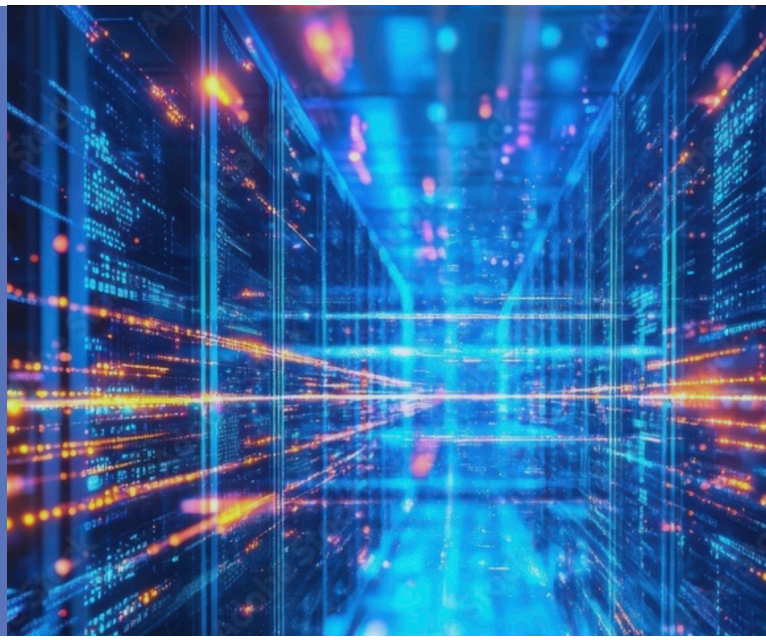
Freeform Dynamics in association with SCALITY

Storage Infrastructure for Enterprise AI

Lessons from seasoned adopters on building scalable sovereign environments



With the appetite for sovereign AI solutions growing, a recent survey of 504 senior IT and data professionals provides insights into what matters when designing storage infrastructure to support today's diverse set of enterprise AI demands.

Key Takeaways

- **Competitive differentiation is the primary driver for enterprise AI investments**
  Most see investment in AI as a strategic competitive imperative. Executive attention is high, with significant funding for AI initiatives, and healthy project pipelines.
- **Interest in LLM-based systems is accelerating demand for all types of AI**
  While LLM-based solutions have been the center of industry attention recently, most of the medium to large enterprises we surveyed are actively developing and deploying across multiple AI genres.
- **Private AI infrastructure is increasingly recognized as critical to success**
  In the cloud vs on-premises debate, many say that sovereignty, compliance and data proximity requirements often mandate running AI workloads on infrastructure that you fully control.
- **Beyond compute, the right storage is needed for fast, robust private AI systems**
  Performance and scalability are clear requirements when specifying AI storage needs, but strong security, risk and resilience capabilities are at least as important.
- **Seasoned adopters tend to look further ahead and build for adaptability**
  Those with greater experience tend to define AI storage needs earlier in the delivery lifecycle, considering requirements across the entire AI-related data pipeline. As a result they prefer flexible storage platforms over point solutions to ensure longer term relevance and ROI.
- **Making effective storage decisions aligns with greater confidence**
  Seasoned adopters are more confident in their storage infrastructure's ability to meet current performance, risk and resilience needs, and to deal with evolving requirements into the future.

Why this research, and why now?

The level of media and industry attention paid to generative AI over the past 2-3 years has been a double-edged sword. On the one hand, it has created interest and a sense of urgency among business leaders, which has helped to get AI onto the corporate agenda. On the other hand, it has created headaches for many IT and data teams charged with delivering on the AI promise.

One challenge here is the expectation gap. Business stakeholders, wowed by services like ChatGPT, often under-appreciate how much skill, time and effort is needed to build more critical AI-powered core business applications.

Adding to this challenge is the growing focus on IT sovereignty as businesses look to maintain control of their application estates, avoid lock-in to cloud service providers, and better deal with regulatory and performance needs that are sensitive to data placement and proximity.

In the context of AI, this increased emphasis on sovereignty and control is driving greater use of private deployments. In this "Private AI" model, AI workloads run on infrastructure that you control, e.g. located in your own datacenter, in a co-lo facility, or accessed via bare-metal hosting.

All of this is against a backdrop of an ever-moving target. While expectations of AI adoption and growth may be high, right now the chances are that no one can tell you with any degree of confidence which use cases will be on the agenda beyond the immediate pipeline, let alone over the longer term.

With this in mind, if AI momentum is building in your organization, it's crucial to have a coherent AI infrastructure plan in place so you can invest wisely, avoiding piecemeal decision-making and the build up of technical debt, dead-end systems and future constraints that stem from this.

A coherent AI infrastructure plan is crucial to avoid dead-end investments and the build up of technical debt.

Focus on storage

When it comes to building Private AI infrastructure, the compute side of the equation often dominates the discussion. This isn't surprising given the industry's tendency to use GPU capacity and performance as a proxy for a system's overall ability to handle training and inference workloads.

But in an enterprise context, success with AI usually depends at least as much on how well you integrate business data into your AI solutions. This in turn shines a spotlight on the storage that supports both AI systems and the data pipelines that feed them, including requirements beyond performance such as resilience, security, data protection and compliance.

It's against this backdrop that the research reported here was conducted. Based on input from 504 medium to large enterprises active with private AI, we explore real-world needs, then go on to look at what we can learn from those with more enterprise AI deployment experience.

Tuning into our research sample

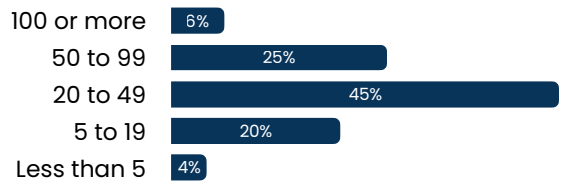
Before we get into the core findings of our research, it's important to understand who participated in the study so you can put the results we will be presenting into context. As mentioned at the outset, our core objective was to gather insights on the implementation of Private AI infrastructure, with a specific focus on storage. When recruiting participants into the study, which was executed in the form of an online survey, we spelled out clearly what was in and out of the study scope to avoid any ambiguity. Here is how the research was positioned to candidates.

Positioning of study with potential respondents

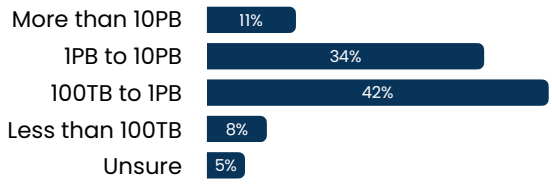
"We are researching storage infrastructure requirements for private AI implementations running in datacenters or hosted environments where you control the infrastructure stack. We're interested in all types of custom AI developments and deployments, from traditional machine learning to LLM based generative AI. The study covers storage performance, security and resilience requirements, plus associated trade-offs. This study is NOT about cloud-based AI services (e.g. from Open AI, Anthropic, Google, Microsoft, Salesforce, etc) or pure cloud platforms, but focuses exclusively on solutions built to be deployed and run in your own private AI environment."

