S&P Global
Market Intelligence

TELEHOUSE
Every connection matters

AI Workload Strategies 2025

Global IT Leaders' Key Considerations for AI Workload Placement

Table of Contents

01 Introduction

02 Key Findings

04 Workload Trends

06 Venue Selection and Decision Criteria

08 AI Training Models and Data Requirements

09 Networking Issues

10 The Importance of Cloud On-ramps

11 Geographic Trends

12 Purchasing Decisions for AI workload Colocation Services

13 Conclusions

14 Considerations for AI Workload Placement

15 About This Research

Introduction

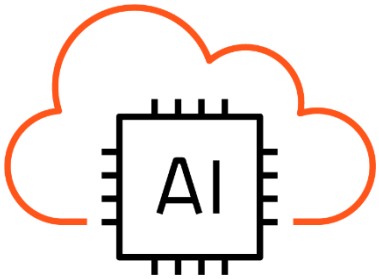
Businesses are increasingly exploring the potential of AI while also seeking to understand the challenges inherent in developing the optimised server and networking infrastructure needed to fully benefit from that potential.

AI deployments are routinely distributed across various computing infrastructures and venues for model training or inferencing. As enterprises scale these workloads, they encounter issues of latency, performance and data security management that may necessitate tailored network and fibre connectivity, as well as GPU infrastructure availability and high-density compute options.

Due to the complex nature of these requirements, there is no one-size-fits-all solution. Most businesses consider cloud hyperscalers, on-premises data centres or third-party colocation providers as potential hubs for AI infrastructure. Approaches vary significantly across industry verticals and regions. This survey delves into the trends influencing AI/ML workload placement, exploring the primary factors driving venue selection and potential near-term changes, as well as regional and sector-specific variations.

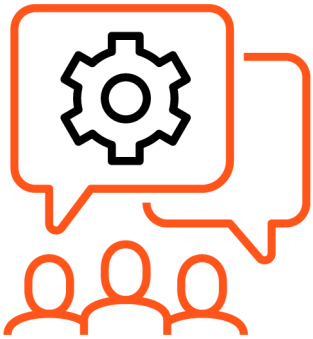
The **AI Workload Strategies 2025** study surveyed more than 900 senior IT executives globally to understand changes needed in their IT infrastructure to address the pressures of evolving AI workloads. It also explored their preferences and pain points in the early stages of AI adoption and how their architecture might evolve.

Key Findings

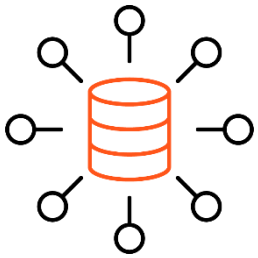


AI initiatives are multi-venue, not just hyperscaler workloads. Deployments are widely distributed across venues, with **35%** allocated to public cloud. On-premises data centres, third-party colocation and specialist GPU clouds are important parts of the mix, each reportedly housing about 10% of AI workloads.

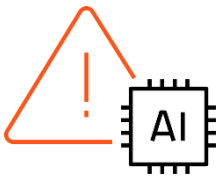
Key drivers for AI workload venue selection include availability of internal IT skills, availability of network/fibre connectivity and application requirements. GPU-based infrastructure availability as well as operating versus capital expenditure considerations are also expected to gain importance. All venues are projected to grow at similar rates, with respondents indicating gradual evolution of their AI strategy and venue selection in the next two years.



55%
of businesses have experienced significant **network issues** with AI



39%
have **abandoned AI** projects

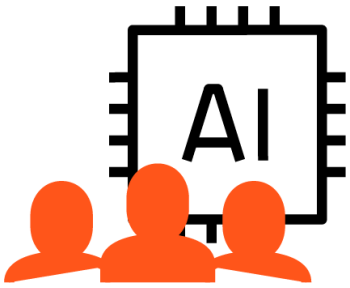


97%

of businesses state that **cloud on-ramp services** are critical or quite important to AI/ML architecture

Geographically, most respondents will consider placing AI workloads outside their operating markets, particularly for model training. European respondents lag in AI workload dedication and are less likely to engage with specialist GPU cloud or edge environments, while respondents in North America and Asia-Pacific are more likely to report networking challenges with their AI deployments.

