



Driving the Generative AI Powered Enterprise with NIMs

Enterprise are on the Generative AI Journey



Explosion

ChatGPT gets announced late in 2022, gaining over 100 million users in just two months. Users of all levels can experience AI and feel the benefits firsthand.



Experimentation

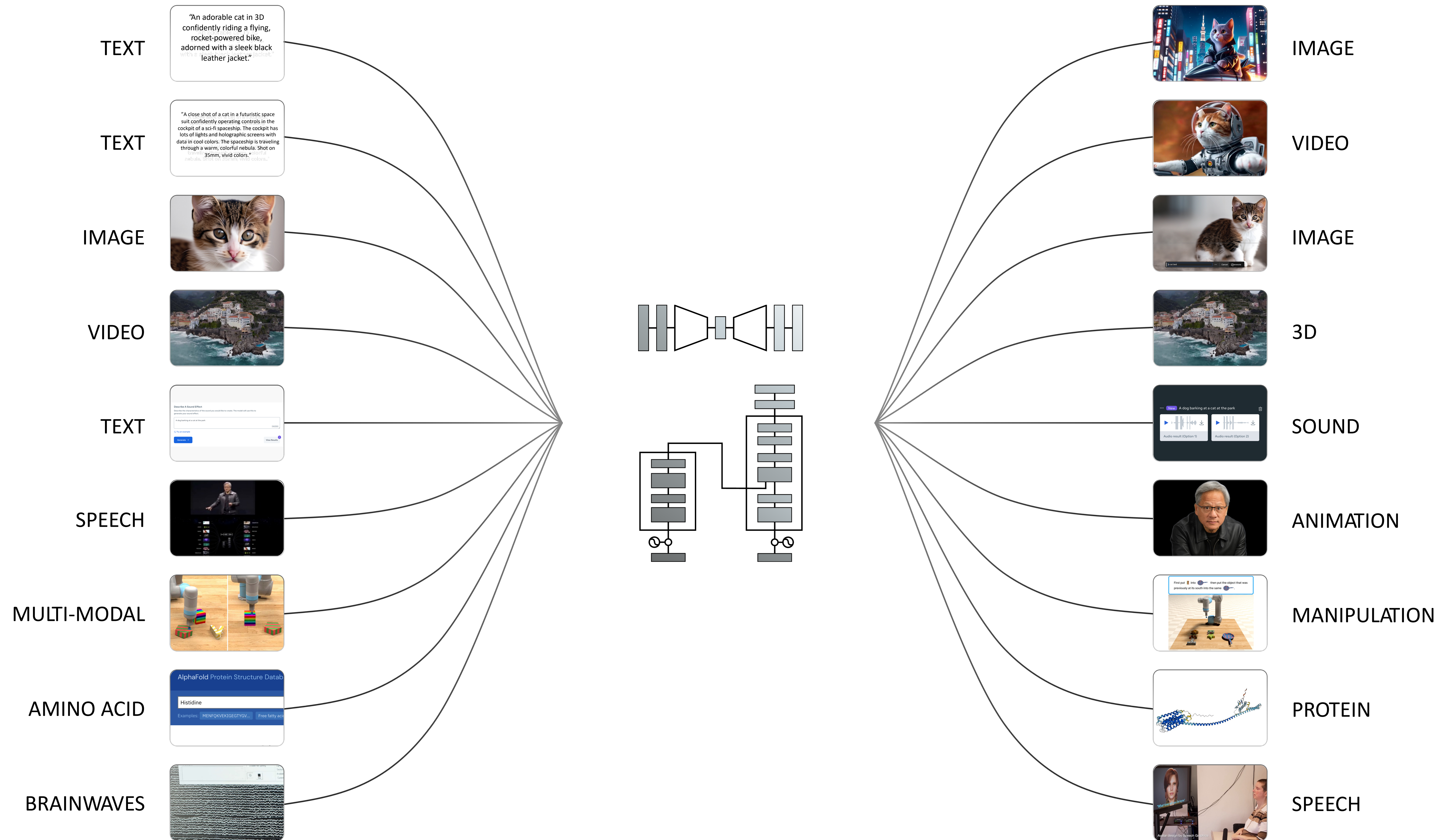
Enterprise application developers kick off POCs for generative AI applications with API services and open models including Llama 2, Mistral, NVIDIA, and others.



