

A woman with curly hair, wearing a dark blue top and large hoop earrings, stands in a server room. She is holding a laptop and looking upwards. The background is filled with server racks and a grid of glowing blue and green dots.

Deloitte.

Together makes progress

AI factory industry trends:
A 2028 outlook

Table of contents

Industry trends in AI adoption, token costs to value and system scalability	03
The AI scale shift: Why enterprise infrastructure is reaching its limits	04
Factors affecting the cost curve for AI	05
The AI factory: Motivations and driving forces for adoption	07
Challenges in AI factory adoption	09
What's the payoff of building an AI factory?	11
The path forward: Scale with discipline	12
Contacts	13
Endnotes	14



Industry trends in AI adoption, token costs to value and system scalability

As organizations across industries move from AI pilots to enterprise-scale workloads, the question many must answer is, “How can we marshal AI in a way that accelerates our most important business strategies with the supporting intelligence and infrastructure, discipline, and governance to scale effectively?”

One emerging answer is to build an AI factory. Why? Because AI factories help organizations turn pilots into governed, enterprise AI capabilities with measurable returns, and they help organizations transform AI into a source of industry advantage by aligning infrastructure, data, and governance to manage token costs, data latency needs, and sovereignty requirements given the realities of their business and the markets they serve. In addition to internal needs, AI factories may also enable organizations to sell GPU as a service, tokens, models, model routing, or their own solutions externally.

Industries covered:

- Financial Services
- Technology, Media & Telecommunications
- Energy, Resources & Industrials
- Life Sciences & Health Care
- Consumer

Explore Deloitte's enterprise AI infrastructure survey: A 2028 outlook¹



The AI scale shift: Why enterprise infrastructure is reaching its limits

The demand to scale AI is straining the IT infrastructure across many organizations and forcing them to plan for a sharp increase in both AI pilots and production deployments.

To better understand industry trends in AI adoption, as well as other factors shaping decision-making about AI, Deloitte conducted the enterprise AI infrastructure survey² in November 2025. Respondents included 515 US enterprise leaders from across the consumer; energy, resources & industrials (ERI); financial services (FSI); life sciences and health care (LSHC); and technology, media, and telecommunications (TMT) industries.

The findings underscore a quick rise in AI workload demand. The percentage of organizations with 31 or more AI pilots is expected to rise from almost 50% in 2025 to nearly 70% in 2028. Further, the share with 31 or more production-ready use cases is expected to rise from 44% to 67% in that same period. As organizations across industries evaluate which pilots should move into production, getting a handle on the cost per task across enterprise functions and applications will be an essential piece to the investment-and-return equation for AI.

This starts with choosing the right business case and redesigning processes across the value chain. While the IT function and software development tend to be early adopters of multiagent solutions, AI is also enabling front-office software reimagination that's ranging from physical product R&D to customer-facing solutions. Therefore, to scale AI effectively, organizations need to *consider the full stack and optimize workloads* across the network so they can deploy the right model in the right place at the right time. That need for greater workload control is prompting renewed interest in hybrid infrastructure strategies that give businesses more oversight.

As AI demand grows, it can increase pressure on infrastructure capacity, spending controls and governance policies.

Moreover, as AI infrastructure expands along with that demand, *power and sustainability* will become additional, crucial constraints as inferencing workloads push power consumption from kilowatts to gigawatts for those running AI at scale. Power availability, cooling costs or upgrades to liquid cooling, and carbon emissions intensity all influence capacity, future build vs. buy decisions, and workload routing.

The upshot? AI infrastructure planning and deliberately thinking through your silicon strategy is no longer a nice to have; it's a business and strategy imperative that must align AI growth plans with capacity, cost and governance across the C-suite.

