Generative AI

ICN Solutions BIM Congres – 24 September 2024

Jean-Pierre van Gastel – HP – NVIDIA tech evangelist

AI Adoption is Growing and is Business Critical

Across the hybrid cloud

Increasing AI Adoption

75% of large enterprises will use AI to enhance efficiency and improve quality[1]

56% average adoption rate of AI by organizations globally

Struggle with Complexity

7.3 months from pilot to production[2]

31% have AI deployed in production

Growing Adoption of Cloud

90% of enterprise cloud infrastructure will be based on public cloud providers

50% of all accelerated infrastructure for performance-intense computing will be cloud based

[1] IDC. "IDC FutureScape, Worldwide AI and Automation 2023 Predictions", 2022; [2] Gartner, "2023 Planning Guide for Analytics and AI", 2023; [3] Gartner, "Forecast Analysis: Cloud Infrastructure and Platform Services, Worldwide", 2021

Terminology Explained

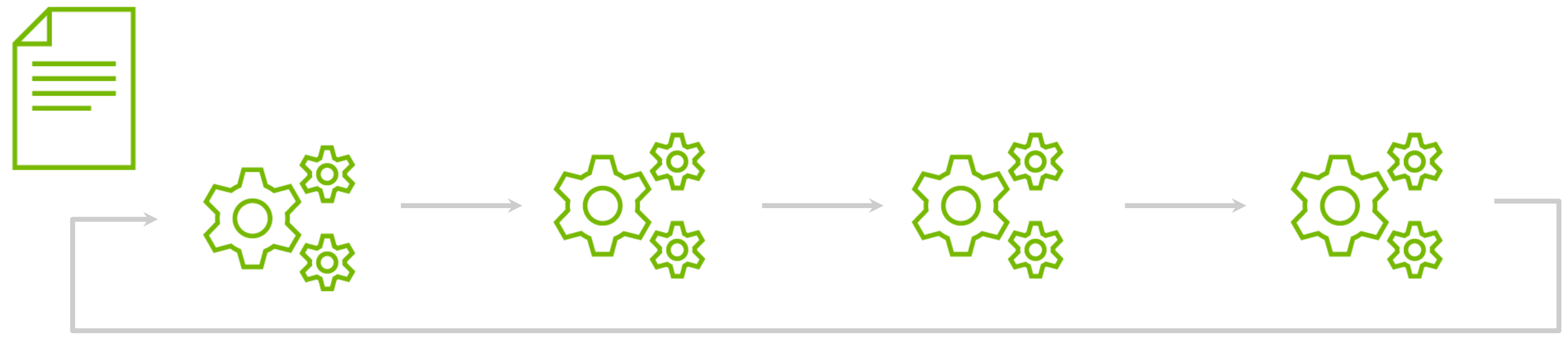
Workload vs. workflow

Workload

Any software program or application, that is standalone or part of a workflow, that uses compute resources to accomplish a task.

Data science, AI, and 3D graphics workloads can be accelerated by libraries and frameworks that leverage NVIDIA GPUs.

Examples: Spark jobs, models doing video analytics, training a large language model, a text-to-speech function, video rendering



Workflow

Multi-step process to get from initiation to completion, where each step is a unique workload. For example, the generic workflow of AI is data prep > training > simulation > inference.

NVIDIA has assembled, tested, and documented reference workflows that can be customized by partners and customers to give them a head start at solving their specific challenge.

Examples: Audio Transcription, Digital Fingerprinting to Detect Cybersecurity Threats, Contact Center Intelligent Virtual Assistants

Workflow Examples

Making Cookies

Chocolate Chip Cookies Recipe

FLOUR
SUGAR

Gather and Prepare Ingredients

Combine Ingredients in Mixer

Scoop Dough onto Baking Sheet

Bake Cookies in the Oven

Cool Cookies on Baking Rack

Intelligent Virtual Assistant

Intelligent Virtual Assistant Documentation

Customer Explains Their Problems

Audio Conversation is Transcribed to Text

NVIDIA-Accelerated With Riva ASR

NLP Service Interprets Problems and Generates Output

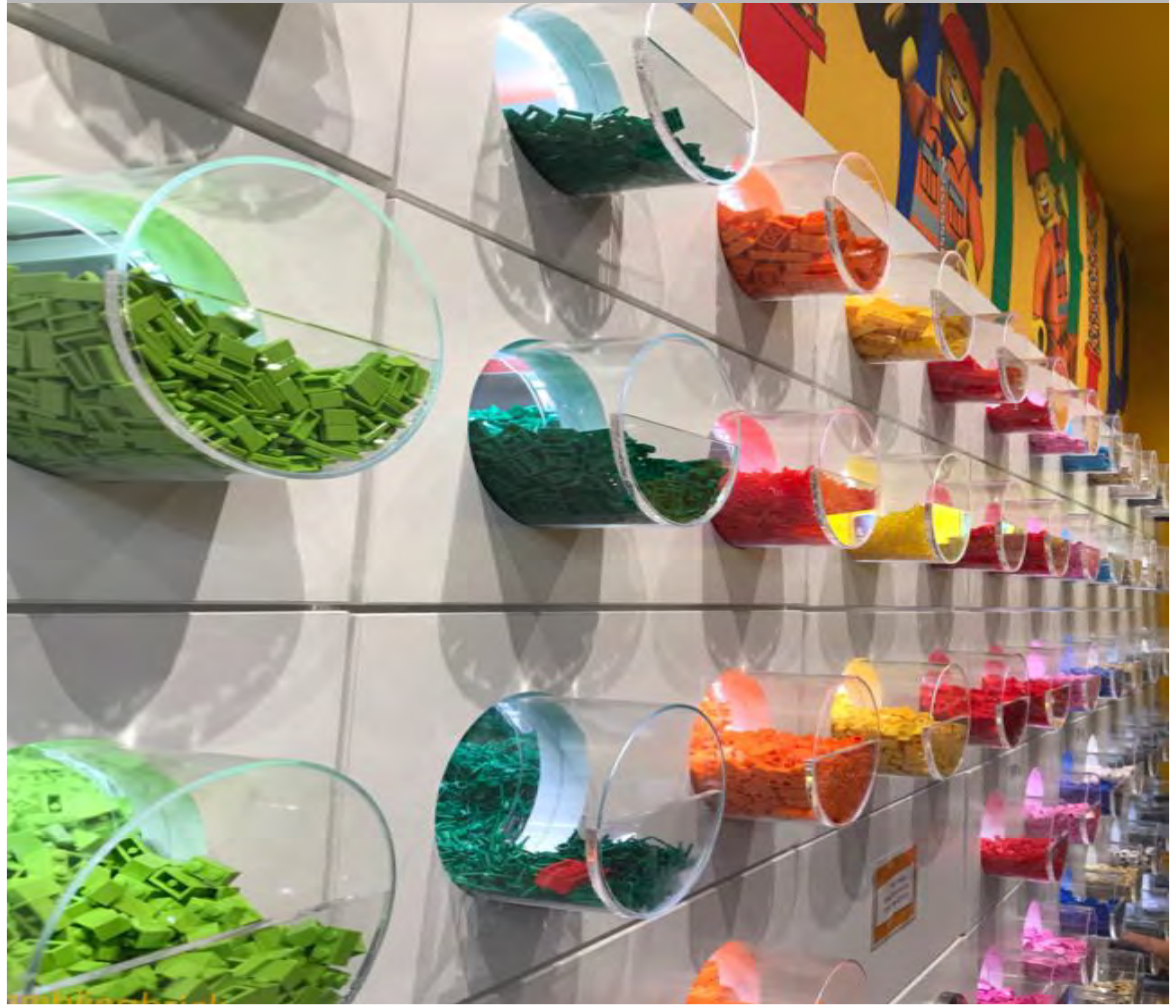
NVIDIA-Accelerated NLP Service

Text is Converted to Audio that Customer Hears

NVIDIA-Accelerated With Riva TTS

Two Ways to Build with Legos

Pick a Brick



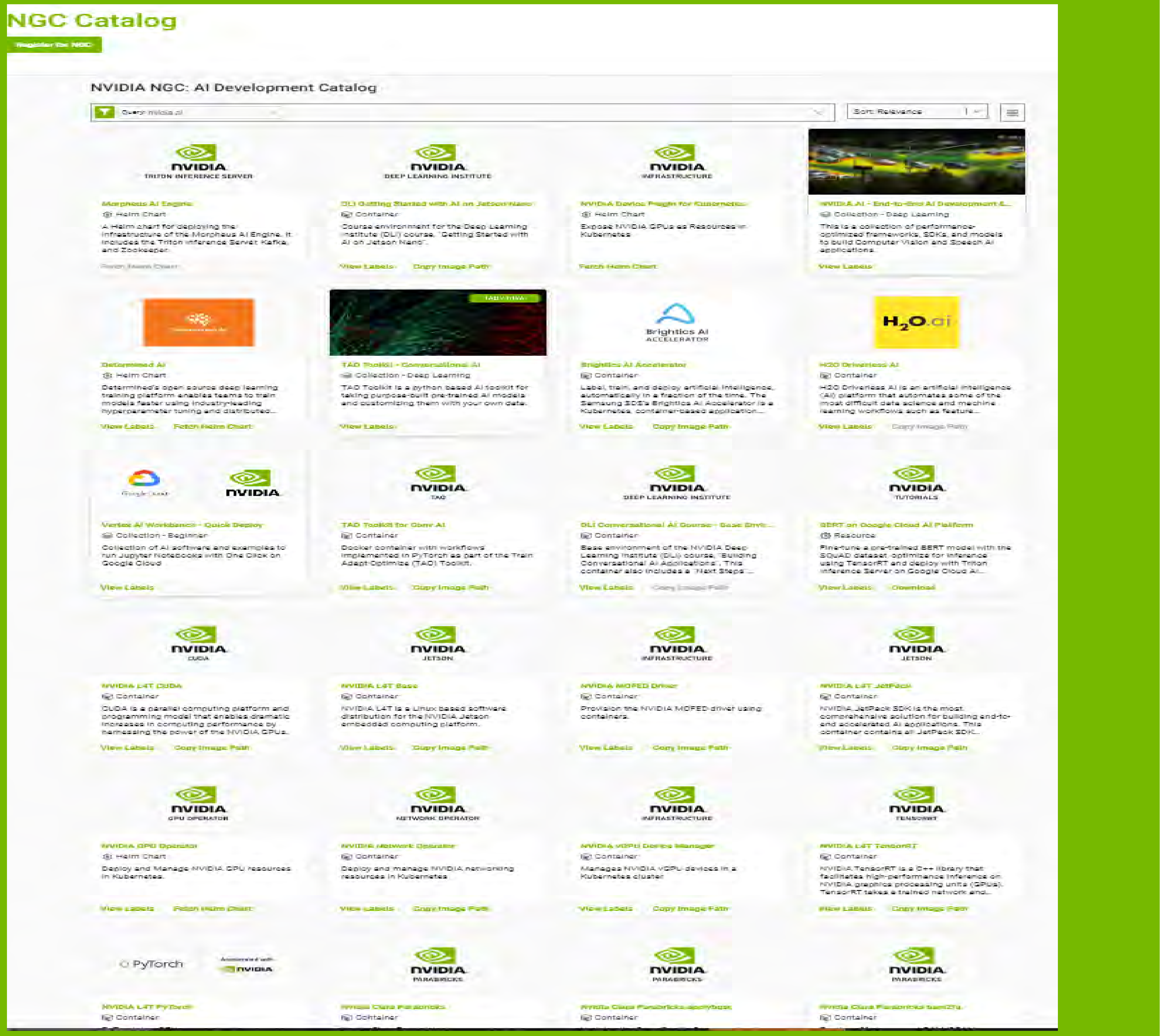
DIY, unlimited options

Lego Kit

Components you need to build the AI solution

Two Ways to Build with NVIDIA AI

Public Catalog on NGC



NGC Catalog

NVIDIA NGC: AI Development Catalog

Search nvidia.ai | Sort: Relevance

Morpheus AI Engine
Helm Chart
A Helm chart for deploying the infrastructure of the Morpheus AI Engine. It includes the Triton Inference Server, Kafka, and Zookeeper.
Fetch Helm Chart

DLI Getting Started with AI on Jetson Nano
Container
Course environment for the Deep Learning Institute (DLI) course, "Getting Started with AI on Jetson Nano".
View Labels | Copy Image Path

NVIDIA Device Plugin for Kubernetes
Helm Chart
Expose NVIDIA GPUs as Resources in Kubernetes
Fetch Helm Chart

NVIDIA AI - End-to-End AI Development &...
Collection - Deep Learning
This is a collection of performance-optimized frameworks, SDKs, and models to build Computer Vision and Speech AI applications.
View Labels

Determined AI
Helm Chart
Determined's open source deep learning training platform enables teams to train models faster using industry-leading hyperparameter tuning and distributed...
View Labels | Fetch Helm Chart

