HPE | NVIDIA

AI FACTORIES IN ACTION

Real-world stories and results

Get started →

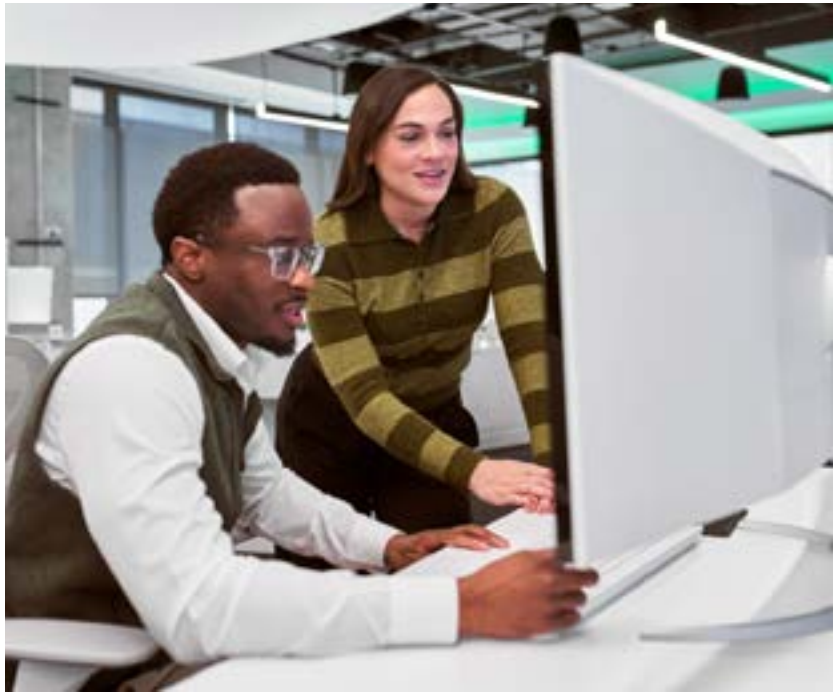
Building trusted AI for the future

The organizations that capture the most value from AI are not simply experimenting with models. They are building the capability to develop, train, fine-tune, deploy, and govern AI reliably, close to the data and at operational scale. The result is AI capabilities that can contribute to faster scientific discovery, more resilient digital services, safer industrial operations, and better results for people and communities.

As AI expectations rise, so do the technological requirements. AI infrastructure has to support performance at scale. That's at a minimum. But for AI to contribute meaningfully to enterprise and societal aspirations, it also has to be trusted, resilient, and governable.

This shift is driving the rise of the AI factory. An AI factory is not a single product or deployment; it's an end to end AI capability under the right operational and sovereignty constraints. Built on accelerated computing and delivered as fully integrated systems available in three configurations,it brings together compute, networking, software, cooling, and services into a repeatable foundation for production AI.

The following success stories show how organizations are putting this model into practice. Across industries, geographies, and missions, these teams are using HPE AI factory with NVIDIA platforms to move faster while reducing risk, design more efficient products, power secure digital services, advance scientific research, and support outcomes that directly improve quality of life. Each story focuses on what it takes to make AI real, sustainable, and impactful today.



TELUS advances sovereign AI innovation

Canada's first fully sovereign AI factory gives businesses, researchers, and public institutions access to a cutting-edge AI factory

TELUS has opened Canada's first fully sovereign AI factory in Rimouski, Quebec, establishing a national foundation for building, training, and deploying AI while keeping data entirely under Canadian control. The facility, powered by an HPE Sovereign AI Factory is 100% Canadian owned and operated, with sovereignty enforced across data residence, security, and operational governance.

Designed as a true end to end AI platform, the TELUS Sovereign AI Factory supports the AI lifecycle from model training and fine tuning to deployment and inference, enabling organizations to move from experimentation to production within a single, trusted environment. The facility is also built with sustainability in mind, running on 99% renewable energy and leveraging TELUS' PureFibre network to deliver low latency, secure connectivity.

