Google Cloud

2025 research report

State of AI infrastructure

Executive summary

While the explosion of interest in AI has sparked widespread experimentation, the focus has now shifted to practical application. Tech leaders are prioritizing ‘how’ over ‘if’—how to integrate, secure, and scale gen AI to deliver sustainable value. This transformative phase demands a strategic approach to platform management and cost-effective deployment.

However, this transformation is far from straightforward. Organizations face significant challenges, including legacy system integration, complex migrations, and the need to tailor AI solutions to their unique business contexts.

Recognizing the critical need for data-driven insights, we surveyed 500+ global business leaders. Our findings reveal a cloud market undergoing a dramatic shift, where companies actively investing in AI-powered cloud infrastructure are leading the charge in the AI era.

This report cuts through the hype, offering a clear-eyed view of the practical hurdles and strategic opportunities. We provide the research and numbers leaders need to build consensus for essential infrastructure investments, empowering them to create truly AI-powered and cloud-driven organizations and confidently navigate this transformative period.

Nirav Mehta
Google Cloud Sr. Director, Product Management

Key findings

Landscape

01 Gen AI adoption is nearly universal

98% of organizations are actively experimenting with, developing, or using gen AI in production

02 Gen AI ROI relies on both internal and external use cases

64% focus on both enhancing customer experience and improving internal operations

Challenges

03 Security and data are top concerns

70% have experienced difficulties with data governance, integrating data into AI models, and having insufficient training data

04 Cost efficiency is a critical consideration

83% prioritize cost-efficiency when adopting gen AI infrastructure solutions

Infrastructure

05 A robust AI platform is a leading criteria when evaluating infrastructure

06 Edge computing is extending AI's reach

73% consider deploying gen AI models in distributed devices or systems to be important

07 Hybrid cloud offers flexibility and control

74% prefer a hybrid cloud approach for gen AI deployments

08 Most organizations rely on gen AI solutions from cloud providers

Optimized infrastructure is crucial for gen AI success

The success of any gen AI initiative hinges on the underlying infrastructure. This isn't simply about adding another workload, it's about building a foundation that can support the unique and rapidly evolving demands of AI—regardless of whether you're aiming to increase customer satisfaction or reduce operational costs.

This foundation needs to do some heavy lifting. It must be secure to safeguard both intellectual property and customer data. It has to be distributed—providing managed services in the public cloud and extending into your data center or to the network edge for latency and data sovereignty requirements.

It needs to be scalable, performant, and reliable to handle the parallel processing of large training datasets, as well as the extended context windows and multi-step processing of inference and reasoning. And of course—it has to be cost efficient.

Sound like an impossible task? Not if you have a plan. Gen AI is redefining organizations' IT strategy, and it isn't slowing down. This research reveals where leading businesses across all industries are on their journey. As you re-examine your cloud infrastructure strategy, consider which providers have capabilities to serve not only AI experts, but also your developers and lines of business.

Finding 01

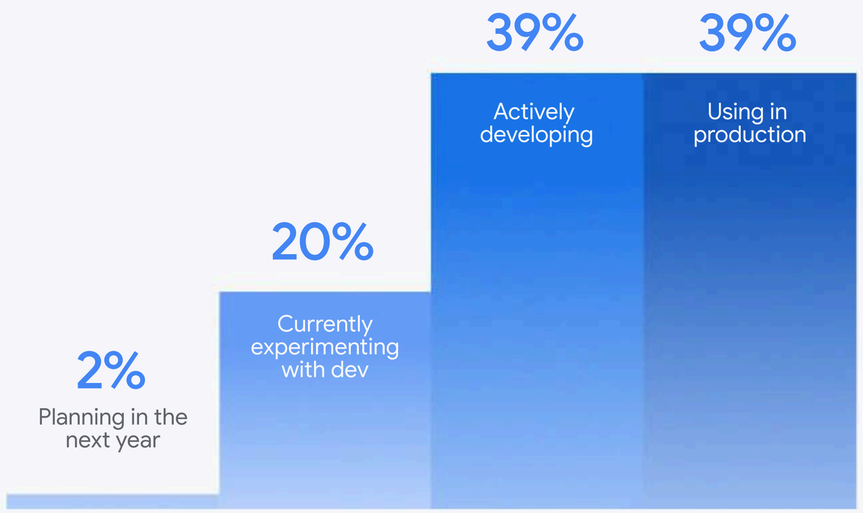
Gen AI adoption is almost universal across industries

The importance of gen AI in today's workplace simply can't be overstated.

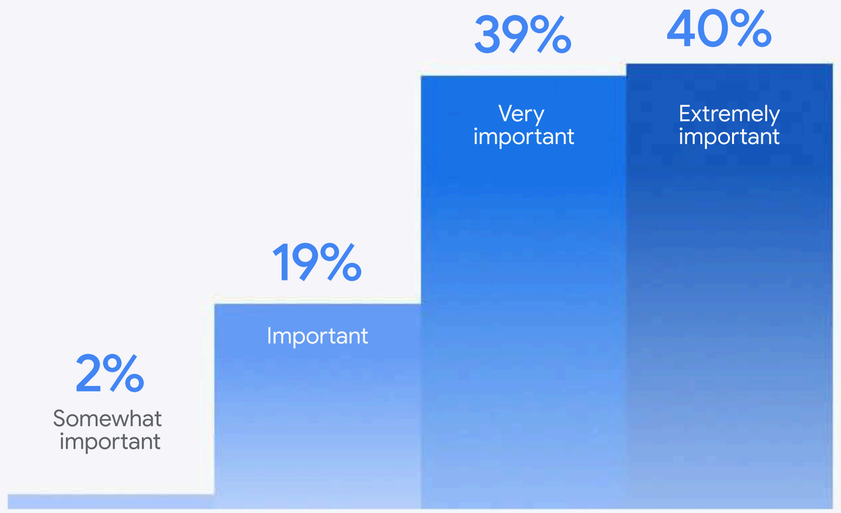
98% of organizations are actively experimenting with, developing, or using gen AI in production

79% of technology leaders consider gen AI to be at least very important to their organization's current and future business operations

At what stage is your organization in terms of gen AI adoption?



How important is gen AI to your organization's current and future business operations?



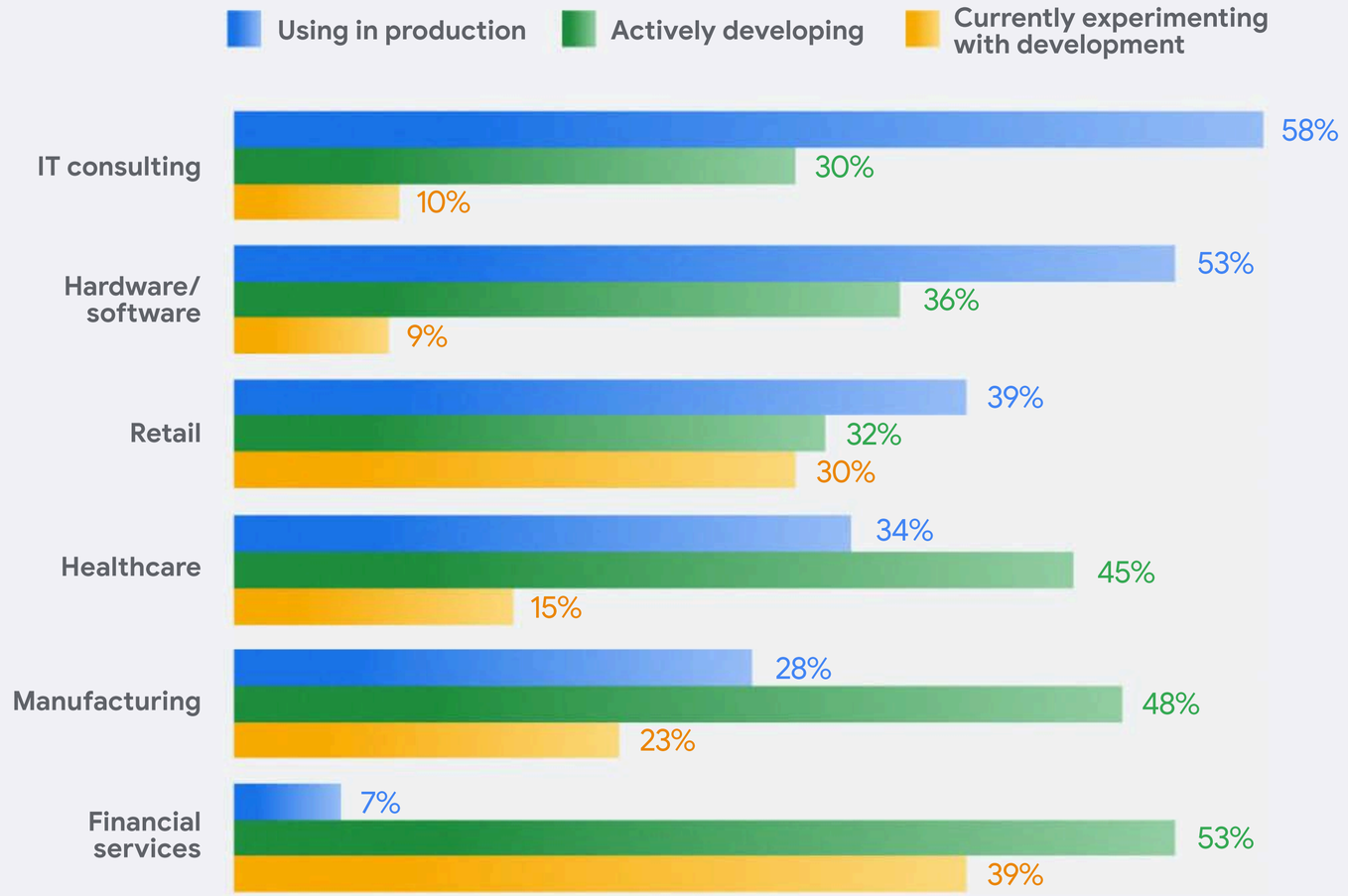
A look across industries

These findings are consistent across industries—technology leaders across the board have overwhelmingly adopted gen AI in their organizations and agree that it's essential to their business. Let's take an in-depth look at the data.

Across industries, the majority of organizations are either actively developing or deploying gen AI in production, showing that companies are already investing heavily in gen AI and are planning to continue to invest. As may be expected, organizations in tech-related industries have been both quicker to move gen AI pilots to production and have higher gen AI adoption rates overall.

However, the gen AI imperative has pushed nearly every company, regardless of industry, to invest in gen AI. Even industries that show lower in-production rates, such as financial services and manufacturing, have strong experimentation and development rates.

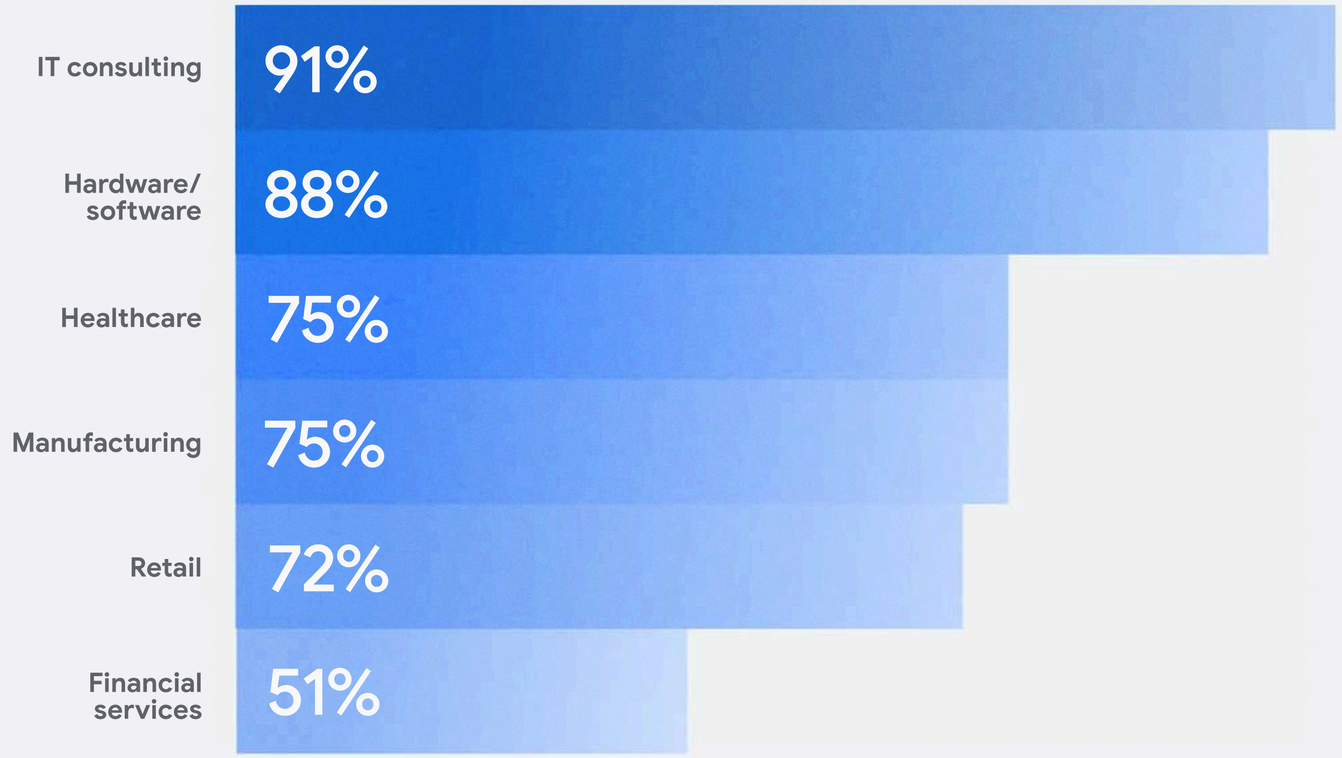
At what stage is your organization in terms of gen AI adoption? (by industry)



Tech leaders in every industry agree that gen AI is essential to their organizations' success.

