Accelerating Enterprise: Tools and Techniques for Next-Generation AI Deployment

Mahan Salehi, Product Manager, NVIDIA
Nave Algarici, Product Manager, NVIDIA

GTC, March 2024

Enterprise are on the Generative AI Journey

2022

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.

2023

Experimentation

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

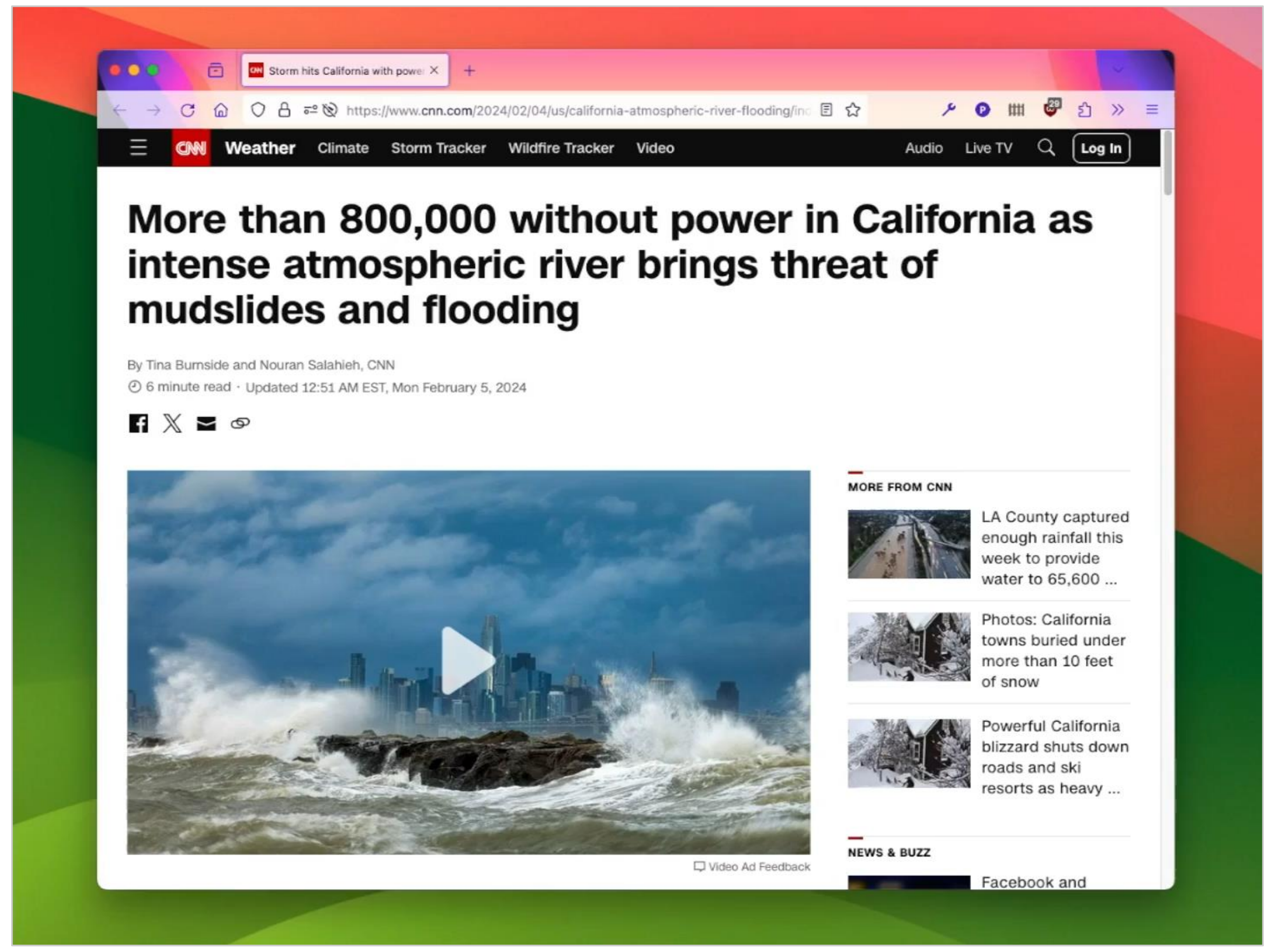
2024

Production
(aka "Inference")

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

Activate the Potential Within Your Organization

Across Every Industry and Job Function


Storm hits California with powe
https://www.cnn.com/2024/02/04/us/california-atmospheric-river-flooding/in
CNN
Weather
Climate
Storm Tracker
Wildfire Tracker
Video
Audio
Live TV
Log In
Video Ad Feedback

More than 800,000 without power in California as intense atmospheric river brings threat of mudslides and flooding

By Tina Burnside and Nouran Salahieh, CNN
6 minute read · Updated 12:51 AM EST, Mon February 5, 2024

MORE FROM CNN

LA County captured enough rainfall this week to provide water to 65,600 ...

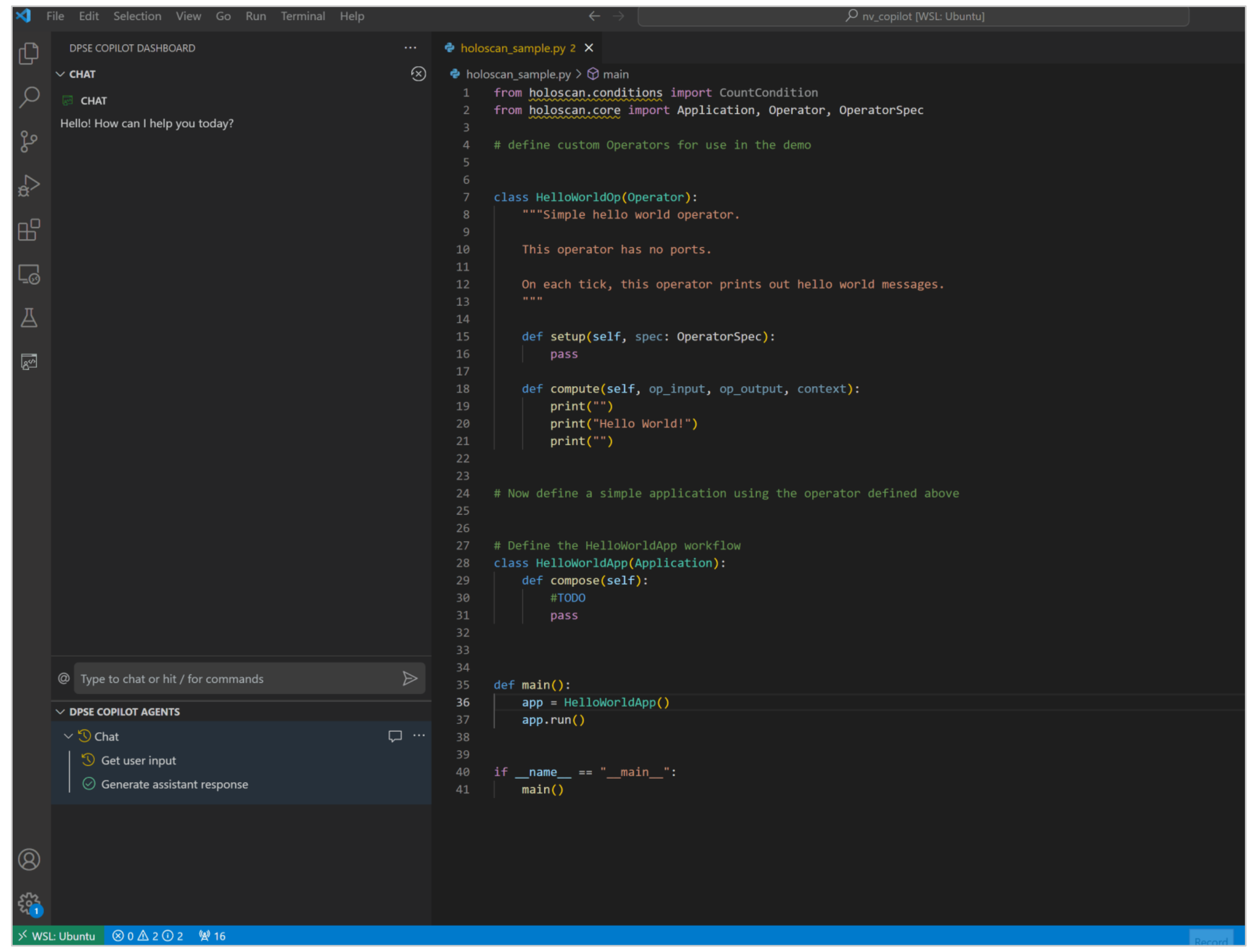
Photos: California towns buried under more than 10 feet of snow

Powerful California blizzard shuts down roads and ski resorts as heavy ...

NEWS & BUZZ

Facebook and

Kinetica Telco Copilot



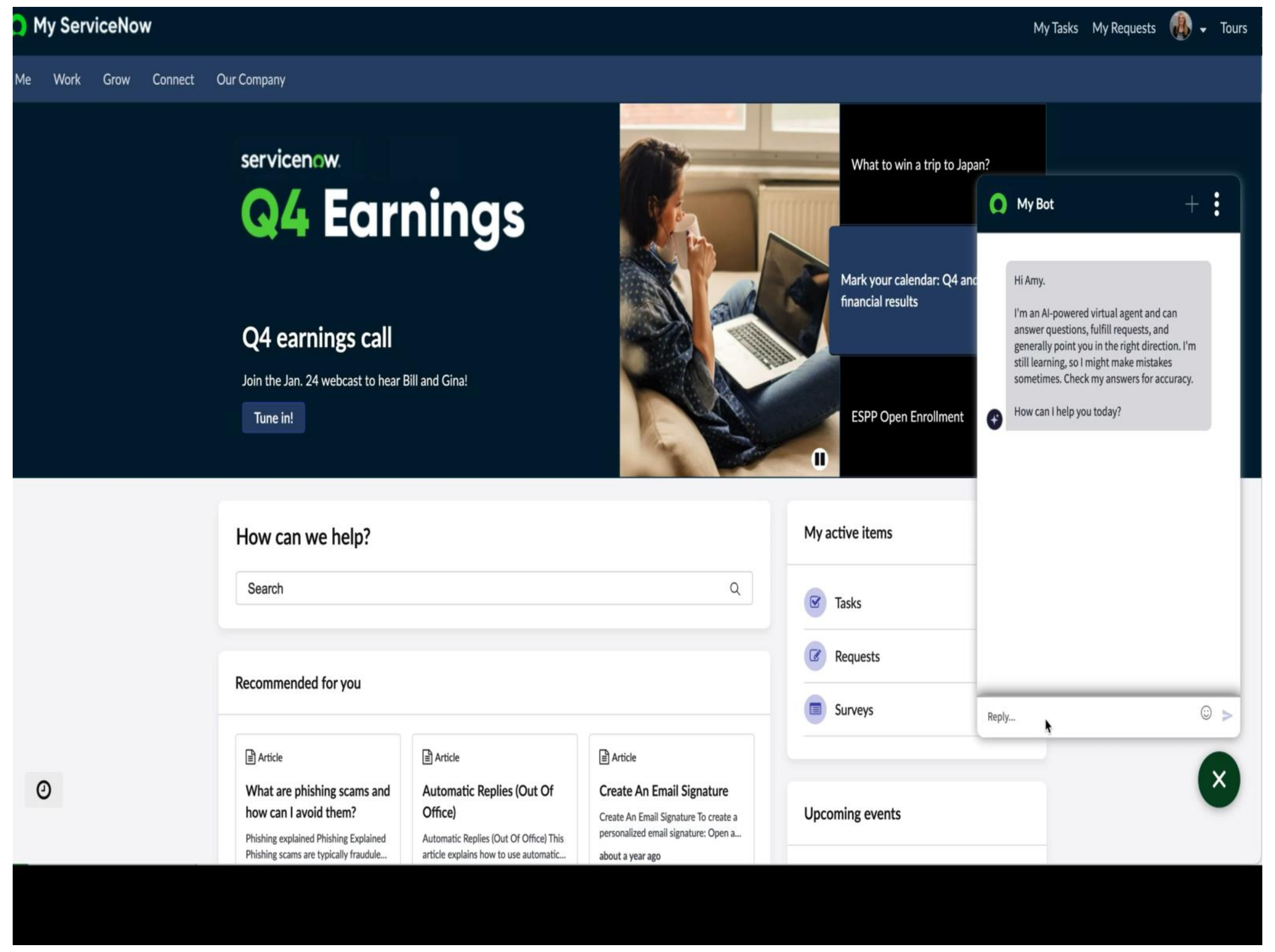
CHAT

Hello! How can I help you today?

