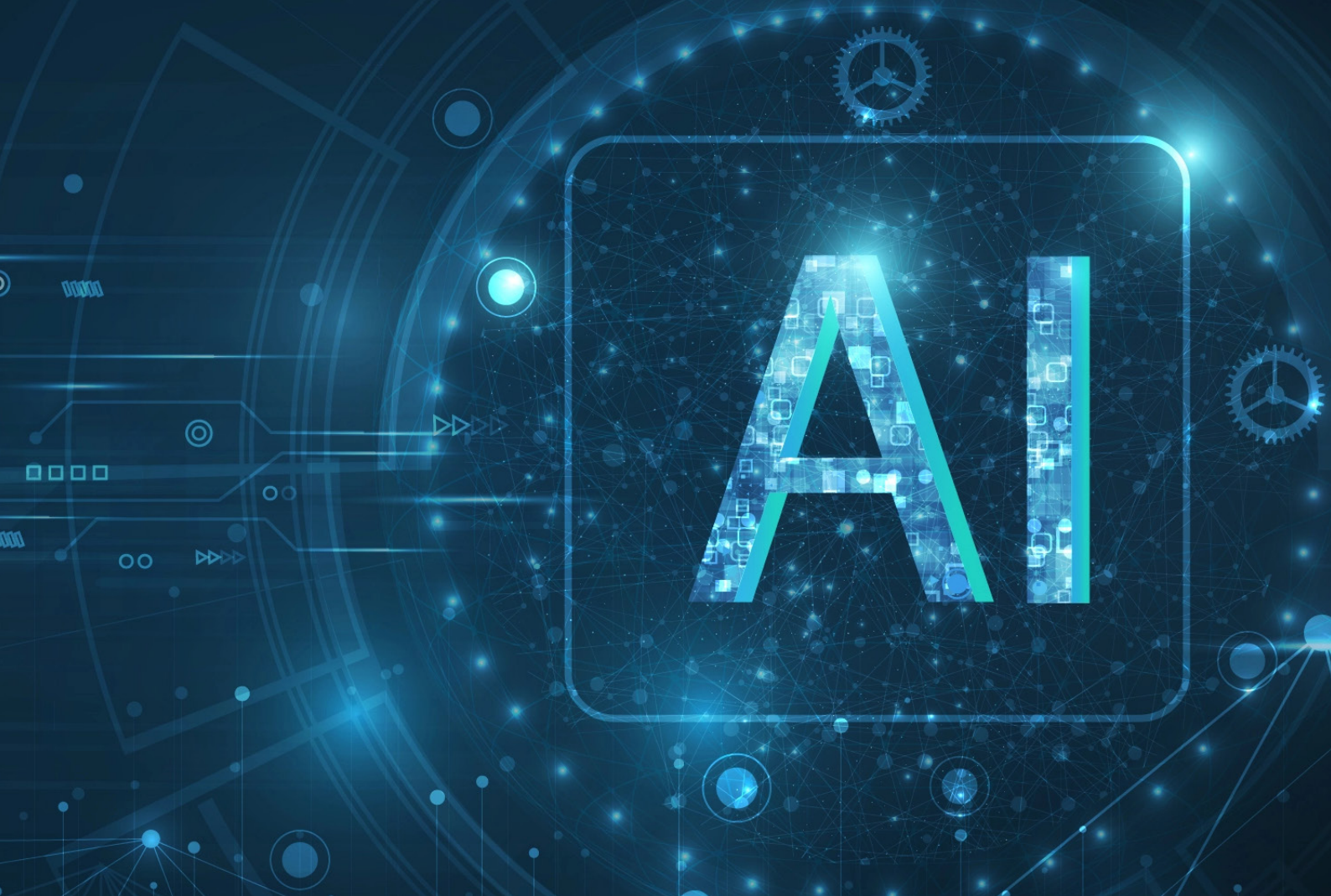
CLOUDIAN®

2026

Enterprise AI Infrastructure Survey

Data Sovereignty, Cost, and Performance: The Growing Case for On-Premises AI Infrastructure



Business Drivers of AI Infrastructure

A new survey of 203 enterprise IT decision-makers reveals growing recognition of the tangible benefits that on-premises AI infrastructure delivers. As organizations move beyond early AI experimentation into production-scale deployment, they are increasingly discovering that on-premises and hybrid approaches offer meaningful advantages in data security, cost predictability, and application performance—advantages that are driving significant adoption.