TAO Toolkit - Conversational AI
Collection - Deep Learning
TAO Toolkit is a python based AI toolkit for taking purpose-built pre-trained AI models and customizing them with your own data.
View Labels

Brightics AI Accelerator
Container
Label, train, and deploy artificial intelligence, automatically in a fraction of the time. The Samsung SDS's Brightics AI Accelerator is a Kubernetes, container-based application...
View Labels | Copy Image Path

H2O Driverless AI
Container
H2O Driverless AI is an artificial intelligence (AI) platform that automates some of the most difficult data science and machine learning workflows such as feature...
View Labels | Copy Image Path

Vertex AI Workbench - Quick Deploy
Collection - Beginner
Collection of AI software and examples to run Jupyter Notebooks with One Click on Google Cloud
View Labels

TAO Toolkit for Conv AI
Container
Docker container with workflows implemented in PyTorch as part of the Train Adapt Optimize (TAO) Toolkit.
View Labels | Copy Image Path

DLI Conversational AI Course - Base Envir...
Container
Base environment of the NVIDIA Deep Learning Institute (DLI) course, "Building Conversational AI Applications". This container also includes a "Next Steps"...
View Labels | Copy Image Path

BERT on Google Cloud AI Platform
Resource
Fine-tune a pretrained BERT model with the SQuAD dataset, optimize for inference using TensorRT and deploy with Triton Inference Server on Google Cloud AI...
View Labels | Download

NVIDIA L4T CUDA
Container
CUDA is a parallel computing platform and programming model that enables dramatic increases in computing performance by harnessing the power of the NVIDIA GPUs.
View Labels | Copy Image Path

NVIDIA L4T Base
Container
NVIDIA L4T is a Linux based software distribution for the NVIDIA Jetson embedded computing platform.
View Labels | Copy Image Path

NVIDIA MOFED Driver
Container
Provision the NVIDIA MOFED driver using containers.
View Labels | Copy Image Path

NVIDIA L4T JetPack
Container
NVIDIA JetPack SDK is the most comprehensive solution for building end-to-end accelerated AI applications. This container contains all JetPack SDK...
View Labels | Copy Image Path

NVIDIA GPU Operator
Helm Chart
Deploy and Manage NVIDIA GPU resources in Kubernetes.
View Labels | Fetch Helm Chart

NVIDIA Network Operator
Container
Deploy and manage NVIDIA networking resources in Kubernetes
View Labels | Copy Image Path

NVIDIA vGPU Device Manager
Container
Manages NVIDIA vGPU devices in a Kubernetes cluster
View Labels | Copy Image Path

NVIDIA L4T TensorRT
Container
NVIDIA TensorRT is a C++ library that facilitates high-performance inference on NVIDIA graphics processing units (GPUs). TensorRT takes a trained network and...
View Labels | Copy Image Path

NVIDIA L4T PyTorch
Container

Nvidia Clara Parabricks
Container

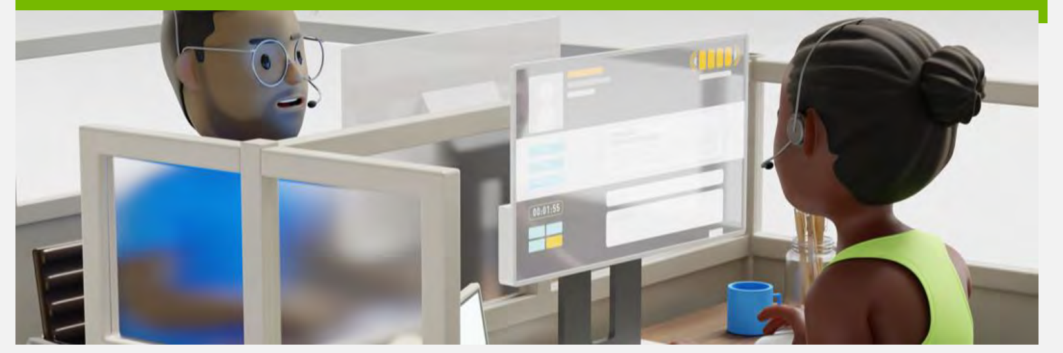
Nvidia Clara Parabricks applybqsr
Container

Nvidia Clara Parabricks bam2fq
Container

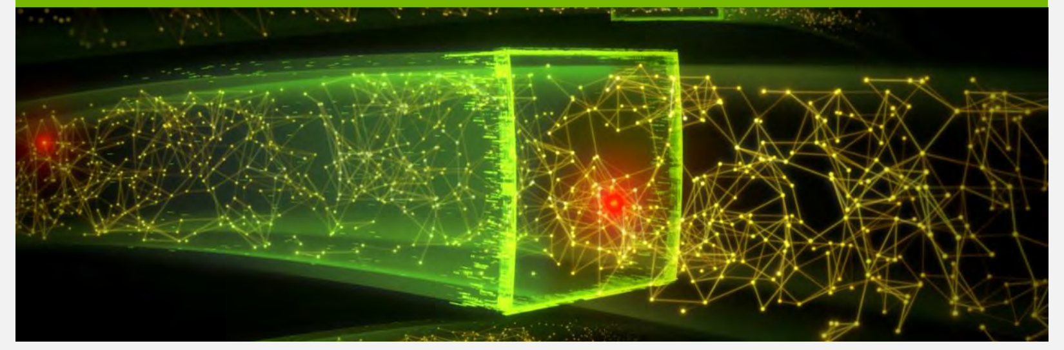
DIY, unlimited options

NVIDIA AI Workflows

Intelligent Virtual Assistant



Digital Fingerprinting Cybersecurity Threat Detection

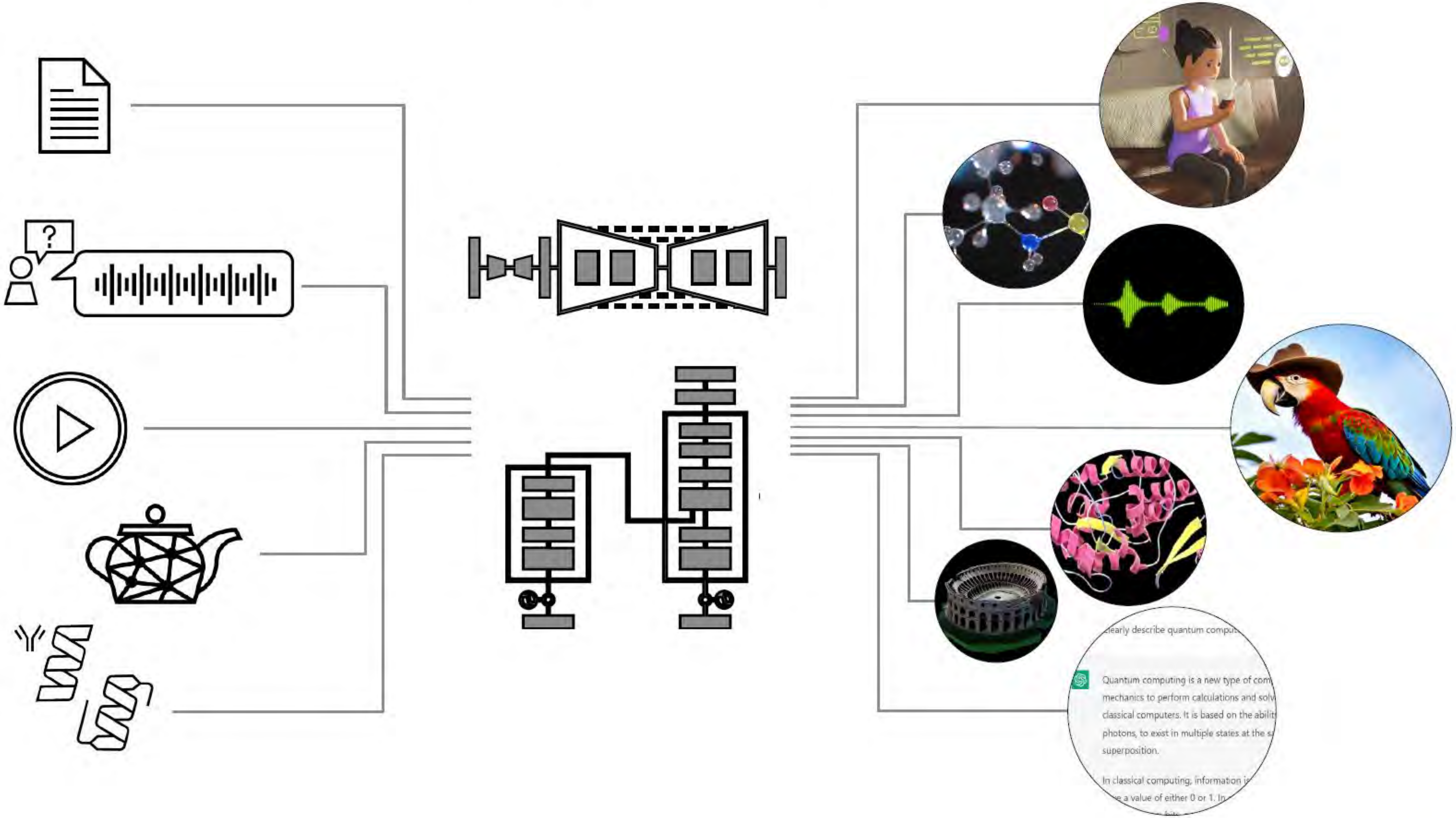


Audio Transcription

Product Recommendations
xxx

Components you need to build the AI application

What is Generative AI?



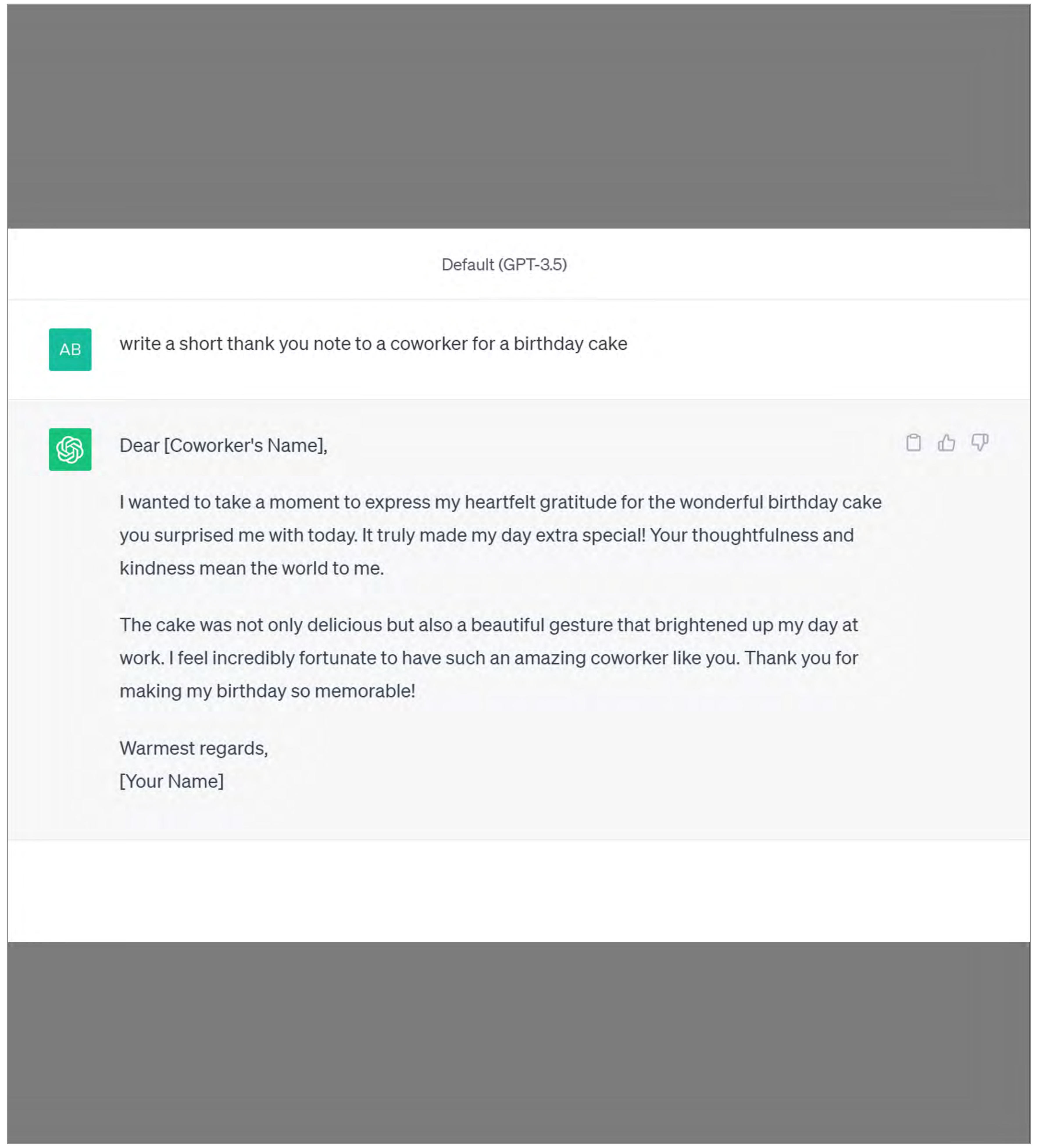
clearly describe quantum compu

Quantum computing is a new type of com
mechanics to perform calculations and solv
classical computers. It is based on the abilit
photons, to exist in multiple states at the s
superposition.