Production

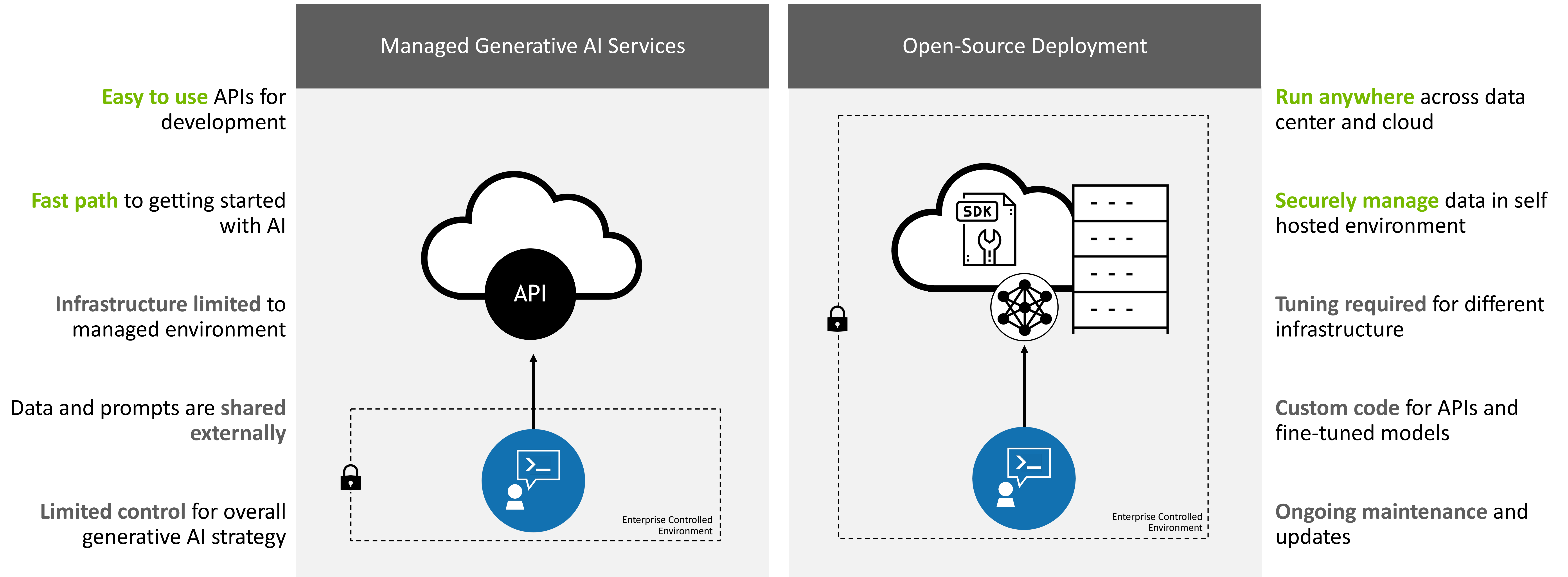
Organizations have set aside budget and are ramping up efforts to build accelerated infrastructure to support generative AI in production.

