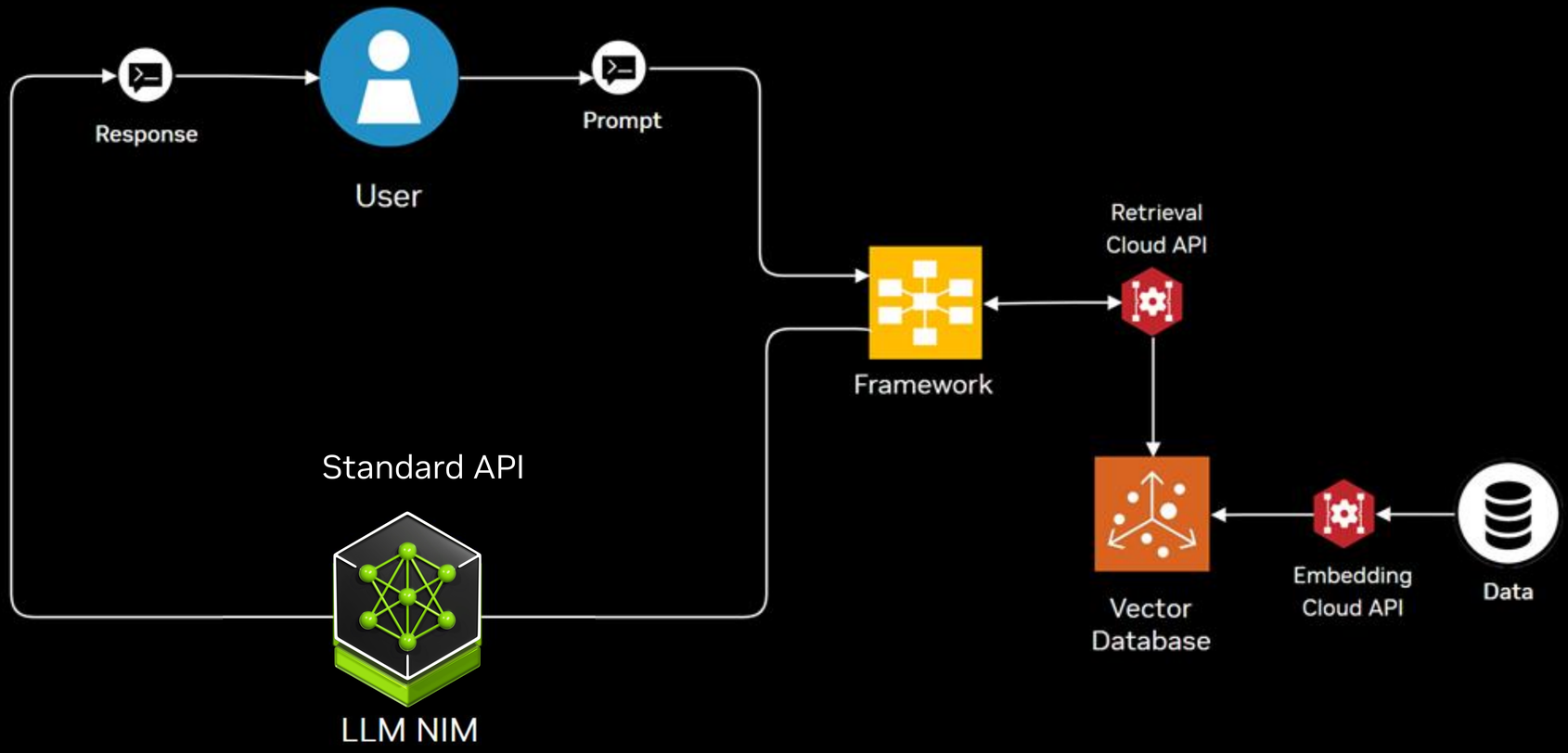
Enter Community Models



How do I keep up with the pace of innovation?



2024 is the Year of LLMs in Production



NVIDIA INFERENCE MICROSERVICE

Pre-Trained AI Models
Packaged and Optimized to Run Across
CUDA Installed Base

NVIDIA Inference Microservices (NIMs) for Generative AI

Accelerated runtime for generative AI, go zero to AI in 5 minutes

Prebuilt container and helm chart tested and validated across infrastructure

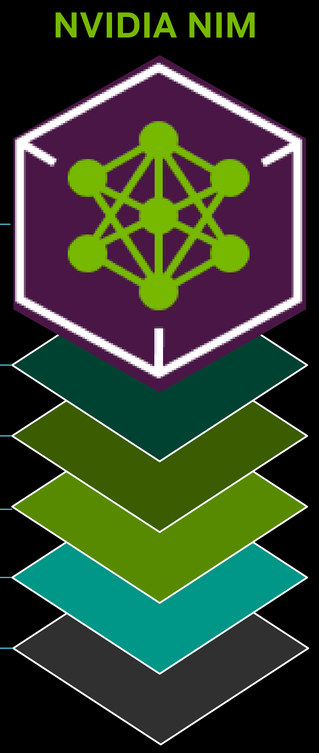
Industry standard APIs with NVIDIA Cloud standards

Domain specific code for each different NIM domain category including LLMs, VLMs, video, healthcare, and more

Optimized inference engines for each model and hardware SKU

Support for custom models build by users targeted use cases

NVIDIA AI Enterprise approved base container



Deploy anywhere and maintain control of generative AI applications and data

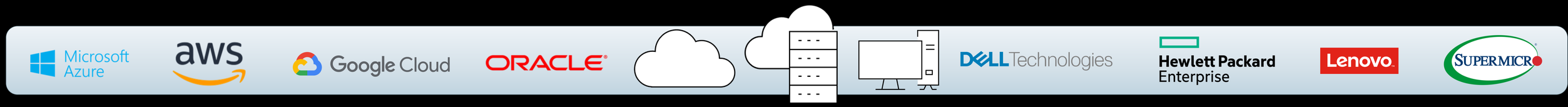
Simplified development of AI application that can run in enterprise environments

Day 0 support for all generative AI models providing choice across the ecosystem

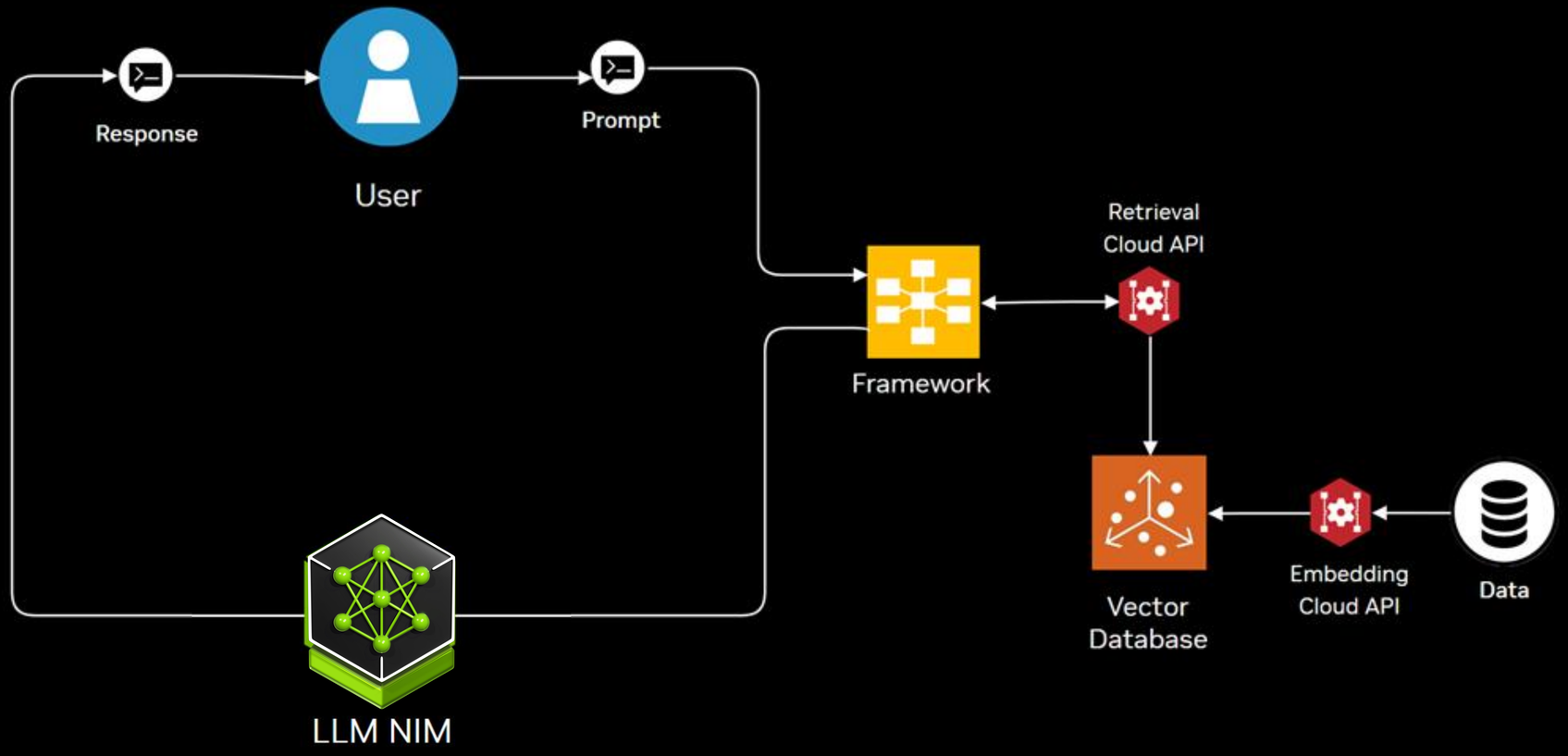
Improved TCO with best latency and throughput running on accelerated infrastructure

