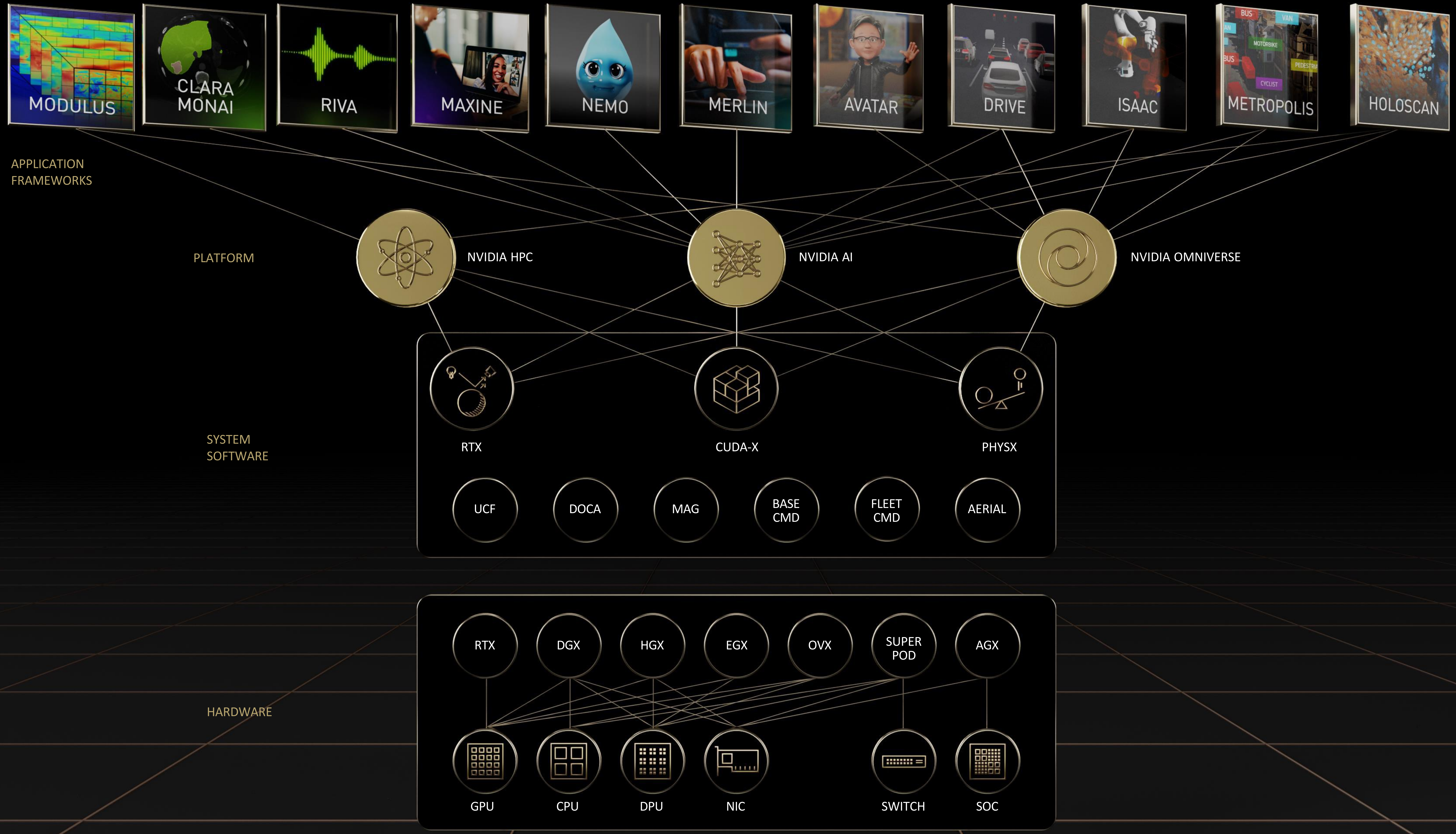
Demystifying Enterprise Generative AI Through Sovereign Cloud

Anissh Pandey | NVIDIA Asia Pacific.