In classical computing, information i
e a value of either 0 or 1. In

Generative AI

From Research To Production In 5 Years



write a short thank you note to a coworker for a birthday cake

Dear [Coworker's Name],

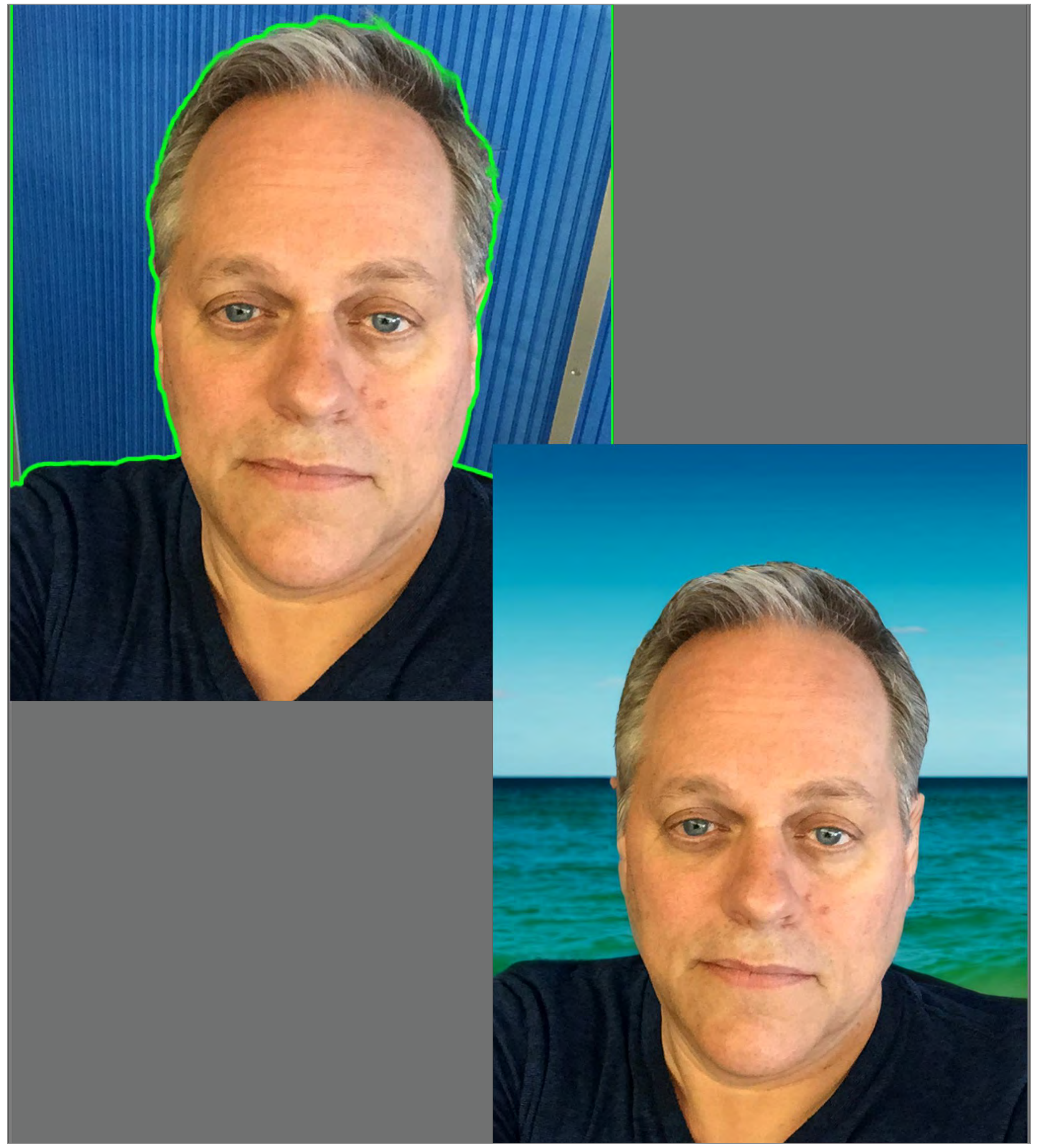
I wanted to take a moment to express my heartfelt gratitude for the wonderful birthday cake you surprised me with today. It truly made my day extra special! Your thoughtfulness and kindness mean the world to me.

The cake was not only delicious but also a beautiful gesture that brightened up my day at work. I feel incredibly fortunate to have such an amazing coworker like you. Thank you for making my birthday so memorable!

Warmest regards,
[Your Name]

ChatGPT, LLAMA2

Stable Diffusion etc



Adobe Photoshop Generative Fill

Generative AI is Transforming Every Industry

3D VFX & Game Design

Generate textures and backgrounds

Architecture & Interior Design

Create floorplans and explore architectural styles

Fashion & Product Design

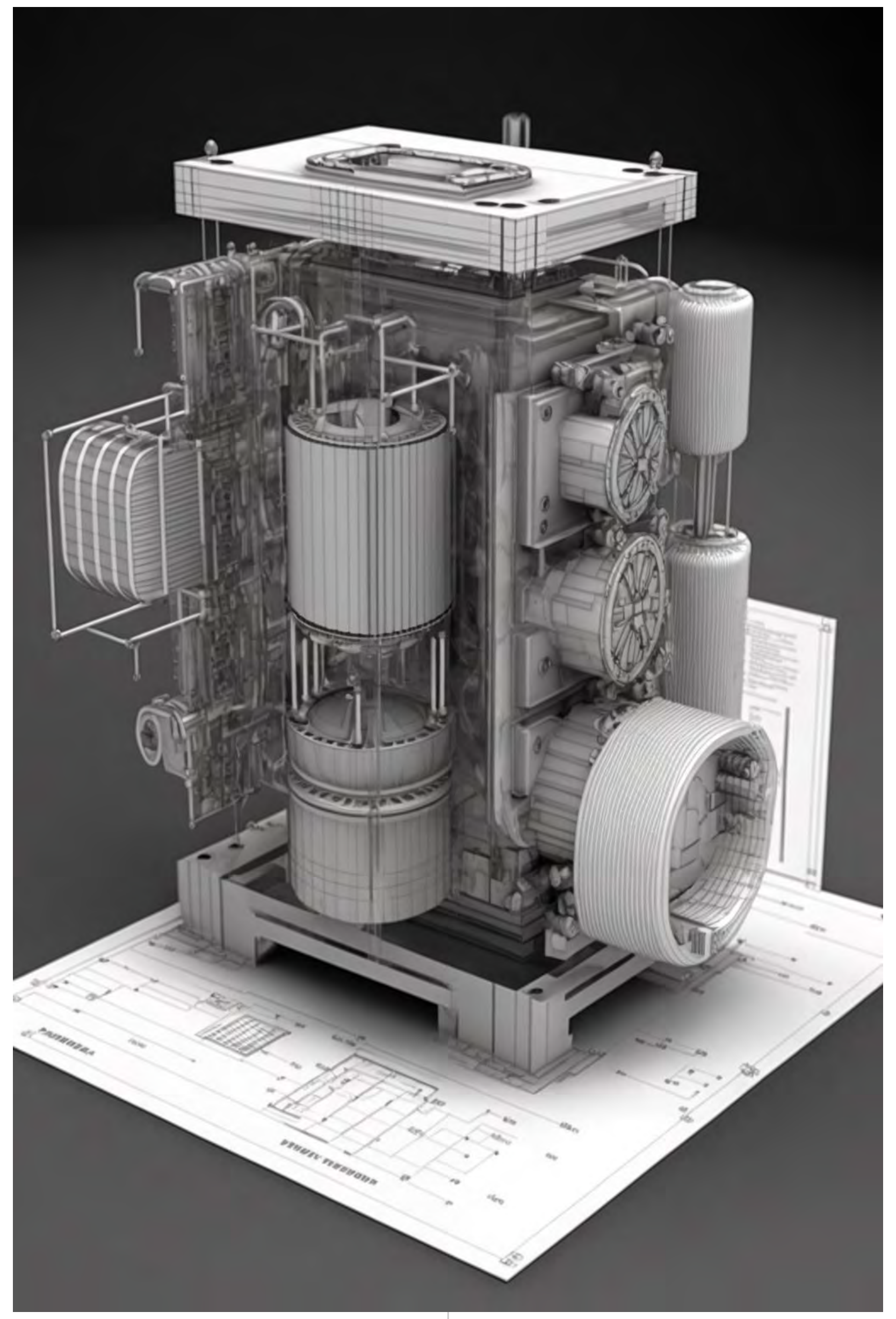
Inspire unique design concepts

Photography & Photo Editing

Background and object replacement

Marketing and Advertising

Create elements & reusable motifs



Manufacturing

Design parts
Explore structures & solutions

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.

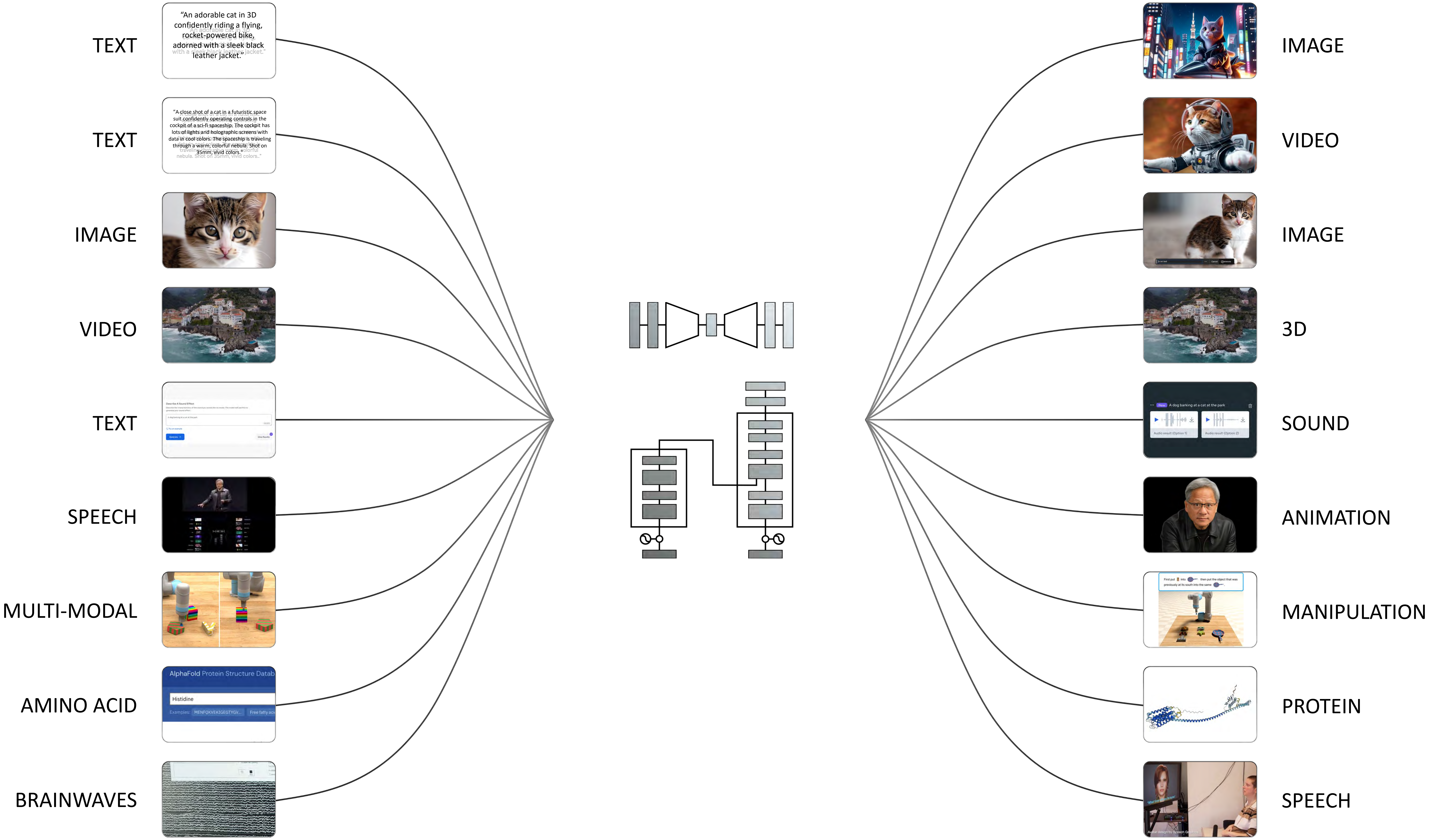
Experimentation

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



TEXT

"An adorable cat in 3D confidently riding a flying, rocket-powered bike, adorned with a sleek black leather jacket."

