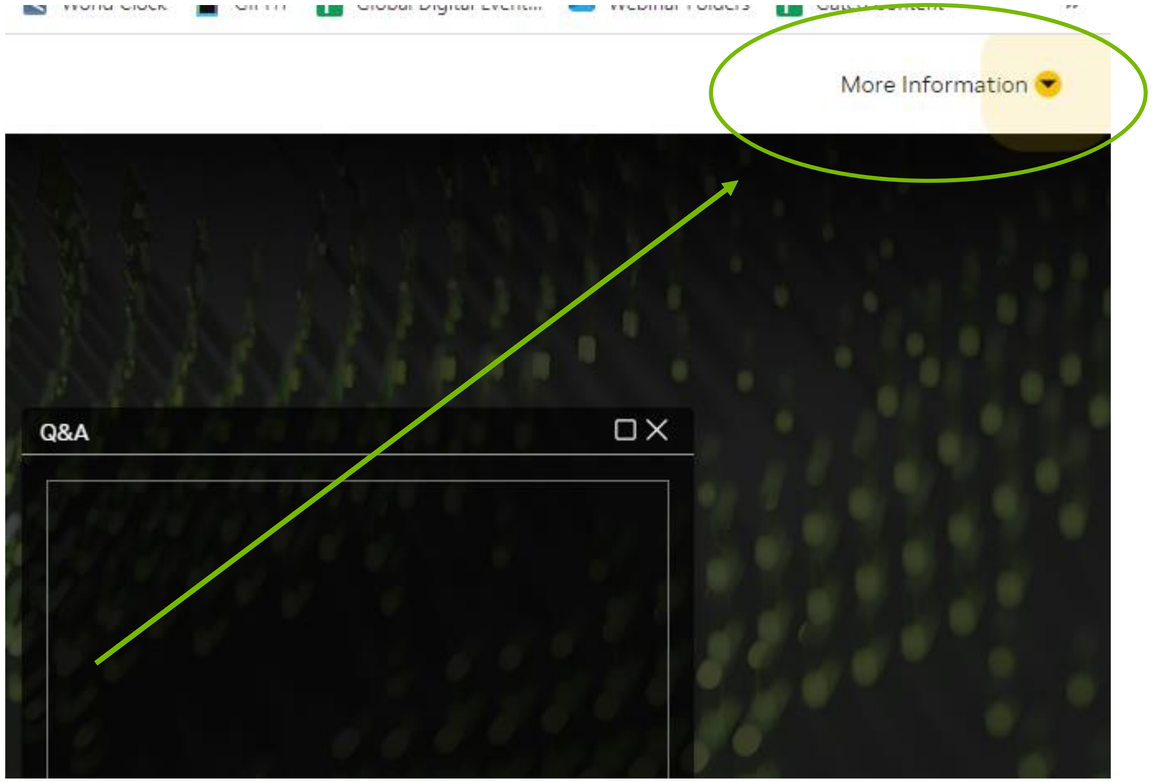
Fast-Track to Generative AI With NVIDIA

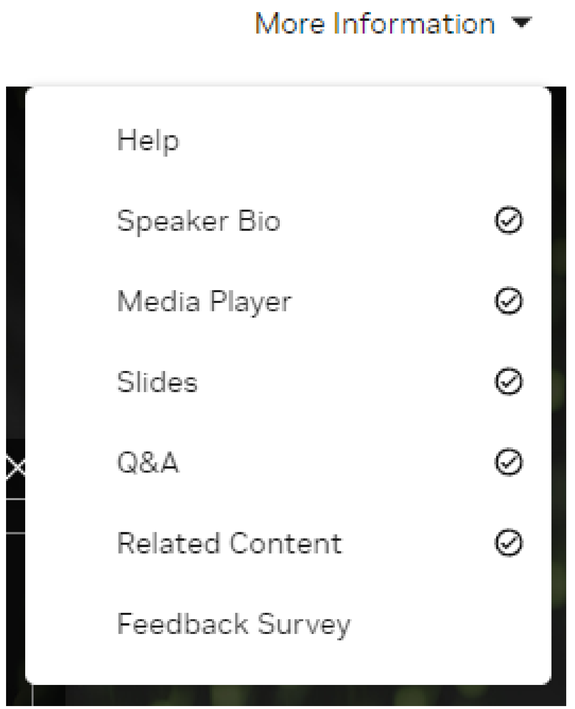
January 2024

Anne Hecht, Sr Director Enterprise Products, NVIDIA

Tony Paikeday, Sr Director AI Systems, NVIDIA

How to Use the Console





What We'll Cover

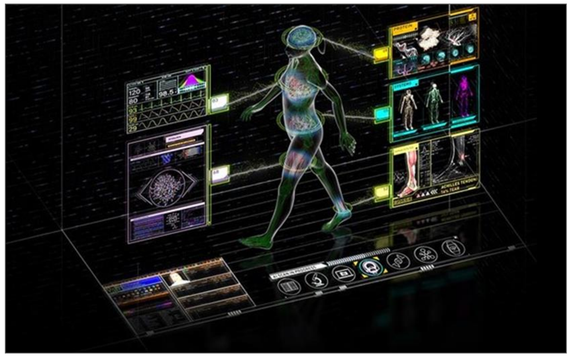
- ✓ How Enterprises are Using Generative AI
- ✓ Model Customization Best Practices
- ✓ Putting Generative AI into Production
- ✓ Getting Started

Generative AI is Transforming Business

Generative AI's impact on productivity could add up to $4.4 trillion annually to the global economy.[1]

Finance

Fraud Detection | Personalized Banking
Investment Insights

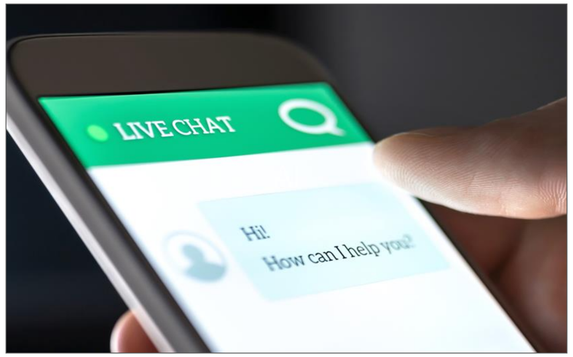


Healthcare

Molecule Simulation | Drug Discovery
Clinical Trial Data Analysis

Retail

Personalized Shopping | Automated Catalog
Descriptions | Automatic Price Optimization