APPLICATION FRAMEWORKS

MODULUS · CLARA MONAI · RIVA · MAXINE · NEMO · MERLIN · AVATAR · DRIVE · ISAAC · METROPOLIS · HOLOSCAN

PLATFORM

NVIDIA HPC · NVIDIA AI · NVIDIA OMNIVERSE

SYSTEM SOFTWARE

RTX · CUDA-X · PHYSX

UCF · DOCA · MAG · BASE CMD · FLEET CMD · AERIAL

HARDWARE

RTX · DGX · HGX · EGX · OVX · SUPER POD · AGX

GPU · CPU · DPU · NIC · SWITCH · SOC

NVIDIA's Generative AI Journey

Generative AI is Transforming Business

TEXT GENERATION

Summarization

GPT-3

Marketing Copy

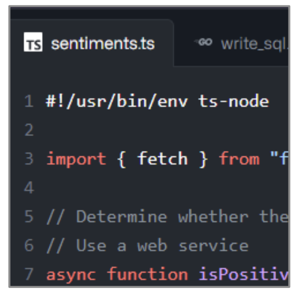
TRANSLATION

Translating Wikipedia

NLLB-200

Real-Time Translation

CODING



```
#!/usr/bin/env ts-node

import { fetch } from "f

// Determine whether the
// Use a web service
async function isPositiv
```