Following this introduction, candidates were asked to confirm that they were active with Private AI and comfortable answering questions on this topic. Those who indicated that all of their AI developments and deployments were cloud-based were excluded from the study. Together with a range of other screening questions, this gave us a sample of respondents who were able to provide informed input on Private AI storage needs. Full demographics are presented in Appendix A, but here are some stats to give you a feel for participants' AI current environments.

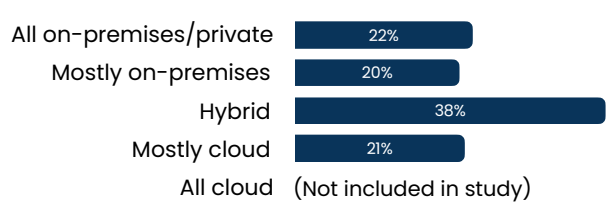
Number of AI workloads currently in production or development



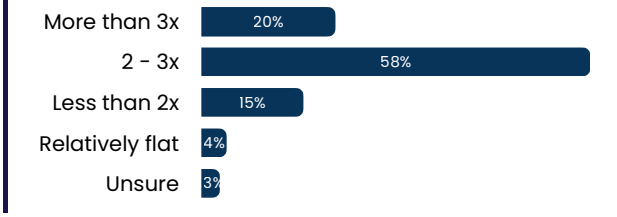
Current AI data volume



Location of AI related storage



Anticipated growth in AI related data over the coming 2 years



On a particular point, the AI data location chart tells us that even within this Private AI adopter community, cloud still generally plays a role, with a lot of hybrid deployments evident.

Level-set on AI drivers and scope of activity

Amidst all of the claims of market disruption and business transformation, it can sometimes be hard to pin down what's really driving AI activity and investment. During the study, we asked respondents to tell us, in their own words, where the impetus for AI is coming from. We then tallied how often the same themes were mentioned, which gave us the following rankings.

Drivers for AI investment (ranked by frequency of mention)

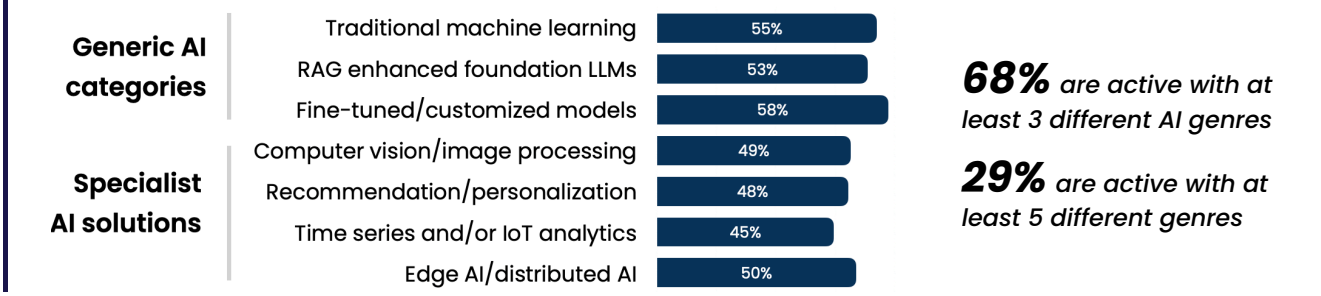
Competitive pressure	Getting ahead in the market, or at least not falling behind
Business necessity	Meeting financial, performance and risk related goals
Operational efficiency	Improving productivity, reducing overheads, better delivery
Tech advancement	Keeping pace with industry developments and opportunities
Customer expectations	Accelerated innovation, and faster/better service delivery
Data driven needs	Dealing with escalating volumes and analytics requirements

This list confirms that many medium and large organizations see AI helping to address market-level imperatives as well as requirements to drive internal transformation and optimization. In line with this, we see many organizations now taking a genuinely strategic approach to AI planning, adoption and deployment, with healthy levels of funding for AI initiatives.

How much do these statements apply to your organization?

But there's an important point to deal with before continuing. While the emergence of generative AI triggered the recent excitement, it's far from the only game in town. Other genres of AI technology are actually much better established and more suitable to deal with many requirements than LLMs. The list below, while not exhaustive, serves to remind us of the richness of the AI landscape.

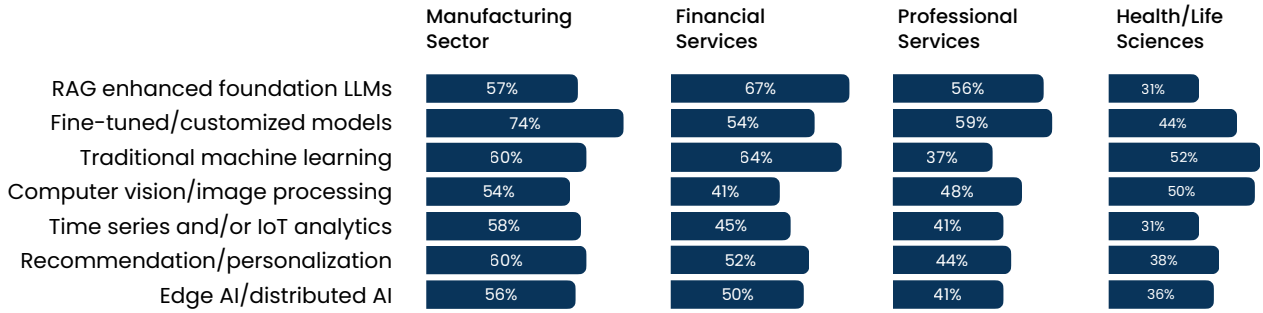
Which of these AI workload types do you have in production or development?



The importance of diversity and specialism

A look at the spread of AI activity underlines the notion that generative AI joins a whole range of other AI genres that have been meeting specific industry needs for a decade or two.

Which of these AI workload types do you have in production or development?



The emphasis on different AI genres across industries reflects a combination of history along with current and emerging use cases. Put these together and the full breadth and depth of activity is impossible to cover given the space available, but here are a few examples to provide a flavor.

A flavor of some industry-specific use cases

Manufacturing has long exploited computer vision in combination with specifically trained models for automated quality control. Meanwhile, time series analytics and machine learning have become integral to advanced planning and execution in areas such as production management, supply chain optimization and predictive maintenance.

Financial Services have been quick to move on RAG-enhanced LLMs for elements of knowledge management, while machine learning models have powered fraud detection, risk modeling and algorithmic trading for years, as well as customer segmentation, marketing and retention.

Healthcare and Life Sciences organizations show notably lower LLM adoption, likely reflecting concerns about variability of output, hallucinations and regulatory constraints. However, computer vision is well-established in diagnostic imaging, often delivering greater accuracy than human interpretation.

In Professional Services, RAG-enhanced LLMs are increasingly being adopted for bid formulation and project execution, while AI more broadly has been used for years for skills matching and resource optimization (with generative AI enhancements now finding their way into the mix).

