Technical Challenges of Generative AI

Meriem Bendris, Senior Deep Learning Data Scientist, NVIDIA

“

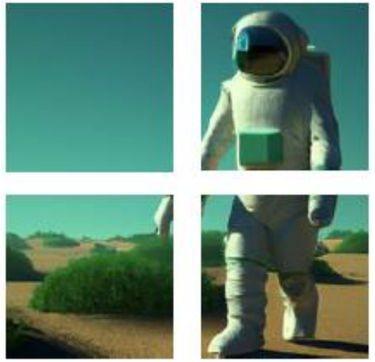
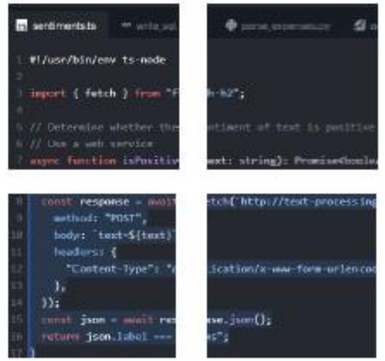
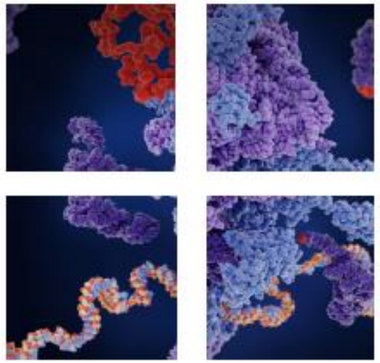
Generative AI?

”

Large Language Models Unlock New Opportunities

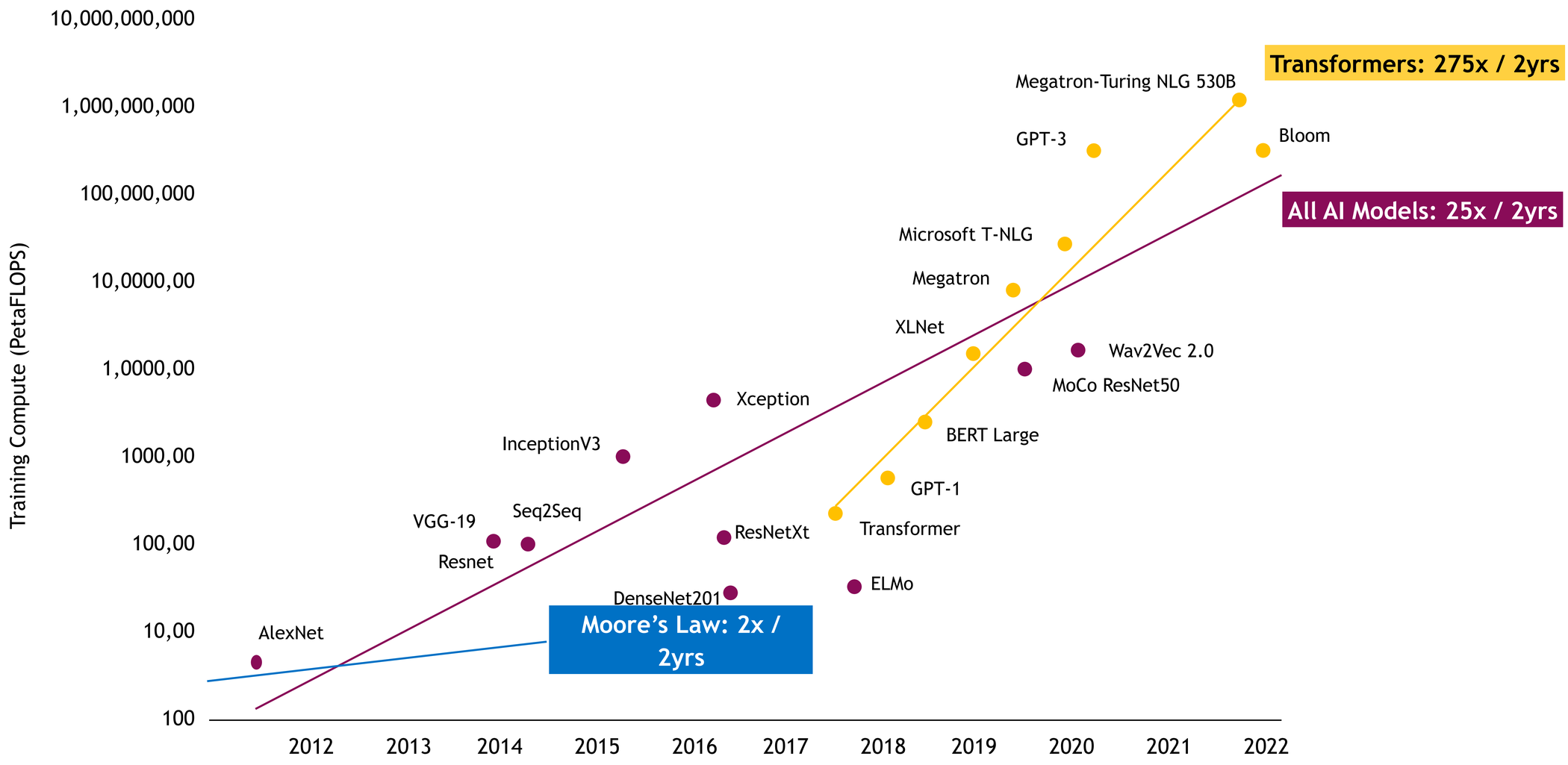
TEXT GENERATION	TRANSLATION	IMAGE GENERATION	CODING	LIFE SCIENCE
SummaCrization	Translating Wikipedia	Brand Creation	Dynamic Code Commenting	Molecular Representations
GPT-3	NLLB-200	DALL-E-2	CODEX	MegaMolBART
Marketing Copy	Real-Time Metaverse Translation	Gaming Characters	Function Generation	Drug Discovery