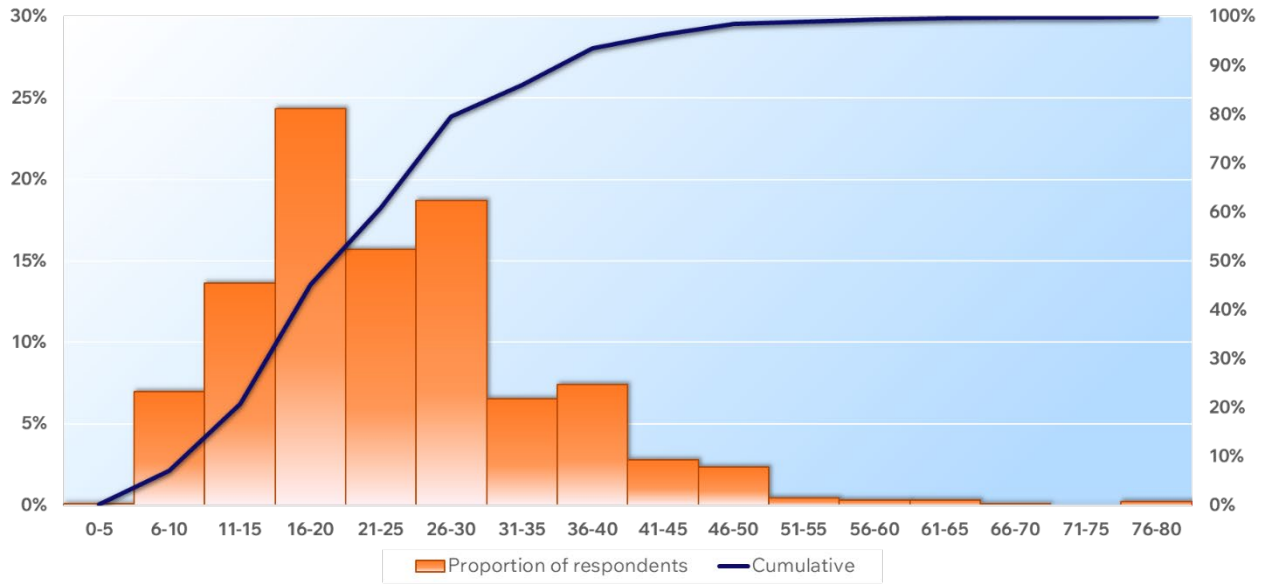
GPU or AI-related services are key factors in selecting colocation providers for AI/ML workloads. Companies show keen interest in AI-related services, suggesting that these services can enhance colocation's appeal as a destination for AI workloads.



Workload Trends: Organisations that are investing in AI are investing heavily

Trends among early AI adopters reveal significant investment in AI. The largest share of survey respondents state that 16%-20% of their overall workloads are AI-related, with a further significant proportion dedicating 21%-30% of workloads to AI. In other words, a considerable percentage of enterprise infrastructure is dedicated to AI deployment.

Percentage of workloads allocated to AI/ML

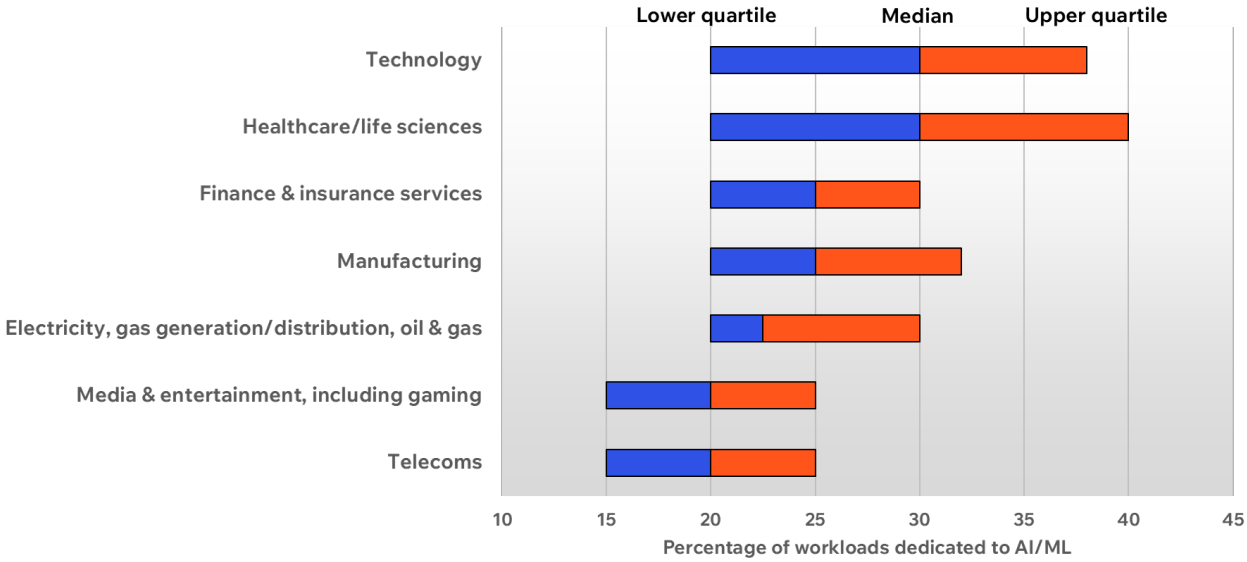


Source: S&P Global Market Intelligence 451 Research, AI Workload Strategies, 2025

Larger companies are more likely to allocate a greater share of their workloads to AI. Respondents from companies with 5,000 or more employees on average report that 31% of total workloads are dedicated to AI, versus 23% among smaller companies with 100-999 employees.