When it comes to AI genres, it's important to think 'and', not 'or'

The use cases mentioned above are clearly just examples of the AI-related activity taking place across these and other industries. An important point to note is that many use cases are served by solutions that blend multiple forms of AI technology. And coming back to the theme of this report, many of them also involve handling large amounts of data, from high-resolution images at one end, to huge volumes of transaction or reference data at the other.

The Private AI imperative

Beyond waking up to the richness of the broader AI landscape, the industry is also paying more attention to delivery models. As part of this, there's been a growing emphasis on Private AI - a.k.a. Sovereign AI - as an alternative to the cloud-based platforms and services that defined the early generative AI market. In the Private AI model, you maintain full control over the entire AI systems stack, including the compute and storage layers. As the cost of infrastructure components has come down (especially the price/performance of GPUs), the barrier to entry has been lowered considerably, in turn making the Private AI approach viable for many more organizations.

As the price/performance of GPUs has come down over time, Private AI is now within the reach of many more organizations.

In terms of ongoing benefits, against the backdrop of a still volatile and immature cloud marketplace for AI services, the Private AI approach also allows you to better manage costs over time. You can avoid the runaway cost and lock-in problems often associated with public cloud services in general. And as your AI infrastructure grows, you can potentially enjoy economies of scale. Just as with the broader public versus private cloud debate, however, it's impossible to generalize on which is more cost-effective in the long-run - it depends on circumstances and requirements.

Application and data related benefits

This brings us to other benefits, and here Private AI helps to overcome many familiar cloud related challenges. Executing AI applications in close proximity to the on-premises data they often rely on helps to maximize throughput and reduce latency. Full control over the location of models and data also has clear advantages from a regulatory perspective. We therefore shouldn't be surprised to see so many Private AI adopters saying this approach is critical for their success.

How much does this statement apply to your organization?

Private AI based on infrastructure we control is critical to our success

43% **Fully applies**

38% **Somewhat applies**

81% *overall recognize the growing importance of Private AI as an alternative to public cloud platforms and applications*

This of course doesn't mean that cloud-based AI is no longer relevant. It's more that Private AI will increasingly become an important part of the mix as AI needs diversify and activity matures.

Infrastructure practicalities

Is it just about buying lots of GPUs?

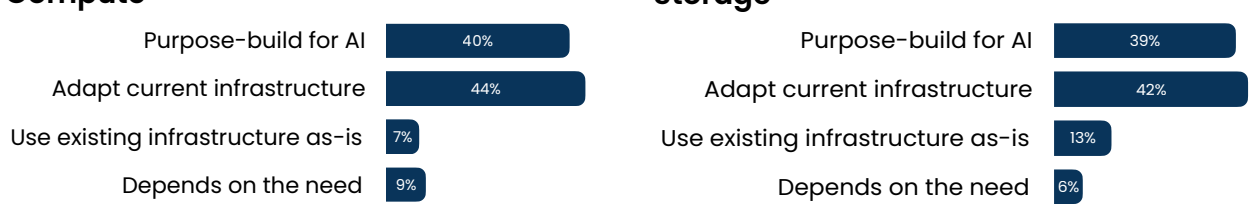
The sums of money mentioned when the media covers the "AI wars" can be pretty scary. We so often see global AI service providers such as OpenAI, Google, Microsoft and others throw about numbers in the billions when boasting about their investments in infrastructure. It's also very easy to get the impression that "winning" is primarily about who can build the biggest GPU farms, sending a message that compute power and capacity is all that matters when it comes to AI. This is all very entertaining, but it's not a great reference point when thinking about your Private AI environment, where success is dependent on a whole range of different factors.

Large service providers are not a good reference point when defining Private AI infrastructure requirements.

A more down-to-earth view of Private AI infrastructure needs

When it comes to Private AI, we have multiple genres of AI in play, together with a growing emphasis on domain-specific models, including SLMs (small language models) in the generative AI space. This reflects the focus on specific business applications and use cases rather than aiming to build systems that try to do everything for everyone (as in the hyper scaler world). That's not to say that some enterprise AI applications aren't extremely demanding, but they still tend to be a lot more focused and live alongside other applications with more modest requirements. The upshot is that there's no single formula for meeting Private AI infrastructure needs. Sometimes you need to invest in specialist equipment and platforms, sometimes you can extend what's already there - e.g. by adding GPUs or more suitable types of storage device.

What's your typical approach to meeting Private AI infrastructure needs?



We'll look at the circumstances in which you might consider more purpose-built infrastructure a little later, but there's another important aspect of Private AI that we need to consider first.

Integration of enterprise data

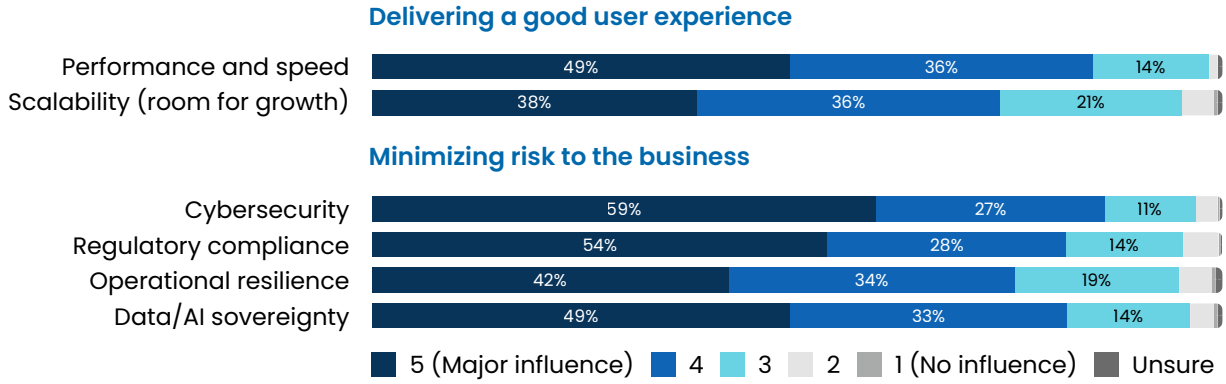
Regardless of the type of AI you are using to drive a particular enterprise application (e.g. generative, machine learning or specialist models), success typically depends on the effective integration of your own business data into AI solutions.

For example, up-to-to-date or even real-time business data is required to meet the needs of many operational and customer-facing AI applications, as well as solutions in areas such as analytics, planning and forecasting.

Spotlight on storage

This emphasis on data shines a spotlight on the storage that underpins AI applications and the data pipelines that feed them. Specifying requirements in this area is far from trivial as it means considering a wide range of performance and risk-related variables.

Factors influencing the selection of Private AI storage infrastructure



An important takeaway from this chart is that risk managements is at least as important as performance and scalability when it comes to AI storage. But there's another perspective on storage that's equally key to consider. This is how needs differ across the AI pipeline, from the data preparation stage (e.g. in a data lake), through staging environments used to hold data for ingestion and training, to hosting of models, vector stores, etc used for runtime inferencing.

How much does the following statement apply?