Positioned as a cornerstone of Canada's AI strategy, the Rimouski factory, officially the TELUS Sovereign AI Factory, demonstrates that sovereignty is not just where the infrastructure sits. It is how data residency, security, operational governance, and AI lifecycle control are built into the platform from training through deployment.

Key achievements

- **First** fully sovereign AI factory in Canada
- **100%** Canadian controlled and operated
- **99%** renewable energy

> "We are creating the backbone for Canada's productivity, competitiveness, and global leadership in the digital era."
>
> —Darren Entwistle, president and CEO of TELUS

Learn more: TELUS opens Canada's first fully sovereign AI factory

KDDI makes AI more attainable and sustainable

The Osaka Sakai Data Center supports startups and enterprises with resources for developing and training AI

HPE and KDDI are collaborating on large-scale AI data center operations at the Osaka Sakai Data Center, expanding Japan's capability to develop and deploy advanced AI at production scale. Built on an HPE AI Factory with NVIDIA, the deployment will feature a rack-scale NVIDIA® GB200 NVL72 by HPE, powered by the latest NVIDIA Blackwell architecture.

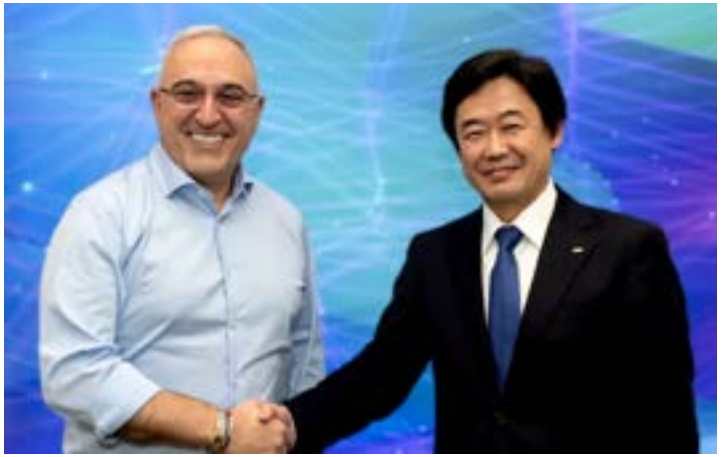
The facility is designed to support startups and enterprises developing AI applications and training large language models (LLMs), while also enabling low-latency inference for real-world deployment.

Beyond infrastructure, KDDI plans to offer cloud-based AI services through WAKONX, its business platform for the AI era, helping accelerate AI adoption across Japan and globally.

To meet the intense performance and energy demands of modern AI workloads, HPE is applying its direct liquid cooling expertise, combined with hybrid cooling that reduces the data center's environmental impact. The Osaka Sakai Data Center reflects a broader shift toward at-scale AI factories that integrate compute, networking, software, and cooling into a single, optimized system, enabling organizations to move AI into production faster while improving efficiency and sustainability.

"HPE's deep expertise in supercomputing and advanced cooling technologies will be instrumental in driving the evolution of next-generation AI data centers."

—Hiromichi Matsuda, president and CEO of KDDI



System architecture

— NVIDIA GB200 NVL72 by HPE

— HPE hybrid cooling innovations to reduce environmental impact

— NVIDIA-accelerated networking, including NVIDIA Quantum-2 InfiniBand, NVIDIA Spectrum-X Ethernet, and NVIDIA BlueField-3 DPUs

Learn more: KDDI and HPE join forces to launch AI data center operations by early 2026

HLRS and Argonne National Laboratory advance AI

AI factories at national research centers drive sovereign AI leadership in Europe and the United States

HPE is advancing sovereign AI initiatives in Europe and the United States by delivering at scale, liquid cooled AI factories for leading national research institutions: the High Performance Computing Center Stuttgart (HLRS) in Germany and Argonne National Laboratory in the U.S.

These systems combine secure infrastructure, accelerated computing, and energy efficient design to help governments and research institutions deploy AI at scale while meeting strict data governance and compliance requirements.