Industry verticals also reflect significant differences, with technology and healthcare/life sciences companies allocating 30% of their total workloads to AI. Financial and insurance services as well as manufacturing companies are not far behind, with a median of 25% of workloads dedicated to AI. Media and entertainment along with telecoms report the lowest median allocations, at 20%.

AI/ML workload allocation by industry



Source: S&P Global Market Intelligence 451 Research, AI Workload Strategies, 2025

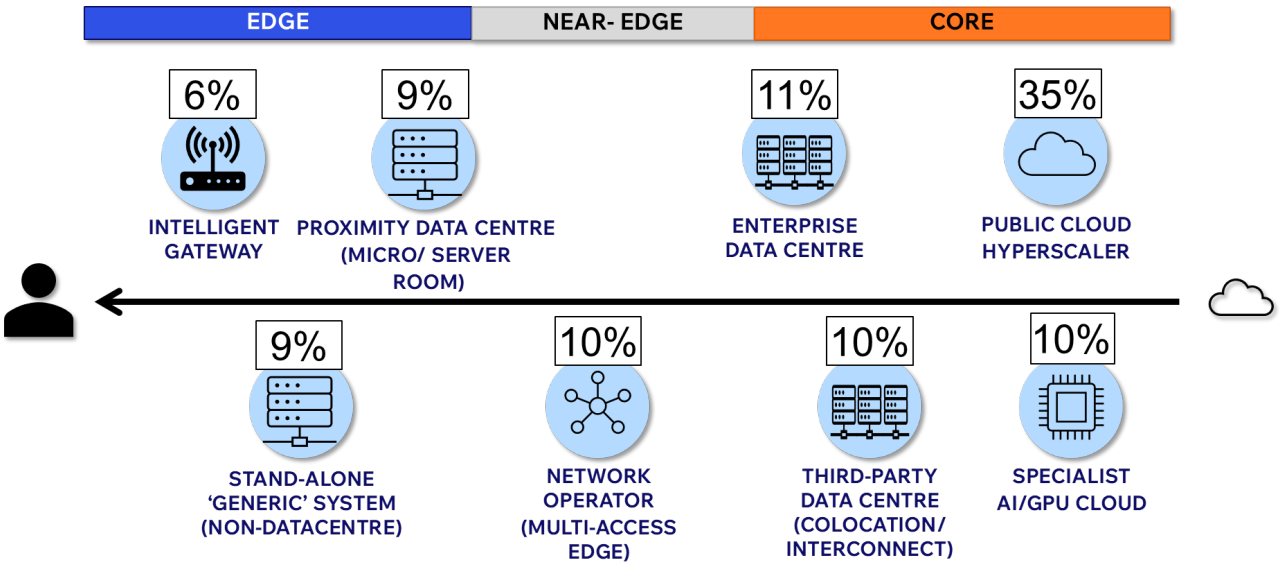
Geographic disparities are also apparent. North America leads with a median 28% of workloads dedicated to AI, followed by Asia-Pacific at 25%, while EMEA trails at 20%. Broadly speaking, companies in Europe are dedicating smaller shares of their workloads to AI.

Venue Selection and Decision Criteria

In addition to indicating the proportion of infrastructure dedicated to AI, the survey results reveal which venues enterprises most often pick for these workloads.

AI workloads are widely distributed across venues. Hyperscalers hold the largest share, with 35% of respondents' AI/ML workloads placed in public cloud. However, the split among other venues is quite even. On-premises data centres are the venue of choice for 11% of these workloads, while third-party colocation data centres, specialist GPU clouds and network operators are close behind with 10% each. Proximity data centres (on-premises micro data centres or server rooms) and stand-alone non-data centre systems account for 9% each, while intelligent gateways rank last at 6%.

Wide distribution of AI workloads across venues



Source: S&P Global Market Intelligence 451 Research, AI Workload Strategies, 2025

This distribution in venue selection is expected to remain stable in the short term. Respondents anticipate a small uptick in public cloud workloads, rising from 35% to 38%, while other venues are expected to grow at similar rates.

Looking at the results by industry, companies in the media and entertainment, telecoms, and electricity/oil and gas verticals are more likely to opt for colocation to house a significant chunk (15% or more) of their AI/ML workloads. In fact, over 40% of respondents in those verticals use colocation in this manner.

Geographically, companies in Europe, specifically in the UK and France, are more likely to adopt colocation for these workloads. Over half of respondents in the UK, just under half in France and more than a third in Germany, place 15% or more of their AI workloads in colocation environments.

Organisations consider several factors when selecting AI workload venues. Key drivers include availability of **internal IT skills** (deemed critically important by 52% of respondents), **network/fibre connectivity** (50%) and **latency/proximity requirements** (42%). Organisations do not anticipate major changes in the factors they view as critical. However, as organisations transition from experimenting with generative AI to deploying it at scale, they report an increased focus on GPU availability, likely reflecting growing resource demands.

By industry, companies in financial and insurance services more often cite application requirements and business considerations as important drivers of venue selection, while healthcare and life sciences companies identify availability of IT skills and expertise as most critical.

