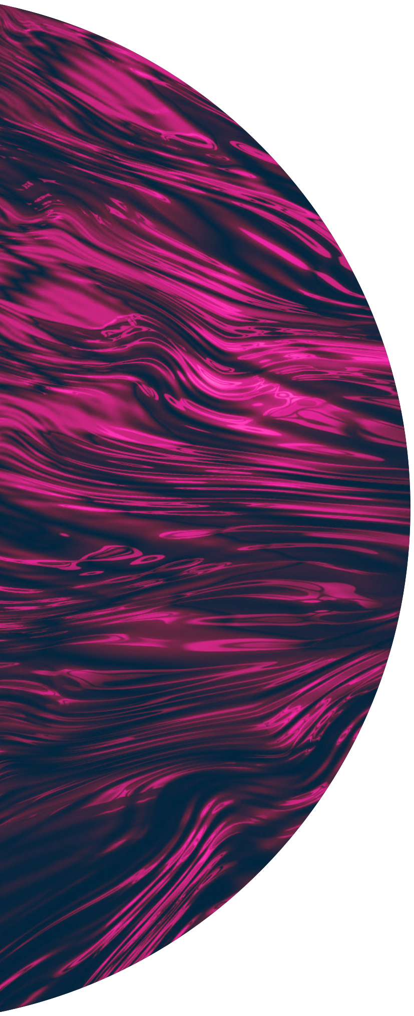
eBook | Jan 2023

The 2023 State of AI Infrastructure Survey



www.run.ai

Table of Contents

Introduction and Key Findings

Introduction

The artificial intelligence (AI) industry has grown rapidly in recent years, as has the need for more advanced and scalable infrastructure to support its development and deployment. The global AI Infrastructure Market was valued at $23.50 billion in 2021 and is expected to reach $422.55 billion by 2029, at a forecasted CAGR of 43.50% between 2022-2029.

One of the main drivers of progress in the AI infrastructure market has been increasing awareness among enterprises of how AI can enhance their operational efficiency, attract new business and grow new revenue streams, while reducing costs through the automation of process flows. Other drivers include the adoption of smart manufacturing processes using AI, blockchain and IoT technologies, the increased investment by GPU/CPU manufacturers in the development of compute-intensive chips, and the rising popularity of chatbots, like OpenAI's recently launched ChatGPT, for example.

The current hype around AI has given rise to a renewed focus on getting it into the enterprise, and organizations are increasingly eager to start using and developing AI applications themselves. But with an abundance of new AI infrastructure tools flooding the rapidly evolving industry, it is akin to a sort of technological "Wild West", with no real best practices for enterprises to follow as they get AI into production. As they begin to invest more heavily in AI, there's a lot riding on how they decide to build their infrastructure and service their practitioners.

This is the second 'State of AI Infrastructure' survey we are running, due to all the new activity in the industry and new AI companies in the AI space, we're keen to see what's changed. We're particularly interested in new insights into how organizations are approaching the build of their AI infrastructure, why they are building it, how they are building it, what are the main challenges they face, and how the abundance of different tools has affected getting AI into production. We hope that the insights from this survey will be helpful to those who both build and use AI infrastructure.

Methodology

To get more insight into the current state of AI Infrastructure, we commissioned a survey of 450 Data, Engineering, AI and ML (Machine Learning) professionals from a variety of industries. This report was administered online by Global Surveyz Research, an independent global research firm. The survey is based on responses from a mix of Data Scientists, Researchers, Heads of AI, Heads of Deep Learning, IT Directors, VPs IT, Systems Architects, ML Platform Engineers and MLOps, from companies across the US and Western EU ranging in size between under 200 and over 10,000 employees. The respondents were recruited through a global B2B research panel and invited via email to complete the survey, with all responses collected during the second half of 2022. The average amount of time spent on the survey was five minutes and fifty seconds. The answers to the majority of the non-numerical questions were randomized, in order to prevent order bias in the answers.

Key Findings

1

Data has been overtaken by Infrastructure and Compute as the main challenges for AI development

A whopping 88% of survey respondents admitted to having AI development challenges (Figure 1), which is telling in itself. But it's also interesting to note that Data, which was ranked by 61% of respondents in last year's survey as their main challenge in AI development, was overtaken this year by infrastructure (i.e., the different platforms and tools that comprise "the stack"), and compute (i.e., getting access to GPU resources, not having to wait for resources, etc.) – chosen by 54% and 43% of respondents respectively as their main challenges. This year, Data ranked as the third biggest challenge in AI development (41%). The fact that infrastructure and compute-related challenges are now the top concern for companies reinforces the importance of building the right foundation, for the right stack, to get the most out of their compute.

2

The more GPUs, the bigger the reliance on multiple third-party tools

As organizations scale and require more GPUs, the more complex it has become to build the right AI infrastructure to get the right amount of compute to all of the different workloads, tools, and end users. 80% of companies are now using third-party tools, and the more GPUs they require, the bigger their reliance on multiple third-party platforms, increasing from 29% reliance in companies with less than 50 GPUs, to 50% reliance in companies with more than 100 GPUs (Figure 2). What would make more sense, is a more open, middleware approach, where organizations can use different tools that run on the same infrastructure, so that they are not locked into one end-to-end platform.