Telecommunications

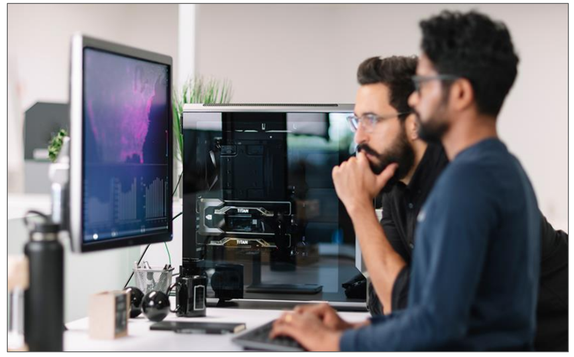
AI Virtual Assistants | Network Performance Tuning
Remote Support Capabilities

Media & Entertainment

Character Development | Style Augmentation
Video Editing & Image Creation

Manufacturing

Factory Simulation | Product Design
Predictive Maintenance



Federal

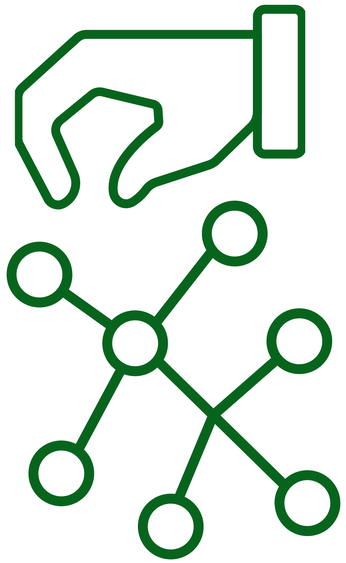
Document Summarization | Audit Compliance
AI Virtual Assistants

Predictive Maint
Knowl

Source: 1. The economic potential of generative AI: The next productivity frontier, McKinsey, 2023

Addressing the need for custom-built AI

Enterprise apps built on industrial-grade models



Built on your brand/vocabulary

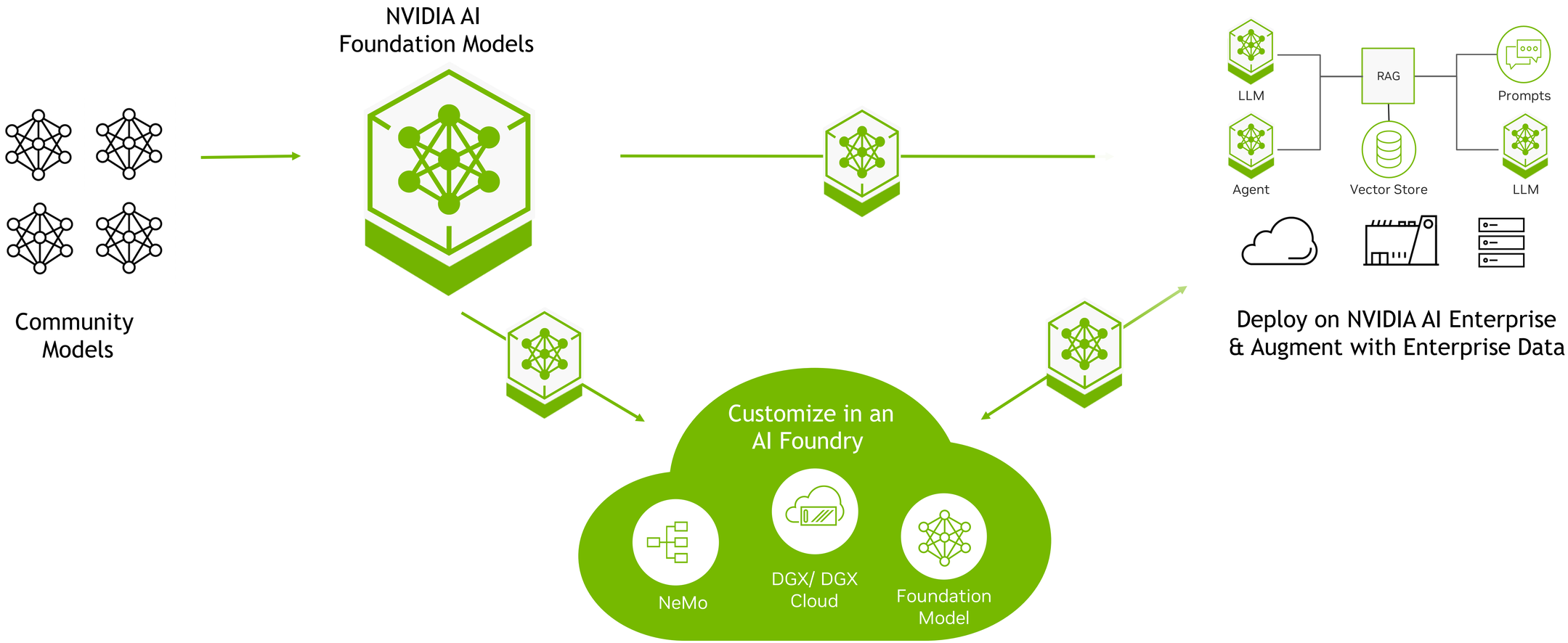
Oceans of training data

Trustworthiness

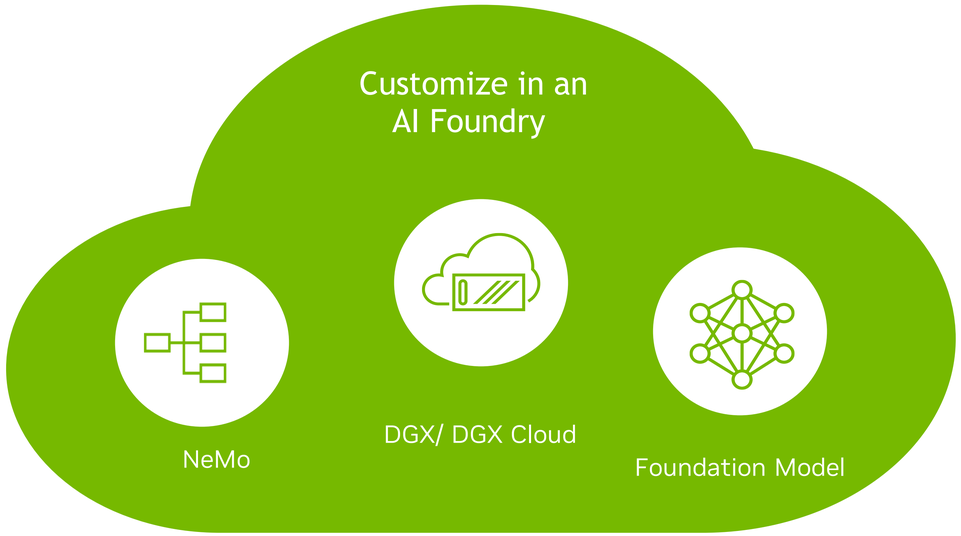
Intellectual property

Building Generative AI for the Enterprise

NVIDIA AI Foundation Models – NeMo with NVIDIA AI Enterprise - DGX & DGX Cloud



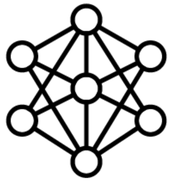
The AI Foundry for Model Customization



Enterprises Need Custom Models to Power Their Business

Businesses need to turn “off-the-shelf” models into proprietary models

“How much storage do I have in my Google Drive?”



