BARC IT MARKET STRATEGY

Optimizing Your Architecture for AI Innovation

Authors: Shawn Rogers and Merv Adrian

Publication: March, 2024

Abstract

This report examines survey results for 335 Artificial Intelligence (AI) leaders and practitioners regarding their use of AI, including adoption status, technology optimization plans, project challenges, risks, and use cases. The report also shares specific insight into how High Readiness organizations are prioritizing their strategies, selecting trusted suppliers, addressing compliance and regulatory requirements, and the role of cloud platforms in AI innovation.

Research sponsored by:

AEROSPIKE

Table of Contents

Preface: Aerospike

Generative AI and large language models (LLMs) emerged in 2023 as key inflection points in technology and have ignited a new wave of innovation not seen since the advent of the Internet. AI applications are now omnipresent across industries and use cases, and there's significant demand for infrastructure to support the scale, speed, cost, and real-time needs of those applications.

An increasing number of companies are seriously considering AI but face many challenges. AI requires the ability to ingest and operate on massive amounts of data effectively. To fully leverage AI, its output must be of consistently high quality at any scale, costs must be controlled, and performance must be reliable. Businesses must be able to augment the basic knowledge of foundation LLMs with "firm-specific" data to produce contextualized, accurate results in real time.

Classical, or "predictive," AI is changing as vectors provide a richer data type enabling semantic similarity searches using approximate nearest neighbor algorithms. These modern techniques help to improve the performance and efficiency of existing AI and ML use cases and open new possibilities.

Aerospike is AI ready, ingesting and persisting massive amounts of streaming data from tens of thousands of sources across diverse workloads to feed AI and ML models. It does all this while operating on a fraction of the infrastructure required by other databases and cutting carbon emissions.

Some of the most successful companies in the world have built large-scale, high-transaction AI and ML applications on Aerospike. These include Adobe, AppsFlyer, Barclays, Flipkart, Myntra, PayPal, and Riskified. With our highly performant vector database, we'll build on this core functionality to make AI applications more accurate, cost effective, and sustainable. Contact us to learn more.

Lenley Hensarling, Chief Product Officer, Aerospike

Research Summary

Enterprises across the globe are grappling with an enormously disruptive “new” technology stack as Artificial Intelligence (AI) implementations are touted (and rightly so) as the next stage in the evolution of information technology. A tsunami of vendor messages, research firm marketing and business pundit proclamations is driving anxious questions about timing, preparedness and possible competitive disadvantage among those firms that have not already ramped up their efforts.

Even corporate positioning is affected; boards and executives want to know that they are moving towards adoption and want to tout their progress to fend off competitors who may or may not be moving faster. But clarity is hard to find: Which technologies are essential? Where can skilled resources be sourced? What is the status of regulatory requirements? How hard will it be to integrate these new technologies?

This research examines the intentions and readiness of potential buyers of these technologies by asking them direct questions about what they know or don’t know; how they propose to build, govern and measure their plan; who their stakeholders, suppliers and partners are; and finally, how specific to their industry they expect their strategies and technology purchases to be. Worldwide survey respondents totaled 335, with the majority split between Europe and North America balanced by a variety of company sizes, job titles (IT, Executive and line of business) and multiple industries.

This is new territory for everyone. Even firms with a history of “AI projects” have been working with highly bespoke technology. The sudden recent ascendance of several nearly relatively standardized components, such as generative AI built on foundation models, vector databases and retrieval-augmented generation, has caught nearly everyone flat-footed. Existing technology providers have varying degrees of readiness for deploying, integrating and governing these technologies, and their service providers, such as systems integrators, are not yet substantially ahead of their prospective clients in experience.

Key Takeaways

- Everything you think you know about the AI market is wrong, perhaps driven by self-interested vendors and research firms with a clickbait approach to marketing. We were surprised many times, and you will be too.
- Over 95% of respondents are currently addressing their plans for AI. However, only 20.5% can be characterized as being in a state of “High Readiness.” Respondents have made the most progress in the area of security standards and compliance, with 56% saying they have either formalized or are reviewing/revising their initiatives.
- Although conventional wisdom claims IT is becoming less involved in computing decisions as decision-making shifts to business units, for this transformation, IT is involved 95% of the time, ahead even of Executive Management.
- The single largest challenge to the successful implementation of AI innovation is the skills gap, combined with formulating a strategy to overcome it.
- Unique and innovative technology is not the first criterion companies use to identify who they want to engage with for AI and GenAI projects: familiarity and credible use case experience lead this strategy.