The findings show strong momentum toward on-premises AI. Nearly 79% of respondents have already moved some AI workloads from public cloud to on-premises or private infrastructure, or are in the process of doing so. Looking ahead, 73% plan to either shift workloads to on-premises infrastructure or expand hybrid deployments with increased on-premises capacity over the next 24 months. This does not signal an abandonment of cloud—rather, it reflects a maturing understanding that different AI workloads have different infrastructure requirements, and that on-premises deployment is increasingly the right choice for many of them.

Three factors are driving this trend. First, data sovereignty has become a boardroom priority—nearly three-quarters of respondents report that shadow AI incidents and data residency requirements have directly impacted their AI initiatives, creating demand for infrastructure that keeps sensitive data within organizational control. Second, cloud cost predictability is proving challenging: over 40% of organizations report that actual cloud AI spending exceeds initial projections, making the fixed-cost model of on-premises infrastructure attractive. Third, performance demands for use cases like real-time processing and low-latency inference are accelerating, with 75% of enterprises identifying workloads that benefit from or require on-premises deployment.

This paper presents the complete survey findings, organized around the three primary drivers of on-premises AI adoption: data sovereignty and security, cost predictability and total cost of ownership, and real-time performance requirements.

93% **of enterprises have repatriated some AI workloads, are doing so, or are actively evaluating repatriation.**

The State of Enterprise AI Adoption

The survey confirms that enterprise AI has moved well beyond the experimental stage. Among respondents, 63% reported that their organizations are either scaling AI across multiple business units (26%), implementing production applications (23%), or optimizing AI that is already embedded in core business processes (14%). Another 25% have completed pilots or proofs of concept. Only 12% remain in the early exploration phase.

This level of maturity is significant because it means the infrastructure decisions these organizations face are not theoretical. They are running production workloads with real data, real performance requirements, and real cost structures—conditions under which the specific benefits of on-premises AI deployment become increasingly apparent.

Data Sovereignty and Security: The Primary Driver for On-prem

A Growing Wave of Workload Repatriation

One of the most notable findings of the survey is the scale of AI workload repatriation already in progress. When asked whether their organization has moved any AI workloads from public cloud to on-premises or private infrastructure in the past 24 months, 79% answered affirmatively. Of these, 26% reported repatriating significant workloads, while 53% have moved some workloads or are currently in the process of doing so. An additional 13% are actively evaluating repatriation. Only 5% indicated their workloads remain in public cloud with no plans to move.

79%

of enterprises have already moved AI workloads from public cloud to on-premises or private infrastructure, or are in the process of doing so.

These numbers reflect a growing recognition among enterprises that certain AI workloads—particularly those involving sensitive data, regulatory compliance, or mission-critical operations—are better served by infrastructure that organizations directly control.

AI WORKLOAD REPATRIATION FROM PUBLIC CLOUD (PAST 24 MONTHS)

