How NVIDIA Accelerates Retailers on their GenAI Journey

NVIDIA Retail SA team

NVIDIA GTC / March 19th, 2024

Agenda

Introduction to NVIDIA Retail – Ward Eldred

How NVIDIA Retail Accelerates the top Retailers

NVIDIA Accelerated Software Platforms – Paul Hendricks

How NVIDIA Helps Retailers on their Accelerated Computing Journey

Generative AI & Omniverse – Dora Li

Exciting emergent use cases across marketing & ads, personalization, product design, and more, utilizing multi-modal generative AI.

Wrap-Up & Questions from Audience

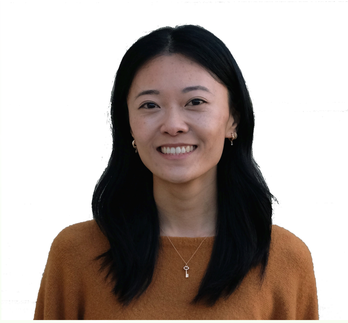
Meet the Team!

NVIDIA Retail Solution Architects

Ward Eldred
Solutions Architect Manager

Paul Hendricks
Senior Solutions Architect

Mark Moyou
Senior Solutions Architect



Dora Li
Senior Solutions Architect

James Clohessy
Senior Solutions Architect

Top Use Cases Across Retail, QSR, and CPG

What we spend our time on

DATA PROCESSING & ML

Customer 360 & Segmentation
Hyper-Personalization & Recommendations
Demand Forecasting
Digital Assortment
Real-Time Customer Interaction
Knowledge graphs

GENERATIVE AI

Avatars for Customer Engagement
Multi-modal Recommendations
Customer Service Chatbot
Marketing asset generation
Product/Packaging Design
Product description generation
Hyper-personalization

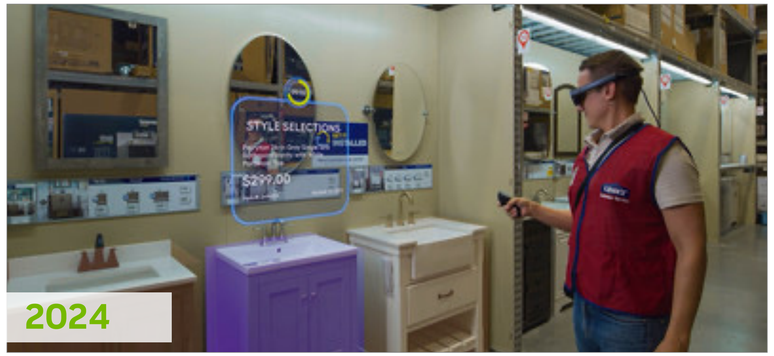
INTELLIGENT VIDEO ANALYTICS

Loss Prevention
Organized Retail Crime
Store Analytics & Stockout
Visual Inspection in Manufacturing
Intelligent Warehouses

ACCELERATED MODEL TRAINING & INFERENCE

GTC Sessions From Our Partners

Companies & teams we work with

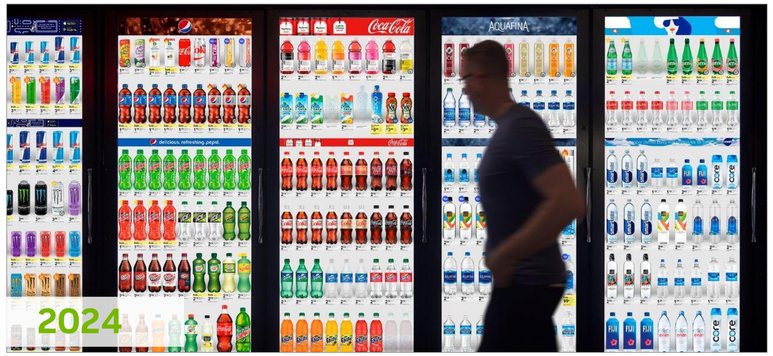


Lowe's

How Lowe's is Driving Innovation and Agility With AI

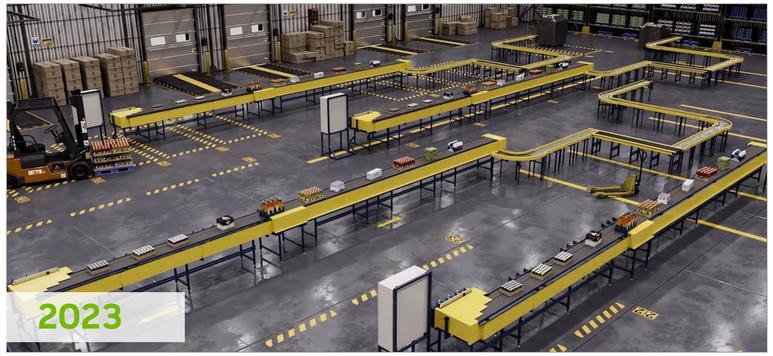
L'Oreal

Augmented Marketing in Beauty with Generative AI



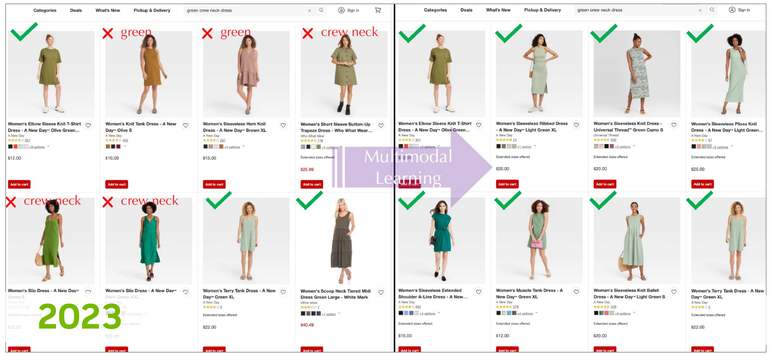
Kroger

Revolutionizing Retail With AI: The Journey of Cooler Screens and NVIDIA in Transforming Kroger Stores



Pepsi

Optimizing Distribution and Fulfillment Center Operations with Computer Vision and Digital Twins



Target

Improving the Guest Experience Using Multimodal Learning for Fashion

Walmart

Product Title Generation for Conversational Systems using BERT

And many more!

Pioneering Accelerated Computing

Accelerated computing requires full-stack optimization, from chip architecture, systems, and acceleration libraries, to refactoring the applications. The global NVIDIA ecosystem spans 4 million developers, 40,000 companies, and over 3,000 applications.

Accelerating Retail GenAI Platforms

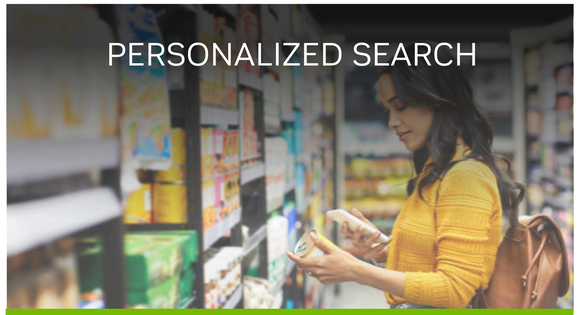
Multimodal Generative AI is Transforming Retail

TEXT	IMAGE	AUDIO	VIDEO	3D
Chatbots or Assistants	Brand Creation	Text to Voice Generation	Dynamic Code Commenting	Product Design & Discovery
NEMO	SDXL	RIVA	PICASSO	NEURAL-ANGELO
Marketing Copy	Personalized Content	Sound Creation	Function Generation	3D Object Generation

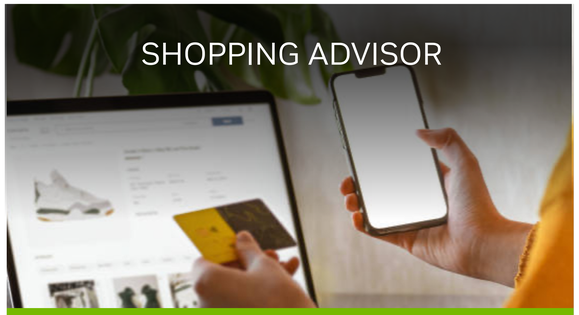
Enterprises that adopt next-generation AI like LLMs and Generative AI are **2.6X more likely to increase revenue by 10% or more** but must invest in their AI infrastructure to fully reap the benefits.

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

Generative AI Impacts Every Function in Retail



Multi-modal* search with natural language understanding to deliver better recommendations

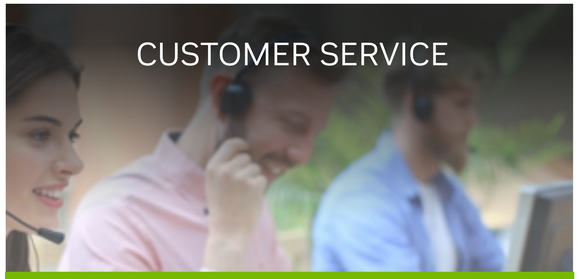


Personalized recommendations and conversation style tailored to each customer's needs

Generating catalogue information like descriptions, titles, and product images

Tools like operator manual chatbots and translators to assist employees

From ideation to creation, simulating products in various environments and use cases



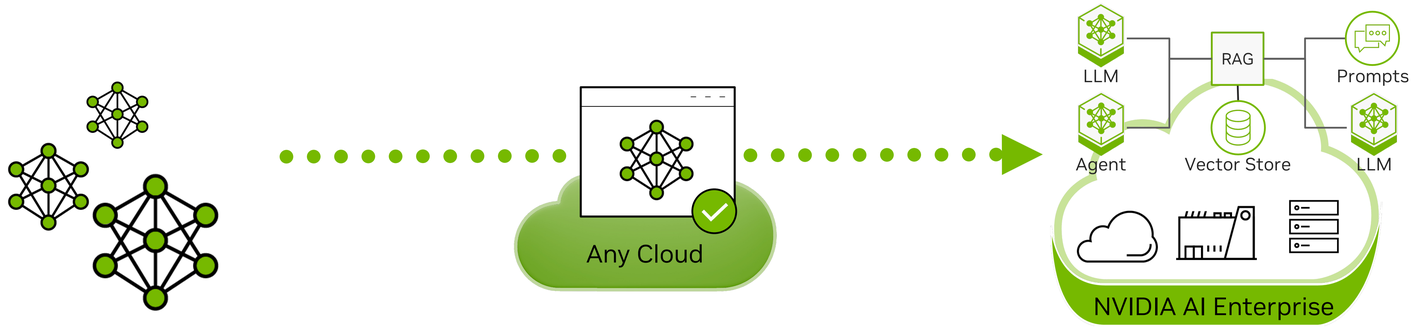
Using chatbots for simple inquiries and routing complex ones to human agents for improved, more efficient customer service

Creating personalized ad copy, email, etc. increasing click to conversion

Generating personalized advertisements or product images

Enterprises Need an AI Foundry to Build Custom Models

An AI foundry *is a new kind of service, that brings together AI Foundation Models, state of the art tools for model customization & model customization know how*



1

Experience
NVIDIA AI Foundation or Open Source Models

2

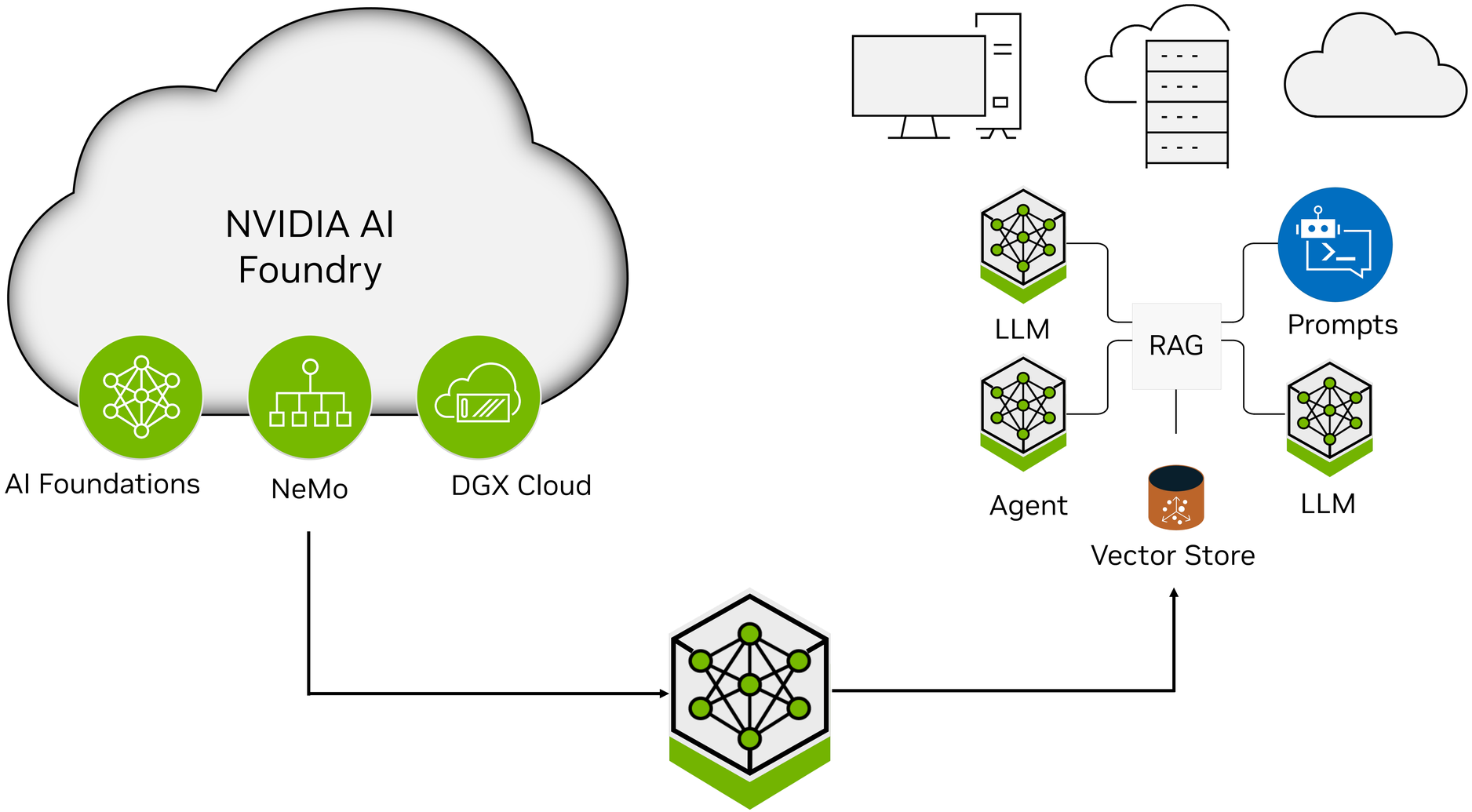
Customize with
NVIDIA NeMo Services

Tune on
GCP

3

Deploy anywhere with
NVIDIA AI Enterprise

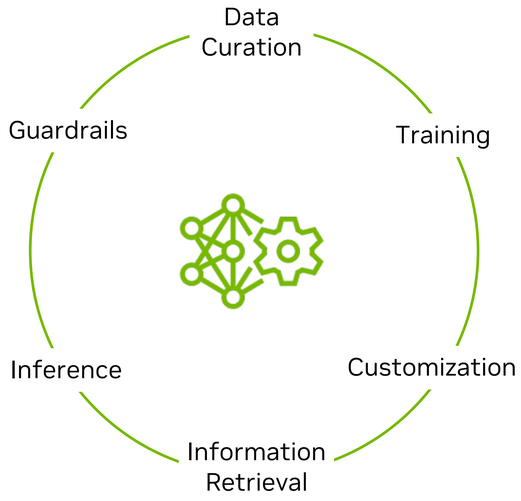
NVIDIA AI Foundry for Custom Enterprise-Ready LLMs