- Chat
  - Get user input
  - Generate assistant response

```python
from holoscan.conditions import CountCondition
from holoscan.core import Application, Operator, OperatorSpec

# define custom Operators for use in the demo


class HelloWorldOp(Operator):
    """Simple hello world operator.

    This operator has no ports.

    On each tick, this operator prints out hello world messages.
    """

    def setup(self, spec: OperatorSpec):
        pass

    def compute(self, op_input, op_output, context):
        print("")
        print("Hello World!")
        print("")


# Now define a simple application using the operator defined above


# Define the HelloWorldApp workflow
class HelloWorldApp(Application):
    def compose(self):
        #TODO
        pass



def main():
    app = HelloWorldApp()
    app.run()


if __name__ == "__main__":
    main()
```

Coding Copilot



Q4 Earnings

Q4 earnings call

Join the Jan. 24 webcast to hear Bill and Gina!

Hi Amy.

I'm an AI-powered virtual agent and can answer questions, fulfill requests, and generally point you in the right direction. I'm still learning, so I might make mistakes sometimes. Check my answers for accuracy.

How can I help you today?

How can we help?

Search

My active items

- Tasks
- Requests
- Surveys

Recommended for you

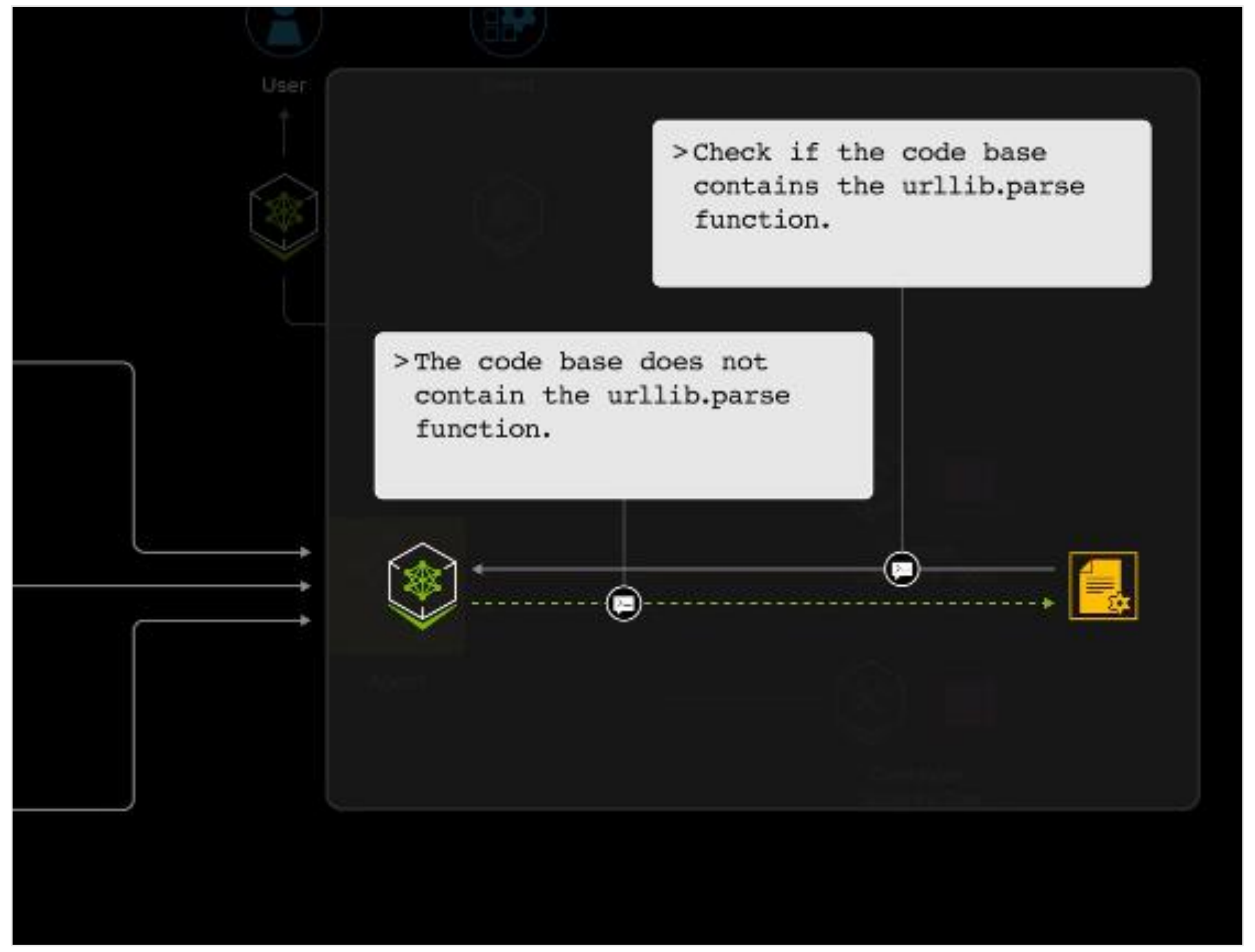
Article: What are phishing scams and how can I avoid them? Phishing explained Phishing Explained Phishing scams are typically fraudule...

Article: Automatic Replies (Out Of Office) Automatic Replies (Out Of Office) This article explains how to use automatic...

Article: Create An Email Signature Create An Email Signature To create a personalized email signature: Open a... about a year ago

Upcoming events

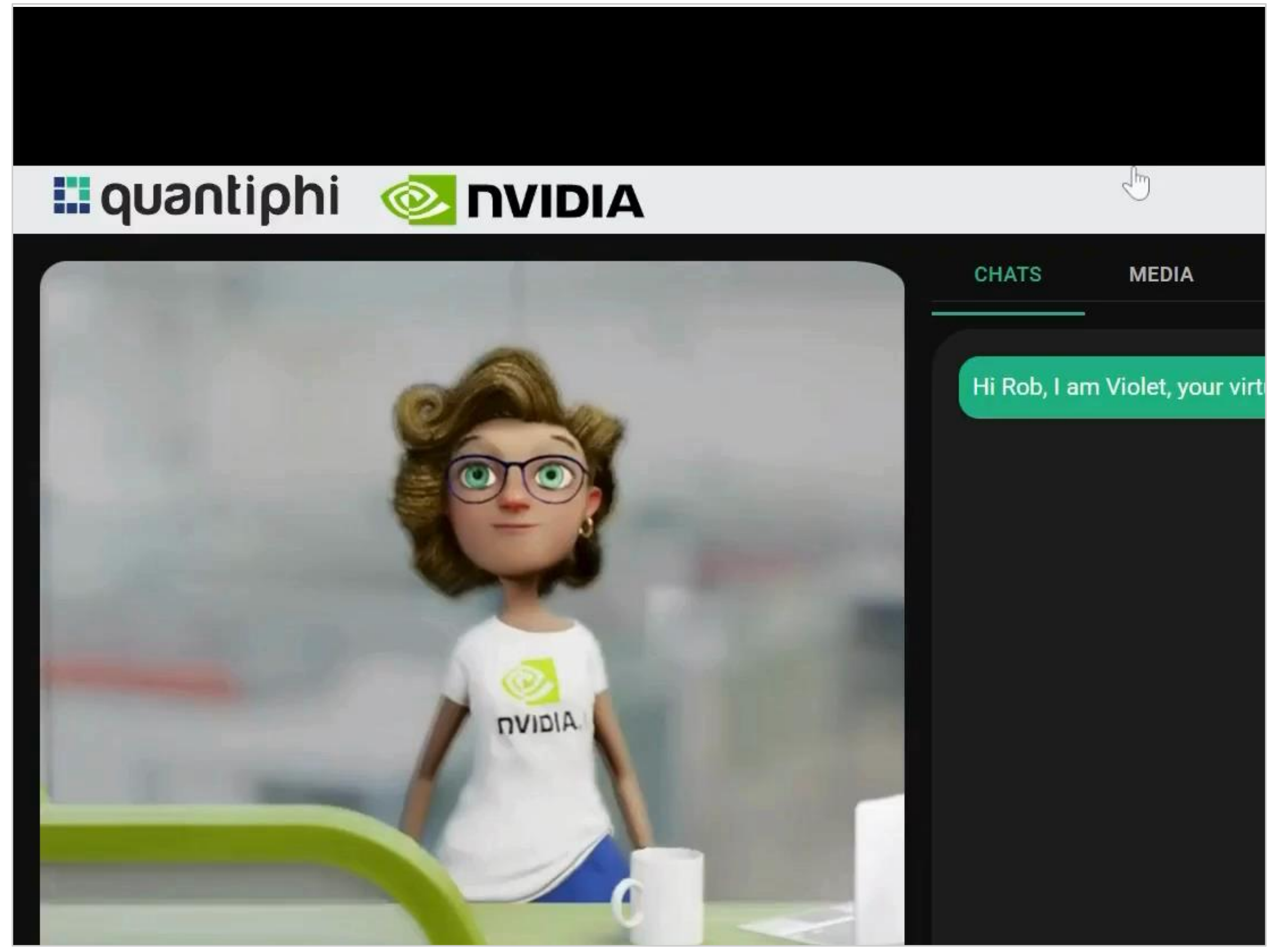
ServiceNow Customer Relations Management



>Check if the code base contains the urllib.parse function.

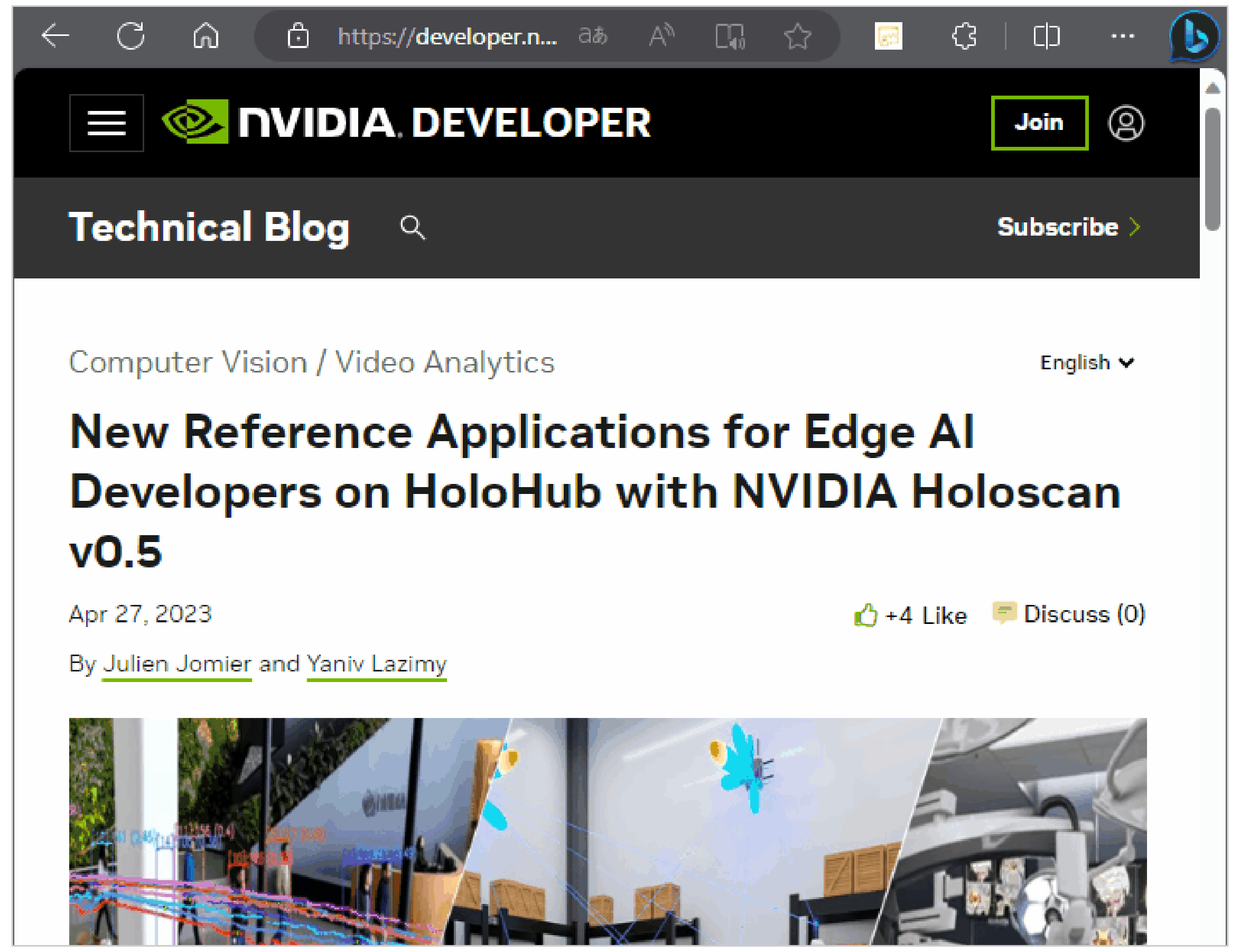
>The code base does not contain the urllib.parse function.