- Hyperscalers are battling to win the war for AI workloads in the cloud. Microsoft Azure is the most utilized platform overall with respondents and is winning the race for AI customers.
- Data of all types is a critical part of a successful AI strategy. Quality, accuracy, access and orchestration have long been challenges for business intelligence and analytics practices. AI has surfaced a renewed requirement for highly performant data systems.

Recommendations

1. Address the skills gap fast and early. Engage your AI workforce with free training and certifications, and embrace knowledge sharing and collaboration at the beginning of your AI journey, not the middle.
2. Keep Calm and Carry On. This is an early-stage market—don't let haste and FOMO (fear of missing out) drive your strategy. And don't be surprised at the inevitable backlash from dissatisfied customers that will appear in late 2024 and well into 2025. Mistakes will be made.
3. Begin with your existing, trusted suppliers. Virtually all of them are seeking to add these capabilities, and thus far few have demonstrated a convincing lead over their competitors. Fit and integration will be key to timely delivery, and your best suppliers will have an advantage—assuming you qualify them correctly.
4. Compliance and regulatory concerns should be at the forefront of your planning but don't make the critical mistake of overlooking a complete Responsible AI strategy. Aligning your AI practice with company ethics and KPIs is a foundational strategy for avoiding common mistakes regarding bias, human collaboration, security and accuracy of enterprise AI.
5. Augment your existing business intelligence and analytics stack with AI. Without massive new disruptive capital investments, you can deliver immediate return on investment. Enable your teams with time-saving automation and serve a wider enterprise community with AI-automated insights, narratives and smart dashboards.

Organizational Readiness

How Mature is Your Approach?

The overwhelming interest in these projects is evidenced in the answers to a series of questions we asked about organizational readiness and plans. We began with a basic planning question*: "At what stage is your organization with these critical AI planning/process initiatives?"* Only 5% are not yet discussing their plans—although over half are still in the early stages. Figure 1 shows the various stages and where our respondents stood in rolling out each of these critical initiatives.