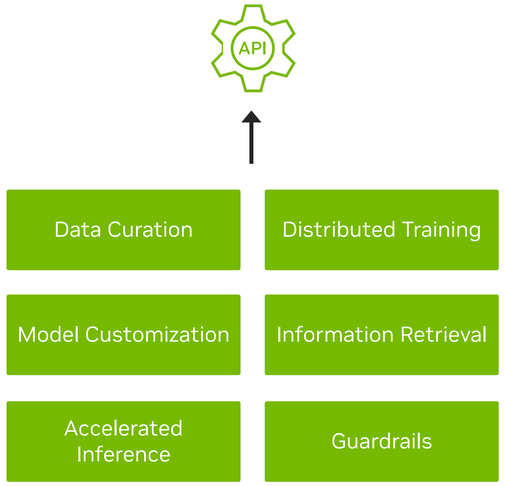
NVIDIA NeMo

Software and APIs to Run an AI Foundry and Create Custom LLM Applications

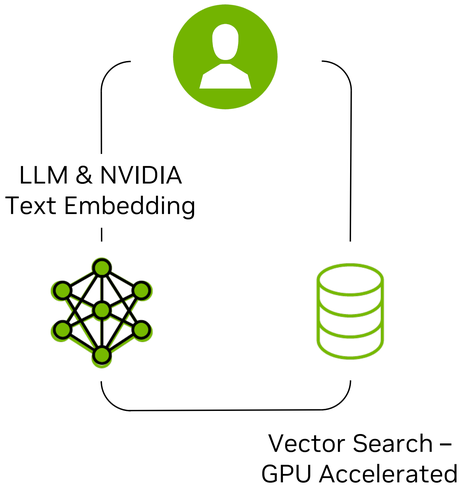
Framework for Customization, Optimization, and Deployment



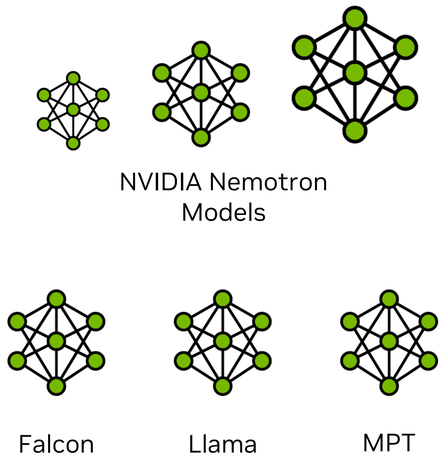
Microservices and APIs that Run Anywhere



Retrieval Augmented Generation for Next Generation Search and Chatbots

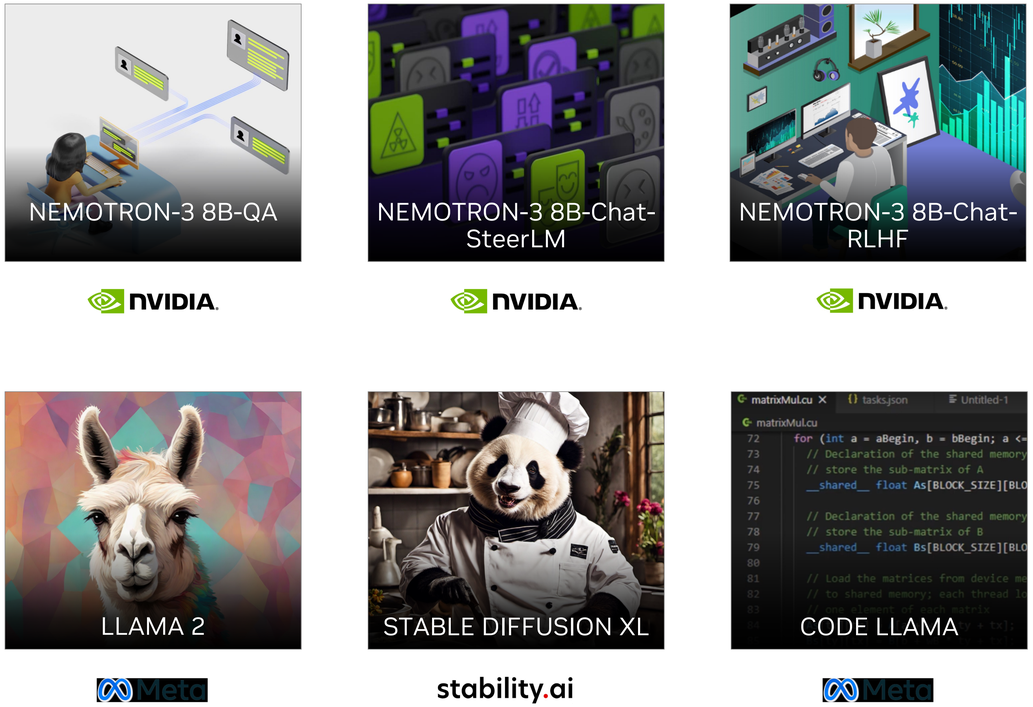


State-of-the-Art Pre-Trained Models

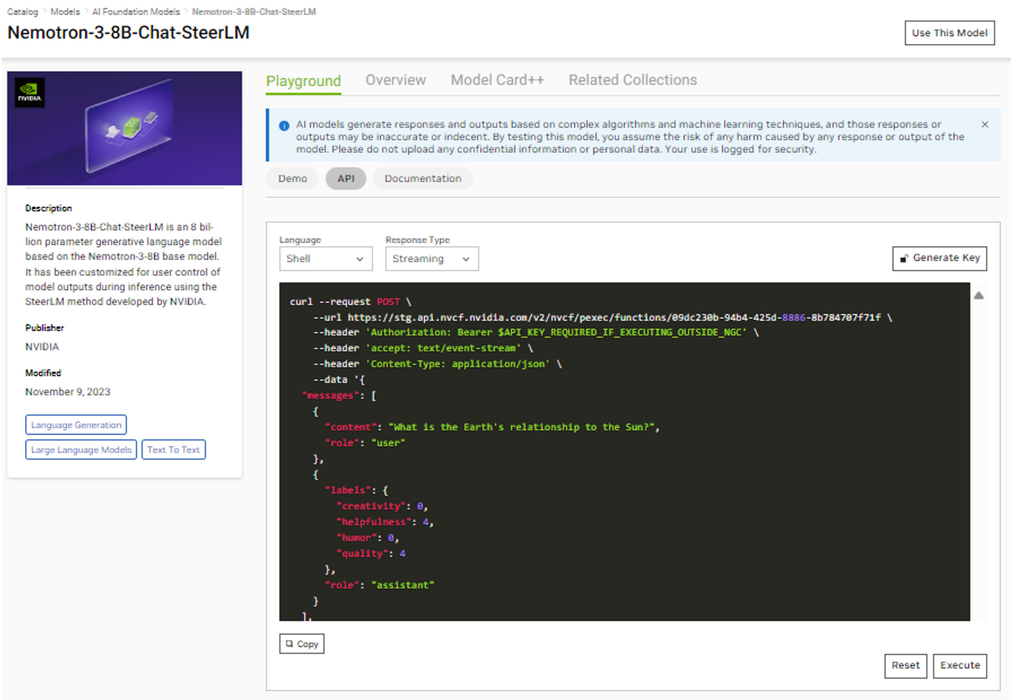


Announcing NVIDIA AI Foundation Models and Endpoints

Fast-track custom generative AI models for enterprise applications



Enterprise-ready NVIDIA models
Performance-optimized community models



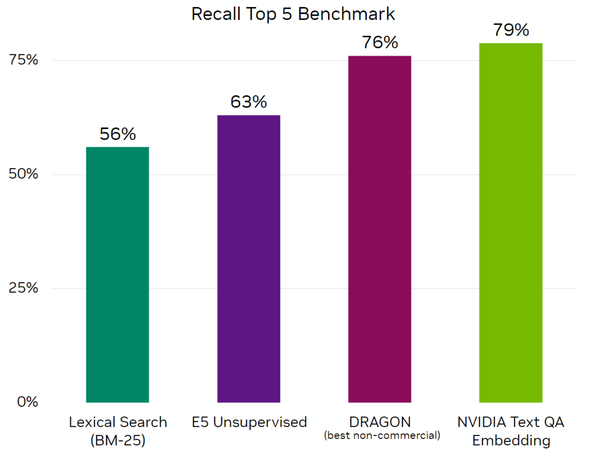
Experience models on NVIDIA AI stack via API endpoints

Two Microservices to Optimize RAG

Real-time Inference for Enterprise Grade Gen AI Deployment

NVIDIA NeMo Retriever Microservice

- **Enterprise Grade RAG:** An embedding and retrieval model, API and microservices that delivers greater accuracy. Leverages GPU accelerated Vector data bases.
- Where to try it:
  - NGC AI Foundations Playground



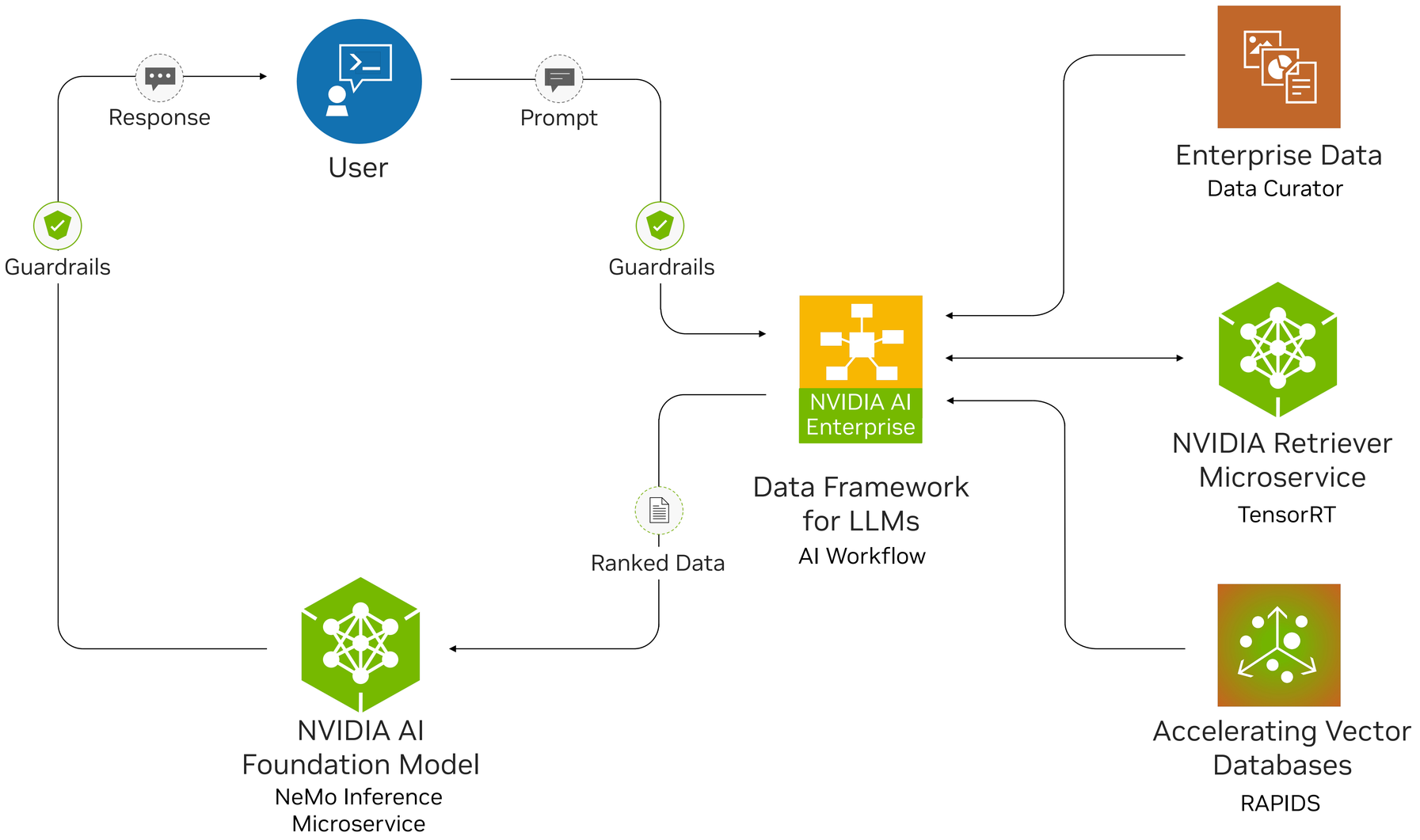
Comparing NVIDIA Text QA Embedding Model vs Other Available Options. Recall Top 5, 300 token chunk size, averaging across representative customer datasets from Telco, IT, Consulting, Energy

NVIDIA NeMo Inference Microservice

- **Real-time Inference for Enterprise-Grade Deployment**
  - The easiest and quickest way to set up an LLM service API endpoint on preferred compute.
  - Zero to LLM Endpoint (on your preferred compute in <5 mins)
  - API Endpoint for LLM to run on your compute
    - Completion API
    - Chat API
    - Healthcheck API
    - Logging and debugging API
  - OOTB conversion-less optimized models expedites deployment
  - NV Inside: Triton, TensorRT-LLM, vLLM, TensorRT
  - Nemotron models 43B, 8B; Llama 2 70B, Starcoder
- Where to try it:
  - NGC AI Foundations Playground

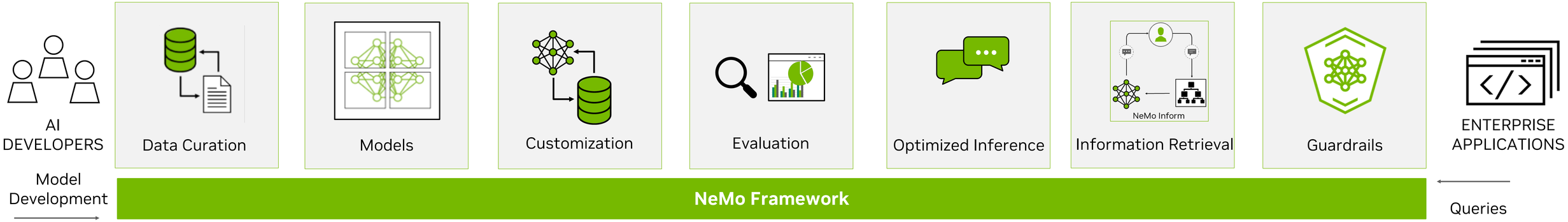
NVIDIA Provides Optimized Retrieval Augmented Generation

Commercially viable, optimized embedding, reranking, and personalization to deliver highest accuracy and performance



AI Foundry: The NeMo Platform

Enterprises Need an AI Foundry to Create, Evaluate, and Deploy Custom 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

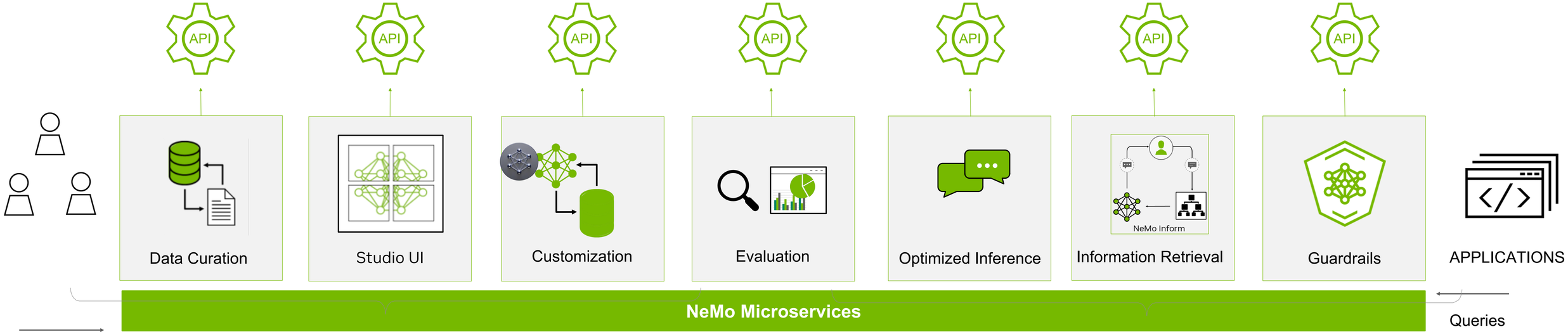
NVIDIA AI Enterprise and experts by your side to keep projects on track