We recognize that different AI pipeline stages (e.g. data prep, ingestion, training, inference) have distinct storage needs

48% **Fully applies**

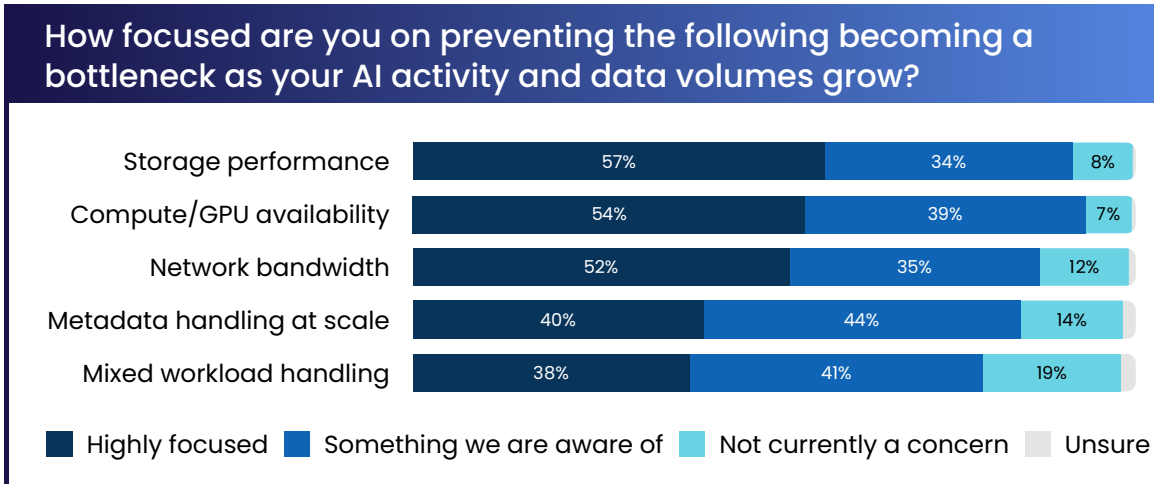
38% **Somewhat applies**

86% *overall recognize that storage needs can vary across the AI data pipeline*

Let's explore some of these performance, risk management and lifecycle requirements in more detail to get a better view of what matters and why when defining AI storage requirements.

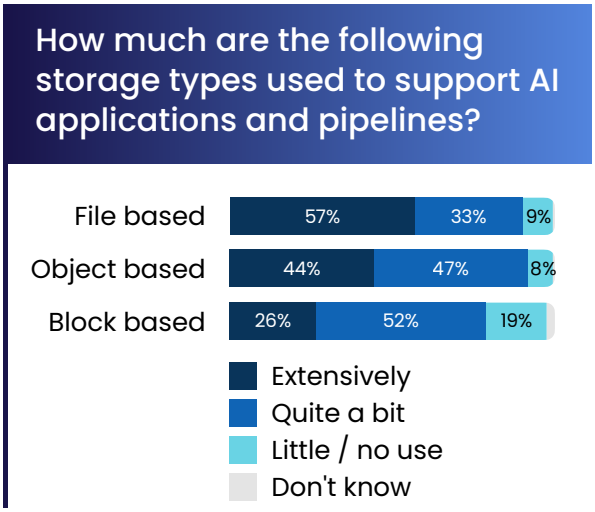
Performance drill-down

When we asked respondents about performance in more detail, specifically where they focused to avoid bottlenecks, we expected to see compute stand-out as the most important factor. What we saw instead was compute, storage and networking ranked very closely together, with storage actually appearing at the top of the list. This reinforces two important notions - firstly that storage is acknowledged as being at least as important as compute, and secondly that AI solutions require the balancing of resources as much as any other type of application.



The metadata factor

Another important takeaway from the above chart is the mention of metadata scaling - a major consideration in an AI environment within which random access to small items of data is fundamental to many operations. It's something to consider given the way training and inferencing works across different AI genres, and how different storage formats are often used across different stages in the AI lifecycle.



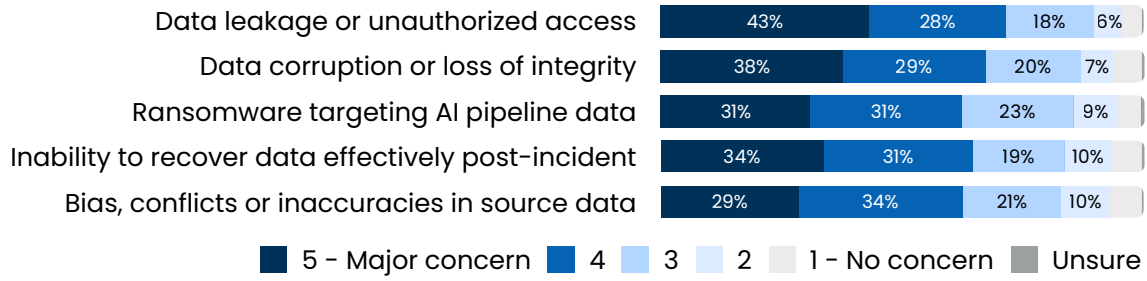
Mixed workload handling

Picking up on that last point, the storage performance characteristics required to optimally support different operations in an AI environment also need to be considered if you are setting up storage pools. Model training demands high-throughput sequential access to massive datasets, while run-time inference requires low-latency random access to model parameters and reference data. Production environments often support both alongside data preprocessing, creating fundamentally different demands, which can create performance challenges. But it's not just performance - security and risk issues can also arise. Let's look at this topic more closely.

Drill-down on security and risk

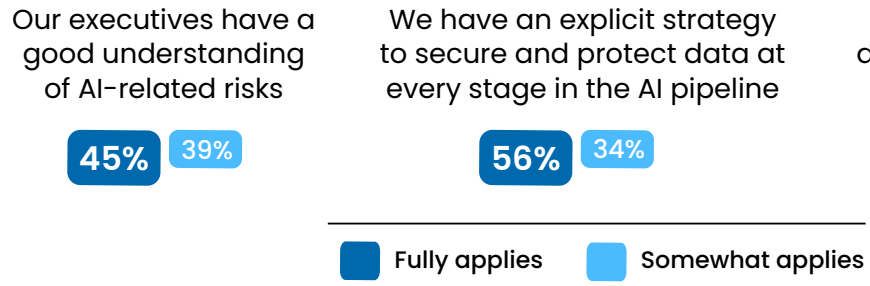
From a risk perspective, AI applications and their data pipelines are no different to other classes of data-driven solution when it comes to protecting data from loss, assuring recoverability, and so on. That said, problems with source data can be harder to spot from the output of an inferencing engine, so it's important to trap bias, conflicts and inaccuracies early, before you get into training.

How much do these occupy your thoughts when deploying AI?



The good news is that many of the respondents in our study have already taken steps to address the risks from a strategy and policy perspective, or are in the process of doing so. This includes making sure that senior executives have an adequate understanding of AI-related risks.

