Special Analysis

AI Factories: An Emerging Option for Mid-Sized Organizations

Jaclyn Ludema and Mark Nossokoff
January 2026

HYPERION RESEARCH OPINION

Organizations that have moved beyond initial AI experimentation now face an infrastructure decision that was largely theoretical just two years ago: whether to continue scaling AI workloads on public cloud platforms to explore emerging alternatives that may offer greater operational control and cost predictability. This decision is no longer confined to hyperscalers and large enterprises. Recent announcements from Hewlett Packard Enterprise (HPE), NVIDIA, and Amazon Web Services (AWS) indicate that managed and shared AI-factory environments are increasingly being positioned as accessible options for small to mid-sized organizations with meaningful AI workloads but that lack the resources or intent to build dedicated infrastructure.

This special analysis examines where managed AI-factory environments may fit within the strategies of smaller and mid-sized organizations. For organizations whose AI workloads are becoming more regular, performance-sensitive, or governance-constrained, managed AI-factory environments may offer advantages in predictability, consistency, and operational alignment. Cloud AI, meanwhile, continues to offer clear strengths for exploratory work, rapidly evolving workloads, and organizations that prioritize flexibility. The relevance of AI factories for any given organization is likely to depend on workload characteristics, regulatory context, and long-term AI objectives.

From a market perspective, these announcements suggest that AI factories are increasingly being offered as one option among several within a broader AI infrastructure landscape. For smaller and mid-sized organizations, managed or shared AI-factory environments may be considered when AI workloads become more regular, performance-sensitive, or governance-constrained. In such cases, the appeal lies less in peak performance and more in predictability, consistency, and operational alignment.

At the same time, cloud AI continues to offer clear advantages for exploratory work, rapidly changing workloads, and organizations that prioritize flexibility over determinism. In many cases, organizations are likely to employ both models concurrently, using cloud services for development and experimentation while selectively adopting managed AI-factory resources for specific production workloads.

Rather than representing a binary choice, AI factories and cloud AI increasingly appear to function as complementary components within hybrid AI strategies.

SITUATION ANALYSIS

Cloud AI for Early Adoption

For many smaller organizations, cloud AI platforms provide an efficient way to initiate AI development. They support exploratory workloads, proof-of-concept efforts, and early deployment with relatively low upfront commitment. The elasticity of cloud resources can be well-suited to intermittent or evolving workloads, and managed services can reduce operational complexity for teams with limited infrastructure experience.

As AI usage increases in scale or regularity, however, some organizations report that certain tradeoffs become more onerous. These may include variability in training performance due to shared resources, increasing difficulty forecasting costs for long-running or data-intensive workloads, and challenges related to regulatory compliance or data residency. These factors do not necessarily invalidate cloud AI, but they may influence infrastructure decisions as AI becomes more tightly integrated into operational workflows.

Distinguishing AI Supercomputers from AI Factories

AI supercomputers are typically optimized to execute extremely large AI workloads, often requiring aggressive aggregate compute performance, high-speed interconnects, and the ability to support model training runs on large data sets. These systems are commonly associated with national laboratories, hyperscalers, or large enterprises conducting high-intensity training workloads. In many cases, their value is derived from peak capability rather than continuous utilization.

Recent announcements of large-scale AI infrastructure initiatives, including the Stargate project, EuroHPC's AI Gigafactories, and the Argonne National Laboratory/Oracle/NVIDIA component of the Genesis Mission reflect continued investment in this category of AI supercomputing. While these initiatives are sometimes described using "AI factory" terminology, they are designed primarily for national research priorities, hyperscaler-class deployments, and large enterprise or public-sector consortia rather than for mid-sized organizations seeking accessible, operationally oriented AI infrastructure.

Managed AI factories as discussed in this special analysis occupy a different position. They tend to emphasize repeatability, integration, and operational continuity. Rather than focusing solely on training a single large model as efficiently as possible, AI-factory environments are structured to support the ongoing production, deployment, monitoring, and retraining of AI models over time. This often includes integrated data pipelines, MLOps tooling for automating and managing the entire machine learning lifecycle, scheduling mechanisms, and governance controls that align AI developments with regular business operations.