Leading Foundation Model

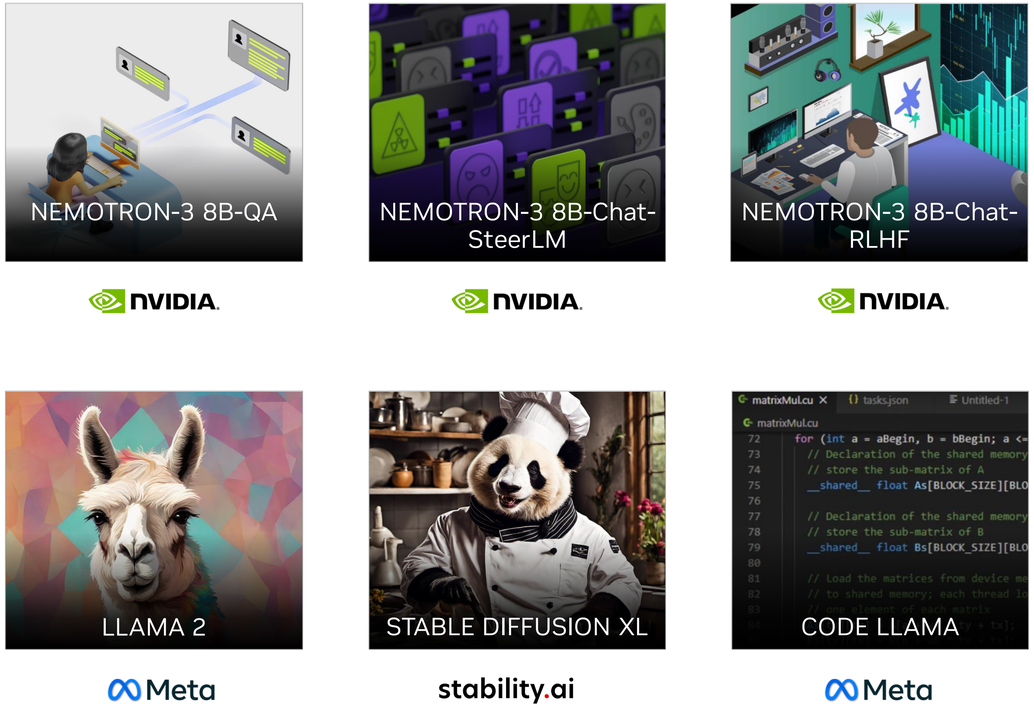
General Response
Every Google account comes with 15 GB of storage.

Enterprise Customized Model

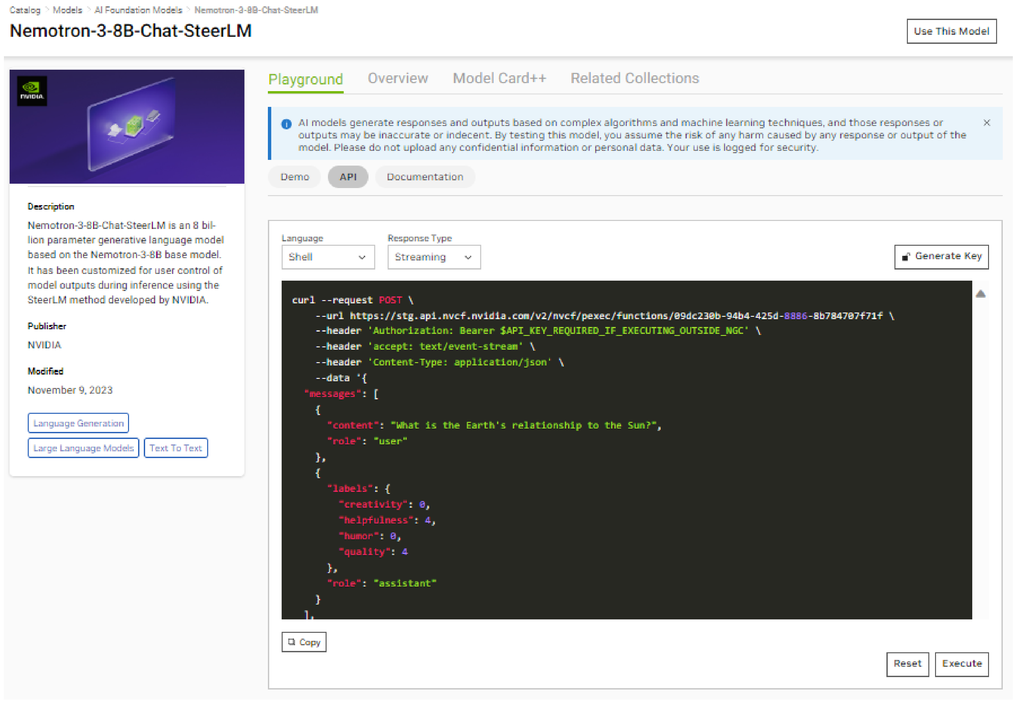
Specific Response
Company employees have unlimited storage in their Google drive.