Best accuracy for enterprise by enabling tuning with proprietary data sources

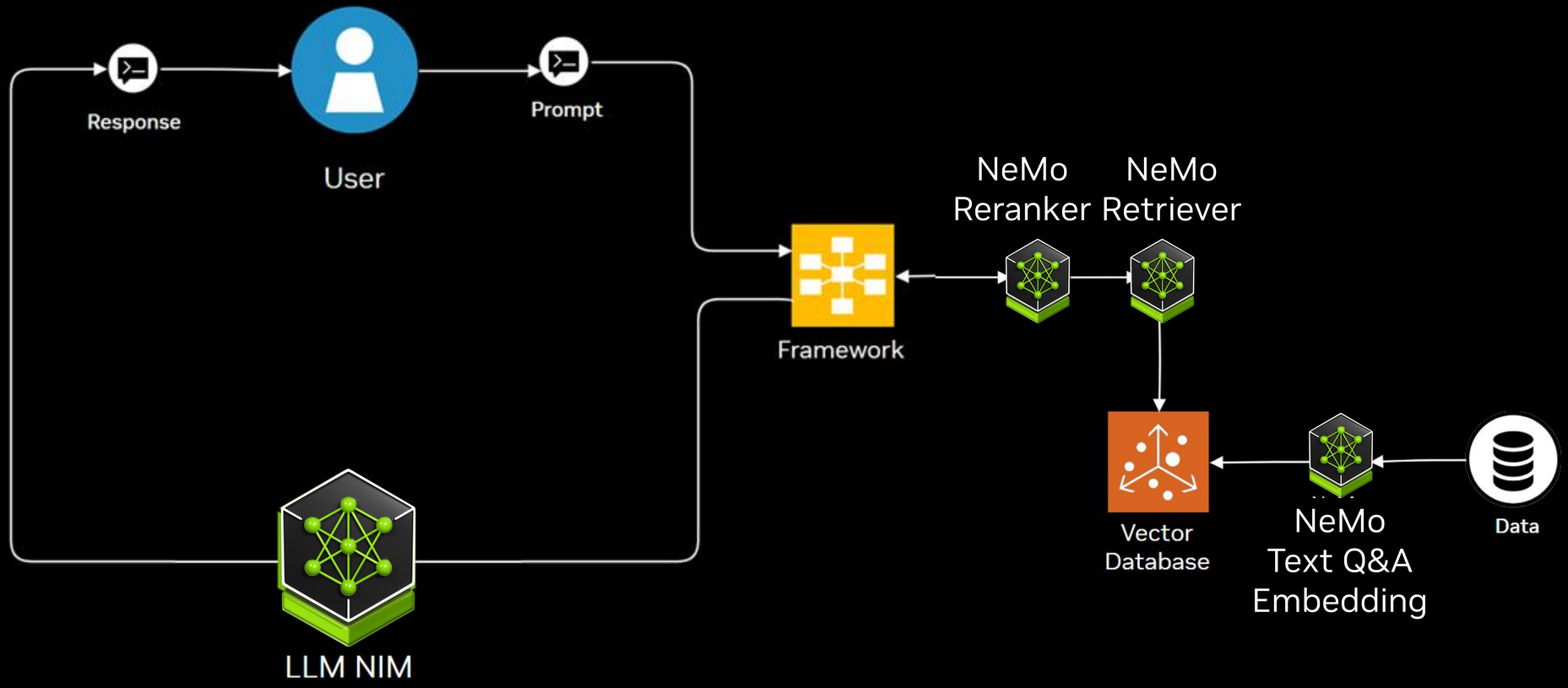
Enterprise software with feature branches, validation and support