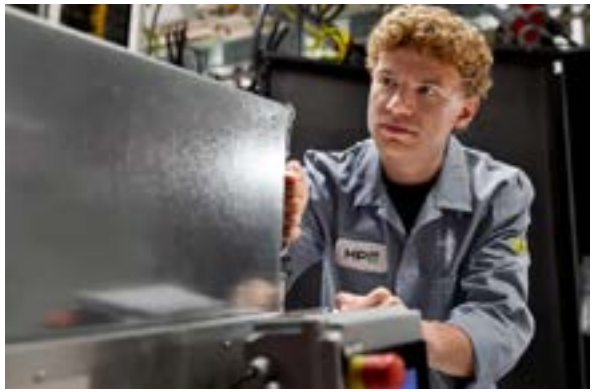
At HLRS, HPE will build and install the system for HammerHAI, the European Union's AI factory, supporting advanced machine learning, engineering, and research across academia and industry. In the U.S., new Janus and Tara systems at Argonne National Laboratory will expand AI training and inference capabilities, accelerating scientific discovery and workforce development.

"This integrated approach will help researchers, startups, and enterprises access AI resources while operating in alignment with European Union data security requirements."

—Dr. Bastian Koller, managing director of HLRS and lead coordinator of HammerHAI

"This integration accelerates discovery at every step, transforming not only the speed but also the way scientists approach their problems."

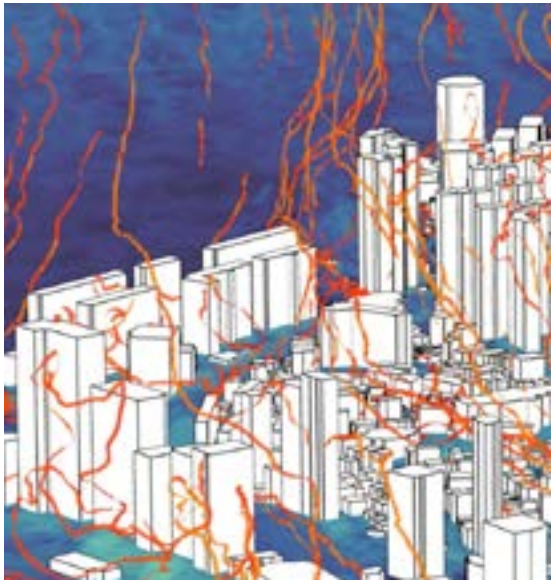
—Rick Stevens, associate laboratory director for Computing, Environment and Life Sciences at Argonne



Project focus

- Provide national research ecosystems with the performance, sustainability, and control needed to move from experimentation to production.
- Deliver some of the world's fastest and most energy efficient platforms.
- Enable organizations to develop, train, and run AI while retaining control of data and IP.

Learn more: HPE drives sovereign AI leadership with advanced systems



Yonsei University and KISTI take steps to make air taxis a reality

Real-time wind modeling made possible

Researchers at Yonsei University and Korea Institute of Science and Technology Information (KISTI) are advancing urban air mobility by developing real-time, high-resolution wind modeling for dense city environments. Using an AI factory, they combine large-scale simulation, accelerated computing, and AI-driven surrogate modeling to analyze complex airflow dynamics across urban terrain.

The platform processes high-resolution simulations down to 4m granularity across multi-square-kilometer city regions, generating hundreds of scenarios in days rather than months. Transitioning from traditional CPU-only based simulations to GPU-accelerated workloads enables faster-than-real-time performance, allowing 24 hours of wind activity to be modeled in significantly less time.