Key Findings

3

On-demand access to GPU compute is still very low, with 89% of companies facing resource allocation issues regularly

Only 28% of the respondents have on-demand access to GPU compute (Figure 8). When asked how GPUs are assigned when not available via on-demand, 51% indicated they are using a ticketing system (Figure 9), suggesting that on-demand access is still lacking. So, it's no wonder that 89% of respondents face allocation issues regularly (Figure 10) – even though some of them (58%) claim to have somewhat automatic access – with 40% facing those GPU/Compute resource allocation issues weekly.

4

In 88% of companies, more than half of AI/ML models never make it to production

While most companies are planning to grow their GPU capacity or other AI infrastructure in the coming year, for 88% of them (compared with 77% in last year's survey), more than half their AI/ML models don't make it to production. On average, only 37% of AI/ML models are deployed in production environments (Figure 13). The main impediments to actually deploying (Figure 14) include scalability (47%), performance (46%), technical (45%), and resources (42%). The fact that they were all mentioned as a "main impediment" by such a substantial portion of the respondents, shows that there isn't just one glaring impediment to model deployment, but rather a multi-faceted one.

5

91% of companies are planning to grow their GPU capacity or other AI infrastructure in the next 12 months

The vast majority (91%) of companies are planning to grow their GPU capacity or other AI infrastructure by an average of 23% in the next 12 months (Figure 4) despite the uncertainty of the current economic climate. Organizations won't invest in AI unless they can actually get value out of it, so this result is a resounding testament to the fact that most companies see huge potential and value in continued investment in AI.

Survey Report Findings

Challenges for AI Development

When asked what their company's main challenges are around AI development, 88% of respondents admitted to having AI development challenges.

The top challenges are infrastructure related challenges (54%), compute related challenges (43%), and data related challenges (41%).

It's interesting to note that infrastructure and compute have overtaken Data as the biggest challenges.

This reinforces the importance of building the right foundation, for the right stack, to get the most out of your compute.

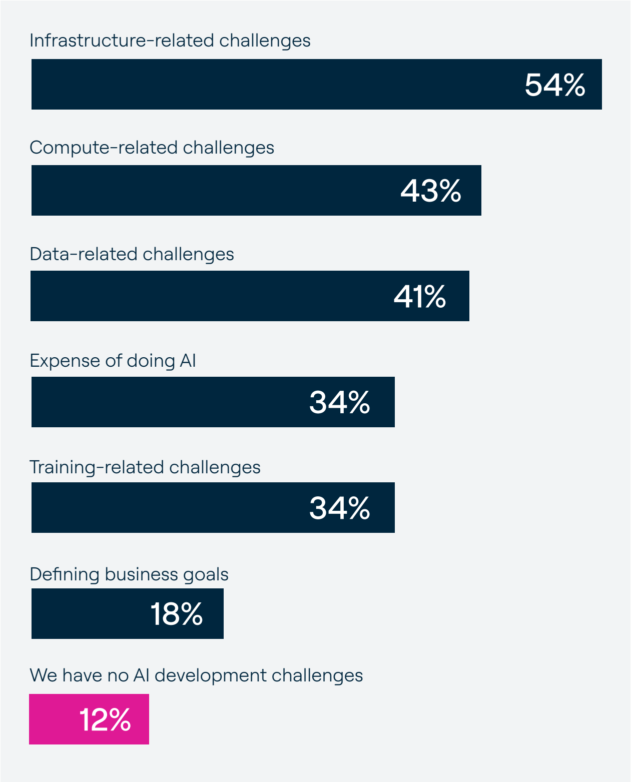


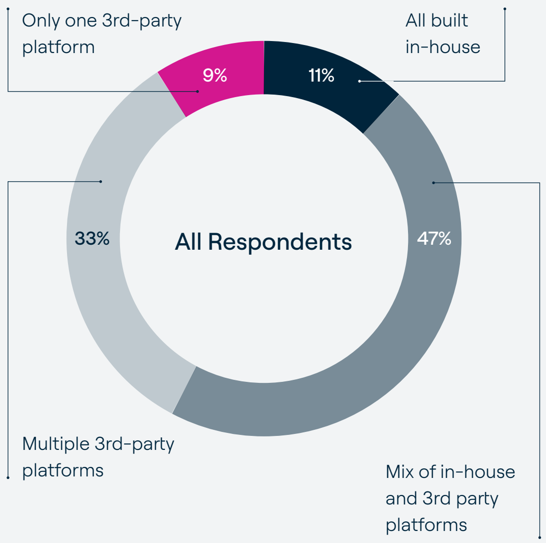
Figure 1: Challenges for AI development

AI/ML Stack Architecture

When asked how their AI/ML infrastructure is architected, 11% of respondents said it is all built in-house, while 47% have a mix of in-house and third-party platforms. We also saw that the use of multiple third-party platforms grows with the number of GPUs (29% for those with < 50 GPUs, and up to 50% for those with 100+ GPUs). This confirms that the practice of taking AI into production and streamlining it (MLOps) isn't a one-size-fits-all process.

Organizations are using a mix of different tools to build their own best-of-breed platforms to support their needs (and those of their users).

The fact that the state of AI infrastructure appears to be somewhat chaotic, with an abundance of tools and no real best practices, is also testament to the growing need among organizations for multiple platforms to meet their various AI development needs, giving rise to new technologies, and new types of users and applications. But this could also overwhelm infrastructure resources, so the more GPUs companies have, there's also an increasingly urgent need for a unified compute layer that supports all these different tools to make sure that the volume of resources and the way they are accessed are aligned to specific end users and the different tools they need.