While a staggering 79% of leaders overall believe gen AI is very or extremely important to their business, this percentage is even higher for tech-related industries like IT consulting (91%) and computer hardware/software (88%).

How important is gen AI to your organization's current and future business operations? (Extremely/very important, by industry)



Why it matters to you

AI is no longer a futuristic concept—it's a core business driver and a fundamental shift in how organizations work. IT leaders have moved past talking about acknowledged potential and have turned to building an infrastructure that can support the growing demands of AI workloads.

With most organizations either actively developing or using it in production, gen AI is enabling organizations to find efficiencies and explore new pathways to success. The infrastructure decisions you make today will determine your organization's ability to compete in an AI-driven future—regardless of your industry or region.

Finding 02

Gen AI ROI relies on both internal and external use cases

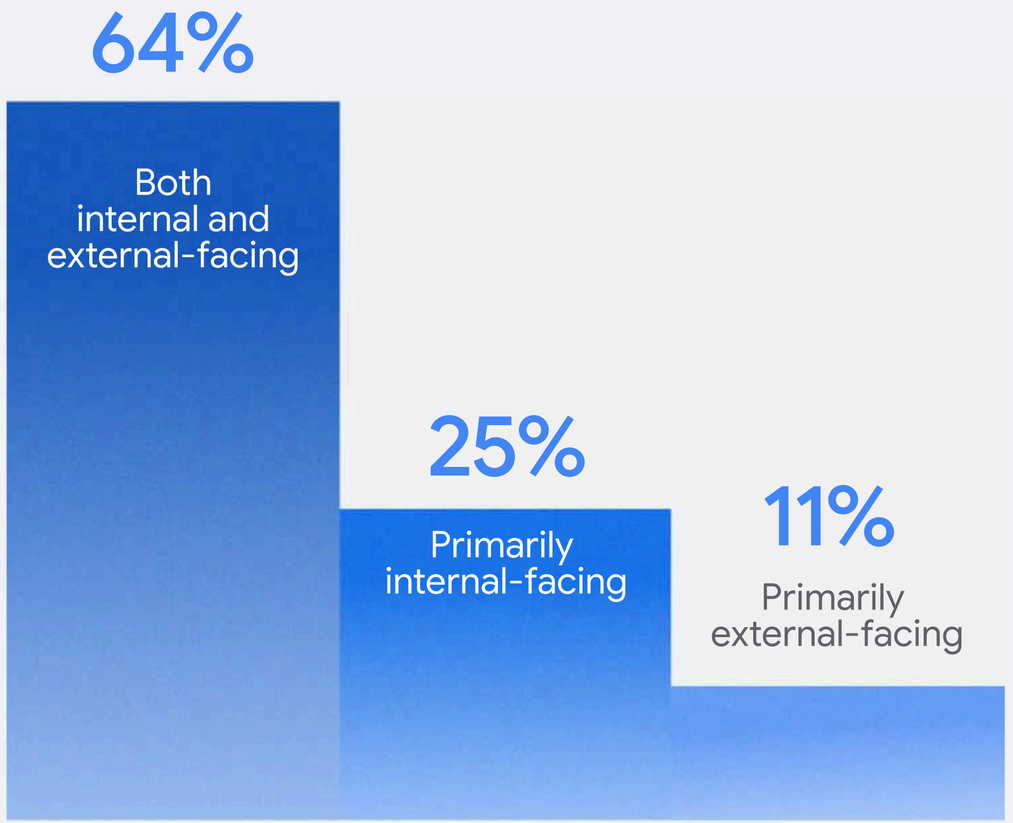
Gen AI delivers results across a number of use cases. While intuitively you may think 'the more the better,' our research shows interesting results about where your gen AI efforts are best placed.

Organizations are using gen AI to improve customer interactions, personalize experiences, and create new value propositions.

Our research shows that many tech leaders (64%) are striking a balance between enhancing customer experience and improving internal operations. Adopting gen AI into external-facing use cases is the difference between providing a personalized, accurate response to a customer versus a generic one.

For example, integrating gen AI with your organization's operational data enables a customer service agent or chatbot to deliver relevant, real-time responses that customers value. While adopting gen AI into internal-facing processes enables organizations to streamline processes, boost productivity, and optimize operational costs.

How would you primarily categorize your organization's current or planned used of gen AI?



Three key use cases are driving gen AI adoption:

01 Data analysis

It’s about more than generating reports. It's about extracting actionable insights from vast datasets, identifying patterns, and enabling data-driven decision-making at scale and speed.

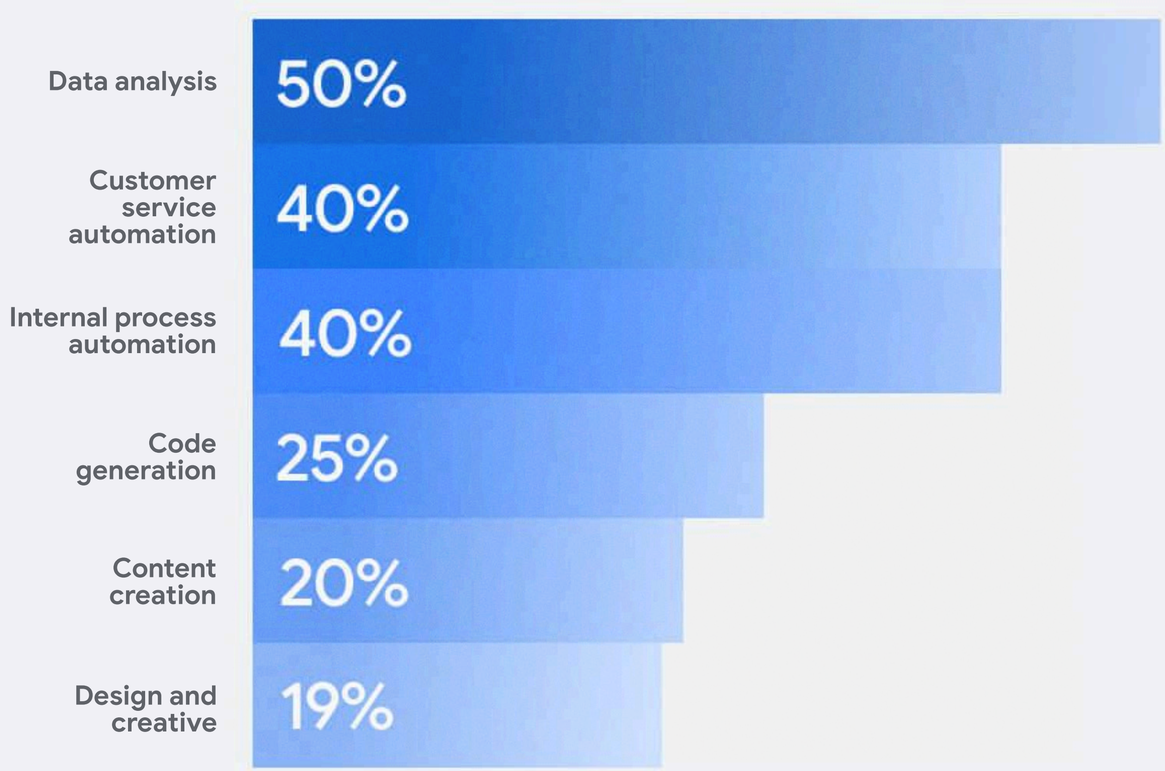
02 Customer service automation

Automating repetitive tasks and streamlining workflows frees up human capital for higher-value activities.

03 Internal process automation

Go beyond basic chatbots. Today’s automation is about creating intelligent systems that can understand nuanced inquiries, provide accurate and personalized responses, and resolve issues efficiently.

Data analysis is the top priority for gen AI implementation, with customer service & internal process automation a distant second



A look across industries

Priorities for gen AI implementation vary significantly across industries.

Data analysis and code generation

60% in IT consulting

69% in hardware/software

These industries also place a higher priority on **code generation** than others (47% and 51%)

Customer service automation

68% in financial services

67% in retail

Internal process automation

55% in manufacturing

54% in financial services

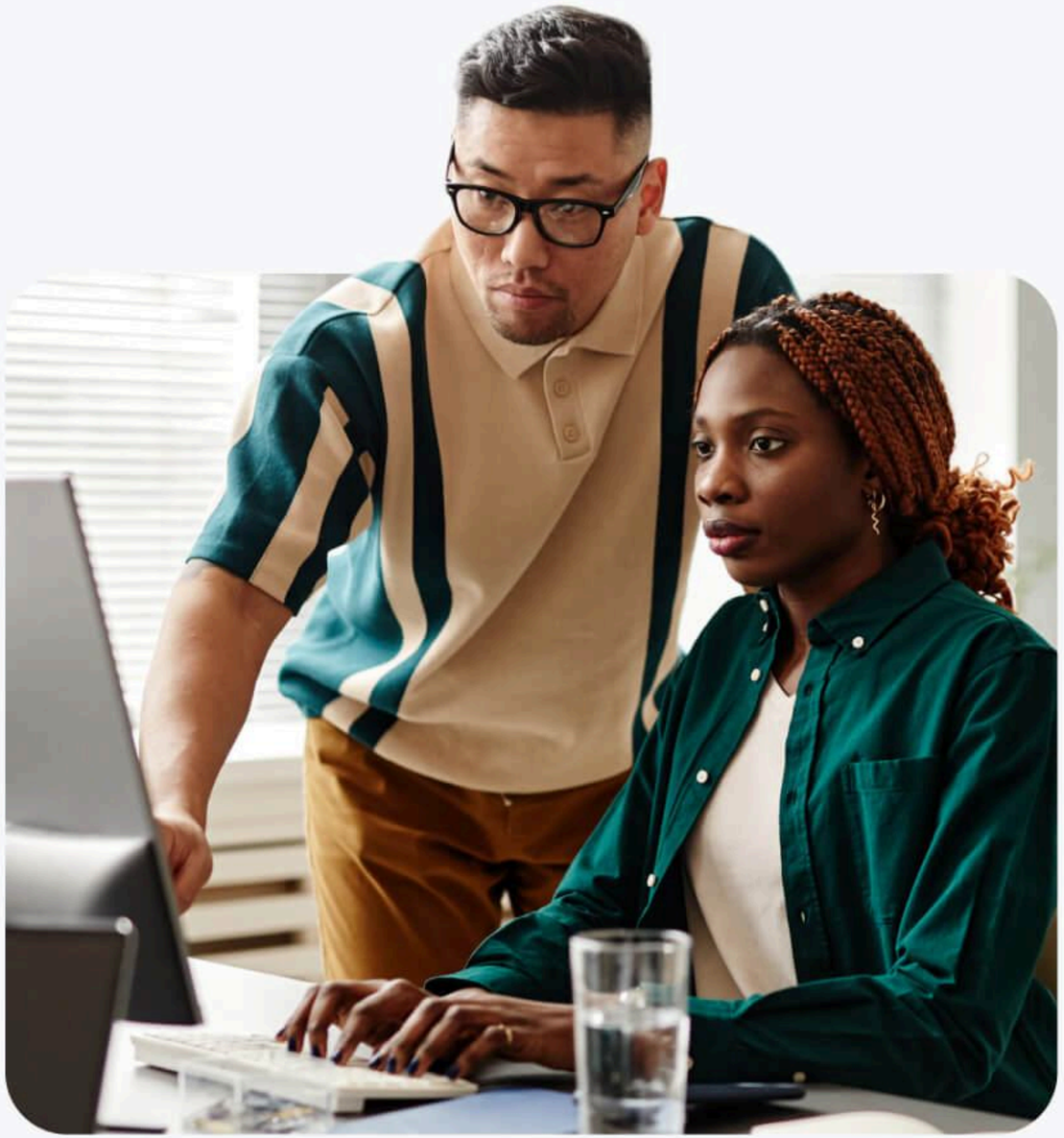
Which of the following gen AI use cases does your organization consider the highest priority for implementation? (by industry)

