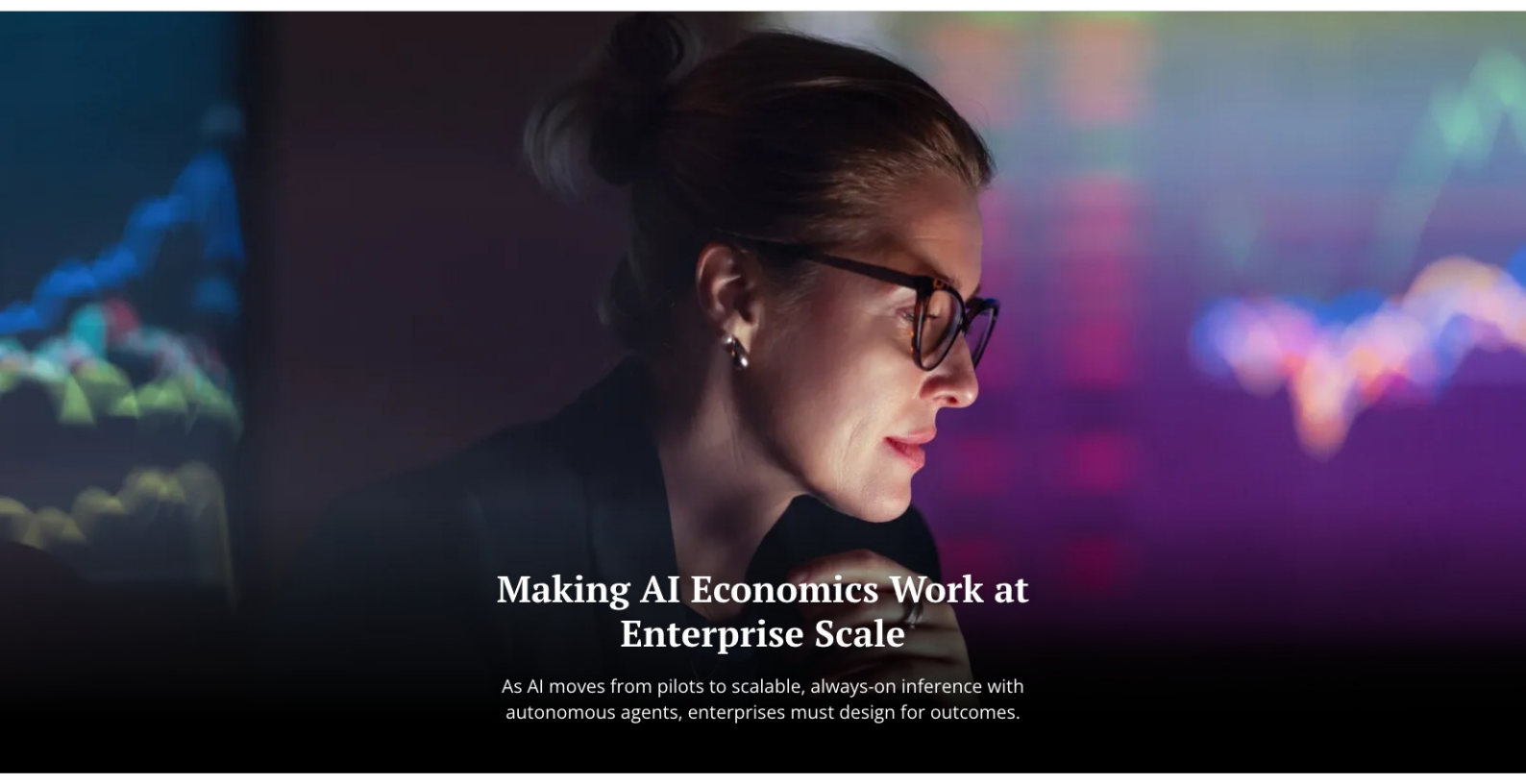


Making AI Economics Work at Enterprise Scale

As AI moves from pilots to scalable, always-on inference with autonomous agents, enterprises must design for outcomes.

Quick Summary Read More

- As enterprises move from AI pilots to always-on inference with agentic workloads, the economics change, requiring leaders to manage AI as an ongoing operating capability tied to business value.

The push to turn pilots into production has investments of time and money pouring into enterprise AI. But as HyperFRAME's 2026 *State of the Enterprise AI Stack* report reveals, that energy isn't translating into returns. Only 23% of respondents say their AI and machine learning projects launched in the last year successfully reached production and met ROI objectives.

Closing this "execution gap" represents a critical challenge for companies across industries, and the starting point shouldn't be what's under the hood, says Steven Dickens, HyperFRAME's CEO and principal analyst. "It's a business and transformation project rather than a technology project, looking at AI holistically and asking what problems in the business we're trying to fix."

This focus on AI outcomes is critical for successful deployments. "It allows organizations to more effectively align technology choices to strategic goals," says Linda Yao, vice president and general manager of hybrid cloud and AI solutions at Lenovo. "We're past the era of adoption and productivity, and we want to think about real outcomes that impact the profit and loss statement (P&L). Companies seeing the most success fund, staff and govern and treat AI like a new capability, not just a portfolio of experiments."

It also helps align two additional qualities that significantly increase the likelihood of AI delivering expected returns. The first, Yao says, is control over your data, IP and core organizational processes. Companies decide the right fit for their needs. The second is cost. As AI moves from periodic model training to continuous inference, autonomous agents and always-on services, the economics change. "Usage compounds across every interaction, agent step and reasoning process, consuming power, compute and money the entire time," says Chris Marriott, vice president of Enterprise Platforms and Solutions at NVIDIA. This makes cost per inference—and ultimately, cost per business outcome—an increasingly important management metric.