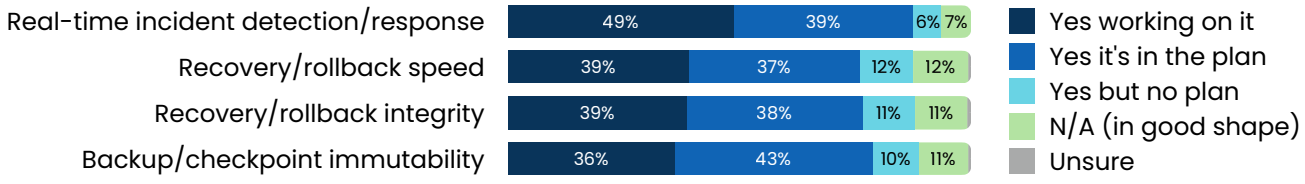
How much do these statements apply to your organization?



A closer look at cyber-resilience

A strong focus on resilience and recoverability is important for any core business application, but in the case of AI, it's not just about business disruption. AI pipelines often involve complex long-running processes, which if interrupted can lead to significant wasted time and resources, and users having to rely on out-of-date models in the meantime. Many have work to do in this area.

Do you need to improve your AI cyber-resilience in any of these areas?



Making AI storage buying decisions

Specialist point solutions versus versatile platforms

When you start out with experiments and pilots, you might legitimately exploit existing storage infrastructure to support your AI projects, perhaps extending or enhancing systems as necessary. As time goes on and you tackle more demanding requirements, you might then procure specialist storage solutions specified to support AI applications and their lifecycle stages individually.

However, while this approach can continue to make sense for very large AI applications, it may not be sustainable as you work through a busy pipeline of projects of all sizes. More versatile storage solutions capable of being configured to support different workload needs can avoid unnecessary cost and overhead, and reduce the startup time and effort for new projects.

How much does this statement apply?

A versatile storage solution that can cater for different AI pipeline needs is preferable to separate, individual solutions for each stage

44% Fully applies

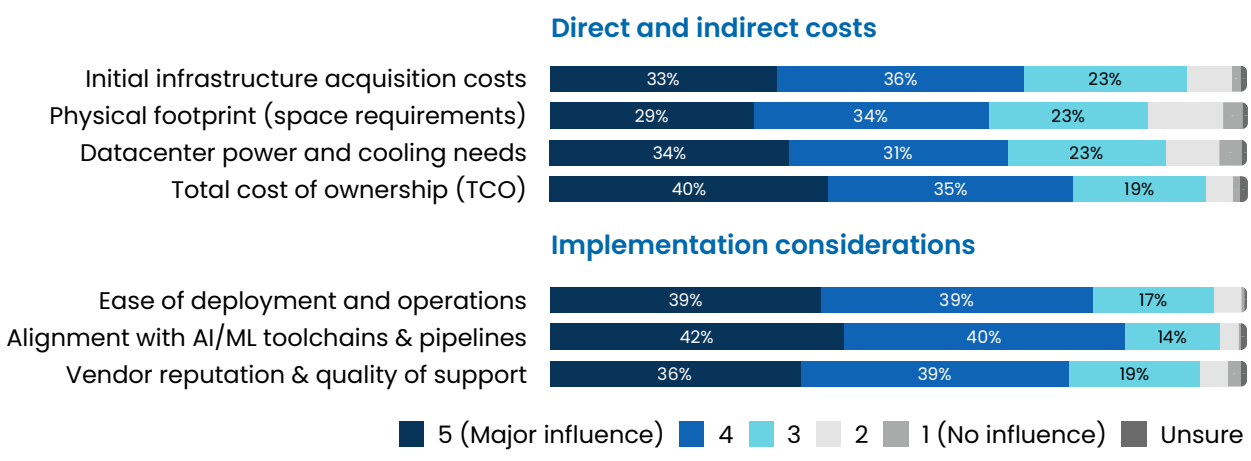
38% Somewhat applies

In practical terms, this translates to a preference for storage platforms based on an architecture that can accommodate components with different performance profiles (e.g. throughput, latency, resilience, security, etc), while providing consistency for application developers and operations staff. Whether this takes the form of dedicated instances for demanding applications, or shared storage pools for more general use, is secondary to architectural consistency.

Cost and practical considerations

When it comes to selecting specific storage vendors and solutions, all of the familiar criteria still apply, perhaps with an additional need to consider alignment with AI tools and pipelines.

Factors influencing the selection of Private AI storage infrastructure



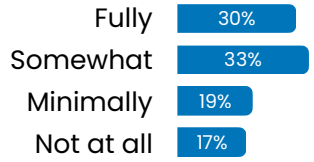
Compromises and trade-offs

A real consideration, but getting easier to deal with

Something we haven't mentioned so far in relation to making storage related decisions is how to deal with the opposing pulls of performance and risk management from a solution perspective. A key question here is when it's acceptable or even necessary to trade off one against the other.

How much does this statement apply?

In practice, we often prioritize speed and scale over cybersecurity and resilience



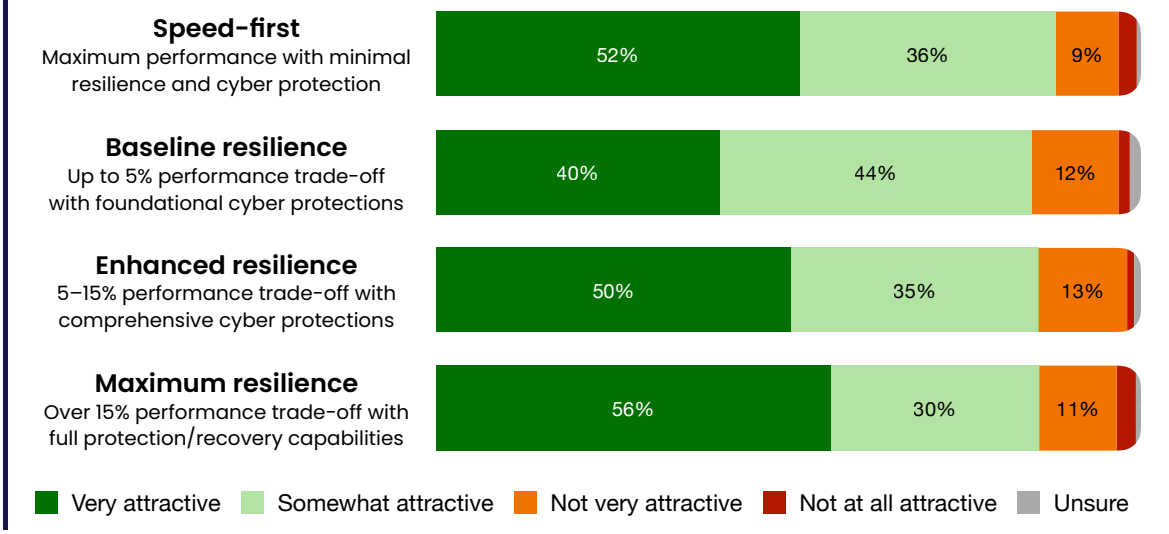
The good news is that advances in storage technology, particularly with regard to configurability to meet different performance and risk management needs, is making a difference in this area. There's a lot less need to compromise today and lock yourself into a particular skew. Indeed, some suppliers offer preconfigured options to potentially provide the best of both worlds - the right profile out-of-the-box, while maintaining underlying versatility.