Dynamic Code Commenting

CODEX

Function Generation

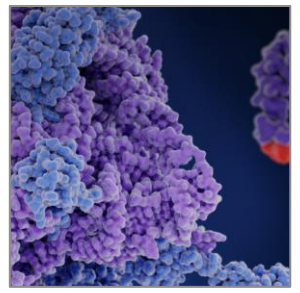
VISUAL CONTENT

Brand Creation

e-Diffi

Gaming Characters

LIFE SCIENCE



Molecular Representations

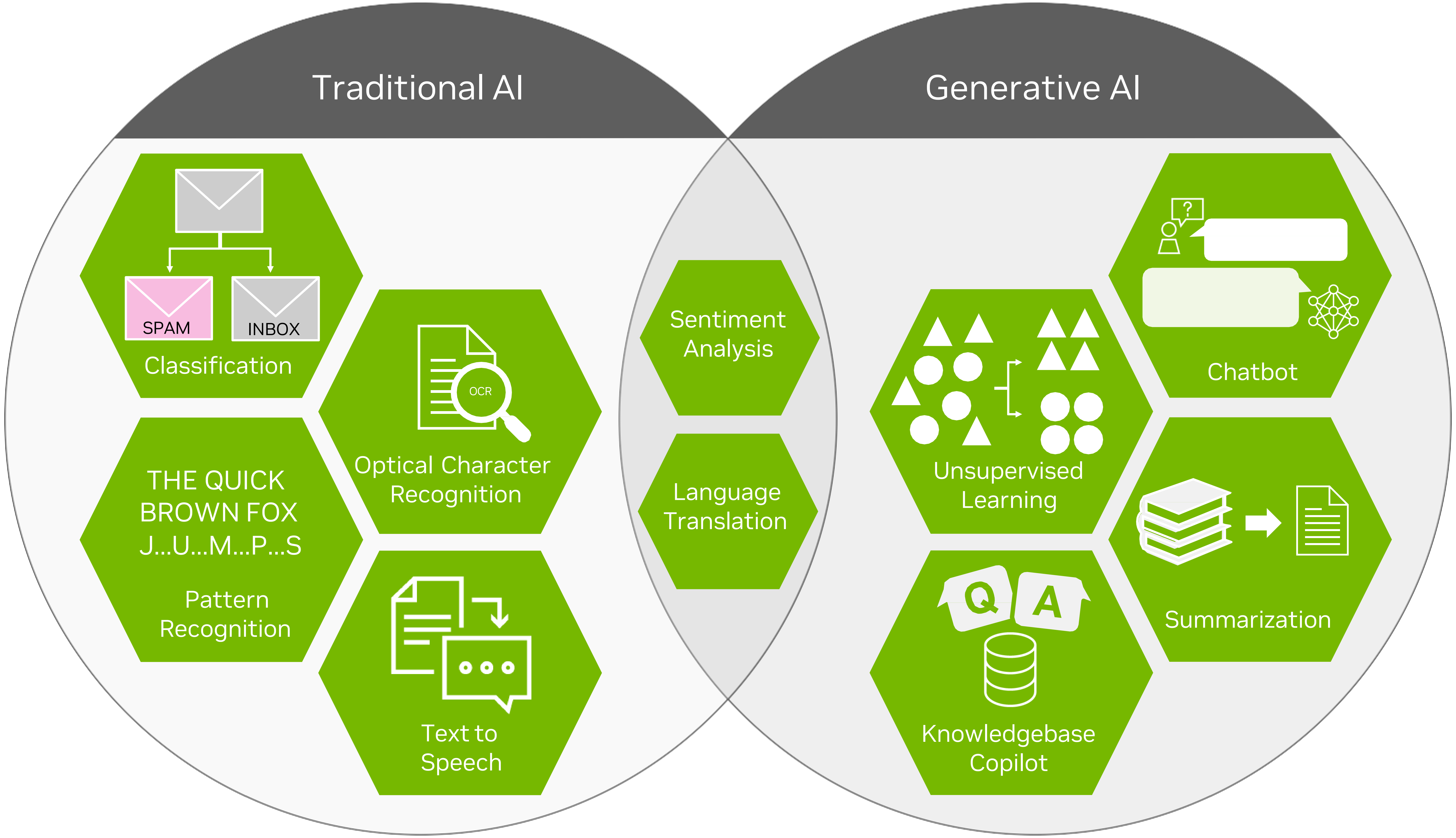
MegaMolBART

Drug Discovery

Enterprises that adopt next-generation AI like LLMs and Generative AI are **2.6X more likely to increase revenue by 10% or more** but must invest in their AI infrastructure to fully reap the benefits.

-Accenture Research. Breakthrough Innovation: Is your organization equipped for breakthrough innovation? WEF 2023.