TEXT

"A close shot of a cat in a futuristic space suit confidently operating controls in the cockpit of a sci-fi spaceship. The cockpit has lots of lights and holographic screens with data in cool colors. The spaceship is traveling through a warm, colorful nebula. Shot on 35mm, vivid colors."

IMAGE

VIDEO

TEXT

SPEECH

MULTI-MODAL

AMINO ACID

AlphaFold Protein Structure Datab

Histidine

Examples: MENFQKVEKIGEGTYGV... Free fatty aci

BRAINWAVES

IMAGE

VIDEO

IMAGE

3D

SOUND

A dog barking at a cat at the park

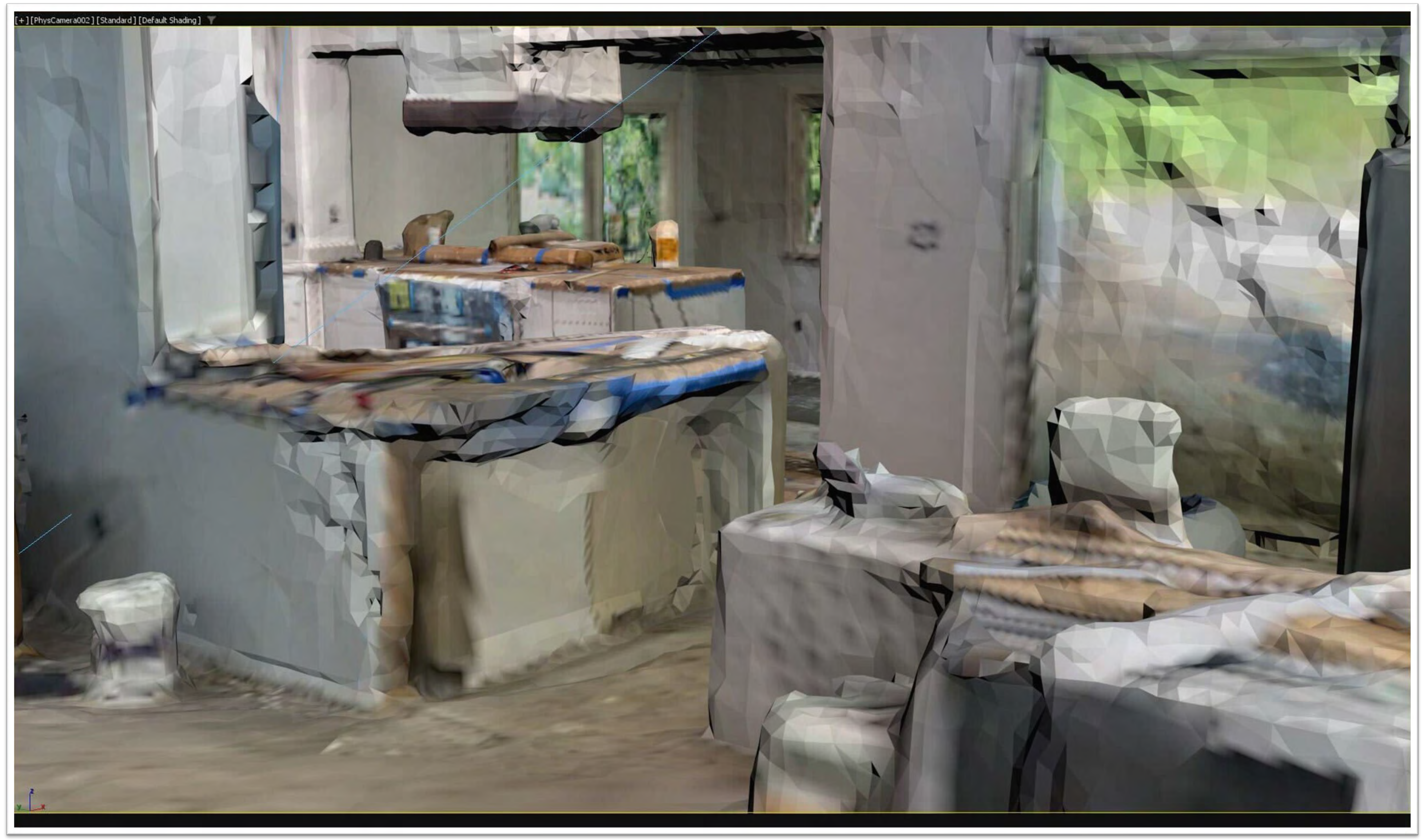
ANIMATION

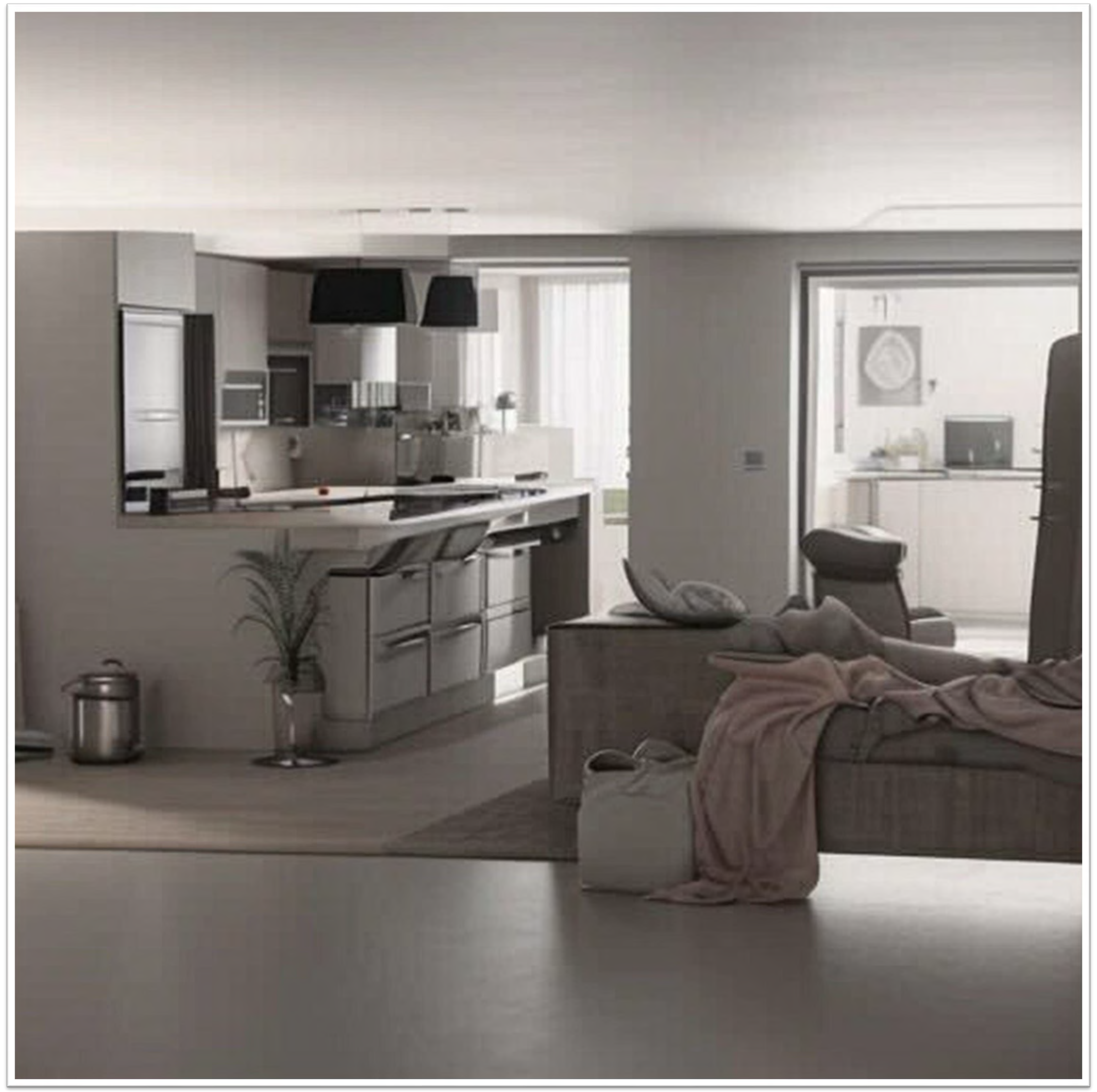
MANIPULATION

PROTEIN

SPEECH

Autodesk 3ds Max – tyFlow demo





Multimodal Generative AI is Transforming Industries

TEXT

Chatbots or Assistants

NEMO

Marketing Copy

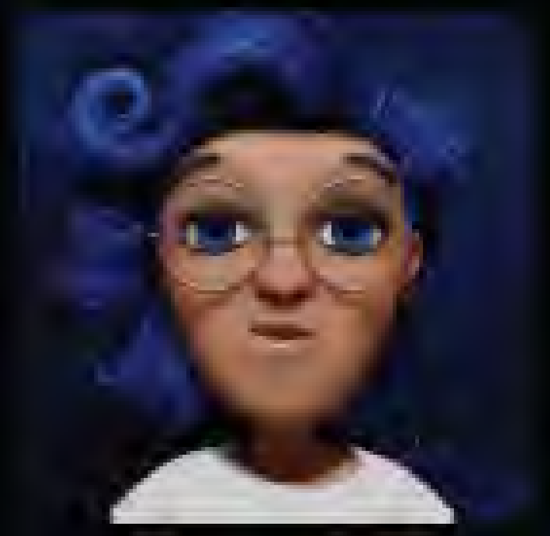
IMAGE

Brand Creation

SDXL

Personalized Content

AUDIO



Text to Voice Generation

RIVA

Sound Creation

VIDEO

Dynamic Code Commenting

PICASSO

Function Generation

3D

Product Design & Discovery

NEURAL-ANGELO

3D Object Generation

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.

How Enterprises are Using Generative AI

Fastest
Fastest Time to Adoption
Slowest
Least Difficult
Level of Difficulty to Develop & Deploy
Most Difficult

Less Customization

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

Moderate Customization

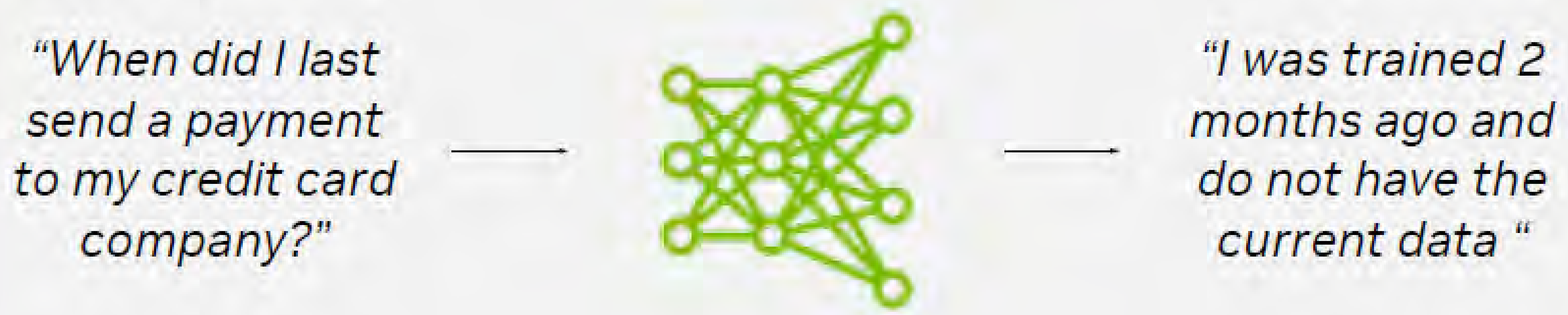
P-tuning and fine tuning of pre-trained model
$M+ for infrastructure and resources
Weeks to months for development