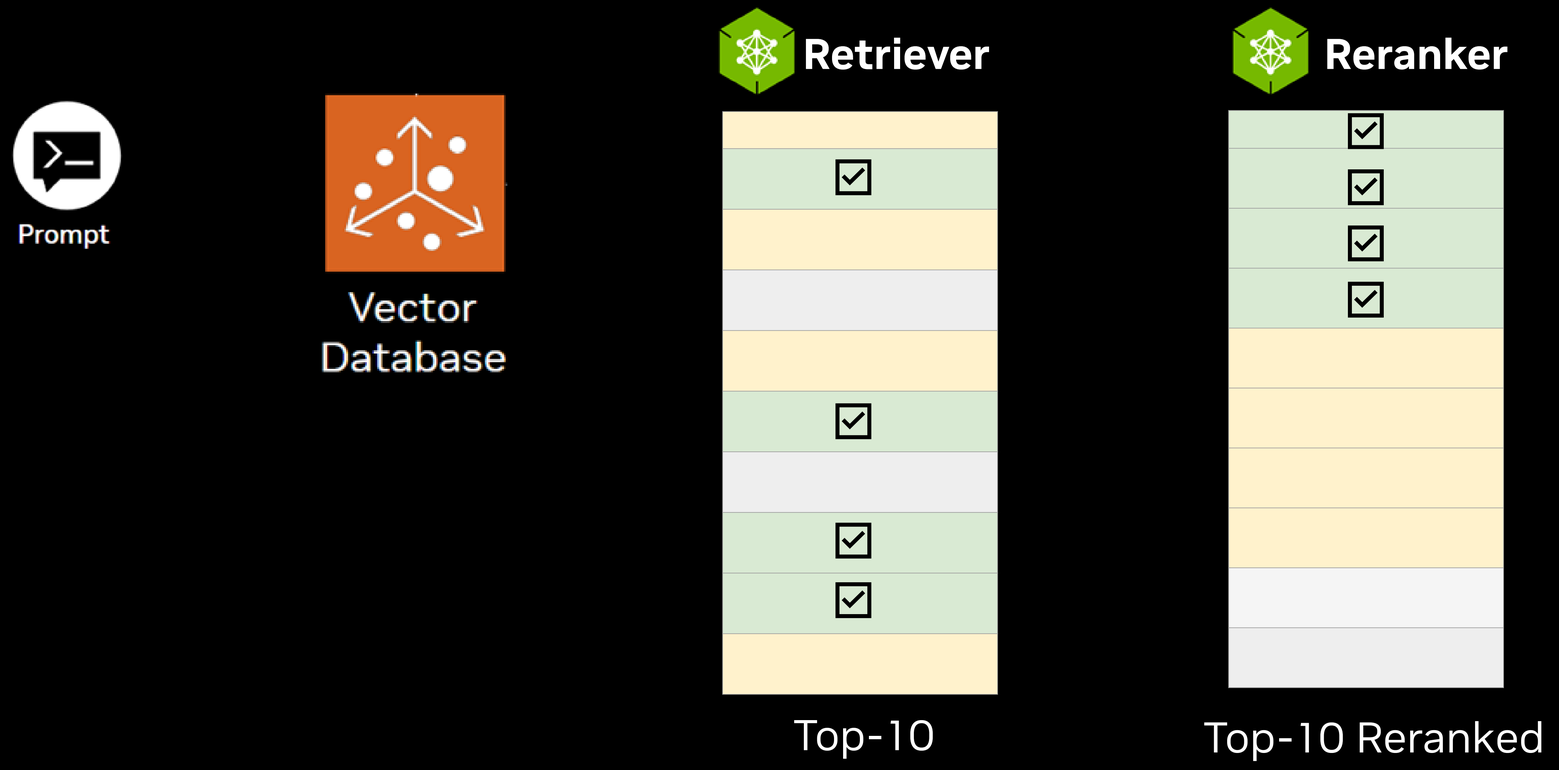
RAG is great... Until it's not



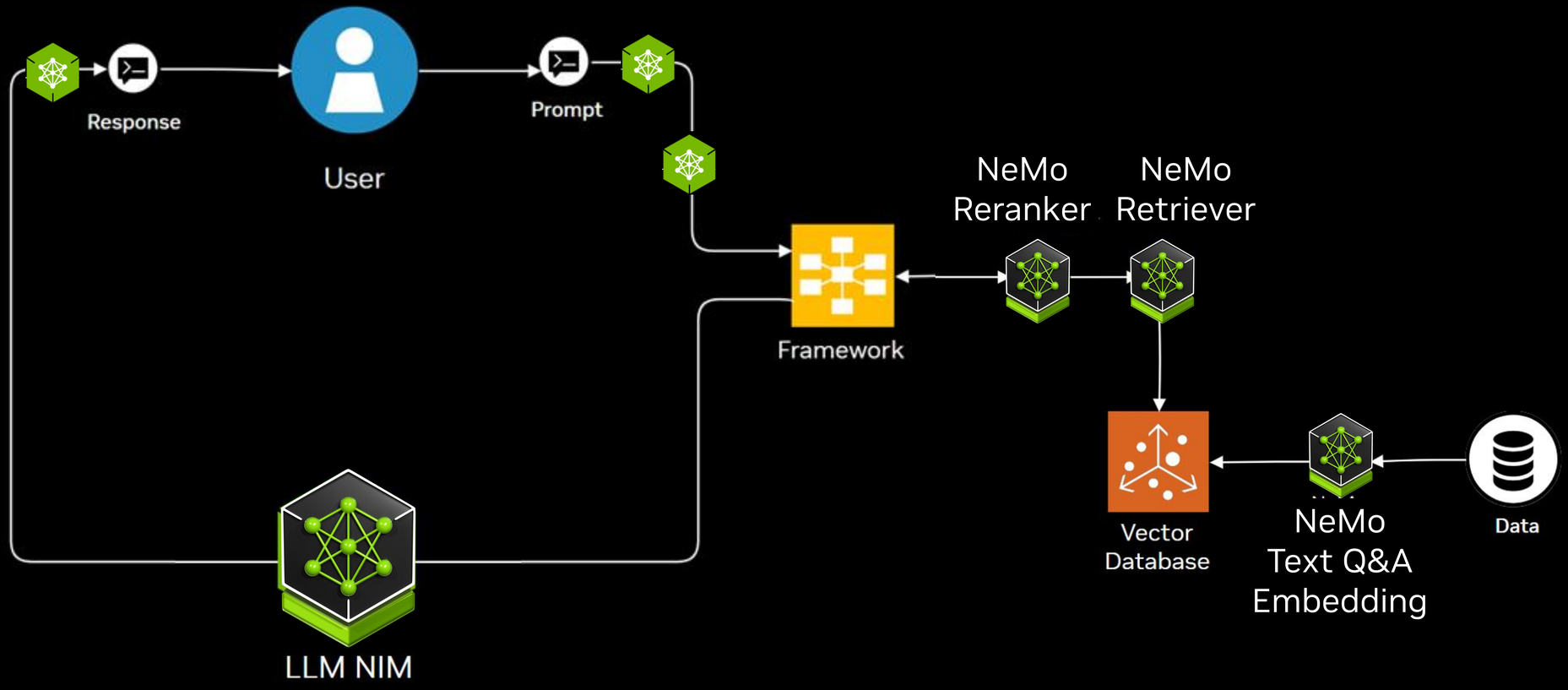
RAG can be improved



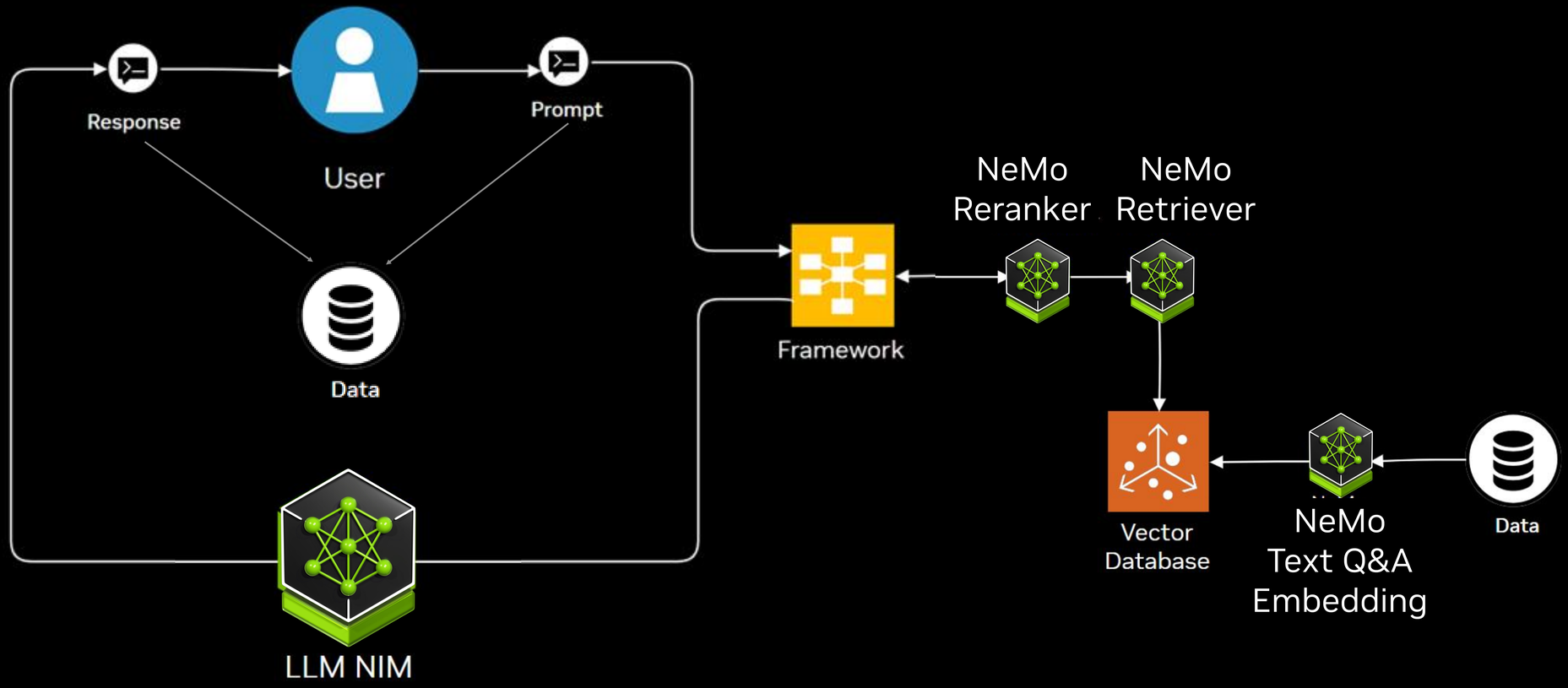
NeMo Retriever and Reranker Microservices



Complex Pipelines



RAG acts a data flywheel



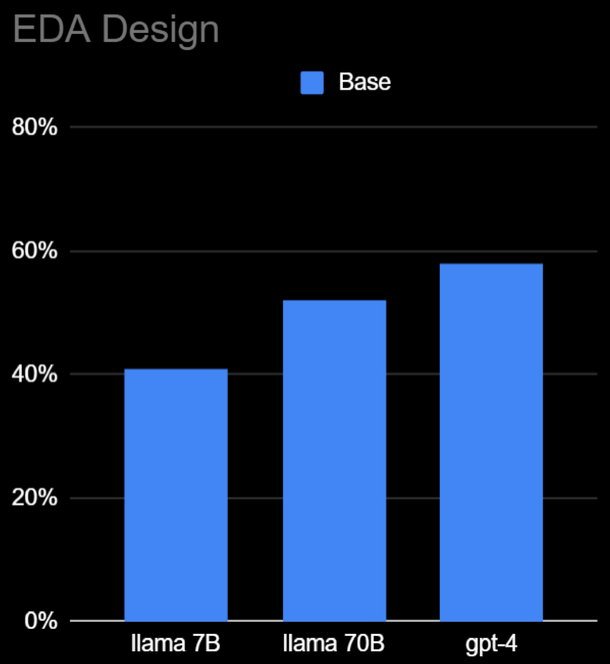
LLM Customizing with NeMo Microservices

NeMo Evaluator

NeMo Curator

NeMo Customizer

LLM Customizing with NeMo Microservices



Write VIVID code to get partitions in a design

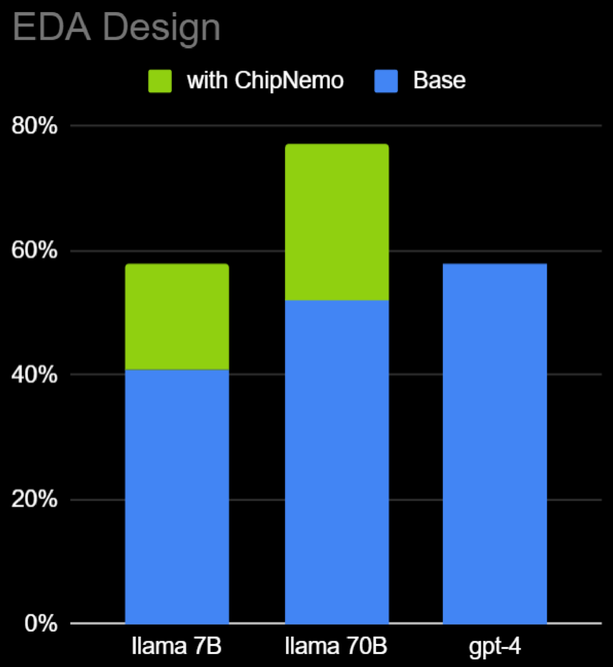
```
def partitions(set):
    if not set:
        yield []
        return
    for i in range(2 ** len(set) // 2):
        parts = [set[j] for j in range(len(set))
  if (i >> j) & 1]
        yield parts

def count_partitions(set):
    count = 0
    for _ in partitions(set):
        count += 1
    return count

# Example usage:
my_set = {1, 2, 3}
print("Number of partitions:", count_partitions(my_set))
```

