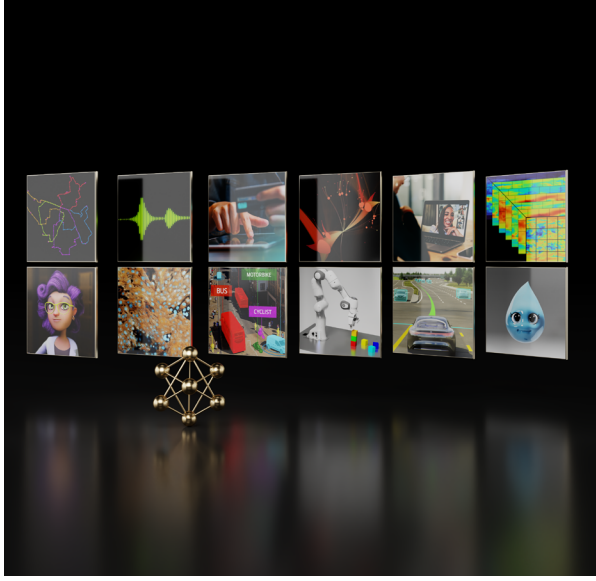
Solution Overview

NVIDIA AI Enterprise

Build production AI with an enterprise-grade software platform.

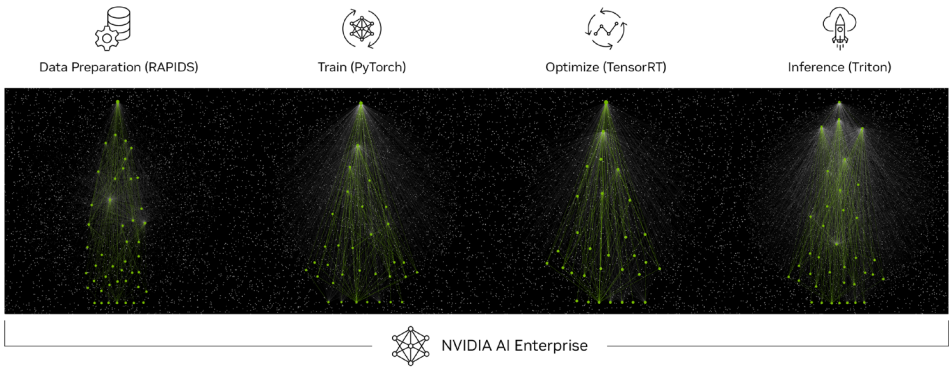


The Challenges of Building and Maintaining an AI Software Platform

AI is profoundly changing how business is done, and while organizations understand the transformative potential of AI, the implementation of this rapidly evolving technology is challenging.

For many enterprises, maintaining a consistent, secure, and stable software platform for building and running AI is a complex undertaking. Consider the foundation software used in AI, which includes over 4,500 unique software packages, 64 of which are NVIDIA® CUDA® libraries and more than 4,471 are third-party and open-source software (OSS) packages.

With this number of software packages, there's a high risk of introducing security vulnerabilities. What's more, maintaining API stability with the 10,000 dependencies between these unique software packages is another challenge, making it nearly impossible for enterprises to reliably use the latest open-source versions for their deployment environments.



NVIDIA AI Enterprise foundation layer. The graph shows unique software (dots) and dependencies (lines).

Enterprise-Grade Software for Accelerated AI

Security, reliability, and manageability are critical for enterprise-grade AI. NVIDIA AI Enterprise is an end-to-end, cloud native software platform that accelerates the data science pipeline and streamlines development and deployment of production-grade AI applications, including generative AI, computer vision, speech AI, and

Challenges of Production AI

- **Complexity:** Pulling together an end-to-end AI software stack from disparate, open-source software—and integrating them with existing infrastructure—is difficult.
- **Risk:** The AI software stack consists of thousands of open-source packages and dependencies, making security patching a challenge.
- **Reliability:** Maintaining a high-performance AI platform and managing API stability across the stack are critical for investment protection and business continuity.

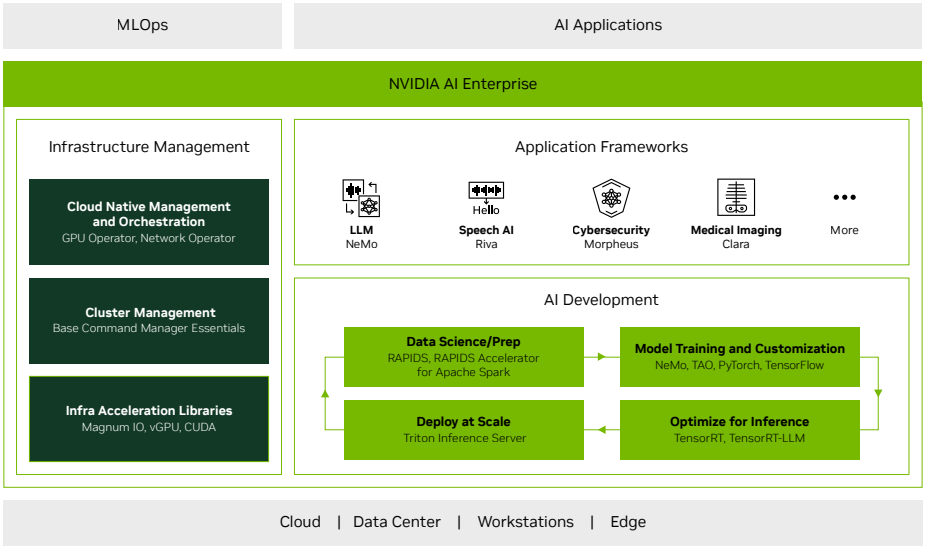
Benefits of NVIDIA AI Enterprise

- Improves productivity and lowers costs with accelerated computing.
- Frees teams to build innovative AI solutions with enterprise-grade security, reliability, and support.
- Is cloud-native and certified to run anywhere and on current and prior GPU generations.
- Speeds time to production with AI workflows and pretrained models.