NVIDIA AI Foundation Models and Endpoints

Fast-track custom generative AI models for enterprise applications



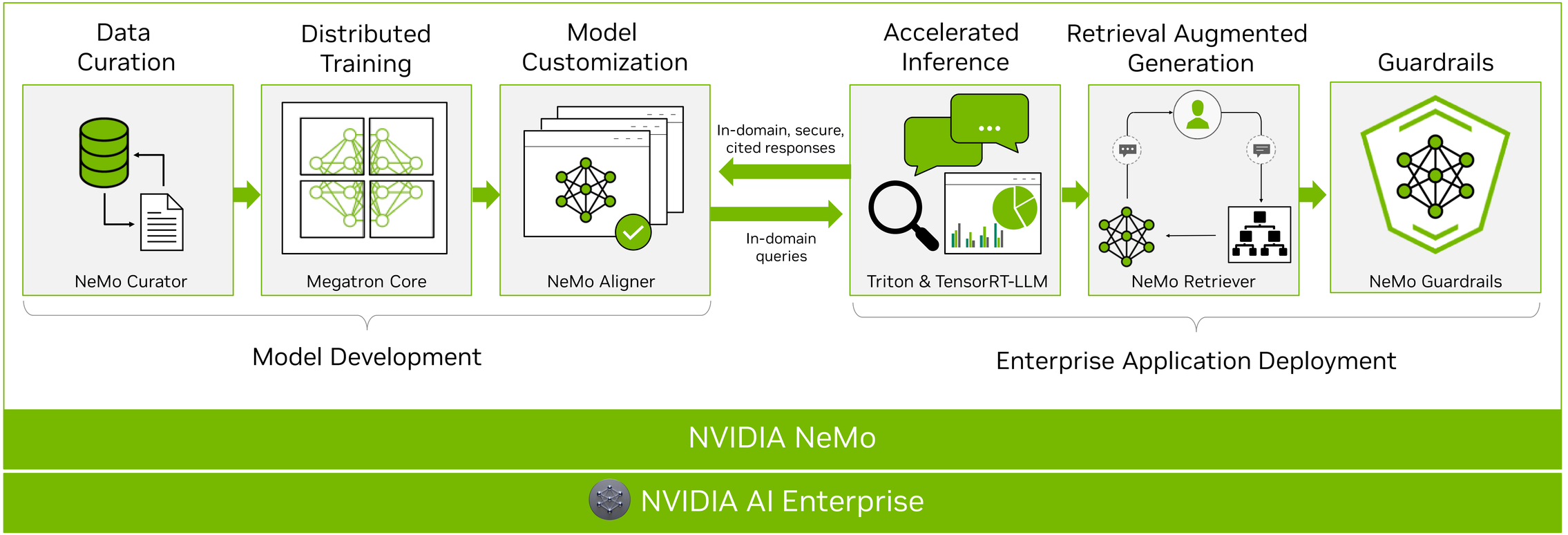
Enterprise-ready, performance optimized models from NVIDIA and the community

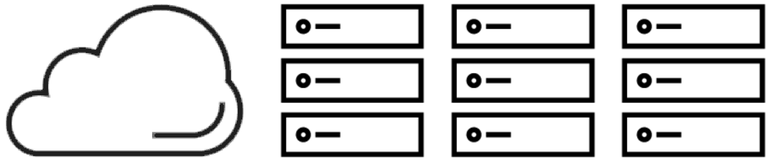


Experience foundation models running on the NVIDIA AI stack via API endpoints

Building Generative AI Applications for the Enterprise

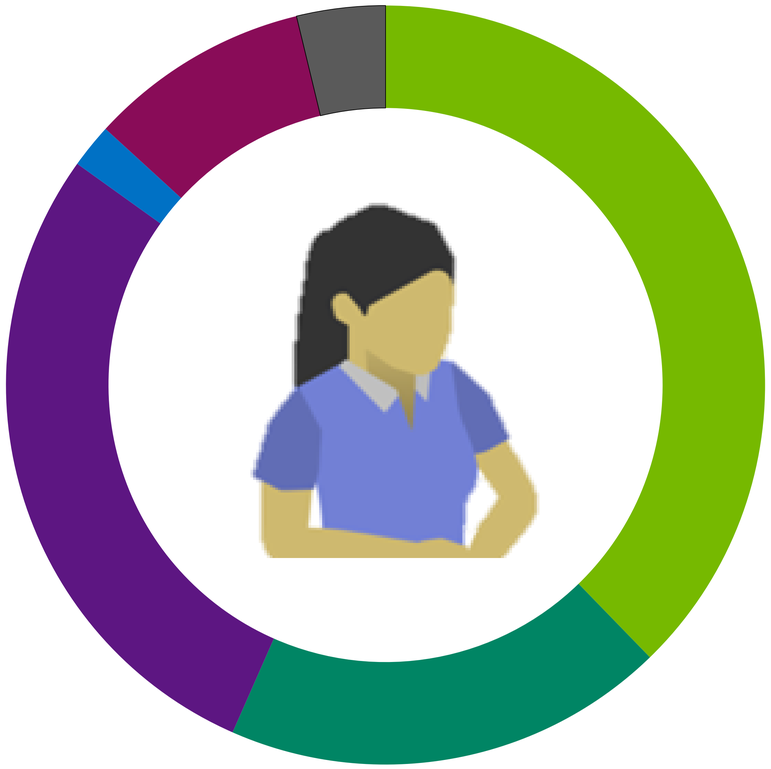
Build, customize and deploy generative AI models with NVIDIA NeMo





Why Are AI Initiatives Costing More Than Anticipated?

The cost of AI developers expending effort on non-development work



Where AI developers are spending their time

- wait on cluster provision
- stack engineering
- model optimization
- wait on resource allocation
- job launch prep
- job monitoring

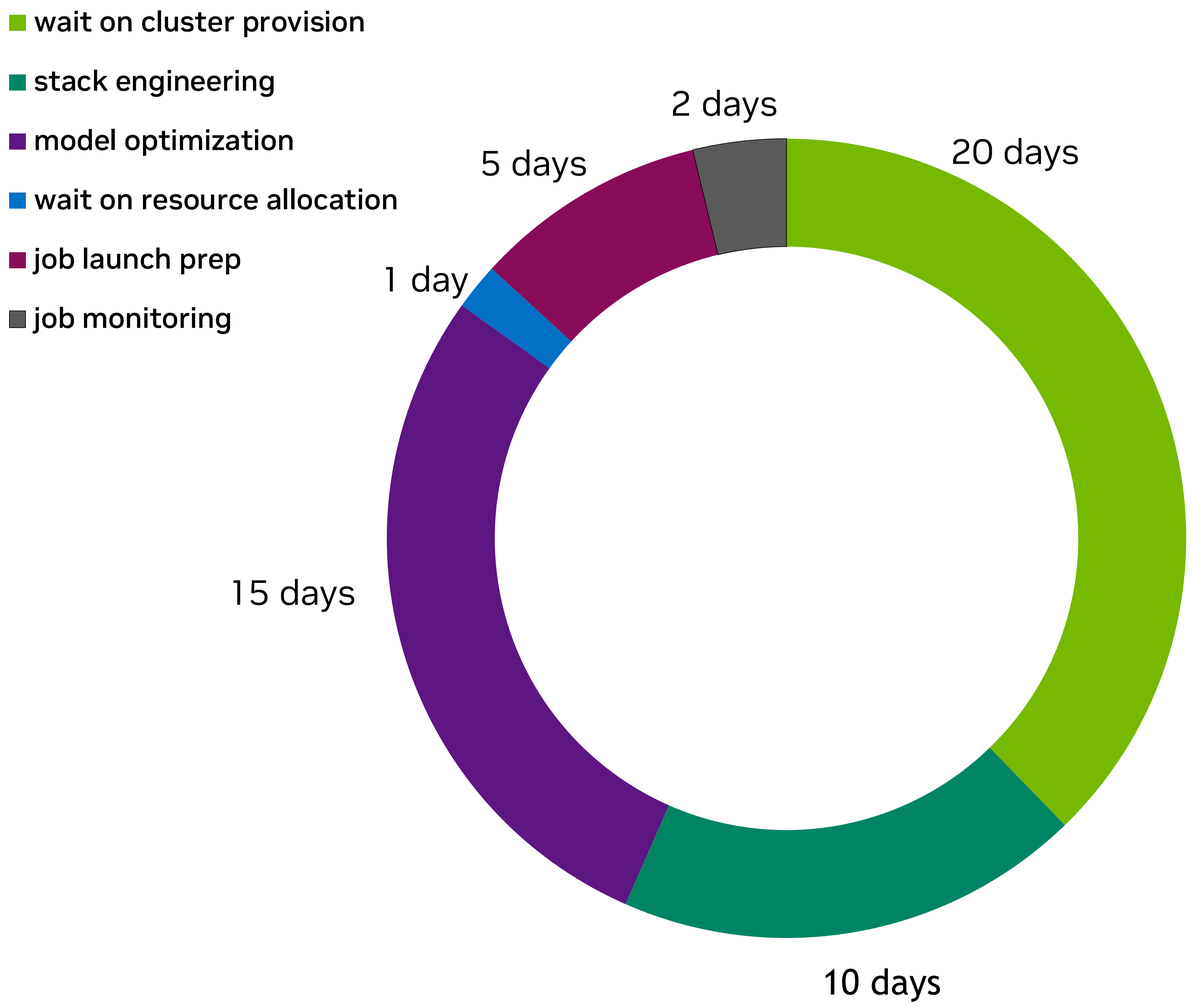
Which of these factors are impeding your develop