Generative AI Can Learn and Understand Everything



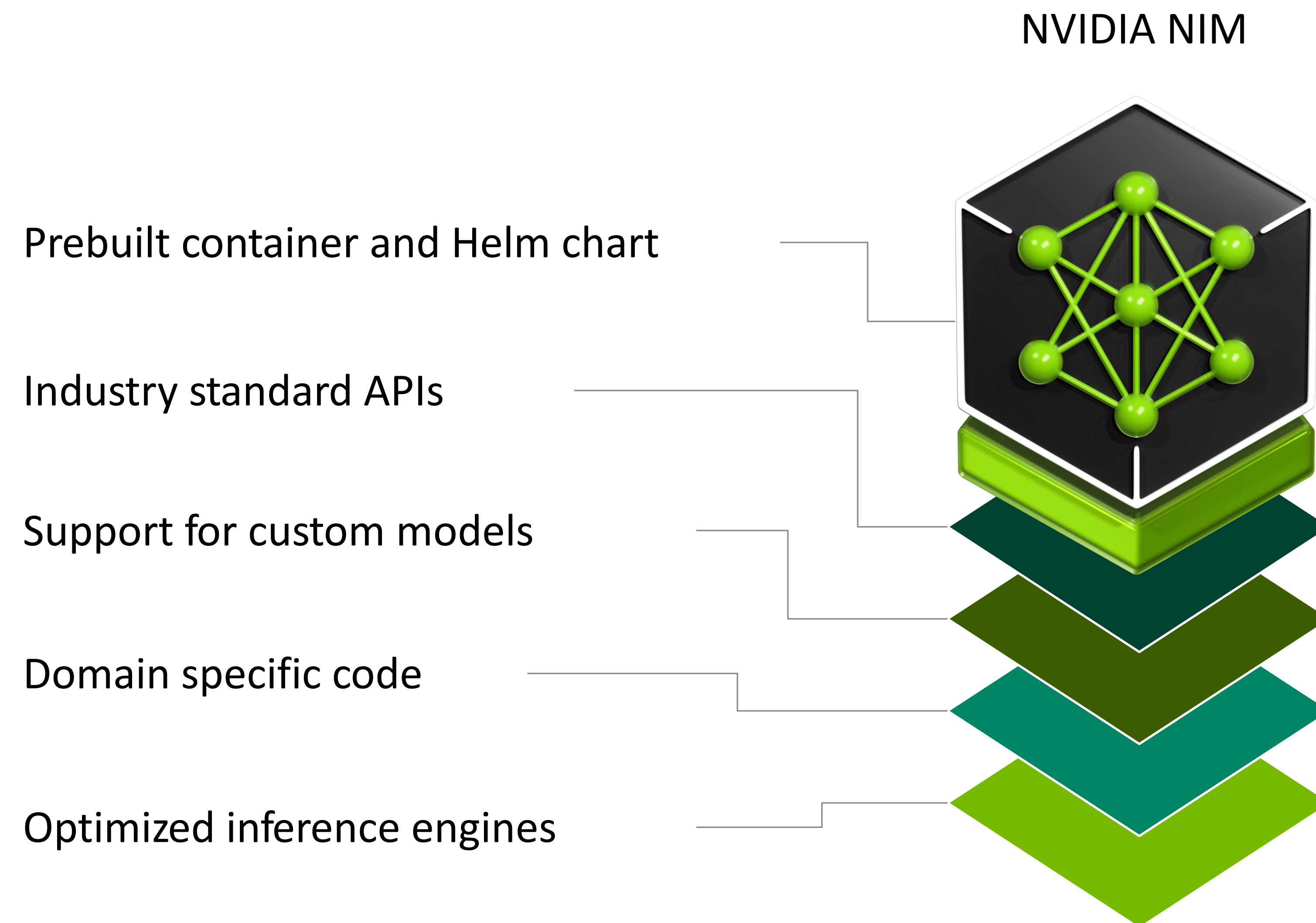
Enterprises Face Challenges Experimenting with Generative AI

Organizations must choose between ease of use and control



NVIDIA NIM Optimized Inference Microservices

Accelerated runtime for generative AI



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




DGX &
DGX Cloud



NVIDIA NIM is the Fastest Path to AI Inference

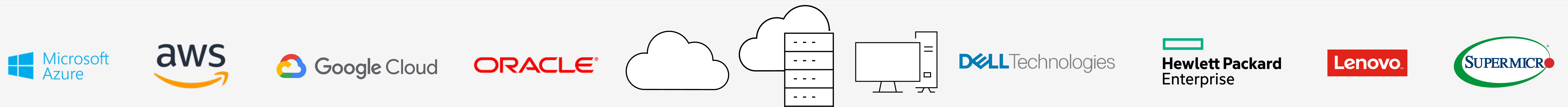
Reduces engineering resources required to deploy optimized, accelerated models