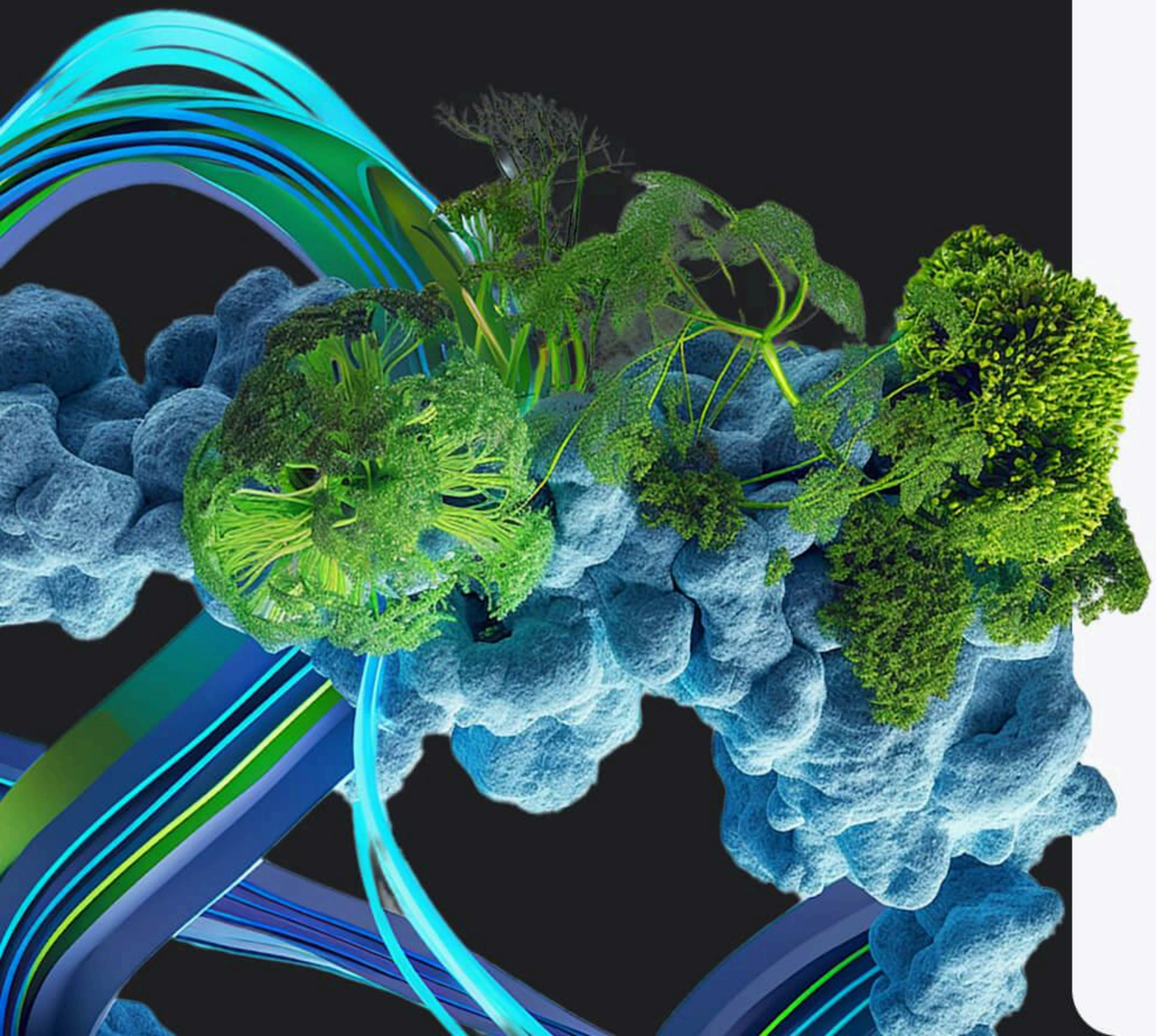
Top implementation priority	IT consulting	Hardware/ software	Financial services	Retail	Healthcare	Manufacturing	Advertising, PR
Data analysis	60%	69% ↑	47%	49%	55%	45%	4% ↓
Customer service automation	22%	25%	68% ↑	67% ↑	47% ↑	30%	8% ↓
Internal process automation	34%	31%	54% ↑	47%	43%	55% ↑	0% ↓
Code generation	47% ↑	51% ↑	7%	9%	19%	18%	0% ↓
Content creation	16%	9%	11%	12%	11%	23%	96% ↑
Design and creative	18%	8%	9%	11%	17%	25%	92% ↑

Why it matters to you

To fully capitalize on gen AI, organizations need to adopt a strategic approach, integrating AI into both internal workflows and customer-facing applications. This requires a flexible and scalable infrastructure that can support diverse AI workloads.

Additionally, while it's valuable to understand the dominant use cases within your industry, cross-industry learning can spark innovation and unlock new value. By exploring how other sectors are leveraging gen AI, organizations can identify novel applications, expand their offerings, and gain a competitive edge.





REGNOLOGY

In the financial services industry, Regnology is leveraging Google Cloud services to develop a regulatory reporting chatbot. This chatbot is designed to expedite the process of obtaining accurate answers to regulatory inquiries, from both internal and external users.

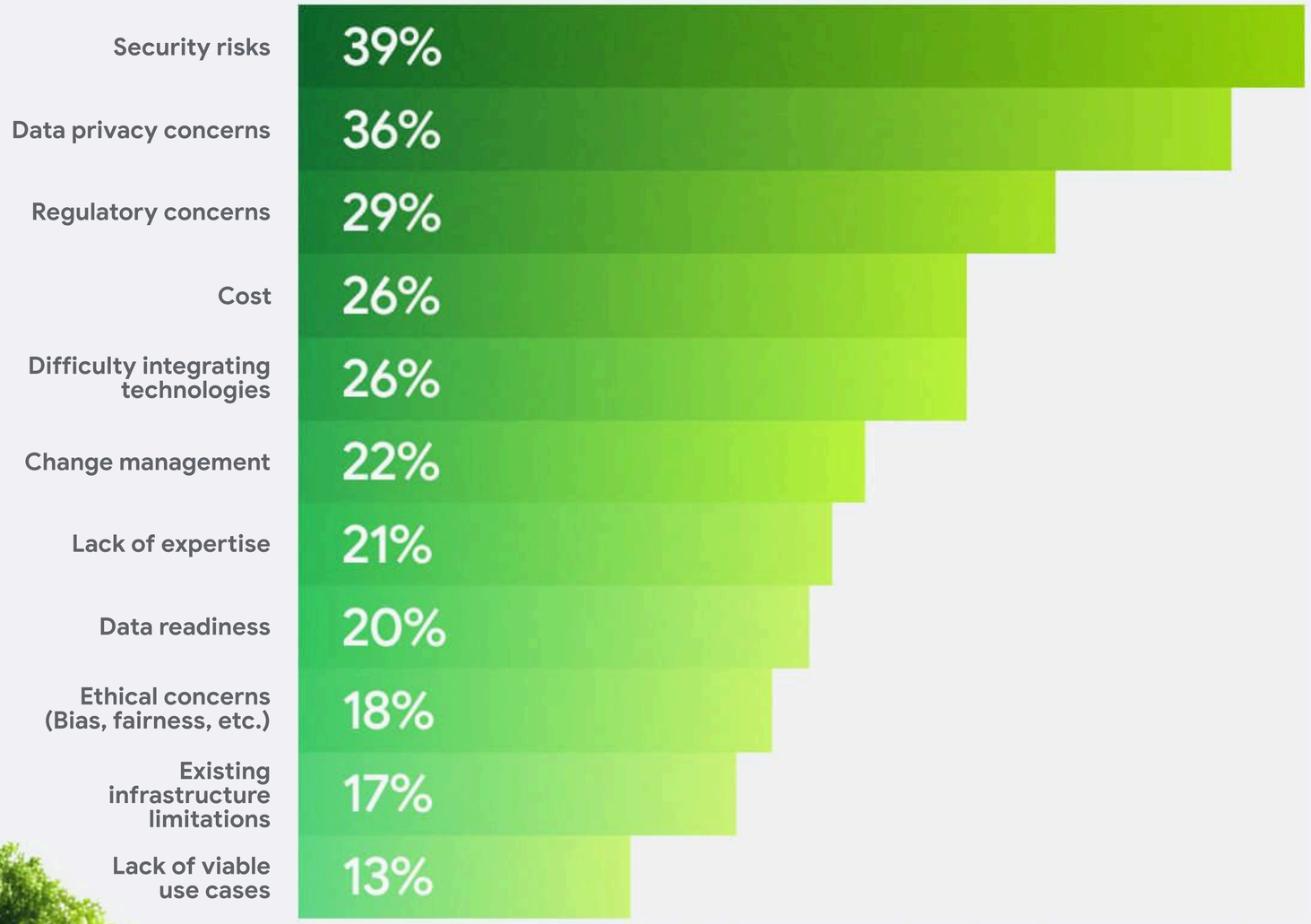
Faced with rapidly growing regulatory demands and increasing data volumes, Regnology helps clients streamline their operations and stay compliant with evolving regulations. As a global leader in regulatory reporting, it struggled with its in-house infrastructure's lack of flexibility. A new chatbot, built on Google Cloud services, expedites the process of obtaining accurate answers to regulatory inquiries, from both internal and external users.

Finding 03

Security and data are key challenges

Safeguarding sensitive information remains a top priority. Organizations want to know how to protect proprietary data when using third-party AI models and ensure compliance with evolving data protection regulations.

Which of the following are the greatest challenges to your organization's gen AI adoption?



Our research underscores the critical security and privacy concerns that technical leaders face.

62% list security risks and/or data privacy as their greatest challenge

The greatest concerns with gen AI adoption are:

01 Security risks

Protecting proprietary data and intellectual property is non-negotiable. The inherent nature of gen AI, involving LLMs and vast datasets, introduces new security challenges that must be addressed proactively.

39%

02 Data privacy concerns

Organizations need to ensure that their gen AI implementations protect sensitive data and comply with regulations like General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), and more.

36%

03 Regulatory concerns

Regulations surrounding gen AI are increasing—the EU AI Act is just one—and it's important for businesses to remain compliant. Some topics under increased regulatory scrutiny include AI explainability, protecting intellectual property, and protecting against misinformation.

29%

70% of orgs adopting gen AI have experienced difficulties with data.

This includes data governance, integrating data into AI models, and having insufficient training data which aligns with key findings from An executive's guide to delivering value from data and AI.

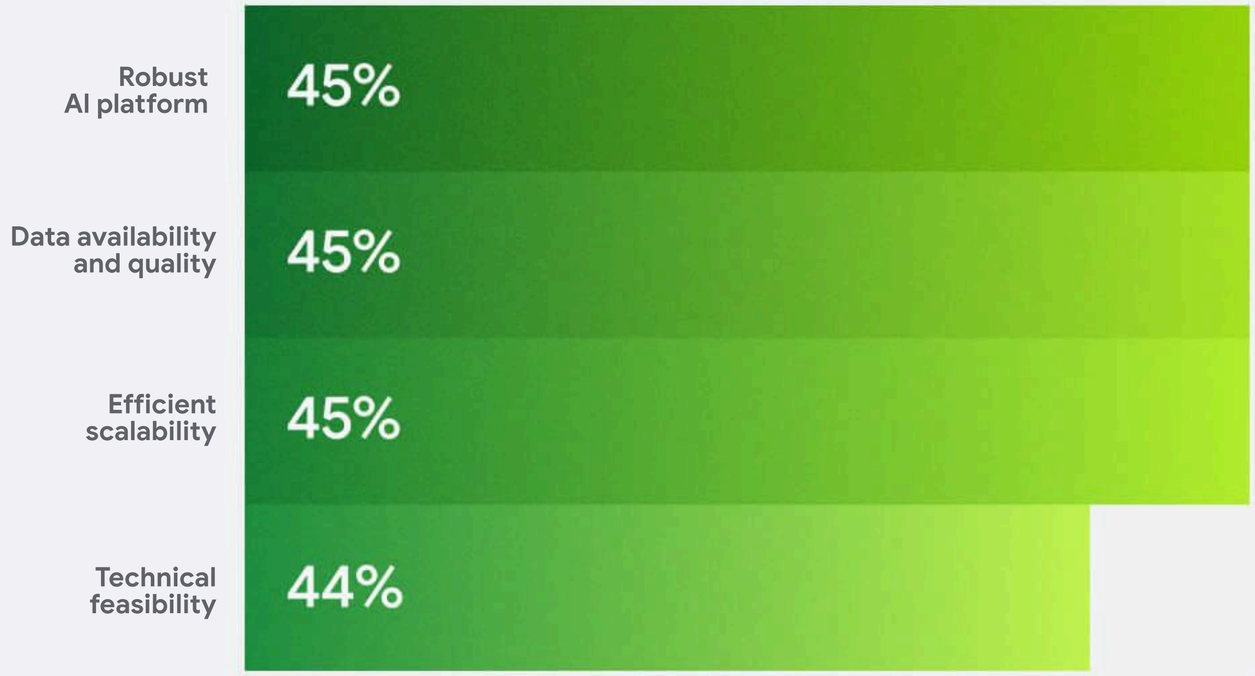
Across the board, organizations are eager to embrace gen AI, but they face challenges in scaling from pilot projects to full production deployments. The key bottleneck? Data. Our research found that data availability and quality, along with the platform storing that data, are the most important factors in gen AI implementation.

