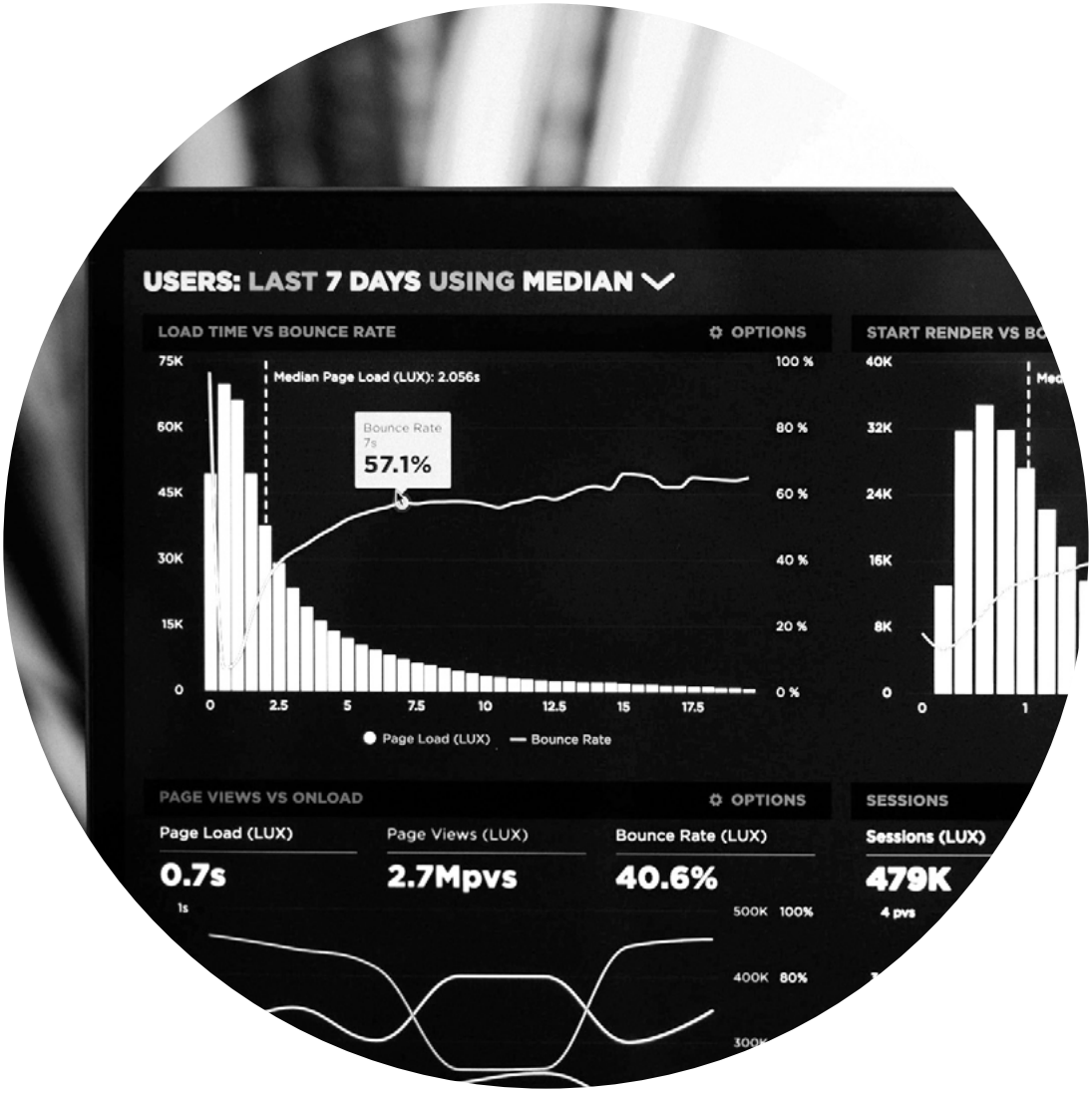
The 2021 State of AI Infrastructure Survey



This Guide Covers:

Introduction

Most research around the state of the Artificial Intelligence (AI) industry talks about the same few facts: AI is still very immature, models rarely make it to production, and challenges remain for data scientists and research teams around creating the right infrastructure and setting up AI for success.

To discover whether these pervasive ideas are still gospel in 2021, we commissioned a survey of 211 data scientists, AI/Machine Learning/ IT practitioners and system architects from 10 countries around the world. We spoke primarily with experts from large enterprise companies with over 5,000 employees, and some with as many as 10,000. We asked these enterprises to open up about the technologies they use, the challenges they face with AI and the size of not only their AI budget, but also their confidence in brining AI into production. The survey was completed by independent research company Global Surveyz and took place in July 2021.

The results are a fascinating look at the true state of AI maturity. We are working in a market with enormous potential. Three-quarters of those surveyed are looking to expand their AI infrastructure, and 38% have more than $1 million in annual budget to make that happen. However, big challenges definitely still exist, and many companies face early-stage hurdles with AI infrastructure setup, data preparation, and even goal setting. With so much invested in making AI successful, and companies looking to forge ahead and make progress, it's clear that early adopters of the right technology have a lot to gain.management, IT, and finance.

Key Findings

- **AI is a cloud-native world**
  AI was clearly born with the cloud in mind, with 81% of companies working cloud-natively (using containers and cloud technologies) for their AI workloads. Along with the use of containers comes adoption of Kubernetes and other cloud-native tools for management of containers. A sizeable 42% of respondents are already on Kubernetes, another 13% on OpenShift, and 2% on Rancher /SUSE. These numbers are considerably higher than container adoption for non-AI workloads, making AI a leader in cloud-native adoption.

- **Big spenders, but a lack of confidence**
  Our study shows that 38% of companies have a budget of more than $1M per year for AI infrastructure alone, and 59% have more than $250k per year. These huge budgets should indicate high confidence among the companies surveyed that they can get AI models into production. However, our survey found that for 77% of companies, less than half of models make it to production. Further, 88% of companies say that they are not fully confident in their AI infrastructure set-up and aren't sure that they can move their models to production in the timeline and budget provided.

- **Infrastructure challenges weigh heavily on AI teams**
  Lack of confidence in AI infrastructure extends to hardware utilization, with more than 80% of surveyed companies not fully utilizing their GPU and AI hardware, and 83% of companies admitting to idle resources or only moderate utilization. Only 27% say that GPUs can be accessed on demand by their research teams as needed, with almost half of those who responded relying on manual requests for allocating compute resources.

- **AI is still a relatively immature market**
  The top challenges for today's AI teams are data collection (61%), infrastructure/compute (42%) and defining business goals (36%). All three of the biggest challenges are early-stage problems for teams working with AI, which speaks to market immaturity. In addition, the tools used to manage infrastructure for AI teams include home-grown tools (23%) and even Excel spreadsheets (16%) again showing that in many ways, AI still lacks maturity.

Budgets are growing, despite challenges

AI challenges are relevant across all respondents, regardless of company size, industry, AI spend, or infrastructure location (cloud, hybrid, or on-premises). Infrastructure utilization is an issue for between 85%-90% of respondents, even among companies that have $10M or more budgeted for AI each year. Despite this, most companies are not limiting their budgets until their challenges are solved, with 74% planning to increase spend on AI infrastructure in the next year.

AI has enormous potential for those who beat the challenges

There is strong pressure on enterprises to launch AI projects and to see value from Artificial Intelligence. While the challenges may still be early-stage issues like goal setting and infrastructure set-up, the spend is far from immature. The financial support is in place to propel AI projects, but it needs to be channelled to the right places, improving the systems used for AI infrastructure management, solving hardware utilization challenges, and supporting research teams in gaining both confidence and access to resources.

Large Teams and Big Budgets Don't Protect from Hardware Utilization Issues

GPU Farm Size and Server Locations

Over half of surveyed companies (53%) have GPU farms of 10 or more GPUs (figure 1), with a full 20% using over 100 GPUs for their AI research. This speaks to the considerable investment that has already been made in AI.