Extensive Customization

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

Enterprise Generative AI Use Cases Require Domain Specific Knowledge

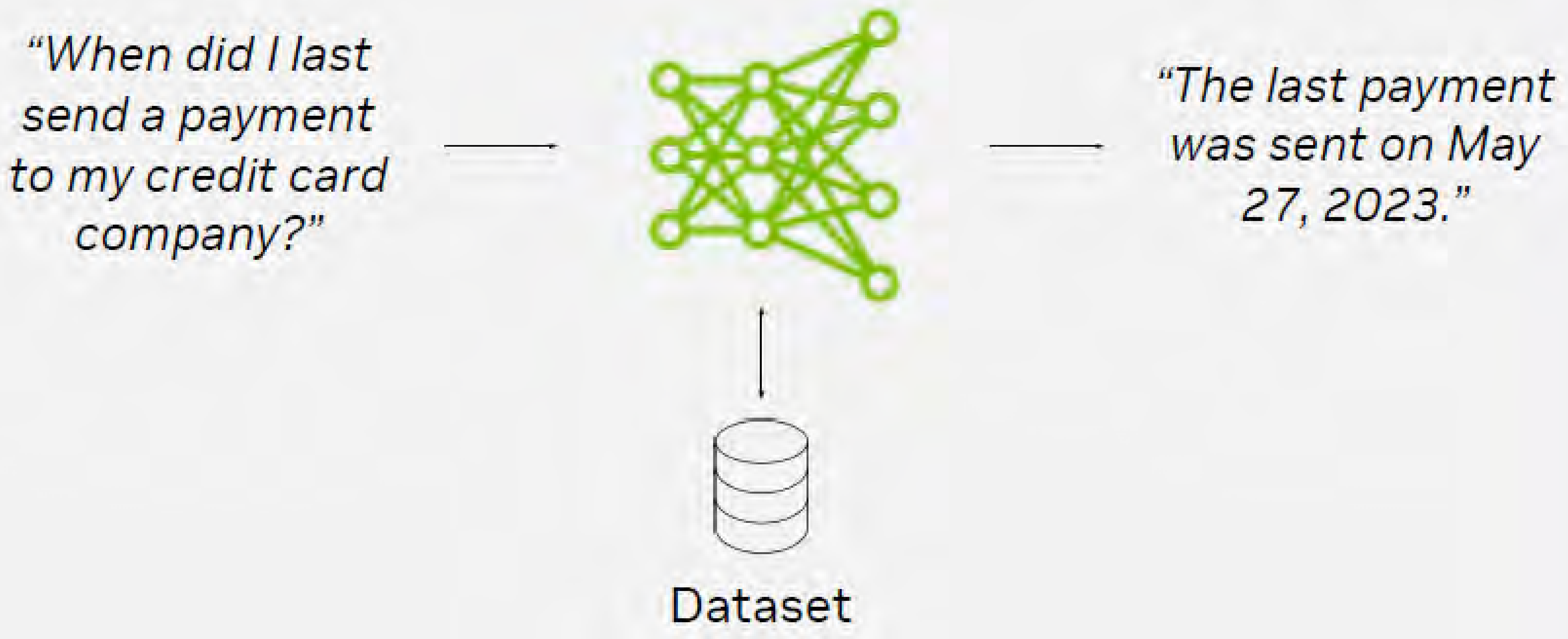
Foundation Model Response



"When did I last send a payment to my credit card company?"

"I was trained 2 months ago and do not have the current data "

Custom Model Response



"When did I last send a payment to my credit card company?"

"The last payment was sent on May 27, 2023."

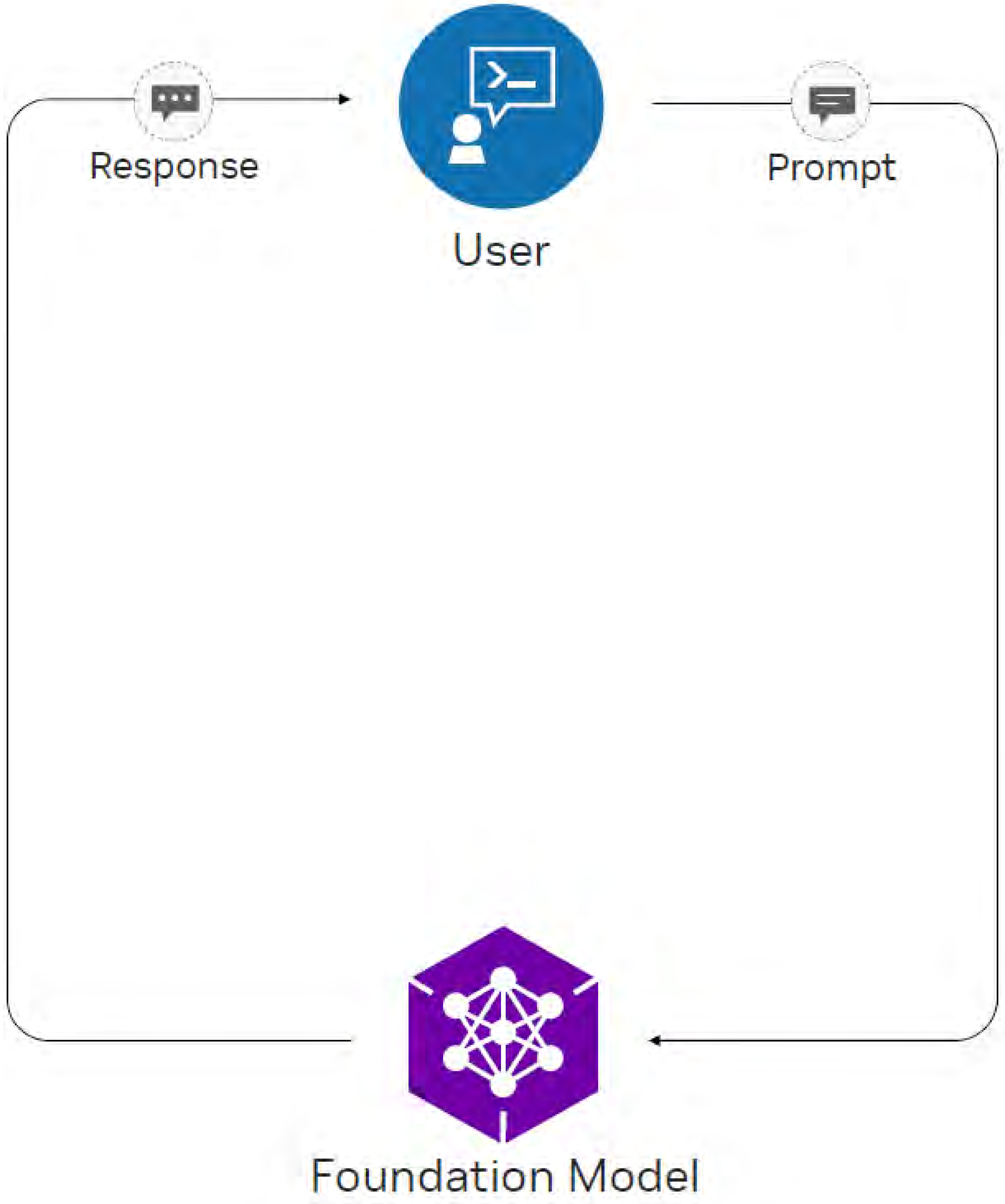
70%

Of enterprise data is untapped
Unlock many new opportunities for greater intelligence

Less frequent re-training
Significant cost and time savings in long-run to maintain LLMs

LLMs are Powerful Tools but Not Accurate Enough for Enterprise

Without a connection to enterprise data sources, LLMs cannot provide accurate information