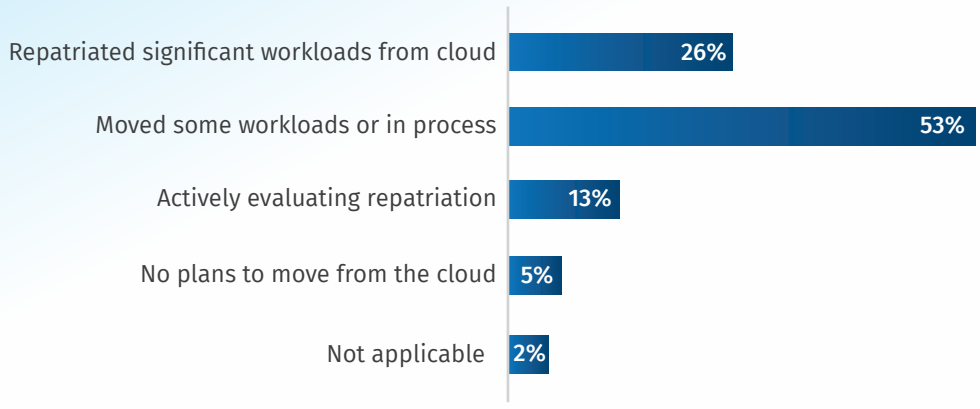


Figure 1: AI Workload Repatriation from Public Cloud (Past 24 Months), n-203

Shadow AI Has Become a Significant Security Concern

The unauthorized use of cloud-based AI tools—commonly referred to as "shadow AI"—has emerged as a significant enterprise security concern and a catalyst for considering on-premises alternatives. A full 74% of respondents characterized shadow AI as either a critical or significant data security concern. Nearly one in four (24%) reported documented incidents of employees uploading confidential data to cloud AI tools like ChatGPT. Half of all respondents (50%) have implemented controls specifically to prevent sensitive data from reaching cloud AI services.

"SHADOW AI" AS A DATA SECURITY CONCERN

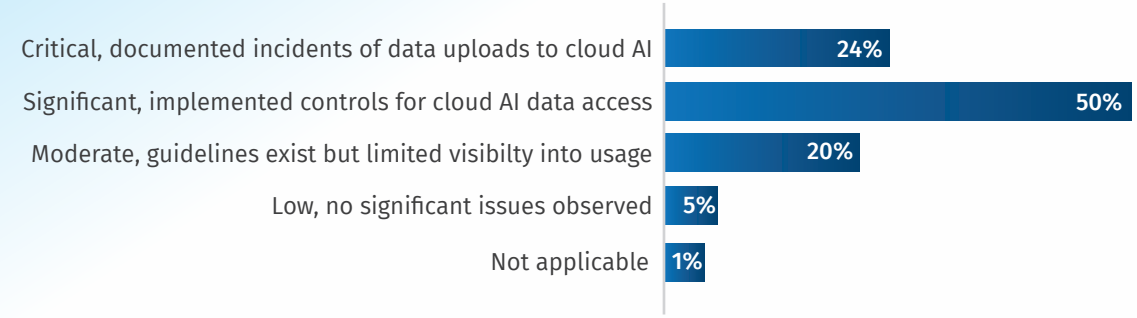


Figure 2: Shadow AI as a Data Security Concern, n=203

These figures illustrate an enterprise landscape where the convenience of cloud-based AI tools has created new data exposure risks that many organizations find difficult to manage through policy alone. On-premises AI infrastructure offers a complementary approach: by enabling AI capabilities within the organization's own perimeter, enterprises can give employees access to powerful AI tools while maintaining control over where sensitive data resides and how it is processed.

Data Residency Concerns Are Constraining AI Adoption

Data sovereignty is not merely a theoretical concern—it is actively shaping AI deployment decisions. When asked whether their organization has declined, delayed, or scaled back an AI initiative due to concerns about sensitive data leaving their premises or jurisdiction, 58% answered yes. Among those, 20% reported that multiple initiatives have been affected, and 37% that at least one significant initiative was impacted. An additional 34% indicated that while no initiatives have been blocked, data residency is a consideration in every AI project.

IMPACT OF DATA RESIDENCY CONCERNS ON AI INITIATIVES

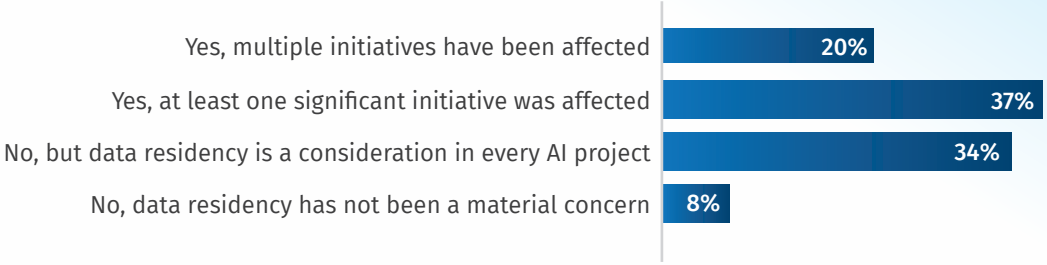


Figure 3: Impact of Data Residency Concerns on AI Initiatives, n=203

58%

of enterprises have declined, delayed, or scaled back an AI initiative due to data residency concerns.