GPT-4 No Fine-tuning

LLM Customizing with NeMo Microservices

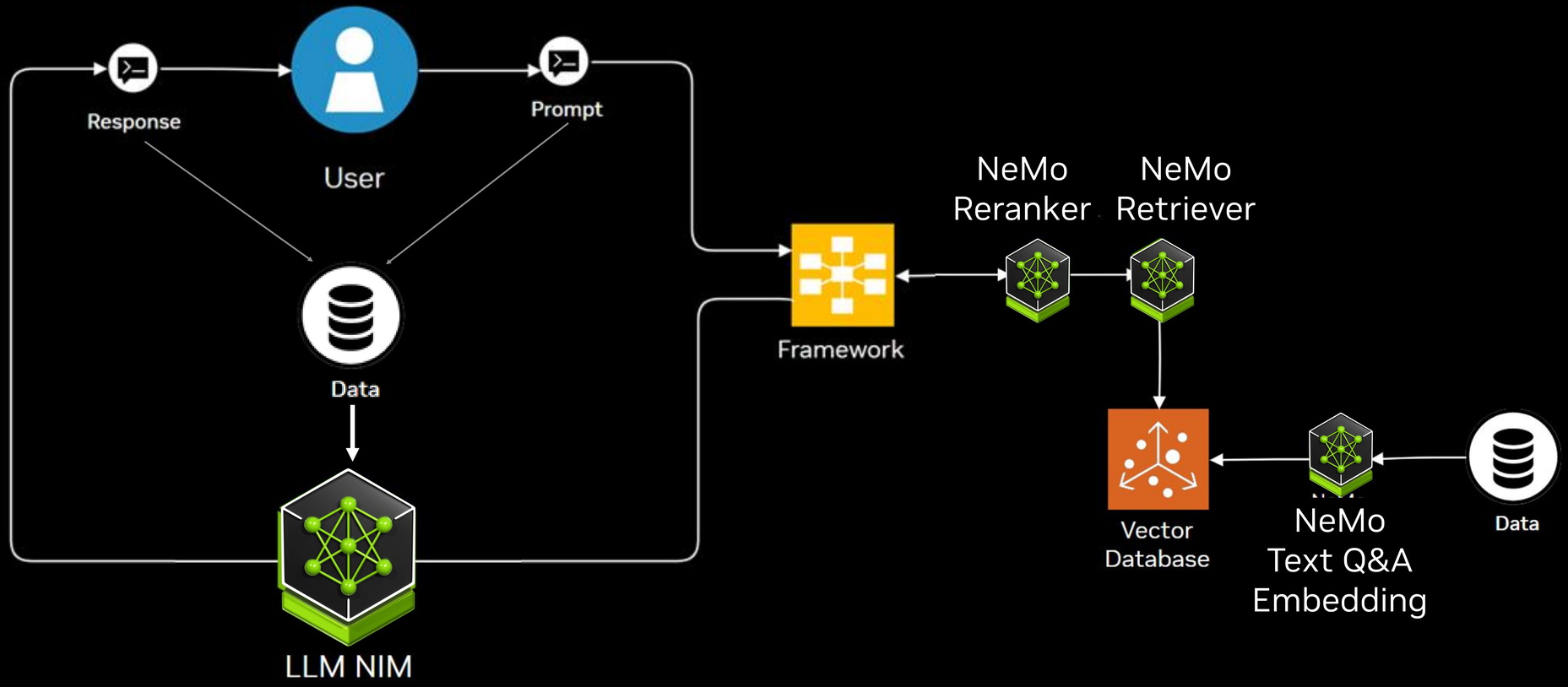


Write VIVID code to get partitions in a design

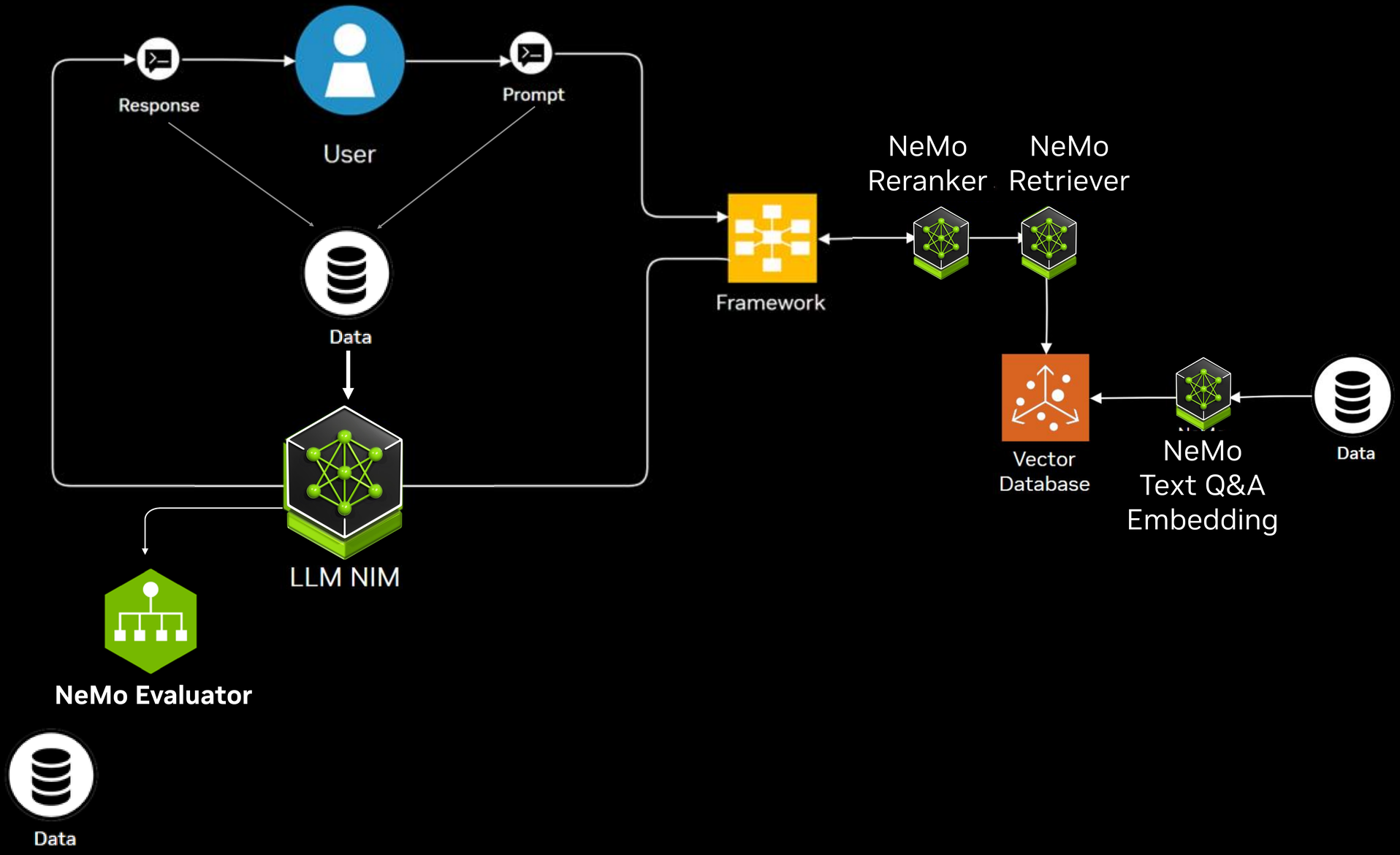
```
get_ref("*").child_refs(Opt.hier).filter("is_partition")
```