Further exacerbating this challenge is the crucial need for data lineage capabilities. Data lineage provides a comprehensive, auditable record of a dataset's origin, transformations, and usage rights—enabling organizations to definitively assert, with legal, ethical, and policy-based justification, their entitlement to utilize specific data for a given AI application. Without verifiable data lineage, organizations face significant risks, including potential legal liabilities, regulatory non-compliance, and reputational damage—effectively hindering the deployment of AI models trained on data with uncertain provenance.

Data availability and quality, a robust AI platform, efficient scalability, and technical feasibility lead in overall importance for new gen AI tech.

Organizations are concerned about keeping their customer and proprietary data private, secure, and compliant. We already know that to ensure quality data governance, organizations need to unify their data—but doing so can be a heavy infrastructure lift. And so it's perhaps no surprise that data quality, a robust platform, and efficient scalability are the most important factors on the minds of tech leaders when it comes to gen AI implementation. Data and its supporting infrastructure still reigns supreme.

What are your top 5 most important factors when choosing gen AI tech?

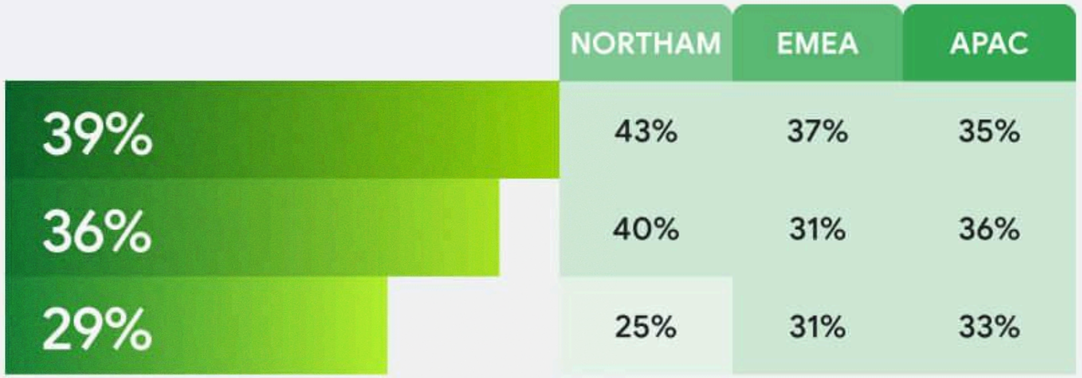


A look across regions

Security and data privacy are the greatest challenges to gen AI adoption, and this finding is consistent across regions.

The regulation landscape varies across global regions, which is of particular concern for companies operating multi-nationally.

What are your top gen AI adoption challenges?



North America

State-level laws are becoming more comprehensive, such as the CCPA, which attempts to secure consumers' rights to access, view, and delete data.

Europe

GDPR maintains strict requirements about how data is stored, collected and deleted. For businesses using cloud resources, in-country cloud centers are important for remaining compliant. The fines for non-compliance are up to €20 million or 4% of global annual turnover.

Asia-Pacific

The landscape varies, with some countries requiring keeping data within their borders, while others have a more fragmented approach.

Why it matters to you

Building a secure and compliant AI infrastructure is non-negotiable. IT leaders must prioritize data security and privacy in their AI strategy to mitigate risks and maintain trust.

No matter your industry, data is often fragmented across different platforms and formats, making it challenging to implement unified security and governance policies. To fully capitalize on AI, it's imperative to create a single source of truth with a unified platform across all forms of data, including open formats.

Additionally, a highly scalable architecture is crucial, enabling the unification of transactional and analytical systems without impacting performance.

This allows for data access between these systems, even with demanding transactional workloads.

And as more data and applications move out of on-premises data centers and away from traditional security mechanisms and infrastructure, data security approaches must be adapted to the cloud.

Ready to learn more?

- Accelerating Secure, Private, and Compliant AI Transformation
- Unlocking gen AI's full potential with operational databases
- Building a Secure Data Platform with Google Cloud

moii.ai

Moii.AI deploys vision AI agents on CCTV to improve safety

With more than a billion CCTV cameras in the world, Moii.AI automates the review and analysis process with AI—turning traditional camera systems into autonomous agents to create safer and more productive workplaces.

> **“Because of BigQuery’s intuitive UI, even engineers without a lot of cloud experience can go into the tool in their first week and start to work on data projects. With another solution, we would need a team twice this size to maintain the infrastructure we’ve built.”**

Lakshman Balasubramanian
Head of AI and Co-founder, Moii.AI

Finding 04

Cost efficiency is a critical consideration

AI demands significant computational resources, and the overall costs for implementing and maintaining AI can vary widely depending on factors such as scale, model complexity, data requirements, infrastructure maintenance, and existing talent.

While the potential of gen AI may be limitless, budgets are not.

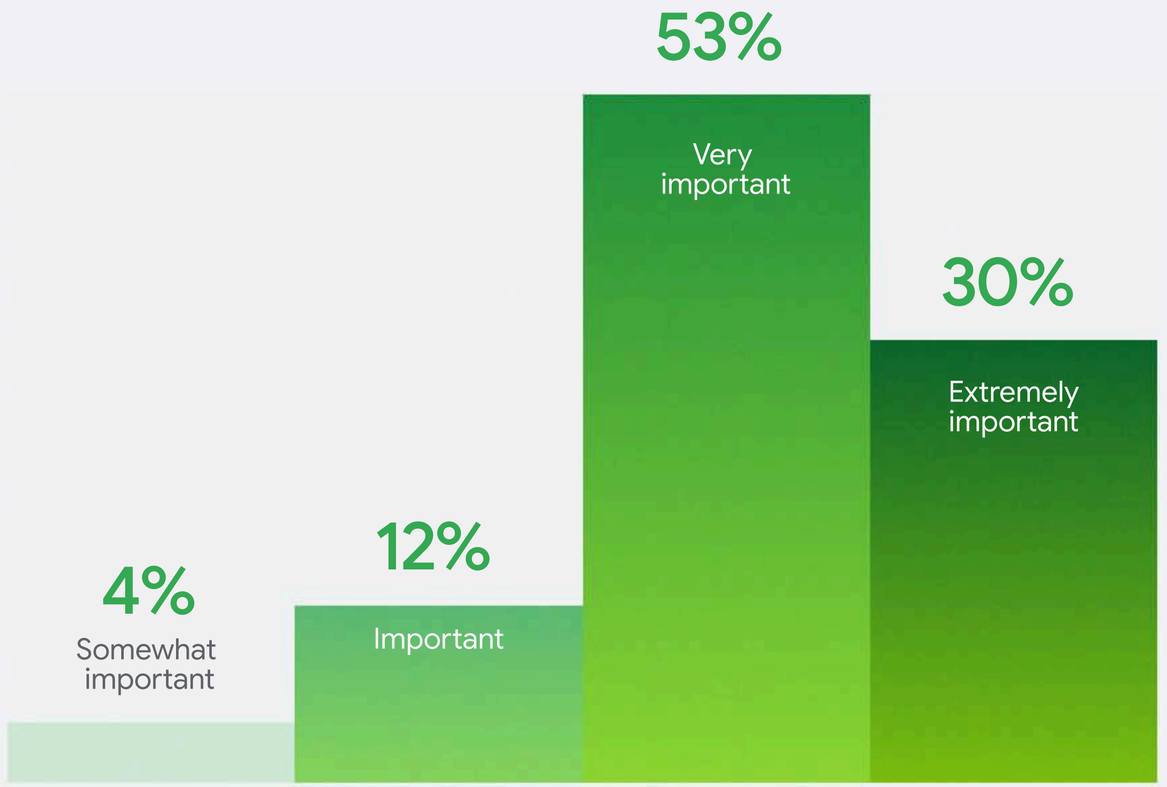
83% of tech leaders say cost is a key factor when evaluating solutions

Harnessing this powerful new technology can be a minefield of unexpected expenditures.

The right cloud provider can offer cost-efficient solutions for training, fine-tuning, and inference of AI workloads—including options for spot instances, reserved instances, and pay-as-you-go.

As we saw in finding 3, the cost of gen AI solutions is also a significant challenge in gen AI adoption, ranking 4th overall.

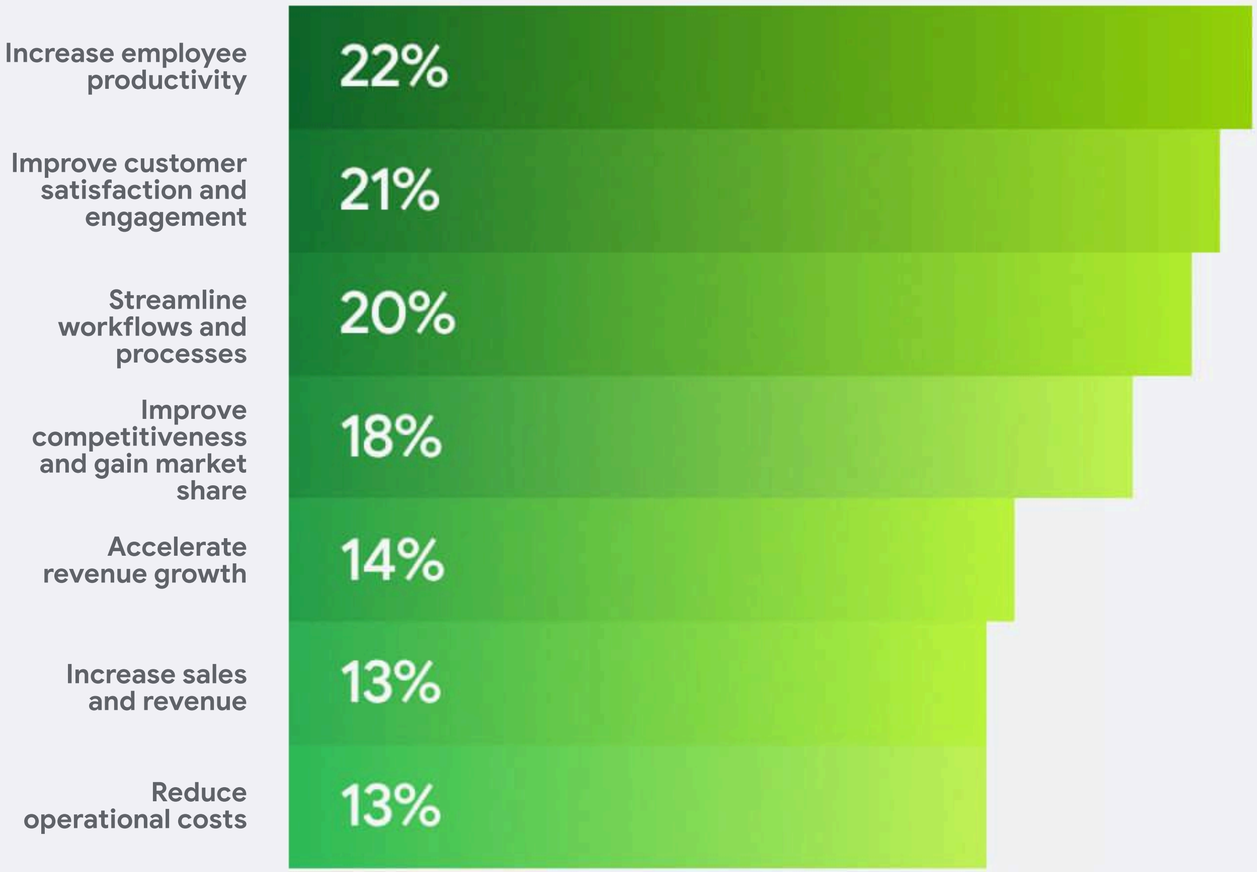
How important is cost-efficiency in your decision-making process for gen AI infrastructure solutions?