Critical criteria for AI workload venue selection, by industry

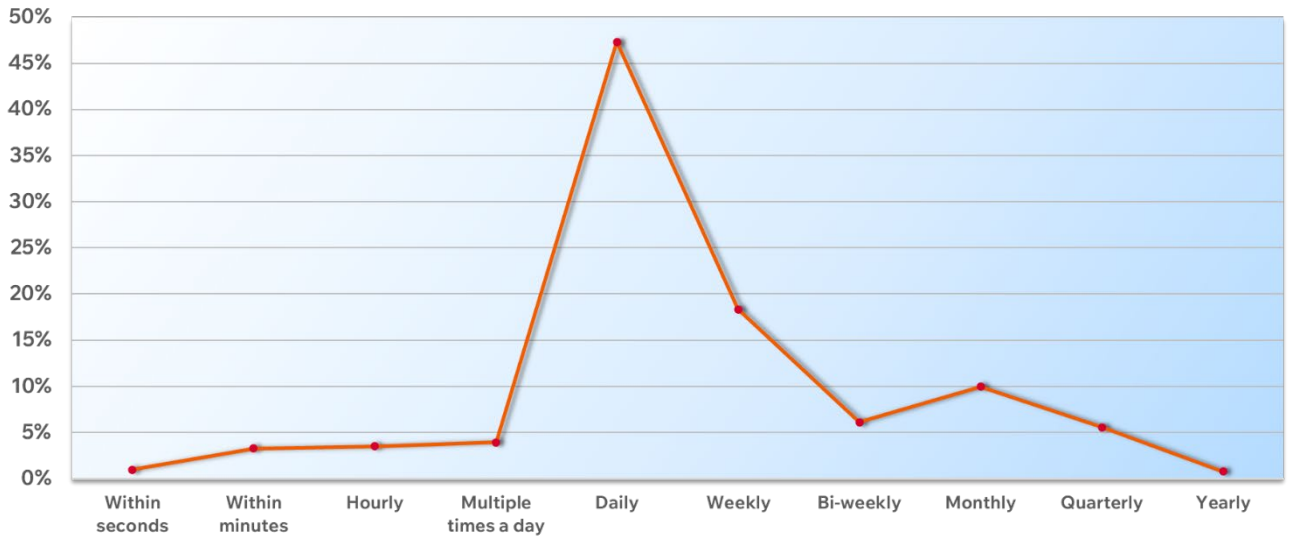
	Application requirements	Availability of GPU-based infrastructure	Availability of internal IT skills and expertise	Latency/ proximity requirements	Line-of-business/ developer preferences	Price/ operating cost	Business considerations	Regulatory/compliance issues	Financial strategy	Global/ regional business requirements	Software vendor requirement/ licensing	Vendor/ service provider recommendations /guidance	Available network/ fiber connectivity	Scalability and flexibility	Energy and sustainability considerations
Manufacturing	52%	38%	48%	46%	39%	48%	44%	44%	39%	48%	42%	41%	51%	45%	44%
Technology	52%	52%	52%	44%	49%	51%	44%	53%	45%	42%	49%	46%	47%	49%	49%
Electricity, Gas Generation/ Distribution, Oil & Gas	44%	39%	43%	39%	36%	36%	35%	39%	47%	36%	44%	26%	41%	48%	37%
Telecoms	36%	54%	45%	32%	48%	45%	42%	41%	52%	43%	50%	49%	42%	46%	41%
Finance & insurance services	67%	45%	56%	47%	43%	45%	61%	54%	50%	51%	48%	54%	57%	55%	53%
Healthcare/Life Sciences	46%	57%	63%	47%	48%	46%	54%	56%	52%	46%	57%	51%	52%	49%	51%
Media & Entertainment	35%	39%	52%	34%	39%	41%	39%	50%	39%	45%	39%	39%	55%	34%	45%

Source: S&P Global Market Intelligence 451 Research, AI Workload Strategies, 2025

AI Training Models and Data Requirements

In addition to altering how enterprises think about their geographic footprint, AI use — particularly for model training — is likely to change how organisations consume data centre power. Frequency of model training is critical in understanding an enterprise's future power demand. The largest proportion of organisations (nearly half) report daily model training, while close to 20% say their models receive weekly training, and only about 10% say monthly.

Frequency of retraining for primary AI model



Source: S&P Global Market Intelligence 451 Research, AI Workload Strategies, 2025