GA with NVIDIA AI Enterprise

Multi-modal available via early access now

AI Foundry: Delivered As Microservices

Run as microservices where your data resides on your preferred compute



Value Prop:
- Accelerated data curation

Value Prop:
- Graphical user interface to simplify using any Nemo Microservice

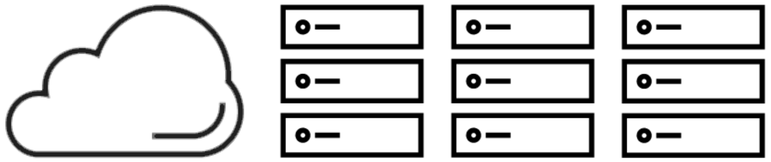
Value Prop:
- Latest tuning techniques for different data / compute restrictions Abstracts complexity of launching training jobs on k8s

Value Prop:
- Easiest way to benchmark and evaluate model quality

Value Prop:
- Easiest and most performant way (latency/QPS) to run LLM models in production

Value Prop:
- SotA Embedding and Retrieval models make sure recall contains the answer or says “I don’t know”

Value Prop:
- Keep LLMs within boundaries and protect from malicious prompts

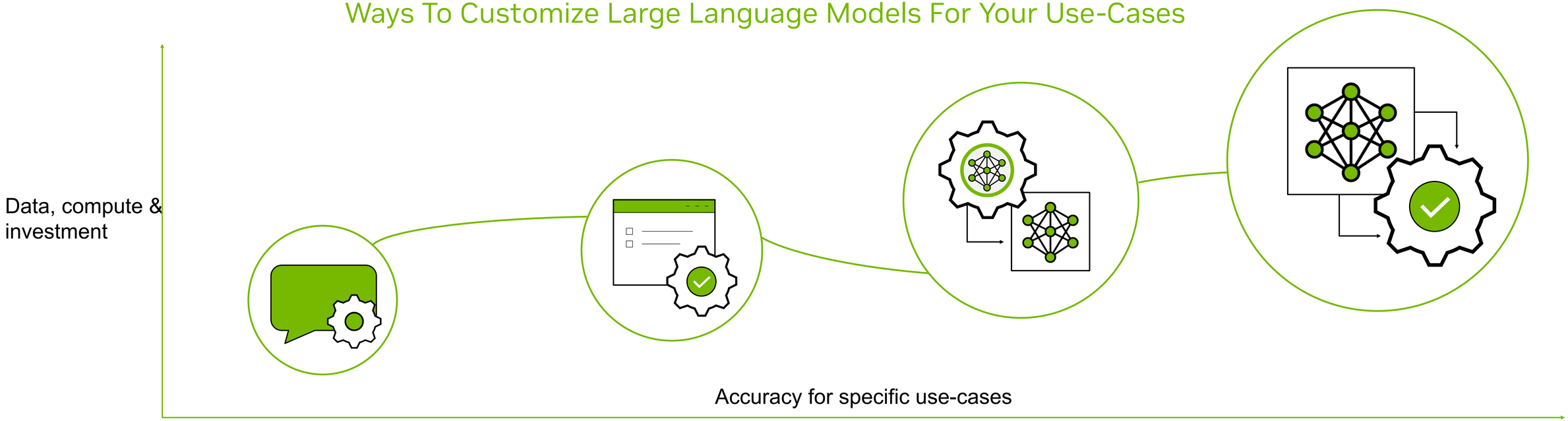


NVIDIA AI Factory

Google Cloud · Microsoft Azure · aws · ORACLE · DELL Technologies · Hewlett Packard Enterprise · Lenovo · SUPERMICRO · NVIDIA

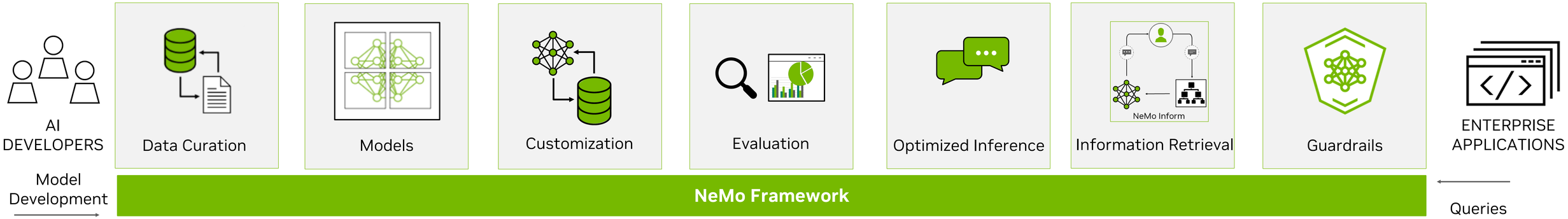
Suite of Model Customization Tools in NeMo

Ways To Customize Large Language Models For Your Use-Cases



	PROMPT ENGINEERING	PROMPT LEARNING	PARAMETER EFFICIENT FINE-TUNING	INSTRUCTION TUNING
Techniques	. Few-shot learning . Chain-of-thought reasoning . System prompting	. Prompt tuning . P-tuning	. Adapters . LoRA . IA3	. SFT . RLHF
Pros	. Good results leveraging pre-trained LLMs . Lowest investment . Least expertise	. Better results leveraging pre-trained LLMs . Lower investment . Will not forget old skills	. Best results leveraging pre-trained LLMs . Will not forget old skills	. Best results leveraging pre-trained LLMs . Change all model parameters
Cons	. Cannot add as many skills or domain specific data to pre-trained LLM	. Less comprehensive ability to change all model parameters	. Medium investment . Takes longer to train . More expertise needed	. May forget old skills . Large investment . Most expertise needed

AI Foundry: The NeMo Platform



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

GA with NVIDIA AI Enterprise

Multi-modal available via early access now

Google Cloud · Microsoft Azure · aws · ORACLE · NVIDIA DGX Cloud · DELL Technologies · Hewlett Packard Enterprise · Lenovo · SUPERMICRO

Product Advisor AI Workflow

Why Product Advisor?

Product Advisor Mission

- Empower Retailers with ability to quickly create LLM/RAG pipelines to answer questions about products.
- Internal and external use cases – Sephora, Shopify, Instacart, Walmart.

Traditional Search vs. Conversational LLM/RAG

- Traditional search – well-solved over 30+ years, hyper-optimized, handles simple queries.
- Performs poorly with long-tail, complex queries.
- Instead, generate and suggest products the user hasn't thought of and retrieve information about those products.
- Answer questions about a product or across several products using the product information.
- Aim to augment – not replace – traditional search.

Product Advisor

Hello! 👋 I'm your dedicated Shopping Advisor created by NVIDIA 🎮, here to answer any questions you might have and help you find anything you're looking for. What can I help you with today?

User

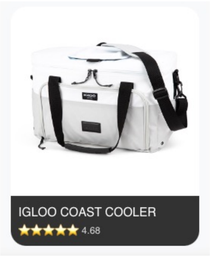
I'm going camping. What are some things I might need?

Product Advisor

When going camping, there are a few essential items you might need to bring along to ensure a safe and enjoyable trip. Here are some suggestions:

....

- Tent
- Sleeping Bag
- Lantern
- First Aid Kit
- Camping Chair
- Water Filter
- Etc.



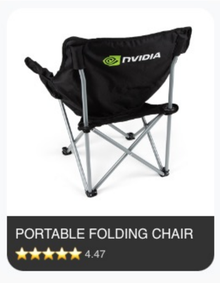
User

Can you recommend a chair?

"camping chair

Product Advisor

The Portable Folding Chair is a great option for camping. It has a steel frame that is strong and durable, and it can support a weight capacity of up to 300 lbs. The chair also includes a mesh gear pouch for storing small items, and it is easy to fold up and transport.

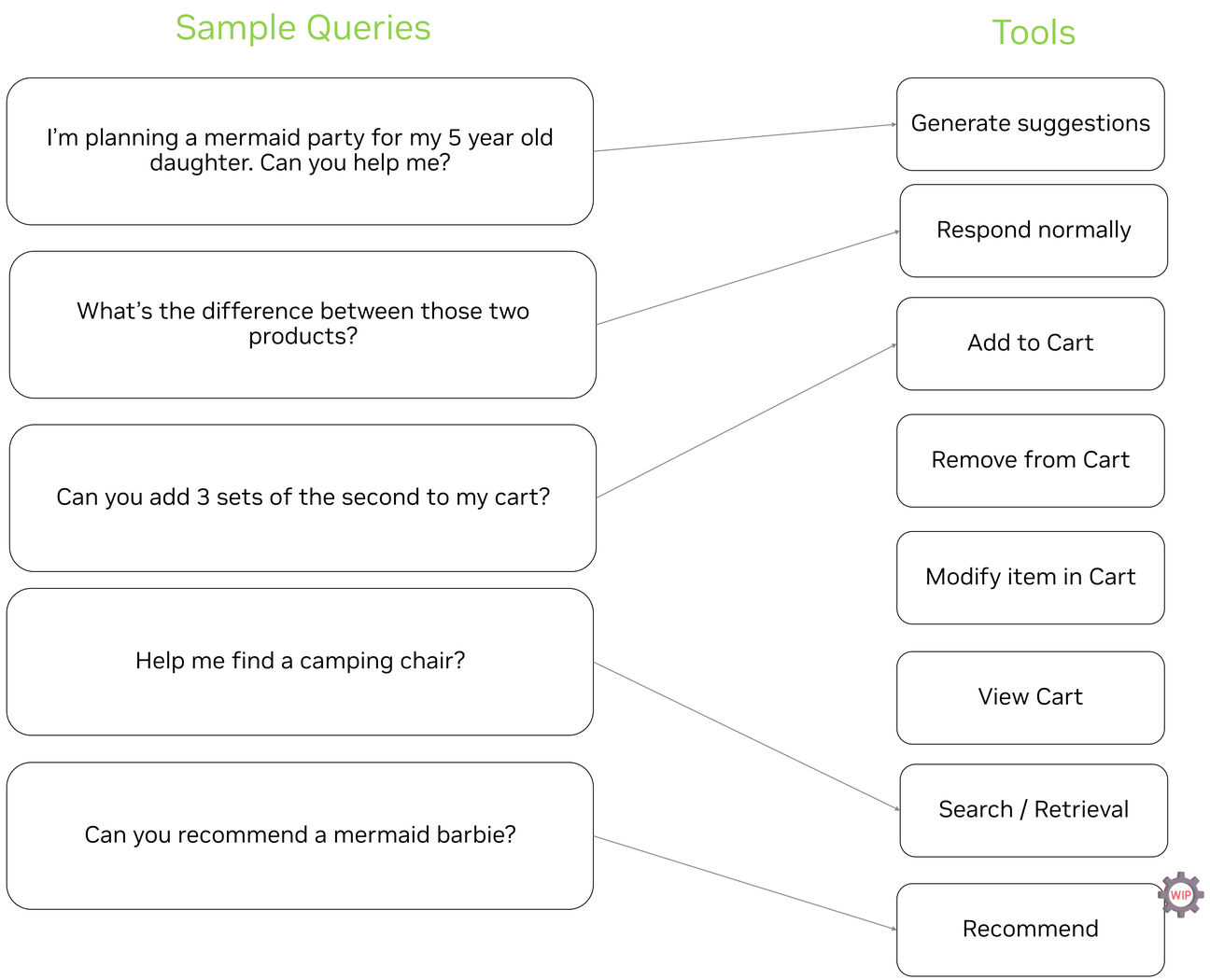


Product Advisor AI Workflow

LLM Decision Making

Empowering LLMs with APIs

- The long pole in the tent is handling the myriad nuances of natural conversation.
- Can't do retrieval every time.
- Creating a prompt to handle every scenario is too wieldy and rigid.
- Retailers have APIs for product information, pricing, inventory, shopping cart, search, recommendation.
- Allowing LLMs to access Tools & APIs and decide which Tool/API to use allows for flexibility and delivers qualitatively promising results.



Product Advisor AI Workflow

Zero to Development in <5 minutes on your own data

Assets

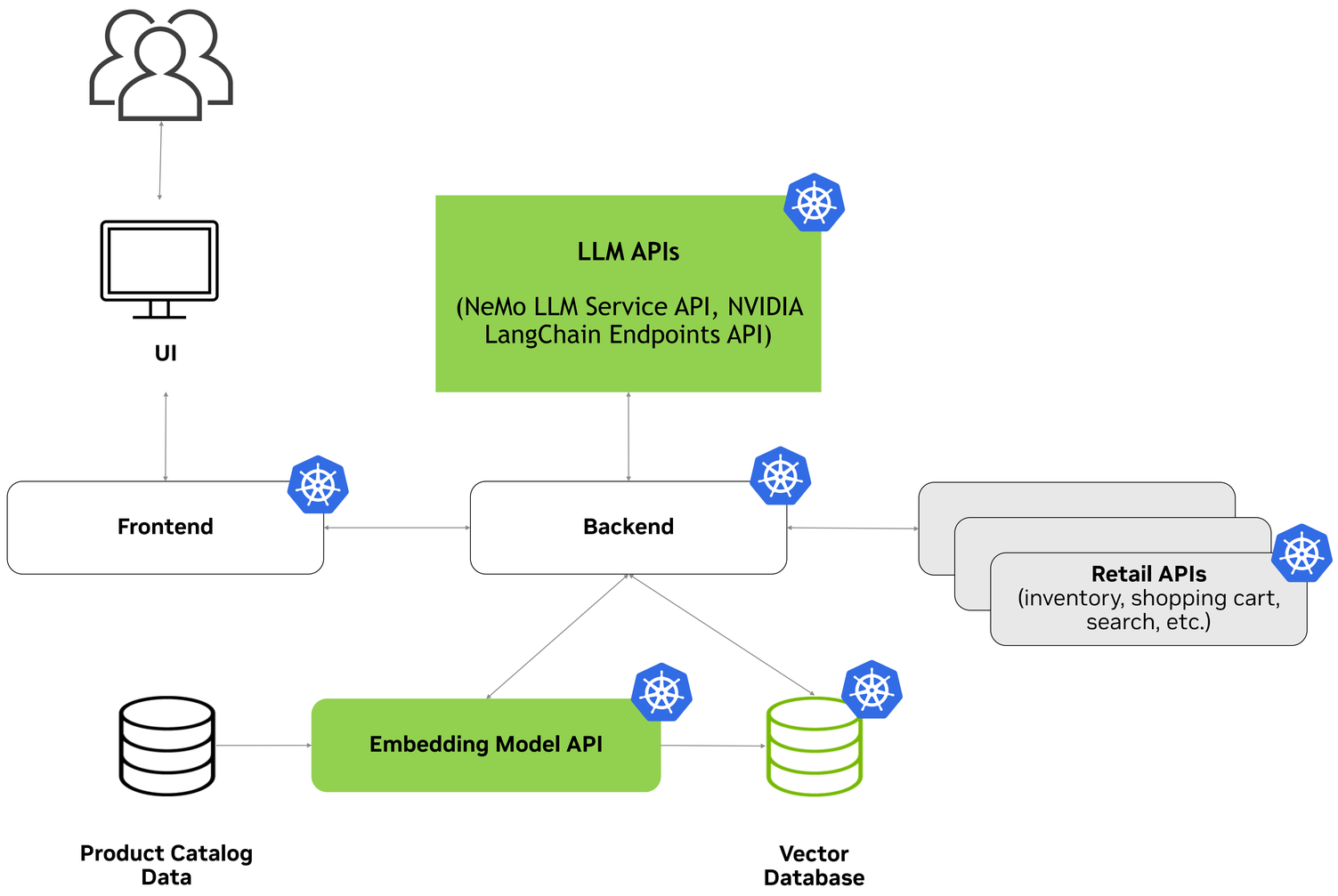
- Codebase
  - Backend – Stateless FastAPI app
  - Frontend – Simple UI in React/TypeScript with modular Components for Chatbot
  - Notebooks – Demonstrate how to use LLMs to parse reviews, summarize products from embedding, extract attributes from from descriptions, multi-modal models for extracting attributes from images, improve product titles for search, propose questions the user hasn't thought of
- Retail Frontend UI - High quality React app with license for demos

Components

- Data - defaults NVIDIA Gear Store with ability to **Bring Your Own Data** in <5 minutes
- LLM – Mixtral 8x7B via NVIDIA Langchain Endpoints (flexibility to use NeMo Inference Microservice)
- Function Calling – OpenAI GPT 3.5-turbo (exploring Mixtral 8x7B)
- Embedding – NVOLVE QA 40K via NVIDIA Langchain Endpoints (migrating to NeMo Embedding Microservice)
- Retrieval – FAISS (migrating to NeMo Retrieval Microservice, get Milvus for free)
- Frontend UI – React application in Typescript
- LLM has access to mocked APIs like **Shopping Cart**, **Search across Catalog**