Appetite for preconfigured options

During the study, we presented respondents with a range of illustrative options for preconfigured AI storage solutions. They were then asked how attractive they would regard each one to be.

Apart from assessing potential demand, this also served to illustrate likely real-world levels of trade-off, e.g. the performance hit that would stem from implementing different levels of resilience, etc.

How attractive would you consider the following options to be?



An AI experience perspective

So far, all of the data we've been looking at has been based on the overall study sample. This has been useful to talk through the context, issues and considerations relating to AI storage. To drive out the next level of insights, however, we segmented our study respondents according to their level of experience based on the breadth and depth of their AI developments and deployments.

Segmenting respondents by experience

Breadth of experience
based on the number of different types of AI currently in use

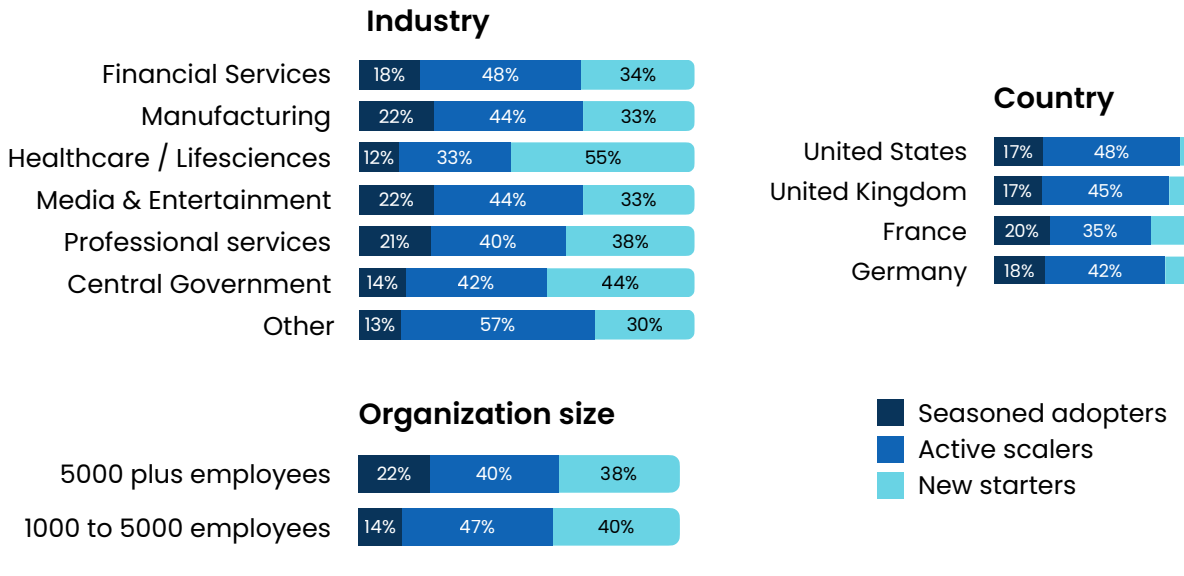
Depth of experience
based on the number of AI workloads in production or development

- **18%** of sample — **Seasoned Adopters**: Significant breadth AND depth
- **44%** of sample — **Active Scalers**: Significant breadth OR depth
- **39%** of sample — **New starters**: Modest breadth and depth

Significant breadth = Active with 4 or more types of AI
Significant depth = 50 or more AI workloads in production or development

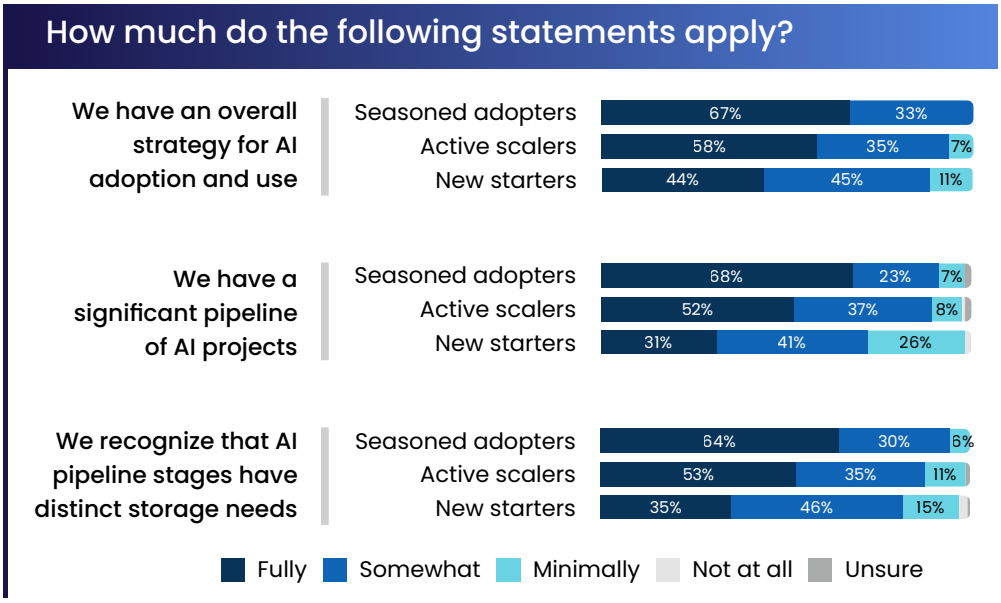
We'll look shortly at what those with the most rounded experience - the "Seasoned Adopters" - have in common. In the meantime, it's interesting to note how AI experience is distributed by industry, country and organization size. Bear in mind as you look at this data that our study was about AI in the broadest sense, not just generative AI, and that other forms of AI (computer vision, recommendation engines, domain-specific machine learning models, etc) have been well-established in some industries for a very long time, sometimes for decades.