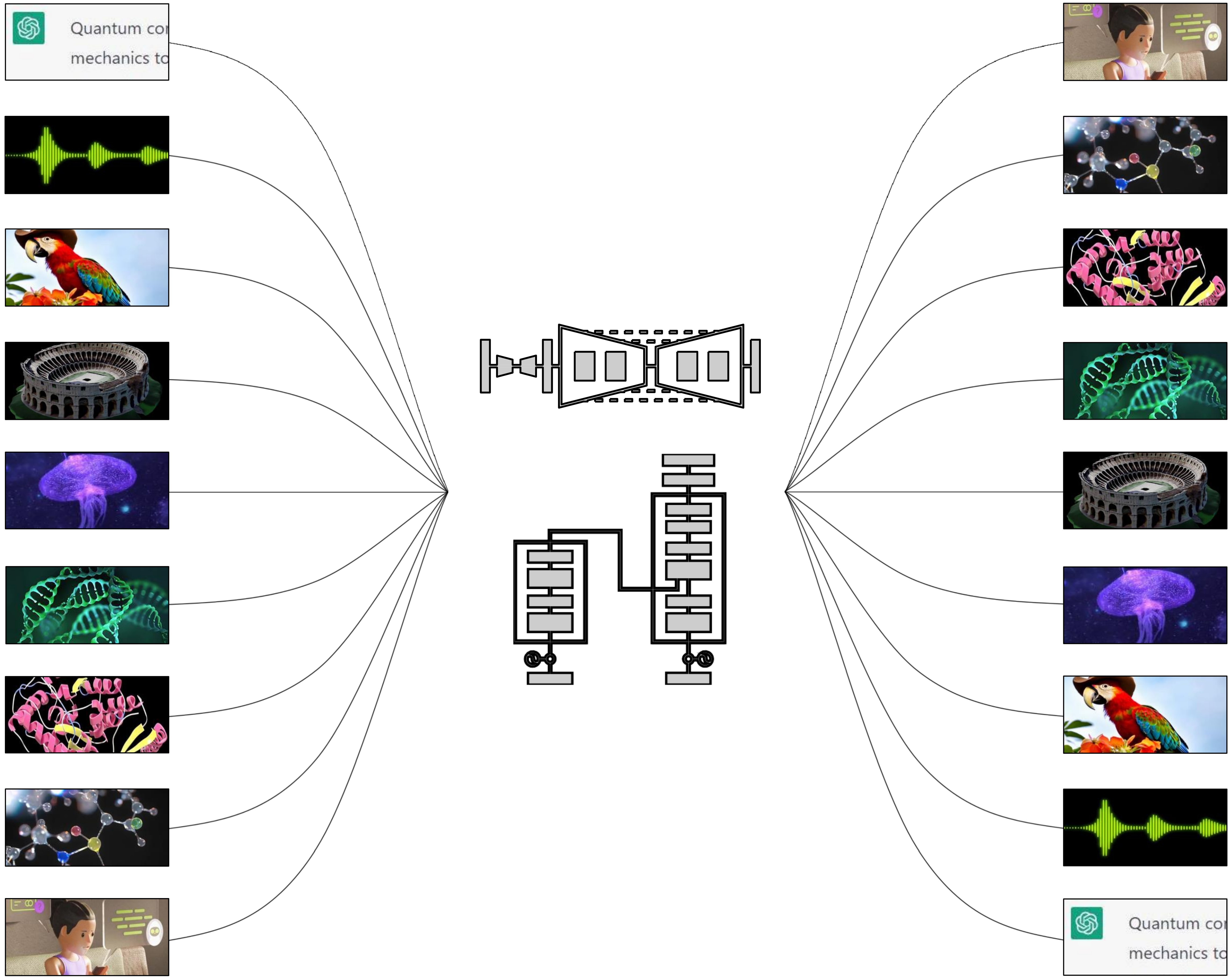
When to Use Generative AI to Solve Enterprise Challenges



Traditional AI focuses on understanding historical data and making accurate predictions

Generative AI creates new data based on patterns and trends learned from training data

What is Generative AI?