- ○ All built in-house
- ○ Mix of in-house and 3rd party platforms
- ○ Multiple 3rd-party platforms
- ○ Only one 3rd-party platform

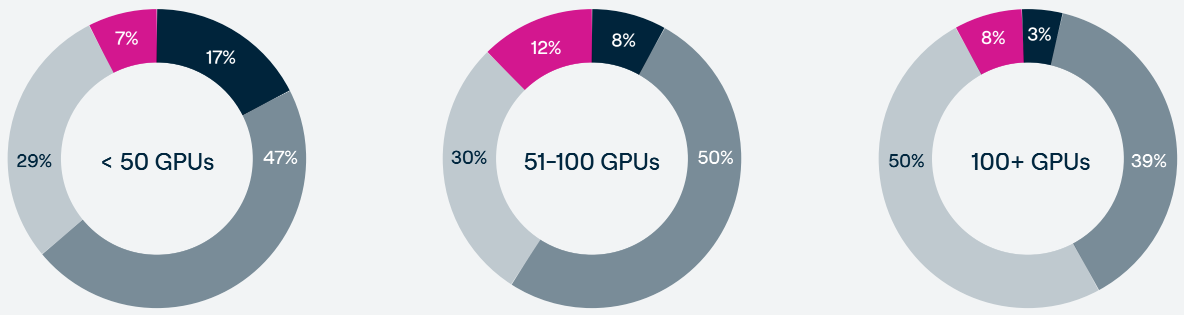


Figure 2: AI/ML stack architecture.

Tools Used to Optimize GPU Allocation Between Users

According to the survey results, basically everyone (99% of respondents) is using tools to optimize GPU allocations between users, with open-source being the most popular choice (36%), followed by home-grown (23%).

Interestingly, the top two tools are also the most challenging for organizations when taking AI into production, because they are both very brittle, indicating there is room for more than half (59%) of companies to move to a more professional way of optimizing GPU allocations between users.

The fact that 73% are using open source, home-grown tools or Excel sheets, also shows that organizations are obviously facing a lot of issues allocating GPU resources, so it appears that despite plenty of options, there is still no clear or definitive way to optimize.

With the majority of respondents still using tools that are not enterprise-grade, these tools will require attention, especially as their organizations scale and need to take more AI into production.

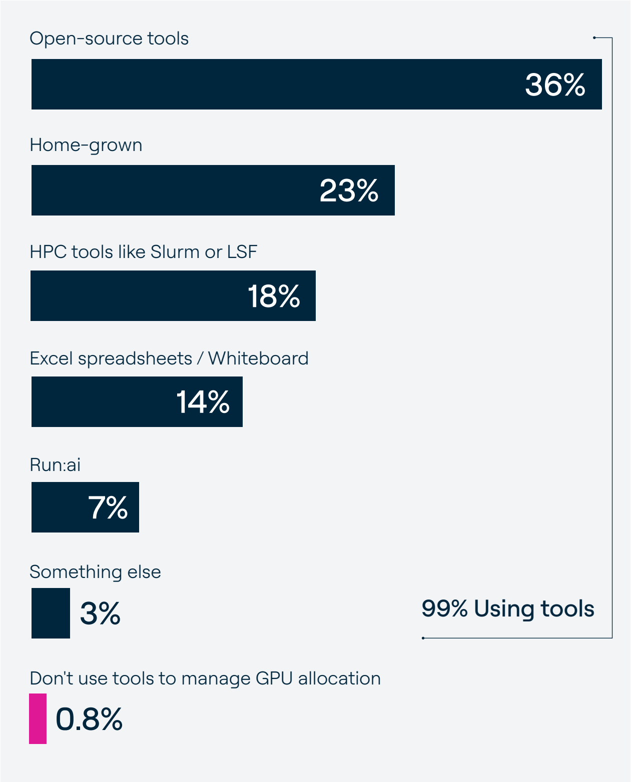


Figure 3: Tools used to optimize GPU allocation between users

Plan to Grow GPU Capacity in the Next 12 Months

It's interesting to see that the vast majority of companies (91%) are planning to grow their GPU capacity or other AI infrastructure by an average of 23% in the next 12 months – despite the uncertainty surrounding the current economic climate.

Organizations won't invest in AI unless they can actually get value out of it, so this slide shows they're still definitely seeing value in continued investment in AI.

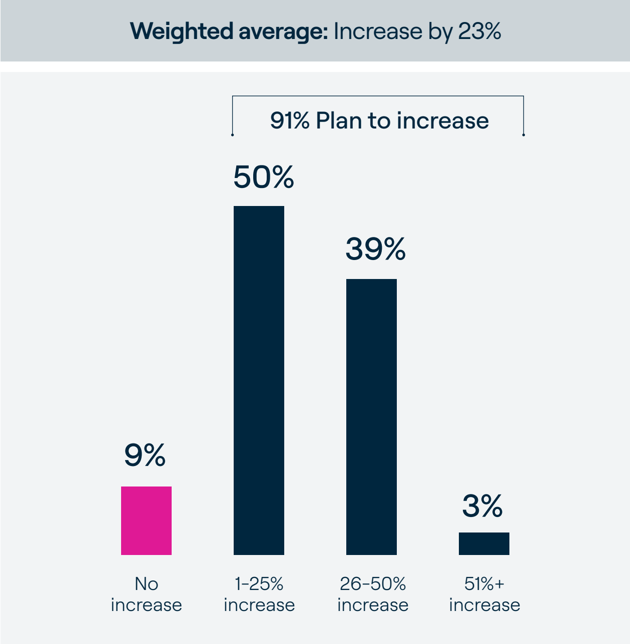


Figure 4: Plan to grow GPU capacity in the next 12 months.