Though the move to cloud from on-premises infrastructure is widely discussed, that trend has yet to be fully realized in practice, with two-thirds (64%) hosting their GPU in the cloud or hybrid and a third running on-prem (figure 2).

Over half (53%) already have all or some of their AI applications and infrastructure in the cloud, with another third (34%) planning to move to the cloud in the coming years (figure 3).

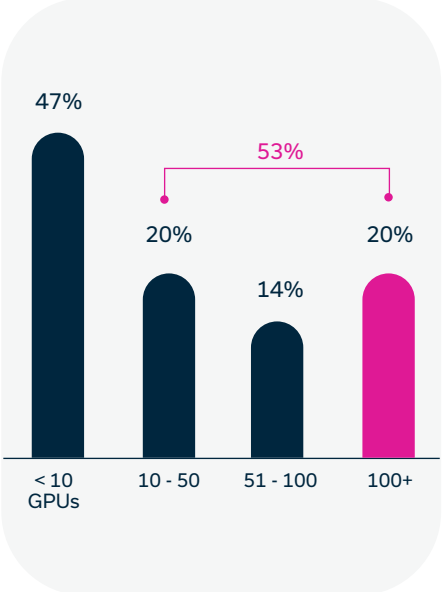


Figure 1: Size of GPU Farm

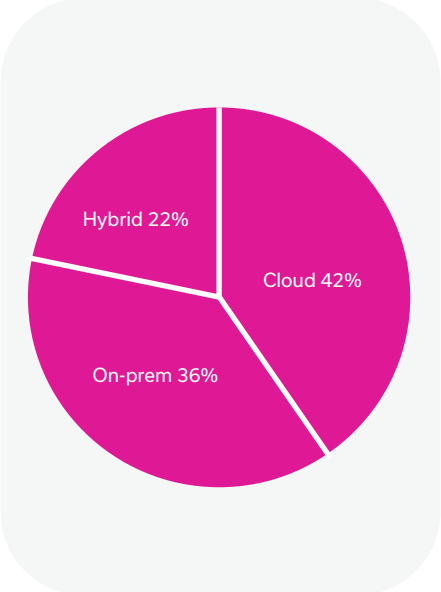


Figure 2: GPU Servers' Location

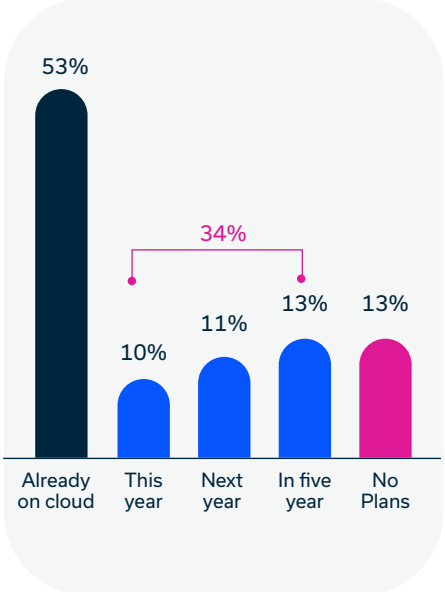


Figure 3: Plans for Moving AI Applications and Infrastructure to the Cloud

Size of Research Teams and Access to On-Demand GPU Compute as Needed

Almost two-thirds (63%) of companies have research teams of 10 or more, and yet only 27% of them have solved the need for fully on-demand access to GPU compute. A larger research team doesn't equate to greater accessibility to compute resources.

Over a third (35%) of research teams access GPU compute only via additional steps or static assignment, and almost half of this group (43%) are subject to waiting for approval of their manual requests (figure 5). Every time they want to run a job, they need to make this manual request, slowing down operations and adding frustration and delay.

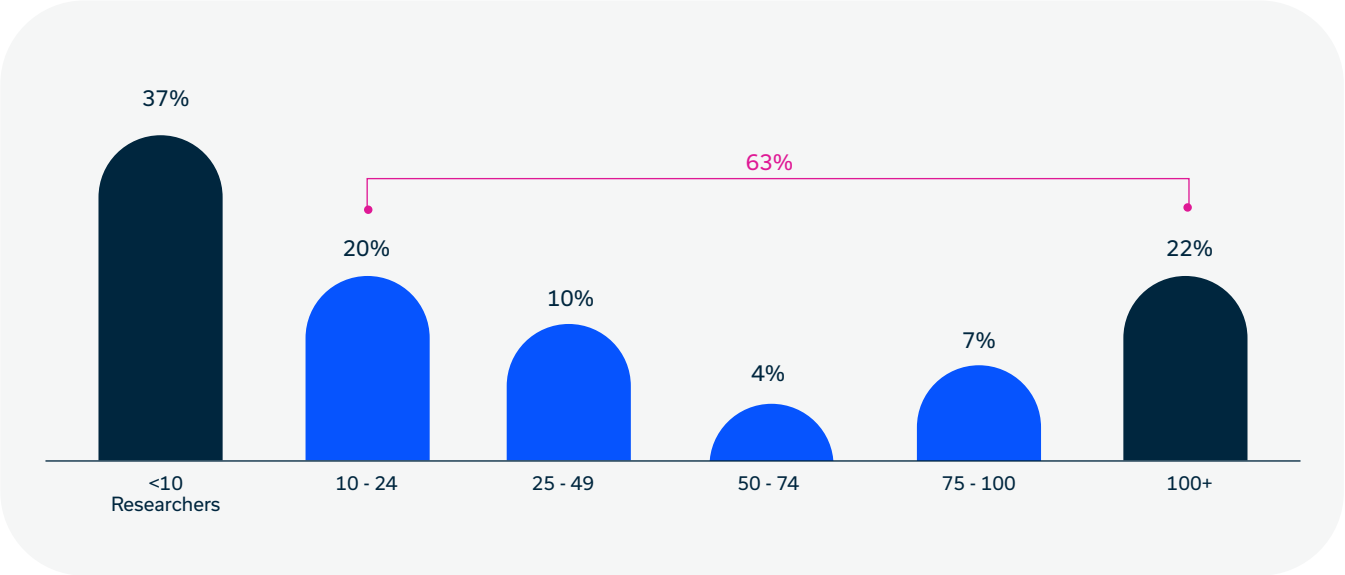


Figure 4: Size of Deep Learning Research Team

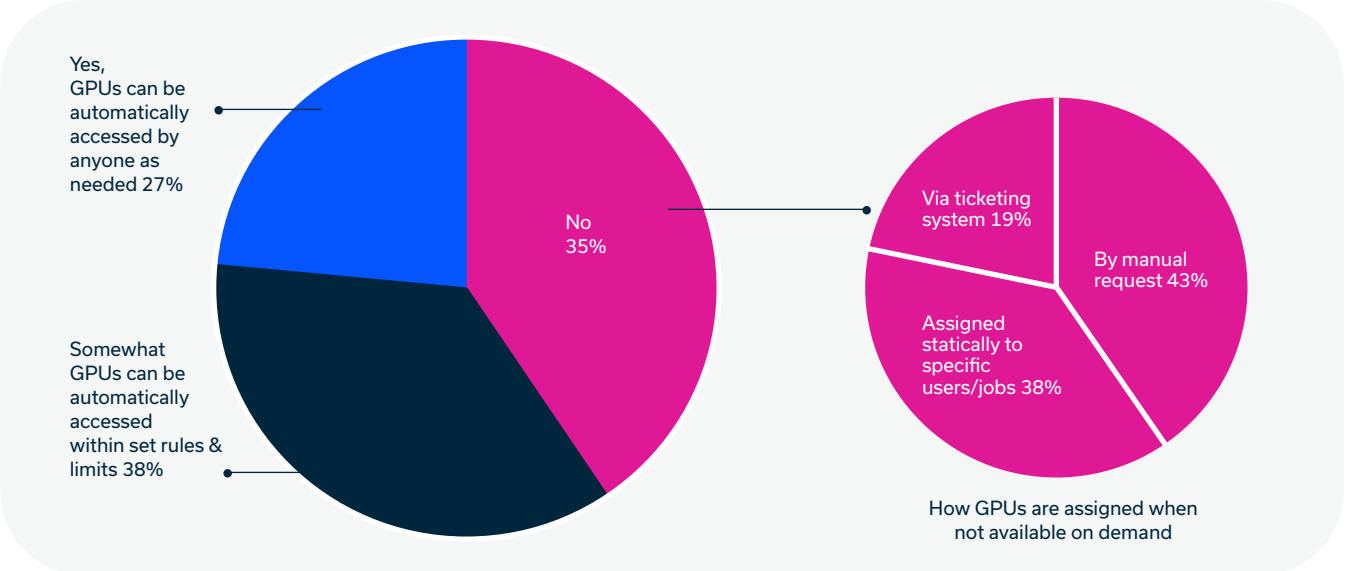


Figure 5: Do Research Teams Have On-Demand Access to GPU Compute?