Higher Productivity | Faster Time to Market | Better Customer Experience

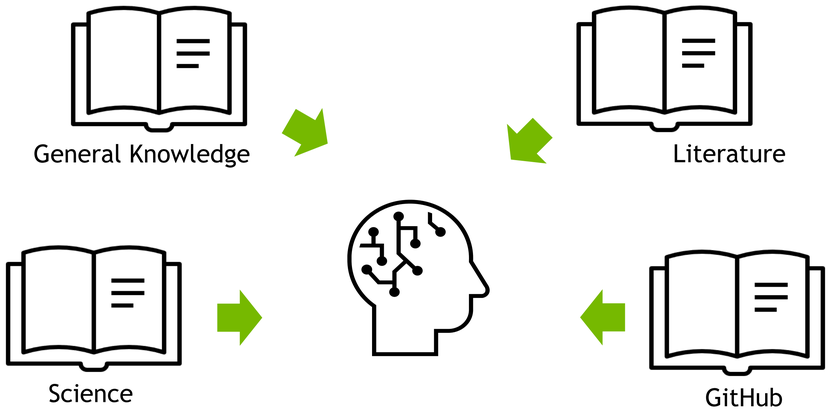
Increase in Model Sizes



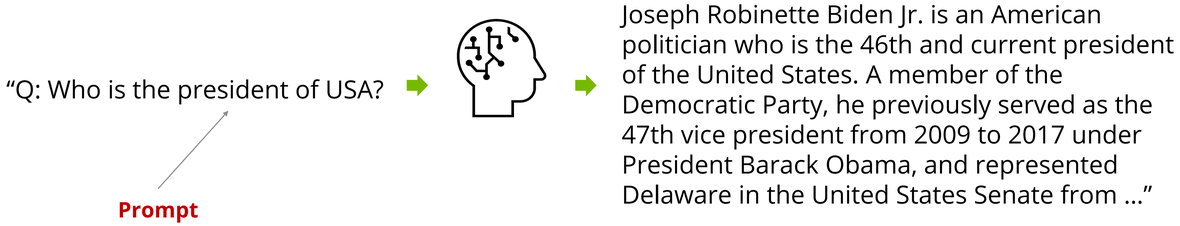
https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model

Fundamental Change in the NLP paradigm

Building Foundation Large Models



Using Foundation Models



Applications Powered by LLMs

Chatbot – Virtual Assistant – Search Engine – Tools Enabled

Examples of Applications Powered by LLMs

Expert, Natural Q&A with NVIDIA Omniverse Avatar for Project Tokkio

“

GPT

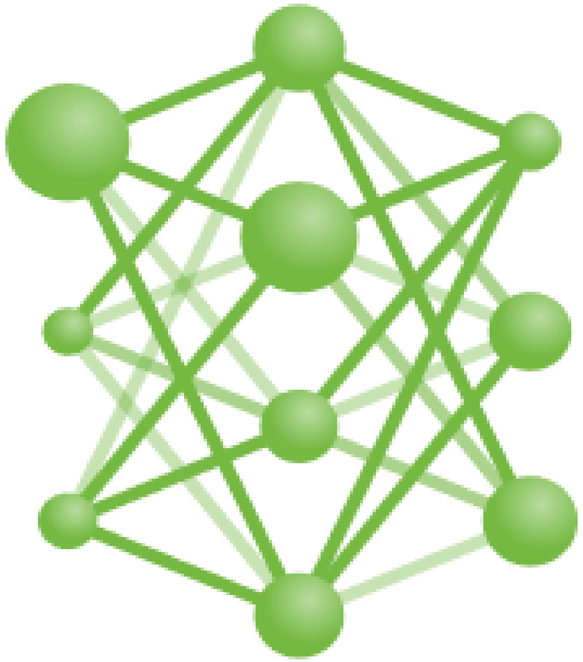
”