Aspects of AI Infrastructure Planned for Implementation (within 6–12 months)

When asked what aspect of AI infrastructure they are looking to implement in the next 6–12 months, respondents indicated that the top aspects relate to challenges in production, including monitoring, observability, and explainability (50%), model deployment and serving (44%), and orchestration and pipelines (34%), so it appears that their answers are focused more on AI models in production and less about model development.

Even the aspect least indicated for planned implementation was mentioned by 19% or respondents, which is still a respectable number, so with all of the options provided indicated as important, it shows that there's "a lot going on": there's no one or two glaringly obvious priorities, but rather multiple aspects of AI infrastructure that companies need to focus on to get their AI into production faster. Many of these aspects involve MLOps, which is interesting, because each aspect requires different tools, and therefore a solid computing platform to plug all of these different tools into.

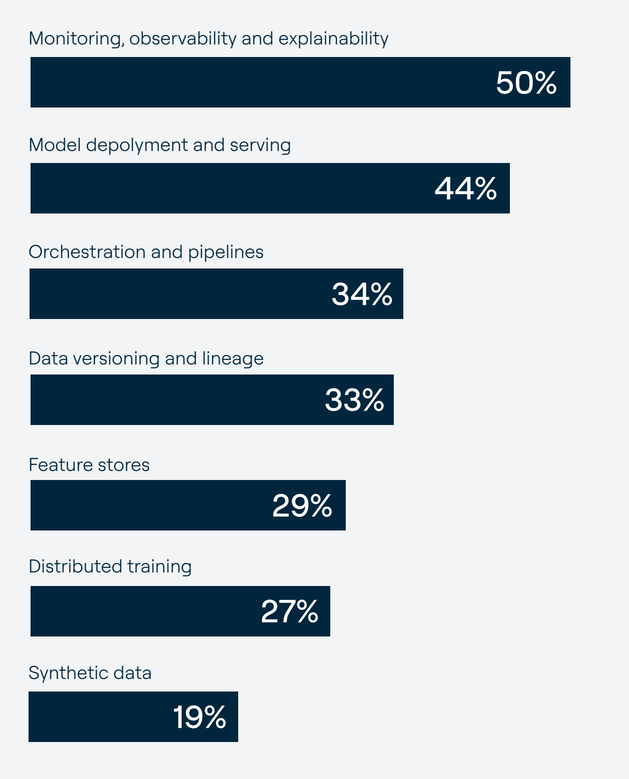


Figure 5: Aspects of AI infrastructure planned for implementation (within 6–12 months).

Measurement of AI/ML Infrastructure Success

When asked how they are measuring the success from their AI/ML infrastructure, 28% of respondents said they are monitoring their reach to new costumers, 21% are measuring their increase in revenue, and 17% measure success by how much time they are saving.

The top measures of success indicated by respondents represent benefits that are both internal and external (both to the companies and their end users), and clearly demonstrate that they are investing in AI because they want to create value, which is especially important during an uncertain economic climate.

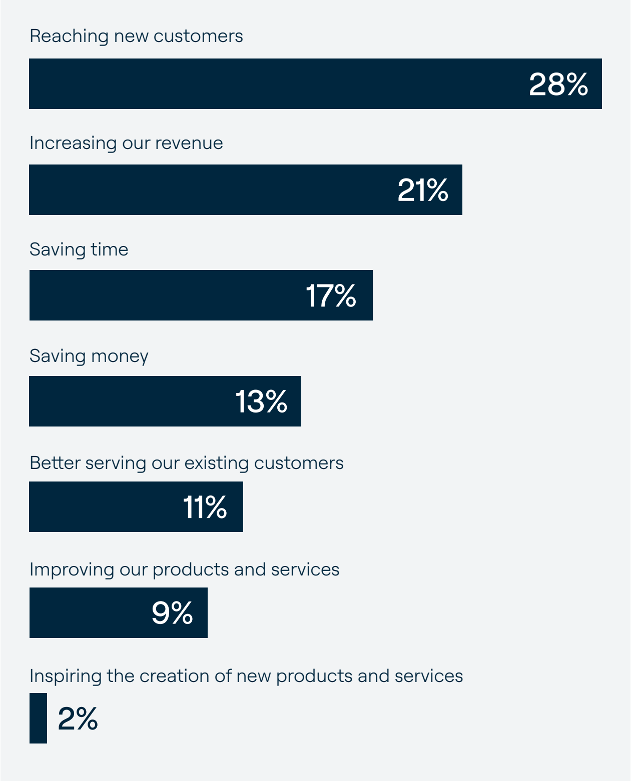


Figure 6: Measurement of AI/ML infrastructure success.