GPU and AI Hardware Utilization and Resource Allocation Issues

Issues with GPU/compute resource allocation were reported by 87% of respondents, with 12% saying issues happen often (figure 7).

As a result, 83% of surveyed companies are not fully utilizing their GPU and AI hardware. In fact, almost two-thirds (61%) indicated their GPU and AI hardware are mostly at moderate utilization (figure 6).

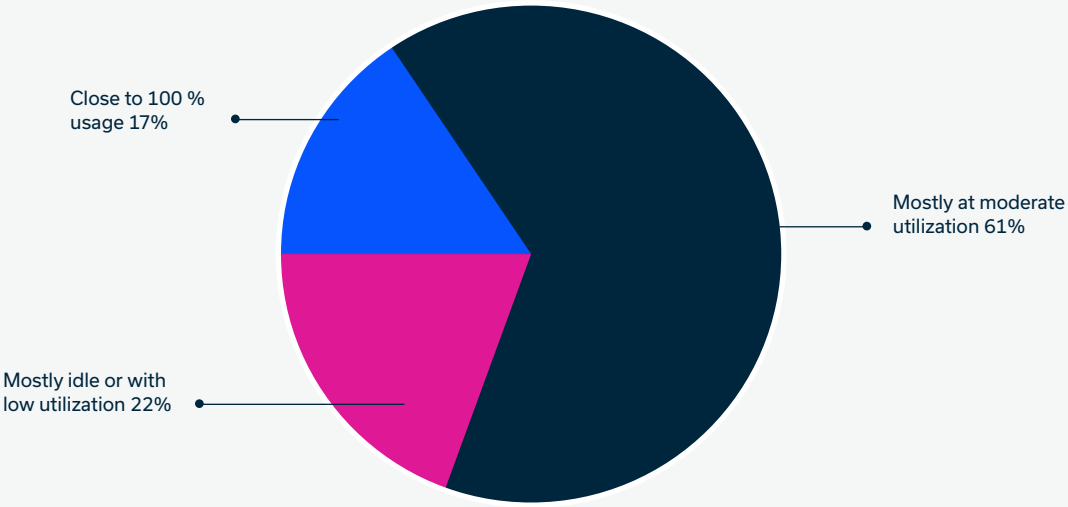


Figure 6: GPU and AI Hardware Utilization

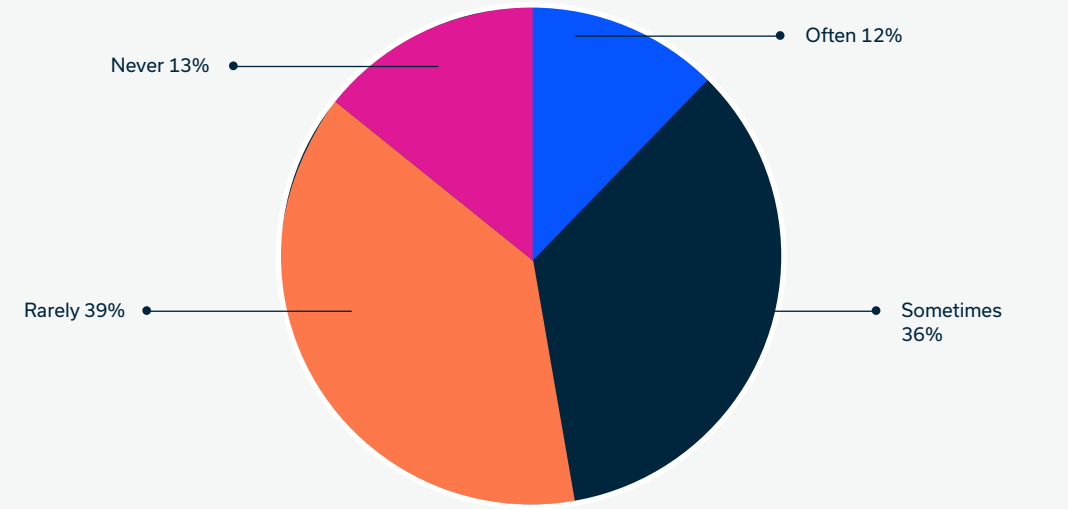


Figure 7: Frequency of Experiencing GPU/Compute Resource Allocation Issues

Companies of All Sizes Struggle with Hardware Utilization

At the high end, 38% of companies have an annual AI infrastructure budget of over $1 million (figure 8). When comparing budget against level of AI hardware utilization (figure 9), we see that companies with smaller budgets of up to $250k suffer the most from having their hardware sitting mostly idle.

Over a third (35%) of research teams access GPU compute only via additional steps or static assignment, and almost half of this group (43%) are subject to waiting for approval of their manual requests (figure 5). Every time they want to run a job, they need to make this manual request, slowing down operations and adding frustration and delay.

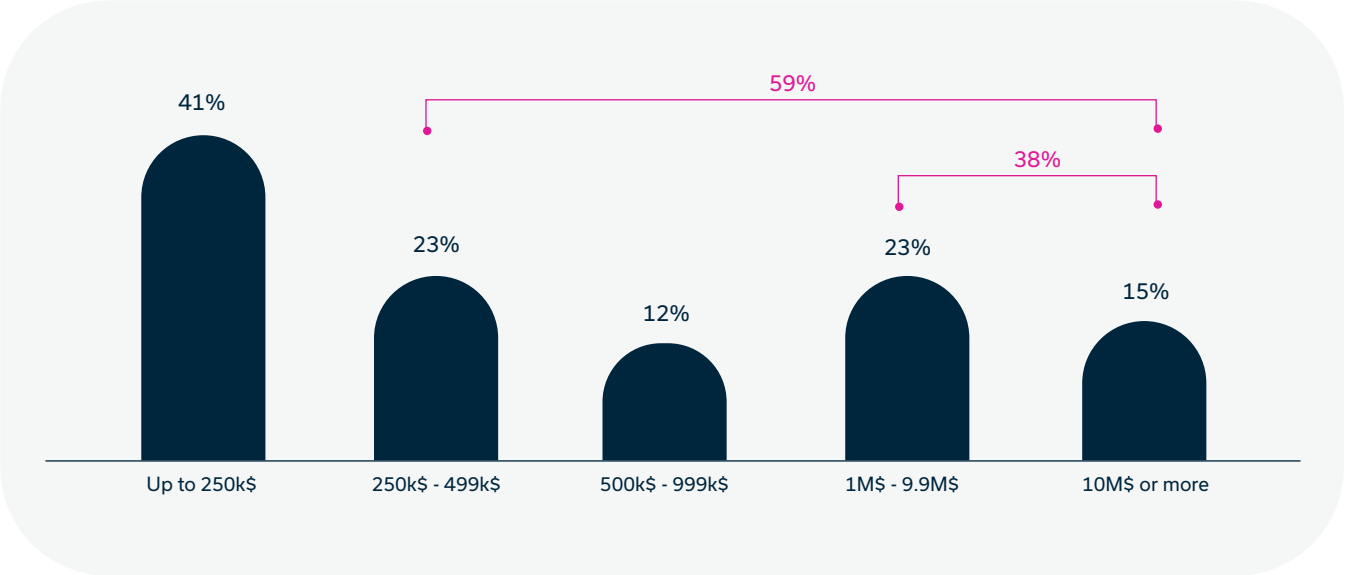


Figure 8: Annual AI infrastructure Budget (Hardware, Software, Cloud)