TEXT

AUDIO

IMAGE

3D

VIDEO

DNA

PROTEIN

MOLECULE

ANIMATION

ANIMATION

MOLECULE

PROTEIN

DNA

VIDEO

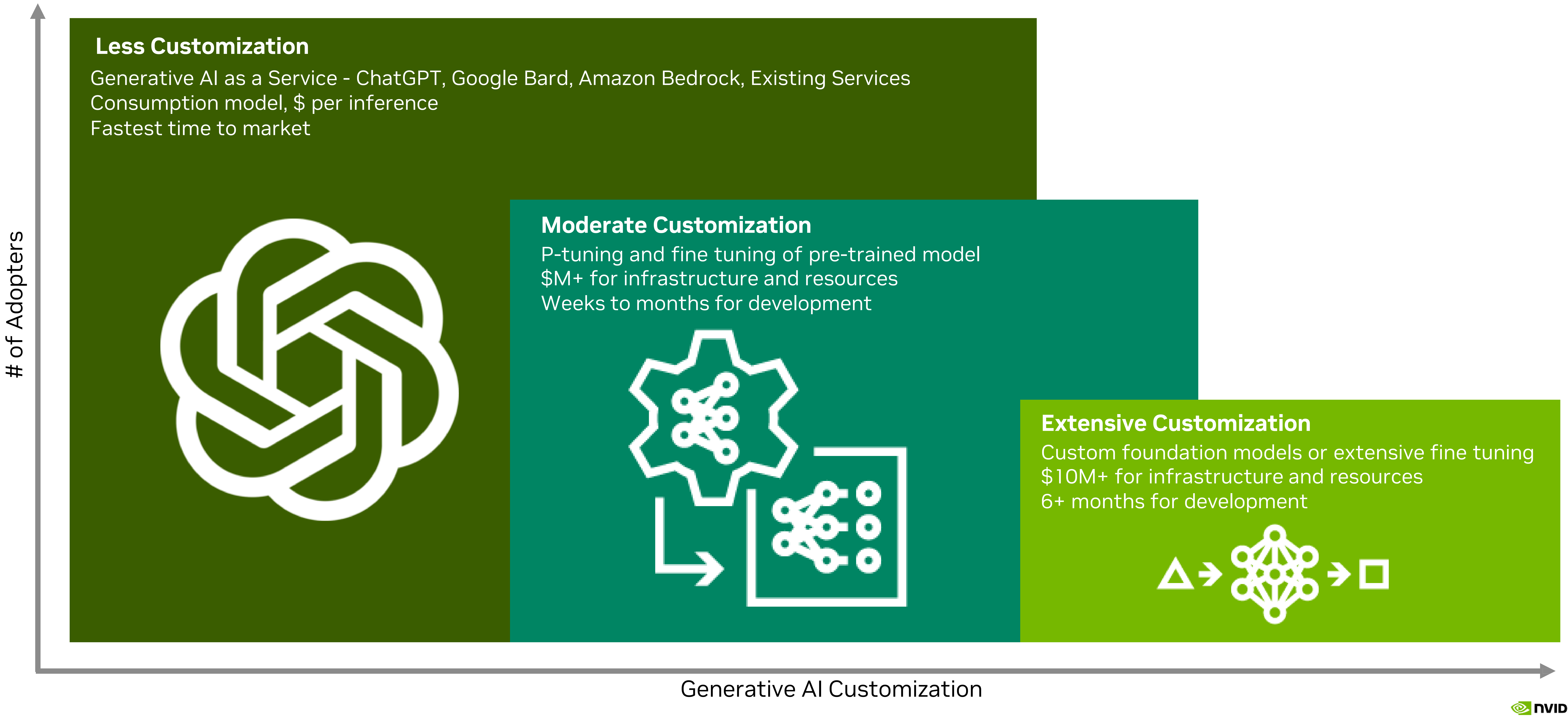
3D

IMAGE

AUDIO

TEXT

How Enterprises are Using Generative AI



Less Customization

Generative AI as a Service - ChatGPT, Google Bard, Amazon Bedrock, Existing Services
Consumption model, $ per inference
Fastest time to market

Moderate Customization

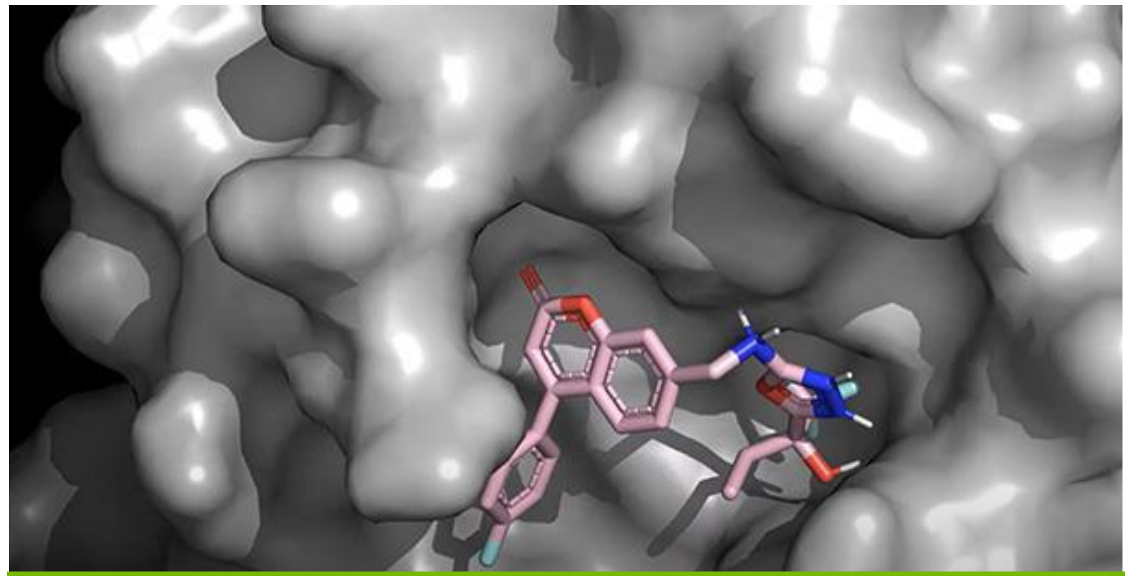
P-tuning and fine tuning of pre-trained model
$M+ for infrastructure and resources
Weeks to months for development