Tools Used to Monitor GPU Cluster Utilization

The top tools used to monitor GPU cluster utilization are NVIDIA-SMI (86%), GCP-GPU-utilization-metrics (53%), and ngputop (49%).

It's very hard to visualize problems with utilization, and therefore very important for companies to get insight into how they are utilizing their GPUs.

But with most respondents using NVIDIA-SMI (86%), GCP-GPU-utilization-metrics (53%), and ngputop (49%) to monitor GPU cluster utilization, it appears it's just as difficult finding the right tool to better understand their GPUs utilization, because they all show metrics from one platform, or even one host, and don't provide a broad overview (or 'the optimal view') of an organization's GPUs utilization across its entire infrastructure. They only provide a 'current snapshot' of a small subset of the infrastructure.

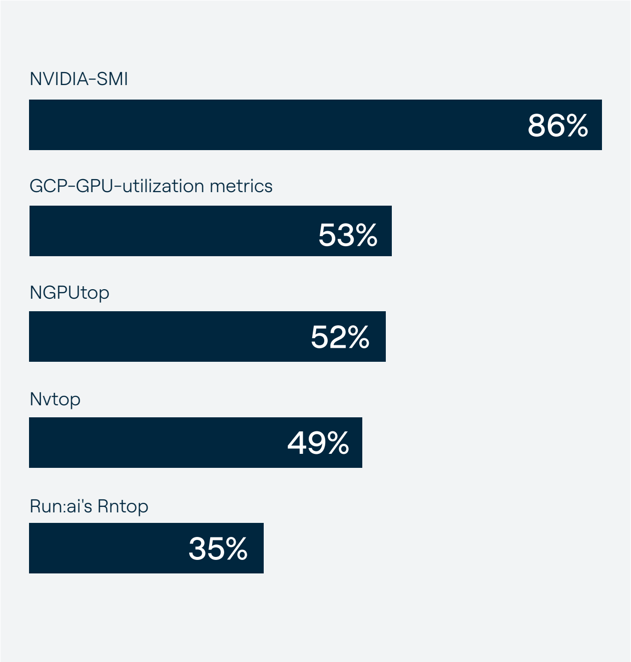


Figure 7: Tools used to monitor GPU cluster utilization.

On-Demand Access to GPU Compute

When asked about availability of on demand access to GPU compute, only 28% of the respondents said they have on-demand access (figure 8).

When asked how GPUs are assigned when not available via on-demand, 51% of the respondents indicated they are using a ticketing system (figure 9), suggesting that there's still no good on-demand access yet. It's no wonder, therefore, that 89% of respondents face allocation issues regularly (Figure 10), even though some of them (58%) claim to have somewhat automatic access.

Only 11% said they rarely run GPU allocation issues, 13% have allocation issues daily, and 40% face allocation issues weekly.

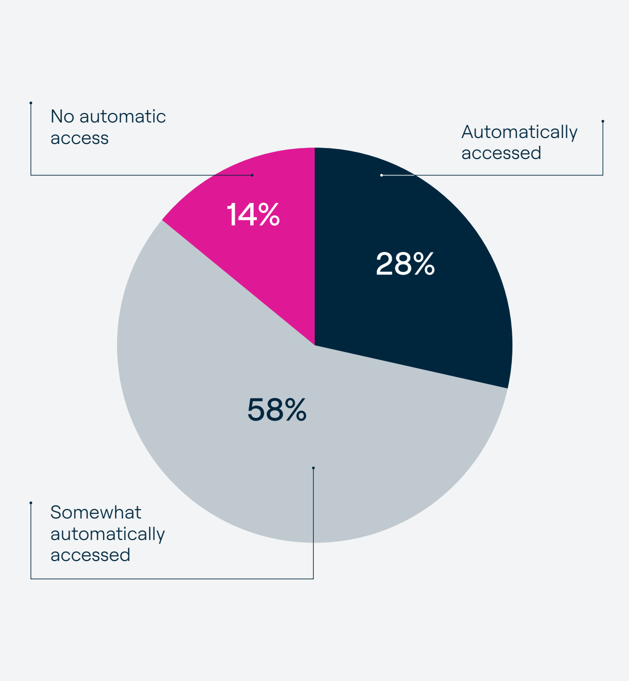


Figure 8: Availability of on-demand access to GPU compute.

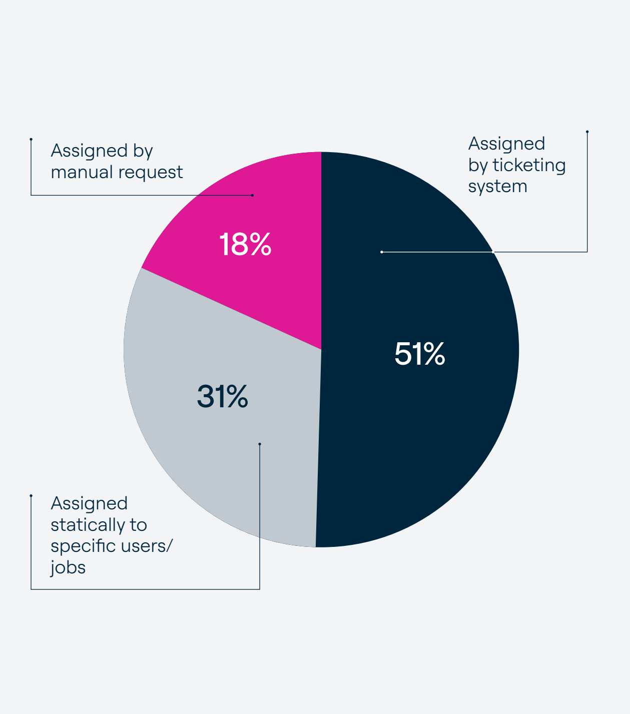


Figure 9: GPUs assignment w/o on-demand access.