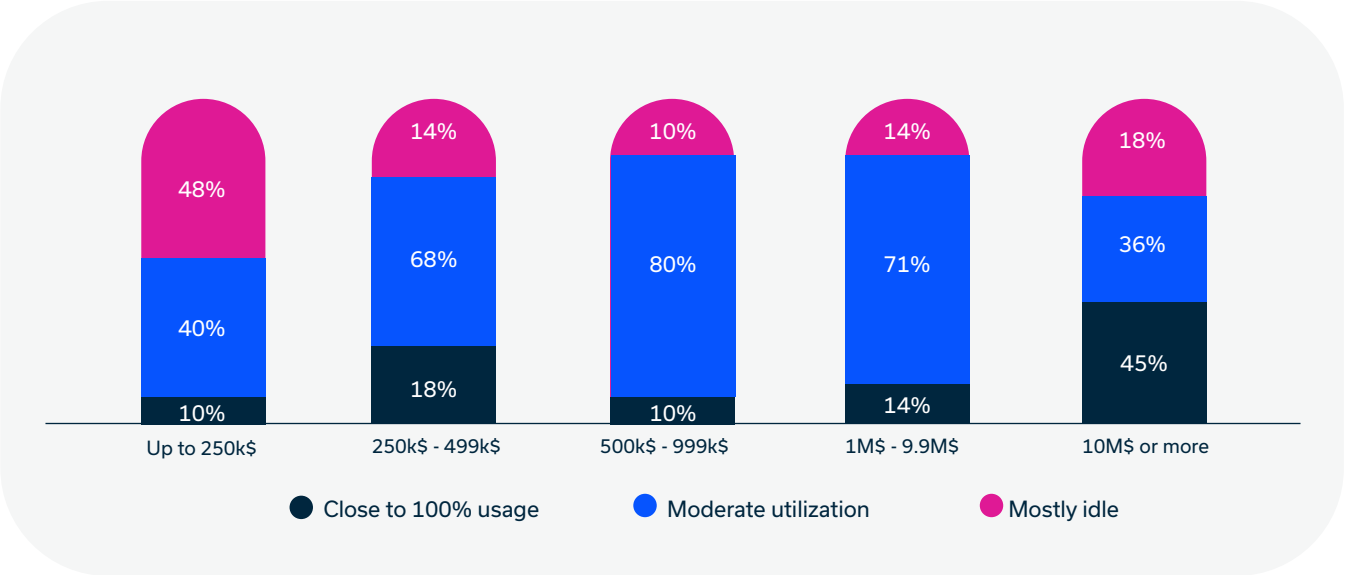


Figure 9: Annual AI infrastructure Budget by AI Hardware Utilization

Tools Used to Optimize GPU Allocation Between Users

The majority (72%) of companies are using one or more tools to optimize their GPU allocation between users. From home-grown solutions (23%), to Excel spreadsheets (16%), these are mostly low-tech solutions, especially when considering the investment that is

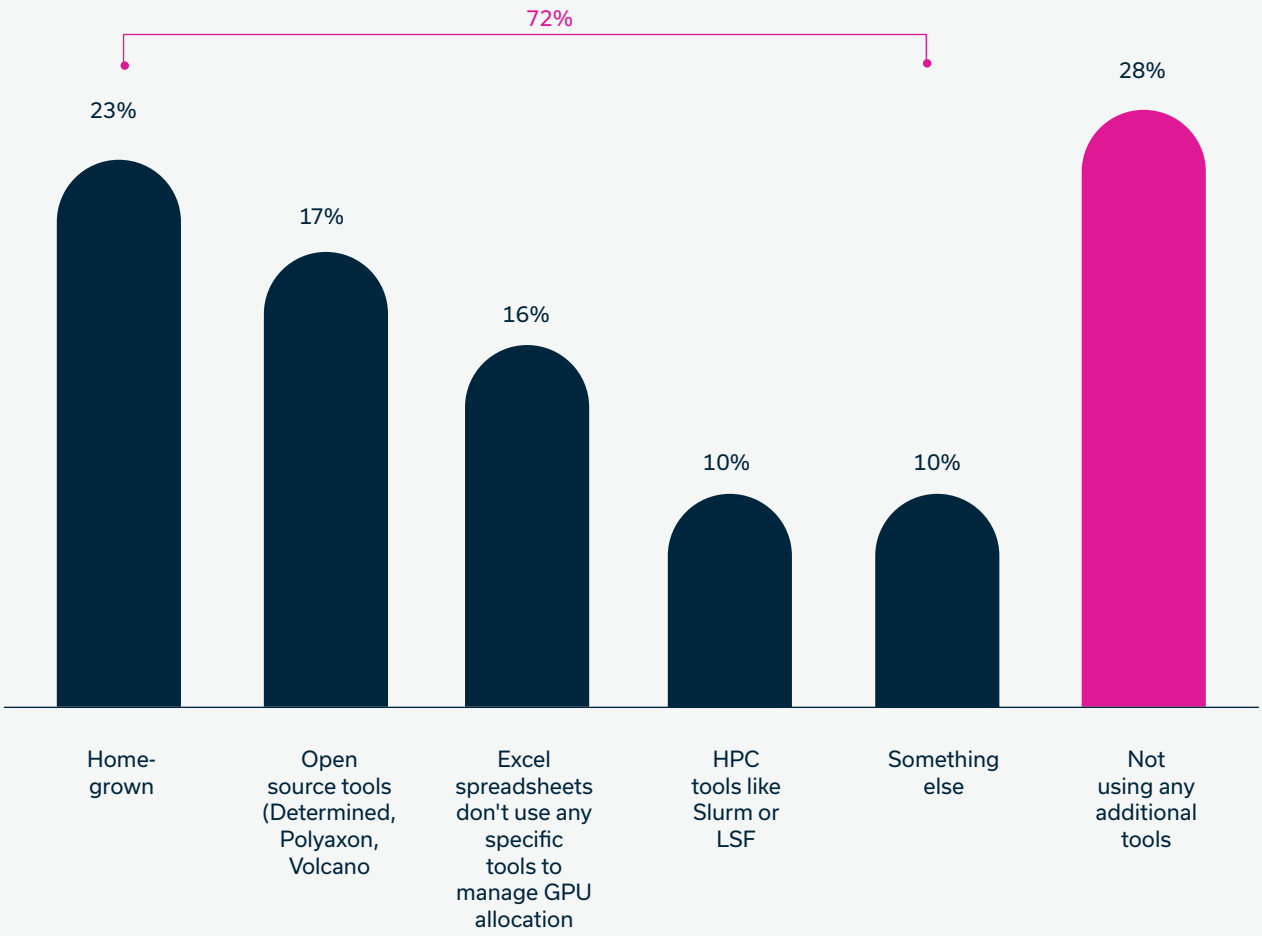


Figure 10: Tools Used to Optimize GPU Allocation Between Users

Containers and Kubernetes for AI Workloads

Containers are being used by 81% of companies for their AI workloads (figure 11) with Kubernetes ranking as the #1 container orchestration system, used by 42% of companies (figure 12).

These numbers show that AI is born in cloud-native infrastructure, and has a far greater adoption of cloud than the broader software world. Kubernetes is also ubiquitous among AI practitioners, with companies either using Kubernetes directly or leveraging managed K8s through a third-party. The use of orchestration tools shows that these companies are confident and mature in their use of containers.