It is within this context that interest in managed AI-factory models has begun to emerge.

EXAMPLES OF MANAGED AI FACTORIES

HPE & NVIDIA: A Managed AI-Factory Model for UK Enterprises

In December 2025, HPE and NVIDIA announced the launch of a Private AI Lab in Manchester, UK, developed with Carbon3.ai. The lab is positioned as a shared, managed AI environment intended to

support UK enterprises that are progressing beyond early AI experimentation but that may not be prepared to deploy or operate dedicated AI infrastructure.

The facility is built on HPE Private Cloud AI and NVIDIA AI Enterprise software, offering access to accelerated compute, networking, and integrated AI tooling in a locally hosted environment. The emphasis on data residency and regulatory alignment suggests that the model may be particularly relevant for organizations operating under governance or compliance constraints. Rather than positioning the lab as a replacement for cloud AI, HPE and NVIDIA describe it as an environment where organizations can evaluate and operationalize AI workloads under more controlled conditions.

This approach illustrates how AI-factory concepts are being adapted to shared and managed formats, potentially lowering the barrier to entry for organizations that require production-oriented AI capabilities without full infrastructure ownership.

AWS: Exploring On-Prem, Managed AI Factories

AWS has also begun to explore AI-factory-style deployments through the introduction of on-premises, AWS-managed AI environments. Announced at AWS re:Invent 2025, these offerings involve deploying AWS-operated AI infrastructure directly within customer facilities, combining local data control with managed services such as Amazon Bedrock and SageMaker.

AWS has framed these deployments as suitable for customers with data sovereignty, latency, or compliance requirements that limit reliance on public cloud infrastructure. Early indications suggest that these offerings are being piloted with enterprise and public-sector organizations rather than broadly positioned for all customers. For smaller organizations, this model may present an option to access cloud-like AI services while maintaining greater control over data location and infrastructure boundaries.

As with the HPE and NVIDIA example, AWS's approach reflects an effort to provide additional infrastructure flexibility rather than a wholesale shift away from cloud AI.

EuroHPC: AI Factories for SMEs and Startups

The European Union has also moved to expand AI-factory access beyond large research institutions and enterprises. Through EuroHPC, the EU has established 19 AI Factories and 13 AI Factory Antennas across member states, with an explicit focus on supporting small and medium-sized enterprises (SMEs) and startups. According to EuroHPC, these facilities offer free, customized support intended to help smaller organizations scale business innovation through access to high-performance AI infrastructure.

This initiative represents a public sector approach to the same challenge that HPE, NVIDIA, and AWS are addressing through commercial offerings: lowering the barrier to production-oriented AI capabilities for organizations that lack the resources to build or operate dedicated infrastructure independently. The EuroHPC model differs in its funding structure and geographic scope but shares the underlying premise that AI-factory environments need not remain exclusive to large enterprises or national research programs.

For European SMEs and startups, the EuroHPC AI Factories may offer a pathway to evaluate AI workloads in factory-style environments without commercial subscription costs, though access models and capacity constraints may vary by location.

FUTURE OUTLOOK

The emergence of shared and managed AI-factory offerings suggests that vendors anticipate a growing need for infrastructure models that sit between public cloud services and fully self-managed environments. For smaller organizations, the relevance of AI factories is likely to depend on workload characteristics, regulatory context, internal capabilities, and long-term AI objectives rather than organizational size alone.

Over time, AI factories may become a more familiar element of enterprise AI architectures, particularly as vendors continue to refine service models and lower operational barriers. However, cloud AI is likely to remain central to AI adoption, especially in early stages and for use cases that benefit from elasticity and rapid deployment.

From this perspective, AI factories may be best understood not as a replacement for cloud AI, but as an additional infrastructure option that organizations may evaluate as their AI strategies mature and diversify.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology, and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). Hyperion Research provides thought leadership and practical guidance for users, vendors, and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com