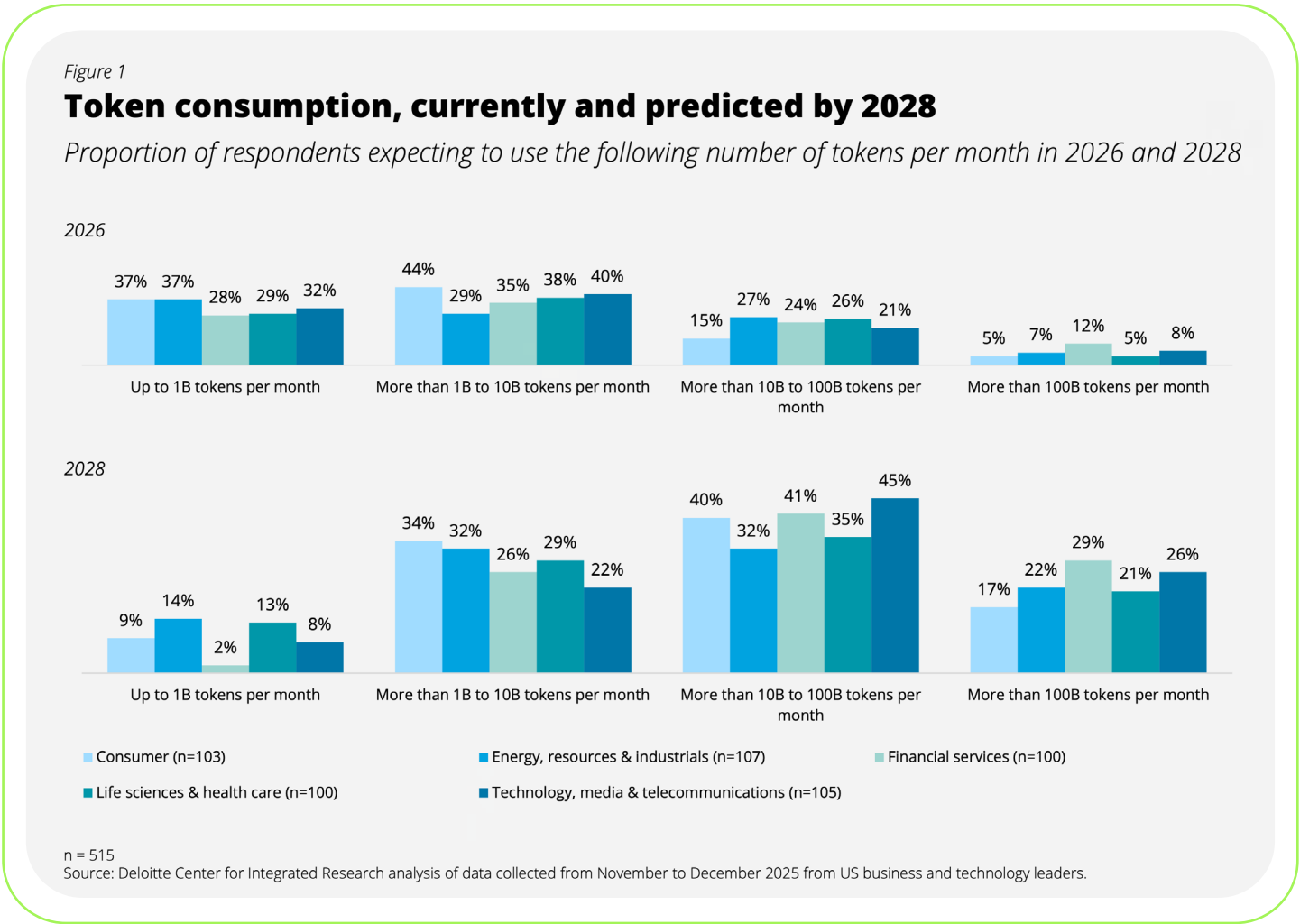
Unless noted otherwise, all statistics and analysis offered here are based on Deloitte's enterprise AI infrastructure: A 2028 outlook survey.

Factors affecting the cost curve for AI

Our survey revealed that two issues are reshaping how companies think about AI implementation costs: AI token consumption and perceived risk.

AI token consumption by industry

AI token consumption is one clear marker that AI scale is reshaping both the technical and business economics of AI. Tokens are a unit of compute and a unit of cost, so accelerating token consumption can serve as an early indicator of stress on organizational infrastructure and economics. While organizations may have some visibility into their token volumes on the cloud, as AI applications become increasingly distributed across business functions and teams, change is abundant. New models with new pricing structures are introduced monthly, vendor pricing models continue to shift, and as user numbers and engagement scale, the costs compound variably. This makes enterprise transparency, cost and value management, and forecasting challenging—especially without industry benchmarks.



As of December 2025, 37% of survey respondents consumed 1 billion to 10 billion tokens per month on average, and 30% consumed more than 10 billion. By 2028, 61% reported they expect to consume more than 10 billion tokens per month on average. A recent Goldman Sachs³ study projected that token volumes across the enterprise market will reach quadrillions by 2030, with some large technology operators already running over 3 quadrillion tokens/month⁴ as of early 2026.