Large Language Model

Generative Pre-Training - GPT

Predicted Next token over the vocab

The cat is playing in the



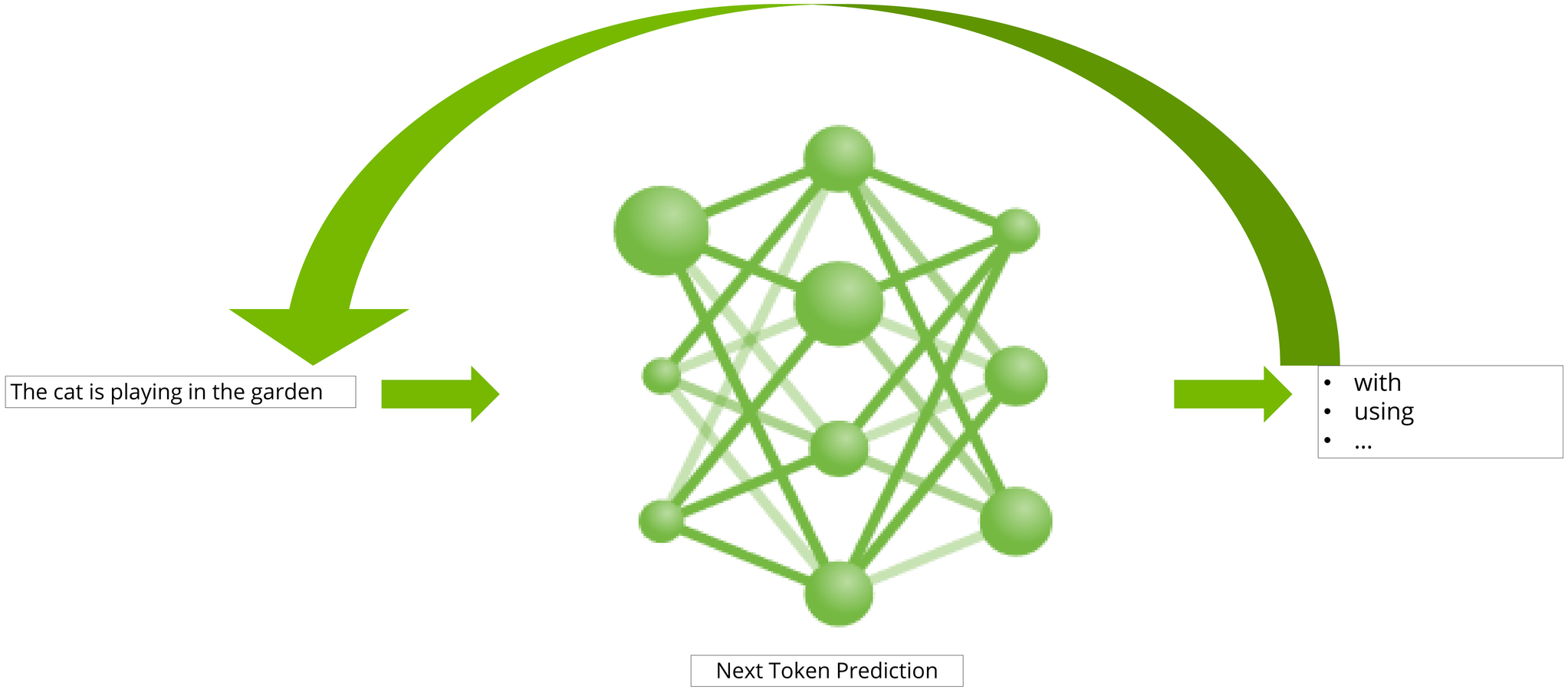
- garden
- kitchen
- ground
- grass
- ...

Prompt

Next Token Prediction

Large Language Model

Auto-Regressive Loop



Challenges of building Generative AI

How Big are LLMs?

Training - Model's Memory Footprint Rough Estimate

GPT3

175B Parameters

Weight

Gradient

Adam optimizer state

$$\frac{175 * 10^9 * (2 + 2 + 12)}{1024 * 1024 * 1024} = 2607.7GB$$

We want GPU memory consumption in GB

*Does not take into account the features maps

NeMo Framework Performance - Training

	Time to train 300B tokens in days (A100) – BF16			
	800 GPUs (5x DGX SuperPod)	480 GPUs (3x DGX SuperPod)	160 GPUs (1x DGX SuperPod)	64 GPUs (8x DGX A100)
GPT-3: 126M	0.07	0.12	0.37	0.92
GPT-3: 5B	0.8	1.3	3.9	9.8
GPT-3: 20B	3.6	6	18.1	45.3
GPT-3: 40B	6.6	10.9	32.8	82
GPT-3: 175B	28	46.7	140	349.9

Building Generative AI Foundation Models

Foundation models are AI neural networks trained on massive unlabeled datasets to handle a wide variety of tasks

Challenges of Building Foundation Models

	Mountains of Training Data
	Complex algorithms to build on large-scale infrastructure
	Deep technical expertise
	Large-scale compute infrastructure for training & inferencing, costing $10 M+ in just cloud costs