Why Are AI Initiatives Costing More Than Anticipated?

The cost of AI developers expending effort on non-development work

“Hidden” IaaS challenges that drive up OpEx

- Delays in infrastructure provisioning
- Effort expended on AI code modification / adaptation
- Training job troubleshooting / resource utilization inefficiency



An AI developer could lose over 30 days on non-value add “effort”

- $30k - $50k lost productivity per developer
- Potentially $1m+ across an AI team of 30 developers
- Unaccounted when observing purely the cost of infrastructure

NVIDIA DGX Cloud

Build Your Models Faster with Serverless AI on NVIDIA DGX Cloud

AI PLATFORM THAT PUTS DEVELOPERS FIRST

Easy-to-use, powerful tools for delivering **production-ready models** sooner

YOUR OWN SERVERLESS AI FACTORY

Dedicated platform for **multi-node training**, optimized for Generative AI

GET UNSTUCK

NVIDIA AI **experts are ready to help** you get better results, faster

FASTEST ROI FOR YOUR AI ENDEAVORS

Superior ROI with **maximized utilization efficiency**

DGX Cloud Solves the Challenges of Scaling AI

Delivering enterprise-scale data science effectiveness and efficiency

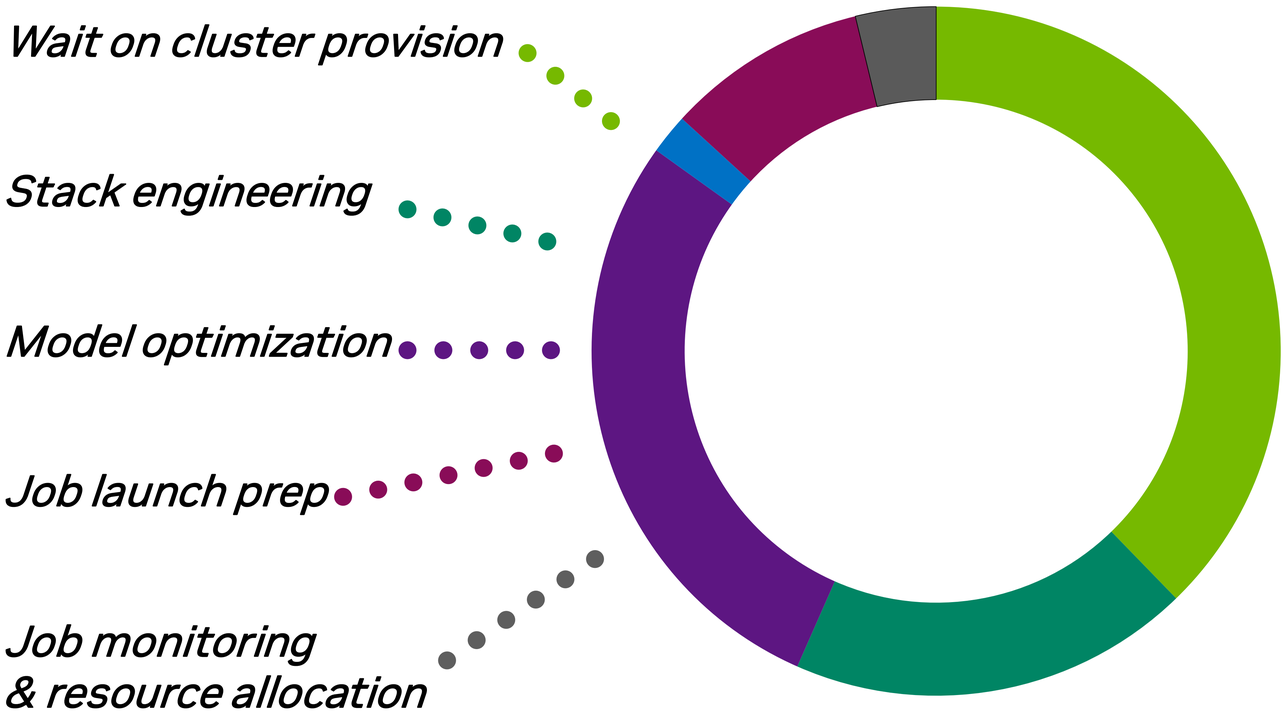
Multi-node clusters are ready, waiting for your developers NOW, not a month from now

Full-stack containers ensure compatibility and performance across layers

Includes pre-trained models that are optimized and ready to use

No refactoring of code to use service APIs

Advanced telemetry and **automated resource management** across all jobs



NVIDIA DGX Cloud for Custom LLMs

Delivering the Premium AI Training Service for the Era of Generative AI

Traditional AI development

DIY tools + open source

Inconsistent access to multi-node scale across regions

Community forums **"sweat equity"**

Escalating costs, add-on fees for reserved instances, storage, data egress etc.