Autonomous AI agents and always-on services have rapidly changed the economics of AI deployments, increasing token consumption—the measurement unit of AI usage—and making "tokenomics" an essential framework for enterprise AI, Marriott says. He believes companies failing to plan for this new reality are in for a surprise they'd prefer to avoid.

More Tokens, More Value?

Customer service is one of the most common uses for agentic AI, and done well, it reflects the strategic advantage in choosing a service outcome approach to AI. Organizations deploy agents in contact centers with different goals in mind, including faster resolution, lower service cost, better customer experience and the ability to redirect employees toward higher-value work. When Lenovo built and deployed AI agents across its customer support operation, the company saw a 25% improvement in resolution time and a 10% reduction in support costs. It's an important demonstration of how costs and value can remain aligned in agentic deployments, even with a considerable rise in token consumption as agents reason with user and application data—everything from purchase and service history to device preferences to location and more.

When guardrails route complex or sensitive cases to a human in the loop, they intervene with much more information at their disposal.

"We're past the era of adoption and productivity, and we want to think about real outcomes that impact the P&L." — Linda Yao, VP and GM, Hybrid Cloud and AI Solutions, Lenovo

Yao notes that companies can tailor the agentic contact center to better meet the needs of different customers across product lines and markets. Healthcare companies, for example, can deploy agents designed to process insurance claims. Financial services companies can design agents to facilitate loan applications. Both might require extra protections around data. Lenovo, she says, developed a Cantonese LLM to help serve its Hong Kong market more effectively. "When we look at outcomes, we were able to get measurably significant customer experience improvements, as well as significant operational improvements," she says.

The inference demands required to bring that much context and reasoning to automated services are significantly higher than those of simple chatbots, dramatically shifting the underlying economics. Companies with greater AI maturity prepare for those changes before deployment. "Budgeting shifts from capital expenditure to operating expenditure. Margin analysis becomes continuous. AI-powered products have direct, variable costs associated with their usage, the same way software-as-a-service (SaaS) products have cloud-hosting costs," Marriott says. "They're treating inference as a production line, building capacity plans, measuring and tuning in relation to strategic goals. The question leaders start to ask is: How fast and efficiently are we producing intelligence?"

Building Choice Into AI

Companies cannot effectively balance cost and performance without a well-designed, well-functioning AI ecosystem. Leaders should design and build their AI capabilities holistically, driven by the strategic outcomes they desire. "What are our requirements around things like performance, availability, scalability, security and data governance?" Dickens says. "What regulatory frameworks are relevant to the business? Should we be on the cloud or on-prem?" The answers to those questions will vary across outcomes, industries and even inside different parts of the organization.

"A bank might need one AI deployment for their coding team. They'll consider deployments for customer and in-branch support. There are the core banking applications and transactional systems that run the ledgers," he says. "They'll be making different choices about AI deployments across all of those environments."

"AI factories are designed to produce intelligence and not simply host applications." — Chris Marriott, VP, Enterprise Platforms and Solutions, NVIDIA

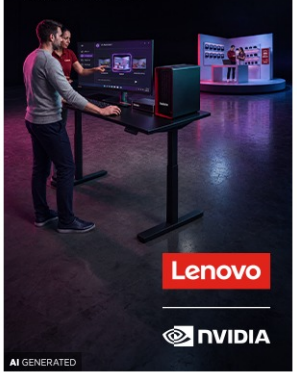
Given the technological complexities and the range of desired outcomes across the organization, enterprises are increasingly looking to the AI factory model to offer prebuilt, customizable architectures, move AI workloads into production and then manage them over time. **Lenovo Hybrid AI Advantage with NVIDIA** applies NVIDIA reference architectures to the factory model across hybrid environments. The solution brings together infrastructure, data, open and proprietary models, AI software and services so enterprises can place workloads according to their business and technical requirements, balancing power and efficiency.

The goal, Yao says, is not to standardize every AI workload on the same architecture, but to determine the right level of intelligence and find the appropriate balance of performance, security, latency, cost and control based on the business value each use case needs to deliver. Lenovo Hybrid AI Advantage with NVIDIA, she notes, can deliver up to eight times lower cost per token than cloud-based infrastructure-as-a-service environments, and up to 18 times lower cost per million tokens compared with model-as-a-service APIs.

This helps organizations deploy and scale AI across the hybrid AI continuum with greater efficiency and discipline: define the business outcome, determine the level of intelligence and performance the workflow requires, place the workload where its economics and governance requirements make sense and continuously measure whether it is producing sufficient value. When requirements or conditions change or benchmarks aren't met, changes can be made with minimal disruption.

"AI factories are designed to produce intelligence and not simply host applications," Marriott says. "The design principle behind the NVIDIA AI enterprise we provide as part of Lenovo Hybrid AI Advantage with NVIDIA is producing the highest value intelligence at the lowest cost and energy per token to help achieve strategic outcomes."

Learn how Lenovo Hybrid AI Advantage with NVIDIA can help move enterprise AI from strategy to production with the performance, control and economics each workload requires.

FAST TRACK AGENTIC AI ADOPTION WITH LENOVO AI LIBRARY.
Lenovo Hybrid AI Advantage™ with NVIDIA.
LEARN MORE >>


This content was created by Custom Content from WSJ, a unit of The Wall Street Journal Advertising Department.

Copyright ©2026 Dow Jones & Company, Inc. All Rights Reserved.