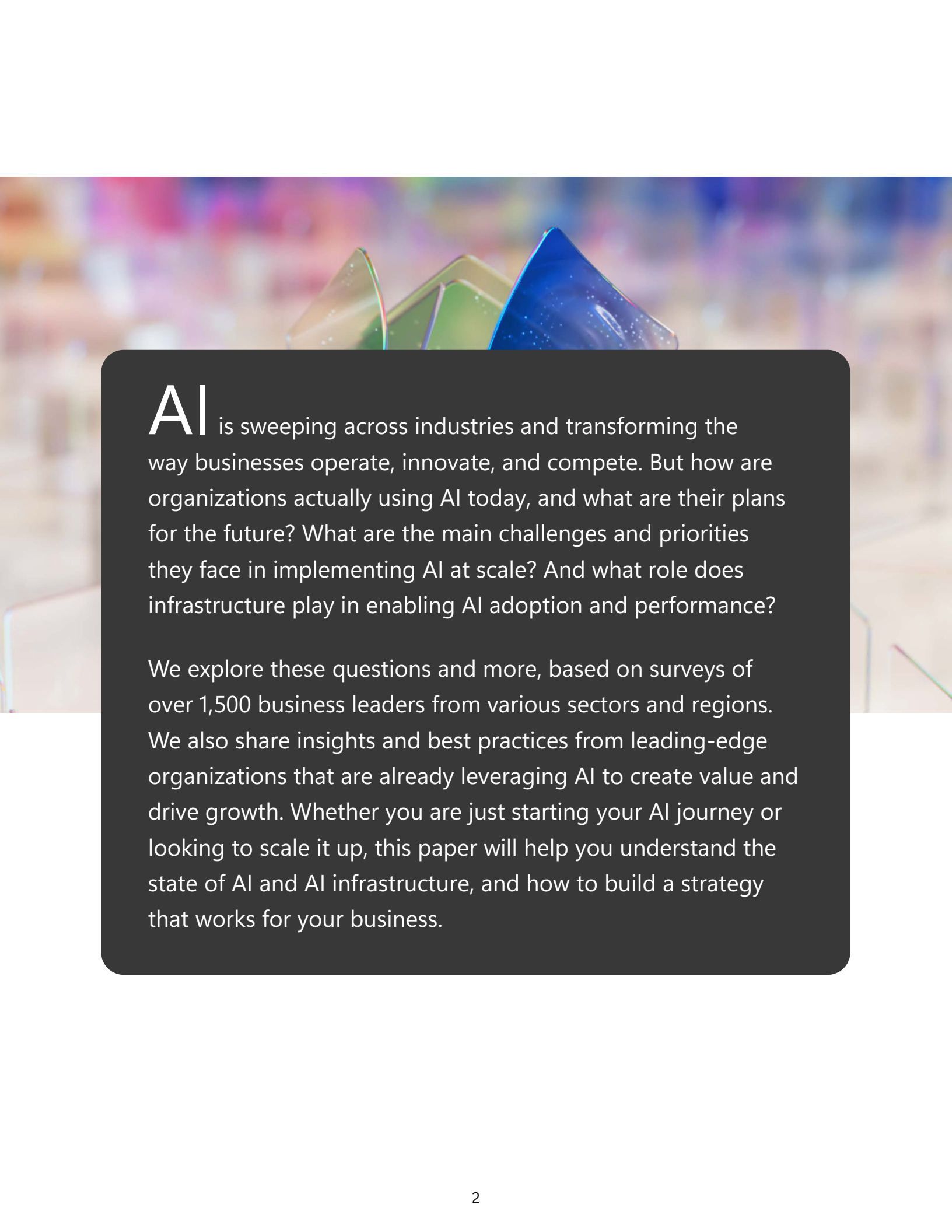Microsoft Azure

The State of AI Infrastructure

2024 Edition

An annual report on trends and developments in AI infrastructure based on Microsoft-commissioned surveys conducted by Forrester Consulting and Ipsos.

AI is sweeping across industries and transforming the way businesses operate, innovate, and compete. But how are organizations actually using AI today, and what are their plans for the future? What are the main challenges and priorities they face in implementing AI at scale? And what role does infrastructure play in enabling AI adoption and performance?

We explore these questions and more, based on surveys of over 1,500 business leaders from various sectors and regions. We also share insights and best practices from leading-edge organizations that are already leveraging AI to create value and drive growth. Whether you are just starting your AI journey or looking to scale it up, this paper will help you understand the state of AI and AI infrastructure, and how to build a strategy that works for your business.

Table of contents

AI is here.
It's just the beginning.

Welcome to the new era, where AI is not only intriguing and engaging consumers but exponentially increasing business productivity, transforming business models, and reimagining customer experiences. From retail to healthcare, there's no doubt AI is making a difference in ways like:

- Aggregating customer data to serve personalized recommendations at scale in retail.

- Powering CT scanners with robust algorithms for more accurate diagnostics and improved medical care.

- Predicting machine lifecycles for real-time maintenance and run-time efficiencies in factories.

- Preventing financial fraud via real-time detection using advanced AI tools and models.

- Using consumer-friendly chatbots to streamline customer service processes.

Recent Microsoft-commissioned research shows most companies are actively ramping up their AI capabilities, with 95% of businesses surveyed planning to increase their AI usage over the next two years. Across industries, AI adoption is believed to be critical for success.

Collectively, there is a consensus over the importance of AI, not just from an organizational standpoint, but also from a personal standpoint. This is an important distinction – these numbers show that it's not only organizations that are driving adoption, but that people see the value personally.

Source: A commissioned study conducted by Forrester Consulting on behalf of Microsoft and NVIDIA, May 2023

AI importance for success

AI is critical to my organization's success

Strongly agree/Agree

Total	75%
Finance	73%
Healthcare	72%
Retail	76%
Manufacturing	80%
ISV	75%

AI is critical to my personal success

Strongly agree/Agree

Total	66%
Finance	68%
Healthcare	62%
Retail	66%
Manufacturing	73%
ISV	64%

Many organizations are at the starting line

More than a third of companies are in the early stages of AI adoption: exploring options, gathering information, and planning various use-cases to strategize implementation, while a quarter are in the early pilot testing stage. With a majority of organizations still figuring things out, business leaders have an opportunity to beat their competition and gain advantages. But to do so, they'll need to act quickly in implementing their own AI strategies.

Organizations are in different stages of AI implementation

38%

Organizations in this **earliest stage** are in the initial stages of exploring options, gathering information, and identifying AI use cases as well as planning and strategizing their AI implementation.