Traditional AI clouds

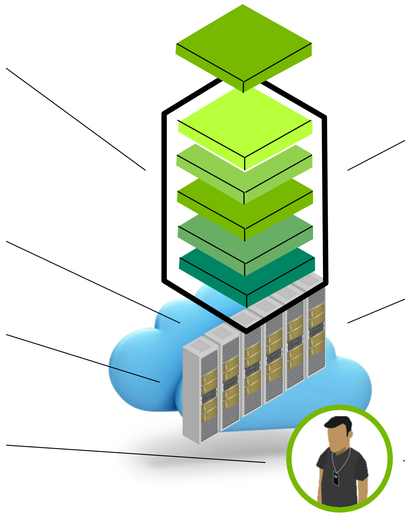
NVIDIA DGX Cloud

NVIDIA AI Enterprise
NeMo, BioNeMo, Picasso
job scheduling
and orchestration

DGX Cloud Instance
(8) A100/H100 80GB GPUs
with multi-node scale

10TB storage / instance
10TB egress

NVIDIA AI Expert



Microsoft Azure

ORACLE CLOUD Infrastructure

aws

Google Cloud

Organizations doubling-down on generative AI need:

Software that unleashes developer productivity

Multi-node training performance + easy scale

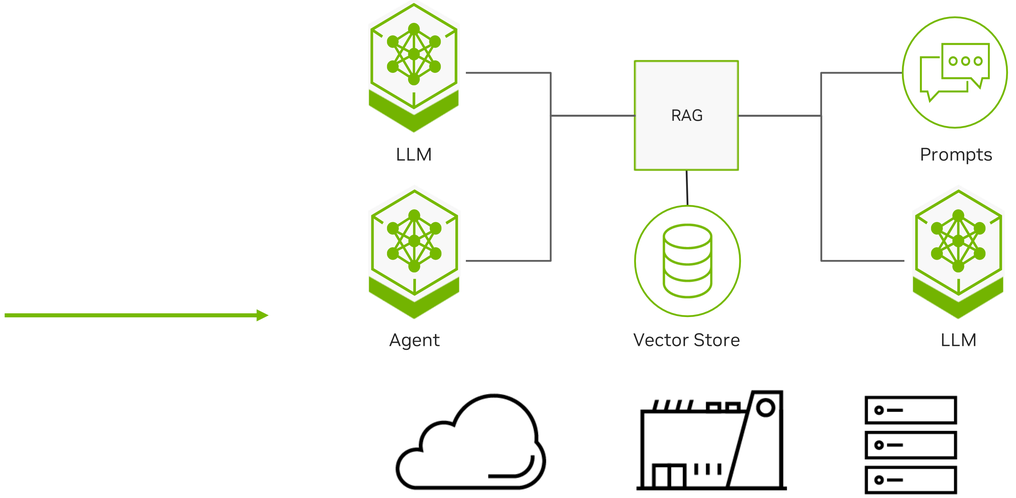
AI practitioners who help maximize performance

Predictable pricing that includes everything

Hosted in **leading clouds**

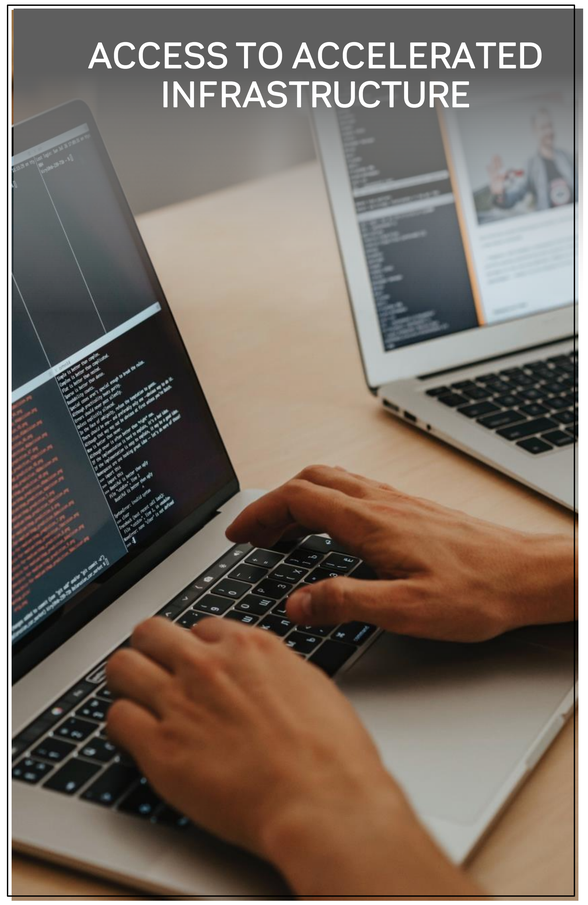
Deploy on NVIDIA AI Enterprise

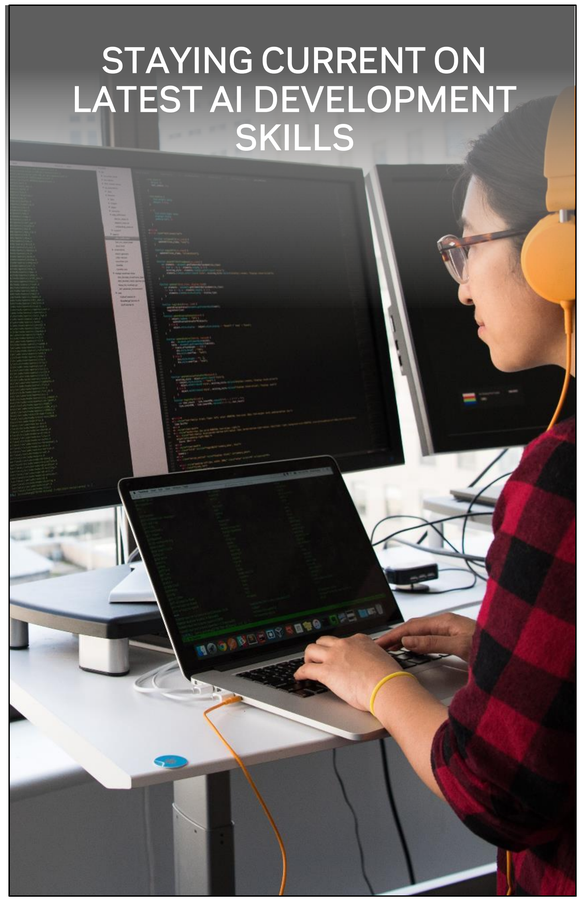
NVIDIA AI Enterprise (runtime)