Software Security Analysis



Hi Rob, I am Violet, your virt

Telco Customer Service Avatar



Computer Vision / Video Analytics

New Reference Applications for Edge AI Developers on HoloHub with NVIDIA Holoscan v0.5

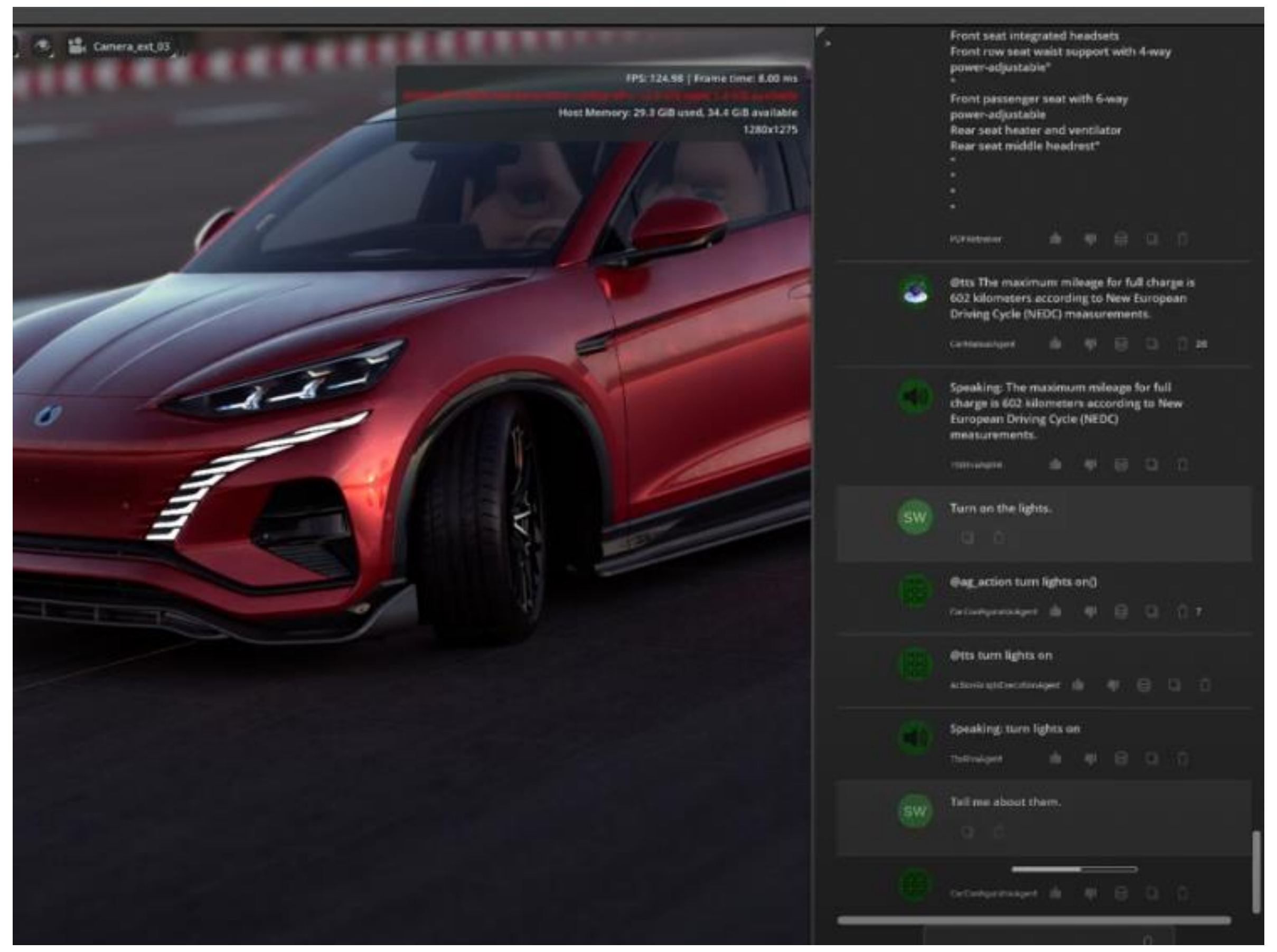
Apr 27, 2023

+4 Like

Discuss (0)

By Julien Jomier and Yaniv Lazimy

Summarization



Front seat integrated headsets
Front row seat waist support with 4-way power-adjustable"
Front passenger seat with 6-way power-adjustable
Rear seat heater and ventilator
Rear seat middle headrest"

@tts The maximum mileage for full charge is 602 kilometers according to New European Driving Cycle (NEDC) measurements.

Speaking: The maximum mileage for full charge is 602 kilometers according to New European Driving Cycle (NEDC) measurements.

Turn on the lights.

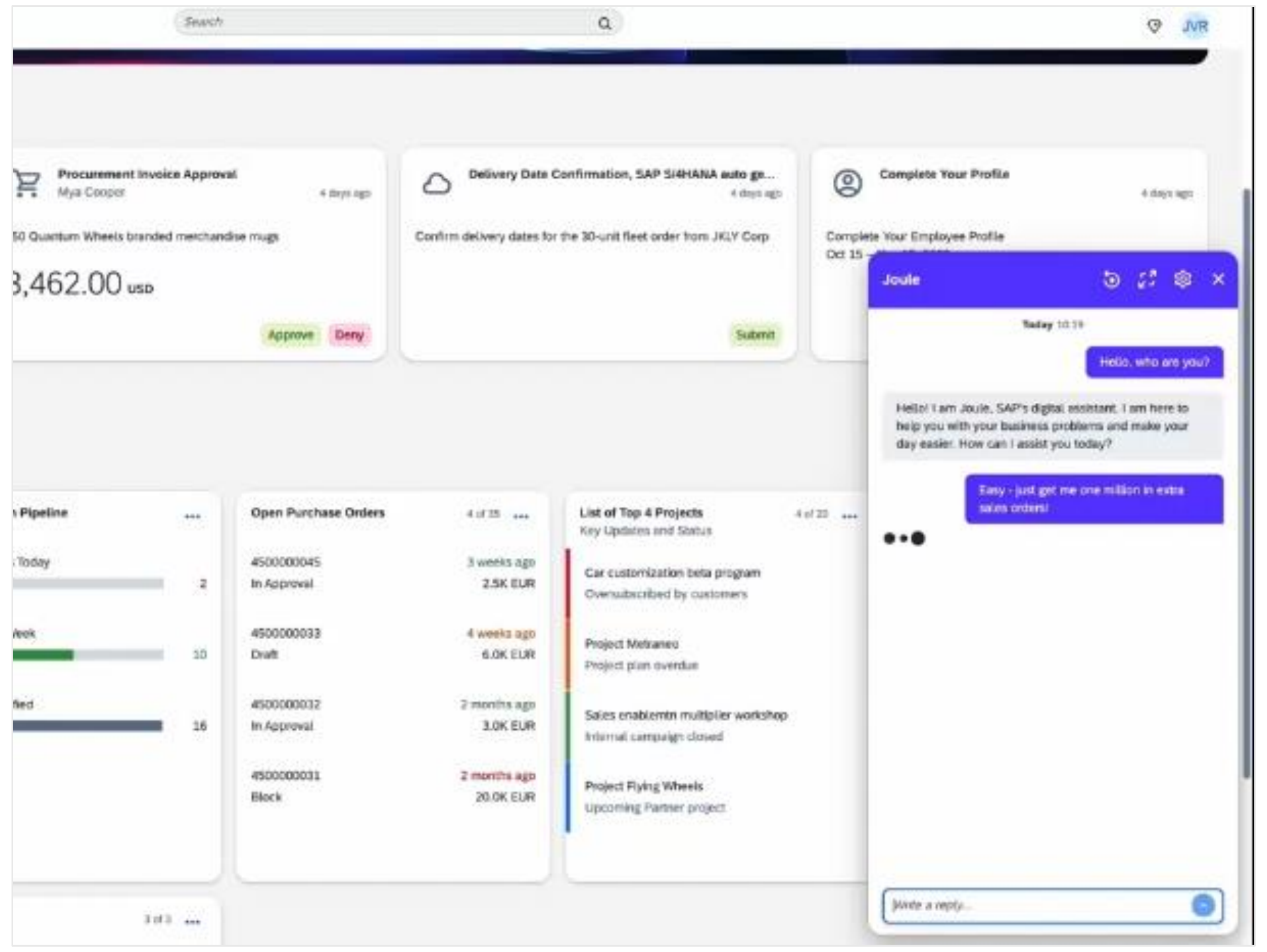
@ag_action turn lights on()

@tts turn lights on

Speaking: turn lights on

Tell me about them.

Car Configurator



Procurement Invoice Approval
Mya Cooper
4 days ago
50 Quantum Wheels branded merchandise mugs
3,462.00 USD

Delivery Date Confirmation, SAP S/4HANA auto ge...
4 days ago
Confirm delivery dates for the 30-unit fleet order from JKLY Corp

Complete Your Profile
4 days ago
Complete Your Employee Profile
Oct 15

Pipeline
Today 2
Week 10
fied 16

Open Purchase Orders 4 of 15

4500000045 In Approval	3 weeks ago 2.5K EUR
4500000033 Draft	4 weeks ago 6.0K EUR
4500000032 In Approval	2 months ago 3.0K EUR
4500000031 Block	2 months ago 20.0K EUR

List of Top 4 Projects 4 of 20
Key Updates and Status

- Car customization beta program – Oversubscribed by customers
- Project Metraneo – Project plan overdue
- Sales enablemtn multiplier workshop – Internal campaign closed
- Project Flying Wheels – Upcoming Partner project

3 of 3

Today 10:19

Hello, who are you?

Hello! I am Joule, SAP's digital assistant. I am here to help you with your business problems and make your day easier. How can I assist you today?

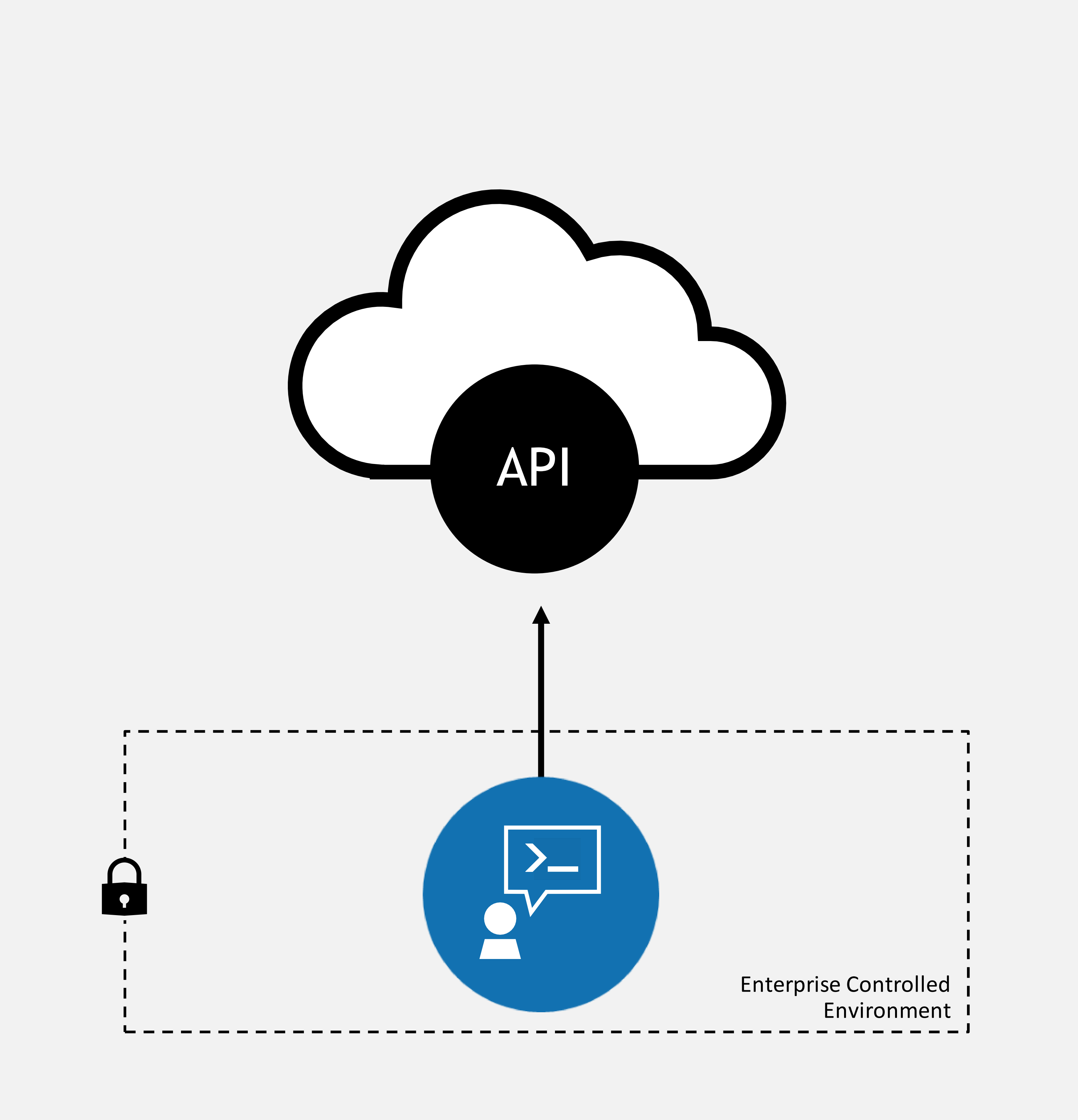
Easy - just get me one million in extra sales orders!