Extensive Customization

Custom foundation models or extensive fine tuning
$10M+ for infrastructure and resources
6+ months for development

NVIDIA Generative AI Platform

NeMo
Language & Multi Modal



BioNeMo
Life Sciences

A photo of a golden retriever puppy wearing a green shirt. The shirt has text “NVIDIA rocks”. Background office. 4k DSLR.

Picasso
Visual Content

NVIDIA AI Enterprise

DGX & DGX Cloud

aws Google Cloud Microsoft Azure ORACLE Cloud Infrastructure

Cloud

DELL Technologies Hewlett Packard Enterprise Lenovo

On-Premises

Accelerated Compute Infrastructure

NVIDIA Approach

- Meet us at Infrastructure, or meet us at the Platform
- Our platform is about: Customization & Freedom

Taking First Steps Now

Steps to Get Started with Generative AI

Leveraging custom LLMs to differentiate your business

Identify Business Opportunity

Target use cases that have meaningful business impact and can be customized with unique data.

Build Out Domain and AI Teams

Identify internal resources and augment them with AI expertise from partners and application providers.

Analyze Data for Training/Customization

Acquire, refine, and safeguard data to build either data-intensive foundation models or customize existing models.

Invest in Accelerated Infrastructure

Assess infrastructure, architecture, and operating model, while considering costs and energy consumption.

Develop Plan for Responsible AI

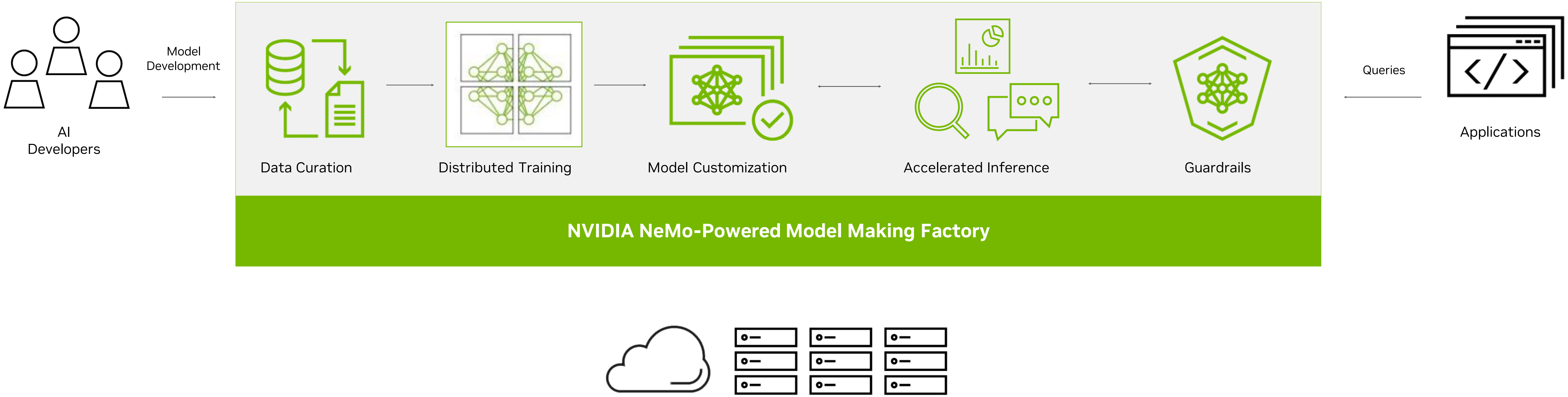
Leverage tools and best practices to ensure responsible AI principles are adopted across the company.