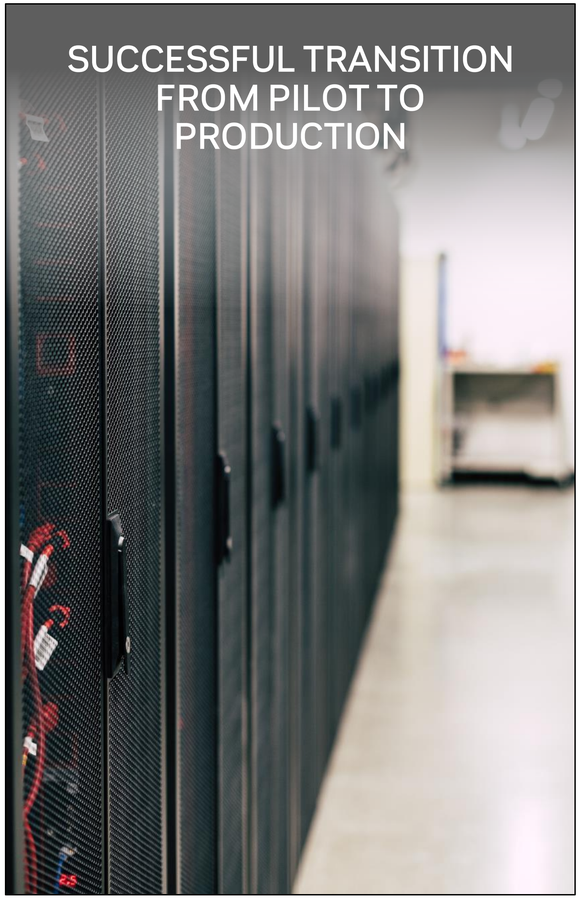


The Challenges of the Enterprise Developer

65,000 public generative AI projects created on GitHub in 2023 – a 248% YoY growth







Designed for Enterprises that Run their Business on AI

NVIDIA AI Enterprise: Production-Grade Software for AI

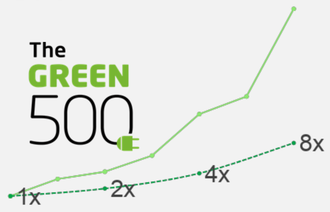
Accelerated Computing
increases productivity while lowering TCO

Generative AI

ETL/Spark

Inference

#1 MLPerf



Enterprise-Grade
security, stability, manageability & support

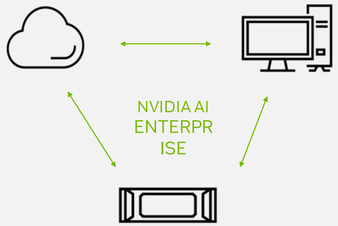
CVE Patching

API Stability

End-to-End Manageability

SLAs with NVIDIA Support

Cloud Native & Certified
to run everywhere



RTX 6000 Ada- H100 - DGX

aws

Microsoft Azure

Google Cloud

ORACLE CLOUD

DELL

Hewlett Packard Enterprise

Lenovo

...

NVIDIA

NVIDIA AI Enterprise

An Enterprise-Grade Software Platform for Your AI Runtimes

MLOps

AI Applications

NVIDIA AI Enterprise

Infrastructure Management

Cloud Native Management and Orchestration
GPU Operator, Network Operator

Cluster Management
Base Command Manager Essentials

Infra Acceleration Libraries
Magnum IO, vGPU, CUDA

Application Frameworks

LLM
NeMo

Speech AI
Riva

Cybersecurity
Morpheus

Medical Imaging
Clara

More

AI Development

Data Science / Prep
RAPIDS, RAPIDS Accelerator for Apache Spark

Model Training and Customization
NeMo, TAO, PyTorch, TensorFlow

Deploy at Scale
Triton Inference Server

Optimize for Inference
TensorRT, TensorRT-LLM

Cloud | Data Center | Workstations | Edge

Flexible Software Branches for All AI Deployments

Security, API Stability, and Peace of Mind for All Your AI Investments

NVIDIA AI ENTERPRISE

Feature Branch

Top of tree SW optimization
Monthly release cadence
CVE patches and bug fixes in roll forward release

v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v12 v13 v14 v15 v16

Production Branch

API Stability
Monthly CVE patches and bug fixes
2 branches/year with 9-month lifetime
3-month overlap between 2 PBs

v1.1 v1.2 v1.3 v1.4 v1.5 v1.6 v1.7 v1.8 v1.9

v7.1 v7.2 v7.3 v7.4 v7.5 v7.6 v7.7 v7.8 v7.9

v14.1 v14.2 v14.3 v14.4

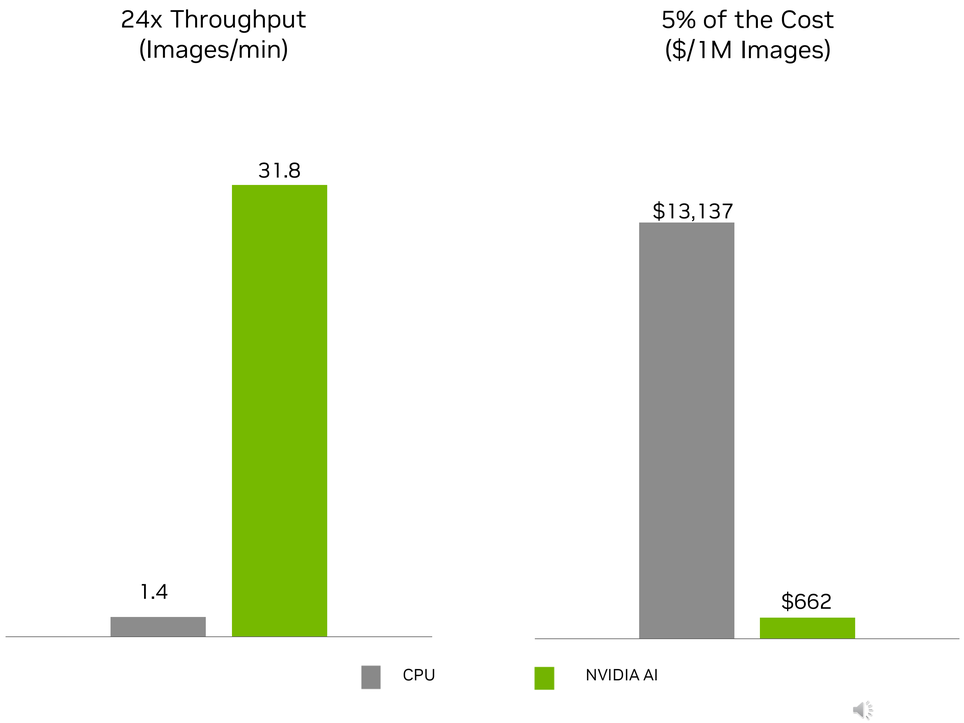
Long-Term Support Branch

For highly regulated industries
Quarterly CVE patches/bug fixes
Up to three years support
6-month overlap period