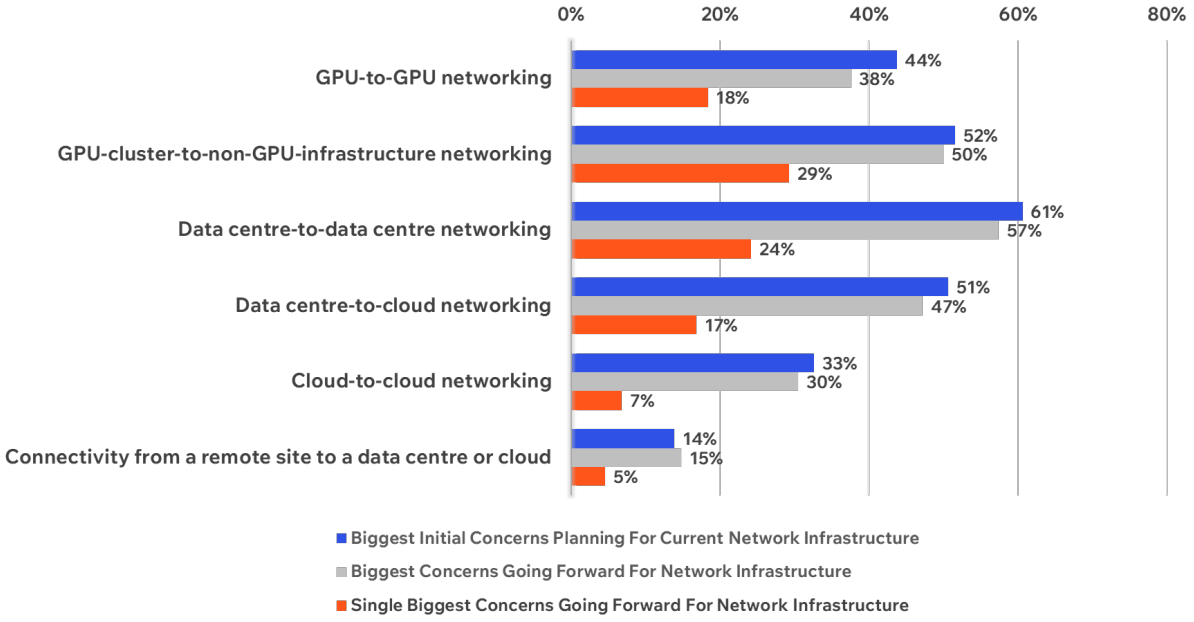
Companies in the healthcare and life sciences sectors are somewhat more likely to train models more than once a day, at 18%, compared to the survey-wide average of 12%. Enterprises in the technology, electricity/oil and gas, and manufacturing sectors are also more likely than average to pursue daily training.

The bulk of organisations report using between 500 TB and 49 PB to build and train their AI models. The amount of data used correlates with company size, larger companies of more than 5,000 employees more likely to utilise greater volume of data compared to their smaller counterparts.

Networking Issues

Networking issues are a notable pain point for enterprises deploying AI workloads: **55%** of respondents report experiencing significant problems, and **39%** have abandoned AI projects as a result. Latency is the primary issue, but bandwidth and proximity issues are also common. These problems affect various networking connections, including cloud-to-cloud, data centre-to-data centre, and GPU-cluster-to-non-GPU-infrastructure.

Network infrastructure issues for AI/ML workloads



Source: S&P Global Market Intelligence 451 Research, AI Workload Strategies, 2025

Looking across verticals, organisations in the technology and media and entertainment sectors are most likely to report severe networking issues (i.e., where projects have been abandoned), and this trend is projected to persist over the next 24 months.

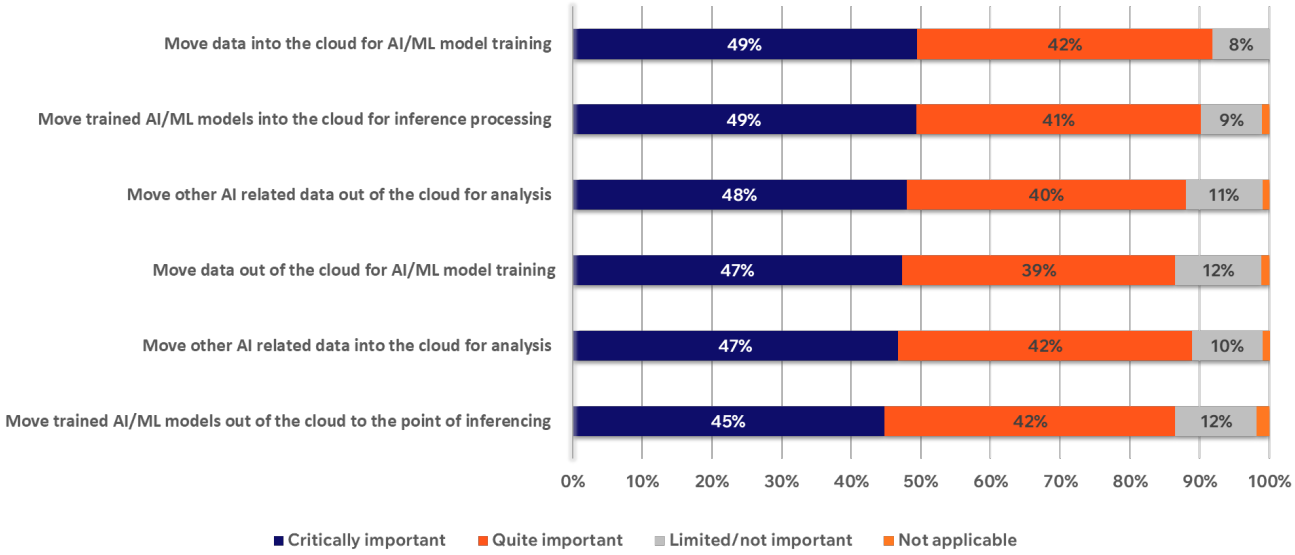
Data centre-to-data centre networking is the single biggest concern for media and entertainment organisations. GPU-cluster-to-non-GPU-infrastructure networking is the single biggest concern for smaller organisations. Generally, smaller organisations are more likely to have experienced severe networking issues leading to project abandonment, although these companies are optimistic that they will face fewer challenges in the next 24 months.