By industry, TMT organizations expect the most growth in AI token consumption, where the respondent share above 10 billion monthly tokens is expected to rise from 29% today to 71% by 2028. FSI organizations are close behind, expected to rise from 36% to 70%. Conversely, only 54% of ERI organizations reported they expect to consume above 10 billion monthly tokens in 2028.

Higher token volumes are a natural consequence of AI scaling, but given that the cost is variable, it's essential for leaders to ensure that they're spending more only when it's strategically differentiating to use more expensive reasoning models because their industry lens and business strategy warrants it and not because of poor system design choices.

Establish the token transparency needed to measure it

As AI initiatives scale, tokens are becoming a critical unit of analysis to understand the cost/benefit of AI-enabled capabilities for customers and work. They, therefore, need to be viewed as a system of value exchange across the enterprise and treated with the same discipline as any other capital investment. Some organizations are starting to build enterprise dashboards to aggregate token consumption across applications and infrastructure, and set token budgets by business unit, use case, and user. These tiered access levels are essential to bring transparency to the enterprise cost profile and top users, and as a diagnostic when scaling.

Optimize it through tech optimization

Once these first two elements are in place, more advanced organizations are exploring showback and chargeback models to make consumption more visible to business owners and to address capital and operating expensing at a product team level as well as additional guardrails to switch off agents when token volumes surge unexpectedly. These steps are reflective of a broader conversation about adopting FinOps for AI because of the essential role it can play in connecting model and infrastructure use to forecasting and business value generation. The objective is to direct token consumption toward value, that is, where AI can deliver the greatest strategic impact.

Perceived risks of AI investment

As hybrid AI environments mature, the dialogue around risk is broadening to include economics and governance, as well as architectural considerations. Increasing spending won't help most organizations scale effectively. The real issue is whether they can turn that investment into durable, disciplined, enterprisewide capability.

Transform cost while maintaining performance with open models and model routing

Organizations can manage down the cost curve with financial and engineering moves. That means token management isn't simply an engineering issue, but a business discipline that is informed by strategy, guided by a philosophy on acceptable tokens per task across different types of projects and users, and governed by FinOps and engineering approaches that enable cost control. Transparency and controls can then be paired with engineering levers like technology optimization (that is, processing, caching, context management), a multitool approach across a portfolio of models at different price points, integrating open models, building in model routing, and rightsizing parameterization with small models.

Factor additional complexities such as disaster recovery and resiliency

Importantly, as business and tech leaders work together to manage token costs, it will be crucial to consider how to ensure that customer-facing and mission-critical applications are always on for business users and customers and with necessary failover and tiering. This can come at a cost as well for those using frontier models, where premium access comes with additional fees. For this reason, more organizations are beginning to explore open models to diversify their cost structures while maintaining performance.

The AI factory: Motivations and driving forces for adoption

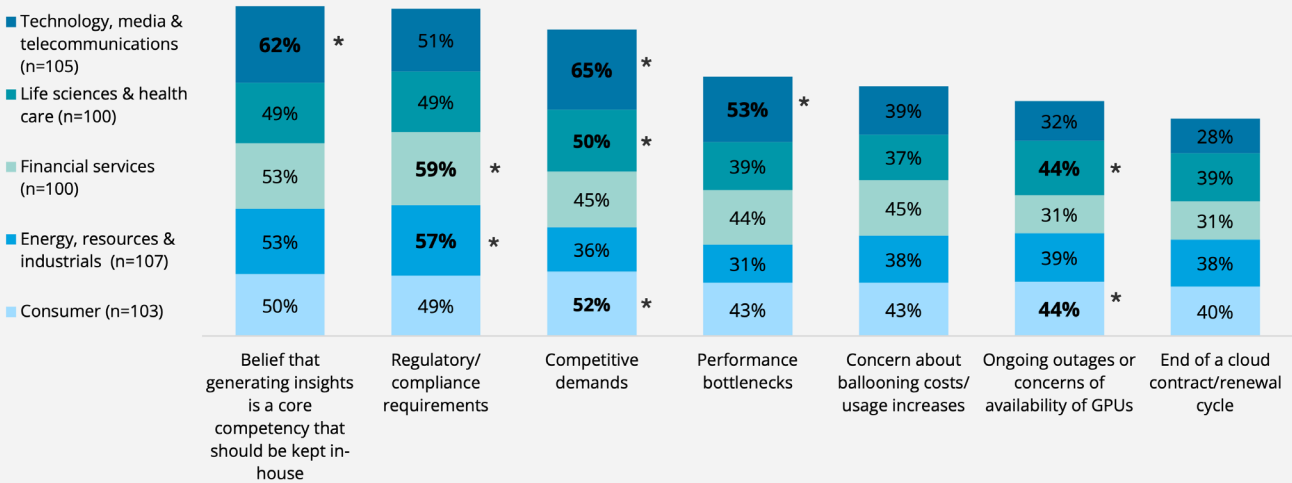
An AI factory can help organizations by reducing cost, accelerating token throughput, and providing more control and transparency as AI scales.