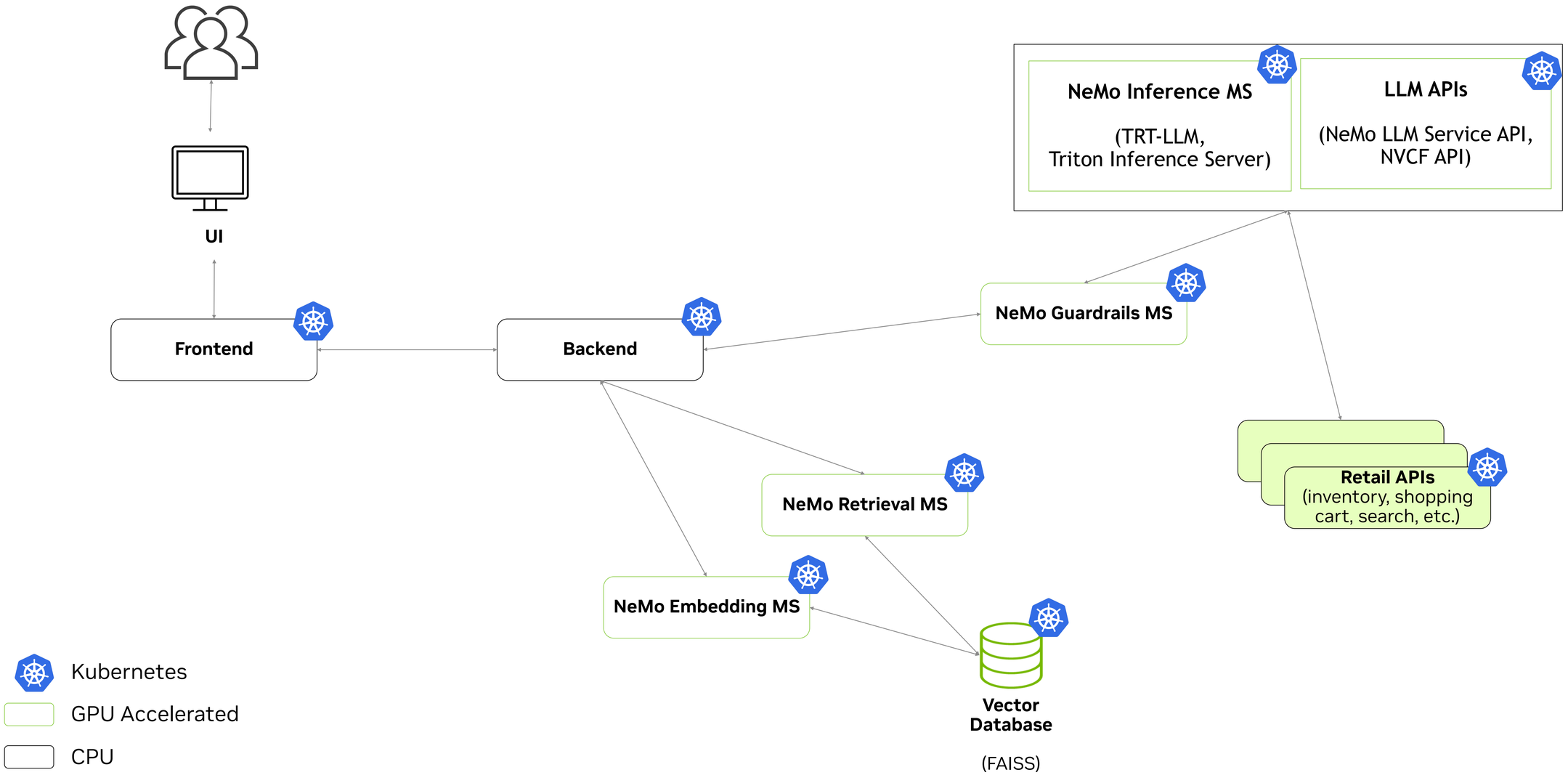
Product Advisor AI Workflow

Inference Architecture



Product Advisor AI Workflow

Inference Architecture



Multi-Modal Generative AI in Retail

NVIDIA Picasso

Foundry for custom generative AI models for visual design

Image

```
{
 "request handler" : {
   "operation": infer,
   "taskType": text2image,
        "data": "A photo of a cute
     cat with Holi colors"
},
}
```

Video

```
{
 "request handler" : {
   "operation": infer,
   "taskType": text2video,
        "data": "Purple
     ioluminescent jellyfish
     swimming in space"
},
}
```

3D

```
{
 "request handler" : {
   "operation": infer,
   "taskType": text23D,
        "data": "An ancient roman
     colosseum"
},
}
```

360 HDRi

```
{
 "request handler" : {
   "operation": infer,
   "taskType": text2360HDRi,
        "data": "Car on the road in dessert"
},
}
```

PBR Material

```
{
 "request handler" : {
   "operation": infer,
   "taskType": text2material,
        "data": "A red, blue and white
     chevron pattern  on canvas material"
},
}
```

Product Design

Design from a sketch

Source Sketch

Imagine Realistic Constructions

Visualize Variations & New Designs

Upscale a Design

Product Design

Explore design variations

Advertising

Create ad concepts

Original Kenzo Image

Image-2-Image
"Poppy flowers on ice"

Added Product Shot in Photoshop

Advertising – 2D

Generating custom backgrounds and product placement

Original Product Image

+

Product Masked Image

Prompt

Stylish brunette fashion model standing on a bridge near sunset

Prompt

Stylish blonde fashion model standing on a bridge near sunset

Advertising – 3D

Photo realistic personalization and localization

Paint Color
With Edify Materials and MDL

Environment
With Edify 360HDRi and Materials

Lighting & Weather
With Edify 360HDRi and Adobe Photoshop AI tools

NVIDIA Omniverse Unifies Cross-Functional Teams

Minimize Data Prep, Accelerate Design Reviews, Enable Full-Fidelity Collaborative 3D Workflows

1 Connect your tools & data

Ae Ps Pt
AUTODESK Alias
AUTODESK Fusion 360
AUTODESK VRED
AUTODESK Maya
AUTODESK 3ds Max
blender
CATIA
SOLIDWORKS
Houdini
KeyShot
CINEMA 4D by MAXON
Rhinoceros
SIEMENS
UNREAL ENGINE

Geometry
Variants
Materials & Textures
Environments
Animations

USD

2 See it in a single view

USD

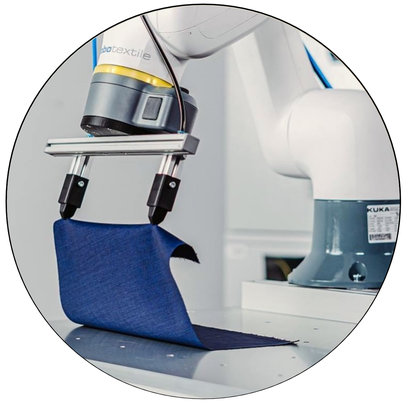
3 Develop new experiences

Viewing App
Variant Presenter
One-Click to XR App
Your Custom App

Unified Digitalization

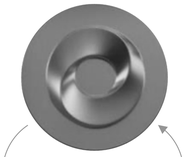
Interoperability from Concepting & Design, to Manufacturing, Marketing & Advertising

Design

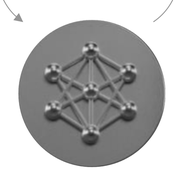


Manufacturing

Advertising



NVIDIA Omniverse



NVIDIA AI

Supercharge with Generative AI

Generative AI Virtual Sets

Minimal photography studio set with natural concrete posing blocks and soft light

A large, white marble table with natural white linen strewn across, soft day light

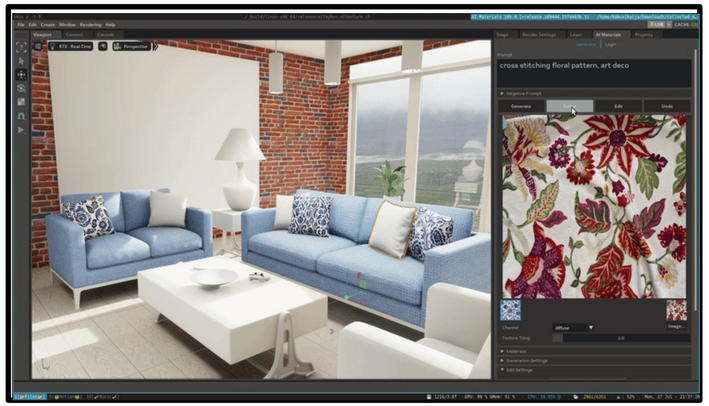
Generate Custom Materials

Edify Image with Omniverse Cloud APIs

Cross stitching floral pattern, art deco

Edify API

PBR Material



Example implementation

Generate Scene Backdrops and Lighting with 360 HDRi

Edify360 with Omniverse Cloud APIs

Snowy winter landscape with a wintery tree with snow covered branches

Edify API

8K 360 HDRi

Example output

Augmenting & Personalizing Product Pages

Data Science & Ecommerce Teams

Product Design and Marketing Campaign Creation

Marketing, Ads, Design Teams

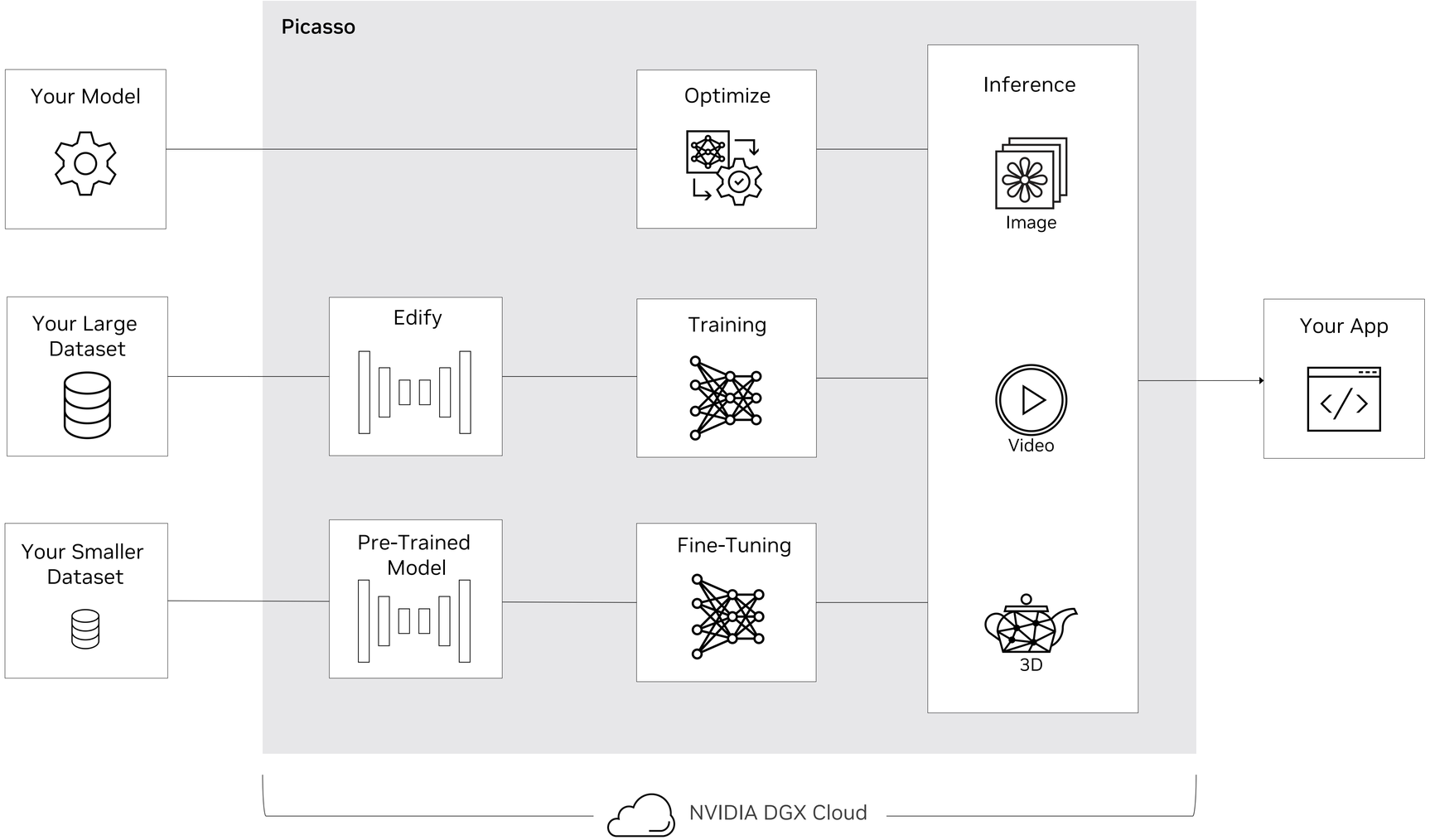
Nemo Multi-Modal

Extension of Nemo Framework for teams who are training and scaling multimodal models

- GenAI

NVIDIA Picasso

Foundry for Custom Generative AI Models for Visual Design



Omniverse + Generative AI

Supercharge 3D workflows with Generative AI

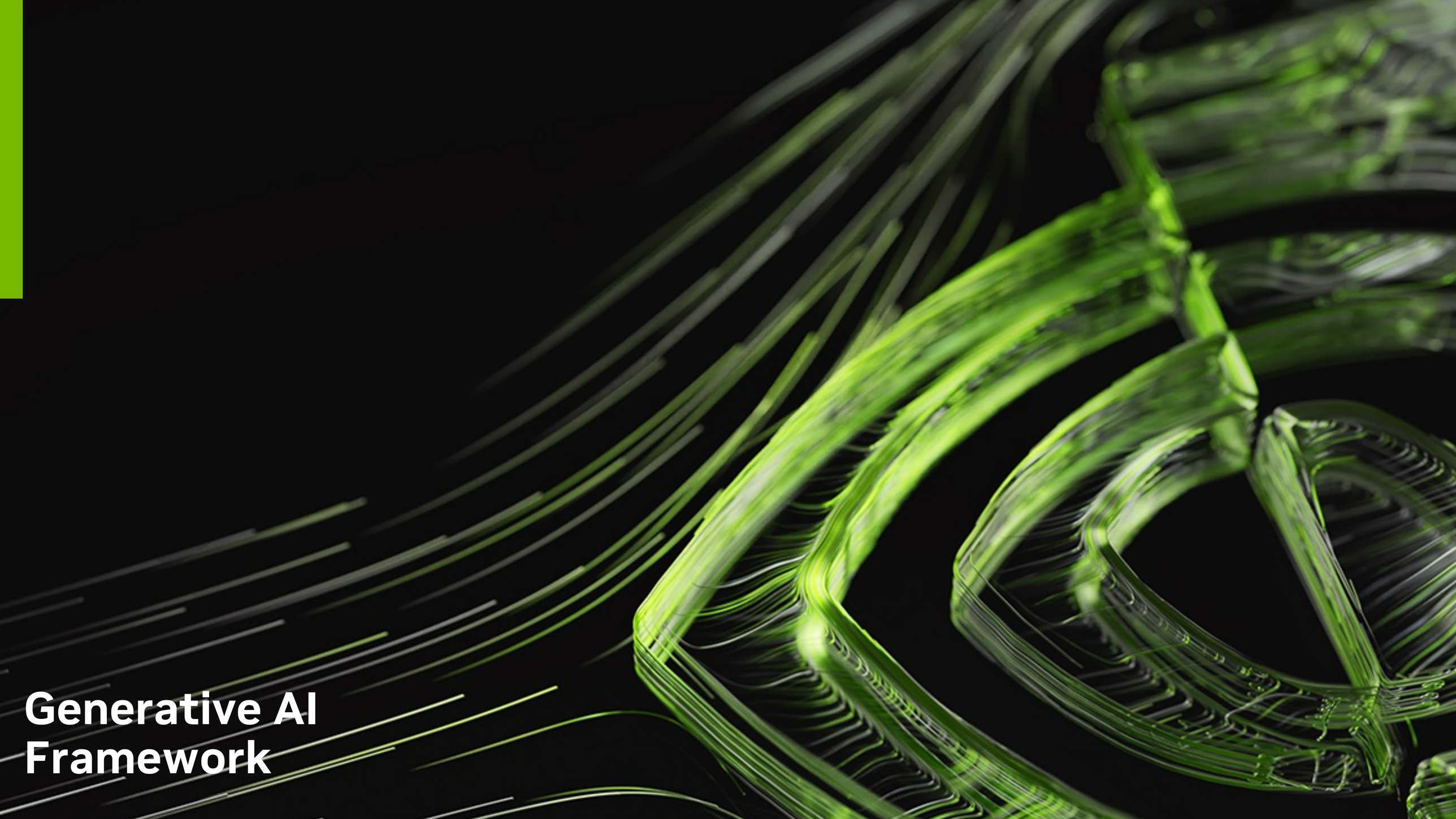
WPP + Omniverse

WPP agency, Horvath built a custom AI content engine as a service on top of Omniverse using L'Oreal products.
NVIDIA Generative AI tools are used for backgrounds and USD 3D models for products
Marketing agency Ogilvy showcased this demo at L'Oréal Digital Acceleration Summit .