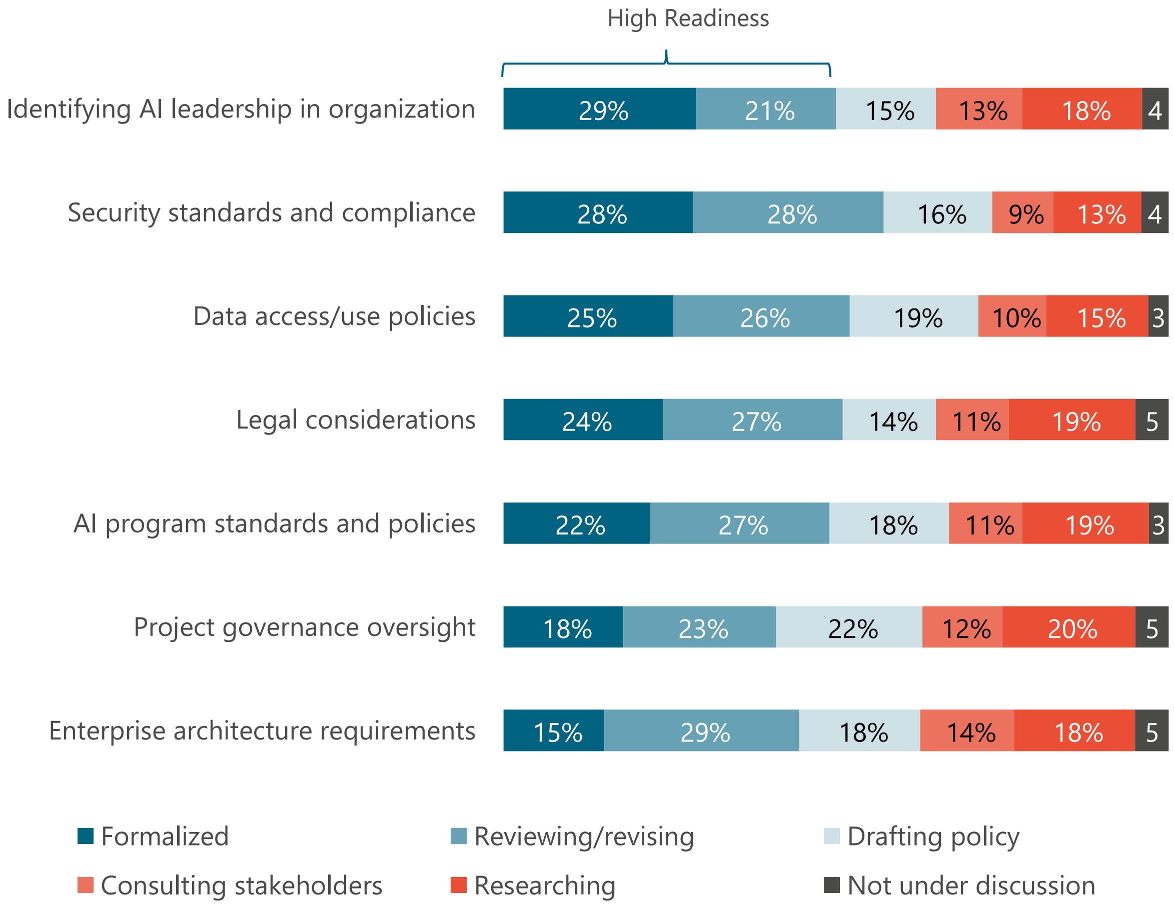

Figure 1: At what stage is your organization with these critical AI planning/process initiatives? (n=335)

In each of the dimensions examined here, roughly half of the respondents have arrived at drafts or final formalized planning/process stages. We refer to those who have done so across all seven dimensions—20.5% of the participants—as "High Readiness." The most mature dimension is Security Standards and Compliance, with over half (56%) saying they have either formalized or are reviewing/revising their initiatives. 73% have moved beyond the first two stages (researching and consulting stakeholders) for this issue.

Scores are fairly consistent for other issues, mostly in the mid-40s in percentage terms for draft or formalized plans. In a theme we will see echoed in other questions, the lowest combined readiness score (41%) is for project governance and oversight. Many organizations have not decided just how they will execute and govern projects, or measure their progress. As we discuss our findings, we will highlight the responses from the High Readiness organizations. Their answers will be less speculative and illuminate the early lessons already being learned.

North American respondents were far more likely than European ones to be part of the High Readiness cohort—87% compared to 10%. The size of the organization had less variability, clustering around 15%. Only firms with revenue between $10 million and $50 million had fewer than 10% High Readiness participants. Those between $1 billion and $5 billion came in at 18%, but those over $5 billion at 15%. From the industry perspective, information technology, retail/wholesale and manufacturing respondents are significantly more likely to be High Readiness than the other industries represented.

The Intersection of Budget and Strategy

Although leadership for the initiatives is typically not yet well defined, it is clear where those decisions are influenced. When asked, "*Which department(s) in your organization are influencing AI projects?*" the findings in Figure 2 were clear: Executive Management and IT are very present in this new round of investment. Combining the scores for Setting Strategy, Providing Budget and Both shows that IT is involved 95% of the time, with Executive Management close behind with a combined total of 88%.