SAP CRM Assistant

Generative AI Deployment Options

Enterprises experimentation with generative AI applications

Managed Generative AI Services



Easy to use APIs for development

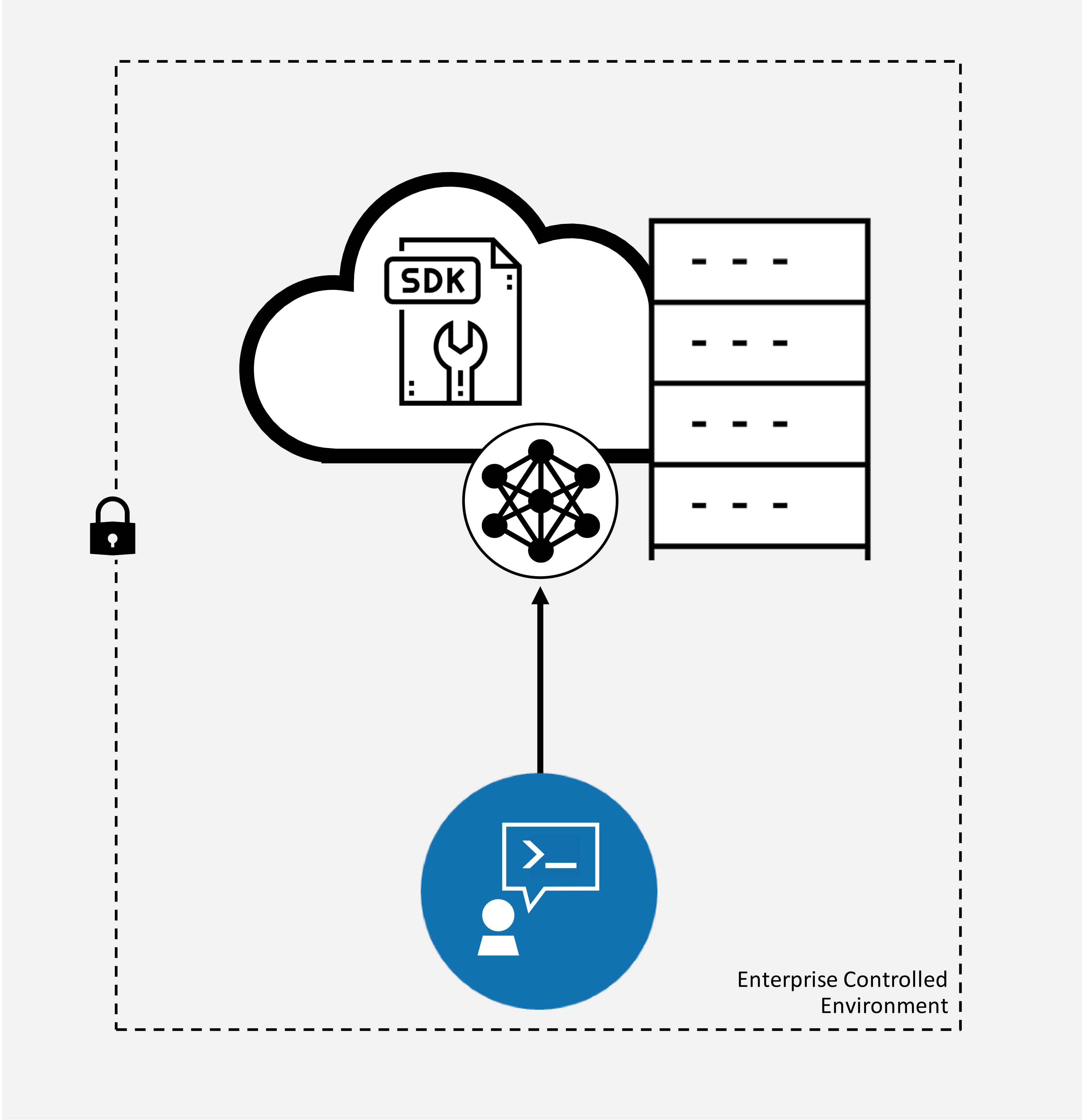
Fast path to getting started with AI

Infrastructure limited to managed environment

Data and prompts are **shared externally**

Limited control for overall generative AI strategy

Open-Source Deployment



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

Inference Microservices for Generative AI

Easiest, fastest, and most portable way to put generative AI models in production. Deploy in 5 minutes.

Deploy Anywhere

Deploy at scale on preferred infrastructure.
Maintain control of generative AI models and data

Easy to use Industry Standard APIs

Easy to integrate. Supports OpenAI API protocol

Day 0 Model Support

Domain specific code for each domain, including LLMs, VLMs, video, healthcare, and more

Improve Cost & Performance

Optimized to provide best latency, throughput, and cost performance

Supports Tuned Models for Best Accuracy

Deploy models that are tuned with proprietary data

Production Ready

Enterprise support with NVIDIA AI Enterprise

Inference Microservices for Generative AI

Easiest, fastest, and most portable way to put generative AI models in production. Deploy in 5 minutes.

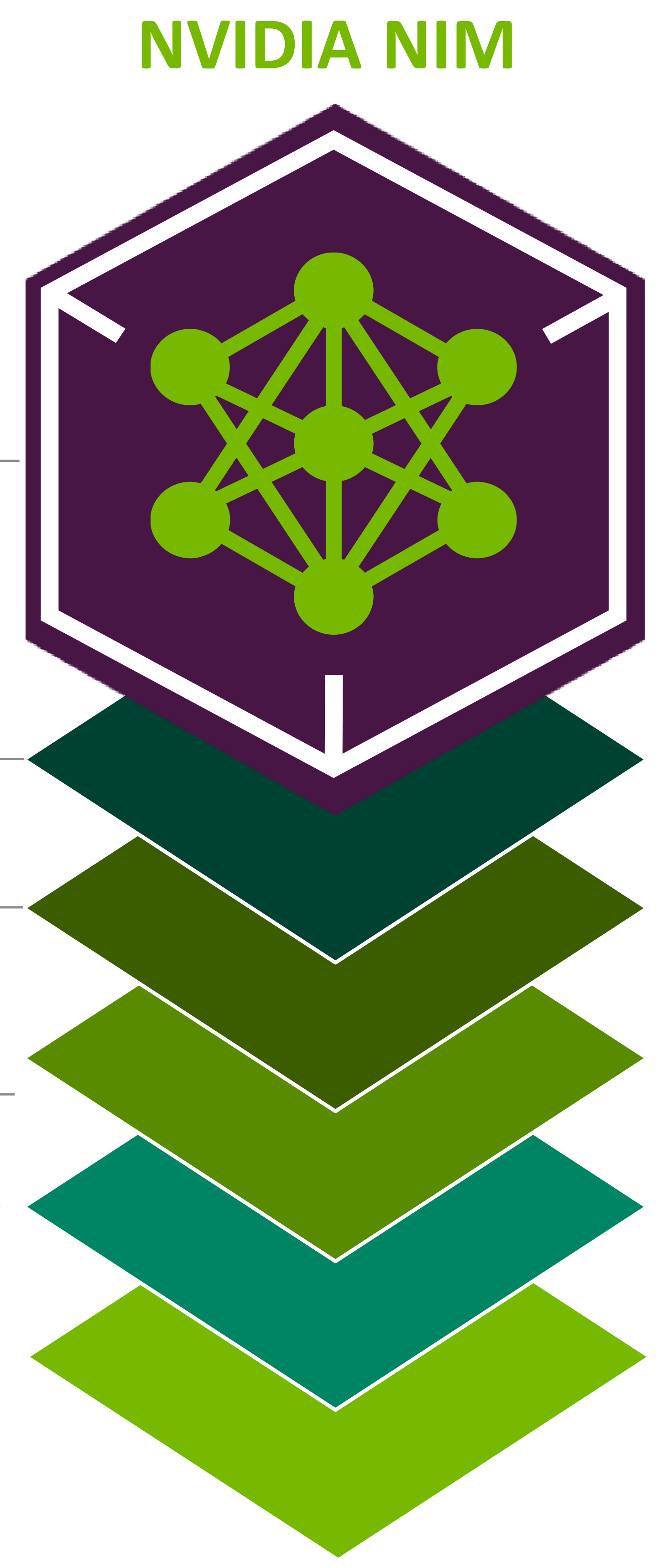
Prebuilt container and helm chart tested and validated across infrastructure

Industry standard APIs
NVIDIA Cloud standards, OpenAI

Domain specific code for each domain e.g. LLMs, VLMs, video, healthcare, and more

Optimized inference models for each model architecture and hardware SKU

Support for customized models build by users targeted use cases (e.g dynamic LoRA, p tuning)



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

NVIDIA NIM is the Fastest Path to AI Inference

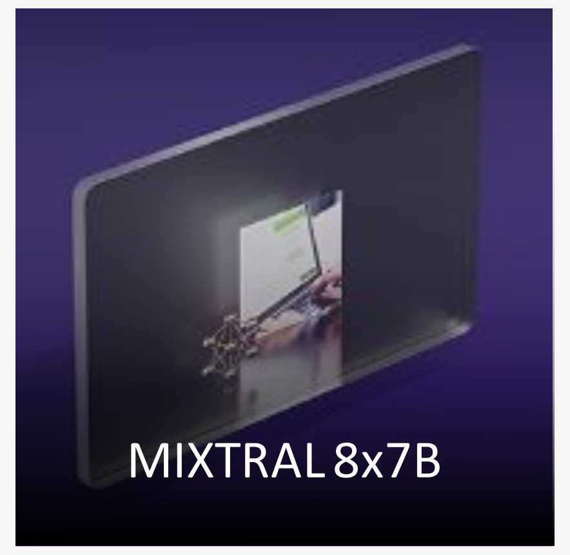
Reduces engineering resources required to deploy optimized, accelerated models

	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 MISTRAL AI_ Llama 2 starcoder NVIDIA Nemotron	User converts checkpoint to TRTLLM 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

Fastest way to deploy AI models on any accelerated infrastructure across cloud, data center, and PC