Our survey shows that while the reasons organizations invest in AI factories vary by industry, the throughline is tighter control over AI deployment and its effects. Compared to other industries, consumer and life sciences & health care were most motivated by outages and GPU computing availability concerns, at 44% each, while energy and financial services were most concerned with regulatory and compliance requirements, at 57% and 59%, respectively. Technology, media, and telecommunications organizations were most motivated to build AI factories due to competitive pressure at 65%, given its an in-house core competency at 62%, and given performance bottlenecks at 53%.

Figure 2

Motivation to build an AI factory by industry

Proportion of respondents expecting to increase their organization’s investments in a future AI factory due to the following factors



* Ranked No.1 factor for each industry as compared to other industries.

n = 515
Source: Deloitte Center for Integrated Research analysis of data collected from November to December 2025 from US business and technology leaders.

Deloitte’s survey results show that as token consumption rises, many organizations are moving quickly toward AI factory adoption. But they’re not progressing at the same rate—and that’s a strength rather than a weakness. Organizations are scaling AI in ways that reflect their industry’s unique operating realities, competitive pressures, and regulatory obligations. Overall, 36% of organizations report at-scale AI factory deployments today, with an expected rise to 73% by 2028.

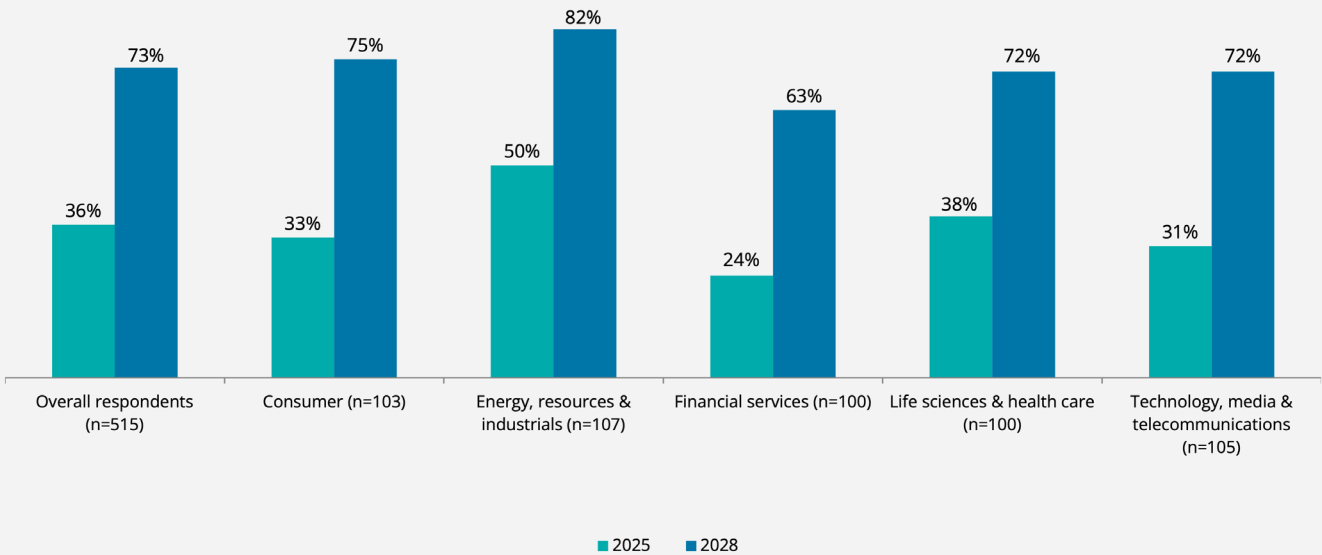
In every industry, we see double-digit growth in AI factory adoption at scale. There were virtually no respondents saying they are not pursuing this strategy today, and two-thirds of respondents or more in every industry expect they will have full-scale deployment by 2028.