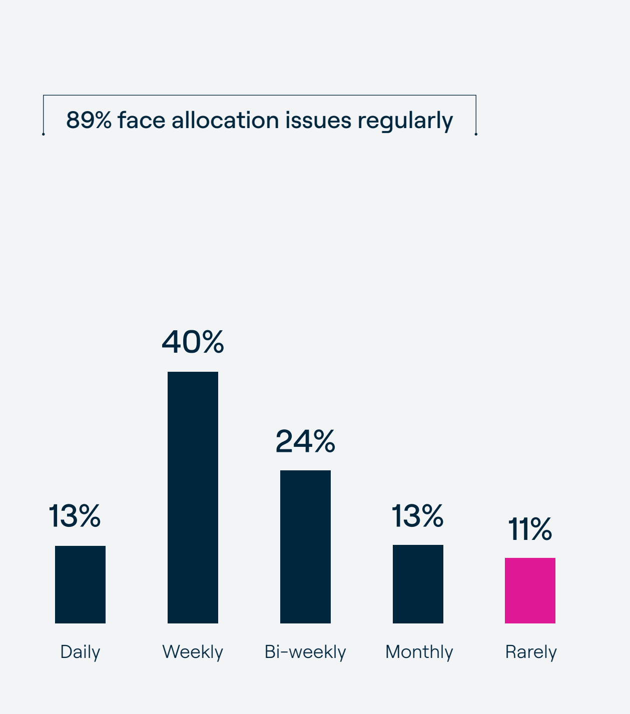


Figure 10: Frequency of GPU/compute resource allocation issues.

Plans to Move AI Applications and Infrastructure to the Cloud

51% of the respondents already have their applications and infrastructure on the cloud, and 33% said they were planning to move it to the cloud by the end of 2022.

When deep diving to see how those already in the cloud differ based on their level of automatic access to their GPU, we saw that of the 51% of companies already on the cloud, the highest level of adoption (78%) is by companies without automatic access.

All respondents indicated that they are either already on the cloud or planning to move to the cloud either this or next year, but according to these results, it doesn't necessarily solve their access to the GPU problem, which is something that those who haven't moved to the cloud yet should be aware of.

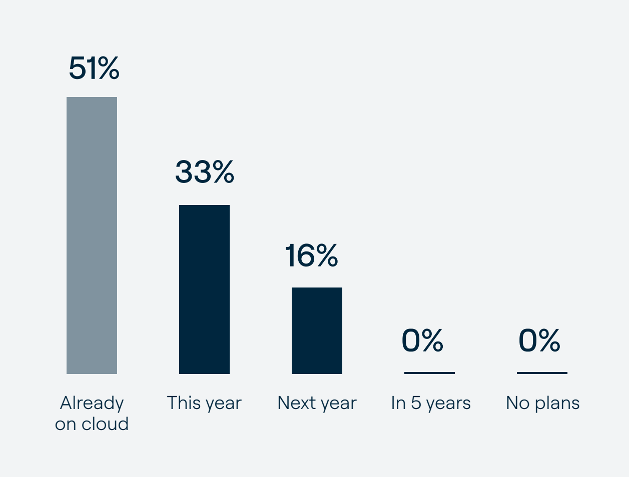


Figure 11: Plans to move AI Applications and infrastructure to the cloud.

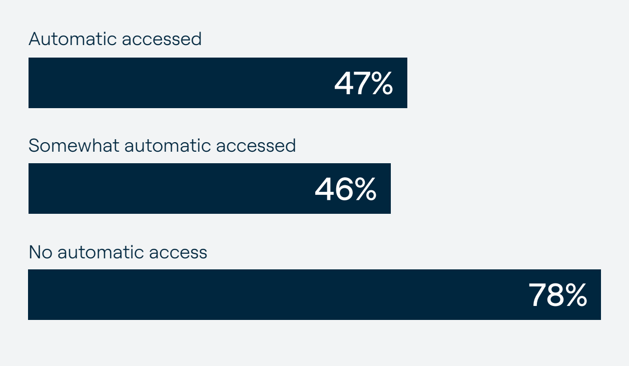


Figure 12: "Already on cloud" by access to GPU.

Percentage of AI/ML Models Deployed in Production

On average, 37% of AI/ML models are deployed in production.

88% of respondents (compared with 77% in last year's survey) are deploying less than 50% of the models, indicating that it's even harder now to get things into production, not just because of infrastructure considerations but also business and organizational ones as well.

It's also why companies are investing so heavily in a variety of different tools (as opposed to just one) to get more AI into production, but as these results show, there's still a lot of room for improvement.