While the cost associated with gen AI solutions is a leading consideration, the benefits of gen AI also positively impact an organization's bottom line.

From improving productivity, increasing sales, and reducing operational costs, a number of leading gen AI use cases are ones that directly deliver cost efficiencies.

Where do you expect the largest ROI from gen AI?



Why it matters to you

Gen AI streamlines processes and automates tasks, creating efficiencies that help organizations reduce costs. Yet, despite the efficiencies up for grabs, quantifying gen AI's business value can be challenging. And that’s because alongside organization-wide cost savings, gen AI can bring ever increasing technology expenditures, including computational costs, model development and management, and data management. That’s why cost optimization is a crucial consideration.

Improve your AI cost efficiency

Effectively managing the costs of gen AI requires a strategic approach that goes beyond simply choosing the lowest-priced services. Here are three critical areas to focus on for achieving optimal cost-efficiency in your AI infrastructure.

By carefully considering these factors and selecting a cloud provider with a commitment to cost optimization, organizations can unlock the full potential of gen AI without incurring unsustainable expenses.

Right-size resources with granular control

Avoid over-provisioning by leveraging cloud platforms that offer fine-grained control over compute resources. Look for capabilities like auto-scaling, which dynamically adjusts resources based on real-time workload demands, and support for heterogeneous compute instances (e.g., CPUs, GPUs, TPUs) to match specific model requirements. This minimizes wasted resources and ensures you only pay for what you actually use.

Leverage AI-optimized hardware and software stacks

Choose infrastructure providers with a proven track record of optimizing AI performance. This includes using specialized hardware accelerators (like TPUs or GPUs designed for AI) and leveraging highly optimized software frameworks (e.g., JAX, TensorFlow, PyTorch) that maximize hardware utilization. These optimizations can significantly reduce training and inference times, leading to substantial cost savings.

Implement intelligent resource management and scheduling

Explore advanced resource management tools that automate the allocation and scheduling of AI workloads. Features like dynamic workload scheduling, preemptible instances (where appropriate for fault-tolerant workloads), and intelligent caching mechanisms can significantly reduce idle resource time and optimize overall infrastructure utilization, leading to lower operational costs.

How gen AI-powered FinOps can help you drive cost-effective returns on technology investments

FinOps brings discipline to the cloud by aligning technology, finance, and business teams with an overarching operational framework focused on driving cost-effective returns on technology investments, with organizations reducing their cloud spend by as much as 30%.

Assess your gen AI adoption readiness with the Cloud FinOps for Generative AI framework.

nuro

Nuro, an autonomous driving technology company, is using Google Cloud to develop the Nuro Driver, its cutting-edge self-driving system, while considerably improving the cost-efficiency of training. Using Google Cloud TPUs, Nuro can now process petabytes of real-world driving data and train AI models for producing city-scale maps and detecting obstacles on the road twice as fast, without incurring incremental costs.

Finding 05

A robust AI platform is a leading criteria when evaluating infrastructure

Respondents want it all—performance, scale, efficiency—and they want it all in one platform that can train, deploy, and manage their AI.

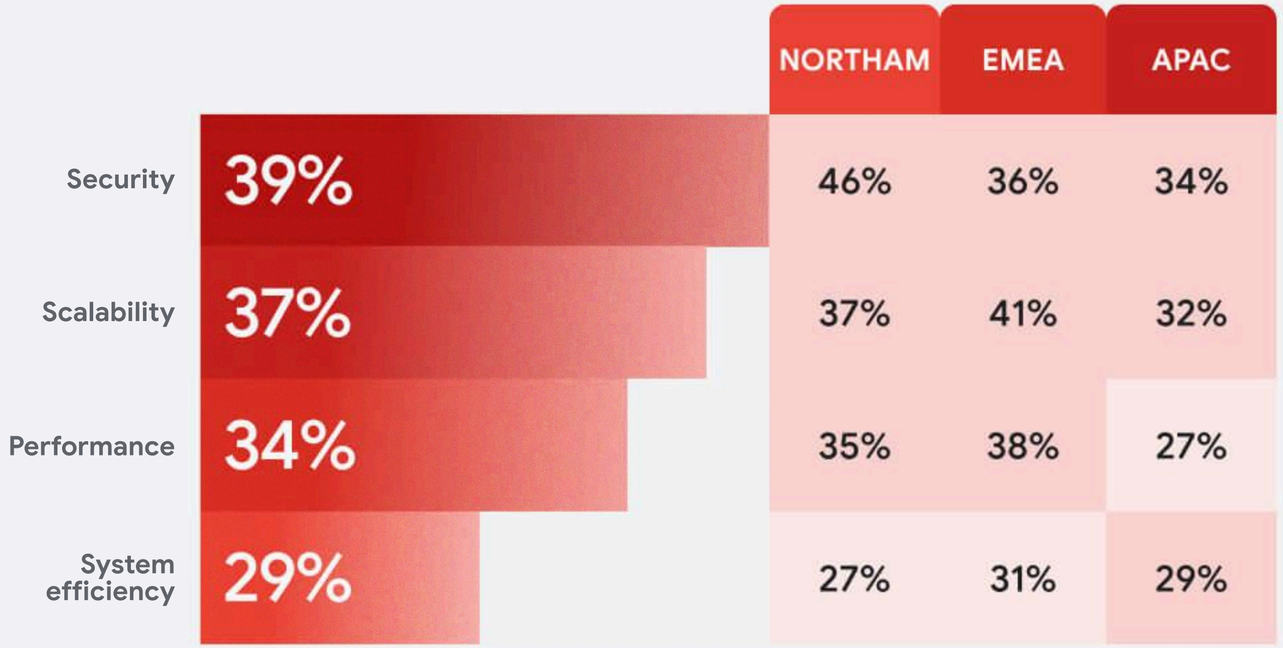
45% of those surveyed want a robust AI platform

It's no wonder that with the breathtaking rate of change in the models, applications, and architecture of gen AI, organizations are prioritizing platforms that can keep pace (see finding 3 for survey results).

Important infrastructure factors

Very few organizations are training large-scale foundation models. But scale and performance are top factors in every geography, whether you're serving models quickly to thousands of internal end users or millions of customers. As AI platforms must handle rapid agent-to-agent communication and an even higher volume of inference requests, the demands on both performance and scalability intensify. A robust platform must support these types of workloads, regardless of whether the architecture is containerized or not, with constant innovation in compute, storage, and networking scaling, resource sharing, access, and routing.

What are the most important factors when evaluating a gen AI infrastructure system?



Why it matters to you

A robust AI platform can support diverse users, from developers to scientists to IT operators. It can scale and optimize evolving architectures and frameworks. And it can handle the full lifecycle of AI, from development to deployment to management.

Innovative infrastructure

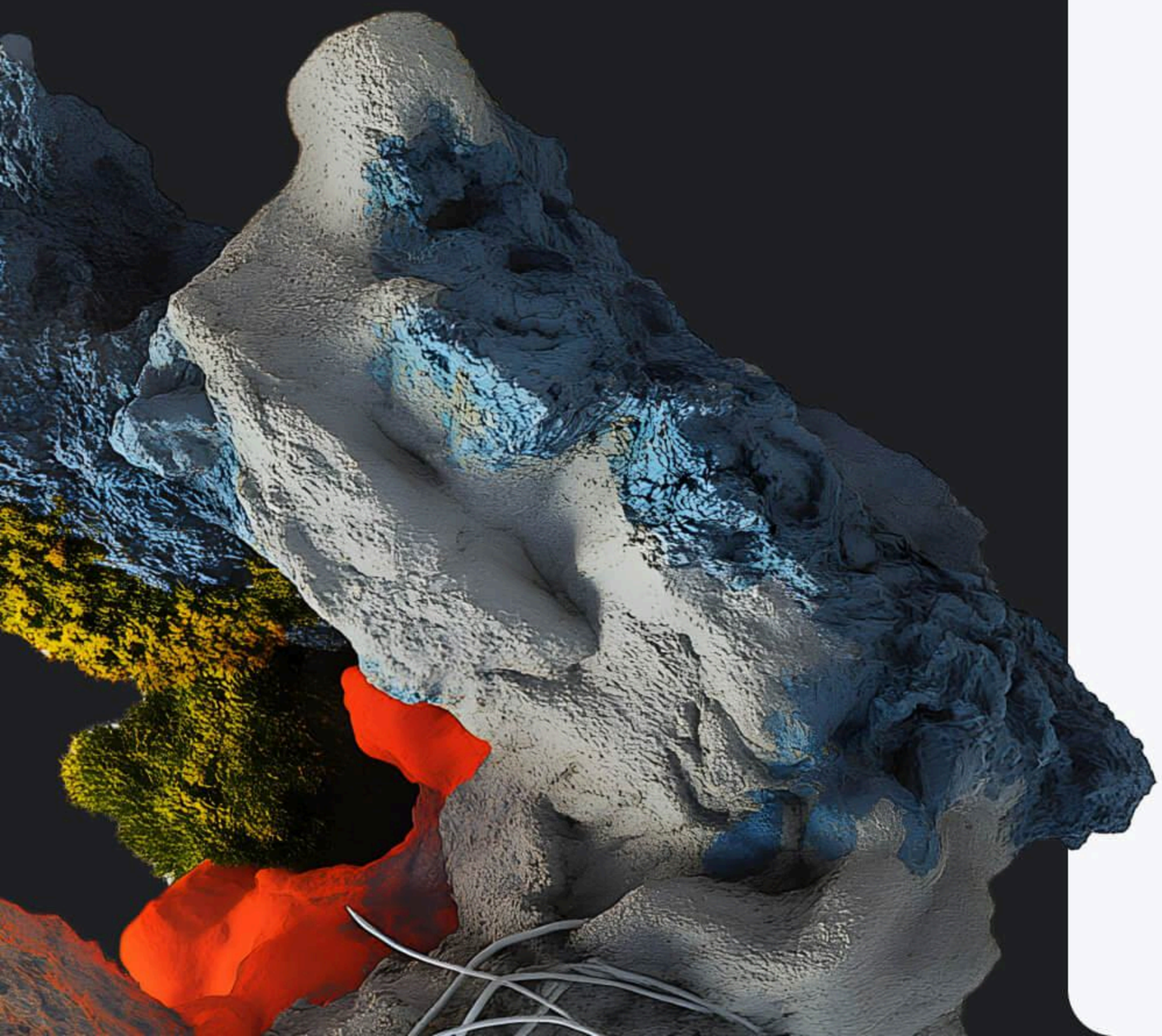
AI has popularized GPUs. However, your core infrastructure should also have the flexibility to provide the best price-performance for every use case, including CPUs and custom silicon. Data is critical for AI, the platform must offer multiple storage mechanisms to ingest, store, and deliver data. For example, parallel file systems are the right storage for certain AI training, but object storage and block storage can be far more efficient when the training data is static. Networking must be considered for each host, across hosts in a cluster, and across cloud, cross-cloud, and private networks.

Dynamic and integrated software

As gen AI implementations evolve, the AI platform's software must evolve as well. For example, variable length context windows require smarter network load balancing to properly route the requests. Multi-step reasoning may require multiple compute hosts to dynamically handle the processing. The sheer infrastructure cost of large AI jobs requires the software to minimize inevitable interruptions with expansive observability, more sophisticated checkpointing, and even for AI clusters to operate in degraded mode.

Serves multiple roles

Some organizations want full access to the AI infrastructure because they want to implement new techniques, frameworks, and models to maximize the hardware resources. Others prefer to take existing models and refine them for their specific use cases—infrastructure control is secondary. Yet another business simply wants to access pre-configured AI models via API, to integrate into an application. All of these organizations benefit when these different modalities operate from the same scalable, performant, and reliable infrastructure.



Moloco

In a congested digital landscape, Moloco offers AI-powered advertising solutions, using first-party data to help companies target and acquire high-value users based on real-time consumer behavior—ultimately, delivering higher conversion rates and return on investment.

Moloco leverages predictions from a dozen deep neural networks with a platform that ingests 10 petabytes of data per day at a peak rate of 10.5 million queries per second. They rely on Google Kubernetes Engine to handle this massive data load and optimize serving efficiency while remaining cost effective.

Finding 06

Edge computing is extending AI's reach

Gen AI is relied upon in an increasing number of contexts, and many occur outside of traditional centralized data centers.

Edge computing includes IoT devices (like smart fridges), autonomous vehicles, and mobile devices. It distributes processing, reduces latency, and requires efficient computational resources. This computational distribution can also improve network performance and enhance adherence to data privacy regulations (e.g., GDPR, CCPA) by enabling local data processing and minimizing the transmission of sensitive information.