Only 8% of respondents said data residency has not been a material concern. This means that for over 91% of enterprises, data residency is a factor in AI planning—and for the majority, it has directly limited their ability to deploy certain AI initiatives. On-premises infrastructure can help remove this constraint, enabling organizations to pursue AI use cases that might otherwise remain stalled due to data governance requirements.

For Sensitive Data, Enterprises Prefer Non-Cloud Infrastructure

The survey posed a direct question: if deploying a new AI application involving sensitive company data today, which infrastructure approach would the organization most likely choose? The results were clear. A combined 91% of respondents would choose on-premises infrastructure (20%), a private cloud or hosted private environment (33%), or a hybrid approach with sensitive data processing on-premises (38%). Only 8% would choose public cloud with enhanced security controls, and just 1% would use public cloud with standard configuration.

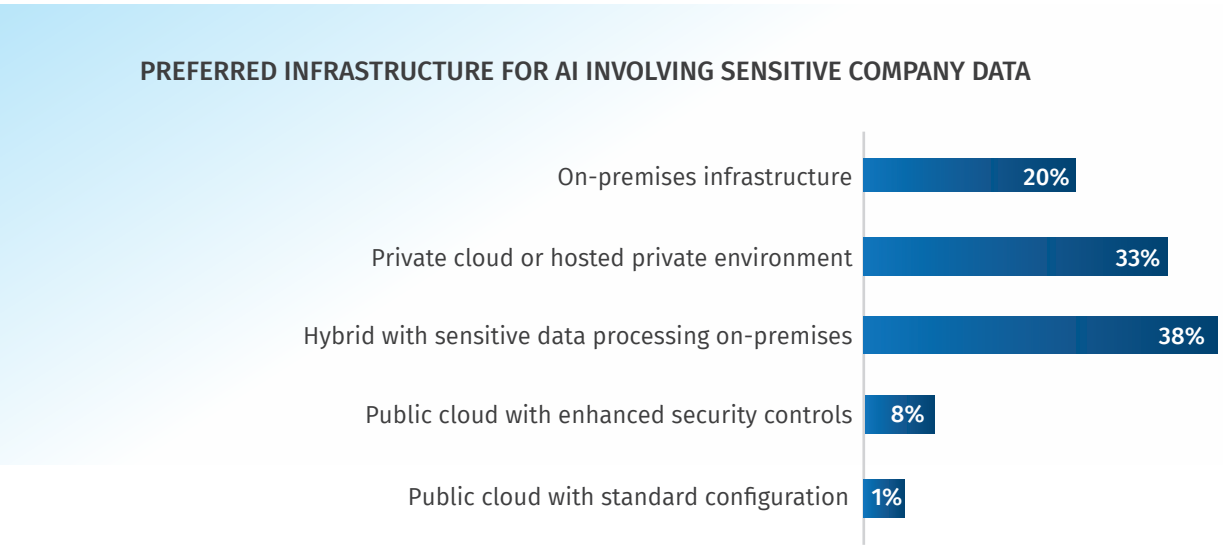


Figure 4: Preferred Infrastructure for AI Involving Sensitive Company Data, n=203

This finding highlights a clear pattern in enterprise thinking. When the stakes are highest—when sensitive data is involved—organizations overwhelmingly prefer infrastructure models that keep data under their own control. Public cloud remains valuable for many workloads, but for AI applications that touch an organization's most sensitive information assets, enterprises see on-premises and private infrastructure as the more appropriate choice.

91%

of enterprises would choose on-premises, private cloud, or hybrid infrastructure for AI involving sensitive data.

Cost Predictability: A Growing Advantage of On-Premises AI

Cloud AI Spending Frequently Exceeds Projections

Cost predictability has become a notable pain point for organizations running AI in the cloud. The survey found that 40% of respondents report their actual cloud AI spending exceeds initial projections. Of these, 35% are over budget by 10–30%, and 5% are significantly over budget by more than 30%. Only 3% report coming in under budget.