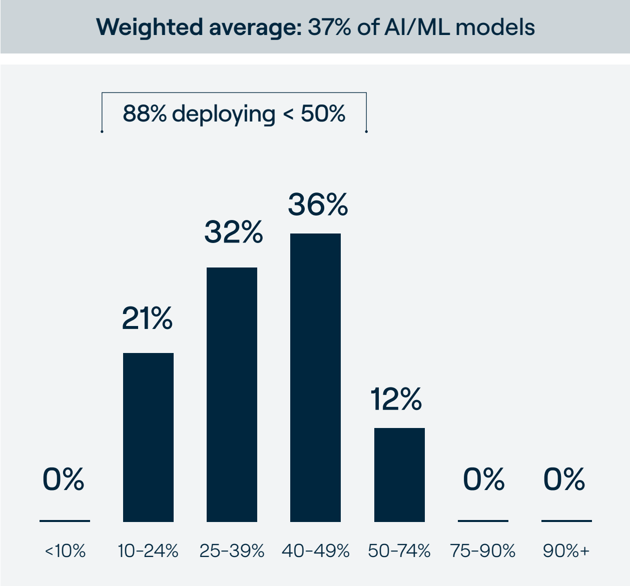


Figure 13: Percentage of AI/ML Models Deployed in Production.

Main Impediments to Model Deployment

The main impediments to actually deploying include scalability (47%), performance (46%), technical (45%), and resources (42%). The fact that they were all mentioned as a "main impediment" by a fairly substantial portion of the respondents, shows once again that there isn't just one glaring impediment to model deployment, but rather a multi-faceted one.

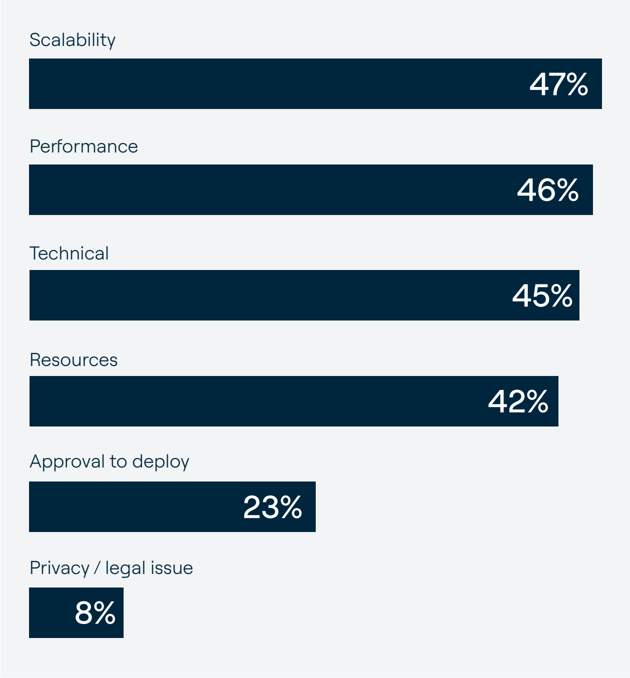


Figure 14: Main impediments to model deployment.