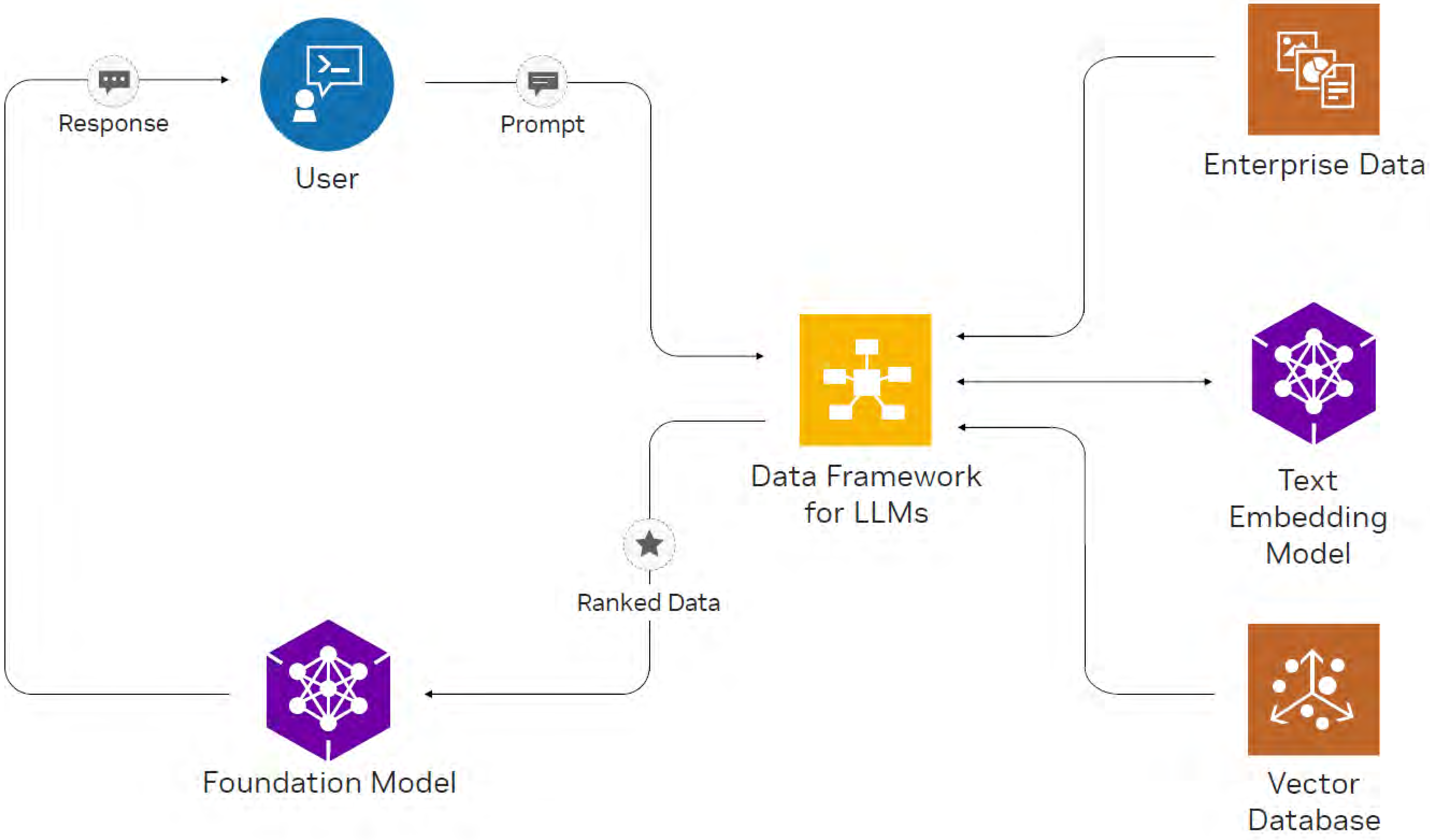
Lacking proprietary knowledge

Risk of outdated information

Hallucinations

Retrieval Augmented Generation Lets Enterprises Talk to Their Data

Enable LLMs to provide up to date and domain specific answers



NVIDIA ChatRTX

Free technology RAG demo that runs on your workstation

https://www.nvidia.com/en-us/ai-on-rtx/chatrtx/

Enterprises Face Challenges Experimenting with Generative AI

Organizations must choose between ease of use and control

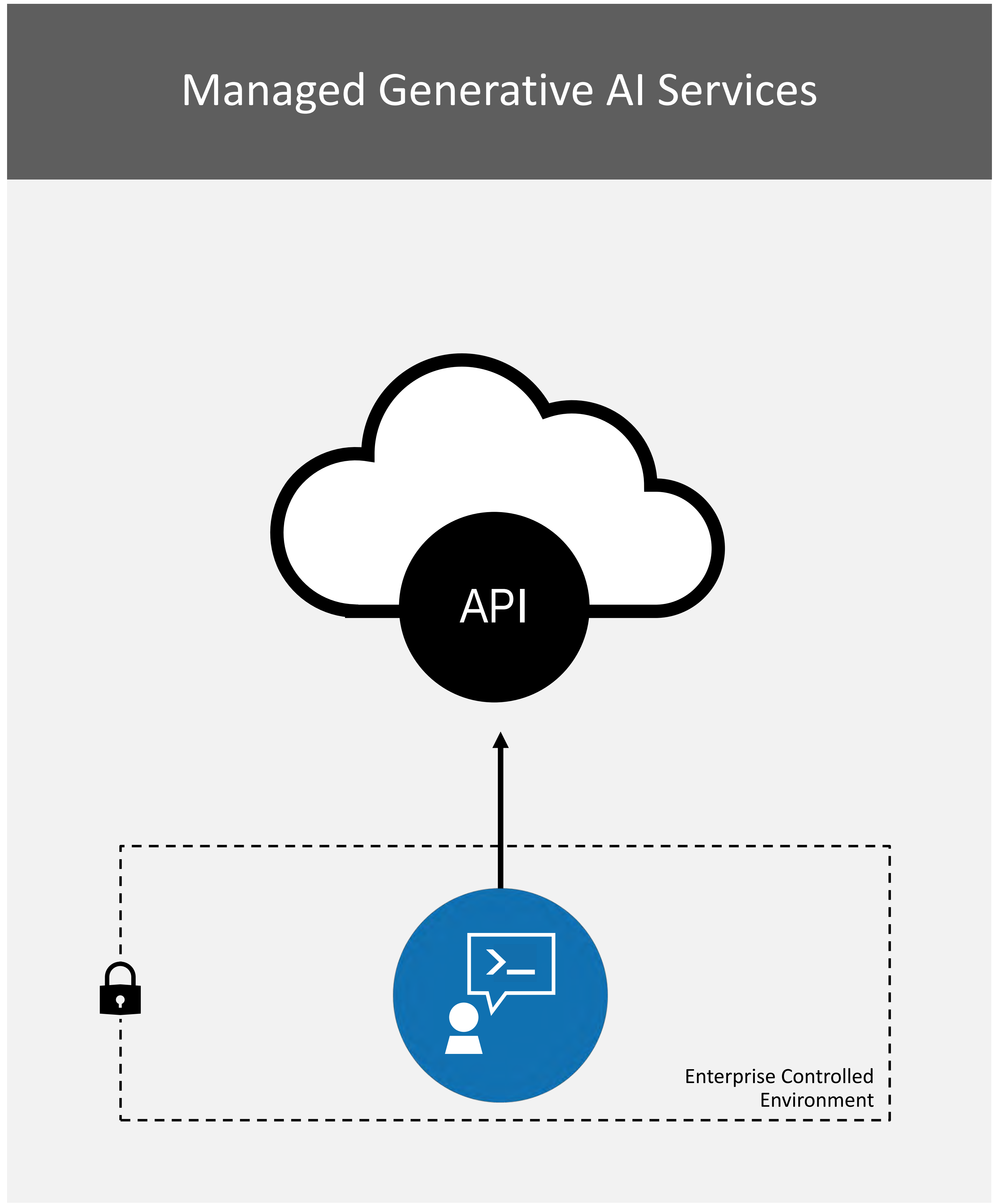
Easy to use APIs for development

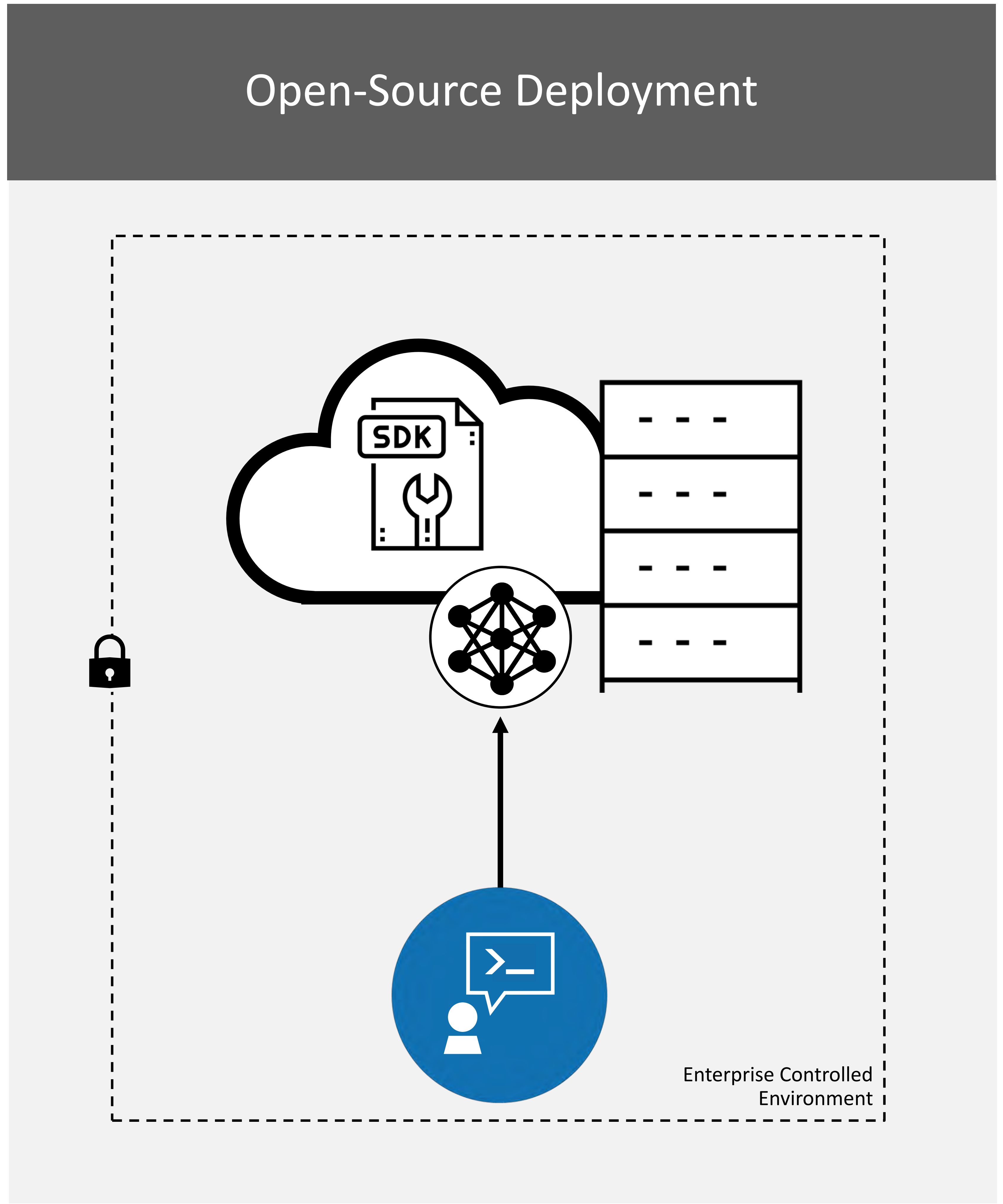
Fast path to getting started with AI

Infrastructure limited to managed environment

Data and prompts are **shared externally**

Limited control for overall generative AI strategy


Managed Generative AI Services
API
Enterprise Controlled Environment



Run anywhere across data center and cloud

Securely manage data in self hosted environment

Tuning required for different infrastructure

Custom code for APIs and fine-tuned models

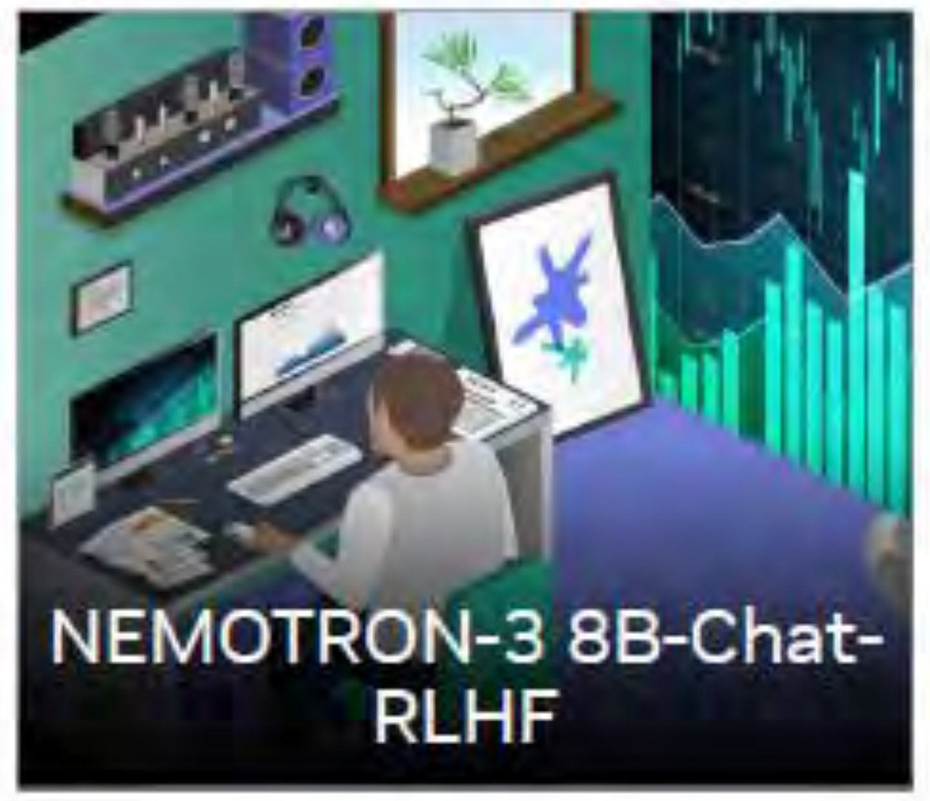
Ongoing maintenance and updates

NVIDIA AI Foundation Models and Endpoints

Fast-track custom generative AI models for enterprise applications

NVIDIA

NVIDIA



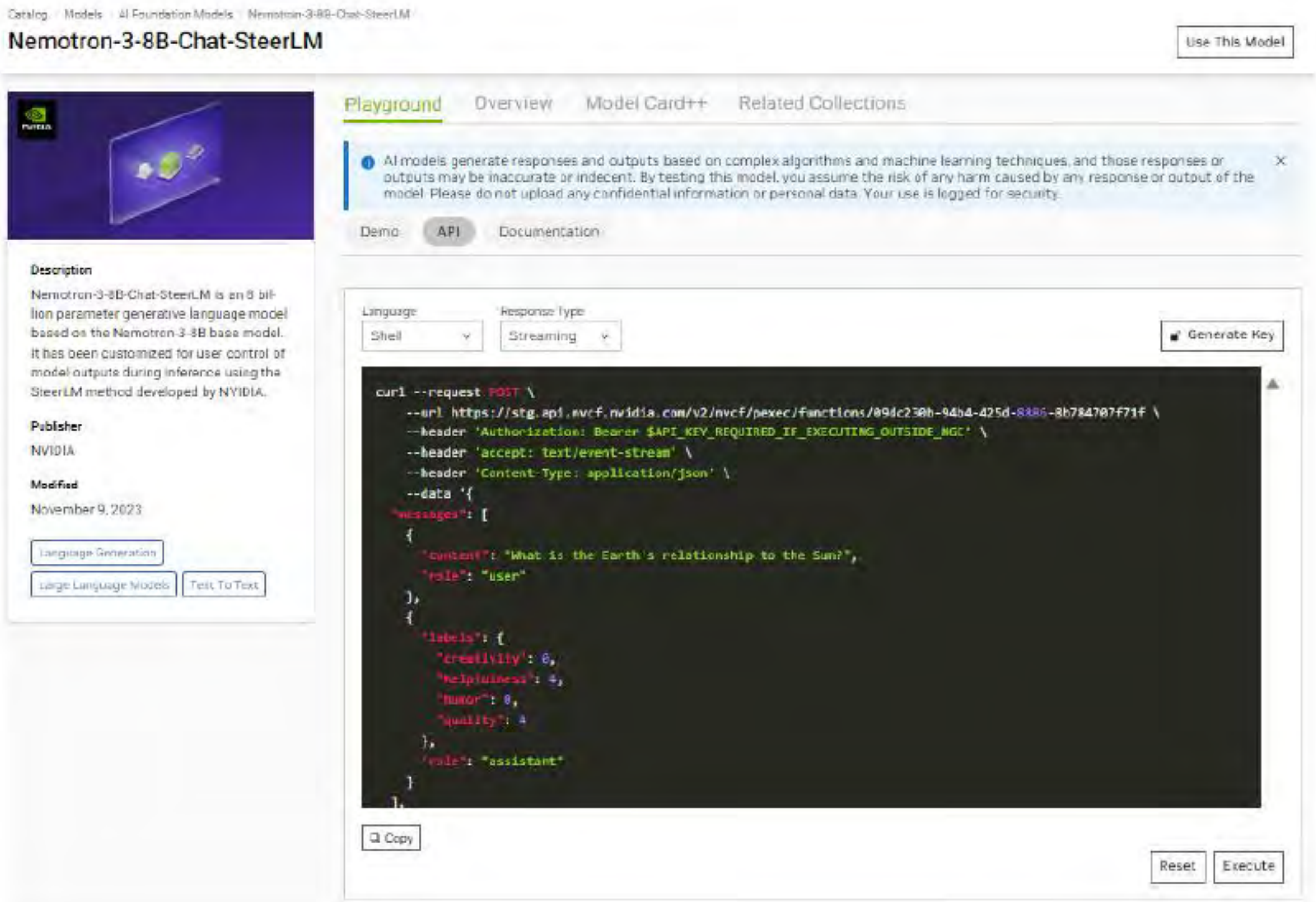
NVIDIA

Meta

stability.ai

Meta

Enterprise-ready, performance optimized models from NVIDIA and the community



Nemotron-3-8B-Chat-SteerLM

AI models generate responses and outputs based on complex algorithms and machine learning techniques, and those responses or outputs may be inaccurate or indecent. By testing this model, you assume the risk of any harm caused by any response or output of the model. Please do not upload any confidential information or personal data. Your use is logged for security.

Description

Nemotron-3-8B-Chat-SteerLM is an 8 billion parameter generative language model based on the Nemotron-3-8B base model. It has been customized for user control of model outputs during inference using the SteerLM method developed by NVIDIA.