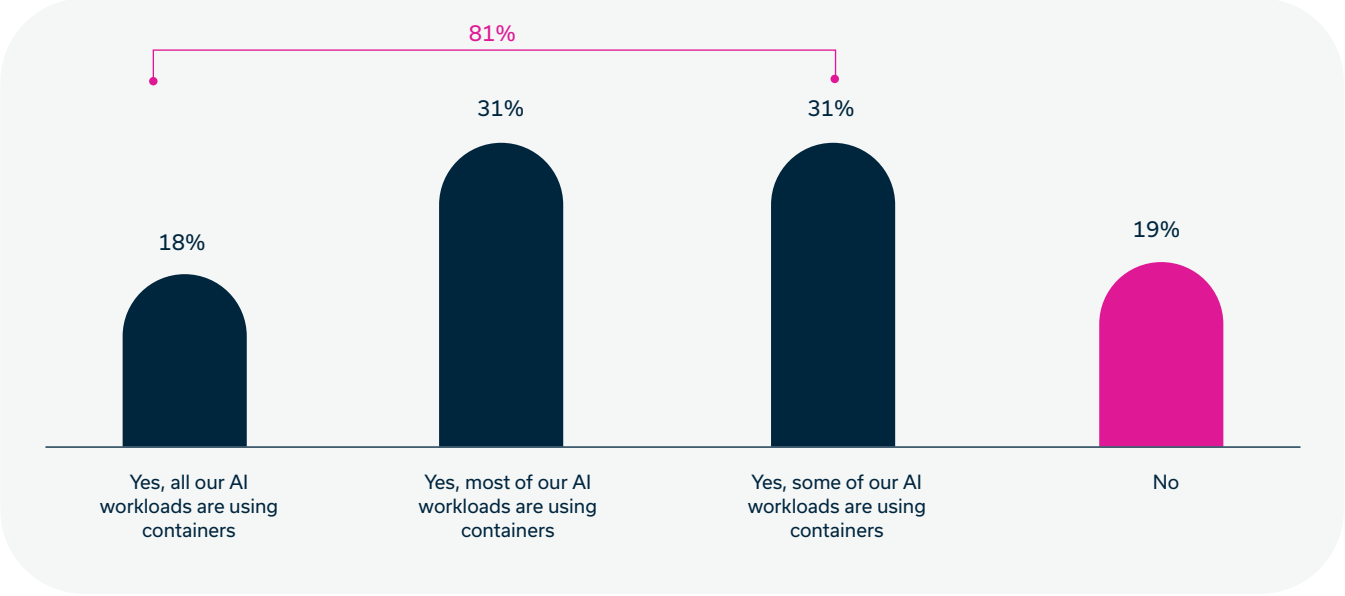


Figure 11: Use of Containers for AI Workloads

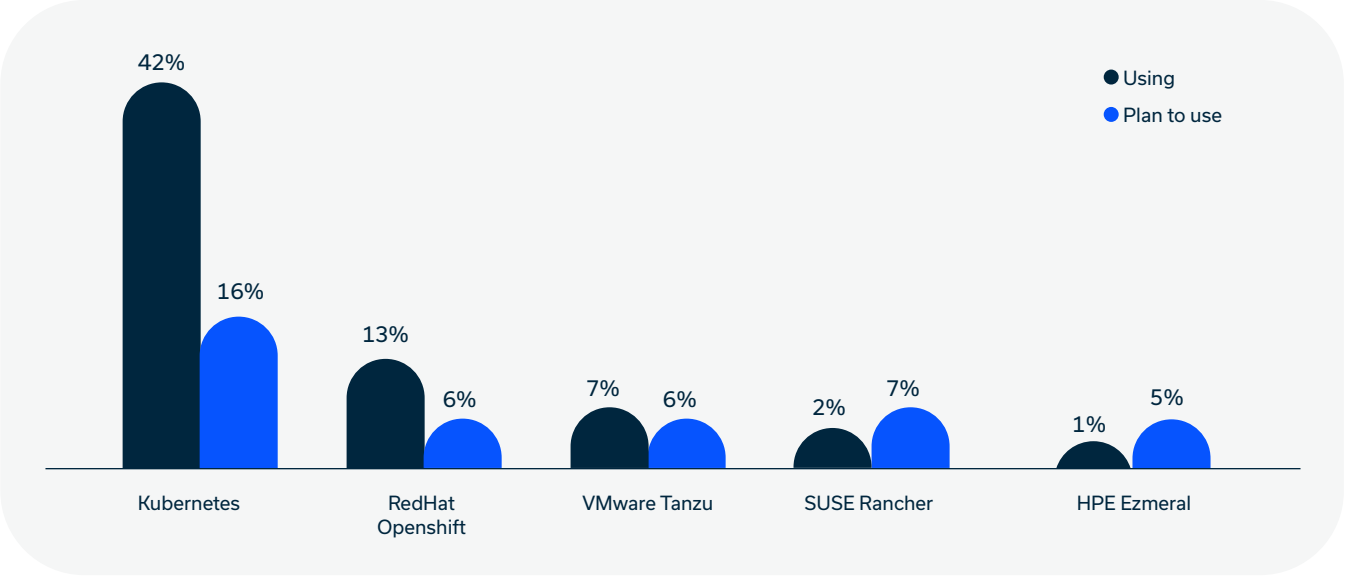


Figure 12: Container Orchestration Tools Used for AI

Big Plans for AI, Despite Multiple Challenges and Limited Confidence

Models Making it to Production

Less than half of AI models make it to production for 77% of surveyed companies.
Only 10% said 90% of their AI models make it to production.

This aligns with the common AI challenges reported in various media outlets. Getting models to production remains an issue of much debate and stymied innovation.