The Importance of Cloud On-ramps

Cloud on-ramps are top of mind for enterprises deploying AI workloads. **More than 90%** of companies view access to cloud on-ramps as critical or quite important to AI/ML architecture. On-ramps are considered important across all regions: 70% of respondents in Asia-Pacific deem them critically important, along with over 60% of respondents in Europe and North America.

Enterprises use cloud on-ramps for multiple AI-related functions, including moving data into and out of the cloud for training and inferencing, as well as moving other AI-related data for analysis. Respondents report using on-ramp data transfers primarily at medium and high frequency.

Importance of cloud on-ramps for various uses



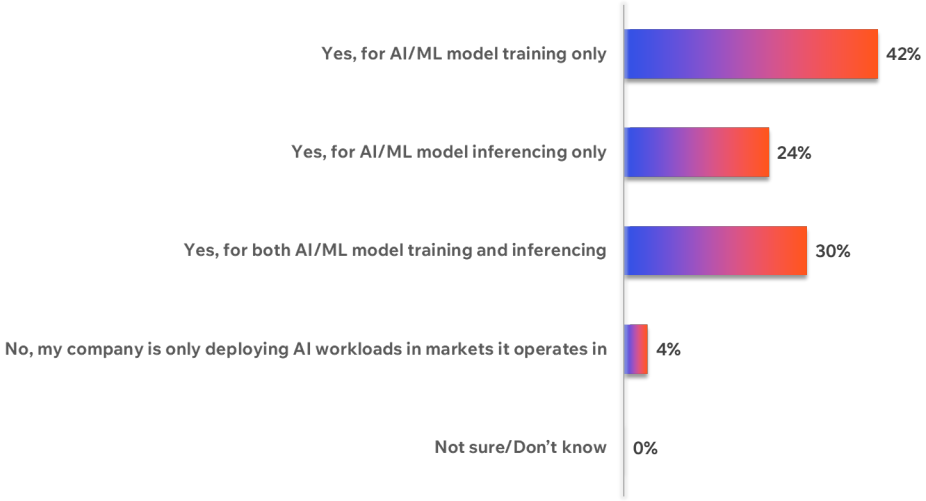
Source: S&P Global Market Intelligence 451 Research, AI Workload Strategies, 2025

Geographic Trends

AI workloads have influenced enterprises' geographic strategies. Respondents report **high geographic flexibility** with model training workloads, but somewhat less flexibility with model inferencing.

More than two in five respondents (42%) consider placing AI training workloads away from markets where they physically operate, while 24% say the same for inference. Notably, 30% express geographic flexibility for both training and inference. Only 4% say their company deploys workloads solely in physical operating markets, meaning most enterprises have at least some geographic flexibilities when placing AI workloads.

Geographic considerations for AI/ML workloads



Source: S&P Global Market Intelligence 451 Research, AI Workload Strategies, 2025

Regionally, Europe stands out once again. Most EMEA organisations (52%) consider placing training workloads outside their physical markets, while a smaller share (17%) considers it for both training and inferencing. Companies in North America and Asia-Pacific reflect similar levels of geographic flexibility for distant placement of AI training (36% North America, 37% Asia-Pacific) and for both training and inferencing (38% North America, 36% Asia-Pacific).

In addition, AI workload deployment is expected to grow in large metropolitan areas with populations of 5 million or more (66% of respondents deploying currently, rising to 77% in two years) and smaller metropolitan areas with populations of less than 1 million (17% currently deploying, rising to 26%).