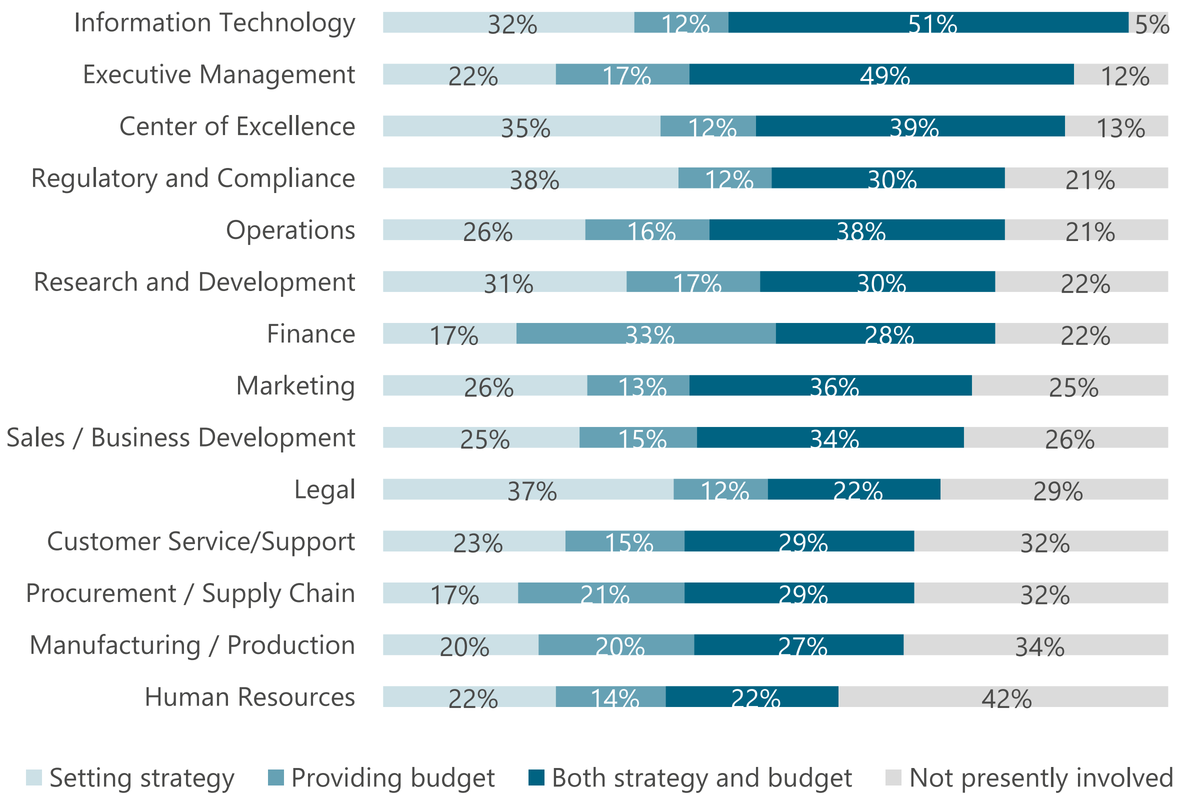

Figure 2: Which department(s) in your organization are influencing AI projects? (n=321)

Not surprisingly, Finance ranked highest for providing the budget, but its impact on strategy lags the leaders. While Regulatory and Legal were quite involved in setting strategy, they are not budget sources. HR is the least involved of all, below Manufacturing, Procurement and Customer Service, which all came in very low. Given the number of use cases associated with improving manufacturing processes and customer-facing systems, these results were a bit surprising.

Obstacles to Success

Unsurprisingly, as seen in Figure 3, lack of skills is the most frequently cited obstacle (39%) to new initiatives. High Readiness firms were even more likely to name this—41% did so. This was followed by costs or limited budgets (36%), although High Readiness respondents ranked it first at 45%. The latter reflects clear awareness that new investments are likely to be required—this is not a routine technology evolution but one that will have far-reaching investment requirements. Moreover, it has become apparent through our conversations with end users and vendors that a new team is likely to be formed alongside existing ones to take on these new projects.

The organizational challenges cited earlier are very present here—nearly a quarter of respondents cite cross-functional collaboration (24%) and lack of leadership and strategy (20%). However, High Readiness firms seem to have firmed up their leadership and strategy plans, and only 10% have named this issue. This was the second-largest single difference between the two cohorts.

The largest gap between cohorts was the issue of model indemnification and trust, which was named by only 7% of High Readiness respondents but 25% of the others.