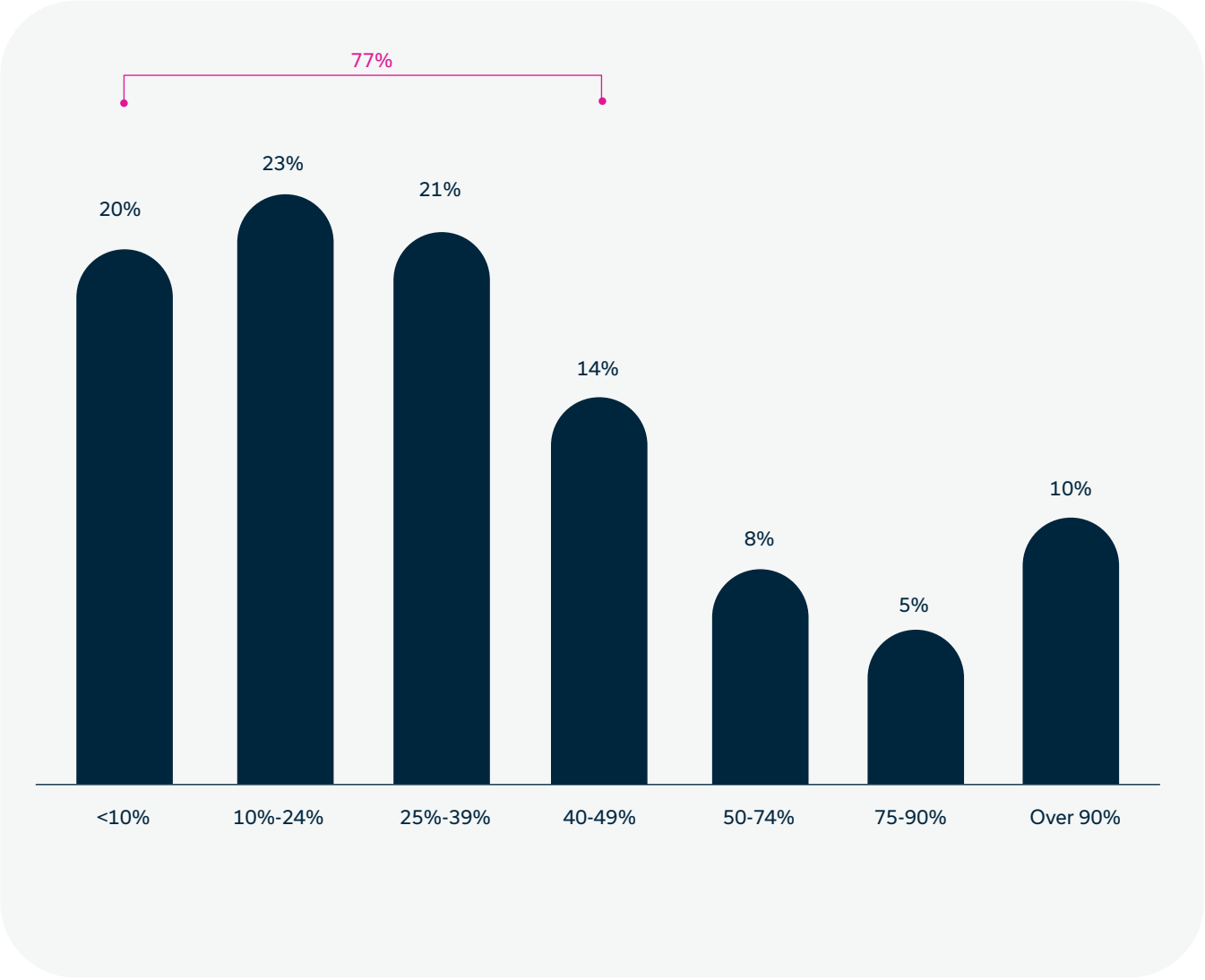


Figure 13: Models Making it to Production

Main Challenges for AI Development

An overwhelming 96% of companies admit to challenges when it comes to AI development.

The top three challenges are data-related (61%), infrastructure/compute related (42%), or related to defining business goals (36%).

These challenges crop up when getting started with AI and reflect the lack of market maturity.

Despite the fact that many companies have a $1M budget in place, they still aren't sure how to measure success or how to collect suitable data.

Infrastructure set-up is a significant challenge, as companies struggle with visibility and control.

In general, the larger the company size, the greater the challenges become

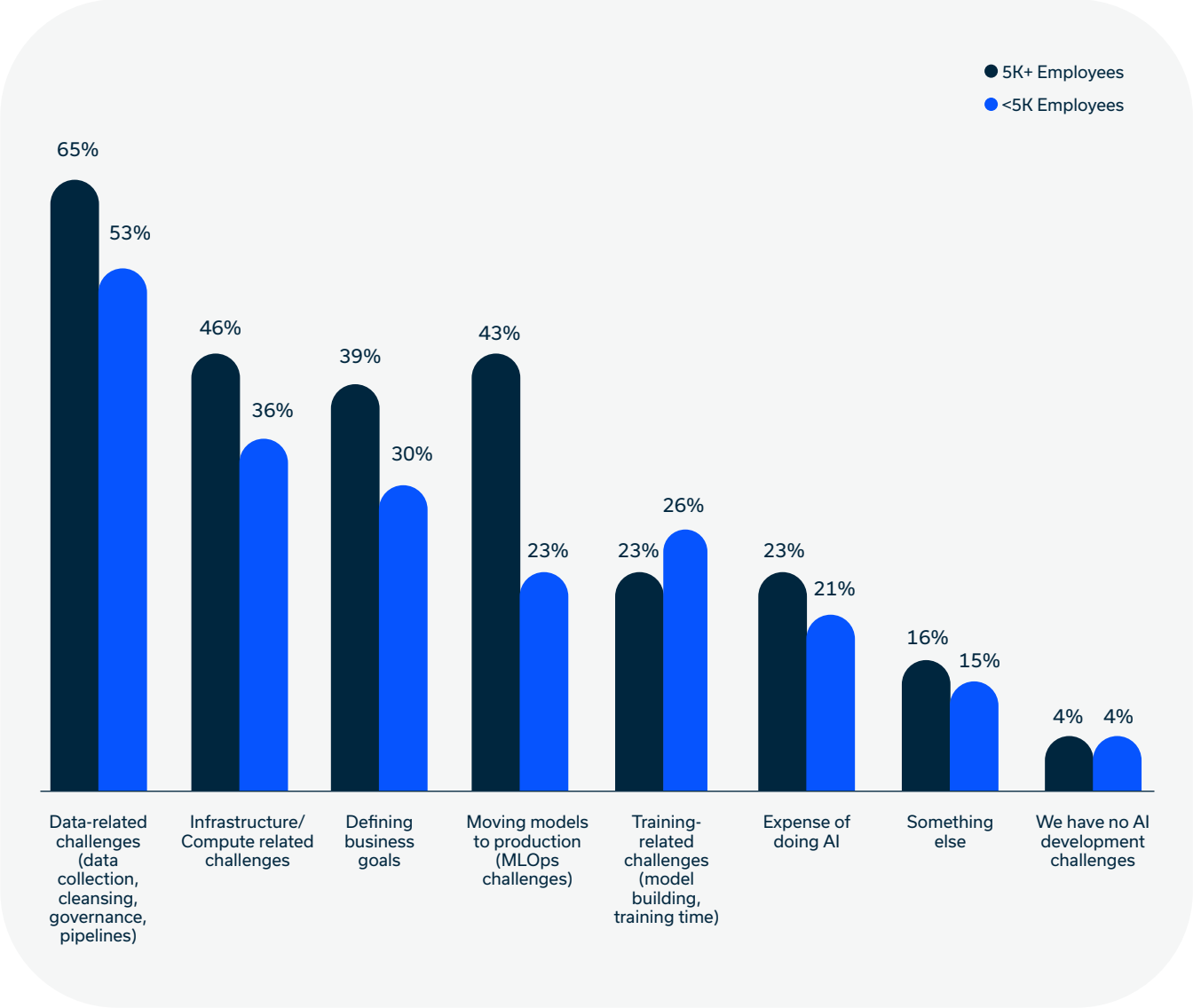


Figure 11: Use of Containers for AI Workloads

Plans to Increase GPU Capacity or Additional AI Infrastructure

Even with all the previously discussed challenges in place, 74% of companies are planning to increase their GPU capacity or AI infrastructure.

Companies are confident that they will ultimately be successful with AI, but will need to address their challenges in order to see a return on their investment.

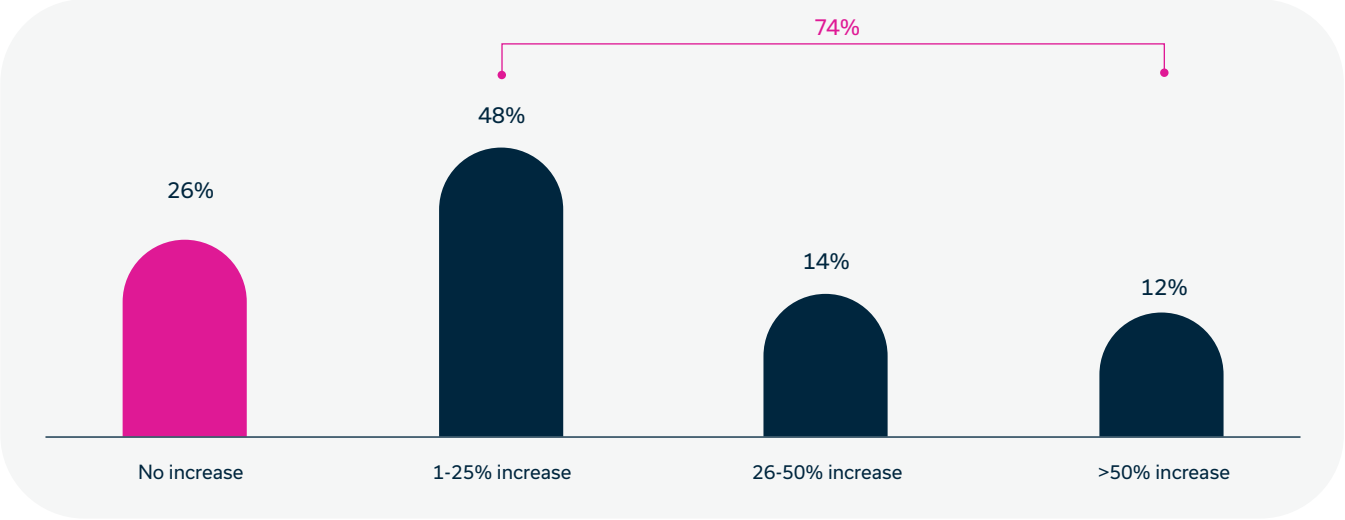


Figure 15: Plans to Increase GPU Capacity or Additional AI Infrastructure

Confidence in AI infrastructure Stack Set-up to Build, Train and Move

Only 18% of companies are fully confident that they have the right AI infrastructure stack to efficiently build, train and move ML models to production on time and on budget.