	NVIDIA NIM	Triton + TRT-LLM Opensource
Deployment Time	5 minutes	~1 week
API Standardization	Industry standard protocol OpenAI for LLMs, Google Translate Speech	User creates a shim layer (reducing performance) or modify Triton to generate custom endpoints
Pre-Built Engine	Pre-built TRT-LLM engines for NV and community models 	User converts checkpoint to TRT-LLM format and creates and runs sweeps through different parameters to find the optimal config
Triton Ensemble/ BLS Backend	Pre-built with TRT-LLM to handle pre/post processing (tokenization)	User manually sets up + configures
Triton Deployment	Automated	User manually sets up + configures
Customization	Supported – P-tuning and LORA, more planned	User needs to create custom logic
Container Validation	Pre-validated with QA testing	No pre-validation
Support	NVIDIA AI Enterprise - Security and CVE scanning/patching and tech support	No enterprise support

Inference Microservices for Generative AI

NVIDIA NIM is the fastest way to deploy AI models on accelerated infrastructure across cloud, data center, and PC

NVIDIA API Catalog



NVIDIA NIM for Every Domain

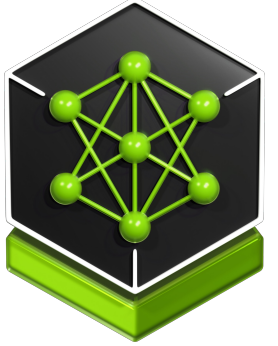
LANGUAGE NIMs



Code Llama 70B



Cohere 35B



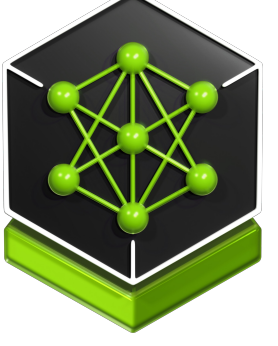
Gemma 7B



Jamba



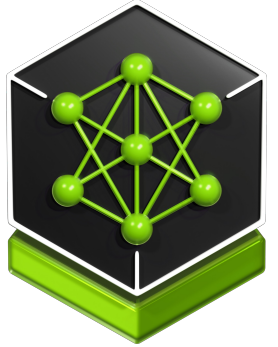
Llama 2 70B



Mistral 7B



Mixtral 8x7B



Nemotron-3 22B Persona



Phi-2

VISUAL / MULTIMODAL NIMs



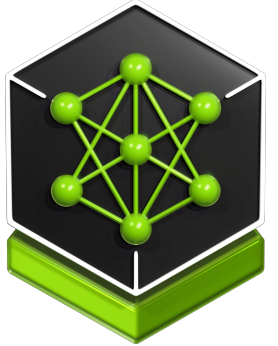
Adept 110B



Deplot



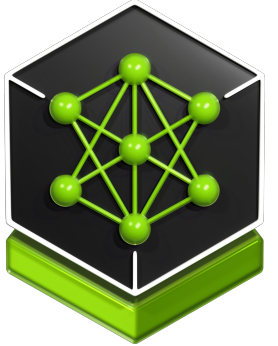
Edify. Getty



Edify. Shutterstock



FuYu 8B, 55B



Kosmos-2



NeVA



SDXL 1.0



SDXL Turbo

DIGITAL HUMAN NIMs



Audio2Face



Riva ASR

OPTIMIZATION / SIMULATION NIMs



cuOpt



Earth-2

DIGITAL BIOLOGY NIMs



DeepVariant



DiffDock



ESMFold



MolMIM



Vista 3D

APPLICATION NIMs



Llama Guard



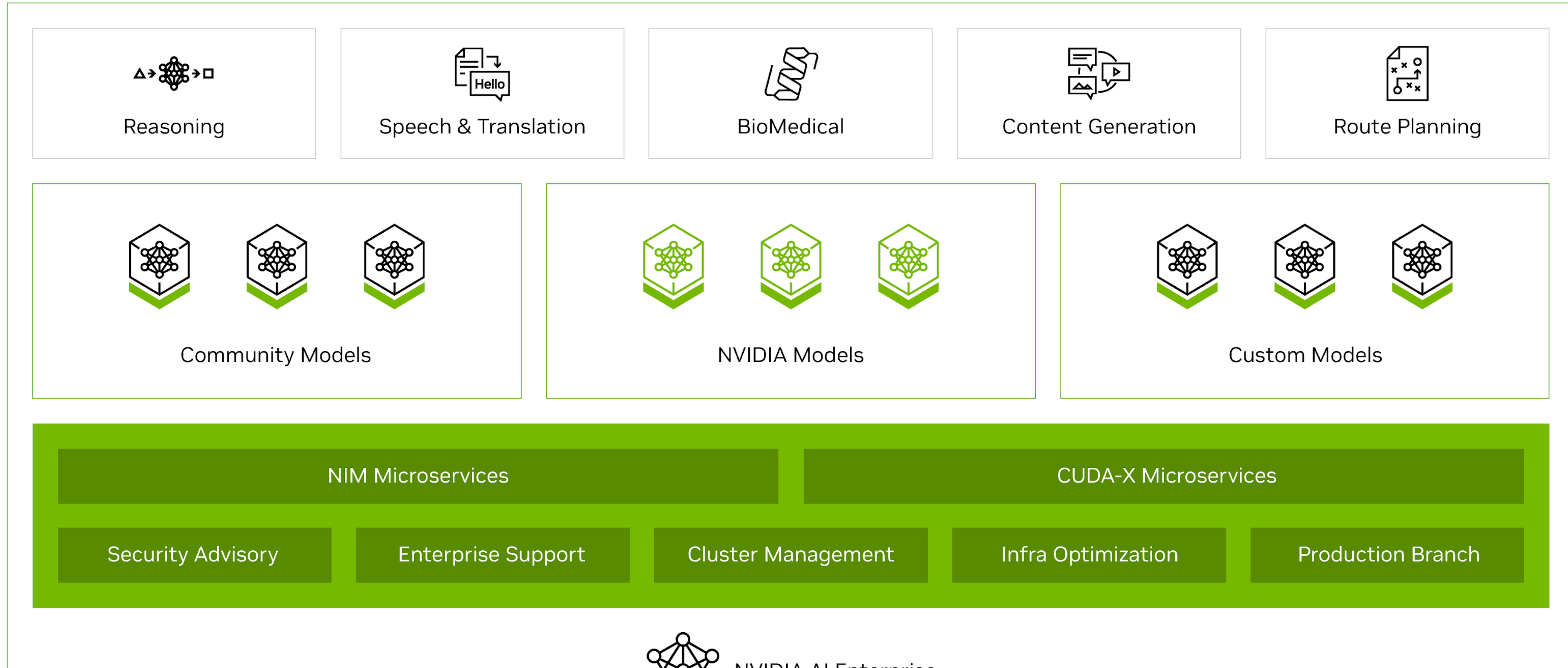
Retrieval Embedding



Retrieval Reranking

NVIDIA AI Enterprise