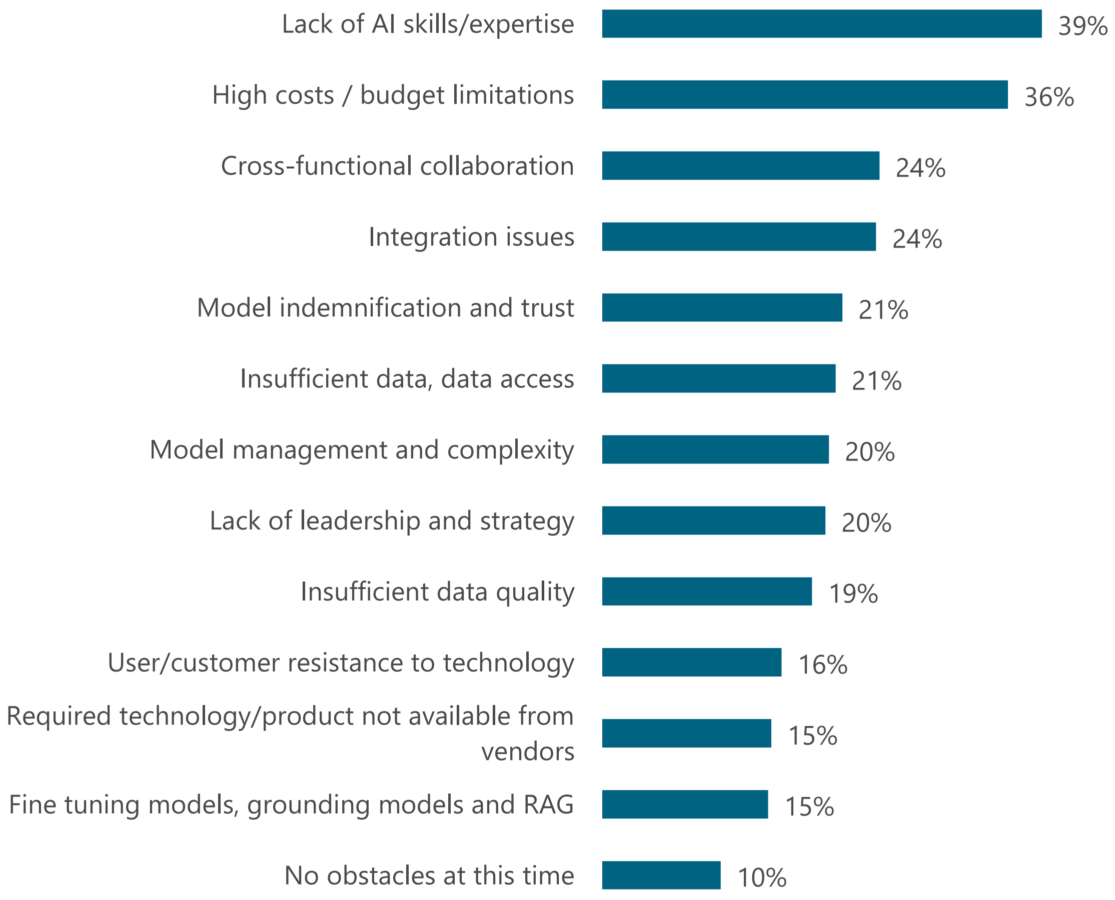

Figure 3: What obstacles are slowing/stopping your organization from delivering on your AI strategy? (n=335)

Technology-related issues, such as model management or tuning, simple availability of technology from vendors, or integration, as obstacles tended to fall at or below 24%—although 28% of High Readiness firms named integration. And, surprisingly, obstacles related to data—both access (21%) and quality (19%)—ranked quite low here as well, and High Readiness respondents were even less concerned, citing each only 13% of the time.

These technology-related responses may reflect faith in the marketing messages respondents have heard, telling them that data identification, access, quality and governance are easily solved if only they buy the products being offered for those purposes. In fact, many of these technologies not perceived as obstacles are relatively unavailable to AI tools, are in early preview or have significant gaps in many vendor offerings today. Clearly, implementers are likely to be hit with surprising technology capability challenges in the next round of deployments.

This soon-to-be-seen "buyer's remorse" is not unusual—in fact, it is a well-understood phase of the technology adoption cycle. But it can be minimized: careful vetting of new technology purchases, development of clear goals for projects and metrics, and governance for their success can all help avoid dead-end investments and activities.

Strategies to Overcome the Skills Gap

The primary reason why companies are struggling to leverage the opportunities presented by AI and GenAI technology fully is the AI skills gap. How will research respondents address their knowledge gap? First, by upskilling and reskilling their existing workforce, selected by 68% of respondents. This strategy is followed closely by internal knowledge sharing and collaboration at 61%. High Readiness organizations were even more likely to cite upskilling (75%) but were less optimistic about internal knowledge (50%) as a strategy.

AI as "the disruptor" is often blamed for reorganizations and reductions in force (RIFs) by companies in all industry sectors. To close the AI skills gap, many companies are refitting their teams with new hiring and talent acquisition. High Readiness firms are far more likely to rely on AI-managed services offerings—they ranked it second at 61%. This was an even greater difference between the two cohorts than the internal knowledge question.