Organizations in this **final stage** have fully implemented and are both actively utilizing and optimizing.

Organizations in the **pilot stage** are currently conducting small-scale implementations to evaluate feasibility and effectiveness.

25%

15%

22%

Deployment means that an organization has deployed in specific areas and is gradually expanding usage.
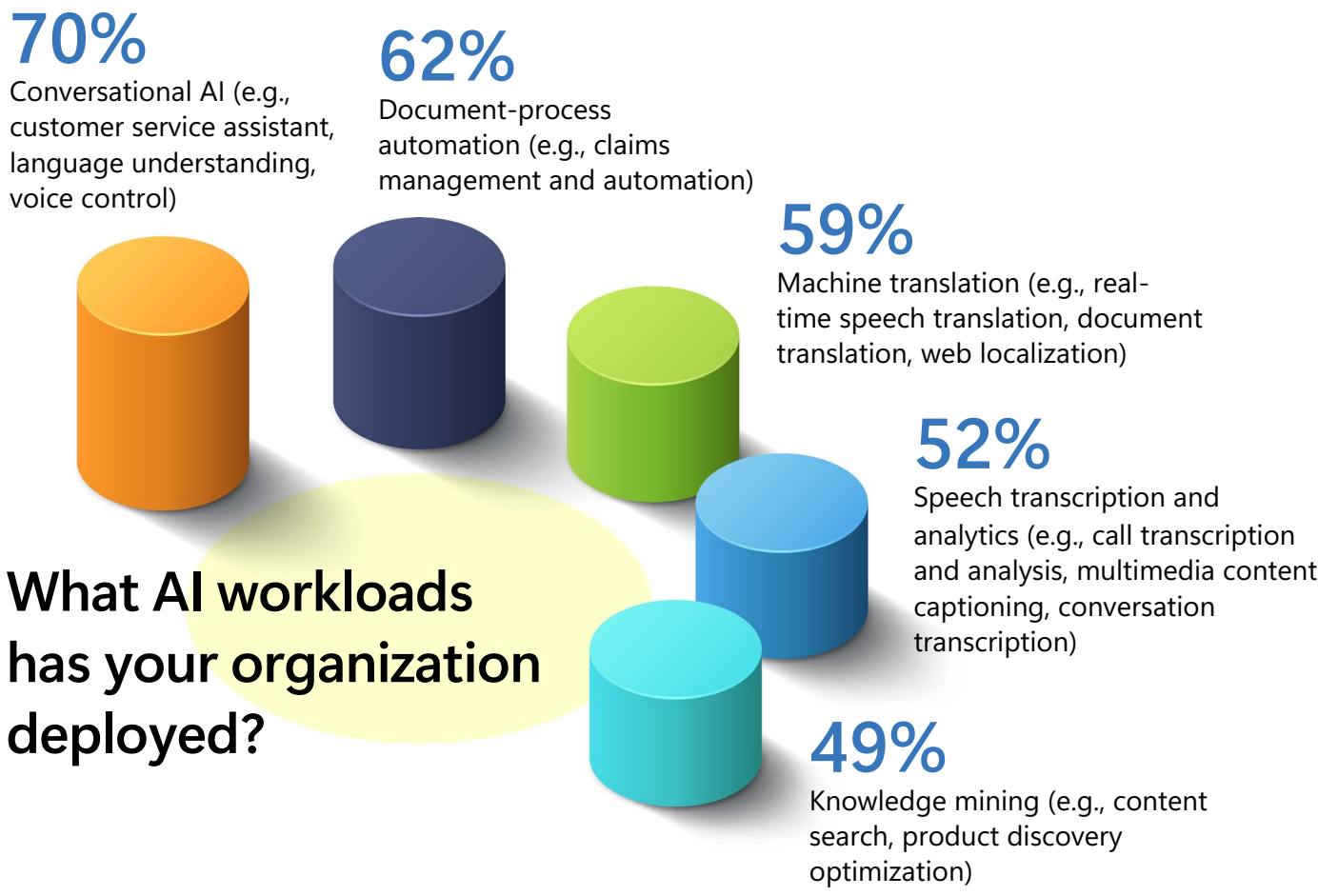
Base: Total (n=900), US (n=500), Germany (n=200), India (n=200).

Source: A commissioned study conducted by Ipsos on behalf of Microsoft, October 2023

Businesses are focused on automation and customers

Those companies that have started integrating AI are focused on supporting customer-facing applications and increasing efficiency through automation. These use cases tend to bring higher ROI, as they focus on getting more value out of their workers by reducing time spent on lower-value tasks. This makes sense, considering respondents, on average, expect a 34% ROI from their AI platforms.

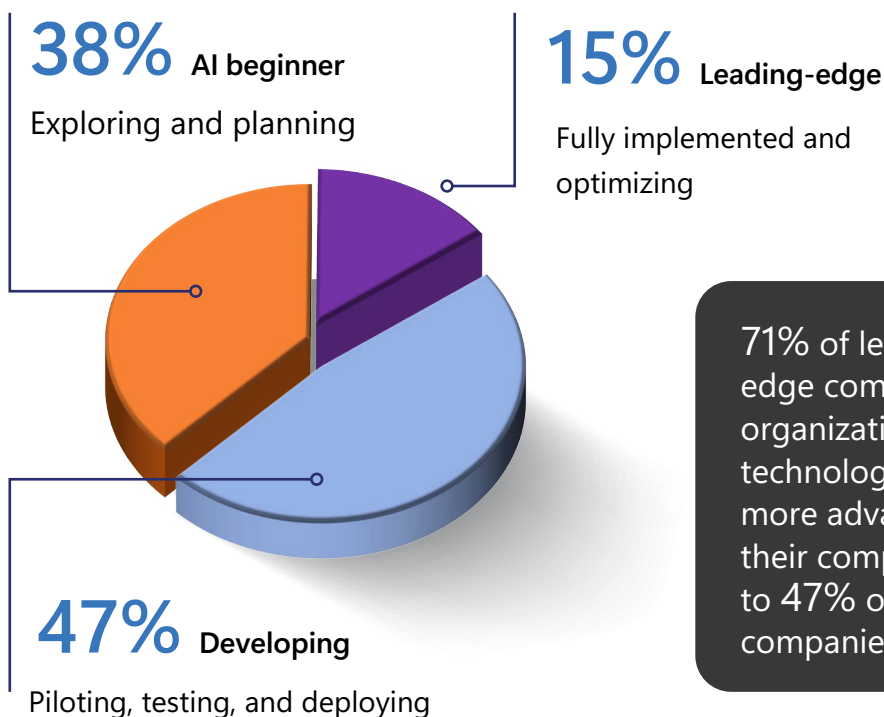On average, 46% of customer-facing applications and 44% of business/core applications leverage AI functionality.