Challenges of Using Foundation models

	Don’t contain domain / enterprise specific knowledge
	Frozen in Time
	Hallucinate and provide undesired information
	Bias & Toxic Information

Building Generative AI Foundation Models

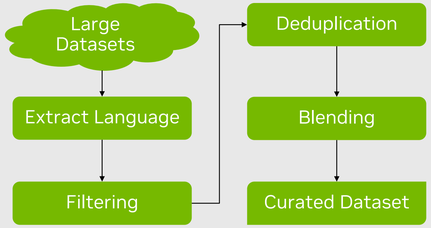
Efficiently and quickly training models using NeMo

Challenges of Building Foundation Models

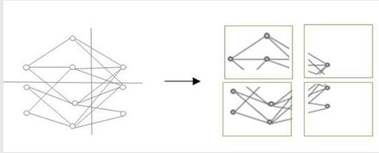
	Mountains of Training Data
	Complex algorithms to build on large-scale infrastructure
	Deep technical expertise
	Large-scale compute infrastructure for training & inferencing, costing $10 M+ in just cloud costs

Accelerated Training With NeMo

Data Curation Tool

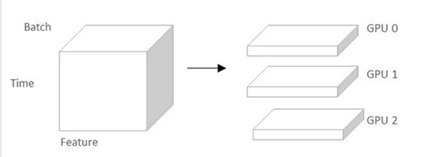


Accelerated Training



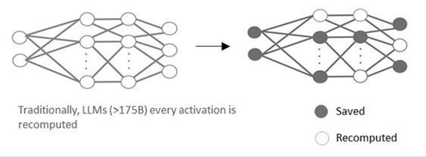
Tensor & Pipeline Parallelism

Accelerated Training



Sequence Parallelism

Accelerated Training



Selective Activation Re-computation

Customization Techniques for Generative AI

Making models useful through SOTA techniques on NeMo

Challenges of Using Foundation models

	Don’t contain domain / enterprise specific knowledge
	Frozen in Time
	Hallucinate and provide undesired information
	Bias & Toxic Information