Nearly every industry now relies on some degree of IoT or mobile devices to support real-time applications and keep operations running. In healthcare, manufacturing, and retail, edge computing uses and processes sensor information to fuel analyses, maintain quality control, and even operate machinery autonomously.

Specific use cases:

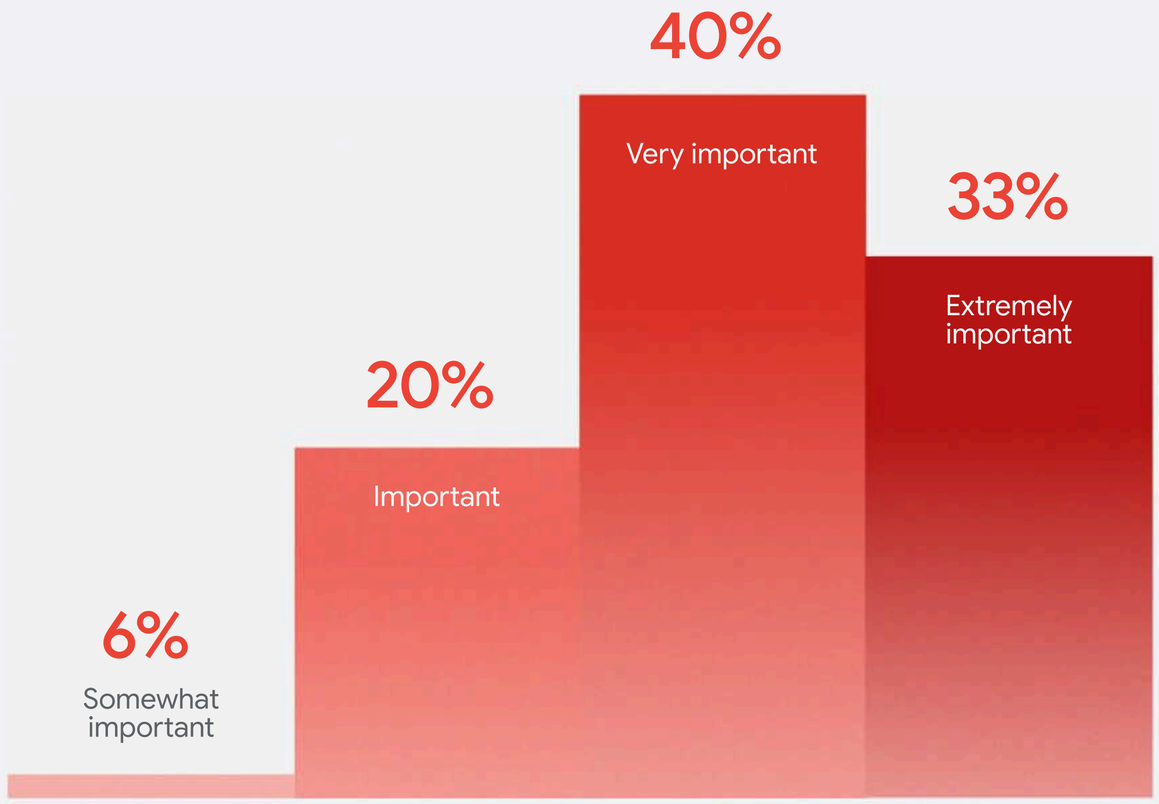
Local processing of patient vital signs from medical devices for immediate anomaly detection and staff notification

Real-time verification of personal protective equipment (PPE) compliance among healthcare personnel by deploying computer vision models at the edge

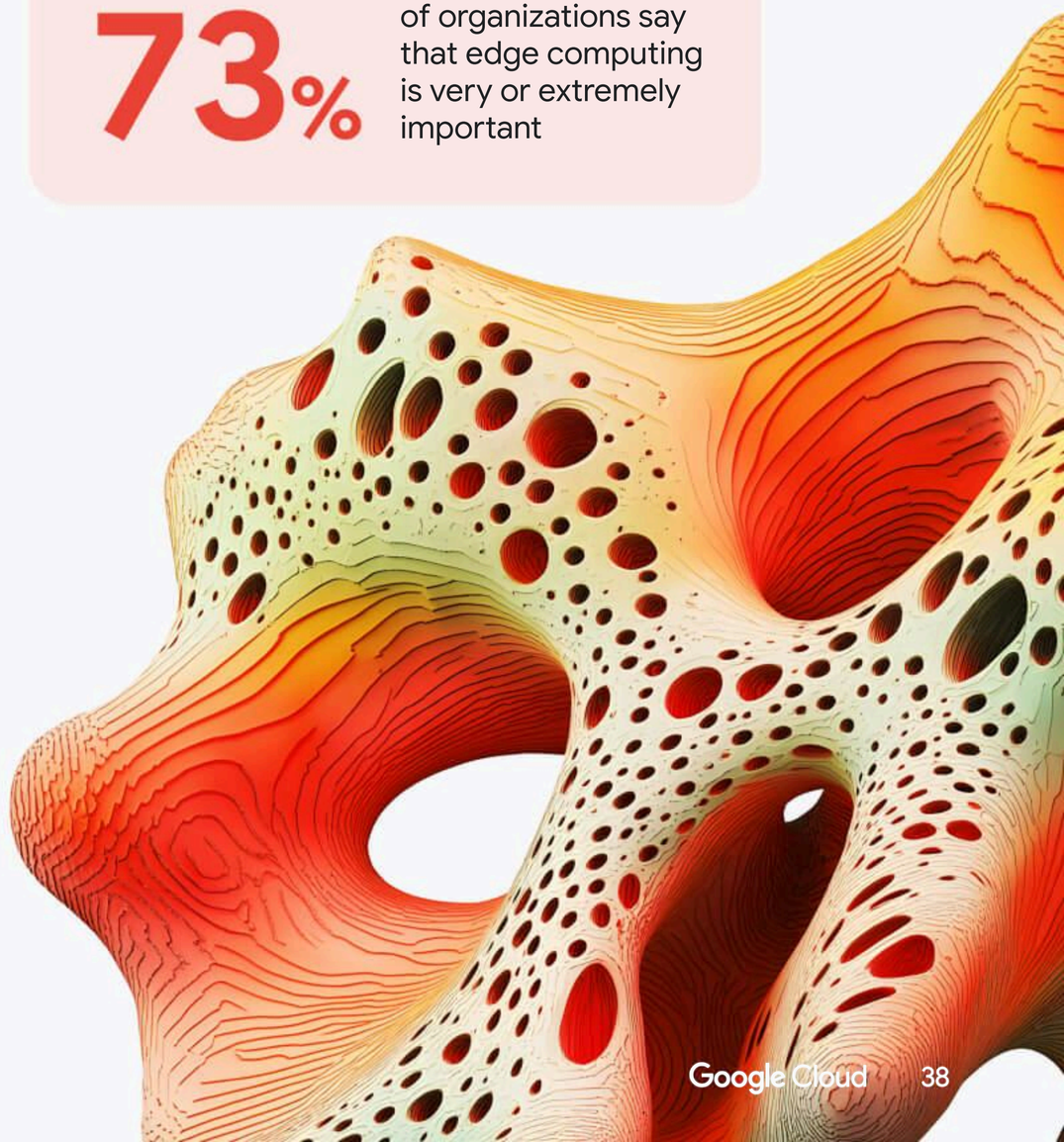
Predictive maintenance of industrial equipment based on sensor data analysis

Real-time inventory management using edge-based video analytics

How important is deploying gen AI models at the edge (e.g., on IoT devices, mobile devices) for your organization?



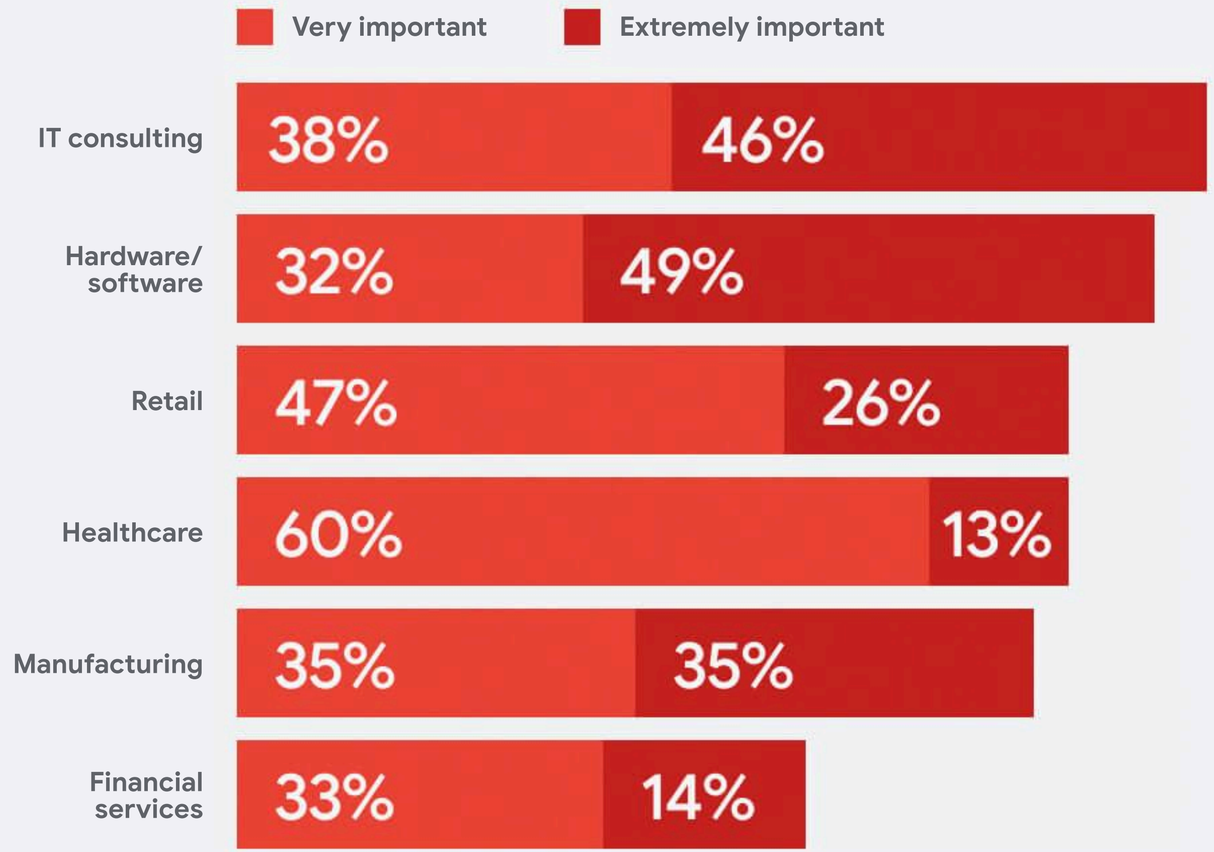
73% of organizations say that edge computing is very or extremely important



A look across industries and regions

As you might expect, deploying gen AI at the edge is important for industries like manufacturing (70%) and healthcare (74%), but more so for companies in the computer hardware or software industry (81%) and especially for computer/IT consulting firms (85%).

How important is deploying Generative AI models at the edge (e.g., on IoT devices, mobile devices) for your organization? (by industry)



Why it matters to you

Despite its rising importance, modernizing and leveraging AI at the edge, maintaining control over mission-critical data, and managing the complexity of multiple edge deployments present significant challenges.

It's no surprise then that there is a growing demand for cloud-based solutions that enable the deployment of gen AI models at the edge for real-time applications.

And as customer expectations for real-time interactions continue to rise, so too will the necessity of edge computing and the infrastructure that makes quick gen AI processing a reliable reality.