70%
Conversational AI (e.g., customer service assistant, language understanding, voice control)

62%
Document-process automation (e.g., claims management and automation)

59%
Machine translation (e.g., real-time speech translation, document translation, web localization)

52%
Speech transcription and analytics (e.g., call transcription and analysis, multimedia content captioning, conversation transcription)

What AI workloads has your organization deployed?

49%
Knowledge mining (e.g., content search, product discovery optimization)

Base: (n=641)

Source: A commissioned study conducted by Forrester Consulting on behalf of Microsoft and NVIDIA, May 2023

"Leading-edge organizations" blaze the path forward

While many organizations are early in their AI journeys, 15% of businesses are advanced in their AI infrastructure and are considered "leading-edge organizations". These leading-edge organizations tend to be early adopters of technology and can provide valuable learnings on effective AI implementation strategies. Our recommendations are based on an analysis of leading-edge organizations plus other key insights to provide AI best practices and recommendations any company can leverage.

Stages of AI readiness

38% **AI beginner**

Exploring and planning

15% **Leading-edge**

Fully implemented and optimizing

47% **Developing**

Piloting, testing, and deploying

71% of leaders in leading-edge companies say that their organization is eager to try new technology and having IT that's more advanced than many of their competitors compared to 47% of non-leading-edge companies.
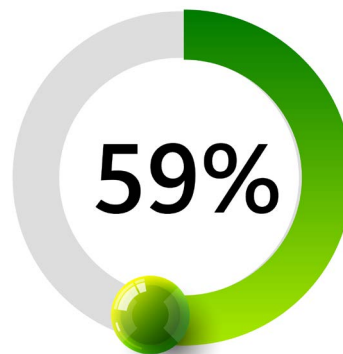
AI is challenging... for everyone

The AI landscape continues to evolve and AI implementation poses many different challenges and obstacles to overcome. Business leaders are faced with the daunting task of figuring out the best path forward.

99%

of organizations have
challenges in scaling and
operationalizing AI

59%

of business leaders
believe the AI market is
growing and evolving

"

My technology environment is very complex and dynamically changing.

We have complex needs and multiple departments use different applications... all with varying needs.

*In just a few years, I believe the AI will be much more advanced than what we have today.*

"

Base: Total (n=900)

Source: A commissioned study conducted by Ipsos on behalf of Microsoft, October 2023

The greatest challenge? Tackling the talent gap

Organizations have an immediate need for AI experience and talent. Addressing this gap by bolstering employees' skills and training today is the key to bridging the gap and moving forward quickly.

39%
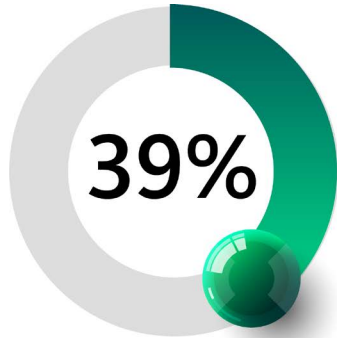
Leaders who rank having the skills required to develop or customize AI models as one of their top 3 technology challenges (out of 13 items).

33%
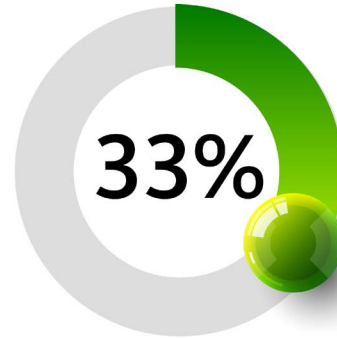
Leaders who rank having enough talent as one of their top 3 organizational challenges (out of 13 items).

Security, capabilities, and ROI considerations

Along with AI talent sourcing, it's not surprising that many of the other challenges businesses face are centered around technological and strategic challenges. Security considerations, having adequate capabilities for designing, implementing, and managing infrastructure, and having the appropriate AI tools are key technology challenges.

And along with employee talent, unclear ROI of AI implementation, the right resources to support AI development and management and collaborating across business functions are top organizational challenges.

Base: Total (n=900)

Source: A commissioned study conducted by Ipsos on behalf of Microsoft, October 2023
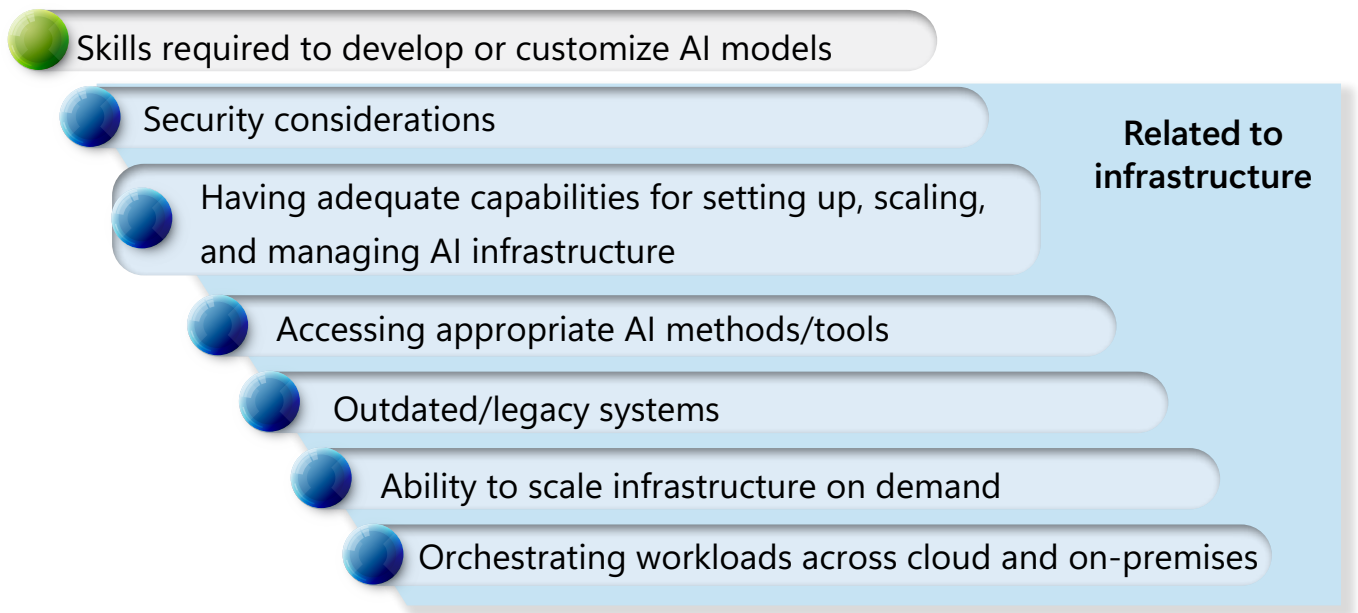
Infrastructure challenges remain top of mind