Distribution of experience across the research sample



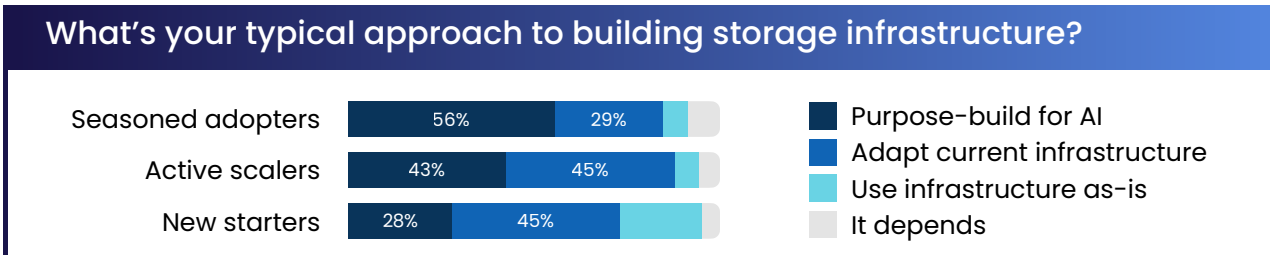
Learning from those with experience

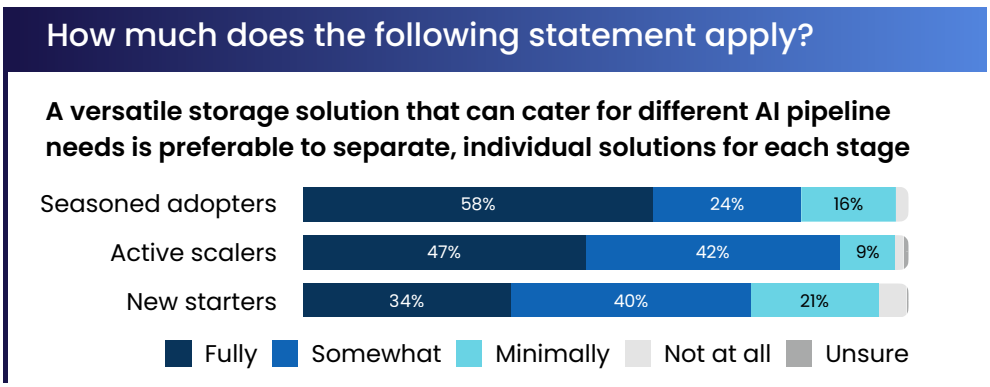
A number of things can be learned by looking at the differences in response between the experience groups. Most of this is intuitive, but it does serve to confirm the importance of certain principles and behaviors, e.g. in areas such as strategy planning and pipeline insights.



That said, we shouldn't infer causation from correlation. Does putting a strategy in place allow you to better define a meaningful project pipeline, for example, or does high demand force you into better prioritization and management? Either way, it confirms the value of standing back to consider where AI will impact your business the most.

Some of the other correlations we see reinforce the notion that as activity and experience grows, you're more likely to appreciate the value of purpose-building AI infrastructure, and prioritizing versatile platforms over point solutions to manage technical debt, avoid dead-end investments and maximize long term ROI and time-to-value.





Lessons on planning and performance

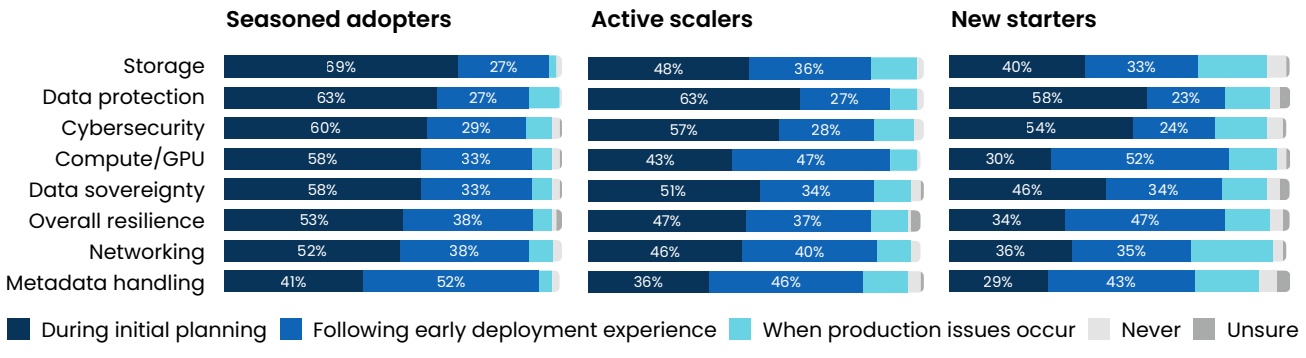
Up front requirements definition versus trial and error

Those with more experience tend to do a lot more upfront requirements definition for Private AI deployments. Their accumulated knowledge through past projects and across different AI genres enables them to assess what matters in the real world with greater confidence.

Those with less experience, meanwhile, are more likely to depend on a trial-and-error approach, optimizing and/or remediating during the early stages of deployment or when problems crop up in the production environment.

The exceptions here are data protection and cybersecurity, undoubtedly because issues with these can cause significant damage, even during pilot projects. At the other extreme, we see a big difference in relation to storage, suggesting that needs here are often underestimated by those lower down on their learning curve.

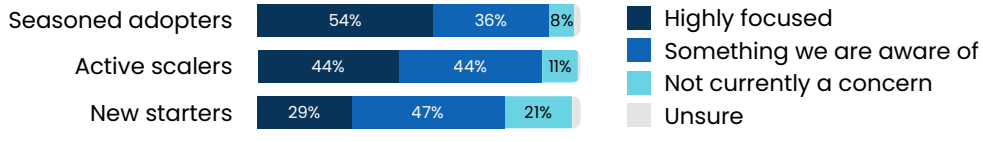
When in the project lifecycle do you typically finalize requirements in these areas?



The metadata blind spot

Given the relative lack of upfront planning seen above in relation to metadata handling, it makes sense to see seasoned adopters more likely to focus on this area. That said, a third of this group still say they haven't moved on from identifying the need to actually doing something about it.

How focused are you are preventing metadata related bottlenecks as your AI requirements grow?



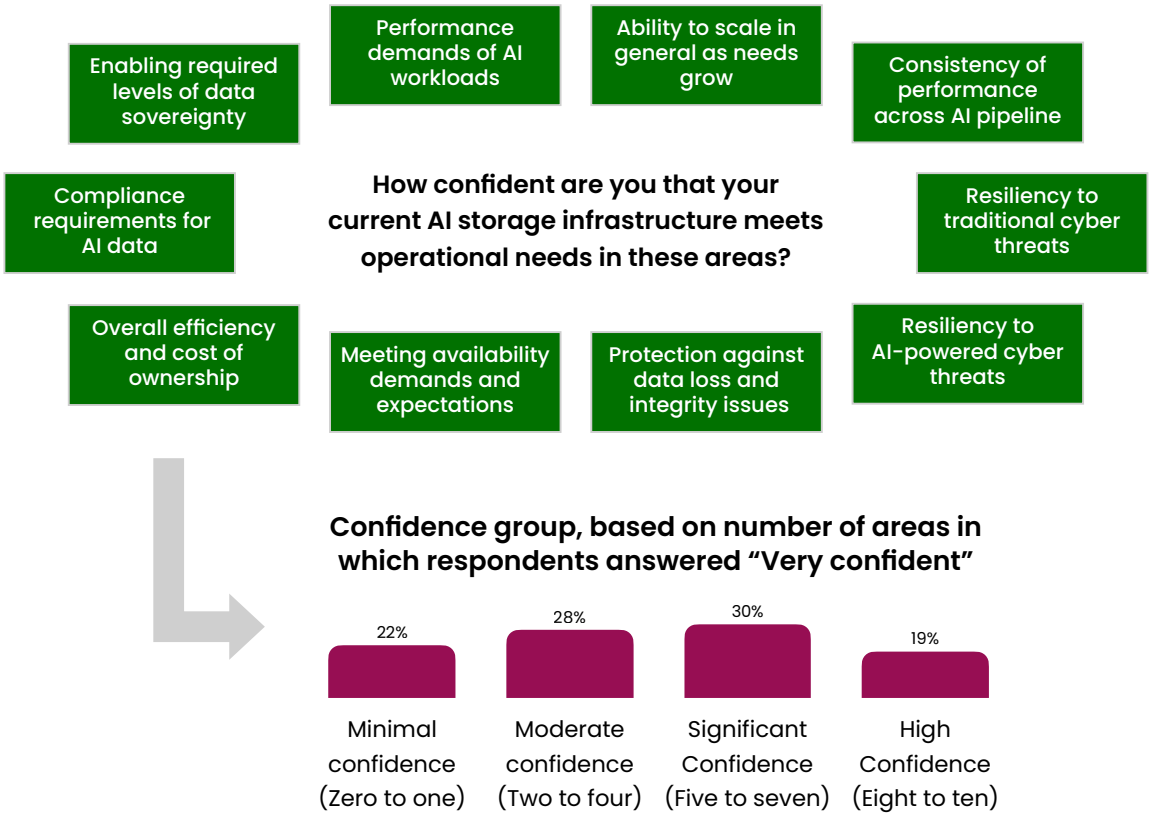
This common metadata blind spot is perfectly understandable given that sensitivity varies so much between different storage formats, as well as across different pipeline stages.

