



# Bringing AI Inference to the Desktop

Accelerate AI inference with NVIDIA RTX PRO AI Workstations.



## The Challenges of AI-Augmented Workflows

AI is bringing profound change across industries, accelerating the adoption of AI-infused technologies at an incredible scale. These AI-powered workflows offer industries the promise of new levels of creativity and productivity and improved efficiency.

Harnessing the capabilities of today's sophisticated generative AI models requires significantly more computing power than ever before. These models are not only massive—often taking weeks or months to train on large server clusters—but also computationally demanding during inference. Unlike earlier AI systems, they process queries by generating numerous intermediate reasoning steps or internal tokens before arriving at a final response. This internal “thinking” can produce hundreds or even thousands of tokens per query, especially for complex tasks, greatly increasing the compute requirements and costs associated with inference. This compute-intensive approach, however, enables the models to deliver more accurate, context-aware, and human-like outputs—unlocking use cases that were previously out of reach. As a result, delivering real-time, high-quality AI outputs—such as natural conversations, realistic images, or personalized content—places immense pressure on underlying infrastructure.

Data center and cloud resources need to be expanded to take on new AI inferencing workloads. Increasing data center capacity or acquiring additional cloud instances can be prohibitive with respect to cost and hardware availability. To harness the power of generative AI on the desktop, businesses are discovering that their traditional desktop computing solutions are inadequate for these new AI-powered tools and applications.

## NVIDIA RTX PRO AI Workstations for Inference

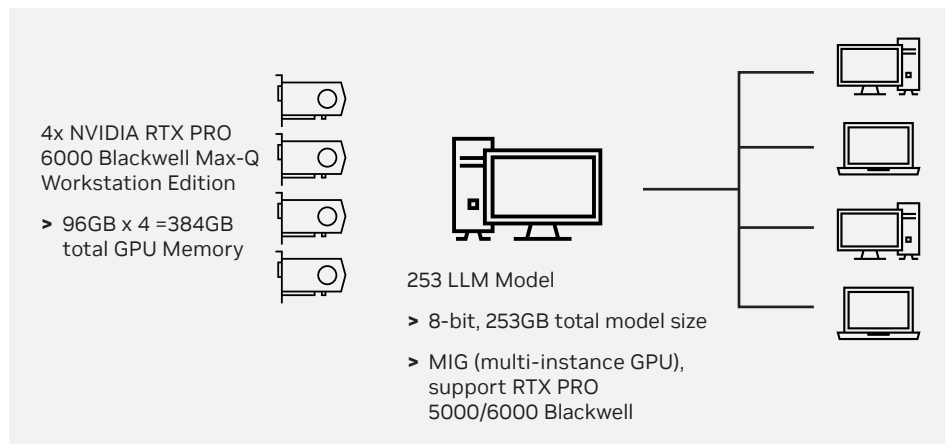
Inference is the stage in the AI workflow where a trained model is deployed to process new data and generate predictions, classifications, or other outputs. Large generative AI models require substantial data center and cloud resources to provide inferencing to large numbers of users with acceptable response times. Significant efforts are being devoted to creating smaller generative AI models that still provide high levels of accuracy. These smaller models can be run on smaller systems outside of the data center or cloud.

### Key Challenges for AI Inference

- **Workflow complexity:** Modern professional workflows require running multiple applications simultaneously to maximize productivity. Adding AI-augmented tools and applications puts additional requirements on current computing solutions.
- **Scaling:** Enhancing workloads with AI-augmented tools and applications requires the latest hardware to take advantage of the latest technology.
- **Network Latency:** Delays from transferring data between distributed servers can slow down AI inference performance and reduce real-time responsiveness, particularly for latency-sensitive workloads.

NVIDIA RTX PRO™ AI Workstations can run smaller model inferencing workloads. The latest generation of Original Equipment Manufacturer (OEM) workstations provides for up to four RTX PRO 6000 Blackwell Max-Q Workstation Edition GPUs per workstation for an incredible 14.2 petaFLOPS of combined compute performance and 384 gigabytes (GB) of total system GPU memory for local inferencing.

A multi-GPU system can also augment data center or cloud resources by efficiently handling the inferencing needs of small teams or workgroups, allowing for local AI model deployment and faster iteration.



AI workstations for workgroup AI inference serving.

## Inference for AI-Augmented Apps

Applications with AI-enabled features have been available for several years, such as Adobe® Photoshop's® Neural Filters, DaVinci Resolve's face tracking, NVIDIA Broadcast's noise and room echo removal, and image denoising in every major rendering application software. Workstations equipped with RTX PRO GPUs have been the platform of choice for modern AI-powered workflows.