Publisher

NVIDIA

Modified

November 9, 2023

Language Generation | Large Language Models | Text To Text

Language: Shell

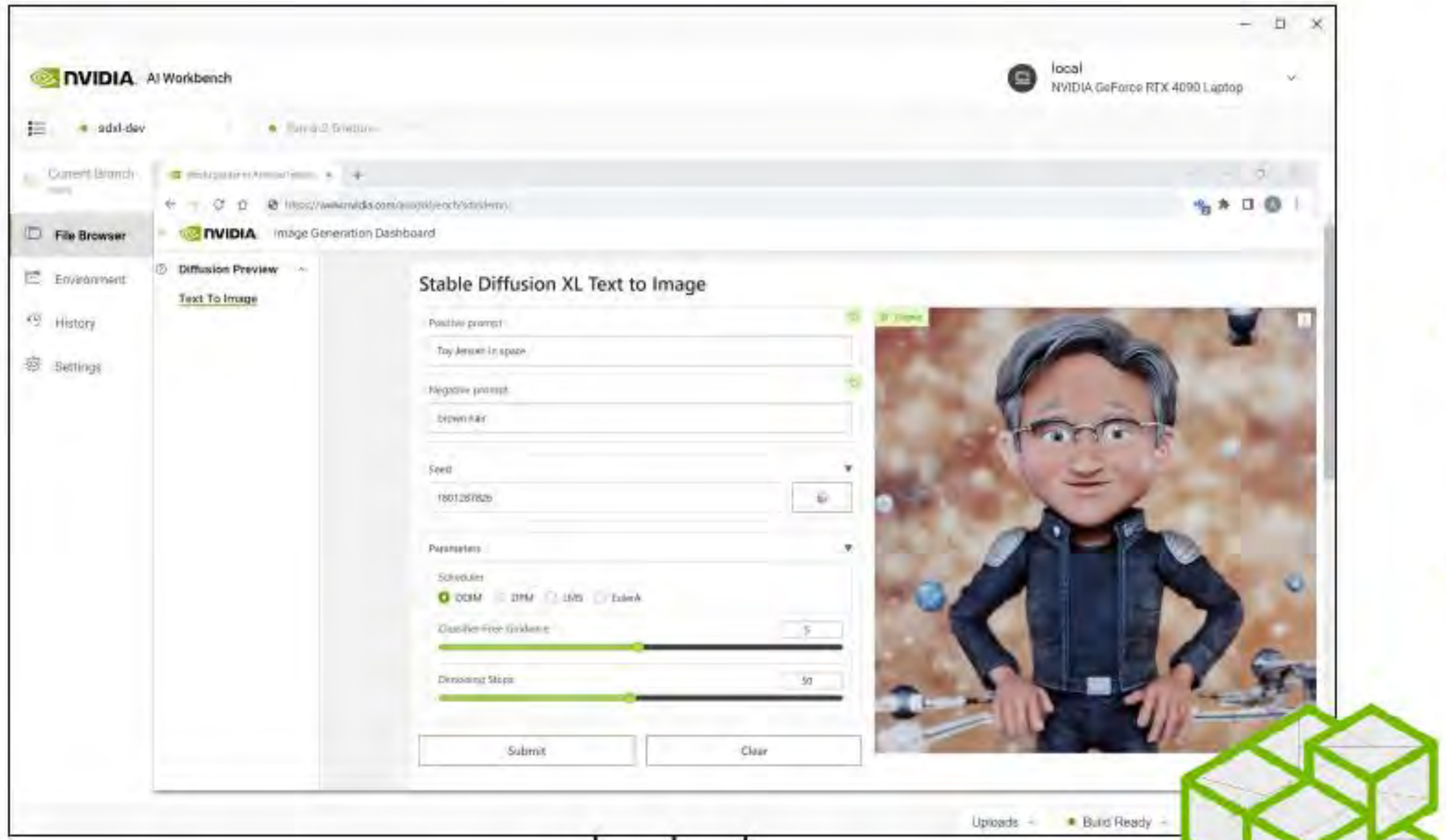
Response Type: Streaming

```
curl --request POST \
    --url https://stg.api.nvcf.nvidia.com/v2/nvcf/pexec/functions/09dc230b-94b4-425d-8886-8b784707f71f \
    --header 'Authorization: Bearer $API_KEY_REQUIRED_IF_EXECUTING_OUTSIDE_NGC' \
    --header 'accept: text/event-stream' \
    --header 'Content-Type: application/json' \
    --data '{
  "messages": [
    {
      "content": "What is the Earth's relationship to the Sun?",
      "role": "user"
    },
    {
      "labels": {
        "creativity": 0,
        "helpfulness": 4,
        "humor": 0,
        "quality": 4
      },
      "role": "assistant"
    }
  ],
```

Experience foundation models running on the NVIDIA AI stack via API endpoints

NVIDIA AI Workbench

Enables anyone with access to a GPU to be a generative AI creator



Stable Diffusion XL Text to Image

Positive prompt: Toy Jensen in space

Negative prompt: brown hair

Seed: 1801287826

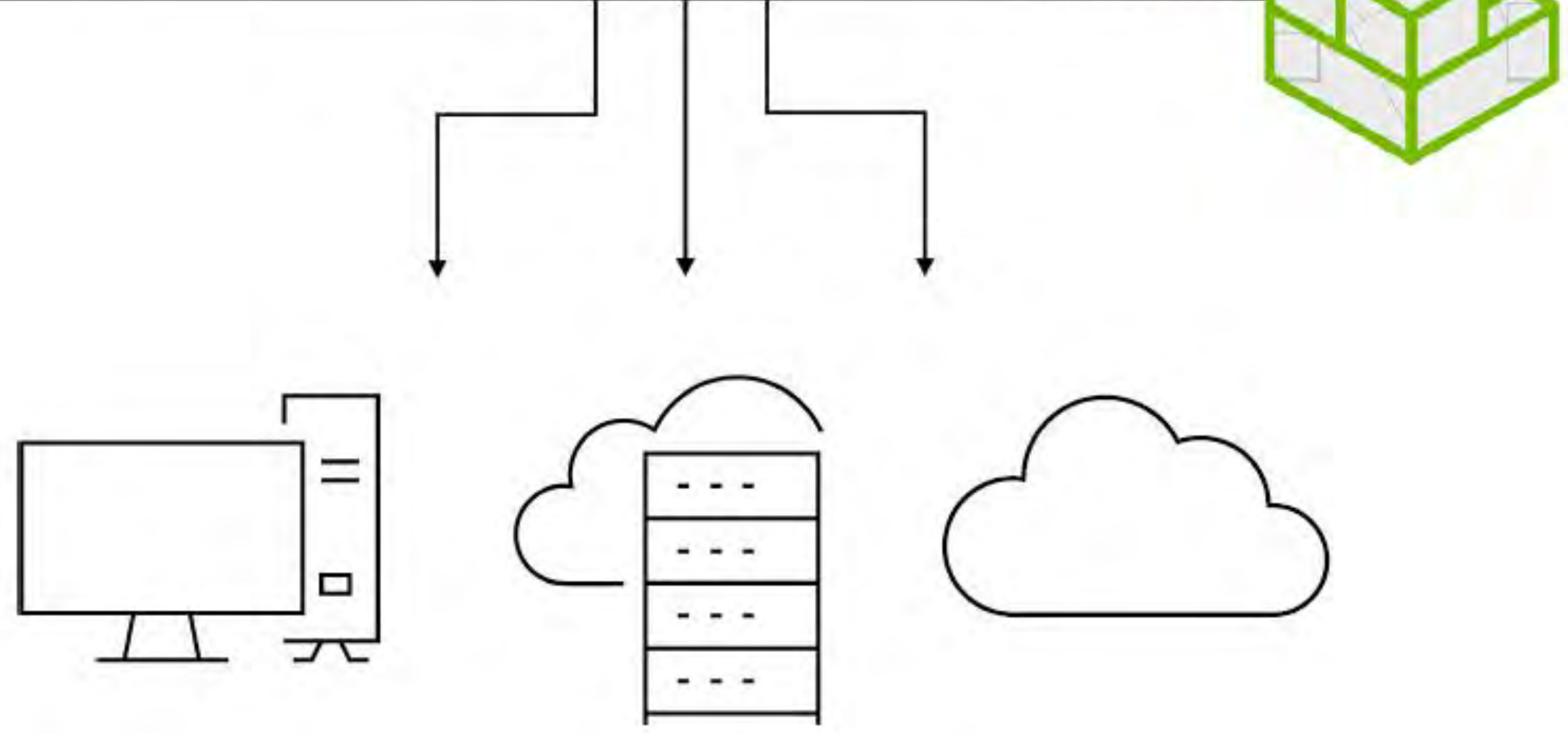
Parameters

Scheduler: ● DDIM ○ DPM ○ LMS ○ EulerA

Classifier-Free Guidance: 5

Denoising Steps: 50

- Create projects for tuning and deployment of generative AI and LLMs
- Move projects between PCs and workstations, data centers, public clouds, and NVIDIA DGX Cloud
- Easily start with pre-built project examples