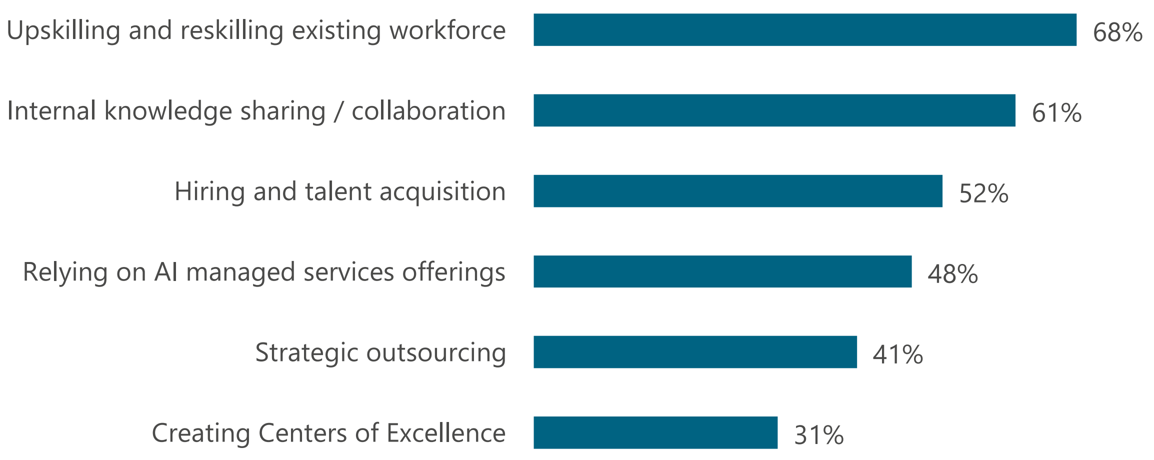

Figure 4: Select the top 3 strategies your organization is utilizing / will utilize to address the AI skills gap (n=130)

From a global perspective, EU-based firms are nearly 10 percentage points more likely to focus on upskilling and reskilling compared to their North American (NA) counterparts. Additionally, there are significant differences in the utilization of strategic outsourcing as NA-based firms are 6 percentage points more likely to bring in skills from this source.

When analyzing strategies by company size, small firms (under $10 million annual revenue) are highly dedicated to upskilling, with 80% of respondents including that strategy in their top 3. This trend is similar to that of global companies having over $5 billion in annual revenue. The trend is lower with companies in between these groups.

Previous BARC research of 529 AI professionals in October 2023 shows that after reliance on IT groups, respondents rank Centers of Excellence (BI, analytics, AI and cloud) at 23% for the group driving AI strategy within their company.

How Enterprises Manage AI Projects

Applying Strategic Resources

Developing a successful strategy for implementing and adopting new technologies like AI and GenAI requires a wide variety of resources. 87% of respondents identified Internal IT Resources as likely or extremely likely to play a role in supporting or enabling their company's AI innovation strategy, reinforcing the leadership we noted above on strategy and budgeting.

IT organizations deliver highly valuable domain expertise that can help fast-track adoption and overcome challenges with innovative technology. Moreover, they are far more likely to have established governance and process metrics to fast-track new initiatives.

Over 83% of respondents ranked reliance on Existing Technology Vendors as the second most popular selection. This is certainly reasonable: there are few true greenfield scenarios. Every new AI project will draw on, integrate with, and often help operate and improve existing systems. If your existing vendors are not in the lead, they certainly will need to be on the team, and you need to let them know that is expected.

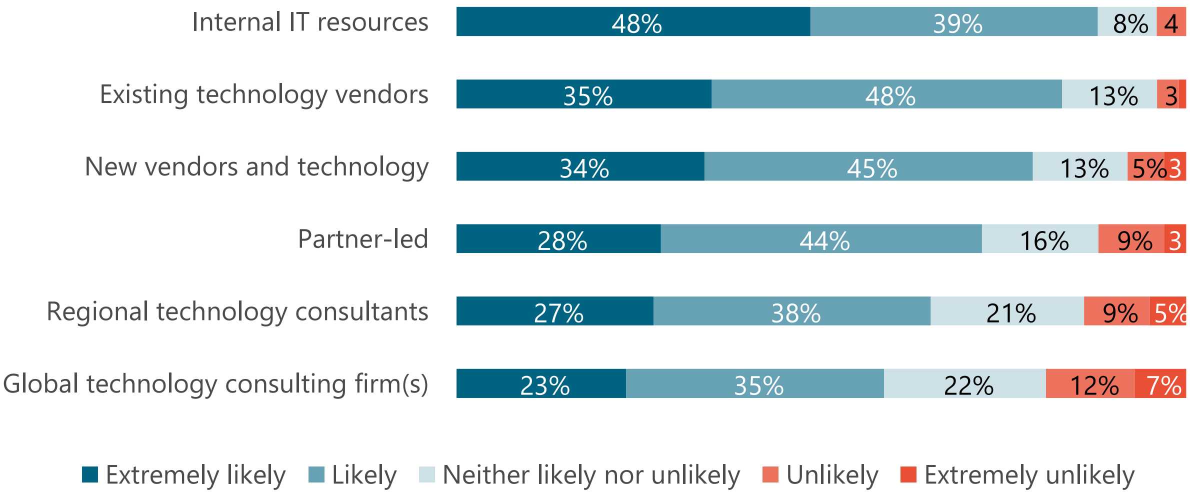

Figure 5: Which of the following resources will support/enable your company's AI innovation strategy? (n=335)

These top responses, coupled with high scores for Partner-Led and Regional Technology Consulting firms, suggest that these existing relationships are perceived as offering less risk for companies. With this said, the door remains open for New Vendors and Technology to enable AI within companies, as 79% of respondents indicated they are likely or extremely likely to include them in their strategy.