Thank you from the Retail SA team!

Ward Eldred
Paul Hendricks
Mark Moyou
Dora Li
James Clohessy

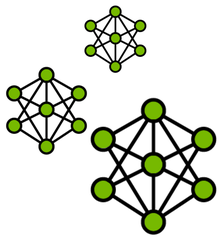
NVIDIA



Generative AI Framework

Enterprises Need an AI Foundry to Build Custom Models

An AI foundry *is a new kind of service, that brings together AI Foundation Models, state of the art tools for model customization & model customization know how*



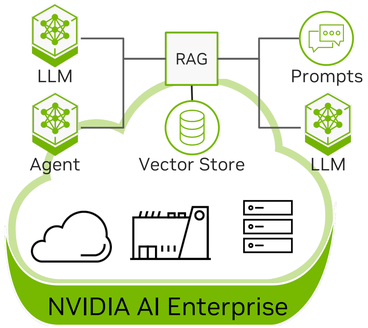
1

Experience
NVIDIA AI Foundation or Open Source Models

2

Customize with
NVIDIA NeMo Services

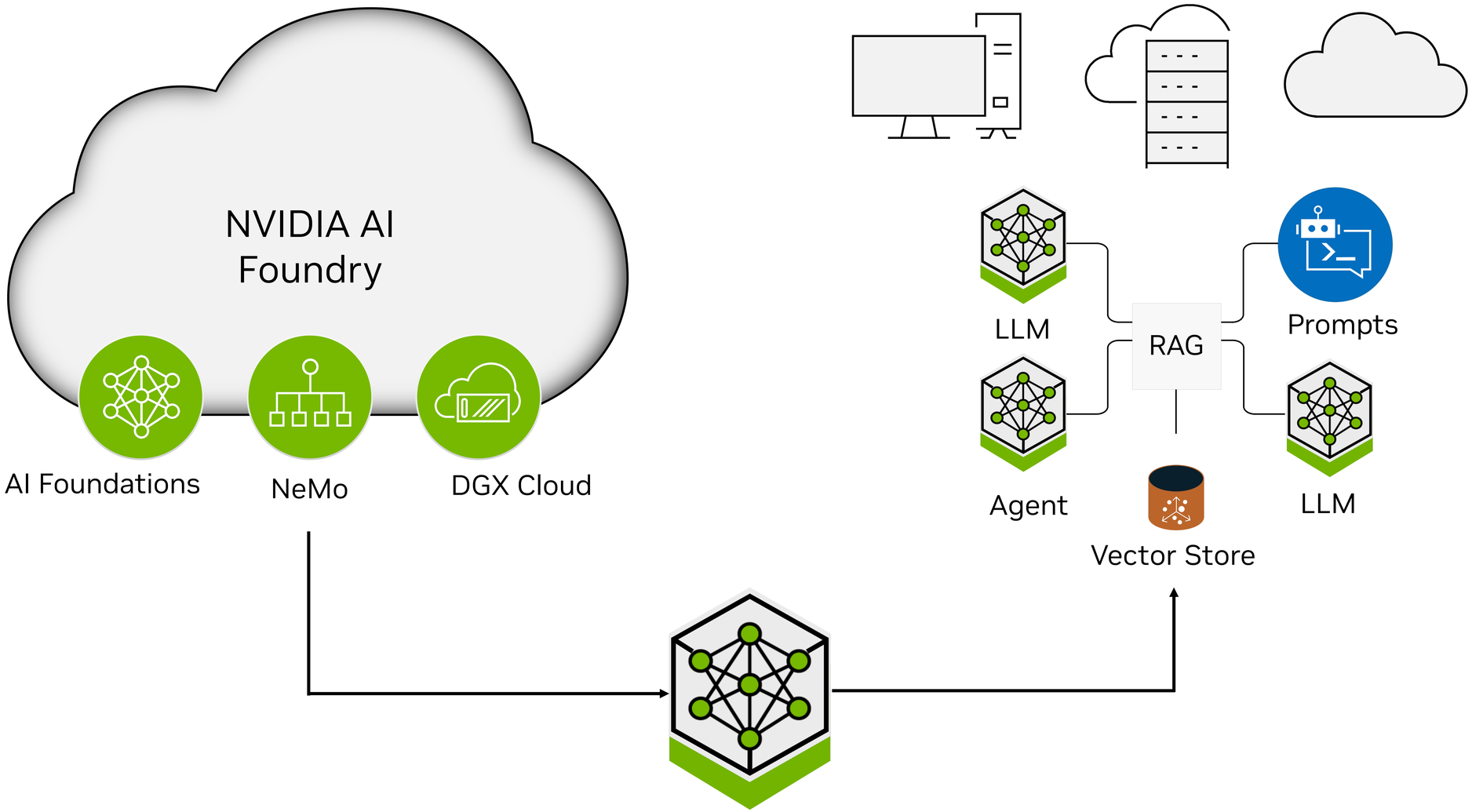
Tune on
GCP



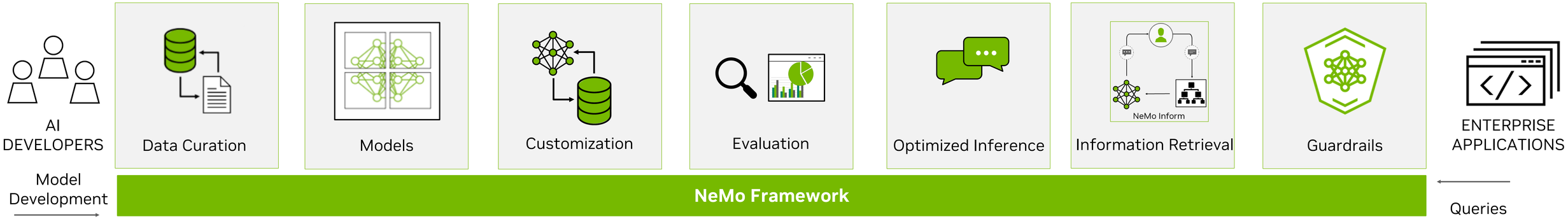
3

Deploy anywhere with
NVIDIA AI Enterprise

NVIDIA AI Foundry for Custom Enterprise-Ready LLMs



AI Foundry: The NeMo Platform



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

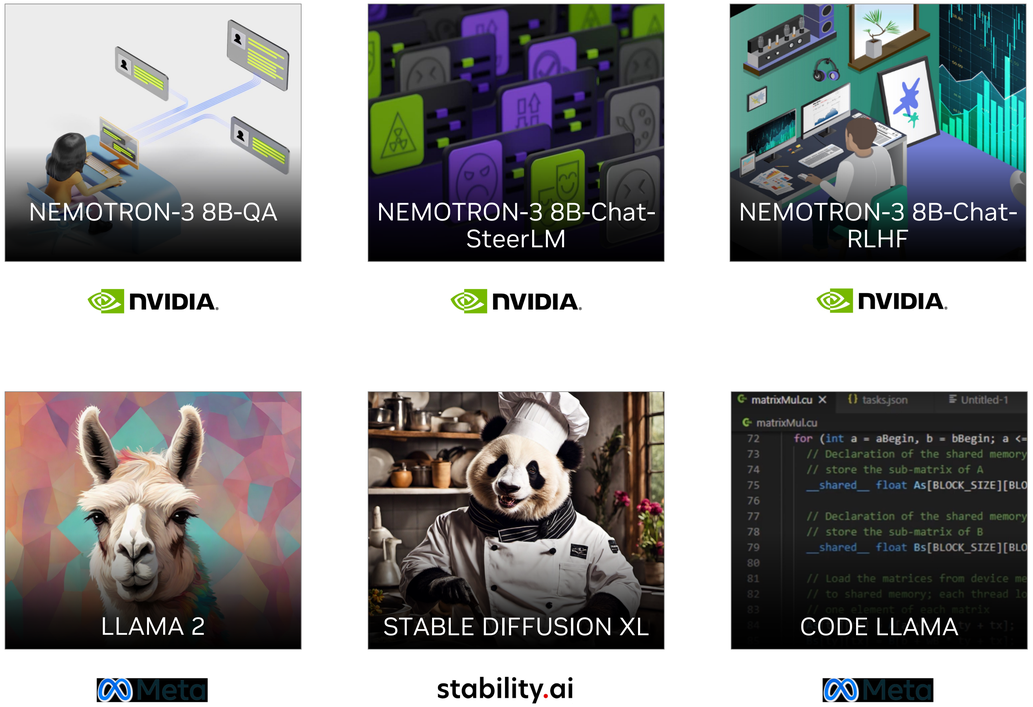
GA with NVIDIA AI Enterprise

Multi-modal available via early access now

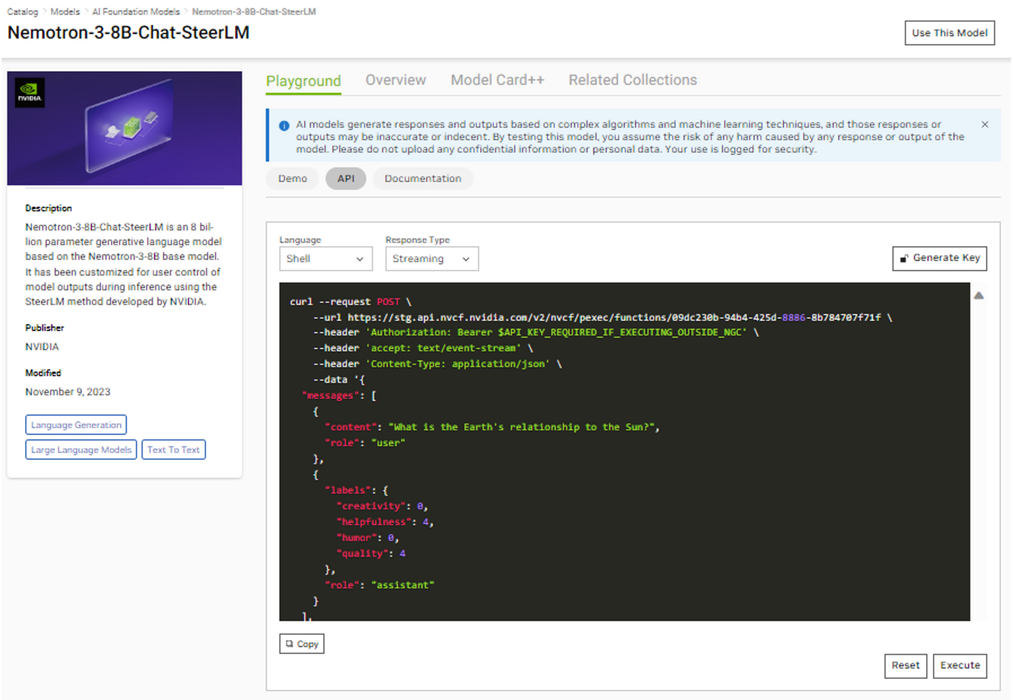
Google Cloud · Microsoft Azure · aws · ORACLE · NVIDIA DGX Cloud · DELL Technologies · Hewlett Packard Enterprise · Lenovo · SUPERMICRO

Announcing NVIDIA AI Foundation Models and Endpoints

Fast-track custom generative AI models for enterprise applications



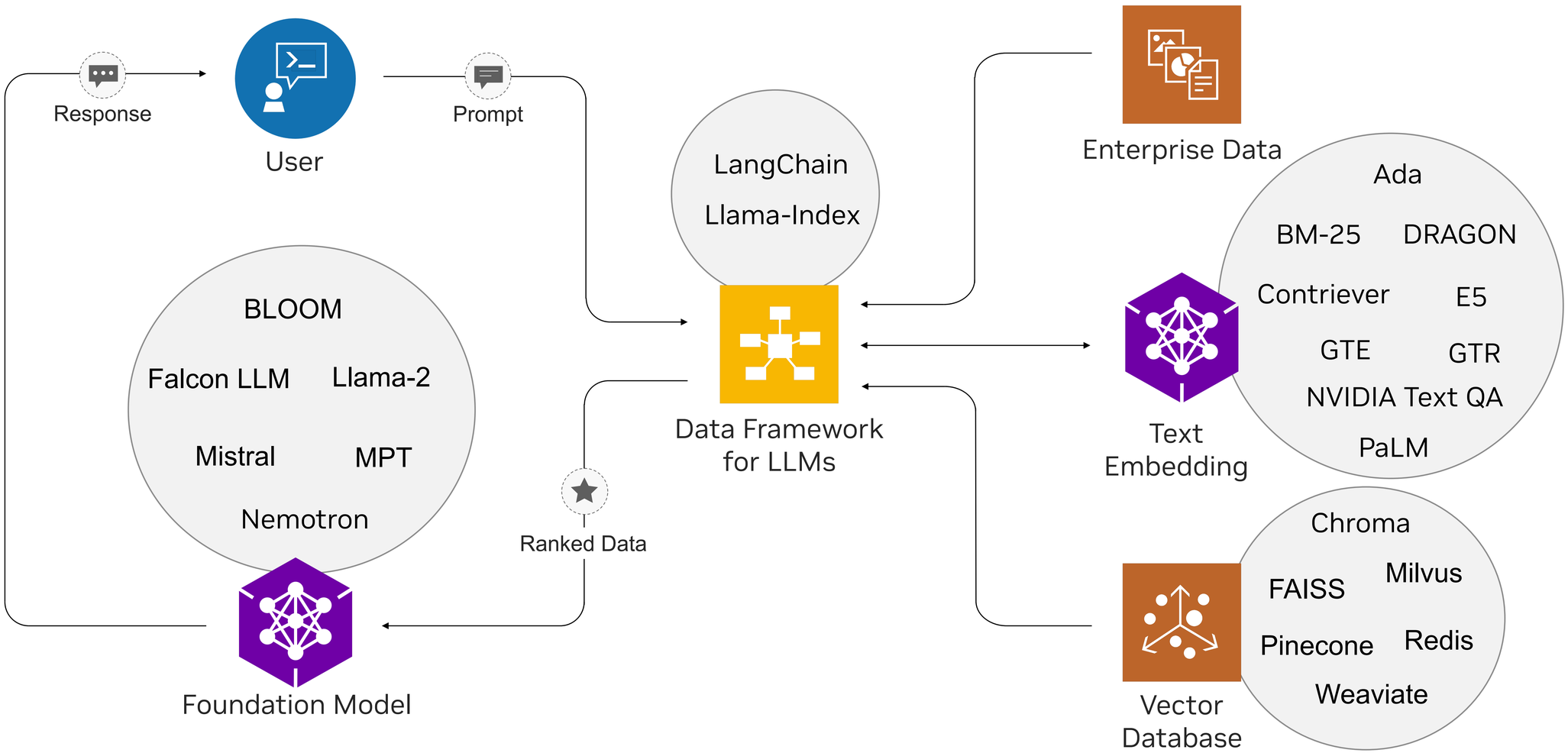
Enterprise-ready NVIDIA models
Performance-optimized community models