NVIDIA API Catalog


MIXTRAL 8x7B

MISTRAL AI_

GEMMA 7B

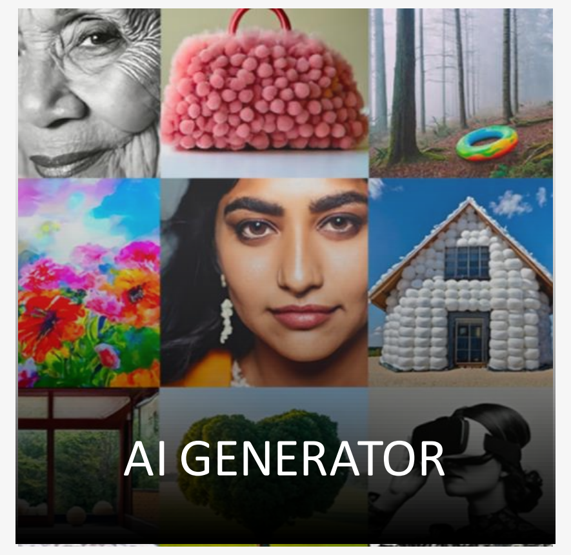
Google

FUYU

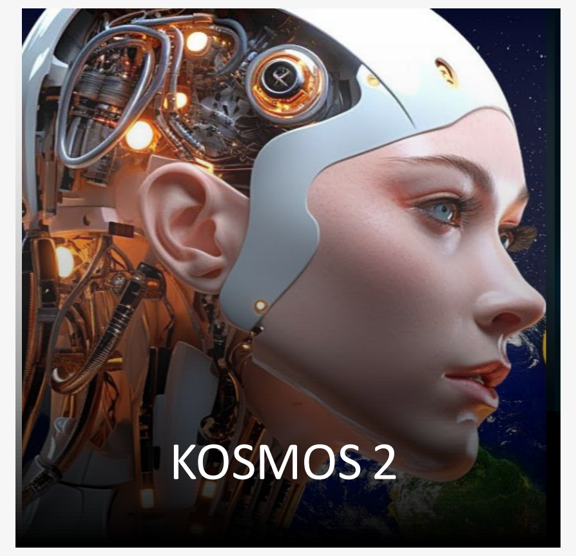
ADEPT

NEMO RETRIEVER

NVIDIA


AI GENERATOR

gettyimages


KOSMOS 2

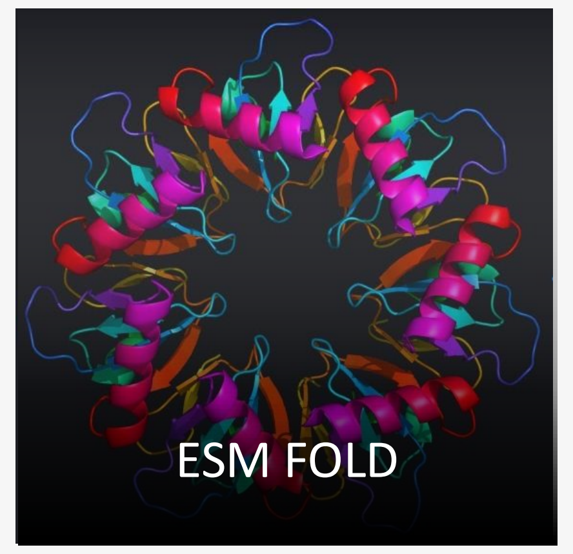
Microsoft

3D GENERATOR

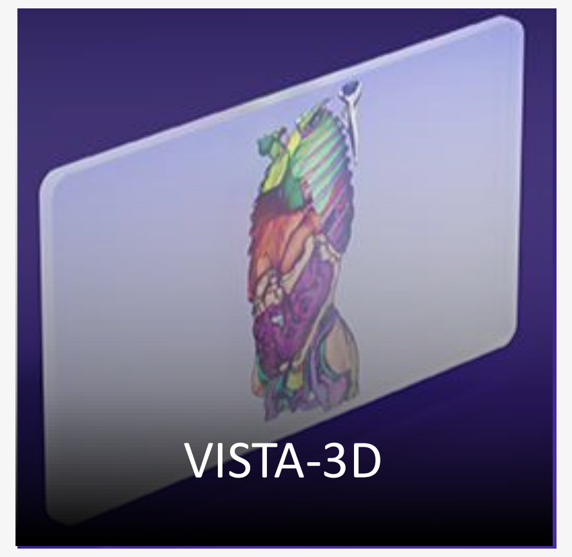
shutterstock

AUDIO2FACE

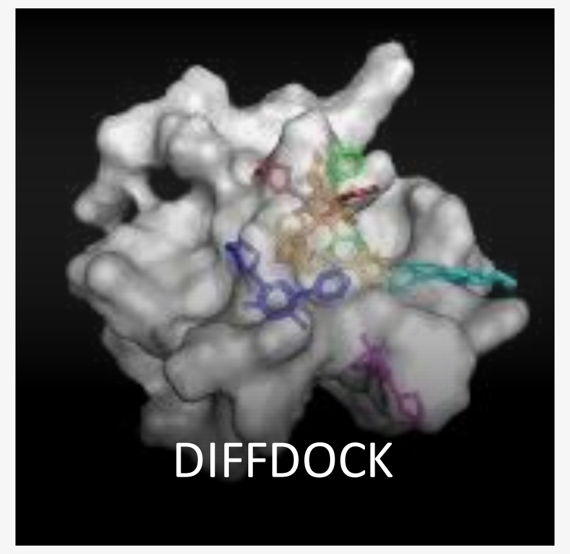
NVIDIA


ESM FOLD

Meta


VISTA-3D

NVIDIA


DIFFDOCK

MIT

MolMIM

NVIDIA

NVIDIA NIM

Microsoft Azure

aws

Google Cloud

ORACLE

DELL Technologies

Hewlett Packard Enterprise

Lenovo

SUPERMICRO

Journey to NVIDIA Cloud APIs

NVIDIA API Catalog - Taking our NIMs to Market

1 Landing Pages

Part of campaign strategy

Platform Pages

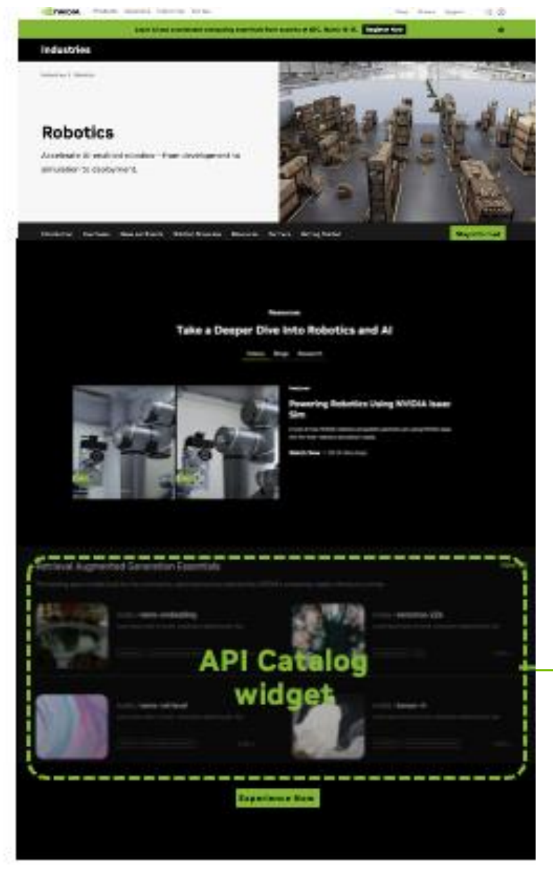
NVIDIA AI

NVIDIA Omniverse

Industry Pages

Powering Breakthroughs in Healthcare and Life Sciences
Advance Medicine and Research With AI
API Catalog widget
Powering Healthcare Solutions With Accelerated Computing

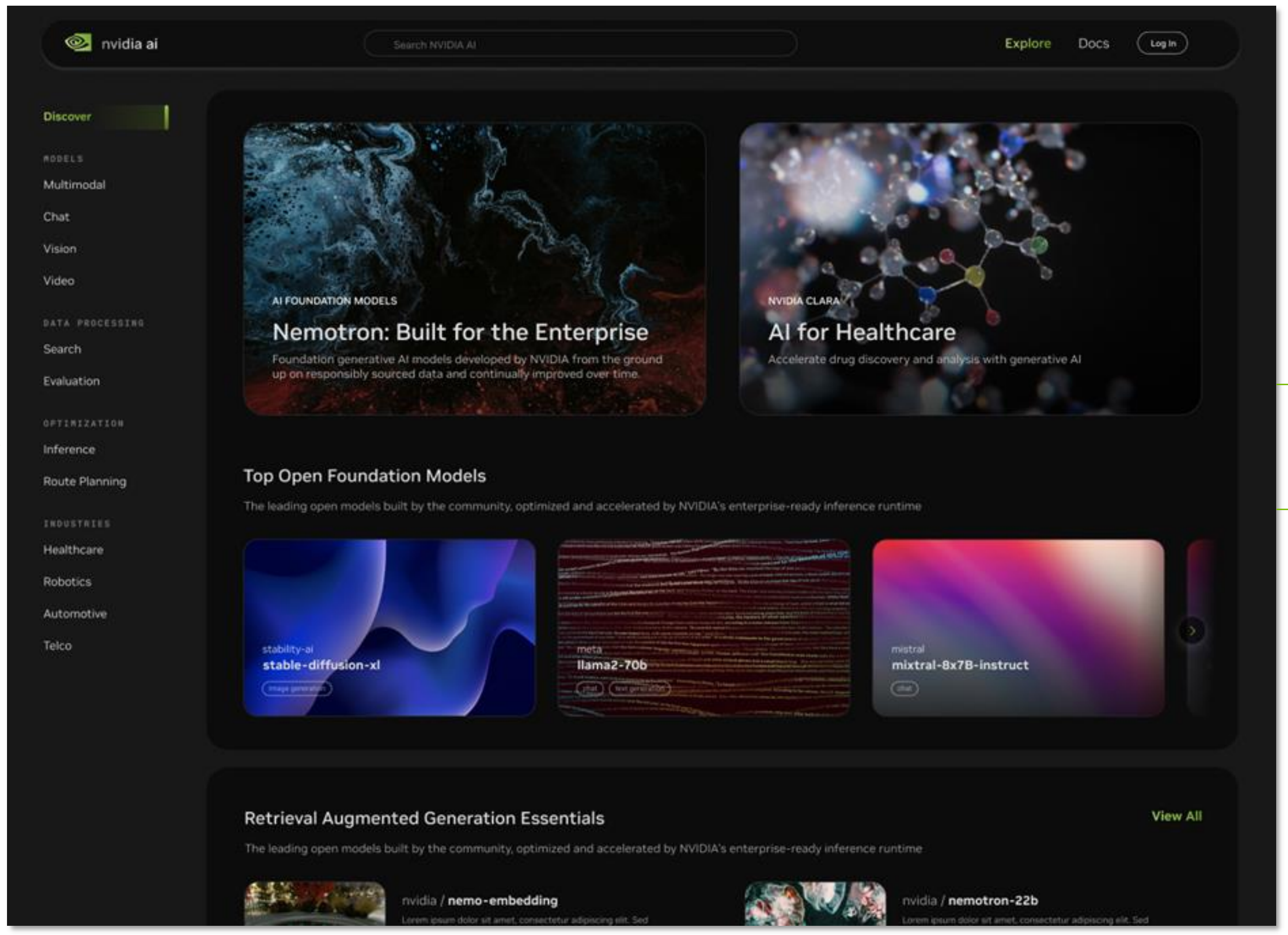
Healthcare



Robotics

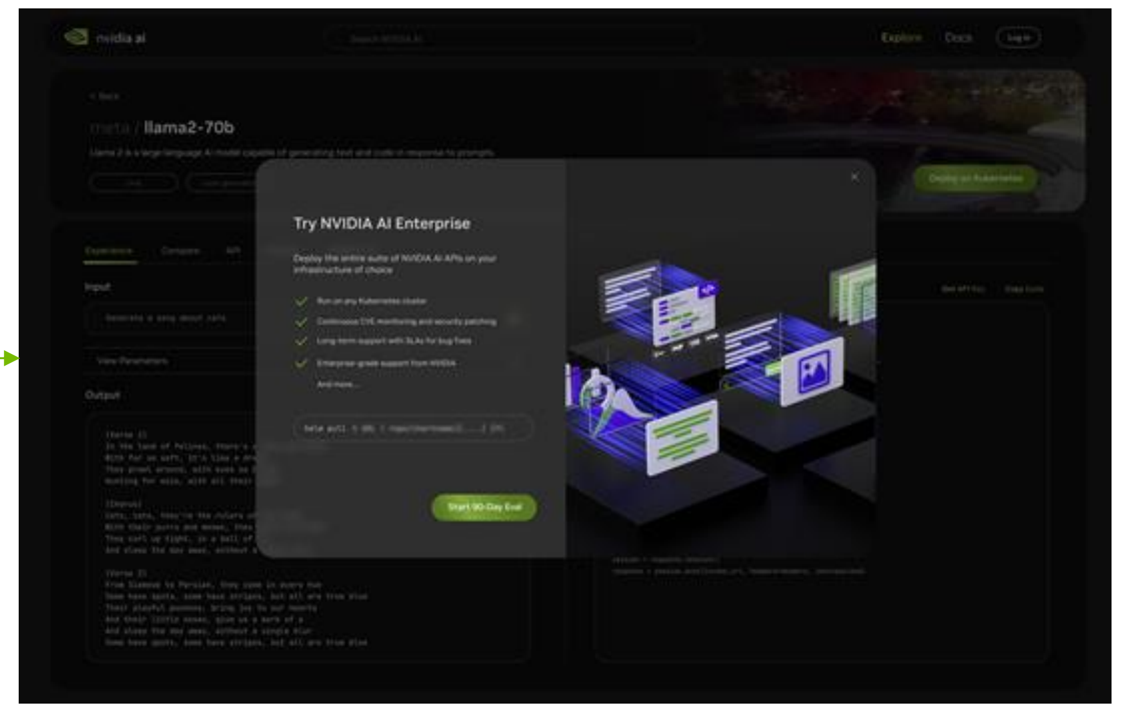
2 NVIDIA API Catalog

Includes all APIs across NVIDIA



nvidia ai
Search NVIDIA AI
Explore
Docs
Log In
Discover
MODELS
Multimodal
Chat
Vision
Video
DATA PROCESSING
Search
Evaluation
OPTIMIZATION
Inference
Route Planning
INDUSTRIES
Healthcare
Robotics
Automotive
Telco
AI FOUNDATION MODELS
Nemotron: Built for the Enterprise
Foundation generative AI models developed by NVIDIA from the ground up on responsibly sourced data and continually improved over time
NVIDIA CLARA
AI for Healthcare
Accelerate drug discovery and analysis with generative AI
Top Open Foundation Models
The leading open models built by the community, optimized and accelerated by NVIDIA's enterprise-ready inference runtime
stability-ai
stable-diffusion-xl
meta
llama2-70b
mistral
mixtral-8x7B-instruct
Retrieval Augmented Generation Essentials
View All
The leading open models built by the community, optimized and accelerated by NVIDIA's enterprise-ready inference runtime
nvidia / nemo-embedding
nvidia / nemotron-22b