But how much do the differences between experience groups matter when it comes to building robust, secure and performant AI storage infrastructure? It turns out quite a bit.

The experience advantage

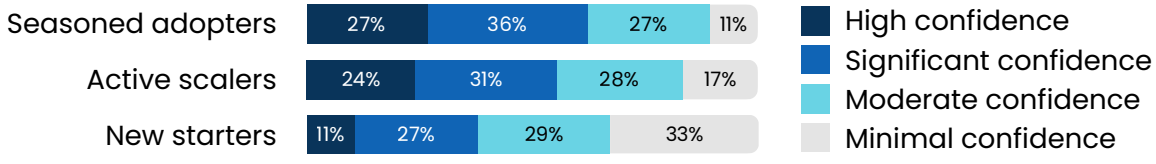
Measuring the effectiveness of an organization's Private AI infrastructure through an online survey is impossible to do in any meaningful way. To provide a rough idea of the impact of some of the behaviors and choices we have been discussing, we therefore used confidence in meeting operational needs as a proxy. We then grouped respondents Into confidence groups.

Assessing Private AI infrastructure strength



It was a short step from here to being able to cross tabulate confidence with experience to examine how the two relate. When we did this, a clear "experience advantage" emerged.

Confidence in AI infrastructure by breadth/depth of AI experience



In reality the difference between seasoned adapters and others is probably greater than suggested here. This is because psychologically, as you gain experience in any area, you tend to be harder on yourself when assessing achievement. In other words, seasoned adopters are almost certainly judging themselves by a higher set of standards than those with less experience.

Final thoughts

The virtuous cycle of adoption and value

At the time of writing, the industry conversation around AI is still dominated by how to exploit cloud-based LLMs. It's this model that generally springs to mind when people discuss AI applications in an enterprise context. However, it wasn't that difficult for us to find qualified participants for our Private AI research study, particularly among larger organizations. And as our study has revealed, those who have started down this route are doing so for good reason. Furthermore, once they are on board with the sovereign approach, momentum builds into a virtuous cycle. The more projects they undertake, the more value they generate, the more demand grows, and the better they get at delivering. It's no coincidence that those with more experience tend to have larger AI project pipelines. As time goes on, and the realities of sovereignty, performance and compliance nudge more and more organizations into their first private AI project, we can expect Private AI to become an embedded part of the mainstream, just like private cloud.

Early adoption of Private AI sets up a virtuous cycle in which more experience and success drives more value and greater demand.

But experience matters, and suppliers have a key role

One of the biggest takeaways from this research is that experience matters. You could argue that this is the same with any emerging technology, but in the AI space, you are often battling against an inflated sense of expectation and urgency, which adds to the challenge. And while AI projects can build on previous experience with other data-driven solutions, the way you specify and build AI infrastructure still benefits from prior knowledge in this specific area. The good news is that if you don't yet have a lot of this in-house, suppliers can help to fill the gap, and this doesn't have to mean big consulting bills. Product companies are usually very willing to help with assessment, sizing and configuration leveraging experience from working with other customers. The bottom line is that you don't have to work everything out from first principles or rely on a trial and error approach.

Now is the time to build or refine your AI infrastructure plan

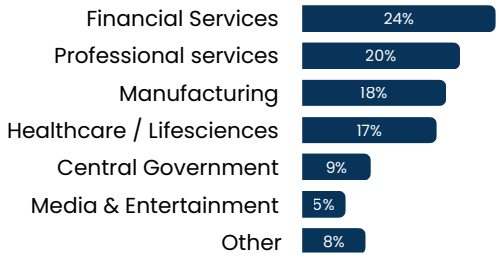
Given that activity is likely to snowball once the value of Private AI is recognized, it's important not be caught out and end up putting infrastructure in place in a reactive, piecemeal manner. Where possible, settle on an small number of versatile architectural options as soon as you can. This will help you establish a solid foundation for both current and future success with Private AI.

Appendix A: Respondent demographics

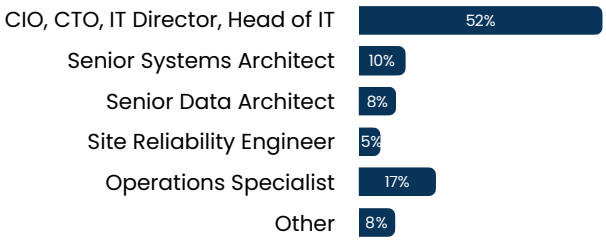
A total of 504 respondents took part in the study, with data gathered via an online survey. The research was designed, analyzed, interpreted and reported on an independent basis by Freeform Dynamics and completed in the Autumn/Fall of 2025. The project overall was sponsored by Scality.

Respondent demographics

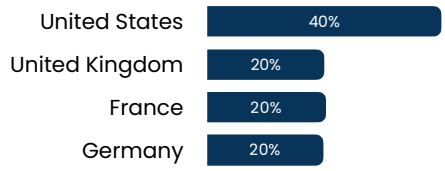
Industry



Role



Country



Organization size

About

Freeform Dynamics

Freeform Dynamics is an IT industry analyst firm. Through our research and insights, we help busy IT and business professionals get up to speed on the latest technology developments and make better-informed investment decisions.

For more information, please visit: www.freeformdynamics.com.

Scality

Scality is a software-defined storage company that specializes in providing cyber-resilient, petabyte-scale object and file storage solutions for cloud, AI, and backup workloads. Its main products include the ARTESCA and RING platforms, which are designed to be highly scalable, secure, and reliable for enterprise-level data management in hybrid or multicloud environments.

For more information, please visit: www.scality.com