High Performance and Efficient Runtime for Generative AI



Cloud | Data Center | Workstations | Edge



Additional slides to be used as needed

NVIDIA NIM

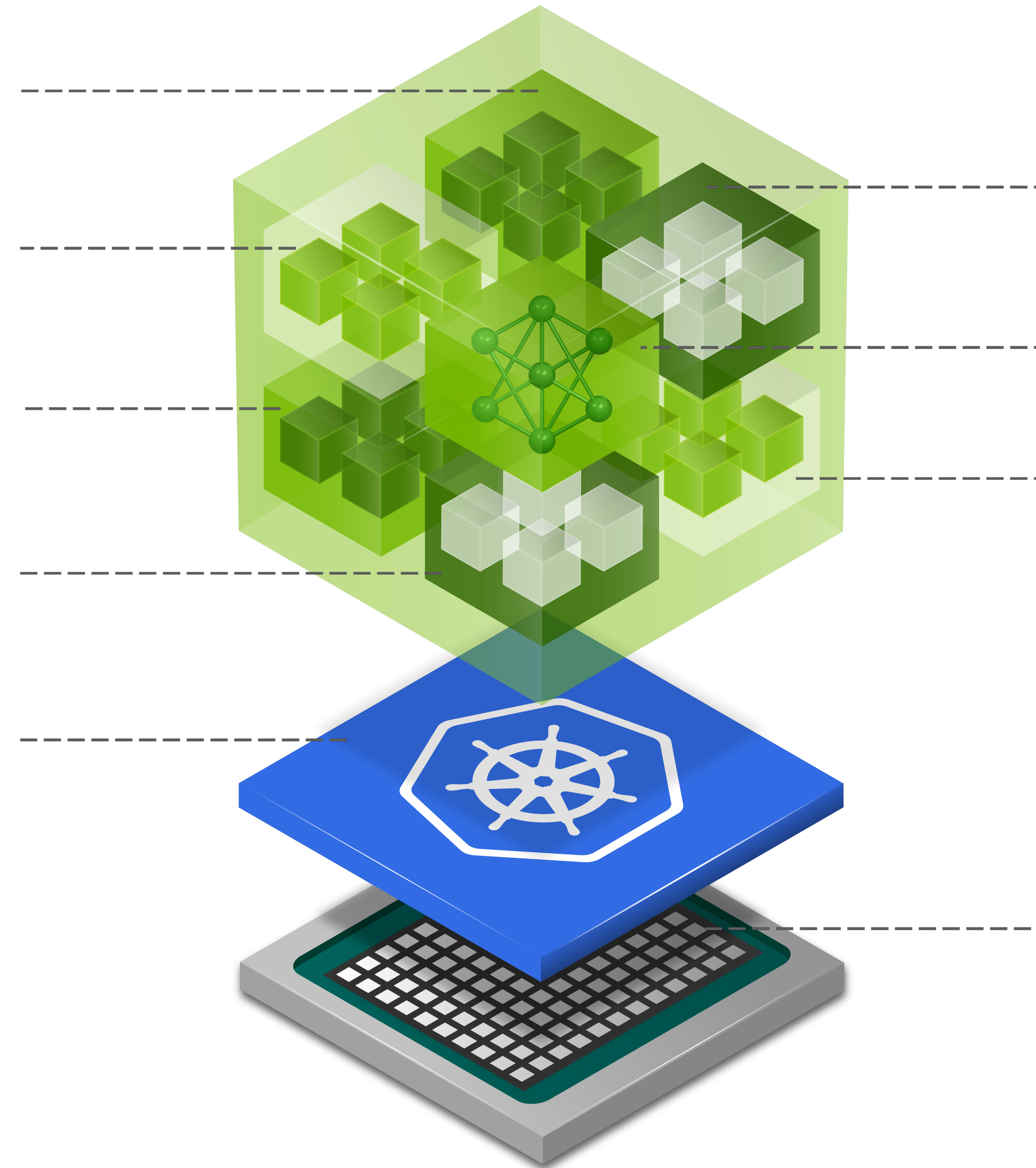
Industry Standard APIs
Text, Speech, Image,
Video, 3D, Biology

Triton Inference Server
cuDF, CV-CUDA, DALI, NCCL,
Post Processing Decoder

Cloud Native Stack
GPU Operator, Network Operator

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

Kubernetes



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

NVIDIA CUDA

100's of Millions of CUDA GPUs Installed Base

Anatomy of a NIM



Prompt



Event

Industry Standard APIs
Text | Speech | Image | Video | 3D | Biology

Cloud Native Container
K8s Support | Metrics & Monitoring | Identity | Secret Management | Liveness Probe

Triton Inference Server
cuDF | CV-CUDA | DALI | NCCL

Pre Processing

Post Processing

Inflight Batching

TensorRT Engine
cuBLAS | cuDNN | In-Flight Batching | Memory Optimization | FP8 Quantization