The right infrastructure can make or break your AI projects. Overwhelmingly, the AI challenges fall within infrastructure (hardware, software, and tools) and remain the most common roadblock in implementing and leveraging powerful AI tools. Prioritizing the right AI infrastructure is key to successful AI implementation, scaling, and innovation.
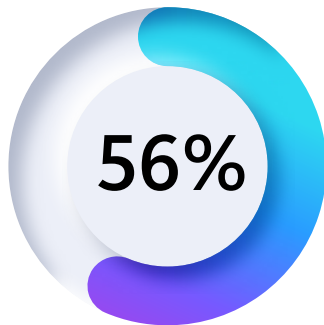
56%   **My organization doesn't have the proper infrastructure to support my organization's desired AI workloads.**

Top technology challenges organizations face
(descending order, showing top 7 of 13 items)

- Skills required to develop or customize AI models
- Security considerations
- Having adequate capabilities for setting up, scaling, and managing AI infrastructure
- Accessing appropriate AI methods/tools
- Outdated/legacy systems
- Ability to scale infrastructure on demand
- Orchestrating workloads across cloud and on-premises

Related to infrastructure

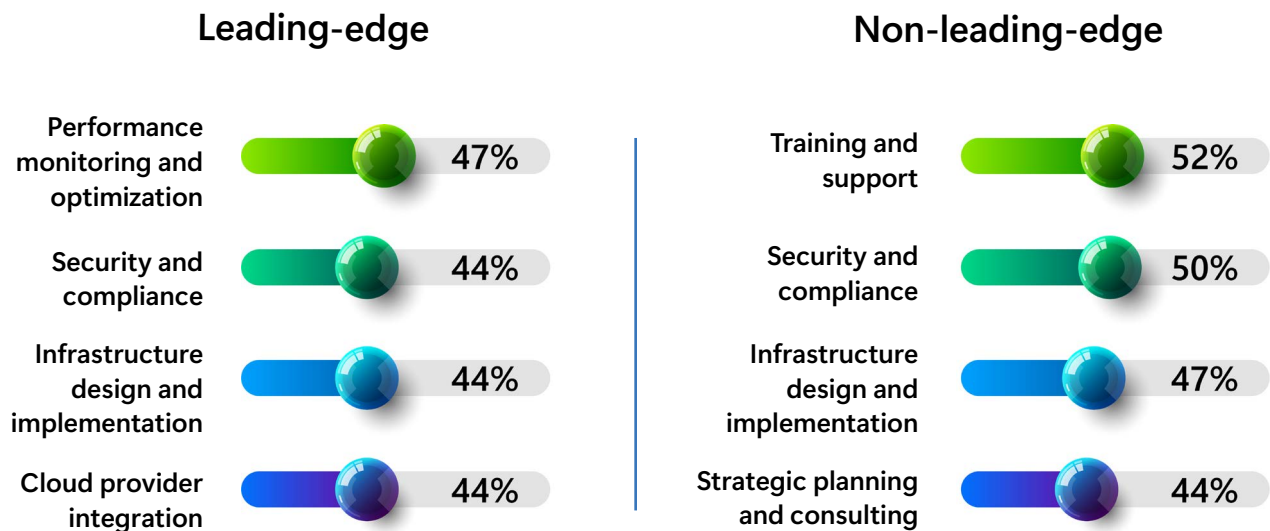Base: Total (n=641), Total (n=900)

Source: Commissioned studies conducted by Forrester Consulting and Ipsos on behalf of Microsoft, May 2023 and October 2023

Leaders are looking to partners for help

For companies not sure how to start leveraging AI, partnering with a solution provider with deep AI expertise and proven AI solutions can help companies accelerate AI production and address AI infrastructure challenges. Business leaders are looking to partners to help with infrastructure design and implementation, training and support, security and compliance, and strategic planning and consultation.

Where it gets interesting is that as companies move further along their AI journey, they start to prioritize things like performance, optimization, and cloud provider integration. Engaging the right partner can help businesses of any size and at any stage of AI implementation accelerate their AI journey. This is both a huge opportunity for partners, and a burden. They must make sure their staff is ready to go and able to help with consulting, strategy, and training.

Where partners are expected to help (top 4)

Leading-edge

Performance monitoring and optimization	47%
Security and compliance	44%
Infrastructure design and implementation	44%
Cloud provider integration	44%

Non-leading-edge

Training and support	52%
Security and compliance	50%
Infrastructure design and implementation	47%
Strategic planning and consulting	44%

Base: Total (n=900)

Source: A commissioned study conducted by Ipsos on behalf of Microsoft, October 2023

AI infrastructure remains elusive

If you're not sure how to approach your AI infrastructure, you're not alone. Robust and scalable infrastructure specifically built for AI is critical to support the complexities of new AI-driven workloads and processes, but AI infrastructure is a key challenge for most leaders, who face many obstacles in implementing and operationalizing AI, such as:
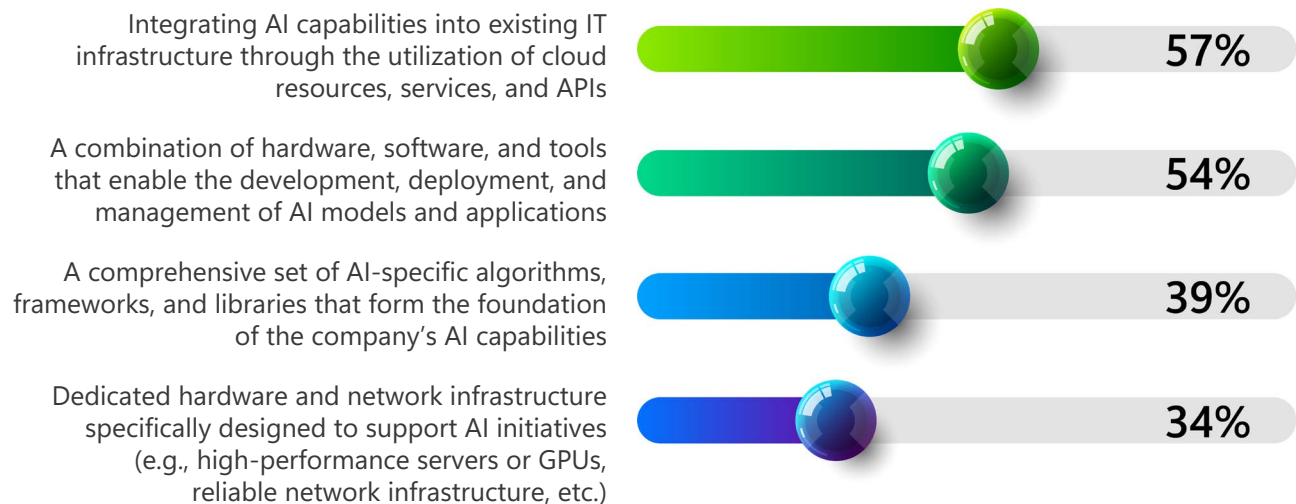
- Outdated and legacy systems that are not designed to handle the complexity and volume of AI workloads.