Requirements for Building Custom LLMs

- Training Data
- Accelerated Computing
- Training and Inference Tools
- AI Expertise

NVIDIA NeMo

Factory for building custom large language models



NeMo Generative Foundation Models

Suite of Pre-Trained Large Language Models built for Enterprise Hyper-Personalization

Fastest Responses

GPT-8

8B w/ 1.1T tokens. SFT w/ FLAN. I/O: 4K tokens

Optimal balance of accuracy - latency

GPT-43

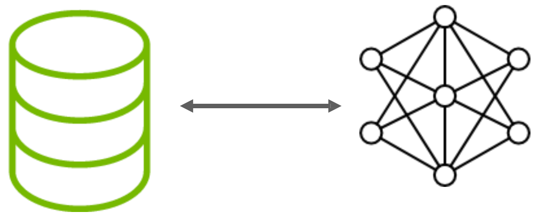
43B w/ 1.1T tokens. SFT w/FLAN. 50 Languages. I/O: 4K tokens

For Complex Tasks

GPT-530

530B w/ 340B tokens. SFT w/FLAN. I/O: 2K tokens

Answers generated from Retrieved models



Inform

Community-built model

BLOOMZ-T0

BLOOMZ-T0-13B w/ 340B tokens. 101 Languages. I/O: 2K tokens. Encoder-only - T5 model .

Customization Techniques for Generative AI

Making models useful for specific use-cases through state-of-the-art techniques on NeMo

Requirements for Custom Enterprise Generative AI Models

- Domain / enterprise specific knowledge
- Up-to-date & factual information
- Protection from bias & toxic information

Customization Techniques with NeMo

Add Domain Knowledge Supervised Fine Tuning	**Add Skills - Incremental Knowledge** Prompt Learning *(p-tuning, Prompt Tuning, ALiBi, Adapters, LoRA)*
Continuous Refinement Reinforcement Learning from Human Feedback	**Retrieve Factual Knowledge At Runtime** Information Retrieval

NVIDIA AI Nations Next Framework

Full-Stack Collaboration Approach

Sovereignty

Sustainability

Safety

National AI Program

AI Initiatives

NVIDIA helps nations advance AI R&D workloads and applied use-cases across every industry

AI Workforce

NVIDIA helps nations upskill local talent and develop the AI-ready workforce

AI Ecosystems

NVIDIA helps nations strengthen their local AI ecosystem and learn from global leaders

NVIDIA AI Enterprise

CLARA	RIVA	TOKKIO	MERLIN	MODULUS	MAXINE	METROPOLIS	CUOPT	NEMO	ISAAC	DRIVE	MORPHEUS
Medical Imaging	Speech AI	Customer Service	Recommenders	Physics ML	Video	Video Analytics	Logistics	Conversational AI	Robotics	Autonomous Vehicles	Cybersecurity

Inception for Start-ups, Deep Learning Institute, Hackathons/Bootcamps, GTC

NV PS, Industry/Domain Experts, HPC + Technical Support, AI Tech Center Research Collaborations

NVIDIA Accelerated Computing Infrastructure through Viettel AI Factory

Cloud

Data Center

Edge

Embedded

NVIDIA
NCP PaaS
Platform
RA+ NVAIE

Hands-on
Labs

NVIDIA®