Demographics

Country, Department, Role, Job Seniority

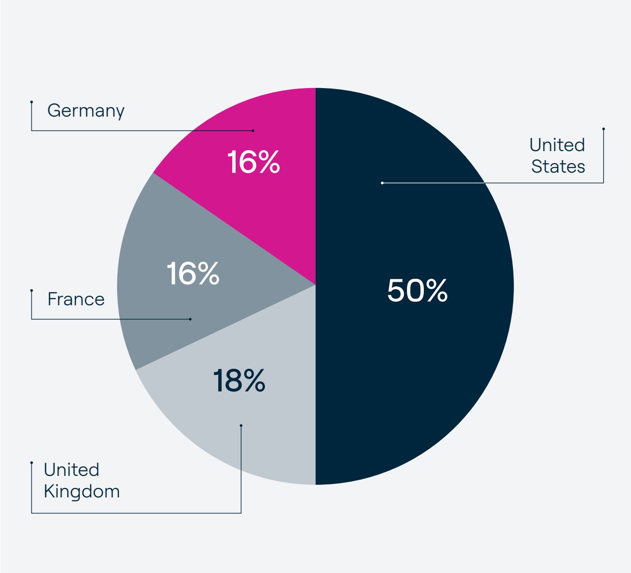


Figure 15: Country

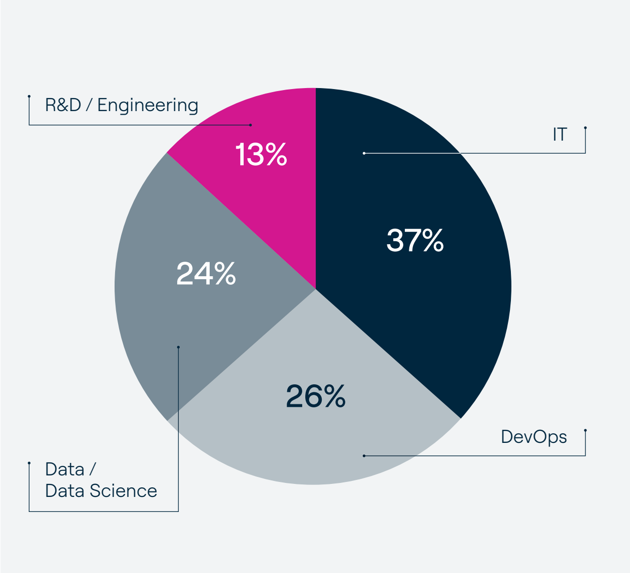


Figure 16: Department

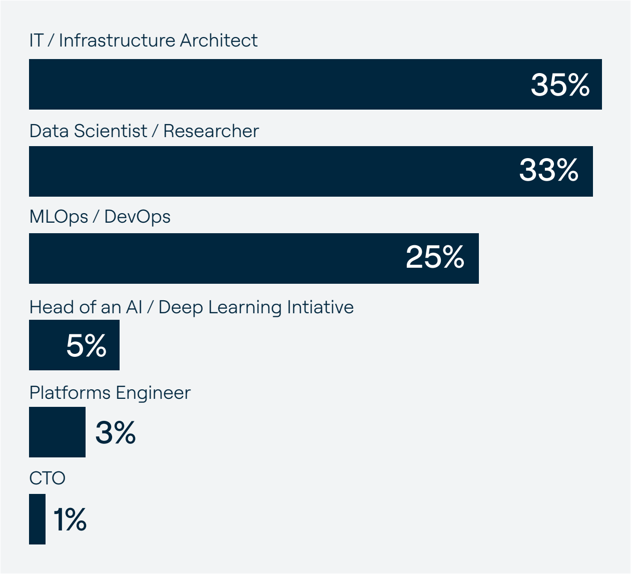


Figure 17: Role

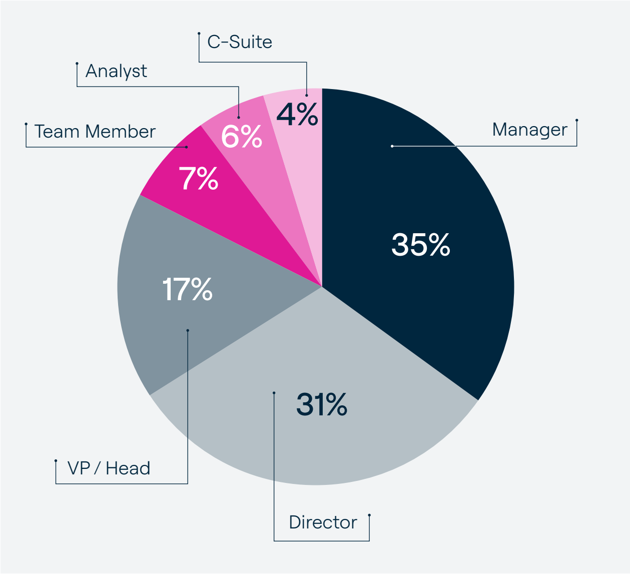


Figure 18: Job Seniority

Company Size, GPU Farm Size

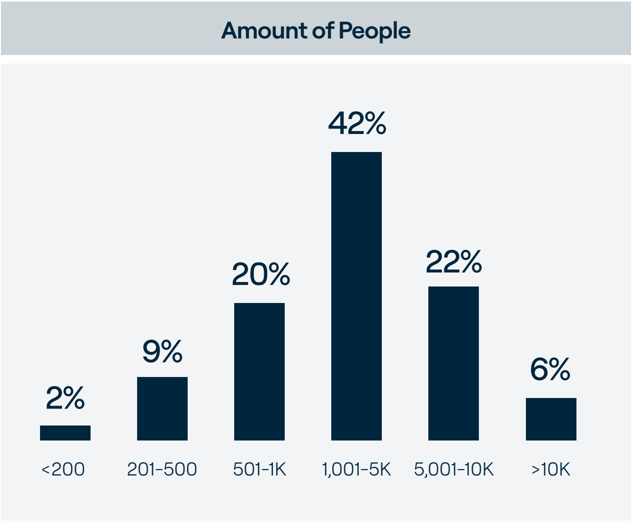


Figure 19: Company size

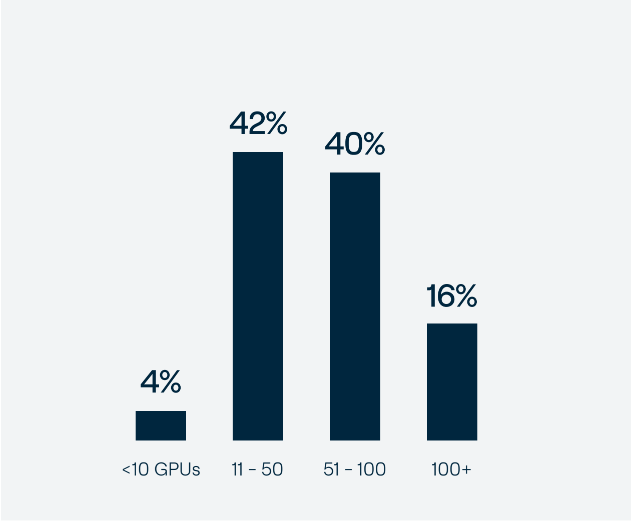


Figure 20: GPU farm size

About Run:ai

Run:ai's Atlas Platform brings cloud-like simplicity to AI resource management - providing researchers with on-demand access to pooled resources for any AI workload. An innovative cloud-native operating system - which includes a workload-aware scheduler and an abstraction layer - helps IT simplify AI implementation, increase team productivity, and gain full utilization of expensive GPUs. Using run:ai, companies streamline development, management, and scaling of AI applications across any infrastructure, including on-premises, edge and cloud.

For more information please visit us:
https://www.run.ai/