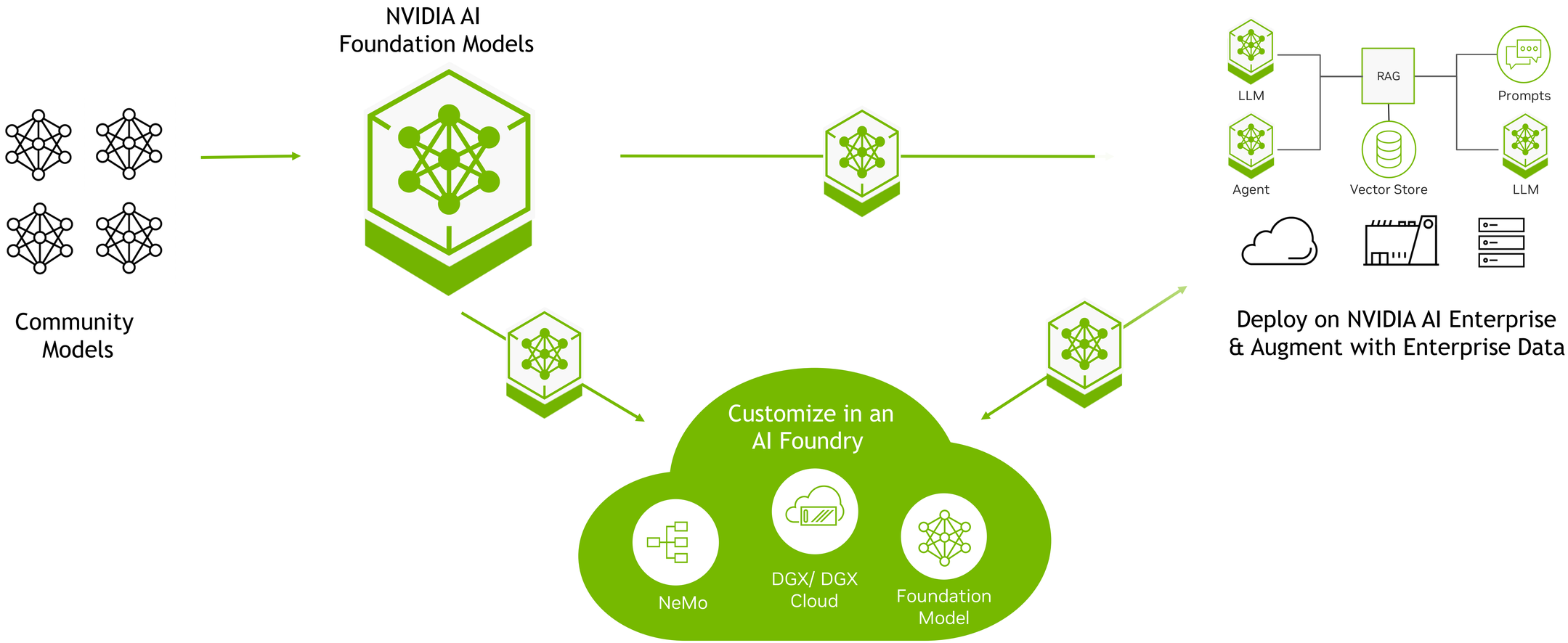
v.next v.next v.next v.next v.next v.next

Accelerated AI Improves Productivity While Lowering Total Cost

Segment Anything Model (SAM) – TensorRT Optimized



Learn More
www.nvidia.com/ai-foundation-models



Enterprises Adopting NVIDIA AI Foundry

amdocs

Dropbox

servicenow

SAP

Improve Spear Phishing Detection with AI

Learn How Generative AI Can Be Used to Detect Spear Phishing Emails Faster

Tuesday, January 30, 9:00 am PT or
Wednesday, January 31, at 10:00 am CET
Register >

Join this webinar to learn how NVIDIA's AI technologies, along with ecosystem partners, can help organizations build powerful solutions to defend against cyber threats.

Move from pilot to production for your spear phishing detection AI solution with confidence with NVIDIA AI Enterprise.

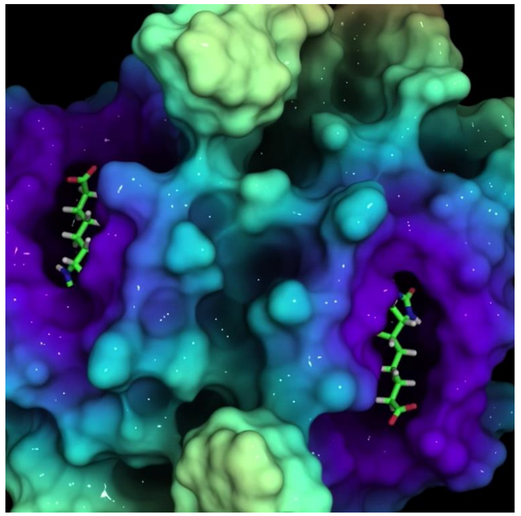
The In-Person GTC Experience Is Back

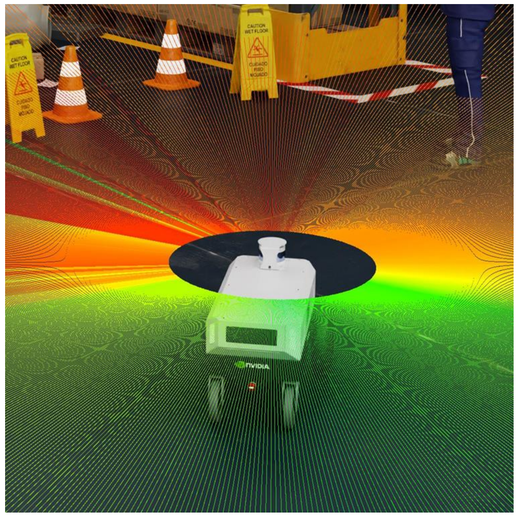
Come to GTC—the conference for the era of AI—to connect with a dream team of industry luminaries, developers, researchers, and business experts shaping what's next in AI and accelerated computing.

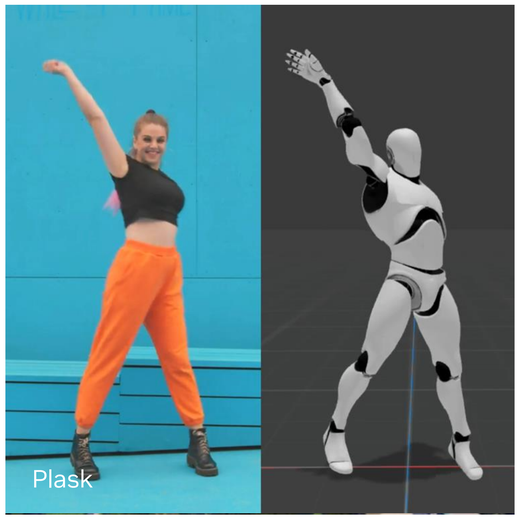
From the highly anticipated keynote by NVIDIA CEO Jensen Huang to over 600 inspiring sessions, 200+ exhibits, and tons of networking events, GTC delivers something for every technical level and interest area.

Be sure to save your spot for this transformative event. You can even take advantage of early-bird pricing when you register by February 7.

March 18-21, 2024 | www.nvidia.com/gtc







Q&A