- Data security and privacy concerns, especially sensitive and personal data, that require robust protection and compliance measures.

- Workload orchestration challenges, such as managing multiple platforms, tools, and frameworks, and optimizing resource utilization and performance.

- Skills gap, as many organizations lack the talent and expertise to develop, customize, and deploy AI models and applications.

- The accelerated rate of technological advancements like GenAI that have large implications on the type and complexity of the infrastructure needed.

Defining AI infrastructure

Infrastructure challenges are amplified due to diverse interpretations of AI infrastructure among organizations. These interpretations range from integrating AI capabilities into the existing IT infrastructure to establishing a dedicated hardware and network infrastructure and developing a comprehensive tech stack that includes algorithms, frameworks, and libraries. This can make something as simple as communicating needs with vendors a real challenge when the same language is not being used.

Not having a clear definition just adds to the challenges of getting started with AI. At the most simplistic, AI infrastructure includes the hardware, software, networking, and tools and services used to develop, implement, and optimize AI. As AI continues to evolve, it'll be more important than ever to settle on a standard definition that can be used across all industries.

The different ways organizations define "AI infrastructure"

Integrating AI capabilities into existing IT infrastructure through the utilization of cloud resources, services, and APIs	**57%**
A combination of hardware, software, and tools that enable the development, deployment, and management of AI models and applications	**54%**
A comprehensive set of AI-specific algorithms, frameworks, and libraries that form the foundation of the company's AI capabilities	**39%**
Dedicated hardware and network infrastructure specifically designed to support AI initiatives (e.g., high-performance servers or GPUs, reliable network infrastructure, etc.)	**34%**

Microsoft's definition of AI infrastructure: "A combination of hardware, software, and tools that enable the development, deployment, and management of AI models and applications."
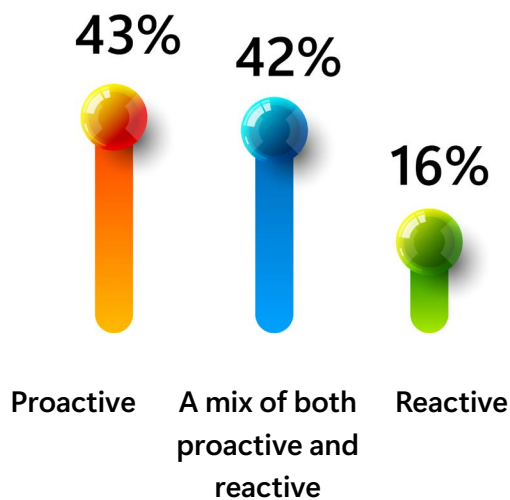
Making AI infrastructure a priority

Businesses are already understanding the urgency and importance of having a solid foundation for their AI initiatives. 41% of leaders agree that infrastructure is the area they need most help with and 39% need help with strategic planning and consulting, citing specialized components like infrastructure or security as well as broader design and implementation. Additionally, 43% are predominantly proactive in developing their AI infrastructure strategy, compared to 16% who are mostly reactive. There's a clear opportunity for partners to provide the consultation and expertise companies need to optimize their infrastructure for AI.

Areas organizations need the most help with

41%
Training and support

41%
Infrastructure design and implementation

40%
Security and compliance

39%
Strategic planning and consulting
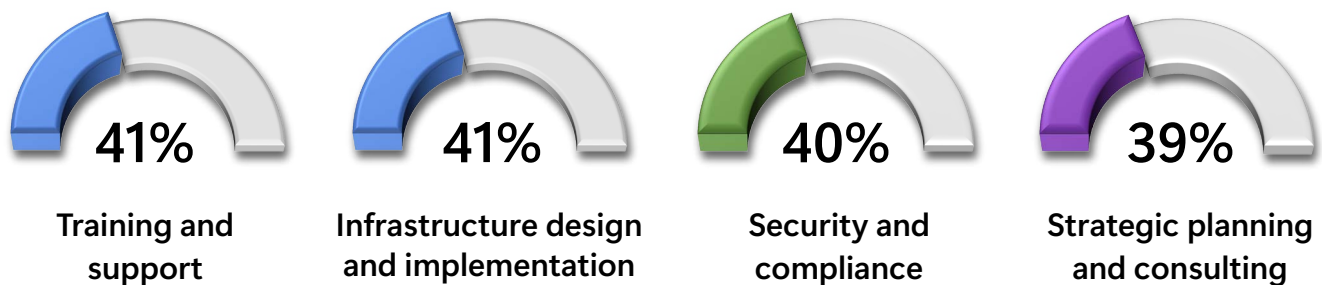
Most leaders aim to be proactive

43%
42%
16%

Proactive

A mix of both proactive and reactive

Reactive

Strategic planning and consulting is desired by many organizations across all industries. Business leaders that were earlier in their implementation were more likely to need help with 42% of AI beginner organizations stating they needed help. As organizations begin their journey, they can leverage the consultation of partners.