Experience models on NVIDIA AI stack via API endpoints

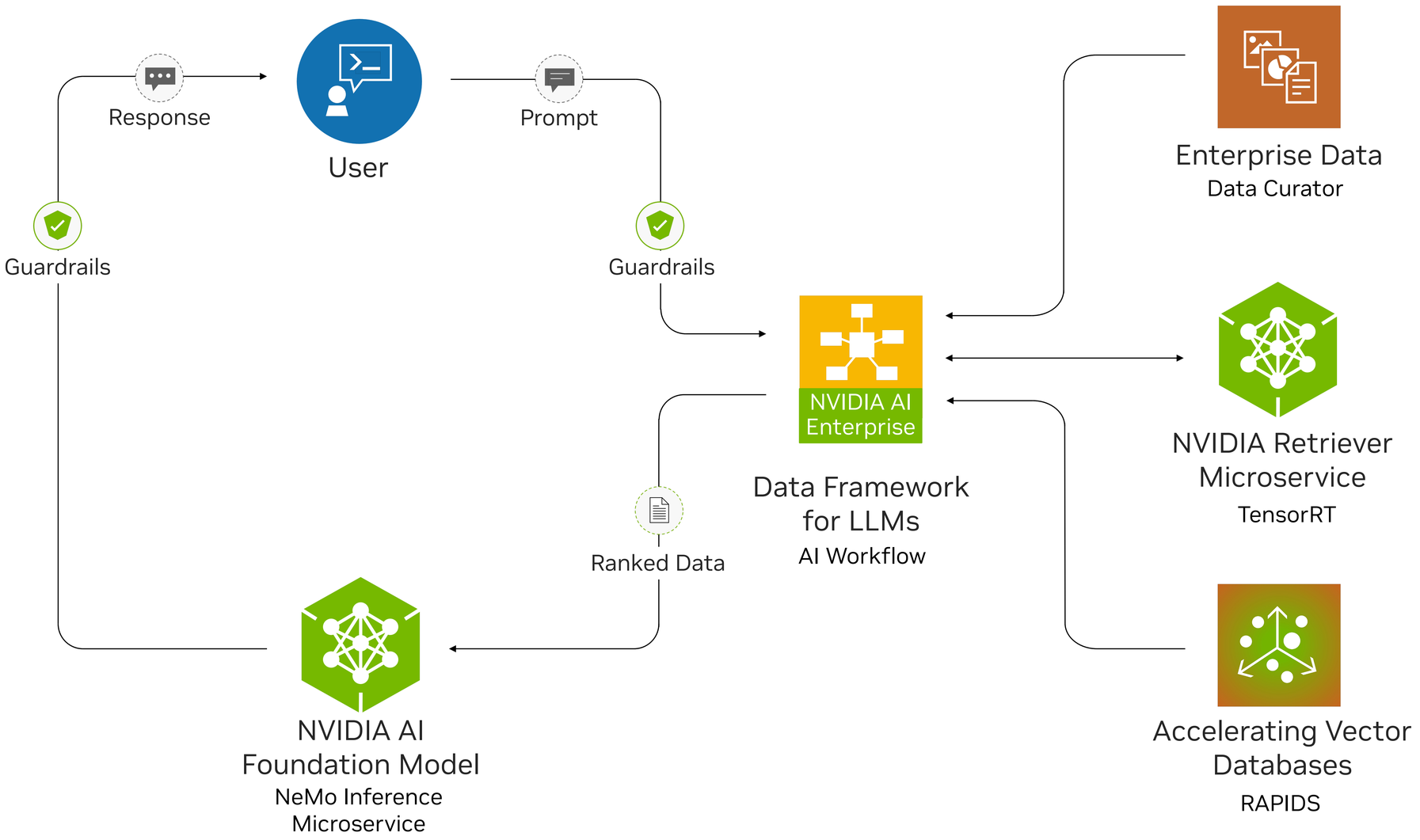
Retrieval Augmented Generation Ecosystem is Complex

Enterprises must piece together solutions from open source, update, and maintain them



NVIDIA Provides Optimized Retrieval Augmented Generation

Commercially viable, optimized embedding, reranking, and personalization to deliver highest accuracy and performance



State of the Art Accuracy and Lowest Latency for Retrieval

NVIDIA NeMo Retriever is a retrieval, embedding, and evaluation microservice for RAG

NeMo Retriever

- Models can be customized for improved accuracy
- Microservice for embeddings
- Microservice for retrieval, reranking, and personalization
- Microservice for evaluation
- Accelerated vector search

Apply for Early Access

State of the Art Accuracy

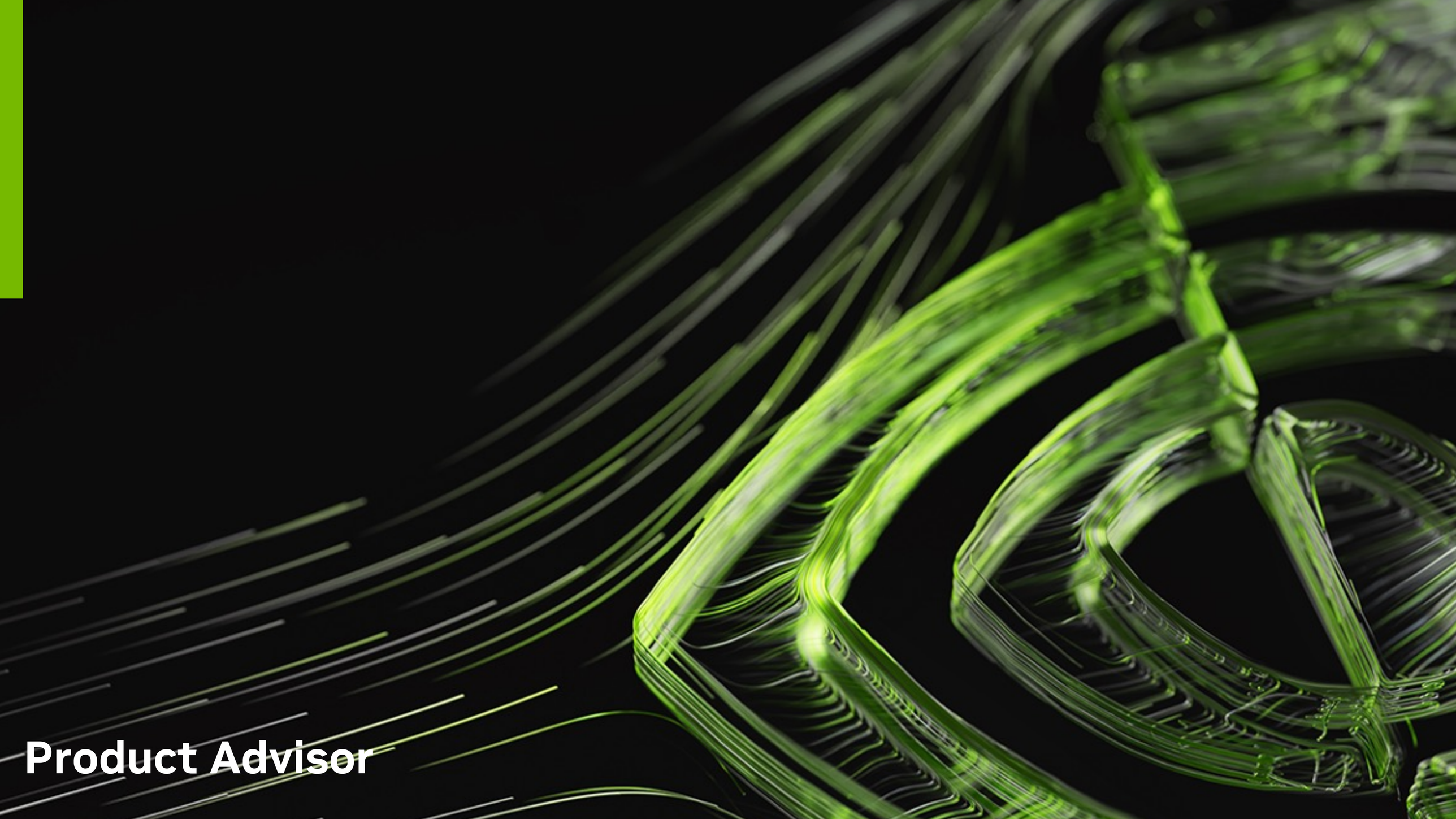
- Latest commercially ready NVIDIA AI Foundation models for embeddings, reranking, and personalization
- Latest techniques for chunking and metadata generation
- All data modalities, such as text, PDFs, images, and videos
- All data sources, such as unstructured (vector databases) and structured (SQL, Tabular)

Lowest Latency and Highest Throughput

- Inference latency and throughput optimizations for optimal user experience and TCO
- Compatible with leading GPU accelerated vector databases
- Latest techniques covering quantization, dynamic batching, and more

Production Ready with Enterprise Support

- Production ready service with API stability, security patches, and enterprise support
- Part of NeMo end-to-end LLM framework with NVIDIA AI Enterprise



Product Advisor

Product Advisor AI Workflow

Why Product Advisor?

Product Advisor Mission

- Empower Retailers with ability to quickly create LLM/RAG pipelines to answer questions about products.
- Internal and external use cases – Sephora, Shopify, Instacart, Walmart.

Traditional Search vs. Conversational LLM/RAG

- Traditional search – well-solved over 30+ years, hyper-optimized, handles simple queries.
- Performs poorly with long-tail, complex queries.
- Instead, generate and suggest products the user hasn't thought of and retrieve information about those products.
- Answer questions about a product or across several products using the product information.
- Aim to augment – not replace – traditional search.

Product Advisor

Hello! 👋 I'm your dedicated Shopping Advisor created by NVIDIA 🎮, here to answer any questions you might have and help you find anything you're looking for. What can I help you with today?

User

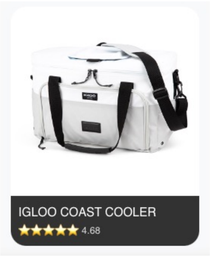
I'm going camping. What are some things I might need?

Product Advisor

When going camping, there are a few essential items you might need to bring along to ensure a safe and enjoyable trip. Here are some suggestions:

....

- Tent
- Sleeping Bag
- Lantern
- First Aid Kit
- Camping Chair
- Water Filter
- Etc.



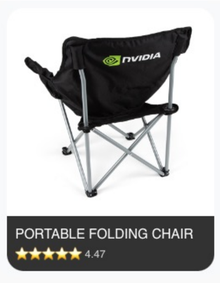
User

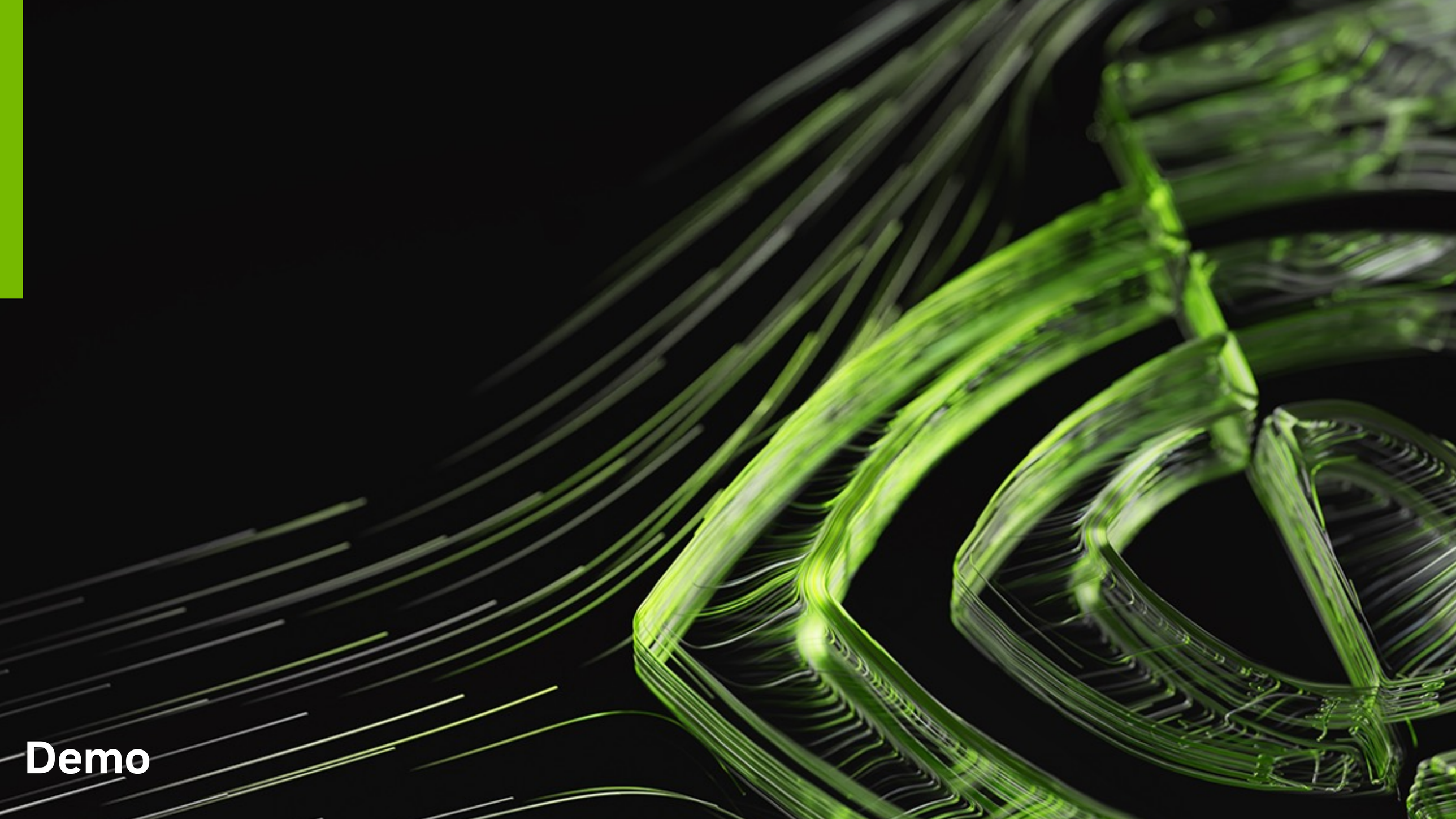
Can you recommend a chair?

"camping chair

Product Advisor

The Portable Folding Chair is a great option for camping. It has a steel frame that is strong and durable, and it can support a weight capacity of up to 300 lbs. The chair also includes a mesh gear pouch for storing small items, and it is easy to fold up and transport.