more. Enterprises that run their businesses on AI rely on the security, support, and stability provided by NVIDIA AI Enterprise to improve productivity of AI teams, reduce total cost of AI infrastructure, and ensure a smooth transition from pilot to production. NVIDIA AI Enterprise includes:

- NVIDIA NeMo™, an end-to-end framework for building, customizing, and deploying enterprise-grade generative AI models; NeMo lets organizations easily customize pretrained foundation models—from NVIDIA and select community models—for domain-specific use cases.
- Continuous monitoring and regular releases of security patches for critical and common vulnerabilities and exposures (CVEs).
- Production branches and long-term support branches that ensure API stability.
- End-to-end management software, including cluster management across cloud and data center environments and cloud-native orchestration.
- Enterprise support with service-level agreements (SLAs) and access to NVIDIA AI experts.

NVIDIA AI Enterprise relieves organizations of the burden of maintaining and securing the complex software platform of AI, freeing them to focus on building AI and harnessing its game-changing insights.



Accelerated AI Improves Productivity and Lowers TCO

With an extensive catalog of AI frameworks, pretrained models, and development tools optimized for building and running AI on NVIDIA GPUs, NVIDIA AI Enterprise accelerates every stage of the AI journey, from data prep and model training through inference and deployment at scale:

- Accelerate data processing up to 5X and reduce operational costs up to 5X over CPU-only platforms with the NVIDIA RAPIDS™ Accelerator for Apache Spark.
- Train at scale with the NVIDIA TAO Toolkit. Create custom, production-ready AI models in hours, rather than months, by fine-tuning NVIDIA pretrained models—without AI expertise or large training datasets.

Certified Platforms

- **Multi-cloud:** Major cloud marketplaces, including **AWS**, **Microsoft Azure**, **Google Cloud**, and **Oracle Cloud Infrastructure**.
- **Hybrid cloud:** VMware Cloud Foundation, Red Hat Enterprise Linux, HPE GreenLake, Ubuntu KVM, and Nutanix AHV.
- **MLOps integration:** Technical preview on Azure Machine Learning and certifications with ClearML, Domino Data Lab, Run:ai, UbiOps, and Weights & Biases.
- **Container orchestration:** VMware Tanzu, Red Hat OpenShift, HPE Ezmeral, Google Kubernetes Engine (GKE), Amazon Elastic Kubernetes Service (EKS), Azure Kubernetes Service (AKS), and upstream Kubernetes.
- **NVIDIA DGX platform:** NVIDIA AI Enterprise license is included with DGX software, part of the NVIDIA DGX platform, to supercharge AI training for large language models.
- **NVIDIA-Certified Systems™:** Over 400 NVIDIA-Certified servers and workstations are available from a wide range of manufacturers. Software license of NVIDIA AI Enterprise is included with each NVIDIA H100 PCIe or NVL GPU and NVIDIA A800 40GB Active GPU.

Explore Licensing Options

- Software licenses can be purchased as perpetual, subscription, or through cloud marketplaces. Learn more about **product pricing and licensing**.

- Accelerate large language model (LLM) inference performance up to 8X with NVIDIA TensorRT™-LLM and inference performance up to 40X with TensorRT over CPU-only platforms, lowering infrastructure and energy costs. .
- Deploy at scale with NVIDIA Triton™ Inference Server, which simplifies and optimizes the deployment of AI models at scale and in production for both neural networks and tree-based models on GPUs.

Reduce Development Time With NVIDIA AI Workflows

AI workflows are cloud-native, prepackaged reference examples for enterprises to jumpstart the building of AI solutions, including: generative AI chatbots that generate accurate responses by retrieving information in real-time from a company's knowledge base, intelligent virtual assistants, cybersecurity solutions for detecting insider threats, using generative AI to improve spear phishing email detection, and more.

They can run as microservices and can be deployed on Kubernetes alone or with other microservices to create production-ready applications. NVIDIA AI workflows can accelerate the path to delivering AI outcomes, reduce time to deployment, lower costs, and improve accuracy and performance.

Cloud Native and Certified to Run Everywhere

NVIDIA AI Enterprise is optimized and certified for reliable performance, whether it's deployed in the public cloud, in virtualized data centers, or on the NVIDIA DGX™ platform. This provides the flexibility to develop applications once and deploy anywhere, reducing the risk of moving from pilot to production that's caused by infrastructure and architectural differences between environments.

NVIDIA's Ecosystem of Partners

- MLOps solution providers for optimizing the AI and machine learning pipeline
- Global solution integrators for customized state-of-the-art AI solutions
- Service delivery partners for applied AI service

Ready to Get Started?

To learn more about NVIDIA AI Enterprise, visit:
nvidia.com/ai-enterprise-suite

To sign up for a free 90-day evaluation license, visit:
nvidia.com/ai-enterprise-eval

To get hands-on experience with NVIDIA AI Enterprise, apply for a free lab through NVIDIA LaunchPad at: **nvidia.com/try-ai**

Contact Sales at: **nvidia.com/ai-enterprise-sales**

NVIDIA