Llama-70B Fine-tuned ChipNeMo

Adding Domain Data to the LLM



Adding Domain Data to the LLM



NeMo Microservices - Evaluator

Custom NIM

NeMo Data Store

Evaluation Results

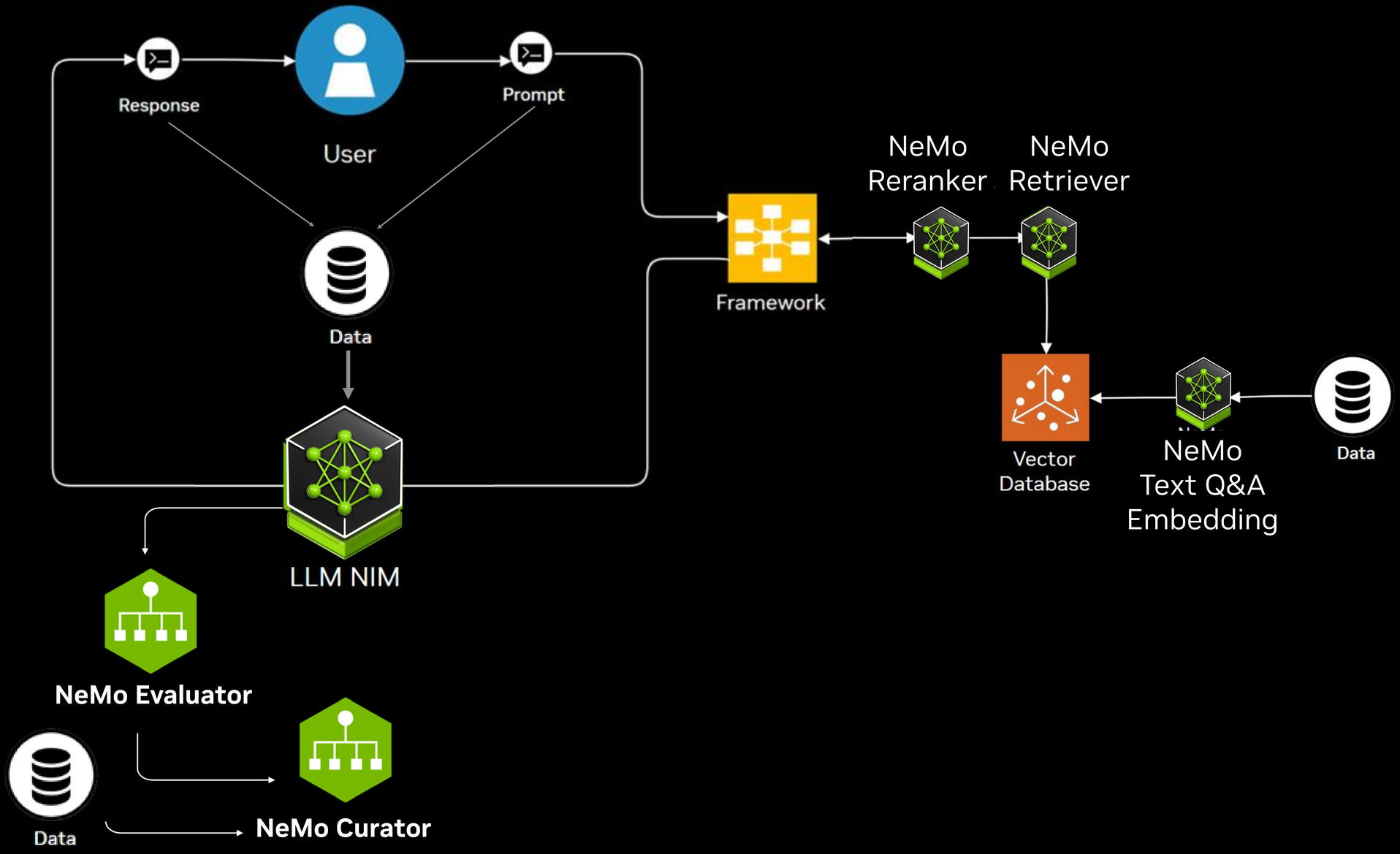
NeMo Evaluator

Automated evaluation of foundation models and fine-tuned LLMs through

- Academic Benchmarks
- Custom Benchmarks
- LLM-as-a-Judge
- Platform for human evaluation and scoring

NIM

Adding Domain Data to the LLM



NeMo Microservices - Curator

Custom NIM

NeMo Data Store

Evaluation Results

Prompts & Responses
Quality filtering

NeMo Evaluator

Automated evaluation of foundation models and fine-tuned LLMs through

- Academic Benchmarks
- Custom Benchmarks
- LLM-as-a-Judge
- Platform for human evaluation and scoring

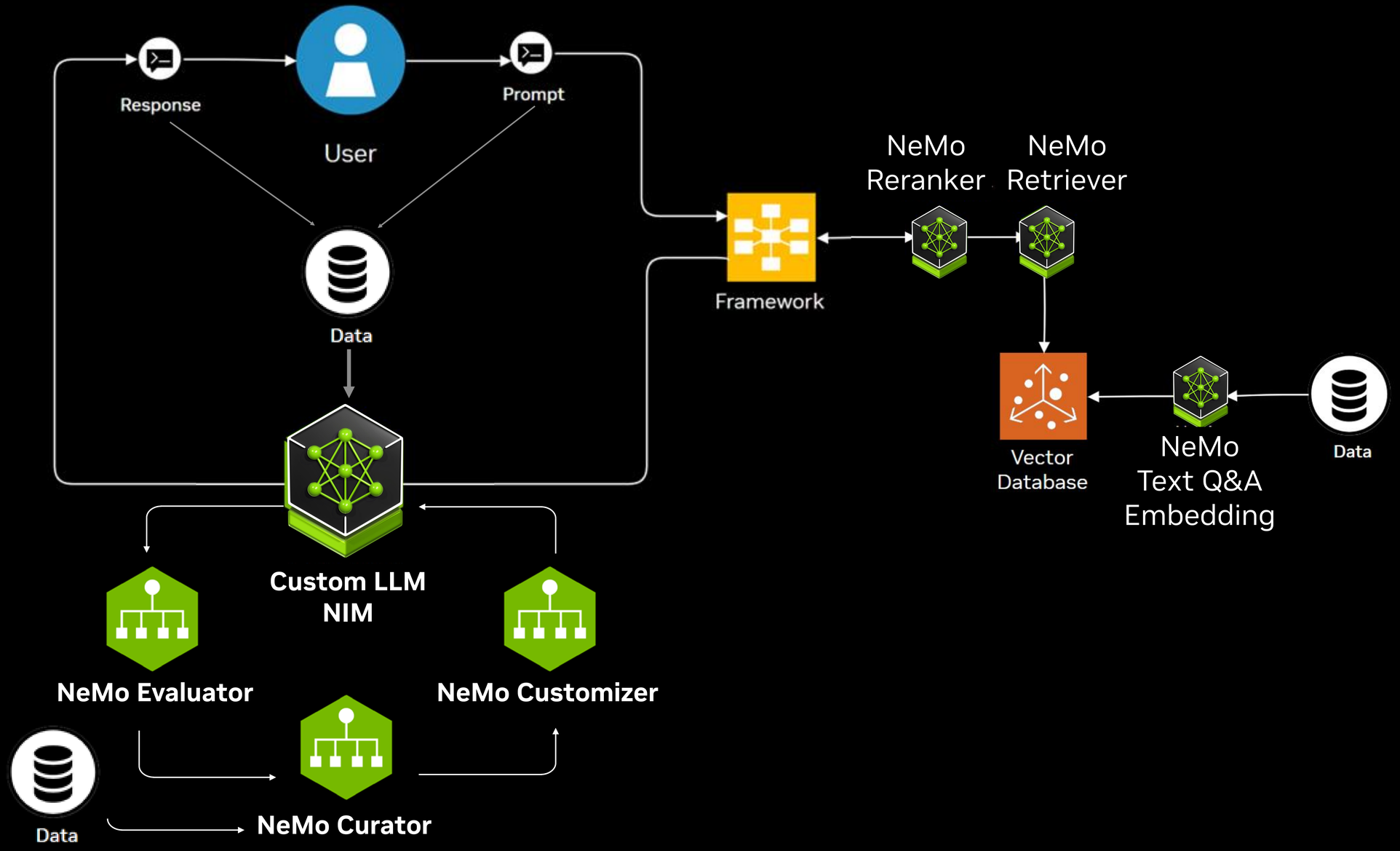
NIM

NeMo Curator

multi-stage curation of high- quality datasets for pre-training and fine-tuning workflows;

RAPIDS

Adding Domain Data to the LLM



NeMo Microservices - Customizer

Custom NIM

NeMo Data Store

Evaluation Results

Prompts & Responses
Quality filtering

Adapters & P-Tokens
.nemo checkpoints

NeMo Evaluator

Automated evaluation of foundation models and fine-tuned LLMs through

- Academic Benchmarks
- Custom Benchmarks
- LLM-as-a-Judge
- Platform for human evaluation and scoring

NIM

NeMo Curator

multi-stage curation of high- quality datasets for pre-training and fine-tuning workflows;