3 Download & self host API



3 Route to Partner

Discover APIs

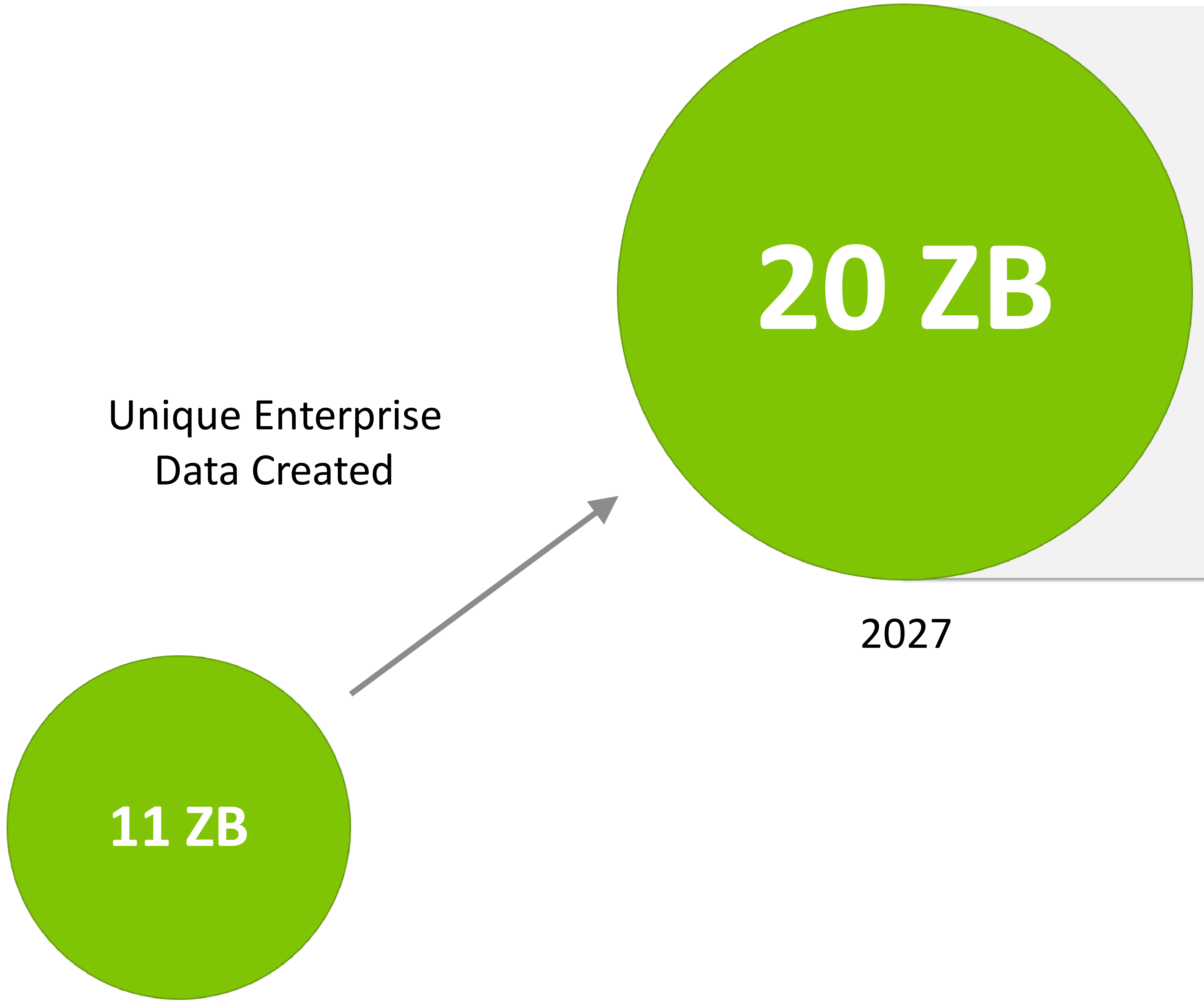
Try & develop with hosted APIs

Consume APIs

Accelerating Access to Enterprise Data with NVIDIA NeMo Retriever

The Amount of Enterprise Data is Massive & Growing

NVIDIA Accelerated Retrieval-Augmented Generation



Unique Enterprise Data Created

800,000 Libraries of Congress

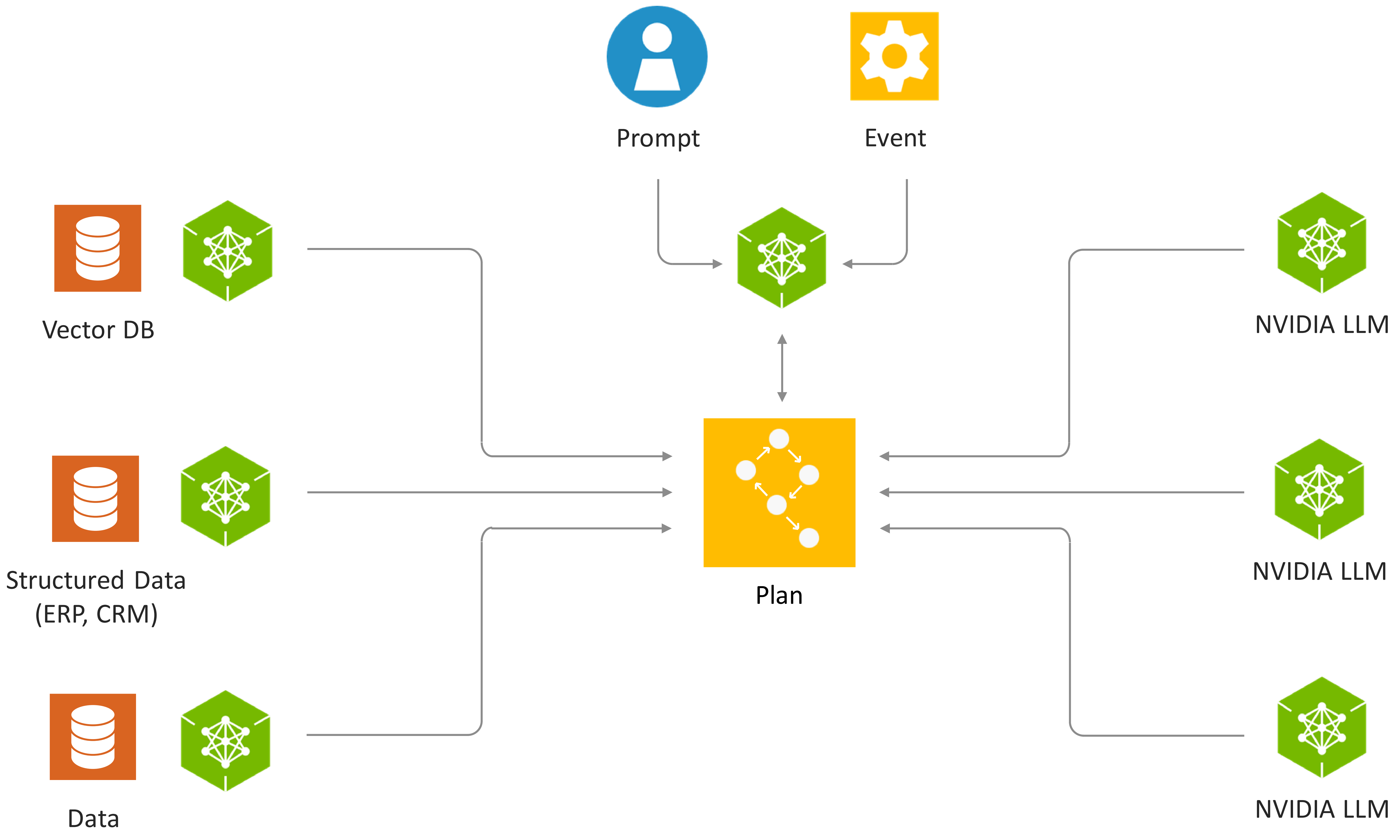
83% unstructured data

50% audio and video

Source: IDC Global DataSphere

What is Retrieval-Augmented Generation?

Enhance the Accuracy & Reliability of Generative AI



RAG Challenges

Difficult to Take a RAG Pipeline from PoC to Production

Accuracy

Accurate generations require retrieval systems tuned to match the data

Data Security

Sending sensitive data to remote endpoints is inherently insecure

Complexity

System builders must piece together and integrate many components

Cost

Automated LLM transactions make transaction costs unpredictable

Innovation Velocity

New models and techniques appear every day

NVIDIA Retriever for World-Class Information Retrieval

Lowest Latency, Highest Throughput, Maximum Data Privacy

Optimized Inference Engines

Built on NVIDIA TensorRT & Triton
Tuned for RAG application workloads

World Class Models & Community Model support

SOTA commercial models
Converts to TensorRT for optimal performance

Flexible & Modular Deployment

Deploy at scale on preferred infrastructure

Customizable Pipelines & Models

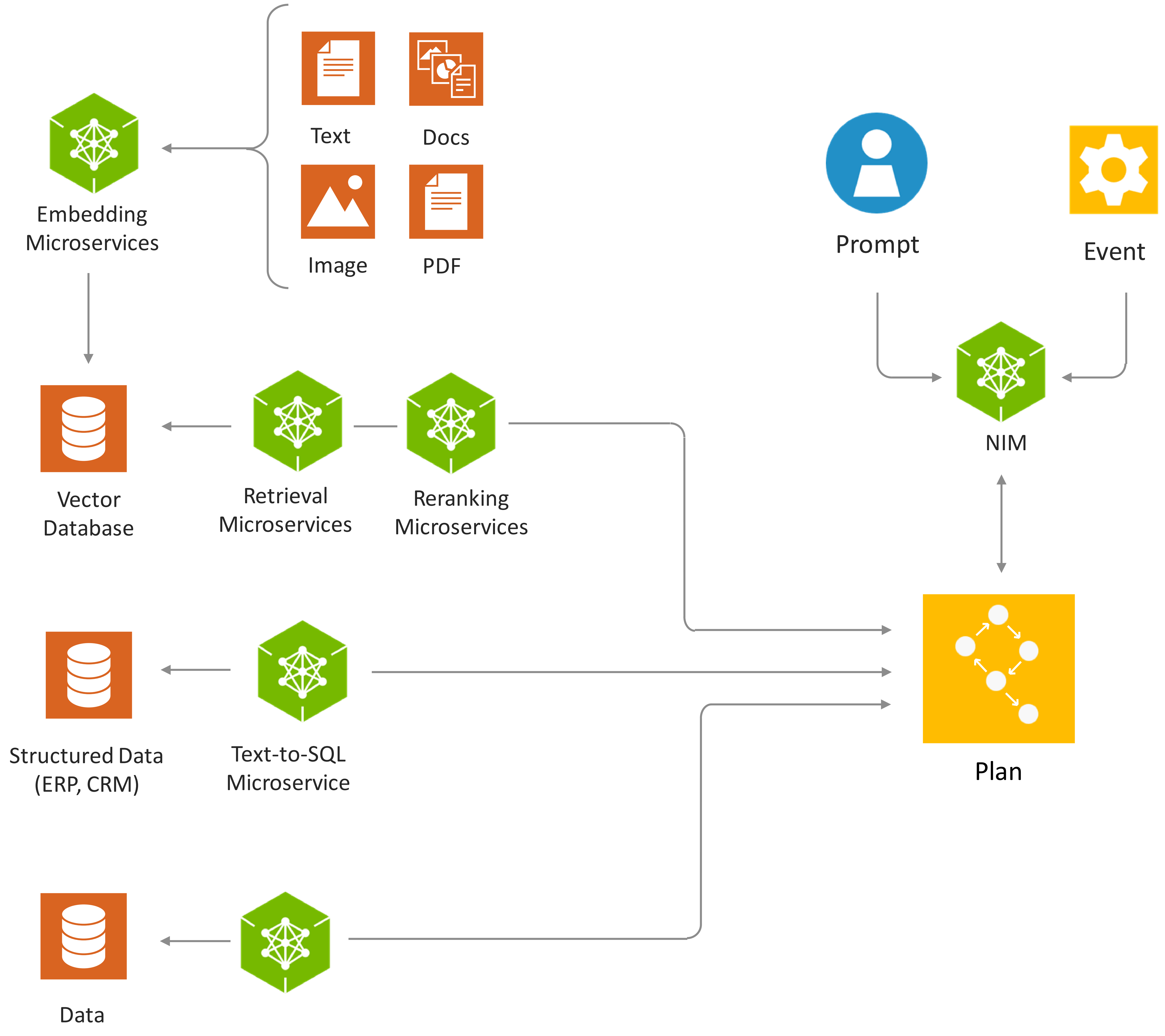
Compose pipelines with microservices
Customize models for target domain

Production Ready

Enterprise support running in any cloud or on prem

NeMo Retriever Supercharges RAG Applications

World Class Accuracy and Throughput



- **2X** World-class accuracy with nearly 2x fewer incorrect answers
- **7X** Faster embedding inference throughput
- Optimized Inference Engines
- World class models and community model support
- Flexible and modular deployment
- Customizable models and pipelines
- Production Ready