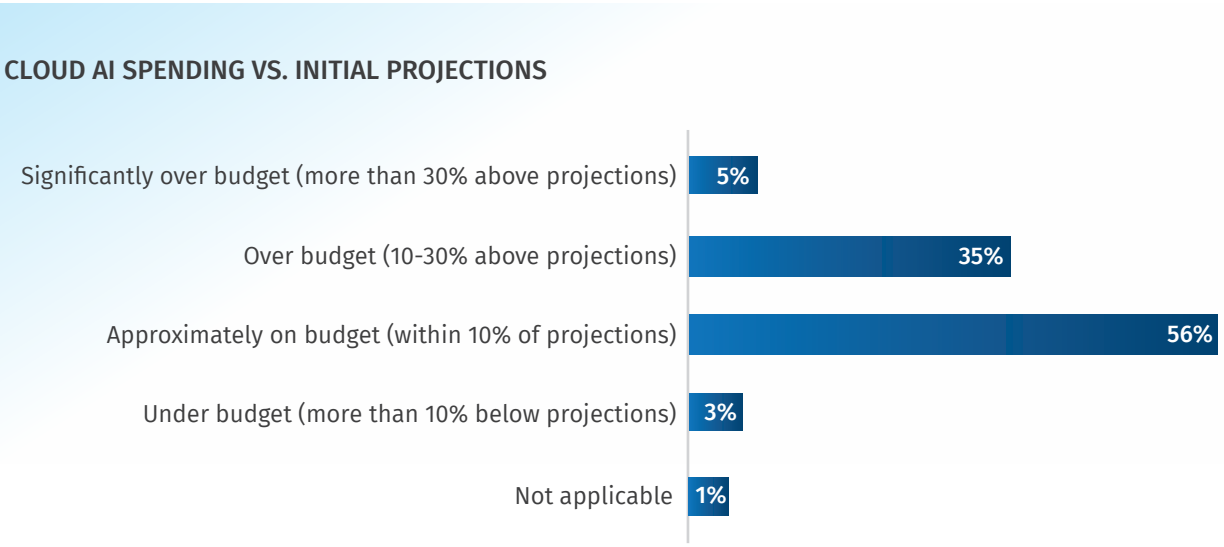


Figure 5: Cloud AI Spending vs. Initial Projections, n=203

When asked which cost-related factors present the greatest challenge to expanding AI adoption, respondents identified data preparation and integration expenses (45%), the high cost of specialized AI talent (34%), software licensing costs (31%), difficulty forecasting total cost of ownership for cloud AI services (25%), and unpredictable consumption-based cloud pricing (23%). These last two factors—representing cloud-specific cost unpredictability—were cited by nearly half of respondents when combined.

On-premises AI infrastructure can help address the cost predictability challenge. With fixed hardware costs and predictable software licensing, organizations can forecast AI infrastructure spending with greater accuracy than consumption-based cloud models that fluctuate with usage patterns, data volumes, and provider pricing changes. For organizations running AI at scale, this predictability can translate into meaningful total cost of ownership advantages over time.

40% of enterprises report cloud AI spending exceeds initial projections.

Performance Requirements: Where On-Premises AI Excels

The third pillar of the on-premises value proposition is performance. AI use cases involving real-time processing—video surveillance, manufacturing quality control, real-time transaction processing, and similar applications—impose latency requirements that can be difficult to meet consistently with cloud infrastructure alone.

75%

of enterprises have production or planned AI workloads that require or would benefit from on-premises deployment for latency performance.

The survey found that 75% of respondents identified current or planned AI workloads that require or would benefit from on-premises deployment for acceptable performance. Specifically, 37% have production workloads that already require on-premises AI due to real-time performance requirements, and an equal 37% are planning workloads that will require on-premises deployment due to latency constraints. An additional 19% are evaluating latency-sensitive use cases that may require on-premises infrastructure.