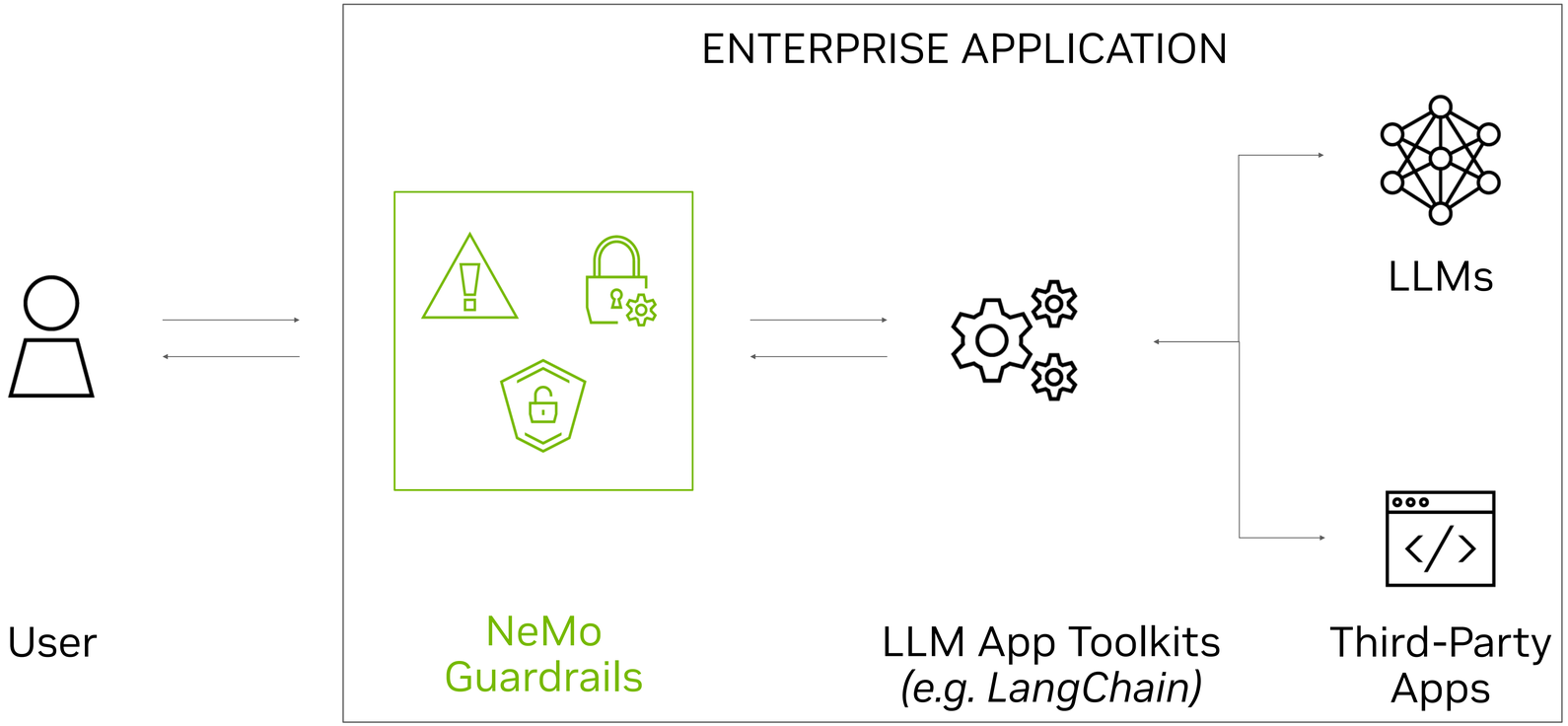
Customization Techniques with NeMo

Domain Knowledge Supervised Fine Tuning	**Incremental Knowledge** Prompt Learning *(p-tuning, prompt tuning, ALiBi, Adapters)*
Continuous Knowledge Reinforcement Learning from Human Feedback	**Runtime Knowledge** Information Retrieval

Building Safe and Secure LLM Applications

NeMo Guardrails for Large Language Models

Add Boundaries Ensure Chatbots Operate According to Use Cases



TOPICAL

Focus interactions within a specific domain

SAFETY

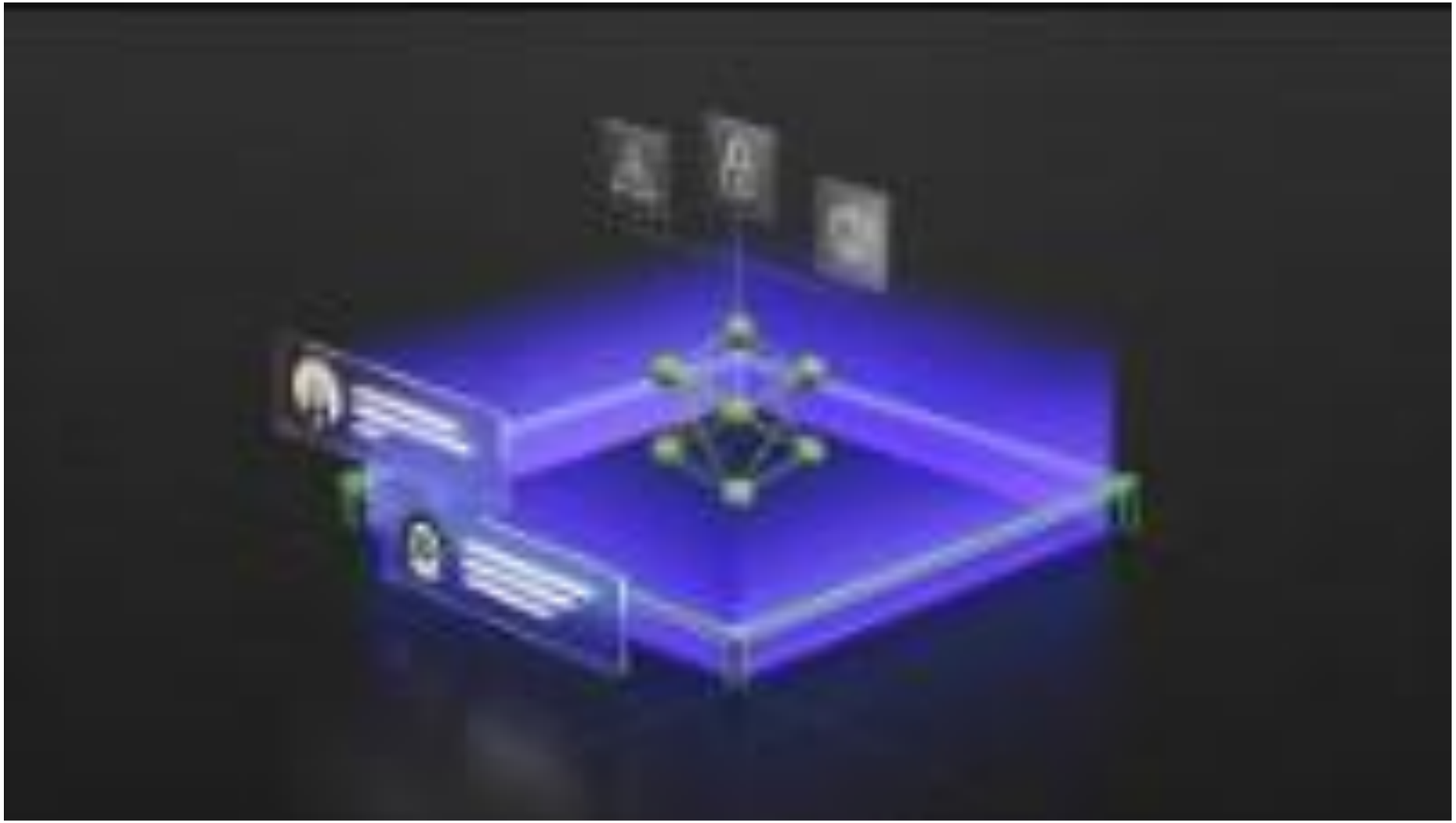
Prevent hallucinations, toxic or misinformative content