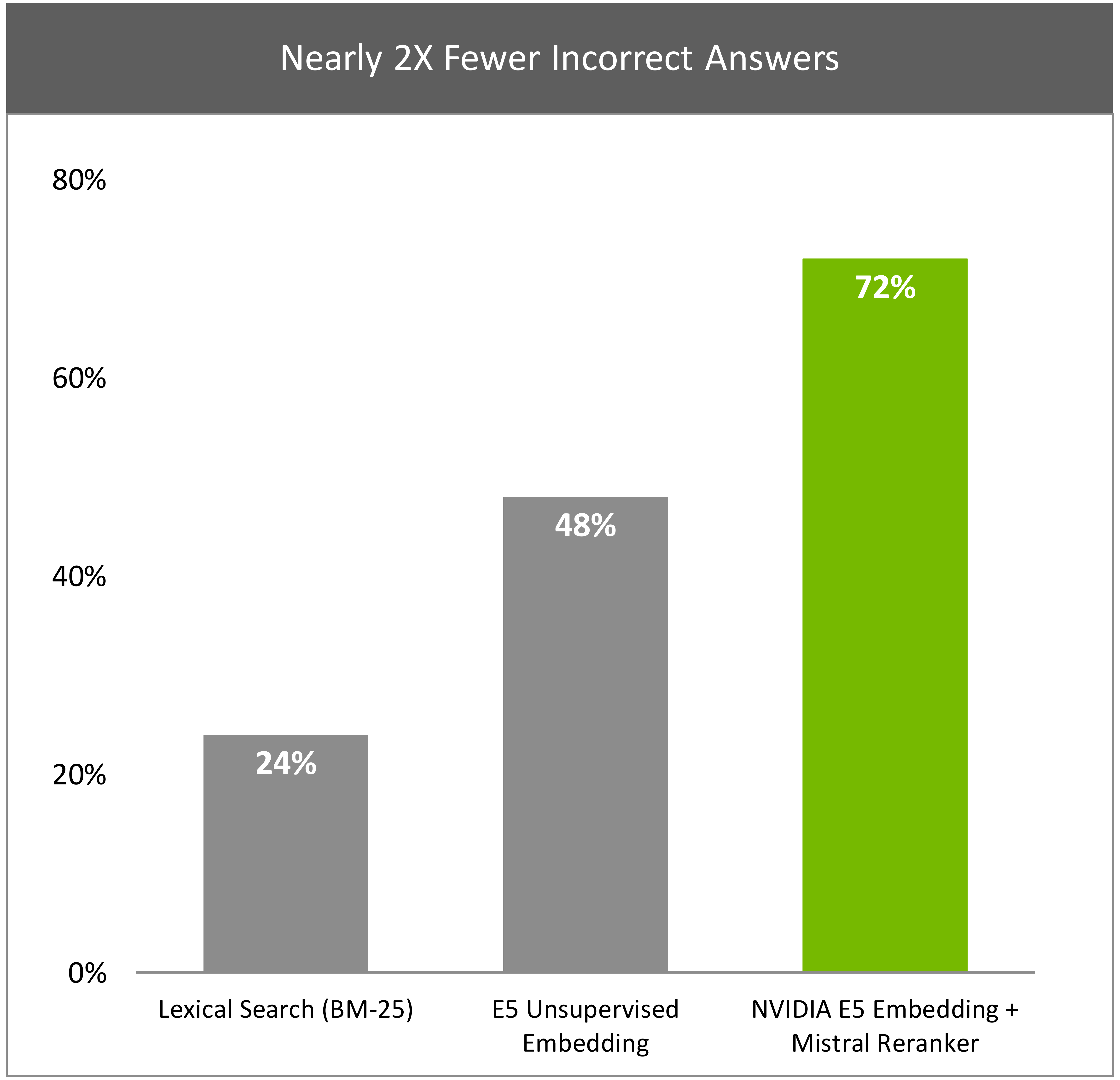
NVIDIA NeMo Retriever: Supercharges your RAG Application

Enterprise-ready microservices for RAG

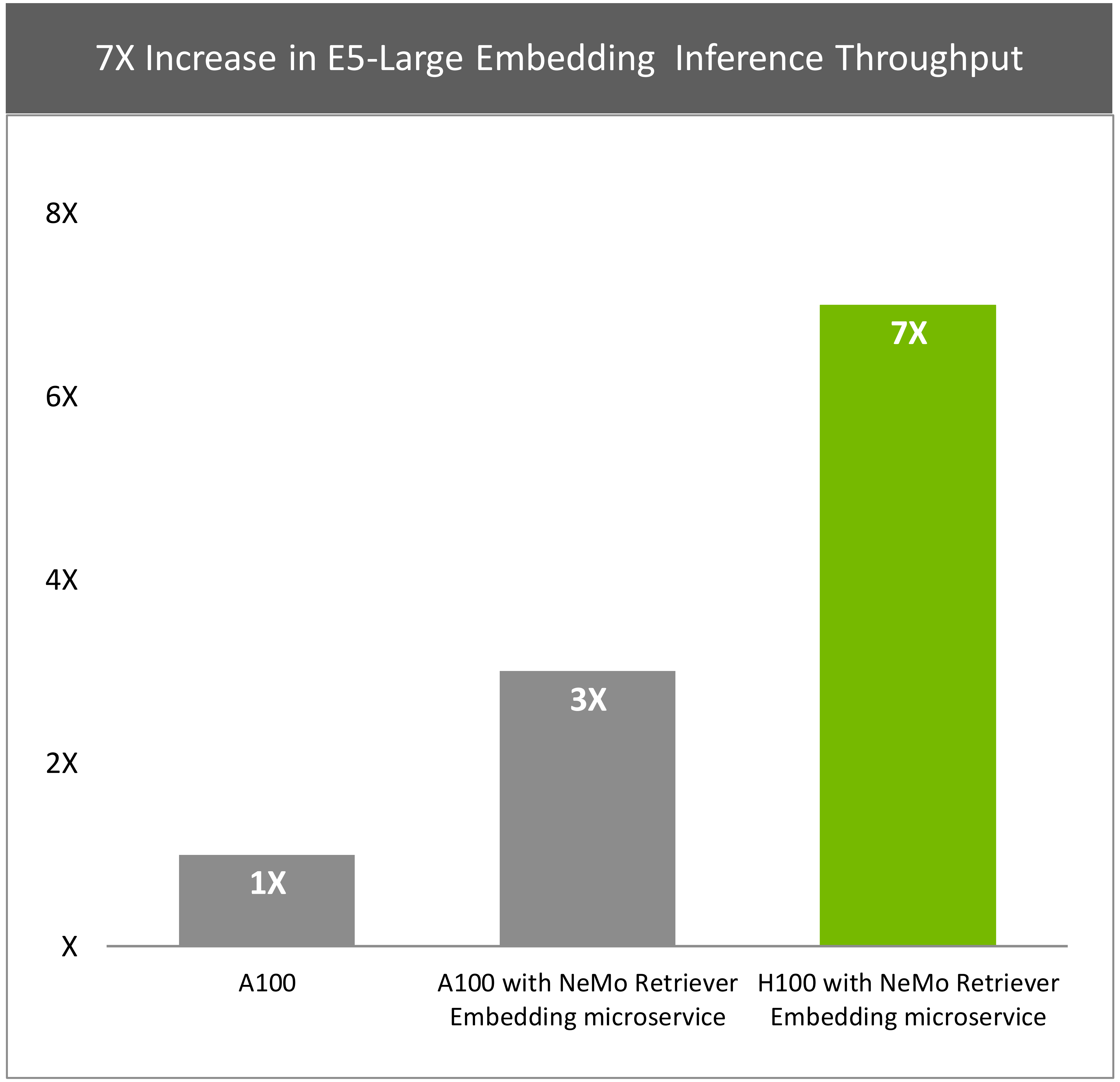
Continuous Data Processing
Text
Office Docs
Image
PDF
NeMo Curator Microservice
Metadata
Data Chunks
Text & Image Embedding Microservices
Knowledge Embeddings
Vector DB
Semantic Embedding of Question
Text & Image Embedding Microservices
Enterprise Data Warehouses
SQL Representation of Question
SQL Retrieval Microservice
RAG-Powered Application
User
Application
"What is the next two quarter pipeline for..."
Question
Agent
Answer
"The next two quarter pipeline of potential closed business is currently ..."
Reranking Microservice
Most Relevant Results
NVIDIA NIM (LLM)

World Class Accuracy and Throughput

Retrieval, embedding, reranking microservices for RAG

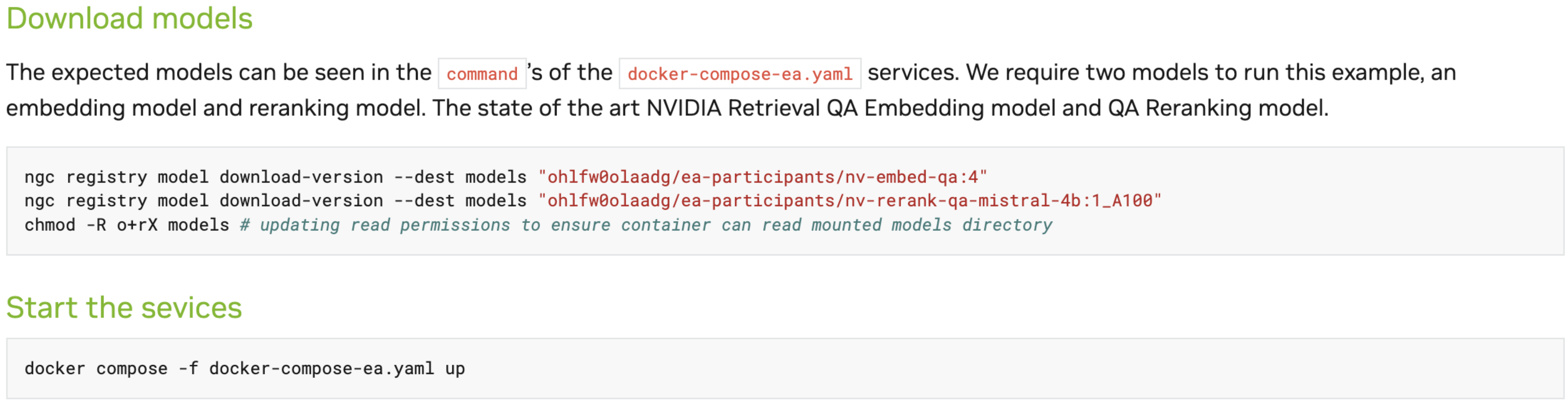


Comparing NVIDIA Text QA Embedding Model vs Other Available Options. Recall Top 5, 300 token chunk size, averaging across representative customer datasets from Telco, IT, Consulting, Energy



Simplifying Retrieval at Scale

NeMo Retriever in Practice



Download models

The expected models can be seen in the `command`'s of the `docker-compose-ea.yaml` services. We require two models to run this example, an embedding model and reranking model. The state of the art NVIDIA Retrieval QA Embedding model and QA Reranking model.

```
ngc registry model download-version --dest models "ohlfw0olaadg/ea-participants/nv-embed-qa:4"
ngc registry model download-version --dest models "ohlfw0olaadg/ea-participants/nv-rerank-qa-mistral-4b:1_A100"
chmod -R o+rX models # updating read permissions to ensure container can read mounted models directory
```

Start the sevices

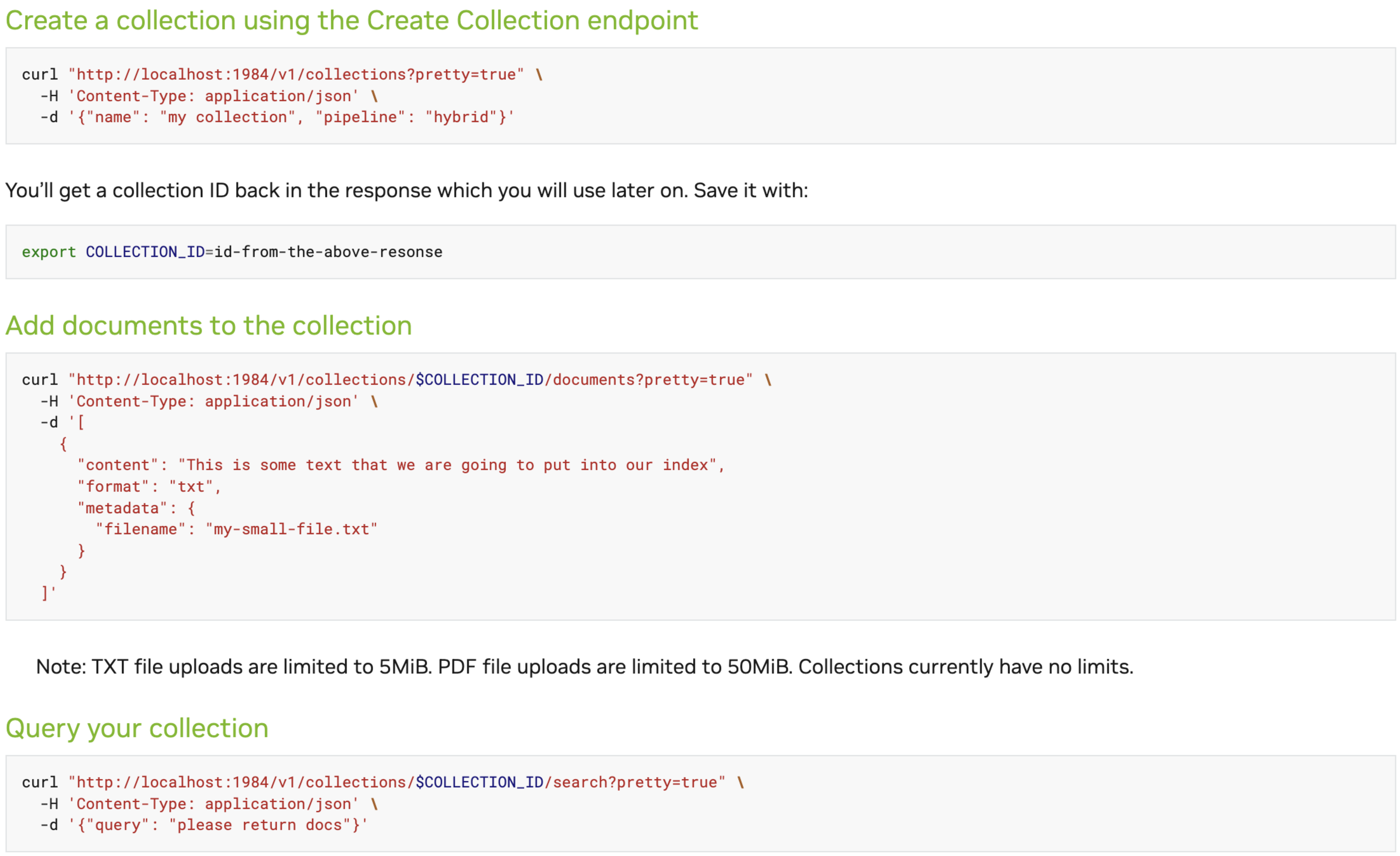
```
docker compose -f docker-compose-ea.yaml up
```