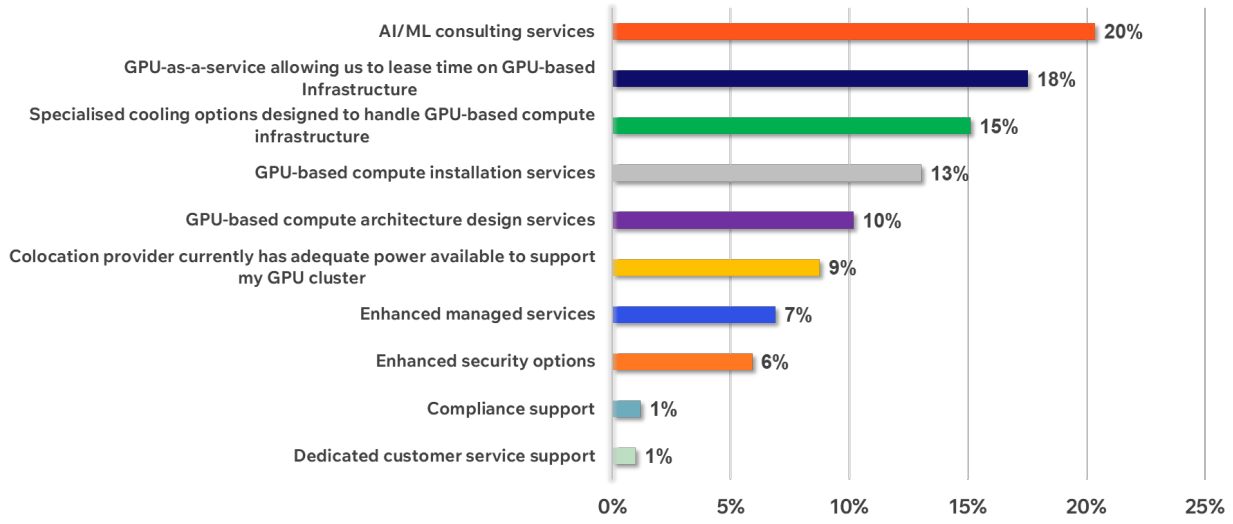
Purchasing Decisions for AI Workload Colocation Services

Survey results also highlight enterprises' preferences and priorities regarding use of colocation for AI architecture. Cost is the top-cited concern when considering colocation for AI workload placement, followed by data and security issues. Power density requirements and GPU infrastructure availability also influence enterprise decisions against adopt colocation.

Companies are broadly seeking AI/ML consulting services, specialized cooling for GPU infrastructure, GPU as a service, and GPU-based compute installation and architecture design. Offering these services could drive new business for colocation providers.

Approximately 20% of respondents cite AI/ML consulting services as the top differentiator that would lead them to consider a colocation provider, followed by GPU as a service (18%), specialized cooling (15%) and GPU-based compute installation (13%).

Most important differentiating service for a colocation provider to offer



Source: S&P Global Market Intelligence 451 Research, AI Workload Strategies, 2025

Conclusions

The AI Workload Strategies 2025 survey provides deep insights into enterprise AI architecture needs. As AI adoption grows, understanding evolving enterprise requirements is critical. Many organisations are ramping up AI efforts in the short term while maintaining their existing IT strategies. Current circumstances illustrate how enterprises can adapt to avoid headaches faced by companies deploying AI.

Four key takeaways for IT leaders

- **AI is cloud-oriented, but multi-platform:** Hyperscale public cloud is the primary destination for AI workloads, but multi-platform strategies will persist.
- **Disparities are evident in AI deployments:** Significant differences exist in AI deployment across geographies and sectors. Companies in North America and Asia-Pacific generally dedicate a larger share of workloads to AI compared to their European counterparts. Enterprises in technology and healthcare/life sciences are investing more heavily in AI than those in other industries. Larger enterprises (5,000+ employees) are investing more in AI compared to their smaller counterparts.
- **Geographic scope is broadening:** AI is prompting enterprises to reconsider where they can place their workloads. Most companies recognize the potential to deploy farther from their physical operations, especially for AI model training.
- **AI-related services are critical:** Cloud on-ramps are viewed as a necessity for AI workloads, as networking issues have plagued AI deployments. Connectivity and networking services are increasingly essential criteria for AI workload venue selection. Further, many enterprises seek third-party AI/ML consulting services, particularly in colocation environments, for customised solutions to their IT needs.

Considerations for AI Workload Placement

- **Determining where an enterprise fits:** Understanding the AI infrastructure strategies of similar-sized companies in comparable industries and geographies can help align a company's AI workload placement strategy with the broader industry. According to this survey, enterprises in the technology and healthcare/life sciences sectors, those based in North America and Asia-Pacific, and those with 5,000+ employees are more likely to dedicate greater proportions of workloads to AI and to invest heavily in AI.

- **Putting the data centre at the centre:** Given the wide distribution of AI workloads, colocation can serve as a central hub for connecting to various venues. Since respondents project that AI workloads will remain distributed, with a slight inclination toward public cloud, cloud on-ramps offered by colocation providers can be particularly valuable for moving data.

- **Understanding AI data consumption:** Analysing trends in enterprise AI model training can help IT leaders anticipate data consumption as companies ramp up AI efforts. The survey results show that companies are using enormous amounts of data and tend to train models relatively frequently, with the plurality doing so daily. As one might expect, larger companies generally use more data to train AI models. For organisations handling massive data volumes and frequent model training, a specialised colocation provider might be the most suitable option.

- **Reconsidering geographic limitations:** Enterprises exhibit increased geographic flexibility for AI workloads. EMEA respondents express a strong openness to placing training models away from their primary business operations, while North America and Asia-Pacific respondents are flexible for both training and inferencing.