Challenges with multiple edge deployments:

Resource and connectivity constraints

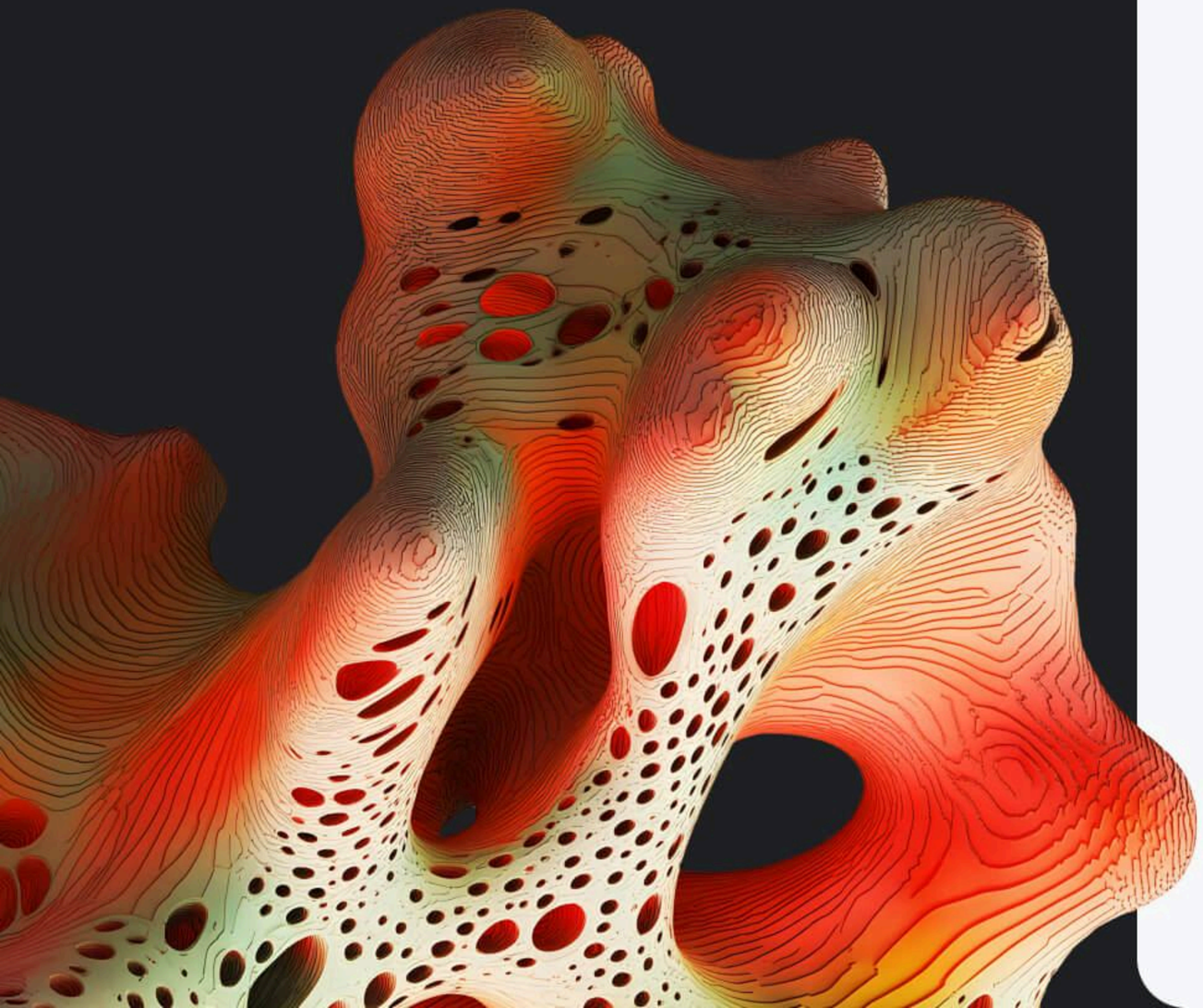
Edge devices have limited compute power and often face intermittent connectivity, making it hard to deploy and manage complex AI models effectively. This requires specialized optimization and robust offline capabilities.

Data security and governance

Protecting sensitive data at the edge, ensuring data quality and compliance, and maintaining control over data sovereignty across distributed locations presents significant security and governance hurdles.

Deployment and management complexity

Scaling AI deployments to numerous, geographically dispersed edge devices, along with monitoring, maintenance, version control, and orchestration, introduces significant complexity that demands automation and sophisticated management tools.



Global retailers like McDonald's plan to deploy Google Distributed Cloud to deliver AI to thousands of restaurants, enabling new customer platforms and bringing information storage and high-powered computing into individual restaurants. With these edge computing capabilities and distributed AI, McDonald's will gain new insights into equipment performance, enact solutions that reduce business disruptions, and simplify operations, allowing restaurant teams to focus on delivering exceptional customer experiences.

Finding 07

Hybrid cloud offers flexibility for gen AI deployments

A hybrid approach involves combining on-premises infrastructure with public cloud resources.

Our research clearly indicates a strong preference for a hybrid cloud approach (74%).

This approach is preferred over on-premises and single public cloud approaches, alone.

The rationale is clear:

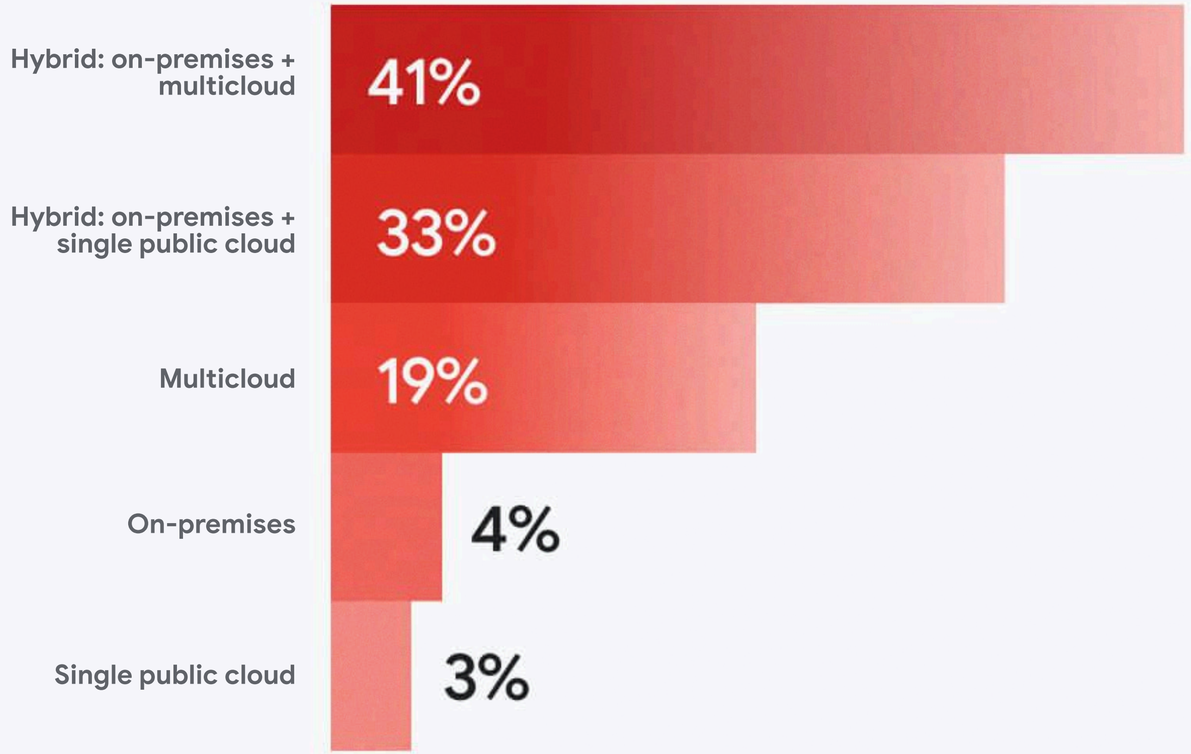
Increased flexibility

A hybrid approach allows organizations to choose the optimal environment for each workload, balancing performance, cost, and security considerations.

Improved data residency and compliance

For certain workloads, regulatory requirements may necessitate keeping data on-premises. A hybrid approach allows for this while still leveraging the benefits of the public cloud.

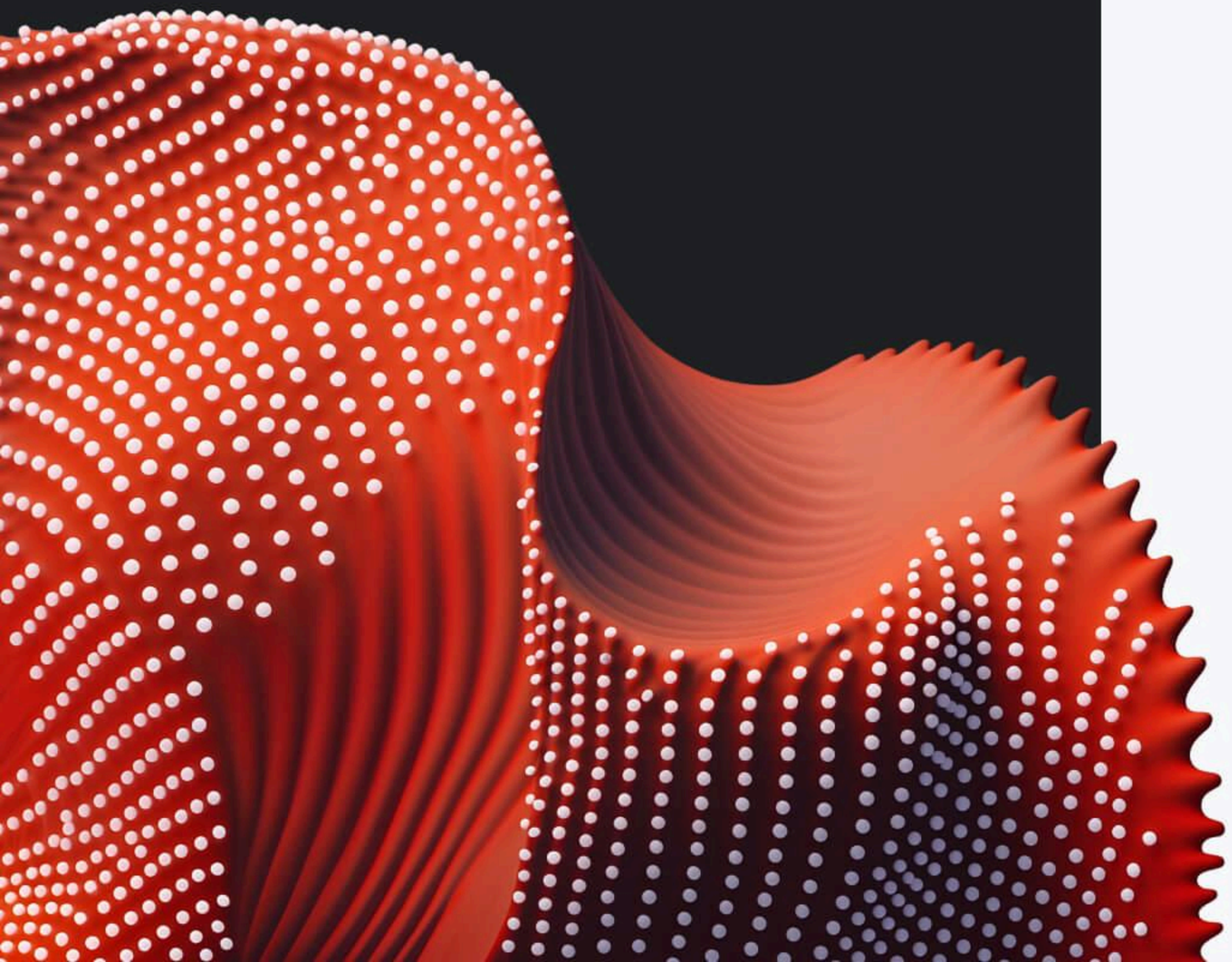
Which cloud infrastructure approach does your organization primarily use for gen AI workloads?



Why it matters to you

The overwhelming preference for hybrid cloud means that a tech leader must understand how to architect, implement, and manage such an environment. Ignoring this trend risks falling behind competitors who are already leveraging the flexibility and control that hybrid offers. They need to evaluate their existing infrastructure and cloud partnerships to see if they can support a hybrid approach.

Flexibility in choosing the right environment for each workload is key. This requires thinking strategically about workload placement, considering factors like performance, cost, and security requirements on a per-application basis.



TOYOTA

Toyota utilized Google Cloud's AI Hypercomputer to build a hybrid cloud architecture to power its innovative AI Platform, which empowers factory workers to develop and deploy AI models across key use cases. Toyota achieved a 20% reduction in learning model creation time, improving the user experience and boosting adoption. With more factory floor employees able to create AI models, Toyota has automated more manual and labor-intensive tasks, saving over 10,000 hours per year through manufacturing efficiency and process optimization.