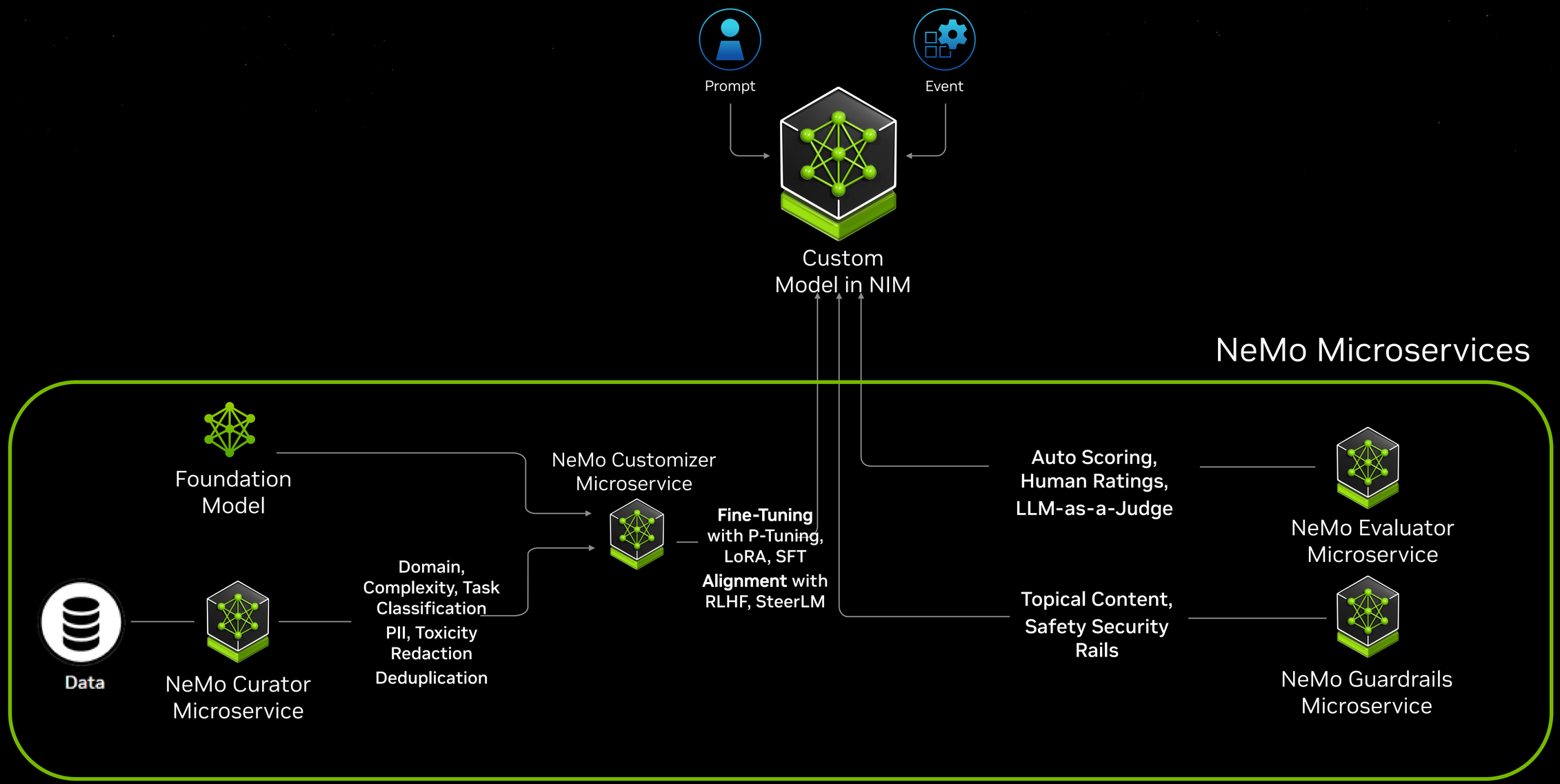
RAPIDS

NeMo Customizer

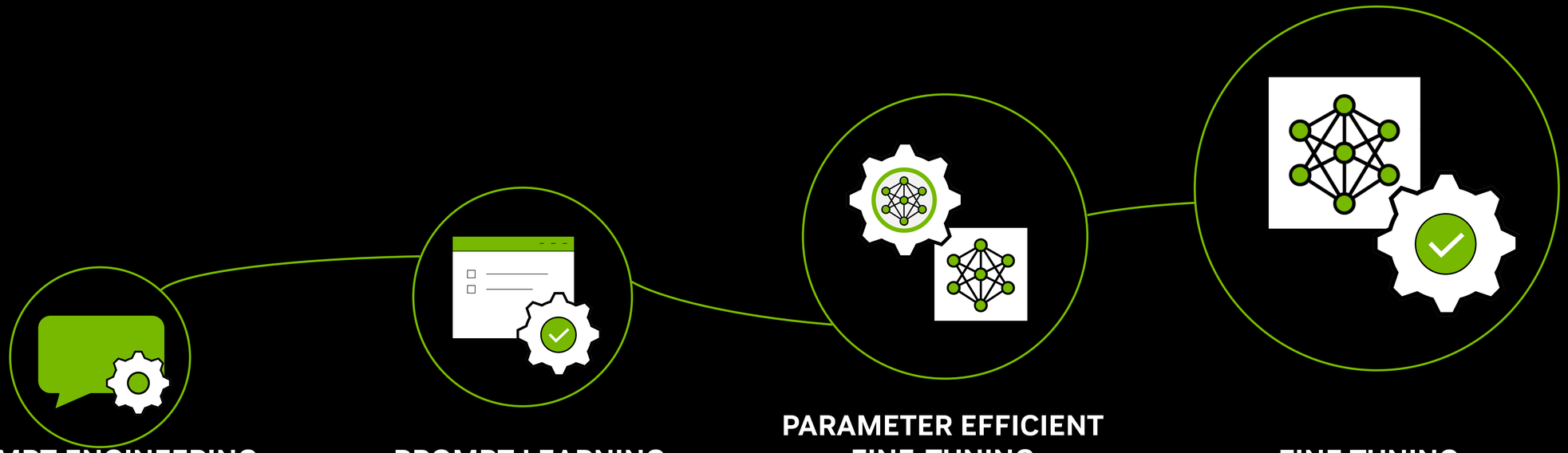
Latest state of the art customization and alignment techniques with easy to use interface, to accommodate users' specific requirements on data/compute availability

NeMo Framework

Adding Domain Data to the LLM



Model Customization Techniques



	PROMPT ENGINEERING	PROMPT LEARNING	PARAMETER EFFICIENT FINE-TUNING	FINE TUNING
Techniques	. Few-shot learning . Chain-of-thought reasoning . System prompting . Prompt Templates	. Prompt tuning . P-tuning	. Adapters . LoRA . IA3	. SFT . RLHF . SteerLM
Data & Compute	None	Limited	Moderate	High
Purpose	Ask Better Questions	Learn new skills	Add incremental knowledge	Add domain-specific knowledge
How	Prompt Questions	Tune (LSTM) companion model	Add “custom” layers to LLM	Tune LLM model weights
Challenges	. Cannot add as many skills or domain specific data to pre-trained LLM	. Less comprehensive ability to change all model parameters	. Building domain specific labeled dataset	. May forget old skills . Large investment . Most expertise needed

LLM NIM

NeMo Retriever & Reranker

NeMo Evaluator

NeMo Curator

NeMo Customizer

AI Foundry Machinery:
DGX Cloud

The World's Best Clouds – Now Supercharged with NVIDIA AI

Co-engineered with leading CSPs

aws Google Cloud ORACLE CLOUD Infrastructure Microsoft Azure

Best-of-Breed architecture at **every layer**

Simplify and **scale** resources

Optional: Your Preferred MLOps

AI and Data Science Development and Deployment Tools

NVIDIA AI Enterprise

Job Scheduling Orchestration

Workload and Infrastructure Management

Co-engineered CSP Infrastructure

Low-latency, high performance networking

A100 | H100 | GB200 (coming soon)