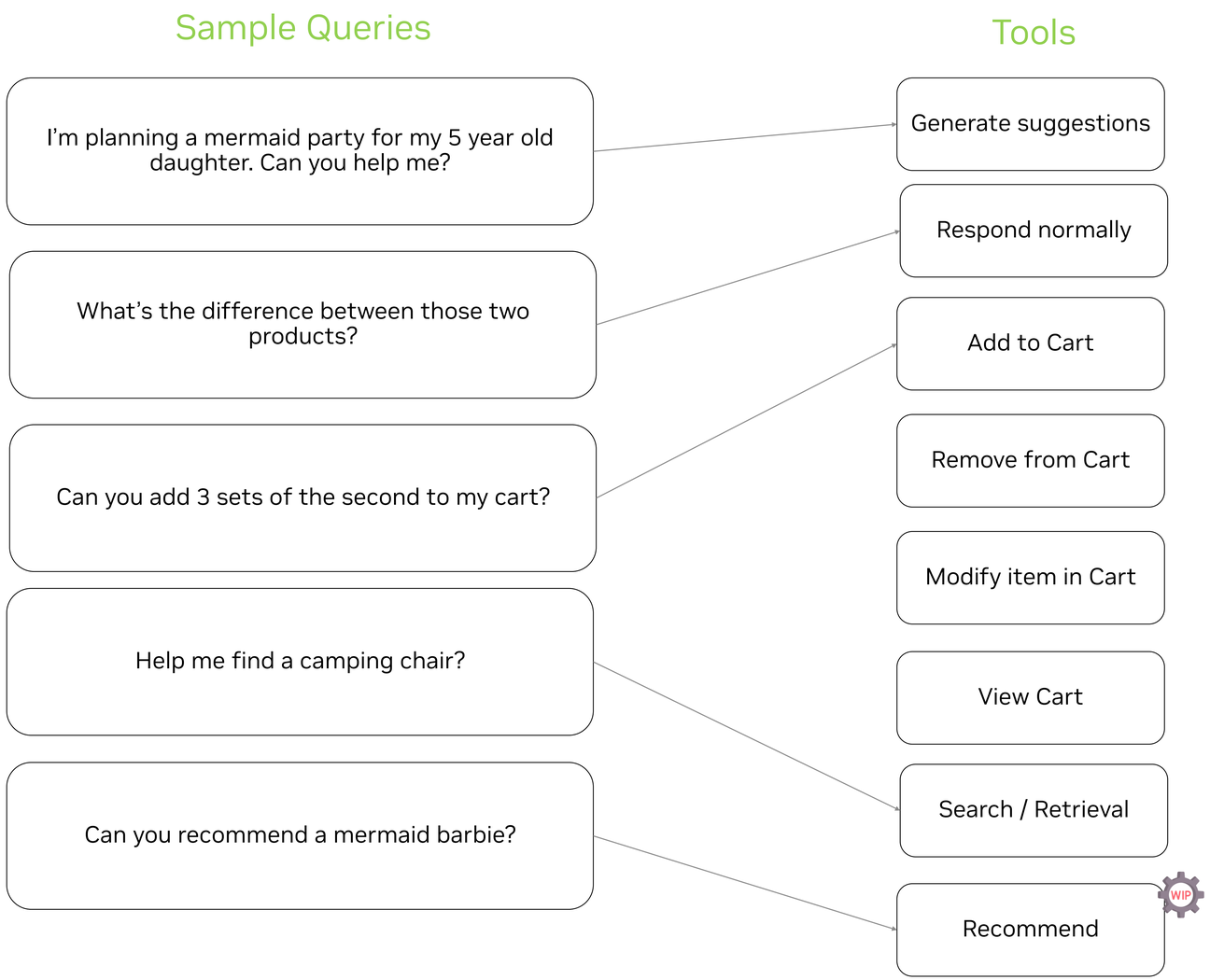
Demo

Product Advisor AI Workflow

LLM Decision Making

Empowering LLMs with APIs

- The long pole in the tent is handling the myriad nuances of natural conversation.
- Can't do retrieval every time.
- Creating a prompt to handle every scenario is too wieldy and rigid.
- Retailers have APIs for product information, pricing, inventory, shopping cart, search, recommendation.
- Allowing LLMs to access Tools & APIs and decide which Tool/API to use allows for flexibility and delivers qualitatively promising results.



Product Advisor AI Workflow

Zero to Development in <5 minutes on Retailer's data

Assets

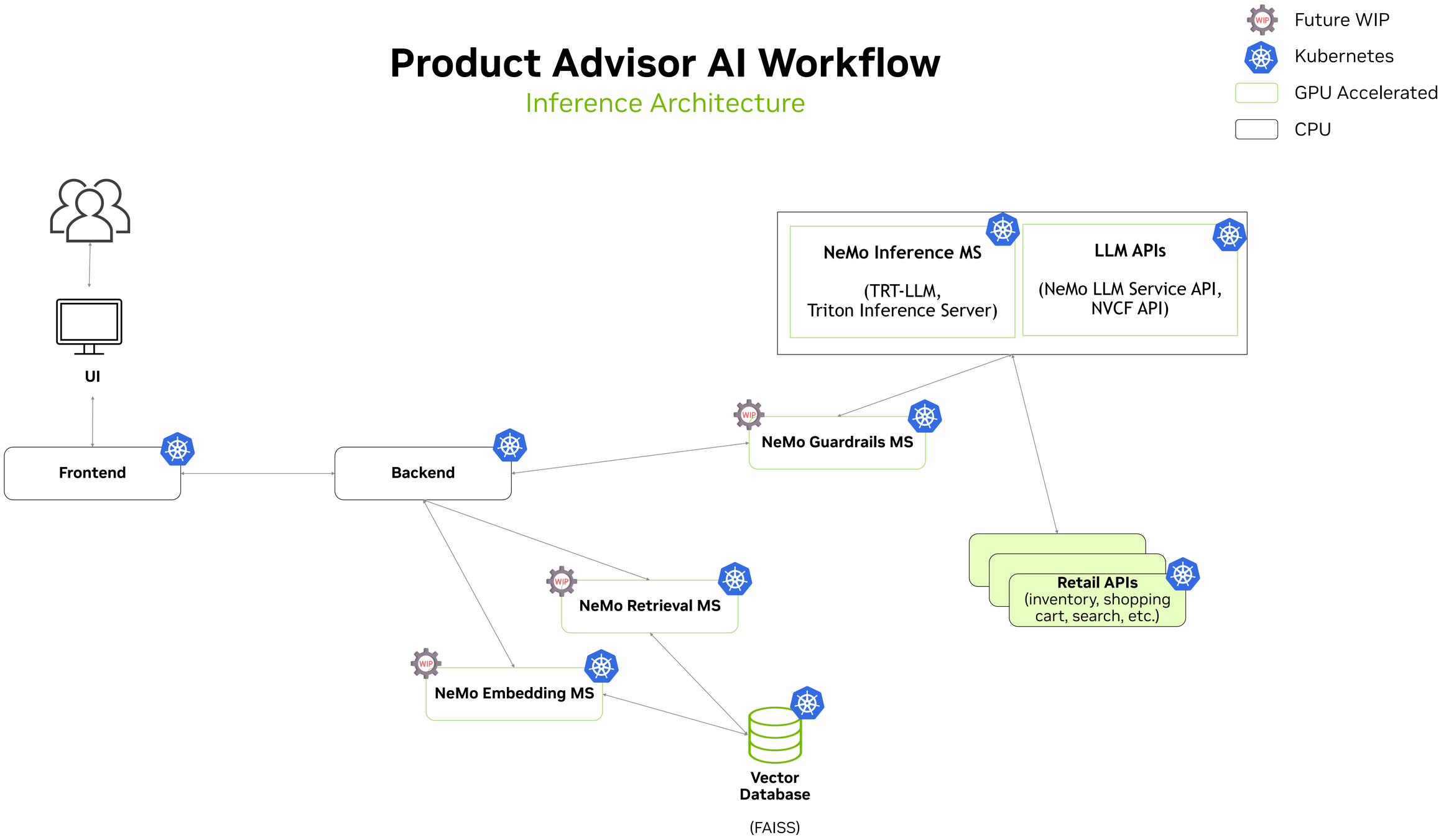
- Codebase
  - Backend – Stateless FastAPI app
  - Frontend – Simple UI in React/TypeScript with modular Components for Chatbot
  - Notebooks – Demonstrate how to use LLMs to parse reviews, summarize products from embedding, extract attributes from from descriptions, multi-modal models for extracting attributes from images, improve product titles for search, propose questions the user hasn't thought of
- Retail Frontend UI - High quality React app with license for demos

Components

- Data - defaults NVIDIA Gear Store with ability to **Bring Your Own Data** in <5 minutes
- LLM – Mixtral 8x7B via NVIDIA Langchain Endpoints (flexibility to use NeMo Inference Microservice)
- Function Calling – OpenAI GPT 3.5-turbo (exploring Mixtral 8x7B)
- Embedding – NVOLVE QA 40K via NVIDIA Langchain Endpoints (migrating to NeMo Embedding Microservice)
- Retrieval – FAISS (migrating to NeMo Retrieval Microservice, get Milvus for free)
- Frontend UI – React application in Typescript
- LLM has access to mocked APIs like **Shopping Cart**, **Search across Catalog**

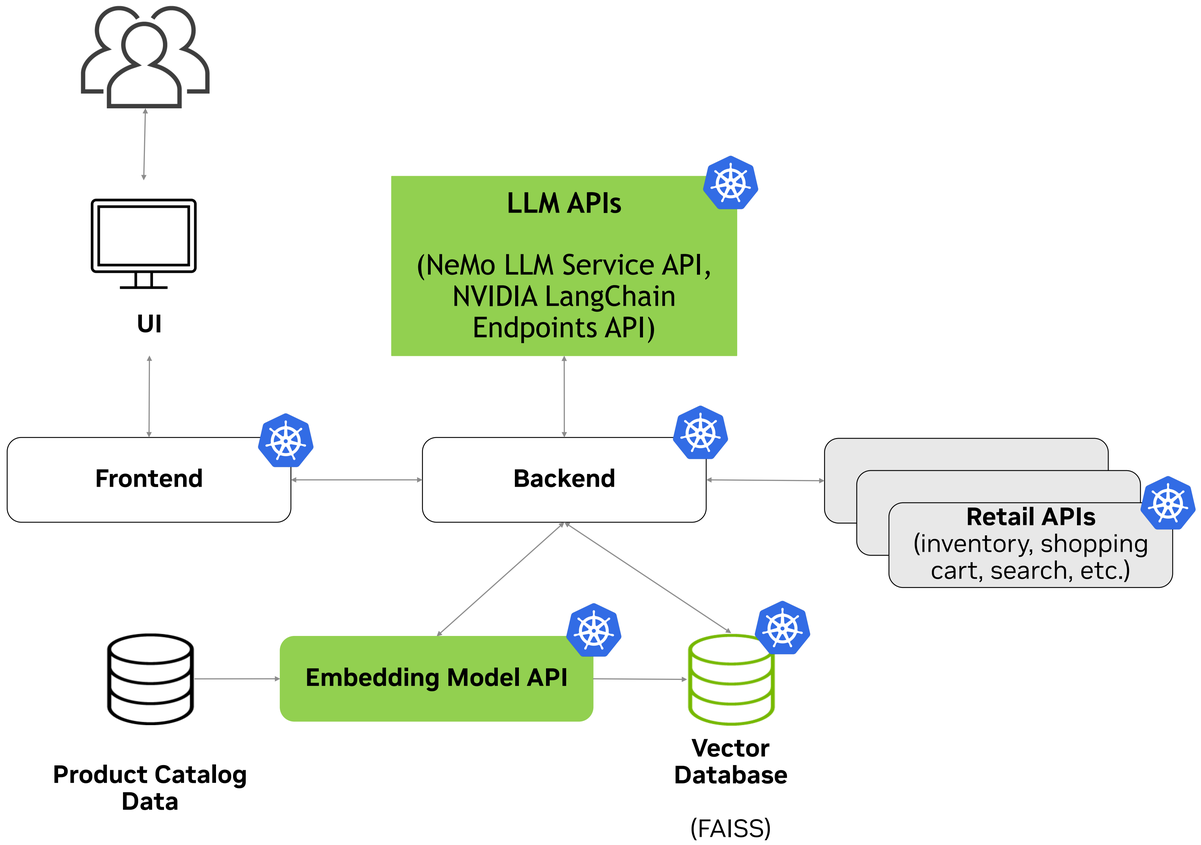
Product Advisor AI Workflow

Inference Architecture



Product Advisor AI Workflow

Inference Architecture



Product Advisor AI Workflow

Next Steps

Next Steps

- Identify Retail LHAs to collaborate with
  - This workflow demonstrates Art of the Possible & likely not “drop into production ready”
  - Will require customizing
- Review with Jacob & ChatLabs team
- Incorporate NeMo Retrieval Microservice, NeMo Embedding Microservice, & NeMo Inference Microservice
- Finish notebooks
- Explore Mixtral 8x7B for function calling to replace OpenAI
- Publish to open-source
- Explore conversational multi-modality – ability to upload image, ask questions, receive pairing suggestions, inpainting

Retail Imagine Copilot

Retail AI Workflow

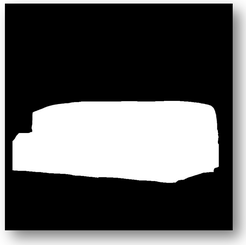
Picasso Inpainting API

Original image

+

Segment Anything

Original image



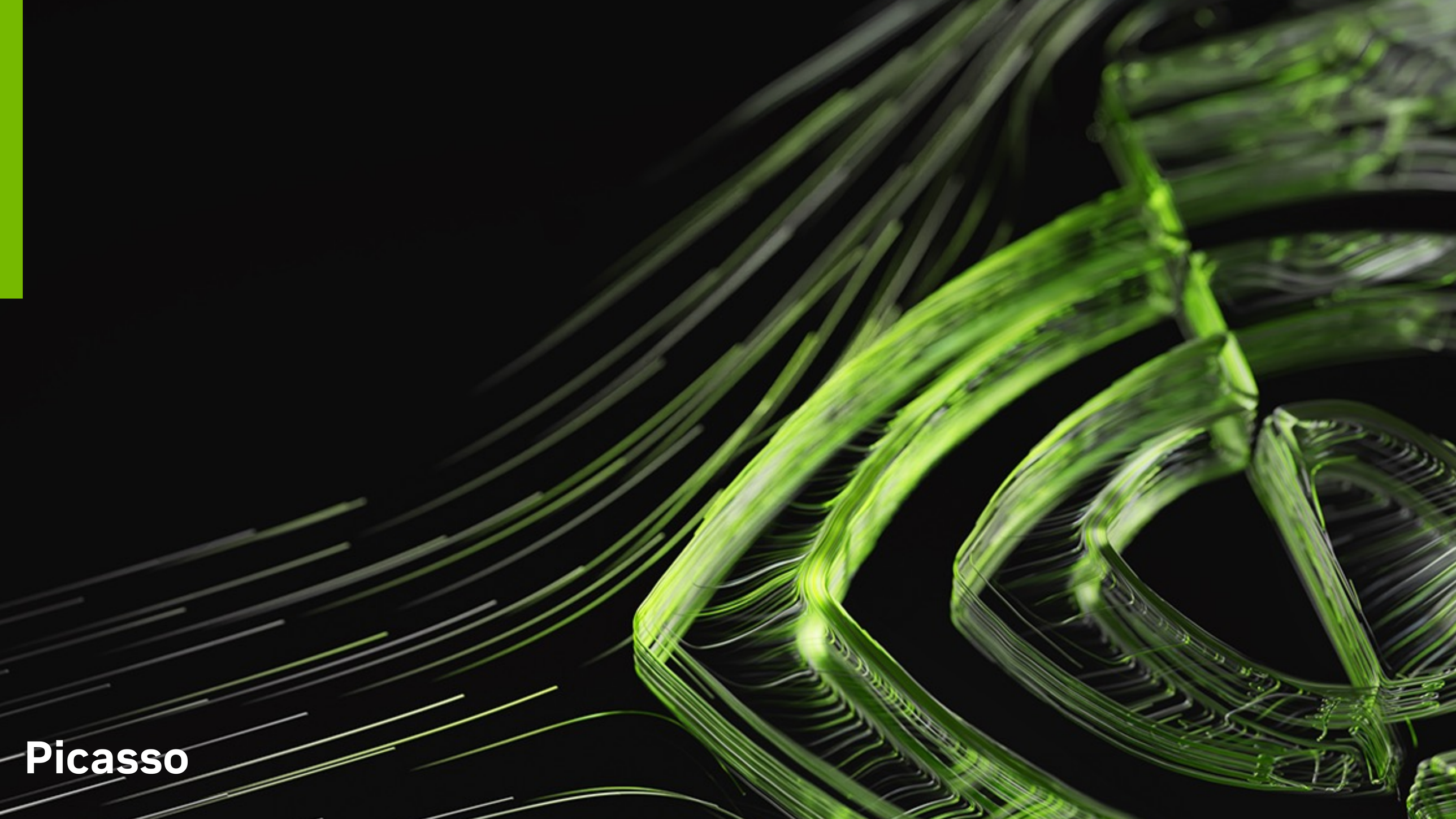
Masked Image

+

Prompt: "Classic brown couch, modern style, metal legs"

Outline (TBD)

- Overview of use cases in large categories
  - Data science & processing
  - Omnichannel -> genAI
    - Text use cases
      - Personalization
      - Product description
      - Chatbot
    - Nemo LLM Platform
    - Image
      - Marketing asset, custom placement
      - Creating moodboards
      - Mockups
    - Picasso
  - Omniverse -> 3D workflows + genAI
    - Use cases – marketing asset generation
    - Product review/design
    - Materials simulation
    - genAI integration
  - Inference ?