SECURITY

Prevent executing malicious calls and handing power to a 3rd party app

NeMo Guardrails for Large Language Models

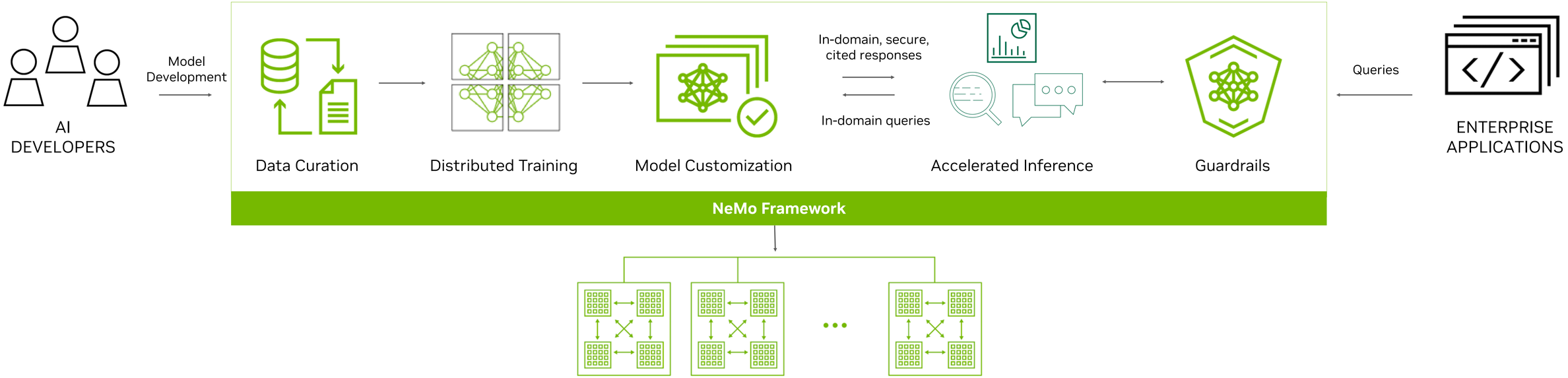
Add Boundaries Ensure Chatbots Operate According to Use Cases



Building Safe and Secure LLM Applications Using NVIDIA NeMo Guardrails

NeMo Framework

End-to-end, cloud-native framework to build, customize and deploy generative AI models



Multi-modality support

Build language, image, generative AI models

Data Curation @ Scale

Extract, deduplicate, filter info from large unstructured data @ scale

Optimized Training

Accelerate training and throughput by parallelizing the model and the training data across 1,000s of nodes.

Model Customization

Easily customize with P-tuning, SFT, Adapters, RLHF, AliBi

Deploy at-scale Anywhere

Run optimized inference at-scale anywhere

Guardrails

Keep applications aligned with safety and security requirements using NeMo Guardrails

Support

NVIDIA AI Enterprise and experts by your side to keep projects on track

Now in open beta, general availability with NVIDIA AI Enterprise in Q2’2023 (LLMs Only)

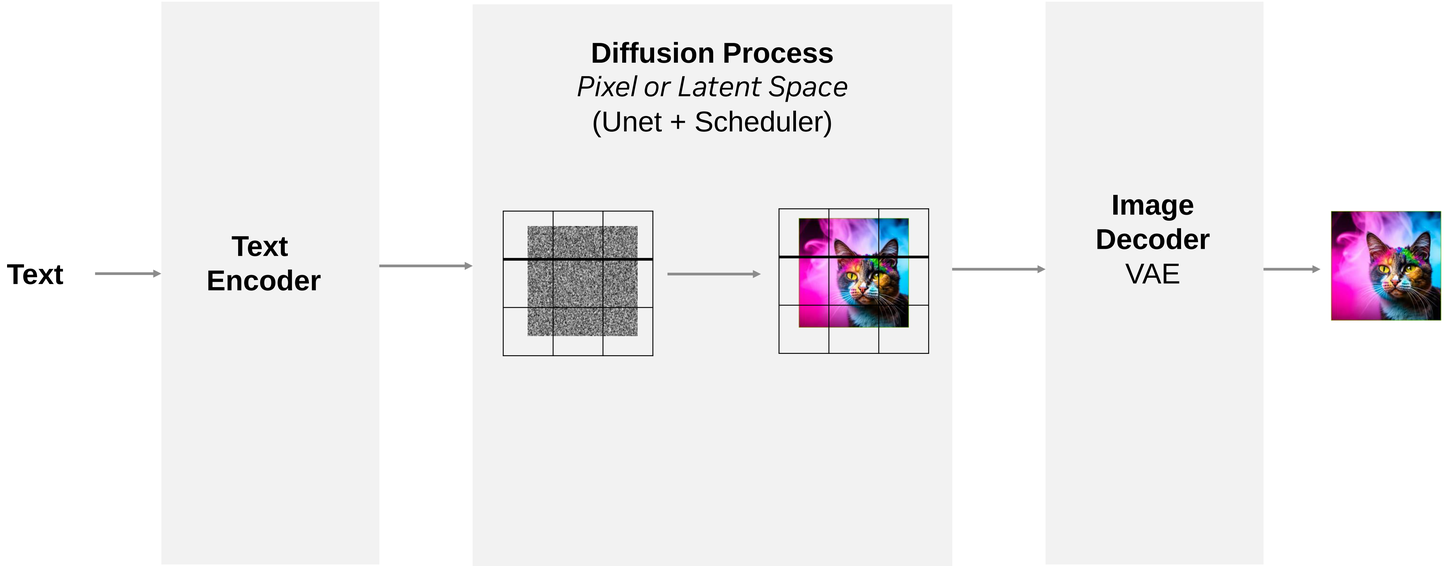
Multi-modal available via early access now