Customization Cache
LORA | P-tuning

AI Model
Text-to-Text | Text-to-Image | Text-to-3D | Multimodal | ASR | Text-to-Speech

```

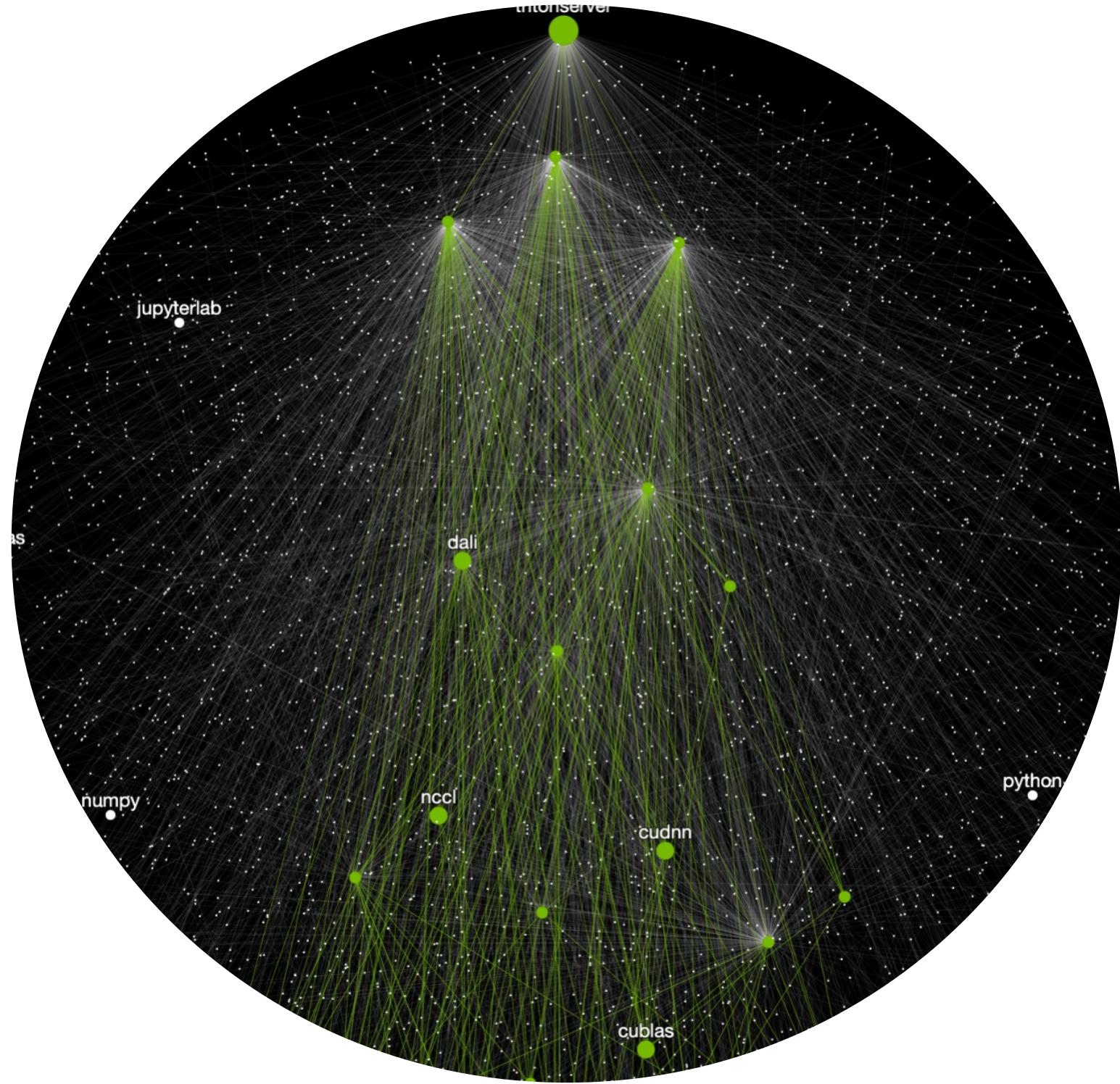
import requests

session = requests.Session()

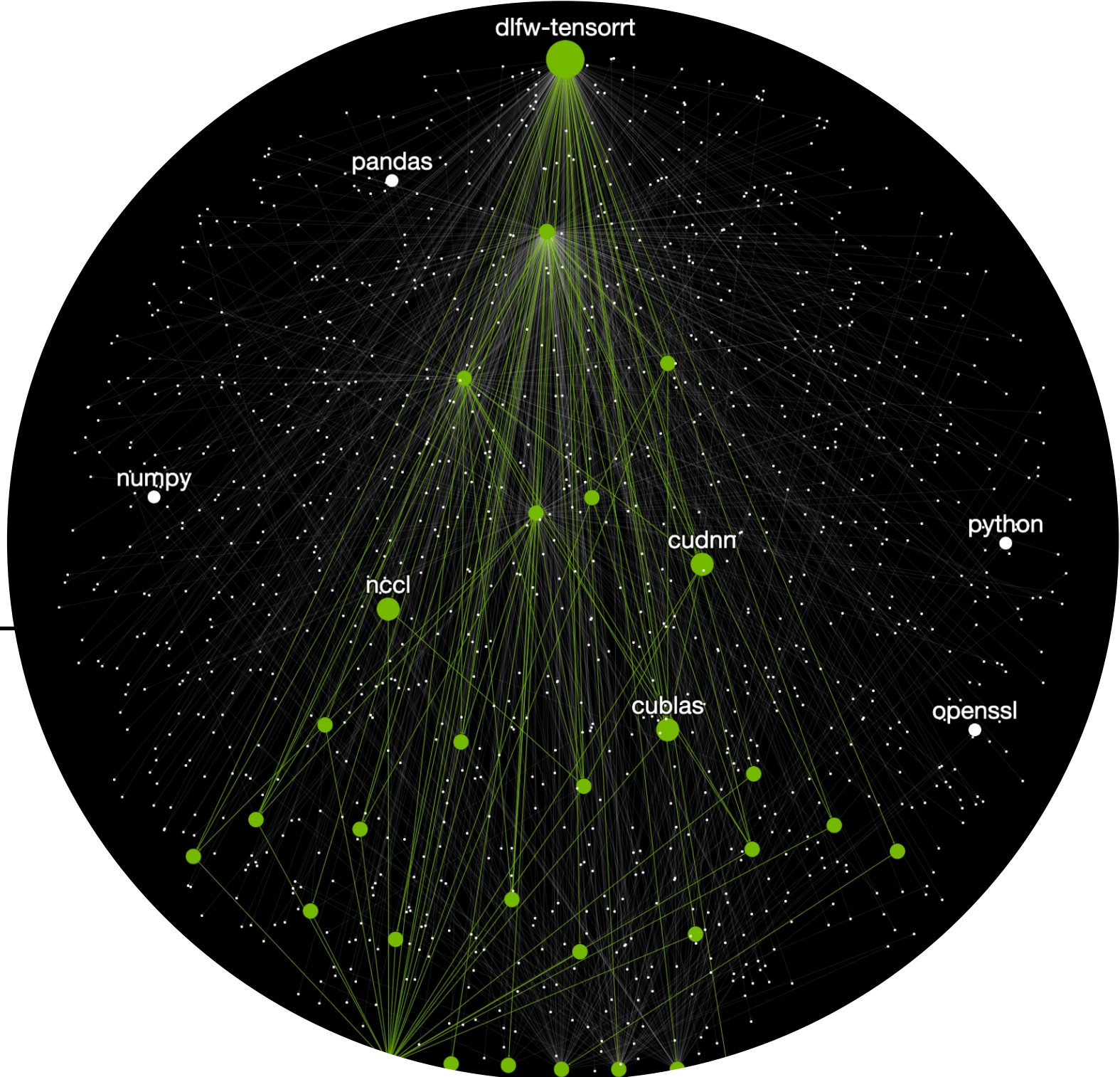
response = session.post(invoke_url,
                        data)

while response.status_code == 202:
    request_id = response.headers.get('request-id')
    fetch_url = fetch_url_format + request_id
    response = session.get(fetch_url)

response.raise_for_status()
response_body = response.json()
return response_body
    
```



Triton has 417 packages/libraries across OSS, 3rd party and NVIDIA



TensorRT has 333 packages/libraries across OSS, 3rd party and NVIDIA