Only 4% of respondents said their AI use cases do not have performance requirements that would benefit from on-premises infrastructure. This finding underscores a practical reality of enterprise AI: for workloads where response time matters, on-premises deployment offers a meaningful performance advantage by placing compute resources closer to the data sources they serve, eliminating the network latency inherent in cloud round-trips.

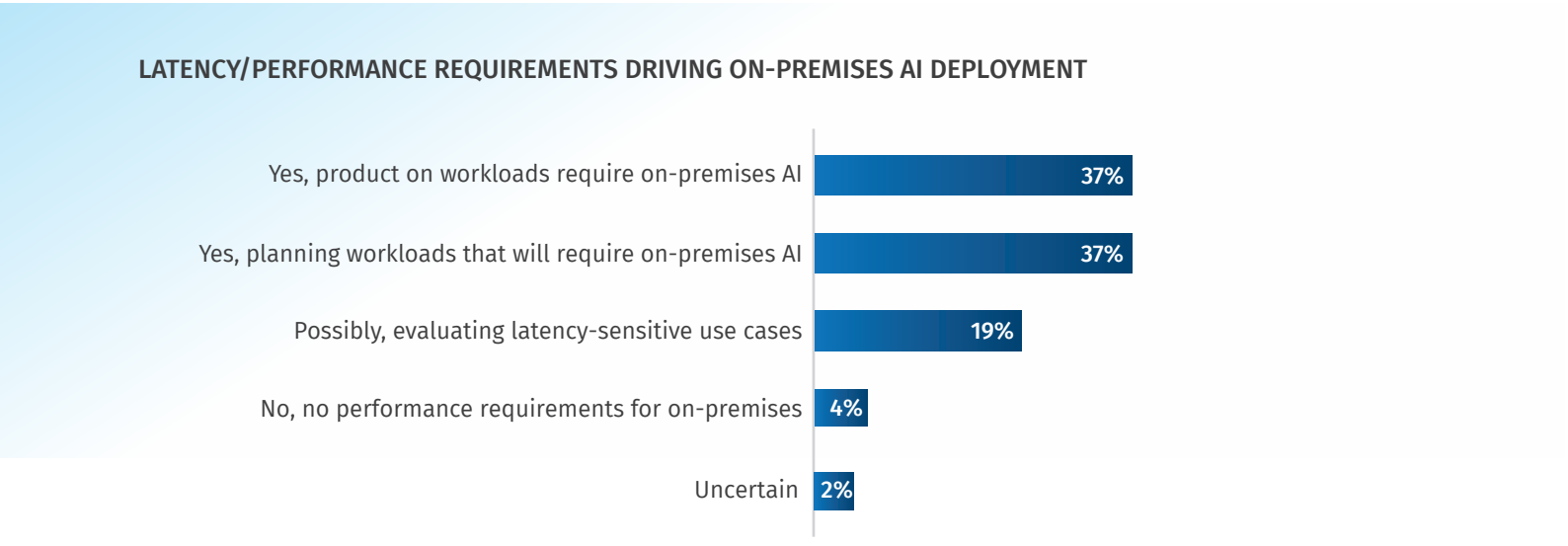


Figure 6: Latency and Performance Requirements Driving On-Premises AI Deployment, n=203

The Path Forward: Growing Momentum for On-Premises AI

Infrastructure Strategies Are Evolving Toward Hybrid and On-Premises

Looking ahead 24 months, the direction of enterprise AI infrastructure strategy is clear. Over seven in ten respondents (73%) plan to shift AI workloads from public cloud to on-premises or private infrastructure (22%) or adopt and expand a hybrid approach with increased on-premises capacity (51%). Just 12% plan to maintain their current infrastructure balance, and only 10% intend to increase public cloud usage.

Importantly, the most popular response—chosen by over half of respondents—was the hybrid approach with increased on-premises capacity. This suggests that enterprises are not abandoning cloud wholesale, but rather evolving toward a more balanced infrastructure model in which on-premises deployment plays a larger and more strategic role, particularly for workloads involving sensitive data, real-time performance requirements, or predictable high-volume processing.