Picasso

Generative AI for Retail Must Be Multi-Modal

NVIDIA Picasso Foundry for generating custom Visual Generative AI

Image

Video

3D

Materials

The image portrays a picturesque Italian outdoor coffee house, exuding charm and tranquility under the warm Mediterranean sun. A rustic, wrought-iron table takes center stage, adorned with an elegant stemmed glass brimming with effervescent Coca-Cola, beautifully reflecting the sunlight and enticing the viewer. Beside the glass, glistening ice cubes add a refreshing touch to the beverage, promising delightful satisfaction.Surrounding the table, vibrant flowerpots overflow with blooming bougainvillea and fragrant lavender, infusing the scene with bursts of color and sweet aroma. Cobblestone streets lead into the distance, inviting exploration and discovery.A crimson-and-white striped umbrella elegantly shelters the table, providing a welcoming shade for visitors to enjoy their drinks comfortably despite the warmth of the day. The coffee house itself boasts timeless Italian architecture, characterized by graceful arches, intricate detailing, and warm earthy tones. The atmosphere is enlivened by the cheerful chatter of locals and tourists, creating a sense of camaraderie and merriment.

Tool Used: PIcasso txt2img

Text-to-Image

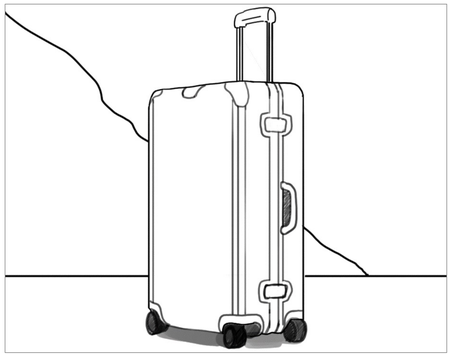
→

Prompt:

Photo of a coca cola glass bottle with a cheeseburger at a campfire, coca cola, glass bottle, coke, golden hour, atmosphere, gorgeous, 8k

NVIDIA CONFIDENTIAL / DO NOT DISTRIBUTE

Sketch-to-Image



Original Image: Sketched Out RIMOWA Luggage

→

Prompt:

A modern golden suitcase with mountains and lake in the background, Gorgeous, Suitcase, Travel, Modern, Sleek, Luxury, High End

Image-to-Image

Original image:
Miss Dior Perfume

→

Prompt:

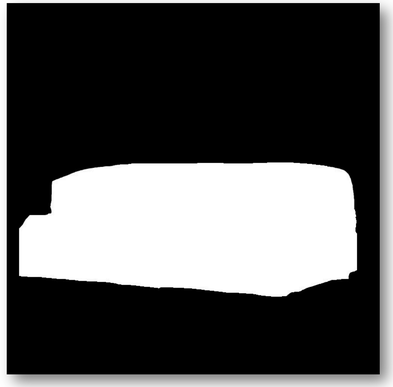
A beautiful pink bottle of perfume in a field of wildflowers, Depth of Field, Gorgeous, Luxury, High End, High Quality

Image-to-Image

"Would a gray or green couch look good in my living room?"

Original image

+



Original Masked Image

Picasso Inpainting API

Prompt: "Classic brown couch, modern style, metal legs"

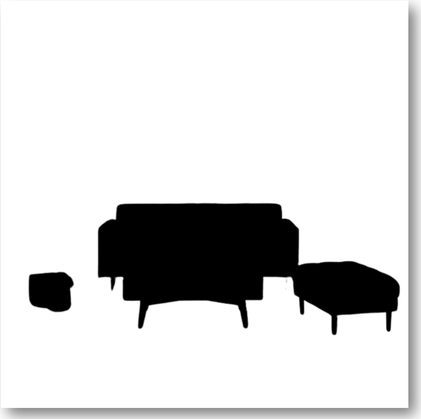
Prompt: "Tufted green couch with wood legs, modern style"

Generating New Product Placements

Take a product and place it in a new environment using REPLACEMENT API

Original Product Image

+



Product Masked Image

+

Prompt

Traditional house with dark wood floors, cream colored walls, white brick fireplace, mantle matching dark wide plank wooden floors, with a olive wood bookcase, fancy aged wooden chandialier

Generating New Product Placements

Take a product and place it in a new environment using REPLACEMENT API

Original Product Image

+

Product Masked Image

Prompt

Warm new York sunset near Brooklyn bridge

Prompt

Warm outdoor italian summer with cobblestones

Image-to-Image

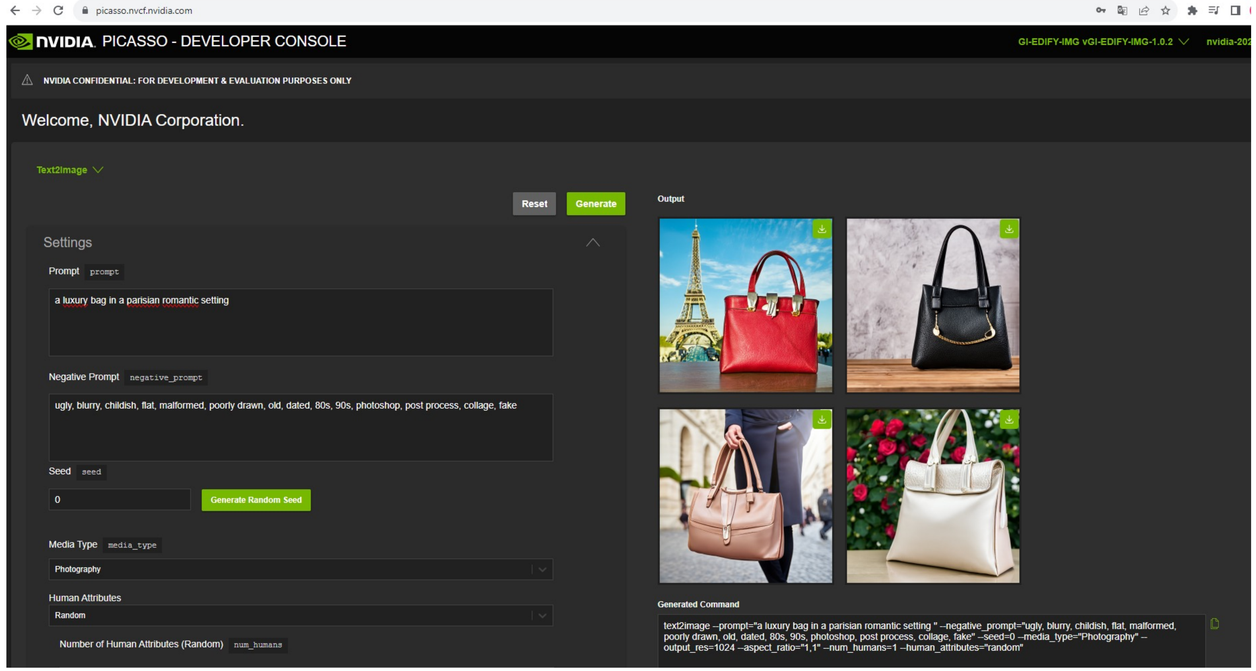
Explore Product Designs

Original Hennessy Bottle Image

Image-to-Image Rapid Design Iteration

Picasso

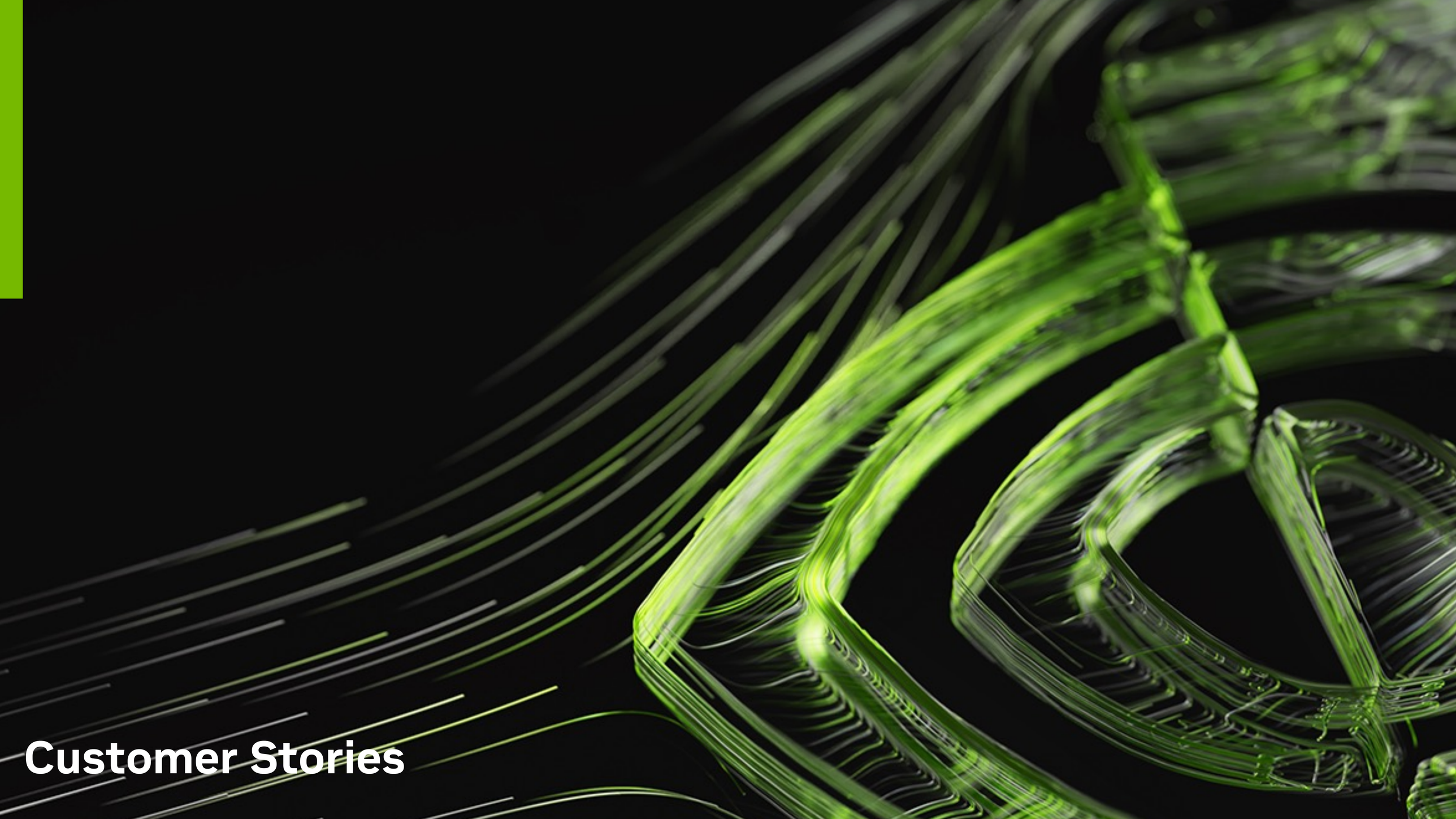
Demo



Text–to-Image

Text–to-Material

Omniverse + GenAI for Product Design, Marketing Asset



Customer Stories

Enterprises Using Custom Generative AI

Leaders in the market are adopting NeMo to power their Generative AI initiatives

IT Service Ticket ChatBot

servicenow

Custom generative AI for automatic issue resolution, knowledge-base article generation based on customer case summaries, and chat summarization for faster hand-off, resolution and wrap-up.

Interactive AI Assistant

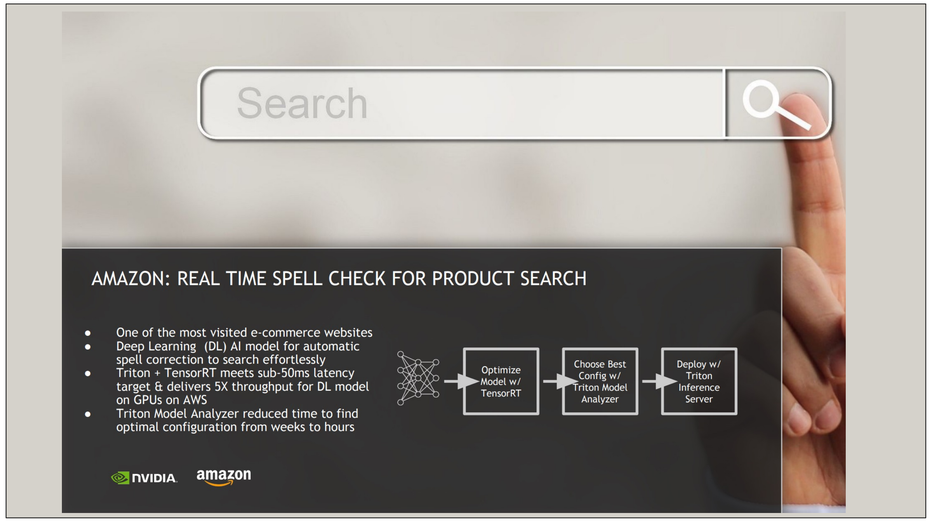
Interactive AI assistant to search and answer product related questions. Generate product descriptions

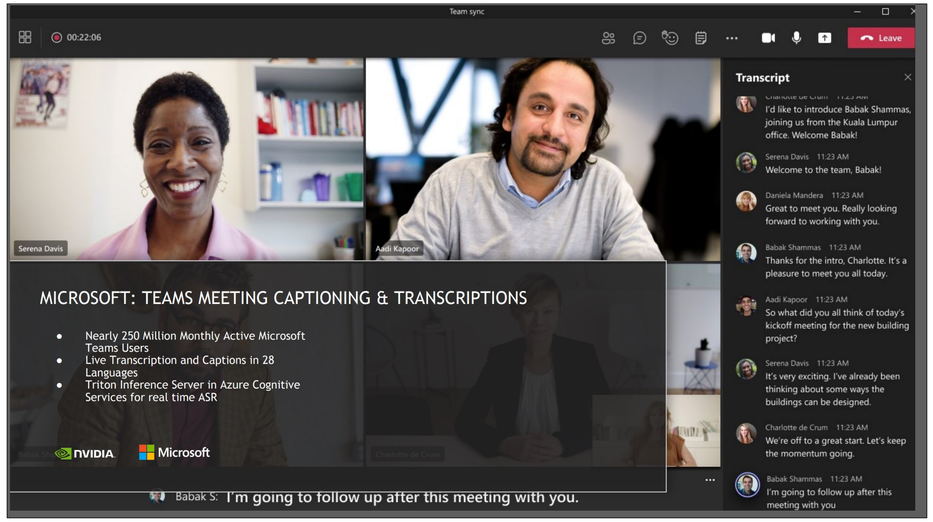
AI Voice Assistant

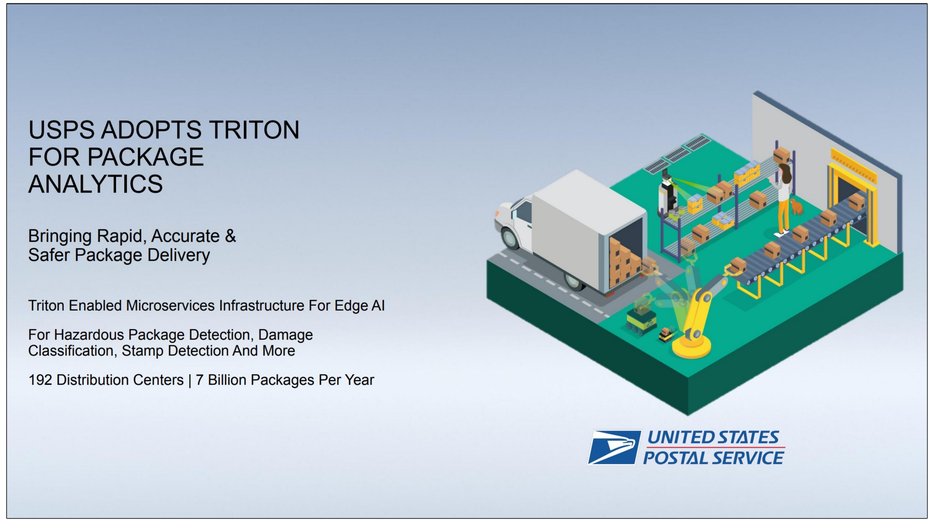
kt

Virtual assistant speaks conversationally in Korean. It can control TVs, offer traffic updates and complete other home-assistance tasks from voice commands.

Triton Inference Server







NVIDIA