Industries bring their own nuances
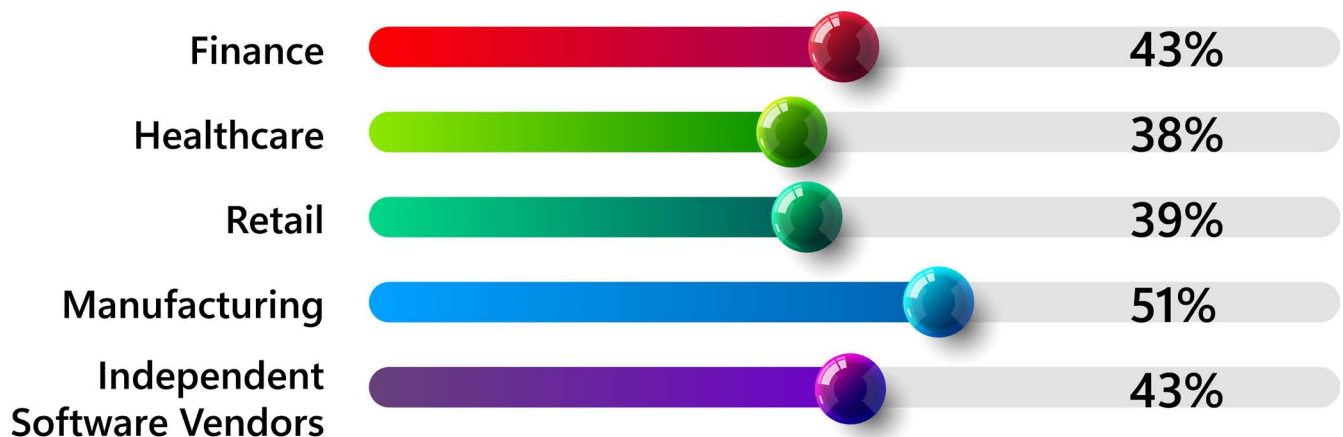
Manufacturing tends to be the most proactive (51%) in planning for AI infrastructure, significantly more so than those in healthcare and retail. In such a process-driven industry, even minor gains in efficiency can bring critical advantages over the competition. AI technologies, even in its current form, brings giant leaps forward in operational efficiency.

The entire industry stands to gain significant benefits through process optimizations, advanced automation, and predictive maintenance. Business leaders across other industries can follow manufacturing's example and move quickly to stand up the right infrastructure for their AI needs. Like any other business strategy decision, these differences highlight the cross-industry nuances that need to be considered when taking on any new technology.

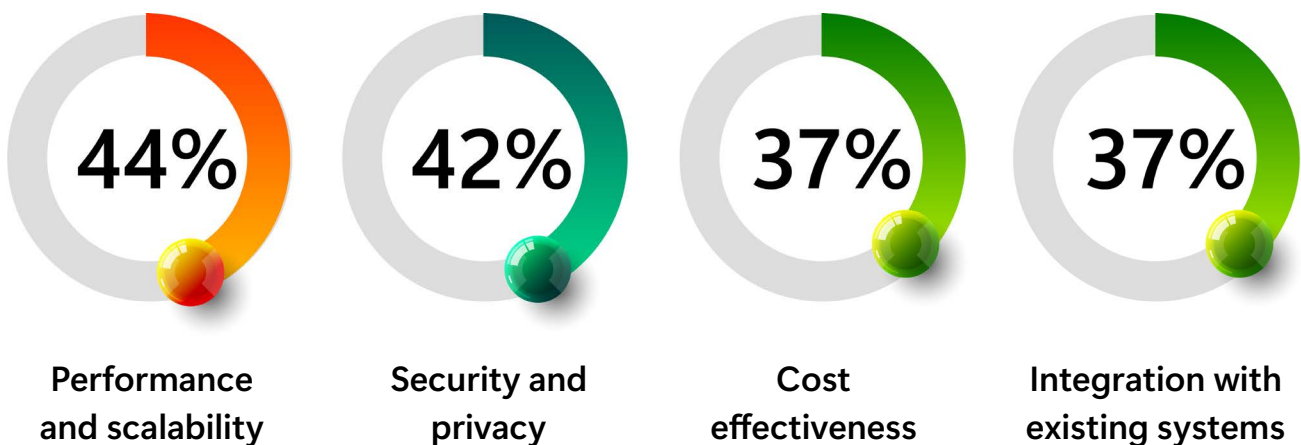## Manufacturing organizations aim to be the most proactive

Industry	Percentage
Finance	43%
Healthcare	38%
Retail	39%
Manufacturing	51%
Independent Software Vendors	43%

Base: Total (n=900)

Start with the trifecta: performance, security, and cost

There is no doubt there's a lot to consider when finding the right AI infrastructure a business needs, and the speed of changes being brought to the market add further complexities. As businesses start analyzing vendors and setups, they can look to these top considerations as a starting-off point: performance, security, and cost.

Top priorities for AI infrastructure

| 44%	42%	37%	37%
Performance and scalability	Security and privacy	Cost effectiveness	Integration with existing systems

Base: Total (n=900)

Source: A commissioned study conducted by Ipsos on behalf of Microsoft, October 2023

Performance and scalability

With 44% of leaders prioritizing performance and scalability, particularly in industries managing high-volume and complex AI workloads like retail, manufacturing, and independent software vendors, the imperative for AI infrastructure lies in providing swift and reliable computing resources to optimize resource utilization, reduce latency, and scale up and out as needed. Focusing on performance and scalability means looking beyond the cost to consider all the benefits that come with AI and maximizing the full impact of AI infrastructure implementations.

52%
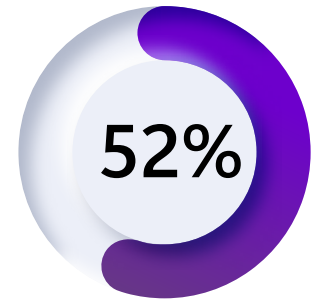
Independent software vendors

Security and privacy

Security and privacy also take precedence, with 42% of business leaders citing it as a top priority. This is especially true for finance and healthcare where they must secure confidential data against unauthorized access, cyber threats, and data breaches while complying with stringent regulations. While security fell slightly in importance for retail, manufacturing, and ISV, these industries still rated it as a key priority to address overall. These factors are essential for ensuring the reliability, efficiency, and effectiveness of AI infrastructure solutions, and addressing the key challenges that leaders face.