Despite a lack of confidence and multiple challenges, companies clearly feel a pressure to keep moving, investing more budget and expanding AI plans to keep up with the competition.

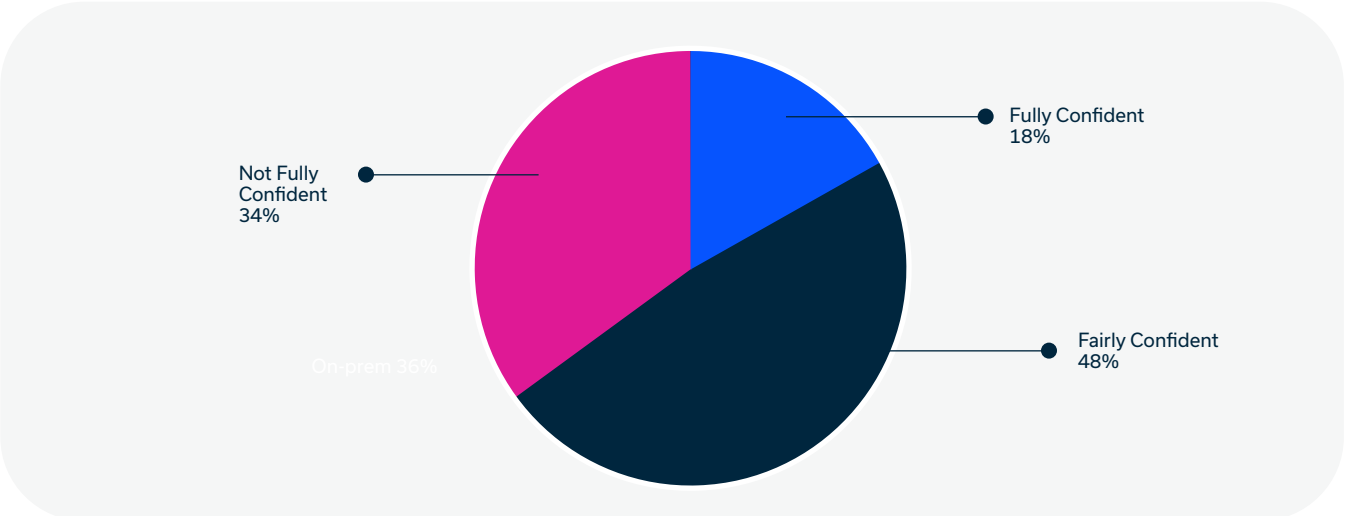


Figure 16: Confidence in AI infrastructure Stack Set-up

Country of Residence

Figure 17: Country of Residence