NeMo Expanding Support Across Modalities

NeMo offers multi-modality support

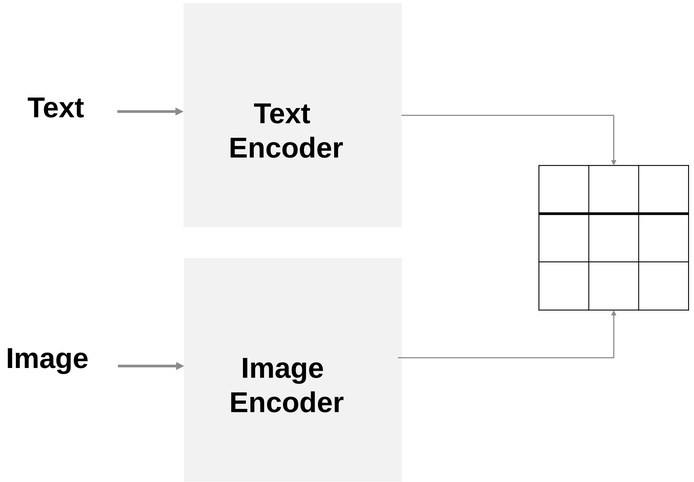
Generative Image Models

Text to Image Generative Models



Discriminative

Suitable for Tasks Like Image Classification, Object Detection



Supported Models In NeMo framework:

Diffusion in Latent Space: Stable Diffusion v1.5

Diffusion in Pixel Space: Imagen

Image-to-Image Models: Instruct-Pix2Pix, DreamBooth

Supported Models In NeMo framework:

Vision-Transformers (ViT)

Multi-Modal: CLIP

Delivering the AI Center of Excellence for Enterprise

Best-of-breed infrastructure for AI development built on NVIDIA DGX

NVIDIA DGX H100

The World's Proven Choice for Enterprise AI

8x NVIDIA H100 GPUs | 32 PFLOPS FP8 (6X) | 0.5 PFLOPS FP64 (3X)
640 GB HBM3 | 3.6 TB/s (1.5X) BISECTION B/W

4th Generation of the World's Most Successful Platform Purpose-Built for Enterprise AI

DGX SuperPOD WITH DGX H100

32 DGX H100 | 1 EFLOPS AI
QUANTUM-2 IB | 20TB HBM3 | 70 TB/s BISECTION B/W (11X)

1 ExaFLOPS of AI Performance in 32 Nodes
Scale as Large as Needed in 32 Node Increments

X-Factors compare performance over DGX SuperPOD with DGX A100 supercomputer configuration with same number of nodes

Thank you!

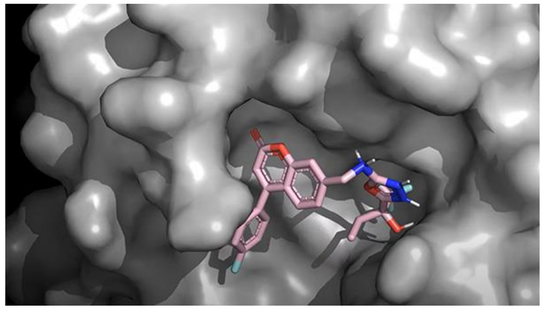
A photo of a cute cat with lots of Holi colors