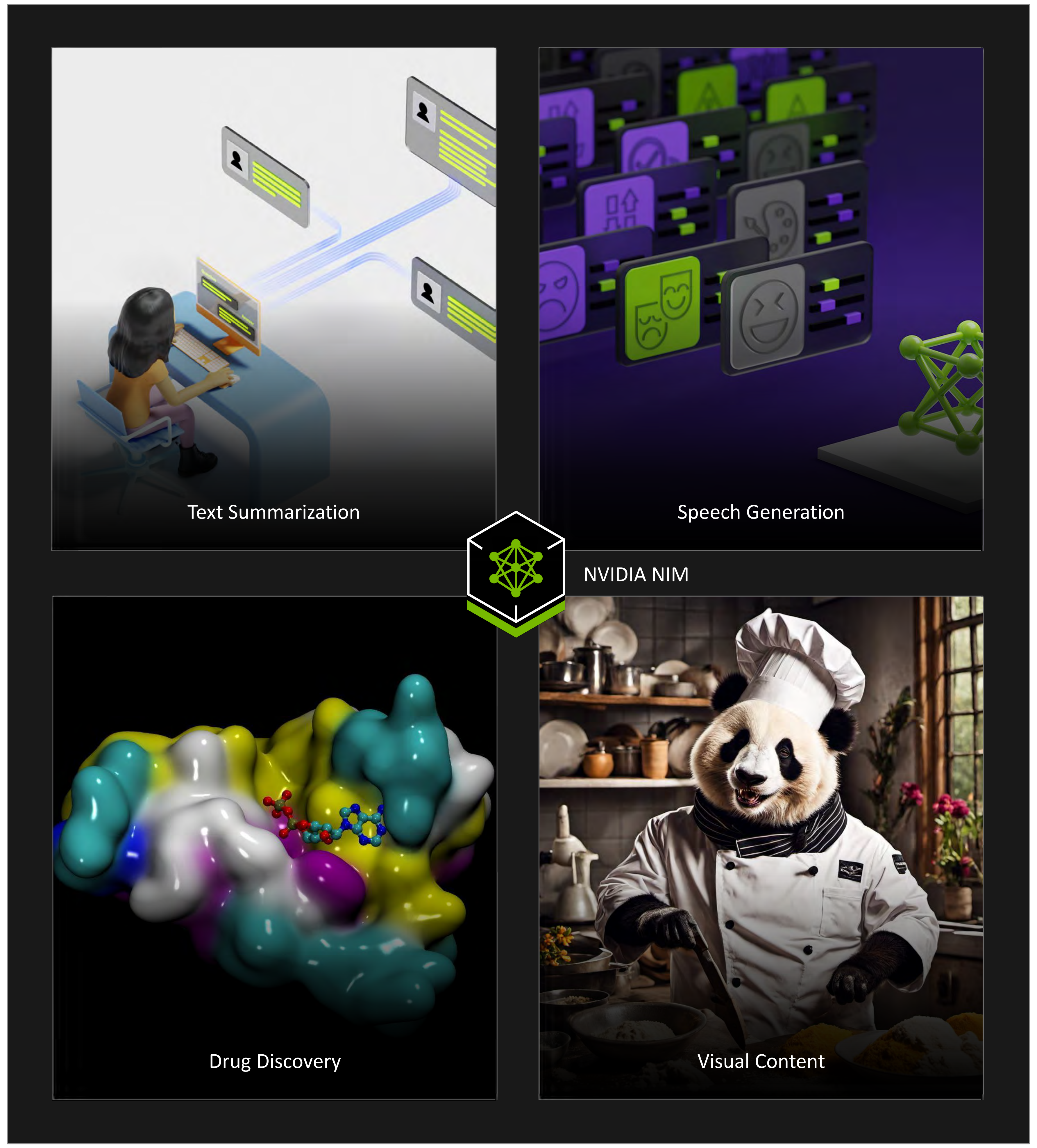
PCs & WORKSTATIONS

DATA CENTERS

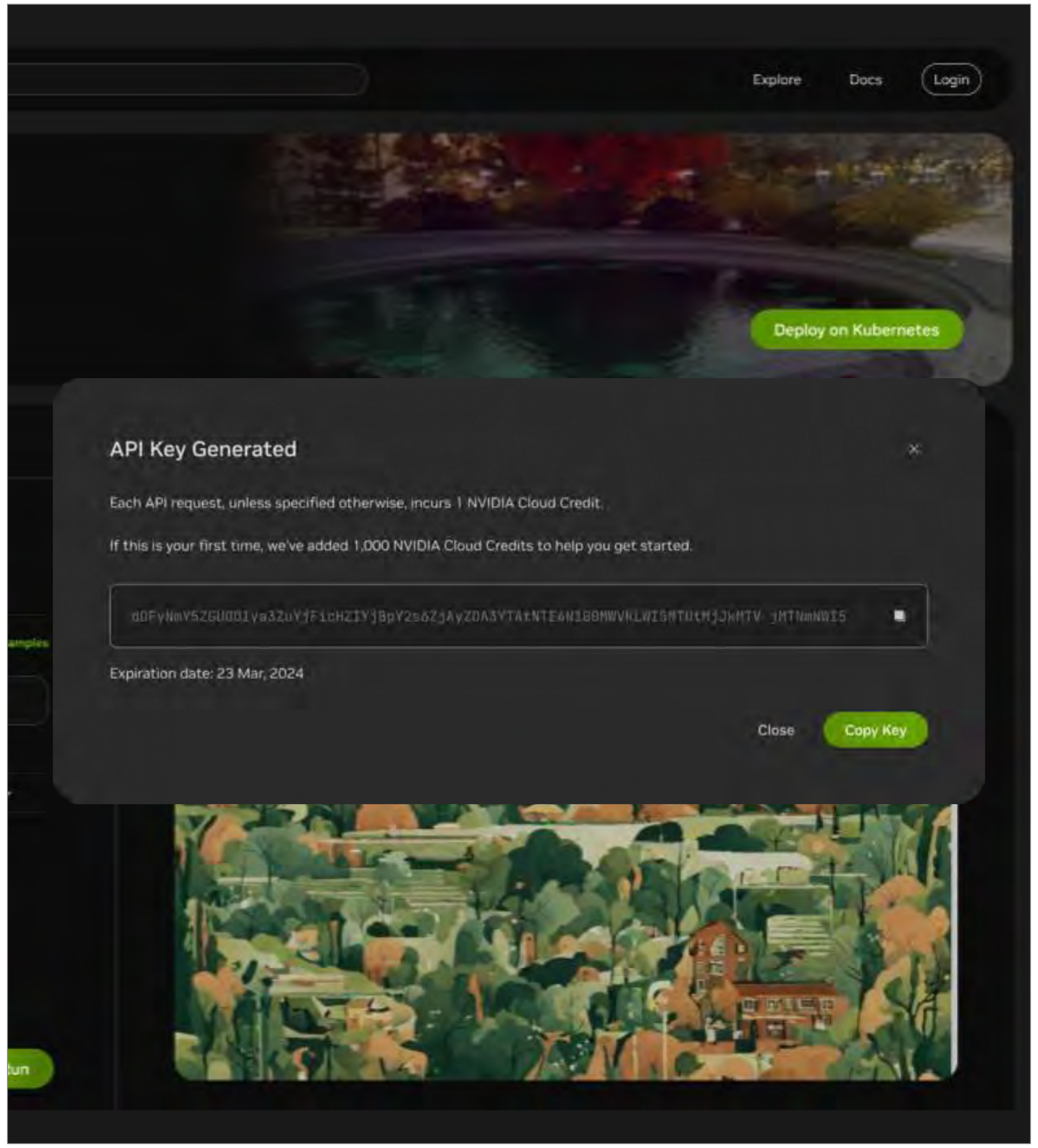
CLOUDS

Experience and Run Enterprise Generative AI Models Anywhere

Use NVIDIA API catalog to get access to NVIDIA NIM



Experience Models



API Key Generated

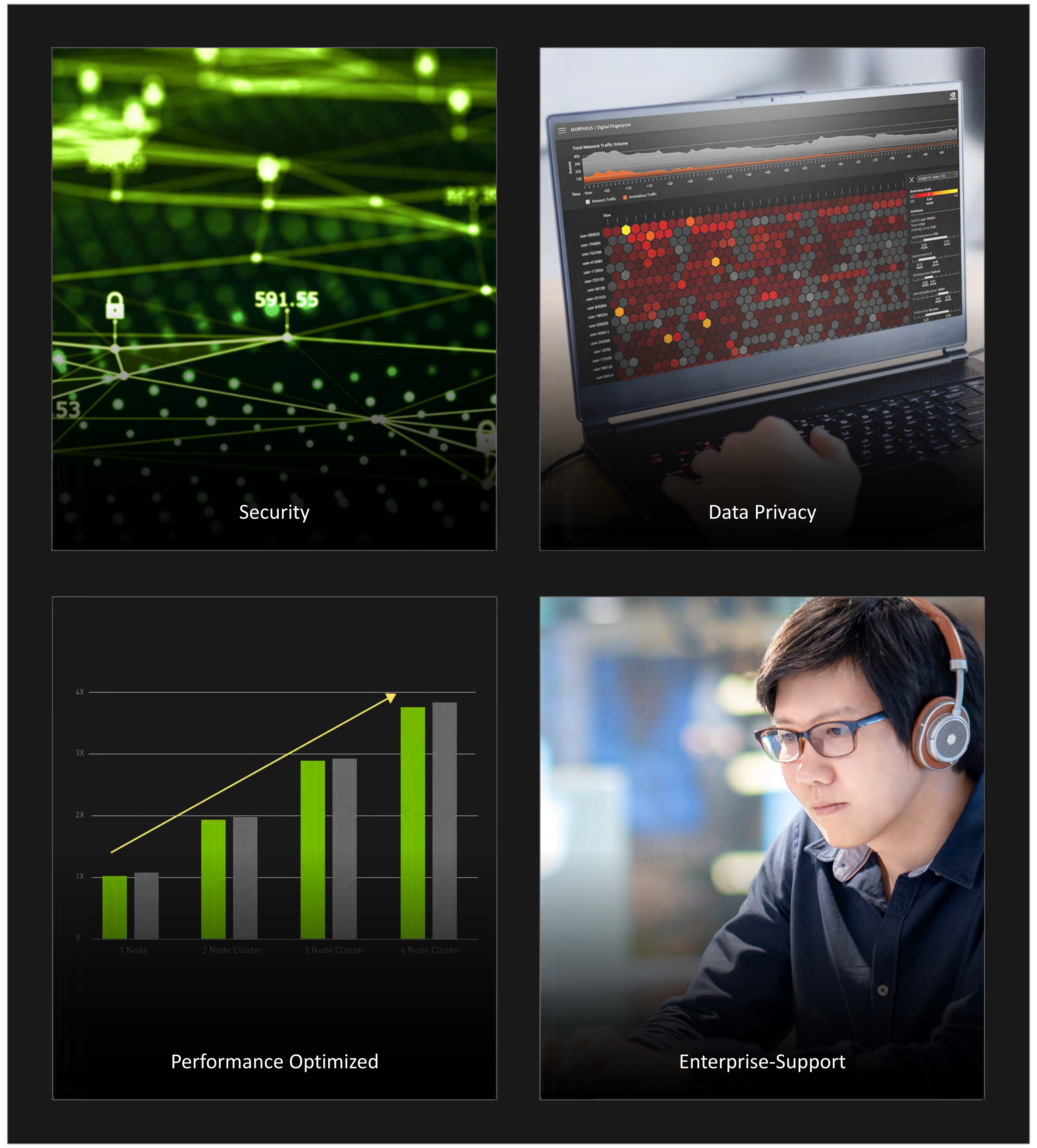
Each API request, unless specified otherwise, incurs 1 NVIDIA Cloud Credit.

If this is your first time, we've added 1,000 NVIDIA Cloud Credits to help you get started.

dDFyNmY5ZGU0OIya3ZuYjFicHZIYjBpY2s6ZjAyZDA3YTAtNTE4NiOOMWVkLWISMTUtMjJkMTV jMTNmNWI5

Expiration date: 23 Mar, 2024

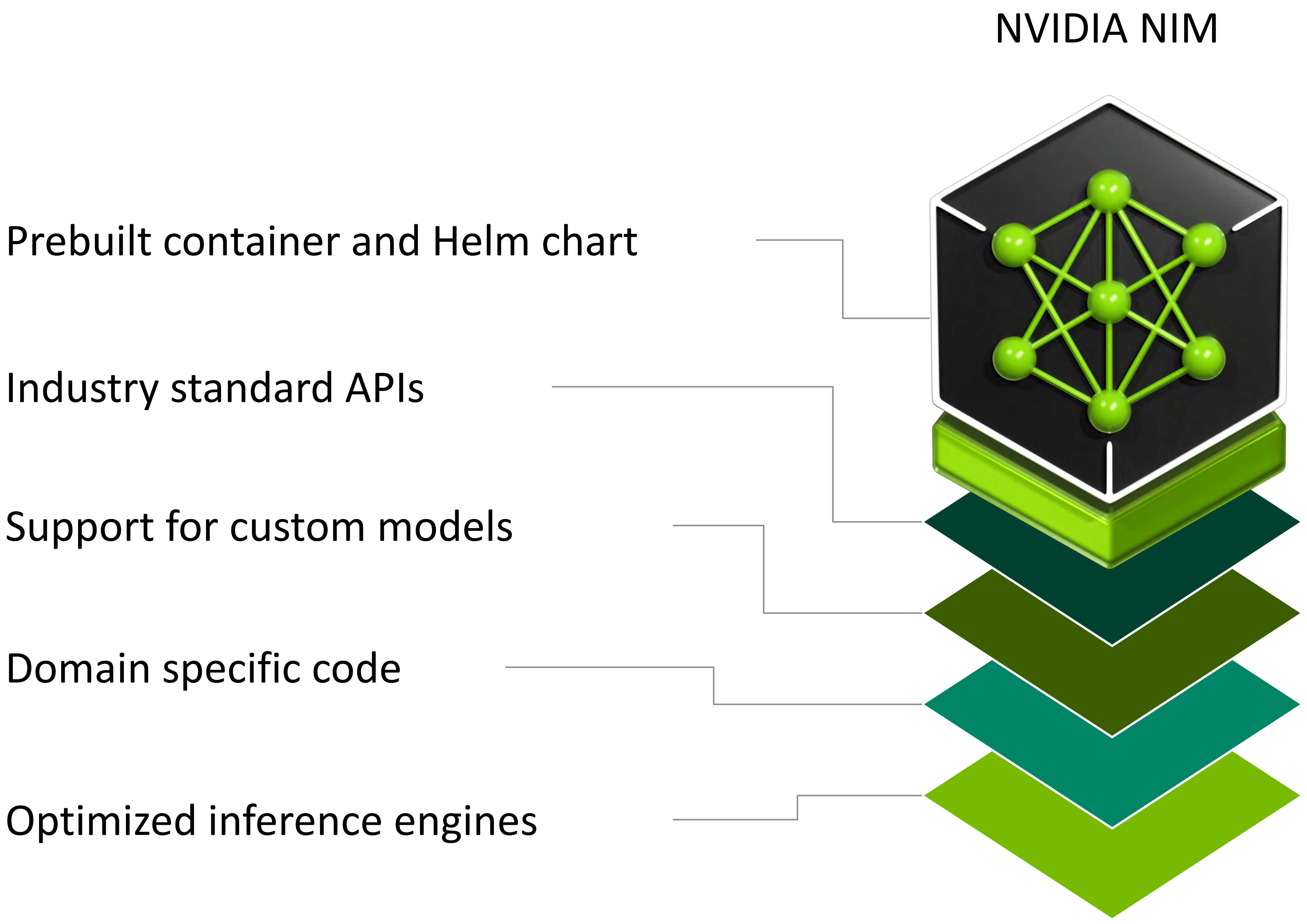
Prototype with APIs



Deploy with NIMs

NVIDIA NIM Optimized Inference Microservices

Accelerated runtime for generative AI



NVIDIA NIM

Prebuilt container and Helm chart

Industry standard APIs

Support for custom models

Domain specific code

Optimized inference engines

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

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

MIXTRAL 8x7B
MISTRAL AI_
GEMMA 7B
Google
FUYU
ADEPT
NEMO RETRIEVER
NVIDIA
AI GENERATOR
gettyimages
KOSMOS 2
Microsoft
3D GENERATOR
shutterstock
AUDIO2FACE
NVIDIA
ESM FOLD
Meta
VISTA-3D
NVIDIA
DIFFDOCK
MIT
MolMIM
NVIDIA

NVIDIA-Certified Systems through leading partners

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

APPLICATION NIMs

- Llama Guard
- Retrieval Embedding
- Retrieval Reranking

DIGITAL BIOLOGY NIMs

- DeepVariant
- DiffDock
- ESMFold
- MolMIM
- Vista 3D

HP Z Captis

Capture All the World’s Materials

Digitize materials from anywhere in minutes,[1] in a portable device equipped with a polarized and photometric computer vision system for efficiency and accuracy.

Learn More