As AI brings new levels of capabilities and efficiency, workflows require more computing power and GPU memory. Professional users work with high-resolution content, utilizing workflows that require the simultaneous use of multiple professional applications. As professional workflows include AI-augmented tools to help with concept development and creation, these compute- and memory-intensive applications will put additional demands on the GPU. Running large language models (LLMs) like chatbots and code copilots locally on a workstation further amplifies these requirements, making powerful GPUs essential to ensure smooth performance and efficient multitasking. NVIDIA RTX PRO AI Workstations are built for these demanding workloads. The NVIDIA RTX PRO 6000 Blackwell Workstation Edition and NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition GPUs, with 96 GB of GPU memory each, have the raw AI computing power and memory necessary to work with high-resolution AI content, iterate, and pass content on to other design or creative applications without needing to shut down other applications or reduce content fidelity.

### Benefits for Inference on RTX PRO AI Workstations

- > AI workloads can be offloaded to RTX PRO AI Workstations for real-time data processing, visualization, and response, augmenting data center and cloud resources.
- > Large GPU memory configurations enable AI-augmented, multi-application workflows that maximize productivity. Workstation solutions let businesses increase system capabilities as their workflows expand.
- > RTX PRO AI Workstations with the NVIDIA AI software stack allow for rapid prototyping, model fine-tuning, and iterative development locally, with the flexibility to deploy at scale when required.
- > Enterprise-grade hardware maximizes uptime with enterprise-level performance, reliability, and support.

# Build and Test AI-Augmented Applications

NVIDIA RTX PRO AI Workstations enable developers to build, test, and iterate on AI-augmented applications with speed and efficiency. Equipped with high-performance GPUs and large memory capacity, these systems support complex model experimentation locally—augmenting cloud or data center resources and accelerating development timelines. Whether developing chatbots, generative AI features, or computer vision modules, developers can prototype and debug AI-powered functionality directly within the workstation environment.

With the NVIDIA AI Enterprise software full-stack solutions, workstations offer an integrated platform for model training, fine-tuning, testing, and deployment readiness. Multi-GPU configurations further enable shared access across teams, supporting collaborative workflows in R&D, application development, and software QA. This allows teams to develop in parallel, test models in real time, and transition seamlessly to production deployments in the cloud or data center—enhancing productivity across the entire AI application lifecycle.

## Local Inference for Edge Development

NVIDIA RTX PRO AI Workstations allow organizations to develop, test, and validate AI inference pipelines locally, simulating real-world edge or on-premises deployments. Developers can run quantized or optimized models using tools like [NVIDIA TensorRT™](#), a full-stack inference SDK with open-source tools and components, allowing for performance benchmarking and compatibility testing before models are deployed to edge devices or production infrastructure.

This localized environment supports real-time performance analysis, latency measurement, and throughput validation under conditions where cloud access is limited or data privacy is critical. By replicating edge deployment scenarios, teams can accelerate iteration, reduce reliance on cloud resources, and confidently deploy across distributed environments. From robotics and manufacturing inspection to medical imaging and smart city applications, RTX PRO AI Workstations offer a scalable, secure, and efficient platform for edge AI development.

## Enterprise-Class Solutions



NVIDIA RTX PRO AI Workstations are based on the latest generation of workstation platforms and are readily available from worldwide workstation partners. AI workstations provide the enterprise-class performance, reliability, and support required for mission-critical enterprise deployments.

With a full stack of enterprise-level deployment, support, and optimization tools, AI workstations easily fit into existing IT infrastructure, providing drop-in solutions for AI inferencing on the desktop.

## Ready to Get Started?

Learn more about NVIDIA RTX PRO AI Workstations at [nvidia.com/ai-workstations](https://nvidia.com/ai-workstations).

Contact NVIDIA sales at [nvidia.com/en-us/contact/sales](https://nvidia.com/en-us/contact/sales).

AI-Enabled Applications		
		
	Desktop GPU	Laptop GPU
Best	RTX PRO 6000 Blackwell Workstation Edition, RTX PRO 6000 Blackwell Max-Q Workstation Edition, Multi-GPU	NVIDIA RTX PRO 5000 Blackwell Generation
Better	RTX PRO 5000 Blackwell, RTX PRO 4500 Blackwell	NVIDIA RTX PRO 4000, 3000 Blackwell Generation
Good	RTX PRO 4000 Blackwell	NVIDIA RTX PRO 2000, 1000, 500 Blackwell Generation