0 Set up the retrieval pipeline in just a few lines

1 Create document collection

2 Upload documents

3 Retrieve relevant data



Create a collection using the Create Collection endpoint

```
curl "http://localhost:1984/v1/collections?pretty=true" \
  -H 'Content-Type: application/json' \
  -d '{"name": "my collection", "pipeline": "hybrid"}'
```

You'll get a collection ID back in the response which you will use later on. Save it with:

```
export COLLECTION_ID=id-from-the-above-resonse
```

Add documents to the collection

```
curl "http://localhost:1984/v1/collections/$COLLECTION_ID/documents?pretty=true" \
  -H 'Content-Type: application/json' \
  -d '[
    {
      "content": "This is some text that we are going to put into our index",
      "format": "txt",
      "metadata": {
        "filename": "my-small-file.txt"
      }
    }
  ]'
```

Note: TXT file uploads are limited to 5MiB. PDF file uploads are limited to 50MiB. Collections currently have no limits.

Query your collection

```
curl "http://localhost:1984/v1/collections/$COLLECTION_ID/search?pretty=true" \
  -H 'Content-Type: application/json' \
  -d '{"query": "please return docs"}'
```

Case Study: Cadence Design Systems

3.3x fewer incorrect answers retrieving from technical documentation

Recall	Top 1	Top 3	Top 5	Top 10
Reference Pipeline	36%	52%	57%	64%
NeMo Retriever Hybrid Search	57%	70%	77%	80%
NeMo Retriever Hybrid Search + Reranker	69%	81%	86%	89%
Improvement Factor	**2x**	**2.5x**	**3x**	**3.3x**

cādence®

NVIDIA.

Unlock Petabytes of Enterprise Data

Transform Data into Business Insights

Adobe

Adobe's proprietary AI will help unlock the knowledge inside the world's more than 3 trillion PDFs worldwide.

CLOUDERA

Cloudera will expand its generative AI capabilities by integrating NeMo Retriever with Cloudera Machine Learning to unlock the potential of 25 exabytes of enterprise data.

COHESITY

Cohesity data platform customers can add generative AI intelligence to their data backups and archives.

DATASTAX

Datastax leverages NeMo Retriever and NVIDIA NIM improving performance of RAG applications. Using NVIDIA H100 GPUs they achieve an embedding and indexing latency of 10 ms.

NetApp

NetApp unlocks exabytes of data empowering customers to securely "talk to their data" to access business insights.

PURESTORAGE®

Pure accelerates time to insight for enterprises using their own internal data for AI training, ensuring the use of their latest data and eliminating the need for constant retraining of LLMs.

SAP

SAP plans to add RAG capabilities that enable generative AI applications to more securely access data running on SAP software to improve accuracy and insights, using NeMo Retriever.

snowflake

Snowflake customers will be able to utilize NeMo Retriever directly on their proprietary data in the Data Cloud, all while maintaining data security, privacy, and governance seamlessly through Snowflake's built-in capabilities.

Next Generation of Enterprise Applications Connect LLMs to Enterprise Data

Retrieval Augmented Generation Improves LLM Performance and Efficiency

Improved Accuracy

Models can answer questions about information without having been trained on that data

Natural Language Interface

Human-readable output texts that are easier for people to understand, raising user trust

Contextual Understanding

AI models better understand context when generating text or other outputs

Reduced Computational Costs

Reduced computational costs from retraining and model size at inference

Improved Efficiency

Models can produce diverse outputs without sacrificing accuracy or efficiency

Building Generative AI Applications for the Enterprise

Build, customize, and deploy generative AI models with NVIDIA NeMo.

API Gateway
NeMo Curator
NeMo Customizer
NeMo Evaluator
NeMo Retriever
NeMo Guardrails
NVIDIA NIM
Data Prep
Training and Customization
Deployment

AI-Based Information Retrieval Unlocks Enterprise Knowledge

NVIDIA NeMo Ecosystem

DATA PLATFORMS

Adobe CLOUDERA COHESITY DATASTAX

Dropbox NetApp PURESTORAGE®

RAG ECOSYSTEM

glean deepset

LangChain LlamaIndex UNSTRUCTURED

APACHE LUCENE™ DATASTAX Faiss kinetica milvus

OpenSearch redis Weaviate

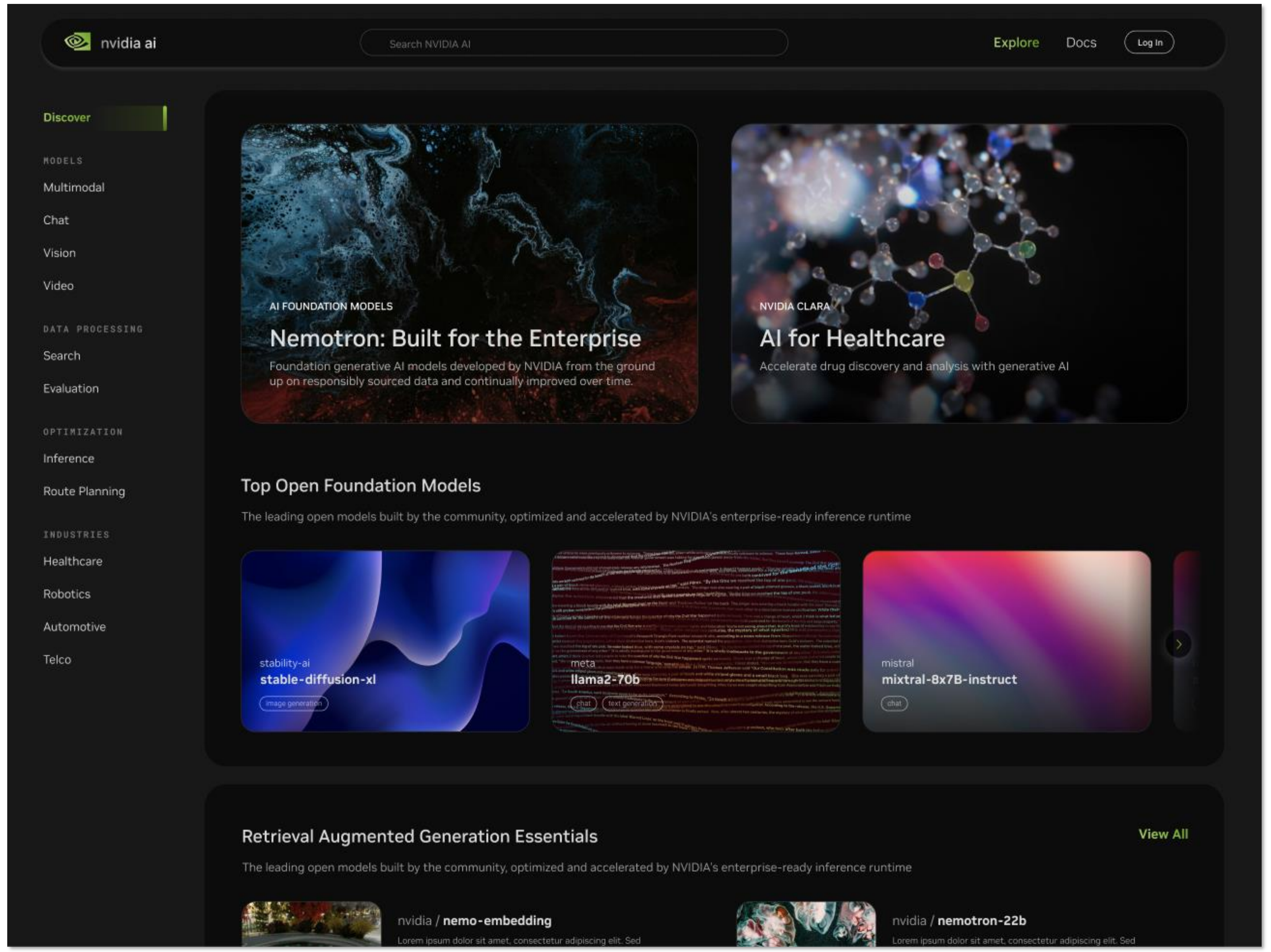
VECTOR DATABASE

Getting Started

Endpoints & EA

1 NVIDIA API Catalog

Includes all APIs across NVIDIA



AI FOUNDATION MODELS

Nemotron: Built for the Enterprise

Foundation generative AI models developed by NVIDIA from the ground up on responsibly sourced data and continually improved over time.

NVIDIA CLARA

AI for Healthcare

Accelerate drug discovery and analysis with generative AI

Top Open Foundation Models

The leading open models built by the community, optimized and accelerated by NVIDIA's enterprise-ready inference runtime

stability-ai
stable-diffusion-xl

meta
llama2-70b

mistral
mixtral-8x7B-instruct

Retrieval Augmented Generation Essentials

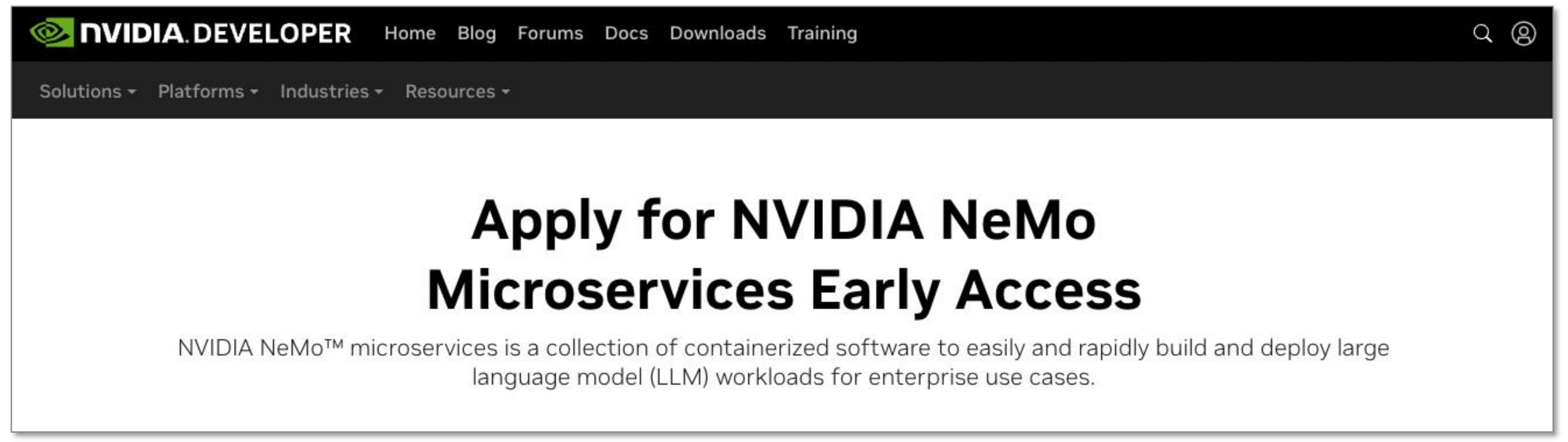
The leading open models built by the community, optimized and accelerated by NVIDIA's enterprise-ready inference runtime

nvidia / nemo-embedding
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed

nvidia / nemotron-22b
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed

2 Apply for Early Access

For NeMo Retriever microservices



Apply for NVIDIA NeMo Microservices Early Access

NVIDIA NeMo™ microservices is a collection of containerized software to easily and rapidly build and deploy large language model (LLM) workloads for enterprise use cases.

NVIDIA®