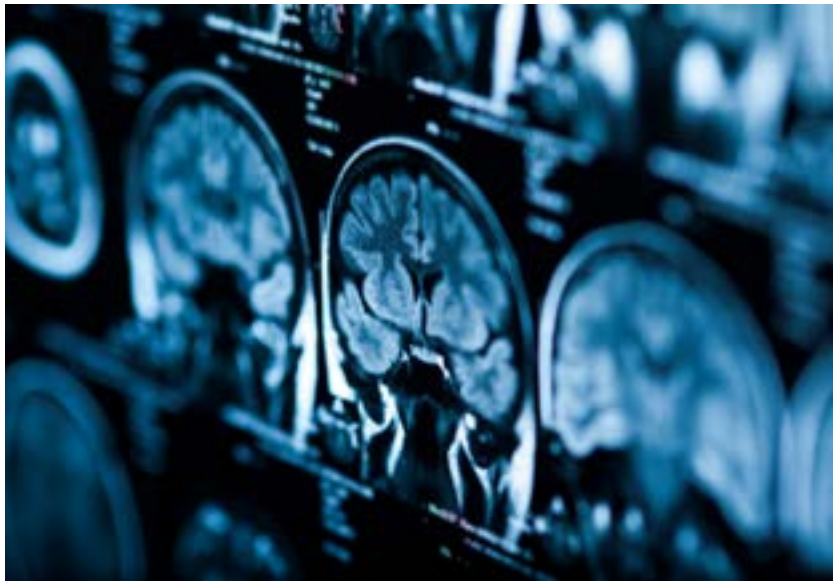
By integrating large simulation data sets with trained AI surrogate models, the system dramatically reduces computation time while maintaining fidelity, supporting rapid iteration across diverse weather conditions. This demonstrates how AI factory architectures can scale scientific computing workloads, optimize performance, and enable real-time predictive modeling. The result is a validated framework for assessing safe takeoff, landing, and flight paths for urban air traffic, showcasing the ability of high-performance AI infrastructure to solve complex, safety-critical challenges in emerging mobility systems.

Outcomes

- Generates **300+ high-resolution** simulations in under a week
- Cuts simulation time by **2.4x**
- Lays the groundwork for **real-time** urban wind forecasting

Learn more: Modeling the wind to bring air taxis to life

> "By leveraging the full capabilities of supercomputers through tailored code modernization, we enable challenging projects such as high-resolution urban wind simulations to advance from concept to real-time feasibility."
>
> —Dr. Ji Hoon Kang, senior research scientist, KISTI



University College London uses GenAI for individualized neurology treatments

Supercomputing transforms brain diagnostics

Researchers at University College London are scaling generative AI (GenAI) beyond traditional statistical models to construct high fidelity, volumetric representations of the human brain. This effort leverages an AI factory to process and train on extremely large, multimodal data sets.

The system supports ingestion and model training across more than 600,000 3D brain images combined with text and structured clinical data, requiring high-throughput compute accelerated with GPUs, and efficient data pipelines. Moving from 2D to 3D generative modeling significantly increases computational complexity, demanding tightly integrated HPC and AI infrastructure capable of handling sensitive, globally sourced data at scale.

By enabling large-scale, fully inclusive data processing and advanced generative model training, the platform demonstrates how HPE AI Factory with NVIDIA can support next-generation scientific workloads. The result is a flexible AI framework capable of synthesizing individualized brain models, illustrating the potential of high-performance AI systems to accelerate precision medicine and complex biological modeling.

Outcomes

- Creates a highly expressive brain model to generate individualized insights
- Improves the fidelity and equity of diagnosis and treatment strategies
- Builds a framework for research that extends to any complex biological context

Learn more: Transforming brain diagnostics with generative AI

"The finesse and sophistication of a human expert are combined with the rigor and objectivity of a machine that neither tires nor retires and can continually improve over time."

—Parashkev Nachev, professor of neurology,
Queen Square Institute of Neurology at University College London

Bringing AI factories to life, at scale, with trust

Across industries and geographies, a clear pattern is emerging: Organizations are moving to deploy AI as core infrastructure. Whether supporting national research priorities, enabling regulated industries, or helping enterprises operationalize AI at scale, these examples show what it takes to move from promise to impact.

What they share is not a single use case, but a common foundation: integrated systems that combine accelerated compute, high performance networking, software, and services into a repeatable model for building and running AI responsibly.

From sovereign deployments and enterprise platforms to data-intensive commercial workloads, this approach gives organizations the performance they need, the control they require, and the flexibility to scale as AI ambitions grow.

Together, HPE and NVIDIA are enabling AI that is powerful, trusted and ready for real world impact.

Learn more at

HPE AI Factory with NVIDIA

Visit HPE.com

Chat now

a00161541enw

HEWLETT PACKARD ENTERPRISE

hpe.com