It's important to note that the substantial number of technology components that will be required are available from specialists, providers of data and analytics products and platforms, and large platform providers, including hyperscalers. For many organizations, a key decision criterion might be current and future scope. Niche vendors will inevitably be acquired—or fail—as the market evolves. Reliance on them for key components of the strategy may be risky. But the readiness of the big players will not be consistent across all layers of the stack and must be assessed thoroughly.

Making the Best Decisions to Drive Success

Optimal selection criteria for users selecting existing or new vendor technology play a critical role in de-risking AI-driven projects while enhancing project success. While Unique and Innovative Technology is part of the scenario at 37%, it is not the first criterion companies are using to identify who they want to engage with for AI and GenAI projects.

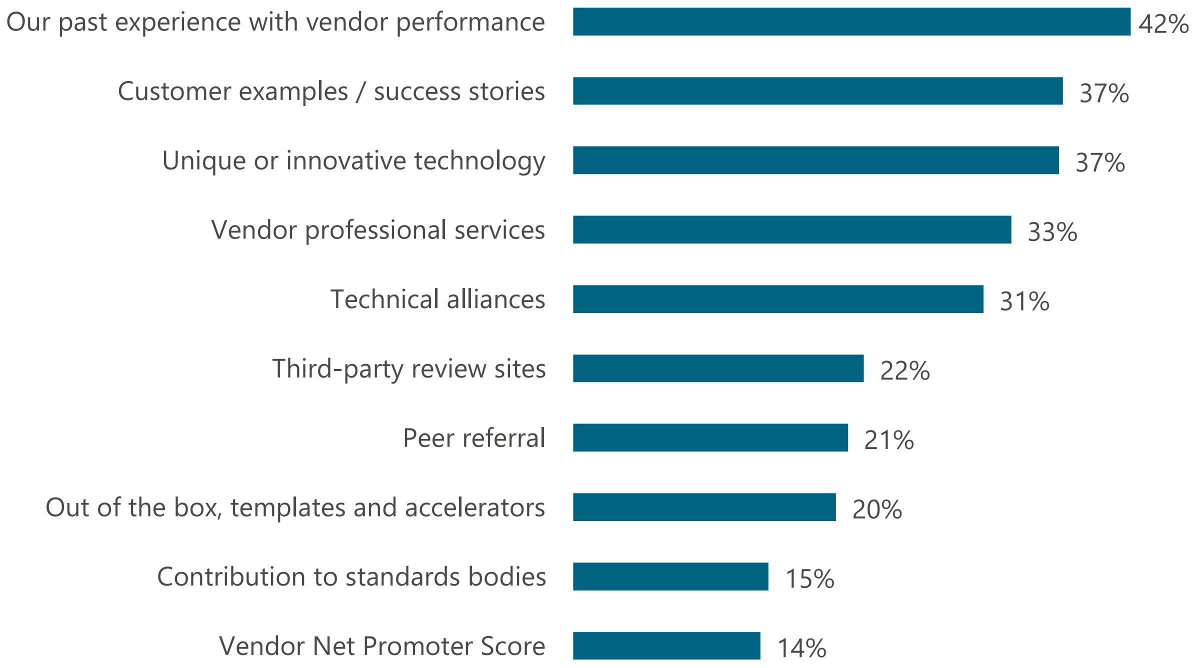

Figure 6: When selecting new and existing vendors for AI projects, what drives your selection and reduces risk? (n=332)

Survey respondents will begin with players they are already working with and are happy with: they identified "tangible experience" as the leading criterion for vendor selection in AI projects. 42% of respondents included Past Experience with Vendor Performance, resulting in the top driver for selection. Buyers are savvy about whether new technology should be considered in isolation: 37% selected Customer Examples and Success Stories. Both selections demonstrate the importance of proof from the customer's perspective.

The Role of Global SIs as a Resource

The list of service providers that will be chosen for these projects is varied. Leading systems integrators like Accenture, Deloitte Consulting and McKinsey & Company came in at the top of the charts. From a regional perspective, 44% of North American respondents identified Accenture as the top choice, followed by Deloitte at 40%, with McKinsey and Company a distant third at 28%. The European respondents selected "Other" as the top choice at 38%. Niche regional solution integrators and consultants varied the write-in answers—a very different trend compared to the North American respondents. Accenture was named second most by European respondents at 27%.

