Omdia
by informa techtarget

Mastering Enterprise AI Tokenomics

As generative AI evolves into continuous, agentic workflows, enterprises face a new challenge: scaling AI without allowing costs, complexity, and risk to scale with it. Omdia examined how a hybrid AI approach can help organizations create a more predictable and sustainable foundation for enterprise AI.

This Omdia Infographic was commissioned by Lenovo and is distributed under license from TechTarget, Inc.

Production-scale AI creates new economic and operational challenges

AI is no longer merely an experimental application sitting next to the business as a proof of concept; it is **an operational capability that lives inside core enterprise workflows, business decisions, and customer interactions.** As organizations scale AI into high-volume production, this transition introduces significant challenges. While the public cloud helped organizations experiment quickly, continuous agentic workflows can create unpredictable consumption costs, network latency, and data sovereignty concerns, alongside the complexity of integrating models, data, security, and governance. These pressures are forcing enterprises to reconsider where different AI workloads should run and how infrastructure decisions affect long-term economics.

As AI moves into production, organizations face growing infrastructure, integration, and resource demands.

6-12 months

can be spent integrating fragmented AI infrastructure and software components.

15 kW per rack

is the average power density of an enterprise data center, with high-density AI pushing power needs even higher.

80%

of enterprise AI effort is consumed by operational challenges, leaving only 20% for model development.

75%

of facilities still rely on air cooling as high-density AI infrastructure increases power and cooling demands.

Build AI around the workload

A hybrid AI strategy gives enterprises greater flexibility to place workloads where they make the most economic and operational sense. Routine, sensitive, data-intensive, or high-volume workloads can run on dedicated infrastructure, while organizations retain access to public cloud and frontier models when those resources better suit the task. The Lenovo Hybrid AI Factory with NVIDIA combines infrastructure, software, models, and validated architectures into a unified platform designed to reduce deployment complexity while improving control over performance, data, and costs.

Omdia's analysis identified several economic and performance advantages that emerge when AI workloads are placed on dedicated infrastructure:

6.5 hours/day

of utilization, which represents the cost tipping point between renting public cloud AI capabilities and using the Lenovo Hybrid AI Factory with NVIDIA.

Sub-50ms

time to first token on dedicated on-premises Lenovo infrastructure, so interactive chat responses start almost instantly.

30–50+ tokens/second

of sustained streaming, well above reading speed, so answers flow naturally. This was measured in Lenovo benchmark tests and consistent with what Lenovo Hybrid AI Factory with NVIDIA customers reported to Omdia.

Zero data egress fees

when AI runs on-premises and lower egress costs in hybrid setups, cutting a recurring cloud expense that's hard to predict.

Improve the economics of AI at scale

As utilization grows, dedicated AI infrastructure can change the economics of enterprise inference by replacing variable consumption costs with a more predictable investment. Omdia's analysis found benefits beyond infrastructure cost, including faster time to value, greater price-performance, reduced power overhead, and improved budget predictability. The result is not an argument for eliminating public cloud but rather for matching each workload with the environment that delivers the appropriate combination of economics, performance, scalability, and control.

Up to 84%

cost savings across evaluated AI workloads compared with public cloud alternatives.

10–15%

improvement in overall AI infrastructure TCO from Lenovo Neptune direct-to-chip liquid cooling.

15 business days

or less order-to-ship fulfillment is available for qualifying configurations through the Lenovo AI Express program, providing an accelerated path to AI readiness.

~8 month

payback on a Lenovo Hybrid AI Factory with NVIDIA ThinkSystem SR650 node, generating $260,000 in savings over a conservative three-year lifecycle and $480,000 over five years per server.

Conclusion

Enterprise AI infrastructure decisions become increasingly important as AI transitions from an experimental application sitting next to the business into an **operational capability that lives inside core workflows.** As this shift to sustained production occurs, workload utilization, tokenomics, data requirements, latency, power consumption, and governance all influence where AI is deployed most effectively. Omdia's analysis indicates that a hybrid strategy can help organizations place workloads according to their economic and operational requirements while maintaining access to cloud resources when appropriate. By evaluating AI infrastructure through the lens of long-term workload economics rather than initial deployment convenience alone, enterprises can build a more predictable foundation for scaling AI.

Learn More

Lenovo | NVIDIA

Source: Omdia Economic Validation, *Analyzing the Economic Benefits of Lenovo Hybrid AI Factory with NVIDIA*, October 2026.
© 2026 TechTarget, Inc. All Rights Reserved. Unauthorized Reproduction Prohibited.