



NVIDIA AI Factory for Agentic AI Workloads

Accelerate AI agent development with the NVIDIA AI Factory

The Next Wave of AI

AI agents are the next evolution of artificial intelligence—combining sophisticated reasoning, iterative planning, and adaptability to autonomously solve complex, multi-step problems. By ingesting vast amounts of data from multiple data sources, AI agents can analyze challenges, develop strategies, and complete tasks independently. They are able to orchestrate resources, collaborate with other agents, and utilize tools to become an integral part of your digital workforce.

AI Agents Across Industries

Organizations in every sector are rapidly adopting teams of AI agents—not just to solve business challenges, but to unlock new competitive advantages in today's fast-paced market.

[NVIDIA AI Enterprise](#) is the software platform that makes it possible to build and deploy AI agents anywhere. Organizations can accelerate their AI journey from prototype to production with tools like [NVIDIA NIM™](#) microservices for rapid model deployment, [NVIDIA NeMo™](#) for enhancing agent performance with data flywheels, and [NVIDIA AI Blueprints](#) to jumpstart agent development with prebuilt, customizable workflows.

AI agents can span a wide range of industries and functions, including:

- > **Supply chain management:** Multi-agent systems can help optimize supply chain operations by analyzing data in real time, monitoring and adjusting inventory levels based on demand, and helping source raw materials by keeping an eye on market fluctuations.
- > **Customer service:** AI agents are improving customer support by enhancing self-service capabilities and automating routine communications. This can significantly improve customer interactions, reducing response times and boosting satisfaction.
- > **Marketing:** AI agents can be leveraged for content creation to autonomously generate, optimize, and manage various types of media, like articles, blog posts, and videos, and personalize content for specific audiences.
- > **Video analytics:** Video analytics AI agents analyze live and archived footage, enable natural language tasking, and support video search, summarization, anomaly detection, incident reporting, quality control, and predictive maintenance.



Key Challenges for Agentic AI Workloads:

- > **Complexity:** model optimization, sophisticated workflows across multiple servers
- > **Production readiness:** cost, manageability, security, maintenance
- > **Infrastructure diversity:** data center, edge, workstation, cloud

These challenges are addressed by NVIDIA's robust, scalable, and flexible AI factories that can run every agentic AI workload—from data retrieval and processing to reasoning and iteration.

From Data Centers to AI Factories

Building and deploying AI agents brings a unique set of challenges—from orchestrating complex reasoning and design-making workflows to managing the sheer scale of data and compute resources required. These agents generate significantly more tokens because they interact, plan, and iterate toward a goal, resulting in higher demands on infrastructure.

To meet the demands of agentic AI workloads, organizations need more than powerful models. They need an AI factory—infrastructure that is purpose-built to transform raw data into intelligence at scale. These factories optimize the entire AI lifecycle and are essential for organizations looking to power mission-critical AI workloads.

Enter the NVIDIA AI Factory

The [NVIDIA AI Factory](#) is a full-stack solution that combines cutting-edge accelerated computing, advanced networking, and NVIDIA AI Enterprise software with enterprise data to maximize efficiency, scalability, and cost-effectiveness. It leverages a broad ecosystem of partners with validated third-party software that drives all feature requirements of an end-to-end platform for agentic AI.

With the NVIDIA AI Factory, organizations can build and deploy AI agents while gaining:

- > **Faster time to value:** Accelerate time to intelligence with composable, day-one-ready building blocks that include pre-engineered rack-level designs, secure AI, and an integrated software stack.
- > **Greater performance and energy efficiency:** Optimize AI performance with NVIDIA accelerated computing, which delivers more tokens per watt and greater energy efficiency across AI factories and applications.
- > **Broad partner ecosystem:** Deploy AI factories on premises and in the cloud, using integrated, full-stack solutions based on NVIDIA Accelerated Computing and NVIDIA AI software.
- > **Scalable AI deployment:** Drive scalable intelligence manufacturing as your organization grows, with modular upgrades that maximize your AI investments and long-term returns.
- > **Reduced complexity and total cost of ownership (TCO):** Streamline AI agent deployment with efficient designs and validated component compatibility for faster, more reliable AI infrastructure setup.

Microway

Microway delivers full stack AI hardware & software deployments spanning GPU clusters, servers, workstations, networking, and storage. As an NVIDIA DGX SuperPOD Specialization Partner and an NPN Elite Provider of NVIDIA DGX and NVIDIA Data Center GPU solutions, we focus on delivering the most complex, high-performance AI & HPC infrastructure to customers throughout North America. Trust in our expertise! We have been delivering NVIDIA Data Center GPU solutions for decades—since before their inception.

Ready to Get Started?

To learn more about NVIDIA AI Factories visit:
nvidia.com/en-us/solutions/ai-factories/

To learn more about Microway visit:
<https://microway.com/platforms/dgx/nvidia-dgx-superpod/>

© 2025 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, NIM, and NeMo are trademarks and/or registered trademarks of NVIDIA Corporation and affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective owners with which they are associated. 4145050. AUG25

“ Microway is one of the best vendors I have ever worked with. Everyone demonstrated the highest level of professionalism throughout the GPU cluster purchase process and installation. I look forward to doing business with Microway again in the future ”

Deng Cao, PhD,
Chair & Professor of Computer
Science, Central State University