"Security and privacy" tops the list for finance and healthcare
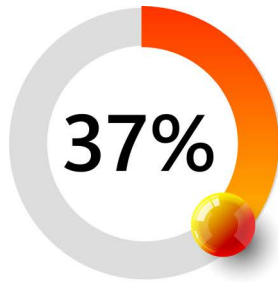
49%

44%
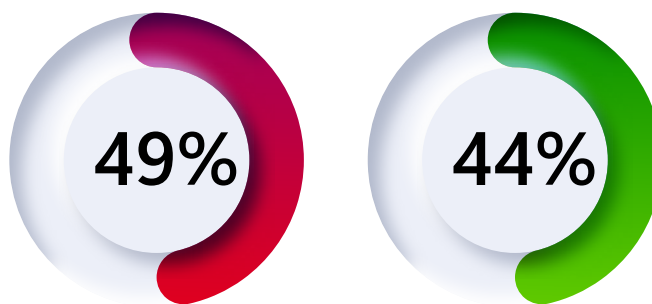
Finance

Healthcare

Base: Total (n=900)

Source: A commissioned study conducted by Ipsos on behalf of Microsoft, October 2023
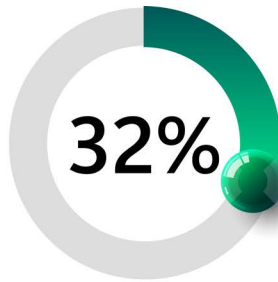
Cost effectiveness

Not surprisingly, cost effectiveness is a priority, with 37% of business leaders across industries citing getting the value they need from their AI infrastructure and meeting ROI goals as important. Cost is even more relevant for retail industries.

AI infrastructure priority	Organizational challenges	Barriers and blockers for AI adoption
37%	**32%**	**20%**
Cost effectiveness	**ROI is unclear**	**Cost and ROI**
(rank 3 of 10)	(rank 2 of 13)	(rank 4 of 20)

Beyond these top three priorities, businesses expect the following from their AI infrastructure:

> Cost effectiveness is more important to leaders in the retail industry (42%) than other industries and tops their list of AI infrastructure priorities.

Other priorities

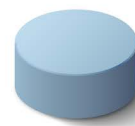| 31%	29%	23%	23%	19%
User-friendly tools and interfaces	**Flexibility and customization**	**Robust data management**	**Continuous performance monitoring and optimization**	**Support and maintenance**

Base: Total (n=900)

Source: A commissioned study conducted by Ipsos on behalf of Microsoft, October 2023

One size does not fit all

Priorities are fluid, shifting to match an organization's constantly evolving context. Factors like industry, market, AI maturity level, and platforms create a constantly evolving environment to navigate. Leaders need to be able to plan for these shifts in priorities by understanding the organization's changing context and impact on implementations.

Level of AI readiness

Leading-edge organizations face different challenges compared to businesses that are earlier in their AI infrastructure implementations. As an organization moves further along, flexibility, data management, maintenance, and support compete with earlier priorities like performance, security, cost effectiveness, and integration.

As new technologies and processes become their standard mode of operation, priorities change, and it's important for leaders to proactively plan for these changes. Understanding that more things become important as your organization progresses will enable you to pivot priorities quickly to meet changing and expanding needs.

Priorities that intensify by readiness level

Robust data management

Leading-edge	34%
AI beginner	17%

Performance and scalability

Leading-edge	38%
Developing	51%

Security and privacy

Leading-edge	37%
AI beginner	47%

Base: Total (n=900)

Source: A commissioned study conducted by Ipsos on behalf of Microsoft, October 2023

Workload type

The AI infrastructure capabilities needed vary based on business model and workloads. One company may need fully customizable software, services, and computing while off-the-shelf AI models, services and platforms may be fine for another. As AI continues to evolve, the needs of companies and even the solutions are rapidly changing. Because of this, providers are ramping up their offerings to reach a wider range of needs and providing multiple points at which a customer can come into their system.

We've identified three different categories of customers for AI infrastructure.

AI leaders

'AI leaders' have a defined AI strategy and want to lead their market by building their own innovative, homegrown AI models and applications. They require highly performant supercomputing infrastructure that can flexibly meet complex storage, network, computing, and security needs. Their AI workloads are incredibly complex, involve massive models, and require control at every layer of their AI infrastructure.

Business drivers: Developing an end-to-end AI service, solution or platform, from the ground up. Require unlimited scalability and the ability to deliver customized user experiences.

AI power users

'AI power users' also have a defined AI strategy and are heavily customizing pre-built AI models, infusing company-specific content and data, and retraining. They need control over each layer of their AI infrastructure but typically don't require massive compute power when working with pre-trained models.

Business drivers: Looking to maximize efficiency and minimize time to market. Save time by using pre-built models and optimizing for their needs.

AI ready

'AI ready' companies want infrastructure that is ready to go so they can focus on defining their AI strategy. They don't want to worry about the ins and outs of infrastructure, they are looking for a scalable, out-of-the-box solution that can support discrete processes now and support AI growth.

Business drivers: Taking their first steps with AI. Need off-the-shelf solutions.

Platform considerations

There is no one-size-fits-all to determining whether a company should be on-premises, hybrid, or on the cloud – every type of solution has their positives and negatives. With a myriad of factors at play, ultimately the decision about what is best lies with the company and their unique situation. For example, on-premises AI infrastructure may offer more control but require more upfront investments and can be expensive to maintain, difficult to scale, and hard to keep up to date with the latest technologies.

AI infrastructure in the cloud offers fast deployment, scalability and flexibility, and typically the best technology available, but some have concerns with security, privacy, and meeting compliance requirements. Hybrid setups offer the benefits (and drawbacks) of both but with higher degrees of complexity. Startups may benefit even more from a cloud set-up due to their small employee size and their naturally increased focus on getting their product to market as fast as possible. Nonetheless, key themes that came across for each of these were security and cost effectiveness.

Top three priorities by solution setup

On-premises	Hybrid	Cloud
• Data security	• Security and compliance	• Data privacy and security
• Cost effectiveness	• Cost effectiveness	• Flexibility and scalability
• Existing IT infrastructure integration	• Scalability and elasticity	• Cost optimization

" *For more advanced models, we need GPUs that will be more useful. But for now, we have some, light-weight models that can be run in CPUs.*

The infrastructure requirements will be different depending on the planned use-case. For example, an early pilot project and actual full-scale implementation have completely different infrastructure. requirements. "
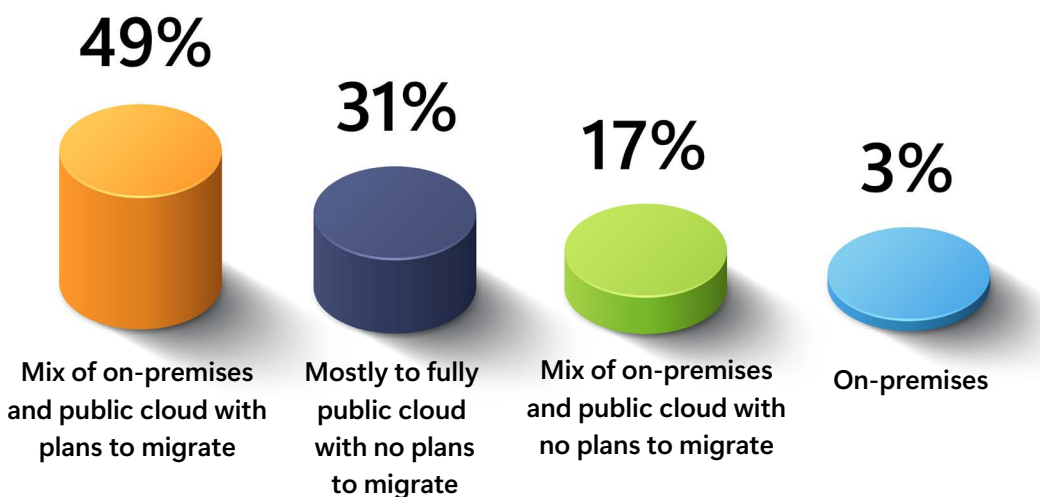
Half of business leaders said they have a hybrid setup with plans to migrate fully to the cloud for better IT governance and security, increased productivity and scalability, more successful deployments, greater innovation, and ROI. For those that are "AI ready," a cloud provider may provide a more comprehensive out of the box solution to get started.