NVIDIA's Generative AI Solutions

Foundations to Create and Run Custom Generative AI Models

NeMo Service

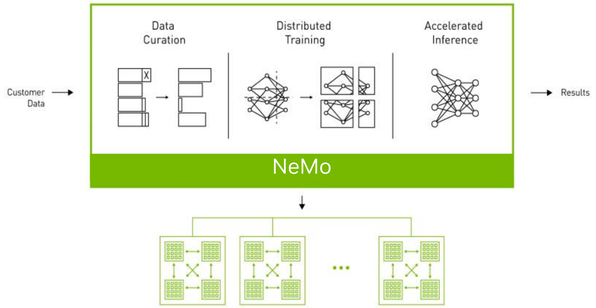
BioNeMo Service



Picasso Service

NVIDIA AI Foundations

NeMo Framework



NVIDIA AI Enterprise

NVIDIA DGX Cloud

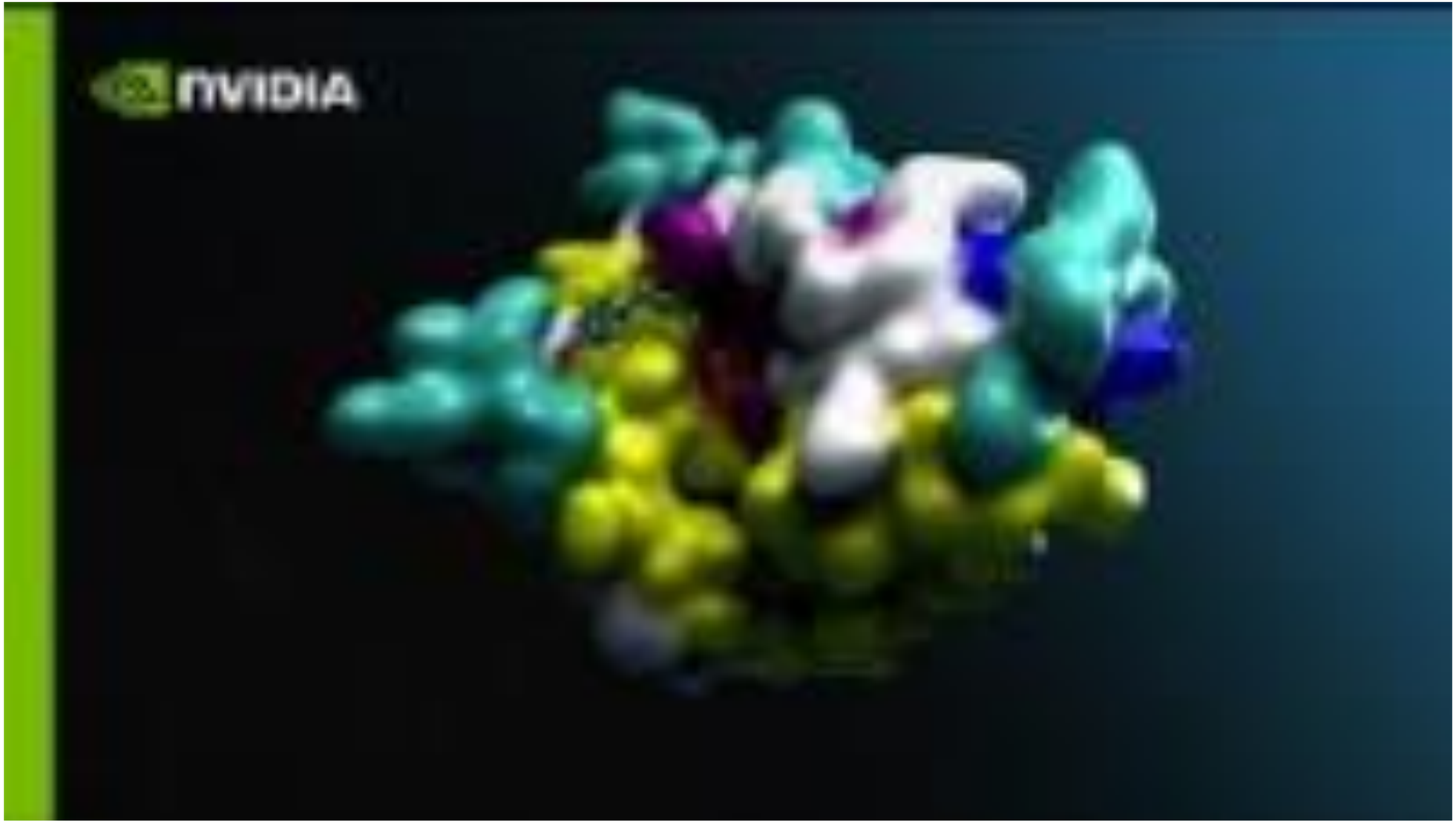
Microsoft Azure | Google Cloud | ORACLE CLOUD

NVIDIA NeMo Service

Enterprise Hyper-Personalization and At-Scale Deployment of Intelligent Large Language Models

NVIDIA NeMo Service | Boosting Enterprise Productivity with Customized Generative AI Models

Accelerate AI-Powered Drug Discovery With BioNeMo



Accelerate AI-Powered Drug Discovery With NVIDIA BioNeMo

NVIDIA Picasso

NVIDIA Picasso, a Cloud-Based Generative AI Service for Creating Images, Videos, and 3D Applications