Increasingly, AI systems will be seen as mission-critical infrastructure that powers the business and unlocks the insights needed to run the business. Respondents across industries recognize that it’s imperative and are recalculating what they own, operate and manage across the stack.

Figure 3

AI factory adoption stage by industry: Today and planned

Proportion of respondents whose organizations are fully scaling their AI factories in 2025 and 2028



n = 515
Source: Deloitte Center for Integrated Research analysis of data collected from November to December 2025 from US business and technology leaders.

Challenges in AI factory adoption

AI factory adoption is not without challenges. Talent and resources including access to power and GPUs remain top areas of concern when scaling AI across industries.

Talent

Given that many organizations have spent the past decade (or longer) moving to become cloud native, many organizations have lost their organizational muscle to own, operate and manage infrastructure. Operating large-scale GPU farms, complex GPU networking, and AI storage with new software stacks requires deep engineering that cuts across domains. It goes beyond hardware. Organizations need talent with intersectional skills to manage the tight coupling with software including how to onboard, serve, benchmark and govern open models that operate on these AI factories. Data engineers (57%), security/compliance specialists (55%), data scientists (51%), and machine learning (ML) Ops/AI Ops engineers (51%) are all important talent. The next challenge becomes where to get them.

While 40% of respondents, on average, reported talent and skills gaps as their biggest issue in managing an AI factory, LSHC respondents indicated other obstacles as well, noting they're more limited by workload variability (38%). In addition to talent challenges, TMT respondents reported lock-in and fast technology cycles (43%) as a primary issue when scaling AI factories given relatively higher maturity in AI capabilities.

Consumer respondents also have above-average needs for talent in data engineering (62%) and security/compliance (60%) skills today.

The takeaway: Industry-knowledge-plus-AI-capability as a hybrid role will be needed to bring together the vast scale of business, financial and technical knowledge needed to achieve this next stage of AI scaling. This goes beyond simply hiring individuals and will require building up systemic capabilities across a hybrid tech ecosystem.

Security and compliance, data engineering, data science, and MLOps are foundational to success. Talent that has both industry and AI knowledge provides organizations with two significant advantages when building an AI factory. First, they can identify the use cases that will have the most impact and be most likely to realize value. Second, they understand the operating and regulatory environment of the business, and that understanding can help smooth the translation of technical investment into business outcomes. It's essential to think of talent not as a commodity but as a carefully considered asset that adds value when designing the AI factory operating model.

Power and facilities

AI also brings significant downstream power and cooling considerations for enterprises building AI factories and technology and the energy operators enabling them. The Deloitte survey shows power demand is shifting toward larger-scale contracts, with strong growth in 500 MW–1 GW (gigawatt) and 1 GW+ segments across industries. Energy and TMT are already concentrated at 100–500 MW today, and shift toward 500 MW–1 GW (27% and 25% respectively) by 2028. Financial services shows the most aggressive move to 1 GW+.