Further than deciding whether to have an on-premises, hybrid, or cloud setup, the vendor choice is fraught with a range of choices. Key features like having high quality AI algorithms, ability to manage networking capabilities, integration with open-source tools, API accessibility, scalability, clear documentation, and multi-cloud/hybrid enablement came through as the highest requirements.

How would you describe your organization's current AI workload implementation?

49%
Mix of on-premises and public cloud with plans to migrate

31%
Mostly to fully public cloud with no plans to migrate

17%
Mix of on-premises and public cloud with no plans to migrate

3%
On-premises

What are the main benefits of hosting AI workloads in the cloud?

Better IT governance

Increased productivity

Higher percentage of AI concepts successfully deployed in production

Increased scalability

Increased data security for ML models and data sets

Industry context

Different industries have different priority nuances. More regulated industries like finance and healthcare put a higher focus on security and privacy, while manufacturing and ISVs require strong performance and scalability. Additional factors like solution set-up and level of AI maturity will also make a difference in both their organizational and technical priorities.

Top 3 AI infrastructure priorities by industry

Finance	Healthcare	Retail
1. Security and privacy	1. Security and privacy	1. Performance and scalability
2. Performance and scalability	2. Performance and scalability	2. Cost effectiveness
3. Integration with existing systems	3. Integration with existing systems	3. Integration with existing systems

Manufacturing	Independent Software Vendor
1. Performance and scalability	1. Performance and scalability
2. Integration with existing systems	2. Security and privacy
3. Security and privacy	3. Cost effectiveness

> *Security is a major concern, with so much data available and protections needed. (Infrastructure let us) use machine algorithms to detect threats in real time and enhance cybersecurity.*
>
> *AI lets us shorten the time required for so many things. And it helps our performance and overall activity.*

Base: Total (n=900)

Source: A commissioned study conducted by Ipsos on behalf of Microsoft, October 2023

Harnessing the power of AI now

To help businesses move forward in their AI journey, we recommend four actions to help navigate the challenges and speed AI production and integration.

Prioritize your AI infrastructure

Infrastructure is at the core of AI innovation. It can determine how fast, how good, how easy, how groundbreaking, and how engaging an AI application, solution, or platform will be. Companies should carefully examine their AI goals and strategy and determine what infrastructure capabilities and platform (on-premises, cloud, hybrid) best fit their needs today and in the future. It is rare that existing infrastructure can power the demands and complexity of AI.

Most businesses will need to make changes, either overhauling their existing infrastructure, choosing a solution provider offering a full-stack AI platform, or something in between. A company's AI infrastructure strategy can shape the future of their business, either accelerating their AI journey or blocking their innovation.

Overcome the skills gap

To overcome the AI skills gap, business leaders need to invest in training and upskilling their current employees and/or consider bringing in outside talent. Partnering with an experienced AI solution provider can also help fill the void and deliver employee training, strategy planning, and AI infrastructure, production, and implementation support.

Make it secure

Security, privacy, and compliance should be at the forefront of any AI and infrastructure plans. Secure AI is the act of securely designing, developing, and deploying AI and GenAI capabilities and systems. Follow these best practices:

- Keep user data private and secure.
- Ensure transparency in procedures and emphasize the significance of clearly communicating sources and criteria for decision-making.
- Ensure security is built-in from inception to deployment of the AI system's lifecycle.
- Keep risk at the forefront when designing interfaces and processes.

Find a partner

Across all industries, business leaders expressed a need for help with strategic planning and consulting as well as training and support from AI solution providers. Leading-edge organizations partner with AI experts to help plan, build, and integrate AI into their business. Companies of any size and at any stage can benefit from a strategic AI solution provider. Forming a partnership with a proven AI solution provider can be key to accelerating AI production and staying competitive.

Take the next steps on your AI transformation journey

Explore how Microsoft Azure is redefining cloud infrastructure to prepare every business for AI by providing the world-class technology for AI workloads and doing so sustainably and responsibly.

Get strategic guidance and insights on AI innovation, tailored for business leaders.

Learn how businesses are balancing performance, efficiency, and cost with Azure AI infrastructure.

Research methodology

In May 2023, we commissioned Forrester Consulting to evaluate the current state of AI among IT directors and decision makers in North America, Europe, and Asia Pacific.

We further explored this topic through a second study with Ipsos in September 2023 among technical and business leaders, developers, and data professionals who are early majority adopters of technology across 3 markets (US, Germany, and India) and 5 different industries (finance, healthcare, retail, manufacturing, and independent software companies).

Forrester Consulting Research

Fielding: May – June 2023

Participants: n=641 Director and above in IT with responsibilities in AI workloads, cloud infrastructure.

Company Size: 1000+ FTE in North America, 500+ in all other countries

Countries: North America (US, Canada), EMEA (UK, Germany, France), APAC (Australia, New Zealand)

Industries: All industries, including automotive, manufacturing, oil and gas, financial services, public sector, universities, bio life sciences.

Ipsos Research

Fielding: September - October 2023

Participants: n=900 ITDMs who are in the early majority of technology adopters at their organizations.

Company Size: ISVs = 50+ employees
All other industries = 500+ employees

Countries: US (n=500), Germany (n=200), India (n=200)

Industries: Finance (US n=100, Germany/India n=40); Healthcare (US n=100, Germany/India n=40); Retail (US n=100, Germany/India n=40); Manufacturing (US n=100, Germany/India n=40); ISV (US n=100, Germany/India n=40)

ISVs are independent software vendors or software houses that develop software for broad commercial distribution, including SaaS.