About This Research

This **AI Workload Strategies 2025** research was commissioned by Telehouse and S&P Global Market Intelligence, surveyed 915 respondents across a wide range of geographies, industries and job functions for this research. The largest share of respondents was from North America, with 29% based in the United States. In addition, approximately 17% of respondents were from the UK, followed by 12% from China/Hong Kong and 10% from Japan.

The highest proportion of respondents worked in manufacturing sector (approximately 20%), followed by finance and insurance (19%), healthcare and life sciences (13%) and technology and telecoms (13%). Nearly half of respondents (46%) worked in smaller companies with 500-999 employees. Approximately one-third (34%) of respondents were from medium-sized companies with 1,000- 4,999 employees, while 16% were from larger companies of more than 5,000 employees.

In terms of job functions, 68% of respondents worked in IT operations, while 32% held infrastructure-related roles. Respondents also came from a range of seniority levels: 12% were at the CIO or CTO level, 14% at the senior vice president level, 59% at the director level and 15% at the manager level.

About Telehouse

Telehouse is a leading global data centre service provider that brings together a diverse range of business partners including carriers, mobile and content providers, enterprises, cloud providers and financial services companies. Established in 1989 by Fortune 500 telecommunications company KDDI, Telehouse operates more than 45 data centres worldwide and provides reliable, secure, and flexible colocation solutions, enabling organisations to accelerate speed to market and create business opportunities through fast, efficient and secure interconnections. For more information, visit: http://www.telehouse.net

TELEHOUSE

About S&P Global Market Intelligence

At S&P Global Market Intelligence, we understand the importance of accurate, deep and insightful information. Our team of experts delivers unrivalled insights and leading data and technology solutions, partnering with customers to expand their perspective, operate with confidence, and make decisions with conviction.

S&P Global Market Intelligence is a division of S&P Global (NYSE: SPGI). S&P Global is the world's foremost provider of credit ratings, benchmarks, analytics and workflow solutions in the global capital, commodity and automotive markets. With every one of our offerings, we help many of the world's leading organisations navigate the economic landscape so they can plan for tomorrow, today. For more information, visit: https://www.spglobal.com/market-intelligence/

S&P Global
Market Intelligence

About the authors

Pedro Schweizer

Principal Research Analyst, Data center Services and Infrastructure

Pedro Schweizer is a research analyst on the 451 Research Data centre Services & Infrastructure team within S&P Global Market Intelligence. His research focuses on data centre market activity across Europe, Latin America and the US. His key research areas include emerging data centre markets; retail, wholesale, and cloud provider activity; industry growth projections; market share analysis; government incentives and regulations; and pricing dynamics.

Alex Johnston

Senior Research Analyst, Data, AI & Analytics

Alex Johnston is a research analyst on the 451 Research Data, AI & Analytics team at S&P Global Market Intelligence. He focuses on emerging technologies and how they can be applied in business contexts. Alex's primary coverage areas are artificial intelligence, distributed ledger technology, event stream processing and data marketplaces. Alex's recent areas of concentration include monitoring the emerging generative AI market, tracking the evolution in blockchain use cases and investigating real-time architectures.

Dan Thompson

Principal Research Analyst, Data center Services and Infrastructure

Dan Thompson is a principal research analyst in the 451 Research technology research group within S&P Global Market Intelligence. He leads the Data centre Services and Infrastructure team, which is charged with keeping tabs on the data centre industry globally to better understand its trends and growth areas. His research includes analyses of data centre providers, market size and supply/demand in key and emerging markets around the world. Dan also covers data centre providers offering services beyond colocation, such as managed and cloud-type services. In addition, Dan provides research on the sustainability of the data centre industry. Beyond renewable energy purchasing and carbon offsetting, he has been investigating full life-cycle emissions, including supply chain emissions as well as efficiency gains and water usage.

TELEHOUSE

Every connection matters

Contact Telehouse:

Asia

Beijing, China
E: info@telehouse.net.cn
T: +86 (0)10 8755 1756

Shanghai, China
E: shanghai@telehouse.net.cn
T: +86 (0)21 5871 8801

Hong Kong, China
E: corporatesale@hkcolo.net
T: +852 39 75 02 00

Tokyo, Japan
E: inquiry@telehouseglobal.com

Singapore
E: info@kddi.com.sg
T: +65 6220 7001

Bangkok, Thailand
E: sales@telehouse.co.th
T: +662 001 0166

Hanoi, Vietnam
E: contact@telehouse.vn
T: +84 24 35 62 60 18

Europe

Paris, France
E: sales@fr.telehouse.net
T: +33 1 56 06 40 30

Frankfurt, Germany
E: sales@de.telehouse.net
T: +49 (0)69 823 79 48 0

London, United Kingdom
E: sales@uk.telehouse.net
T: +44 (0)20 75 12 00 80

Other EMEA Countries
E: sales@uk.telehouse.net
T: +44 (0)20 75 12 05 50

North America

Toronto, Canada
E: sales@ca.telehouse.com

New York/Los Angeles, USA
E: sales@telehouse.com
T: 844 518 0026
(Only available inside the U.S)

T: +1 718 355 2500
(From outside of the U.S)

S&P Global
Market Intelligence