The AI Platform for Enterprise Developers

Integrated full-stack solution co-engineered with leading CSPs

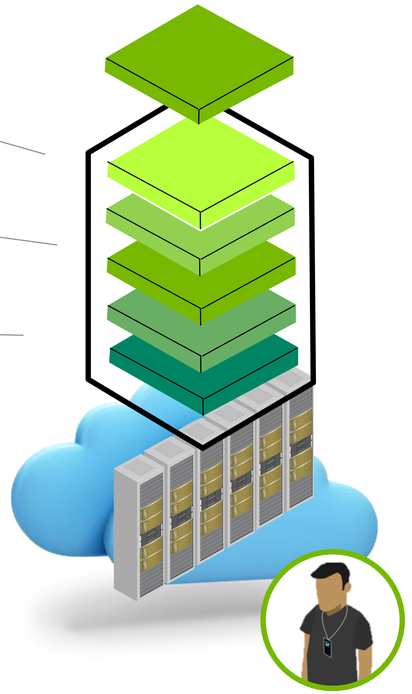
NVIDIA AI Enterprise

Train, fine-tune, inference in one service

CSP co-engineered technology stack

Direct access to AI expertise

One price with no long-term commitment

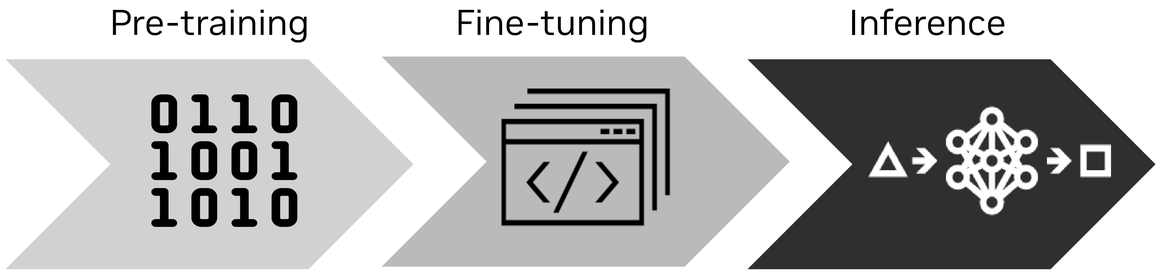


Software that unleashes developer productivity

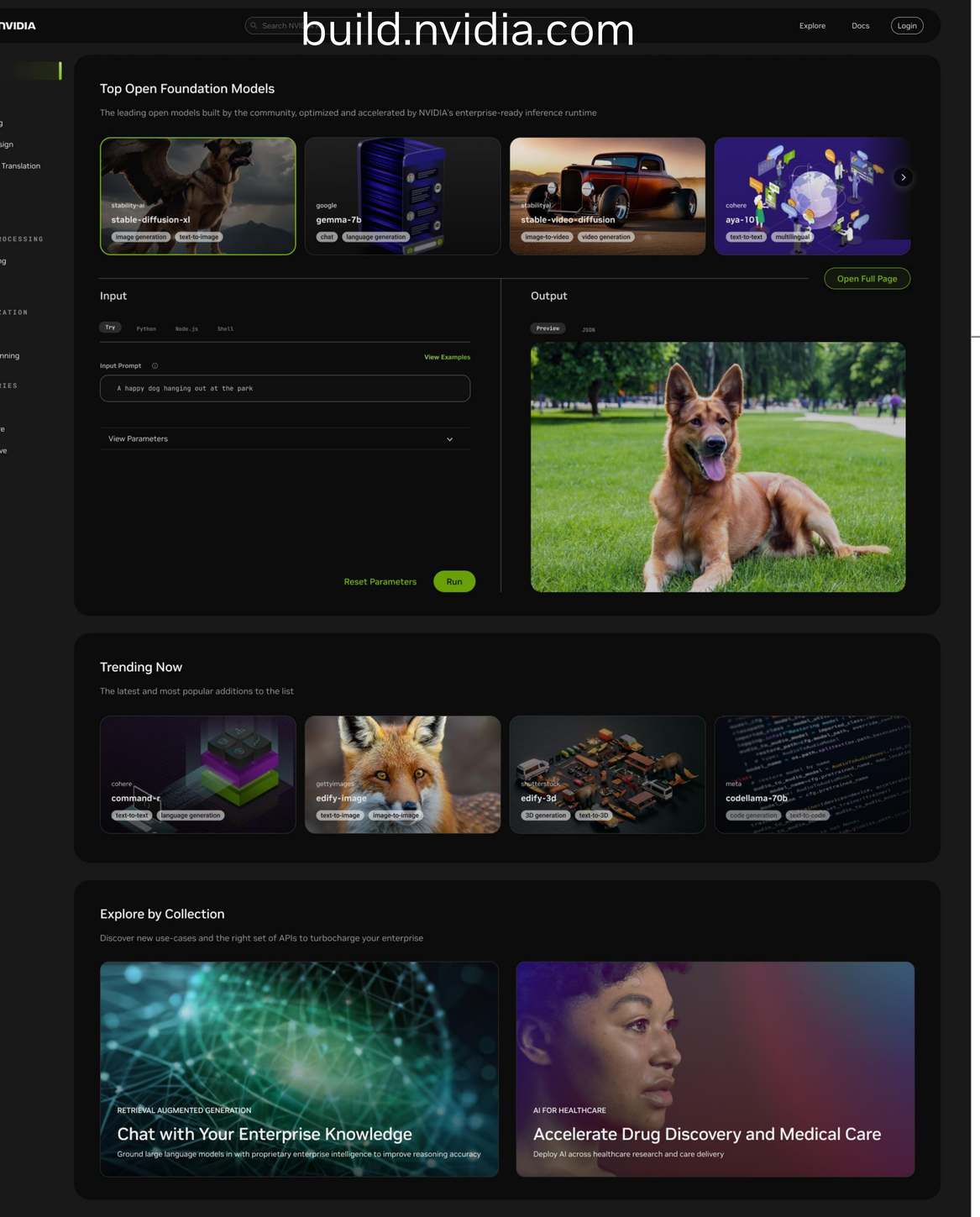
Access AI practitioners who help maximize performance

Best of breed performance

Hosted in **leading clouds**



Ready to try it?



AI Foundry Takeaways

NVIDIA AI Foundry components:

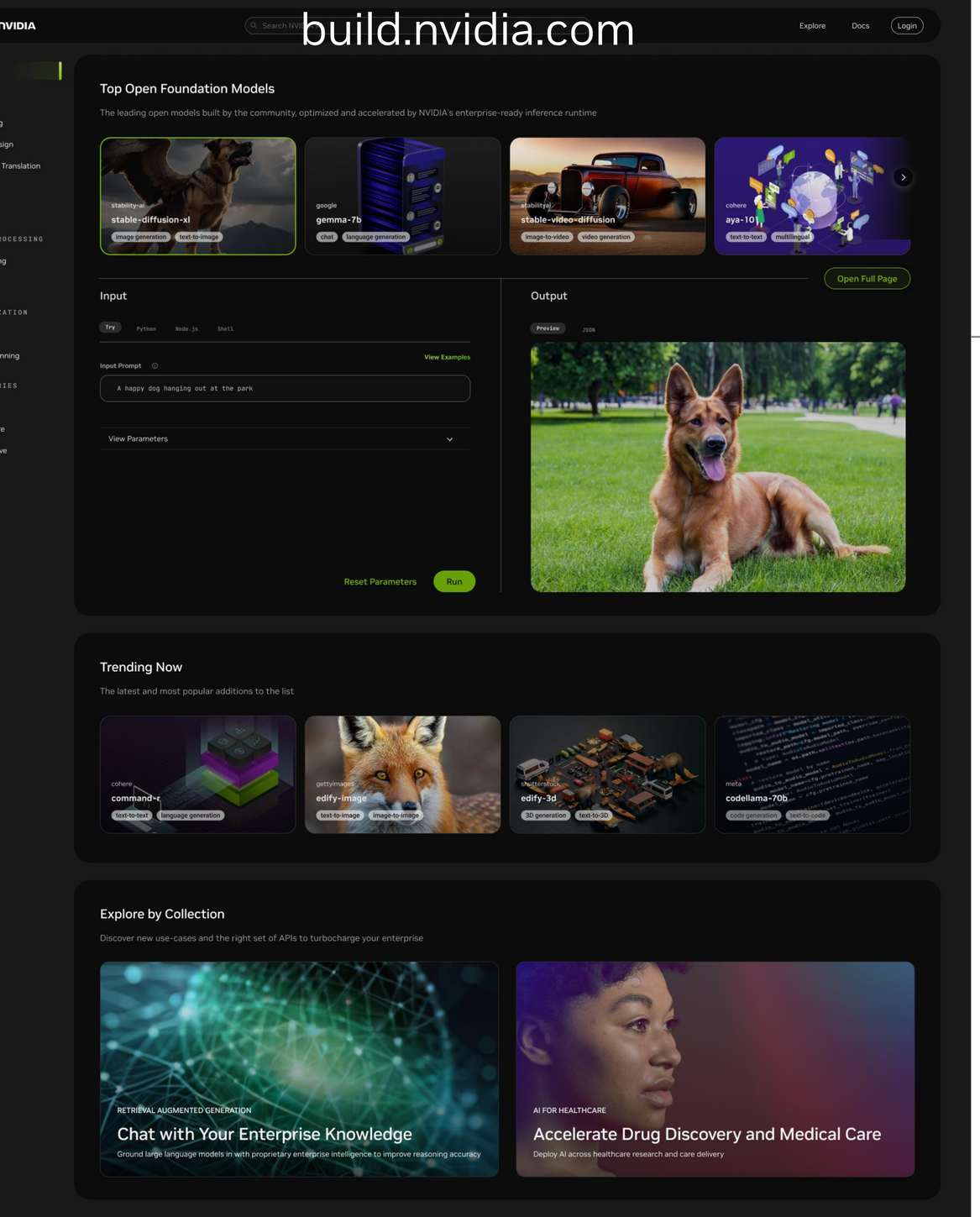
Input: Proprietary Data

AI Foundry:

1. NIMs
2. NeMo Microservices
3. DGX Cloud Compute
4. AI Expertise

Output: Domain Specific Models

NVIDIA



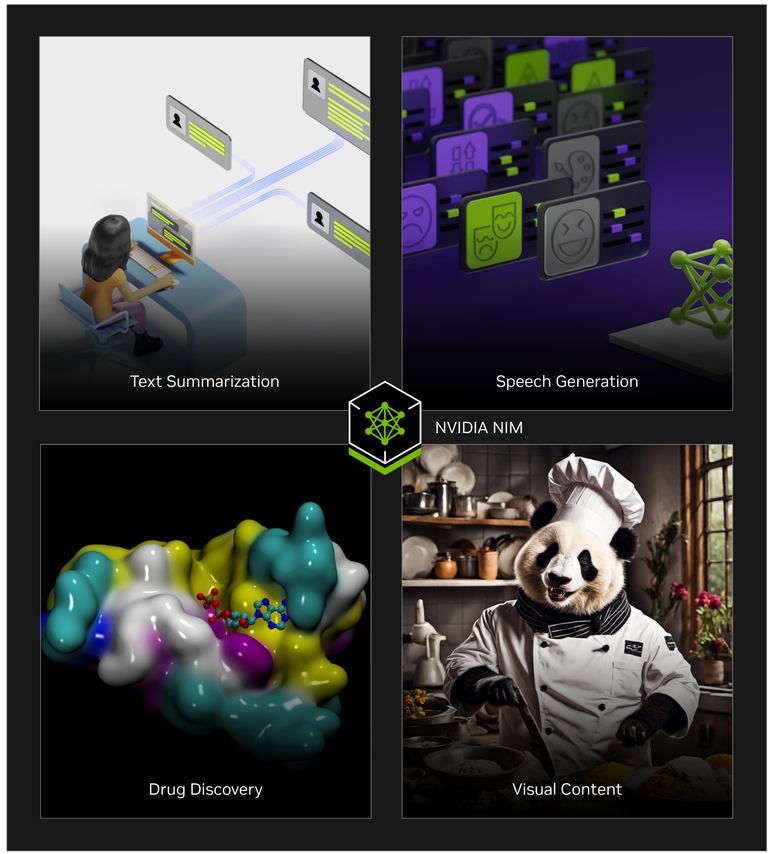
Resources to Get Started

- Explore **NVIDIA API Catalog**: **https://build.nvidia.com/**
- **NVIDIA RAG:**
  - https://build.nvidia.com/explore/retrieval
  - https://github.com/NVIDIA/GenerativeAIExamples
- **NeMo Microservices:**
  - Apply for Early Access: developer.nvidia.com/nemo-microservices-early-access
  - https://developer.nvidia.com/docs/nemo-microservices/index.html

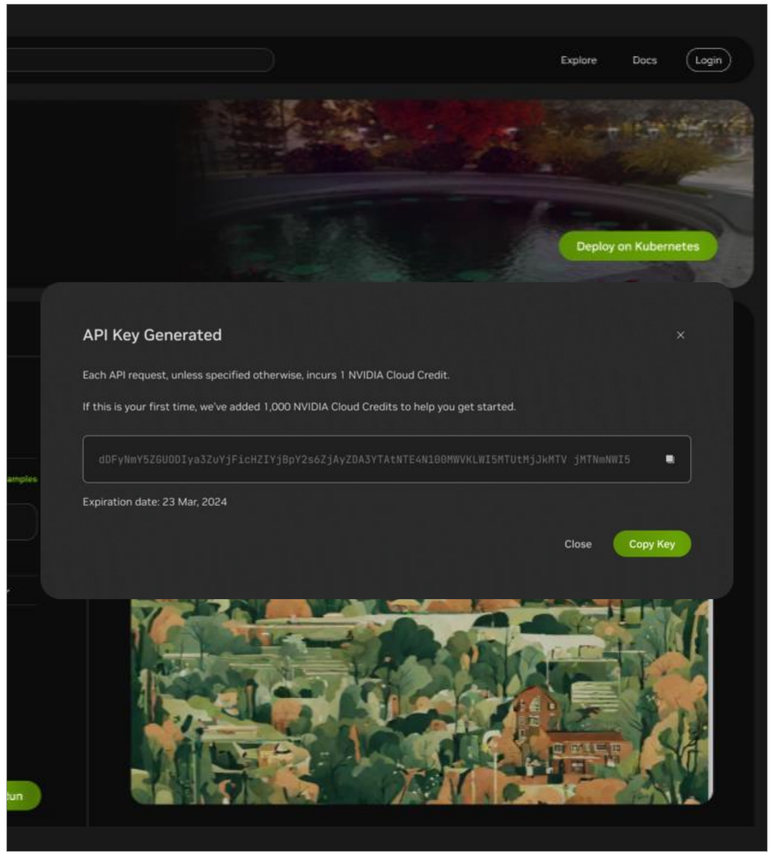
NVIDIA

ai.nvidia.com

Call To Action



Experience Models

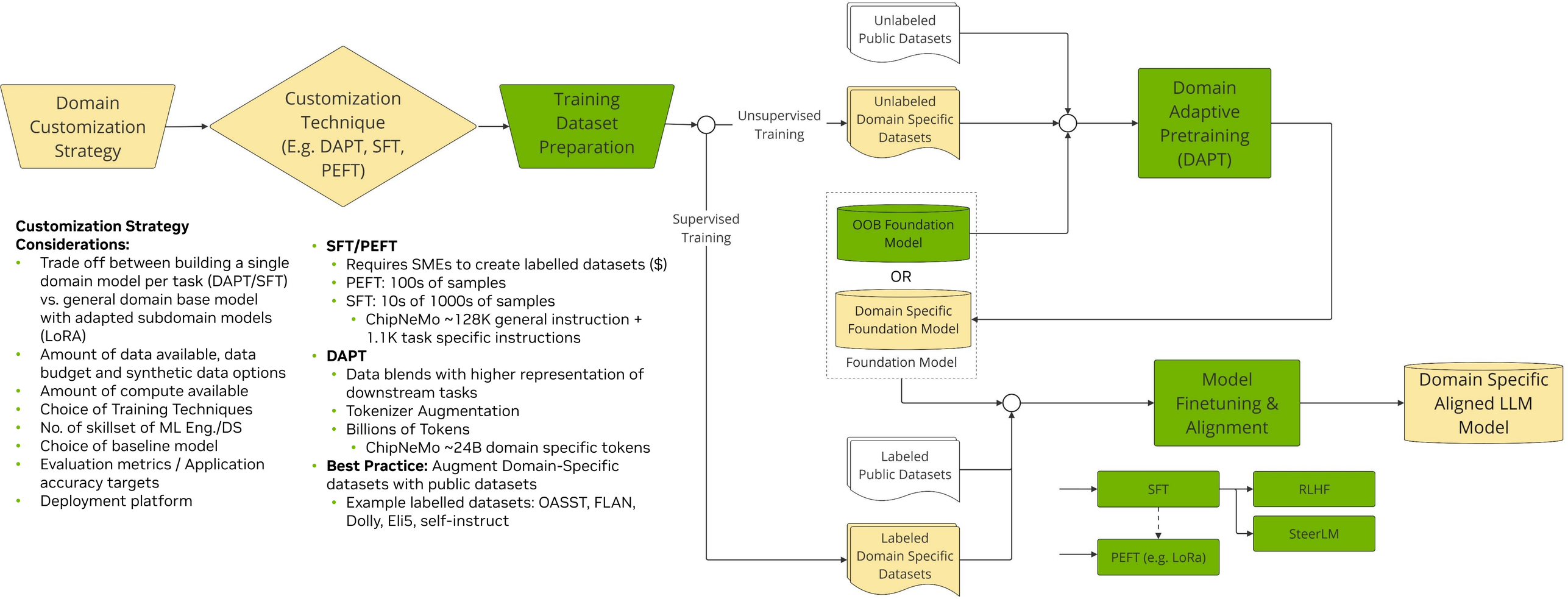


Prototype with APIs

Deploy with NIMs

Domain Adapted LLM Generator Model

NeMo Framework



Customization Strategy Considerations:

- Trade off between building a single domain model per task (DAPT/SFT) vs. general domain base model with adapted subdomain models (LoRA)
- Amount of data available, data budget and synthetic data options
- Amount of compute available
- Choice of Training Techniques
- No. of skillset of ML Eng./DS
- Choice of baseline model
- Evaluation metrics / Application accuracy targets
- Deployment platform

- **SFT/PEFT**
  - Requires SMEs to create labelled datasets ($)
  - PEFT: 100s of samples
  - SFT: 10s of 1000s of samples
    - ChipNeMo ~128K general instruction + 1.1K task specific instructions
- **DAPT**
  - Data blends with higher representation of downstream tasks
  - Tokenizer Augmentation
  - Billions of Tokens
    - ChipNeMo ~24B domain specific tokens
- **Best Practice:** Augment Domain-Specific datasets with public datasets
  - Example labelled datasets: OASST, FLAN, Dolly, Eli5, self-instruct

Legend:
NVIDIA Tooling
Customer's Work & Decisions