Company Size, Job Functions, Seniority and Industry

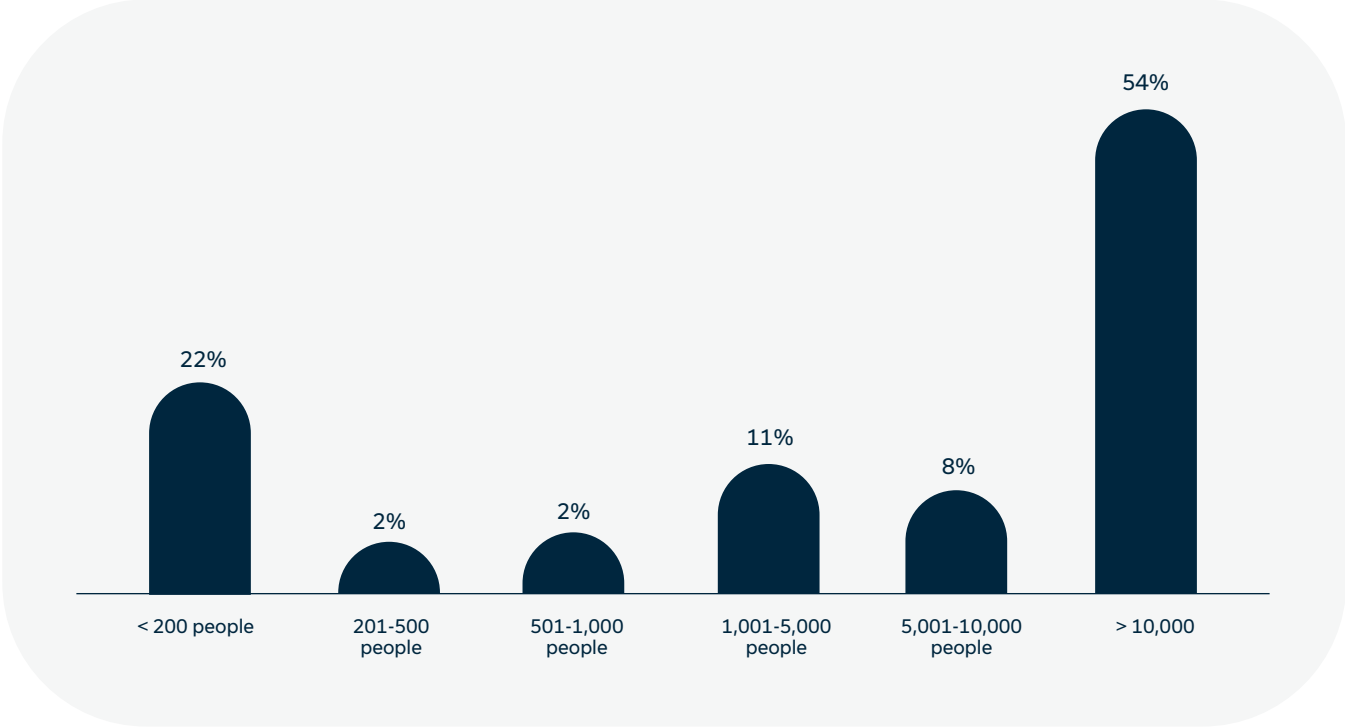


Figure 18: Company Size

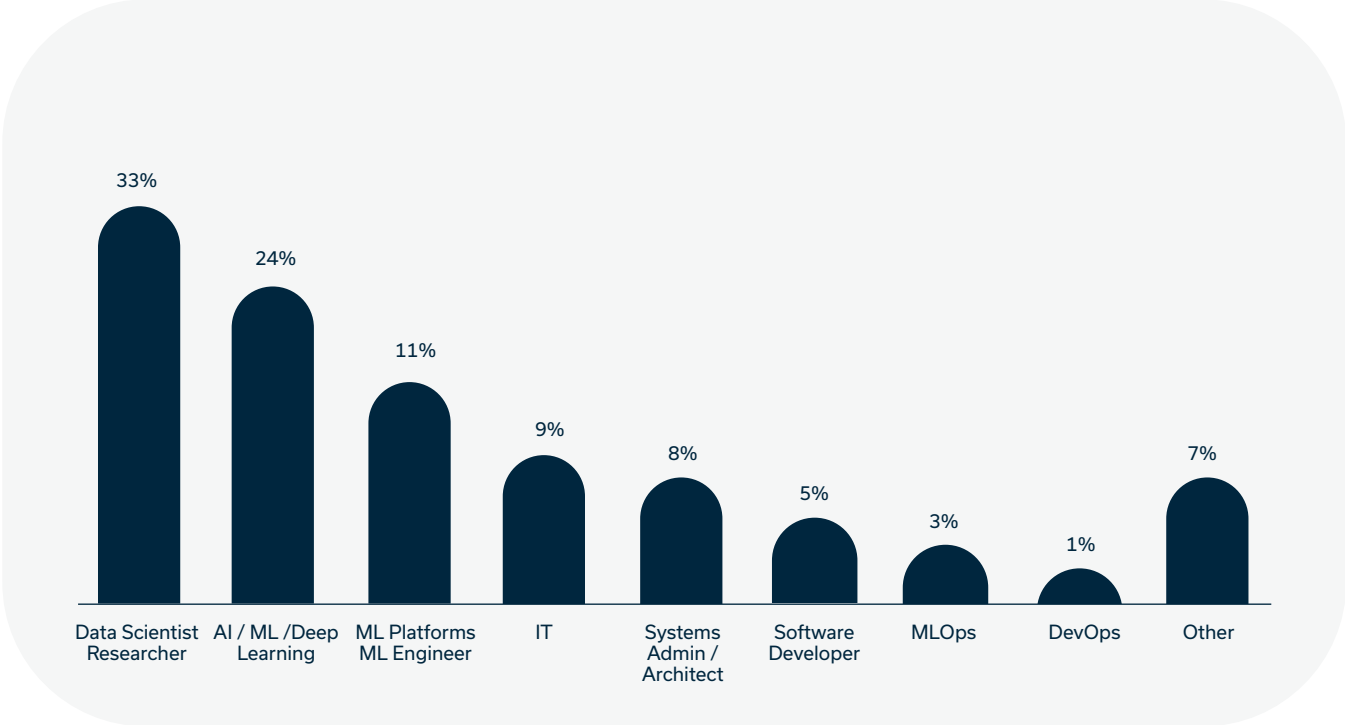


Figure 19: Job Function

Company Size, Job Functions, Seniority and Industry

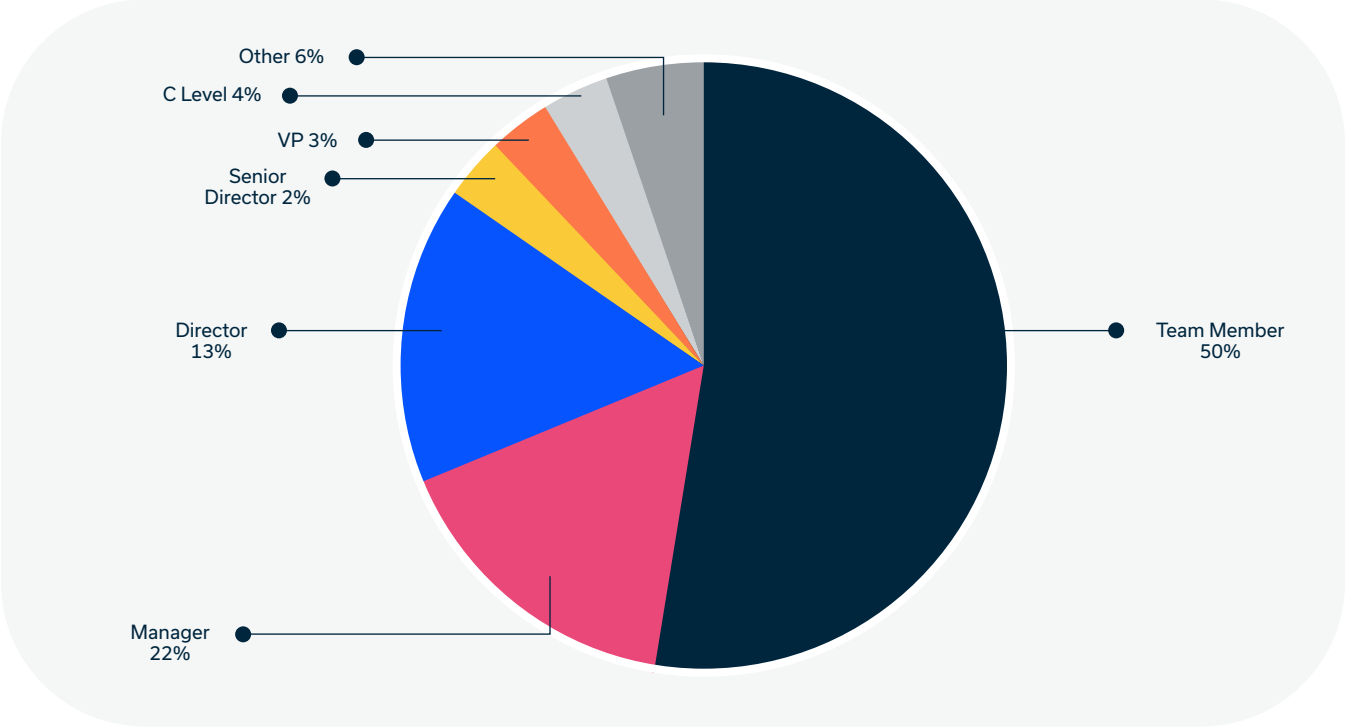


Figure 20: Job Seniority

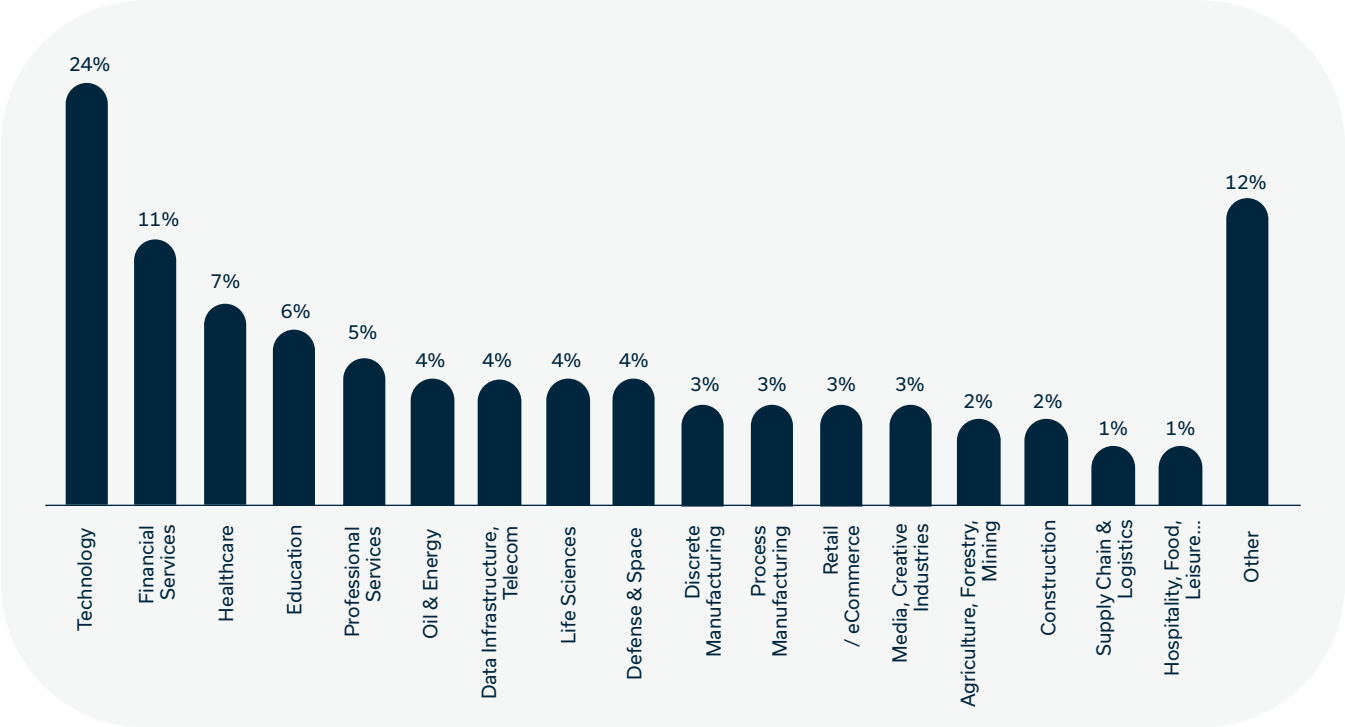


Figure 21: Industry