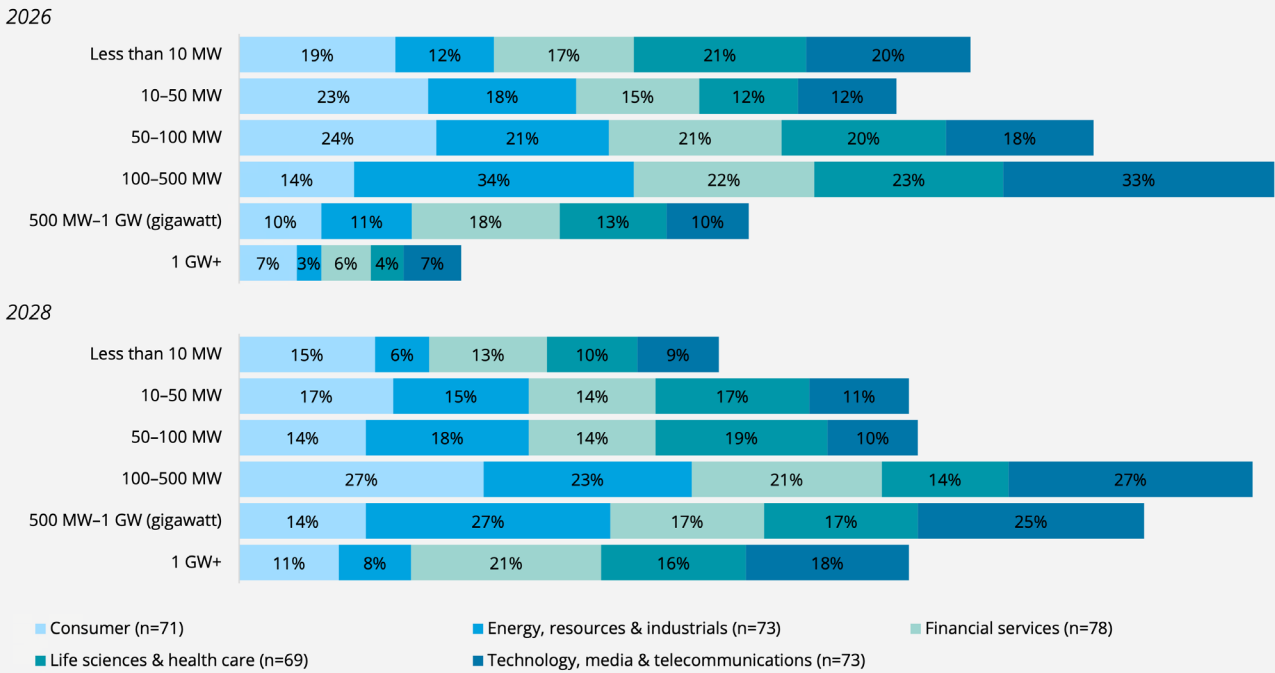
- **Power and cooling strategies** are diverging by industry, reflecting very different infrastructure operating models. Organizations overall are pursuing a mix of sourcing strategies, but Energy, Resources & Industrials stands out with a stronger preference for building its own power generation, while Financial Services, Life Sciences, and Technology, Media & Telecom are more inclined toward preconfigured compute facilities. Consumer appears to follow a more mixed model.
- **Planned power demand contracts** are expected to shift upward, with growth into 500 MW–1 GW and 1 GW+ segments; Financial Services shows a particularly aggressive move toward 1 GW+.

Figure 4

Contracted power capacity by industry, currently and predicted by 2028

Proportion of respondents contracting power for their AI factories in 2026 and 2028

(considers only those respondents who are “leasing/renting new colocation space” or “adopting pre-configured, portable, and modular compute solutions”)



n = 515

Source: Deloitte Center for Integrated Research analysis of data collected from November to December 2025 from US business and technology leaders.

AI factory adoption success depends on addressing both technical and resource challenges, and on bringing in industry knowledge to translate technical investments into business outcomes. This will take engaging with partners across an increasingly hybrid set of AI infrastructure choices to rearchitect the enterprise as the technology and energy ecosystem is evolving.

What’s the payoff of building an AI factory?

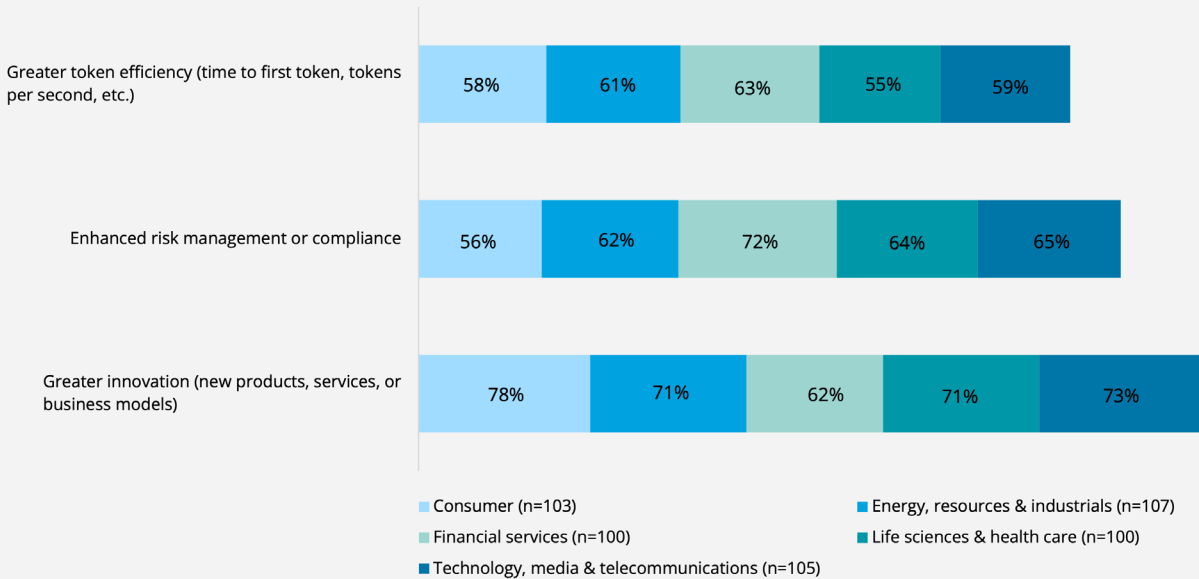
The value of an AI factory isn't just that it can help organizations deploy and operate a larger portfolio of enterprise AI use cases, but that it can help enterprises run them with more consistency, resilience and strategic intent. An AI factory can help organizations:

- Maintain data security, compliance and sovereignty;
- Launch, scale and manage more AI initiatives reliably across the enterprise;
- Support innovation with scalable, reusable AI capabilities;
- Strengthen risk management with better governance;
- Improve resilience through controlled, repeatable AI operations; and
- Improve cost discipline with AI resource optimization.

Figure 5

Expected outcomes of AI factory investments

Proportion of respondents measuring the business impact of AI factory investments using the following metrics



n = 515
Source: Deloitte Center for Integrated Research analysis of data collected from November to December 2025 from US business and technology leaders.

According to our survey, the three clearest expected outcomes for all five industries are increased capacity for innovation, answered by 71% respondents, improved risk management by 64%, and token optimization by 59%, which positions AI factories as both growth and control assets. Innovation is the top measure of business impact in every industry except FSI, where enhanced risk management and compliance are the preferred measures of success. Consumer also stands out as the industry most likely to use innovation as the key gauge of AI factory value, while FSI’s investment rationale is driven more by compliance. Taken together, these findings show that the business case for an AI factory isn’t about simply adding infrastructure. It’s about creating a more governed, repeatable and economically viable environment for enterprise AI.



The path forward: Scale with discipline

AI is maturing across industries, but businesses are finding unique paths based on their unique industry ambitions and constraints. Those furthest ahead have moved beyond point solutions and examined how AI can drive business transformation across core business functions and domains.

Organizations that scale AI effectively will treat consumption, infrastructure, governance and talent as a cohesive operating model and scale with discipline. They'll track token use effectively, link infrastructure choices to business value, and build the governance structure to scale responsibly.

Bottom line: Over the next two to three years, success won't be determined solely by how much AI organizations deploy, but by how effectively they align AI infrastructure, governance, and talent to the unique needs of their industry. Organizations that do so will be positioned to transform rising AI demand into lasting IndustryAdvantage™.

Contacts

Chris Thomas

Principal
Hybrid cloud infrastructure offering leader
Deloitte Consulting LLP
chrthomas@deloitte.com

Nicholas Merizzi

Principal
Silicon2Service US leader
Deloitte Consulting LLP
nmerizzi@deloitte.com

Matt Jacobs

Managing Director
Deloitte Consulting LLP
mattjacobs@deloitte.com

Steve Hebert

Managing Director
Deloitte Consulting LLP
stevehebert@deloitte.com

Diana Kearns-Manolatos

Senior Manager
Subject matter specialist
Deloitte Services LP
dkearnsmanolatos@deloitte.com

Iram Parveen

Specialist Assistant Manager
Deloitte Support Services India Private Limited
iparveen@deloitte.com

Learn more about AI factory infrastructure, powered by Silicon2Service.



Endnotes

1. Kavitha Prabhakar et al. "[Deloitte's enterprise AI infrastructure survey: A 2028 outlook](#)," *Deloitte Insights*, March 30, 2026.
2. Ibid
3. Goldman Sachs, "[AI agents forecast to boost tech cash flow as usage soars](#)," May 20, 2026.
4. Sundar Pichai, "[I/O 2026: Welcome to the agentic Gemini era](#)," Google, May 19, 2026.



This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor.

Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.

Copyright © 2026 Deloitte Development LLC. All rights reserved.