Finding 08

Most organizations rely on gen AI solutions from cloud providers

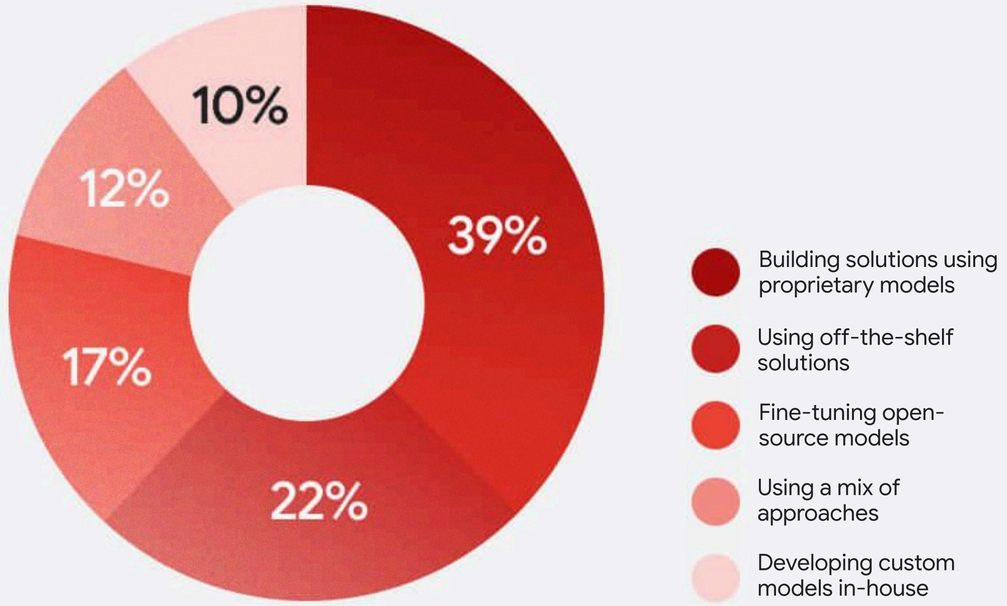
Building and running AI workloads is complex.

Organizations need managed services to streamline operations and fully leverage the cloud.

When building gen AI solutions, the most common approach is leveraging proprietary models like Gemini, Claude, and AI21—prioritizing speed, ease of use, cost-effectiveness, and scalability. Building custom in-house models is the least common approach as it requires considerably more time and resources, suggesting that for many organizations, proprietary models are sufficiently performant for their needs.

Which approach does your organization prefer for managing gen AI infrastructure?

Which of the following best describes your organization's current approach to utilizing gen AI?



Cloud providers play a pivotal role in gen AI implementation.

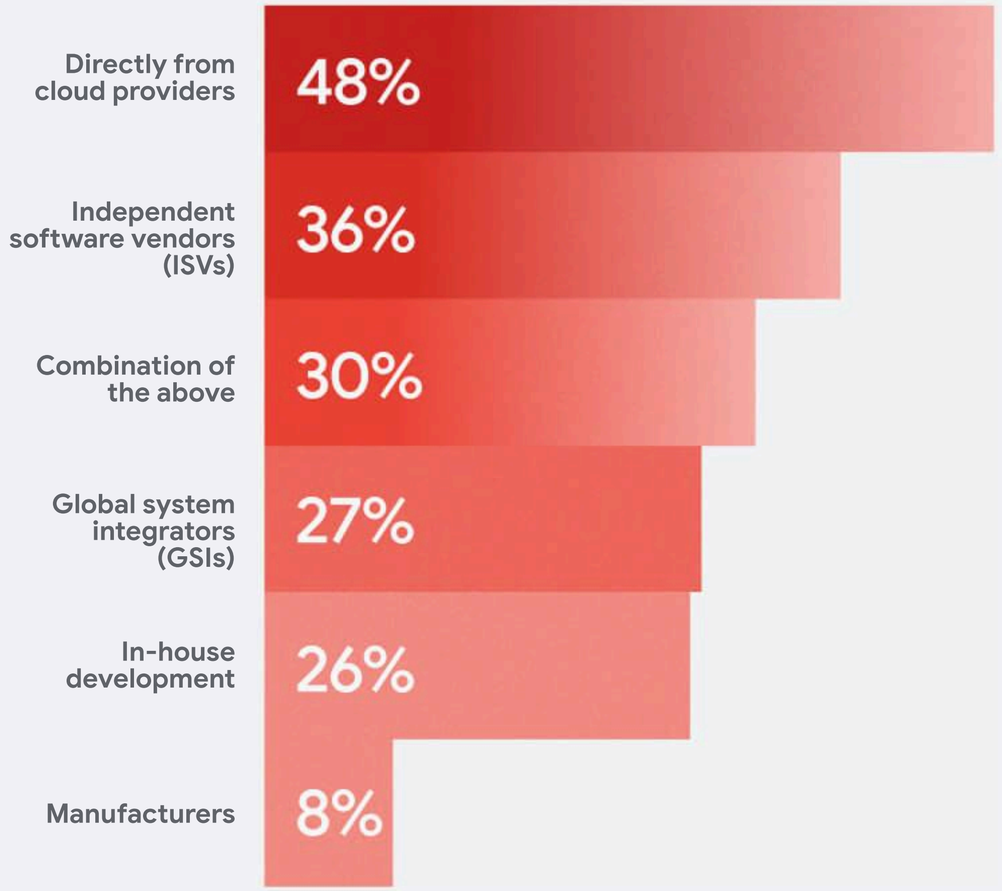
Architecting for long-term success requires a delicate balancing act—utilizing new gen AI functionality for productivity, optimizing core applications for customers, and modernizing existing infrastructure to keep pace. We are rapidly approaching an inflection point where traditional cloud computing and infrastructure will no longer be capable of keeping up with current capacity and performance demands.

Given the complexity and resource intensity of building and managing AI-ready infrastructure, organizations are increasingly turning to cloud service providers (48%) and, to a lesser extent, independent software vendors (ISVs) (36%). This shift is largely about forming strategic partnerships to leverage specialized expertise and resources.

This is driving considerable increase in enterprise infrastructure market size. As outlined in the 2025 Google Cloud AI Business Trends report, adoption in enterprise infrastructure is expected to increase by over 30% by 2026.[1] This rapid adoption rate has created global demand for data center space capable of handling the high computational power and power density required for AI workloads. Demand for AI-ready data center capacity is expected to rise at an average rate of 33% per year through 2030,[2] and spending on data centers is expected to double in the next five years.[3]

[1] Credence Research (2024). All-In-One Infrastructure Market Size, Share and Forecast 2032.
[2] McKinsey & Company (2024). AI power: Expanding data center capacity to meet growing demand.
[3] National Bureau of Economic Research (2024). The Rapid Adoption of Generative AI.

Through which channels does your organization primarily acquire and implement gen AI solutions?



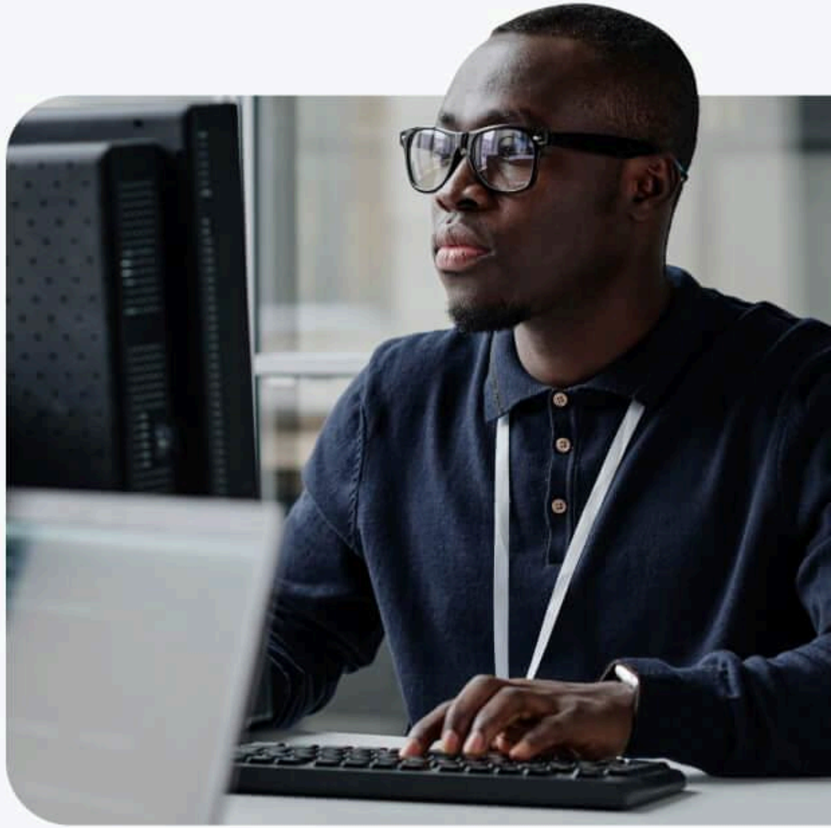
Why it matters to you

Building solid AI infrastructure is foundational for success. And so future success will hinge on an organization having infrastructure that can provide the scale, reliability, computational performance, efficiency, security, ease of use, and cost-effectiveness to support the combination of large-scale AI workloads, enterprise application estates, and new application development.

Many businesses and governments lack the skills, time, and resources to develop these technologies and tools independently.

It's no surprise, then, that they are relying on managed service providers to manage, scale, and secure their infrastructure. Finding the best AI solution for your business helps you keep ahead of the curve.

Google Cloud's global partner ecosystem has been specially trained and certified to deliver cloud solutions that address industry-specific needs.



Cloud providers are strategic partners in the AI era

This shift towards cloud providers isn't just a trend. It's a strategic imperative for leaders who are grappling with the complexities of AI development, deployment, and management—while navigating budget constraints and competitive pressures. The research confirms that gen AI is a present-day reality for many organizations. Increasingly, businesses are seeking ready-to-go solutions that incorporate AI and data technologies to simplify decision-making when operating, scaling, optimizing, and securing infrastructure.

And as gen AI use cases proliferate, organizations will require a wide range of models to satisfy these different needs. It can be overwhelming to adopt the best models for your use case without drowning in complexity.

Leading organizations are leveraging the expertise of cloud providers to develop a comprehensive solution. Addressing key considerations such as:

Security

Implementing robust security measures to protect sensitive data and ensure compliance

Scalability

Designing an infrastructure that can scale to meet the growing demands of AI workloads

Data quality

Ensuring access to high-quality data to train and deploy effective AI models

Cost

Capping expenditures associated with the significant computational resources and maintenance AI requires

Flexibility

Choosing the right environment for each type of workload on a per-application basis

By carefully selecting the right partners, tools and frameworks, organizations can position themselves for success in the age of AI.

The Rearchitecting your infrastructure for generative AI guide helps technical leaders architect robust, scalable, and cost-effective gen AI systems.

Google Cloud provides a comprehensive platform for gen AI success

The imperative for technology leaders is clear: build an AI-ready infrastructure that can meet your business needs while supporting the demands of this transformative technology.

AI Hypercomputer is a fully integrated supercomputing architecture designed explicitly for AI workloads.

Our research shows that success isn't about point solutions, but deploying a robust and integrated AI platform.

Integrating AI into your business should be your focus. Google Cloud can do the heavy lifting with AI infrastructure that offers the performance, scale, and efficiency already powering the most advanced AI products in the world. We're tuned to the market, already running cutting edge reasoning models, serving AI at the edge, and helping customers securely deploy AI in their own data centers. From enterprise data scientists using Vertex AI, to developers incorporating AI with CloudRun and Gemini endpoints, to AI companies building foundation models on AI Hypercomputer, our full stack is ready for today and the rapidly evolving future. Google Cloud isn't just a cloud provider, we're your strategic partner for AI transformation. Our commitment to open-source principles, our continuous innovation driven by Google DeepMind and serving AI technology to billions of users daily, and our focus on enterprise-grade security and governance position us as the preferred choice for organizations seeking to achieve leadership in the era of gen AI.

Let's explore how Google Cloud can help you secure your future, and lead the way in AI.

Build AI-ready infrastructure today.

Contact us

Methodology

This report is based on quantitative research conducted in August and September of 2024.

Respondents have been with their organizations for at least a year and have a significant influence or decision-making authority on the purchase or use of technology solutions or the selection of technology vendors. The companies represented in this research have either a workload currently utilizing AI or they intend to leverage AI in the next 12 months.

In addition, the technology leaders in this study are familiar with several gen AI applications and have experimented with or implemented some for their respective firms. They believe gen AI is at least somewhat important to their organization's current and future business operations and have at least some influence on decisions related to AI and machine learning (ML) technology adoption in their organization.

Respondents did not know Google was the research sponsor and the identity of participants was not revealed to Google.

Technology leaders from companies in:

North America
United States
Canada

Europe
United Kingdom
France
Germany

Asia-Pacific
Singapore
Australia
Japan

Companies from North America and Europe have at least

1,000 employees

Companies from Asia-Pacific have at least

500 employees

513

technology leaders were surveyed for our research in the United States, Canada, the United Kingdom, France, Germany, Singapore, Australia, and Japan with the U.S. making up more than a quarter of the sample (29%).

Region		Company size		Role level	
North America	40%	500-999	5%	C-level executive	19%
Europe	30%	1,000-4,999	54%	VP / GM	14%
Asia-Pacific	30%	5,000+	41%	Sr Director / Director	33%
				Manager level	35%

Time in role		Age		AI familiarity	
1-2 years	9%	18 to 34	32%	Intermediate	21%
3-5 years	41%	35 to 44	42%	Advanced	55%
6-9 years	38%	45 to 54	21%	Expert	24%
10 years+	12%	55-64	6%		