EXPECTED AI INFRASTRUCTURE STRATEGY EVOLUTION (NEXT 24 MONTHS)

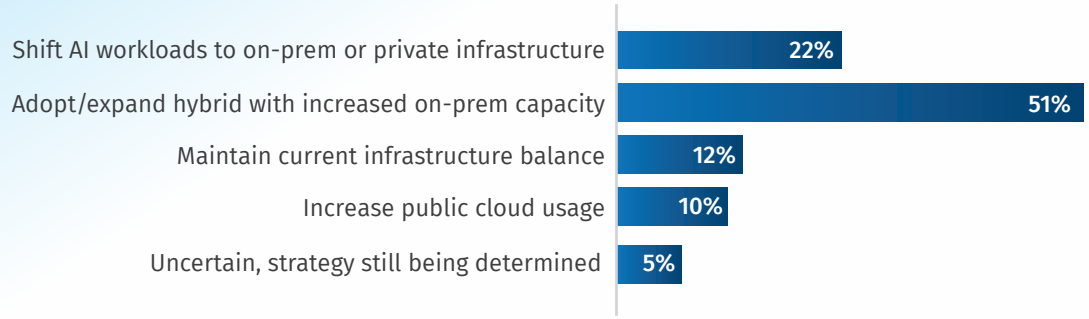


Figure 7: Expected AI Infrastructure Strategy Evolution (Next 24 Months), n=203

What Enterprises Value in On-Premises AI

When asked which factors would most increase their likelihood of deploying AI infrastructure on-premises, respondents ranked data privacy and sovereignty guarantees first (53%), followed by better performance for latency-sensitive applications (43%) and improved total cost of ownership compared to cloud alternatives (40%). Vendor-managed support that reduces internal staffing requirements (22%) and turnkey solutions that reduce implementation complexity (17%) rounded out the list. Only 4% expressed no interest in on-premises AI infrastructure.

TOP FACTORS THAT WOULD INCREASE LIKELIHOOD OF ON-PREMISES AI DEPLOYMENT

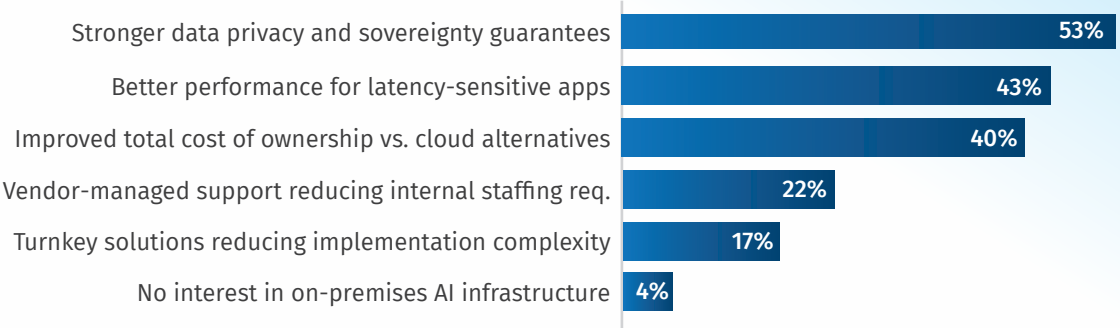


Figure 8: Factors that would Increase Likelihood of On-Premises AI Deployment, Select Up to Two, n=203

These results confirm that the three core value propositions of on-premises AI—sovereignty, performance, and cost—align closely with what enterprise buyers are seeking. Notably, the secondary factors of vendor-managed support and turnkey solutions point to a market opportunity for infrastructure providers that can reduce the operational complexity traditionally associated with on-premises deployment.

What Enterprises Value in On-Premises AI

The budget environment for AI remains strongly favorable. A commanding 86% of respondents expect their organization's annual AI budget to increase in 2026 compared to 2025. Of these, 7% anticipate increases of more than 50%, 33% expect increases of 25–50%, and 46% project increases of 10–24%. Only 1% expect a decrease.

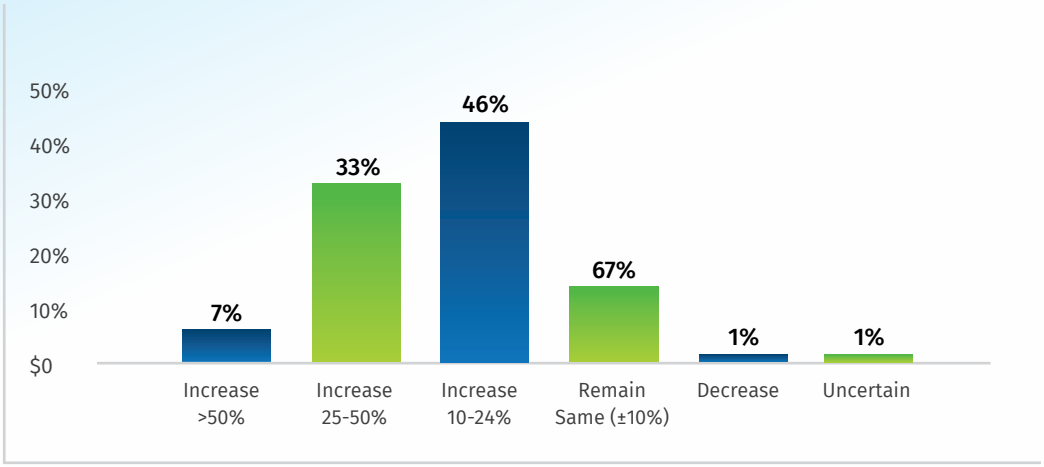


Figure 9: Expected AI Budget Change for 2026 vs. 2025, n=203

With budgets growing and infrastructure strategies increasingly incorporating on-premises capacity, the conditions are favorable for significant enterprise investment in on-premises AI infrastructure over the coming year.

Conclusion

The results of this survey paint a clear picture of an enterprise AI landscape in transition. Across 203 IT decision-makers spanning multiple industries, three findings stand out.

First, enterprises are recognizing the data sovereignty benefits of on-premises AI. Nearly 79% have already begun moving some AI workloads on-premises, driven by direct experience with shadow AI incidents, evolving data residency requirements, and the recognition that sensitive data warrants infrastructure under their own control. This does not mean the cloud is being abandoned—rather, organizations are making more deliberate choices about which workloads belong where.

Second, the economics of on-premises AI are proving attractive for many use cases. Over 40% of organizations report cloud AI spending above projections, and nearly half cite cloud-specific cost unpredictability as a barrier to AI expansion. On-premises infrastructure offers a more predictable cost model that enables organizations to invest confidently in AI at scale.
Third, the performance demands of enterprise AI workloads—real-time processing, low-latency inference, and edge deployment—naturally favor on-premises infrastructure for many applications. Three-quarters of respondents identified workloads that benefit from or require on-premises deployment for acceptable performance.

Taken together, these findings describe an enterprise market that is evolving beyond the cloud-first assumptions of early AI adoption toward a more nuanced, workload-driven infrastructure strategy. On-premises AI is not replacing the cloud—it is emerging as an essential complement to it, offering distinct advantages in security, cost predictability, and performance that enterprises increasingly recognize and value. Organizations that develop a thoughtful on-premises AI strategy today will be well positioned to deploy AI more securely, more predictably, and with better performance across the workloads that matter most.

Learn more about Cloudian's HyperScale AI Data Platform, go to www.cloudian.com/hyperscale

SURVEY METHODOLOGY

**This survey was conducted in February 2026 via the Centiment research platform. A total of 203 qualified respondents completed the survey. All respondents were screened to ensure decision-making authority over AI strategy, investments, or infrastructure: 60% identified as primary decision-makers or budget owners, 32% as key influencers or members of the decision-making team, and 8% as individuals who are regularly consulted on these decisions.*

The respondent pool represented a cross-section of enterprise industries, with technology and telecommunications (25%), manufacturing (16%), retail and e-commerce (13%), financial services and insurance (12%), and healthcare and life sciences (9%) comprising the largest segments. Government and public sector (6%), professional services (7%), and energy and utilities (4%) were also represented.

Respondents were actively engaged with AI: 66% reported having at least one AI application running in production, 25% had conducted pilots or proof-of-concepts, and 9% were actively planning or budgeting for AI initiatives. The survey included an attention-check question to ensure data quality, which 100% of respondents passed.