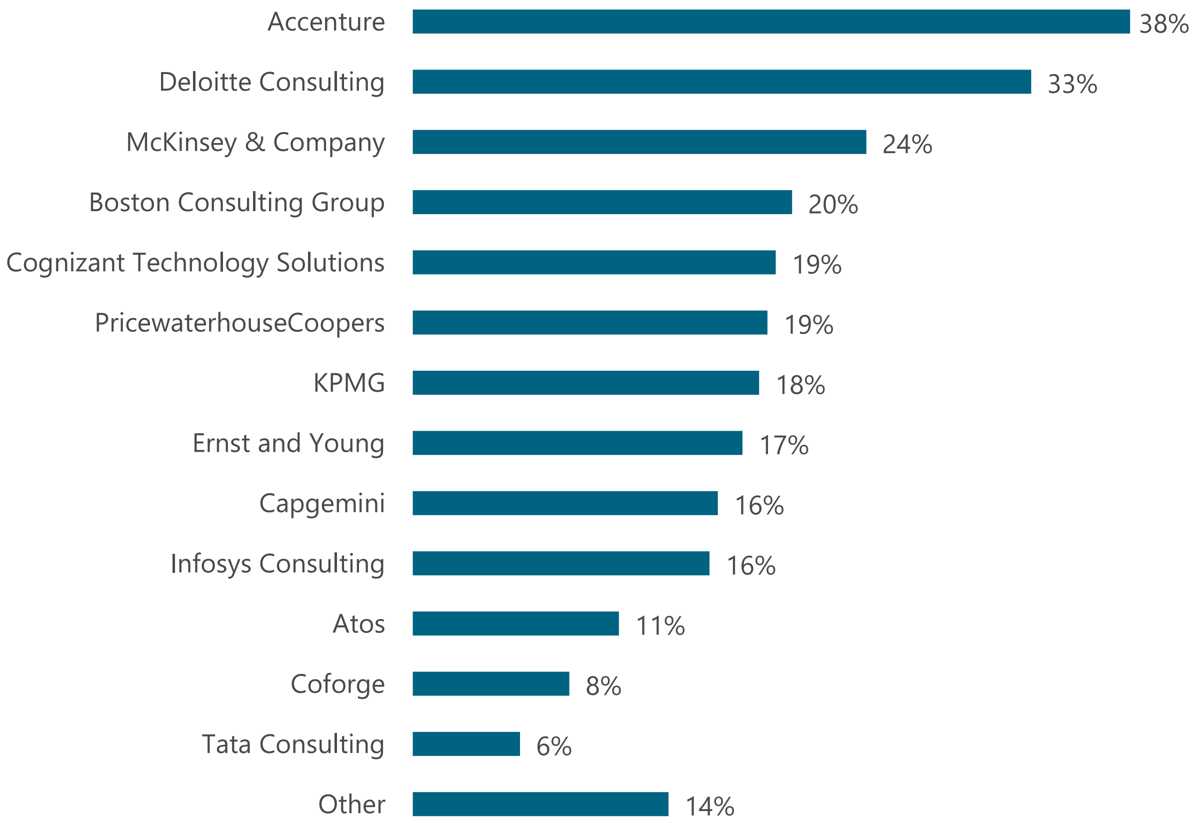

Figure 7: Which of the following is your organization likely to work with on AI strategy and projects? (n=229)

Large platform providers who have their own consulting offerings—most notably, leading hyperscalers AWS, Google, IBM, Microsoft and Oracle—were not listed among the choices for this question, in part to test whether respondents in the space provided would write them in. They did not—only a handful of references to these vendors as providers for strategy and projects were made. This is another surprise since those vendors are all aggressively marketing their offerings in this market. Future planned research will include more focus on this issue.

Responsible AI

Informed companies preparing to leverage AI-driven innovation are monitoring the quickly evolving landscape of global AI regulations. While that terrain can prove complicated with the EU AI Act, the US NIST Guidelines and others, it's important to include Responsible AI as an equally critical part of the overall success strategy.

Responsible AI acts as a guiding set of principles for how your company will manage data, train models, and deploy and leverage AI. It's important that these principles align with your corporate ethics and culture. These policies need to be shared transparently across your organization and with customers and partners.

Setting goals and policies for data and model bias, accuracy, human-AI collaboration, job impact, security and privacy are all topics responsible companies need to explore and align with.

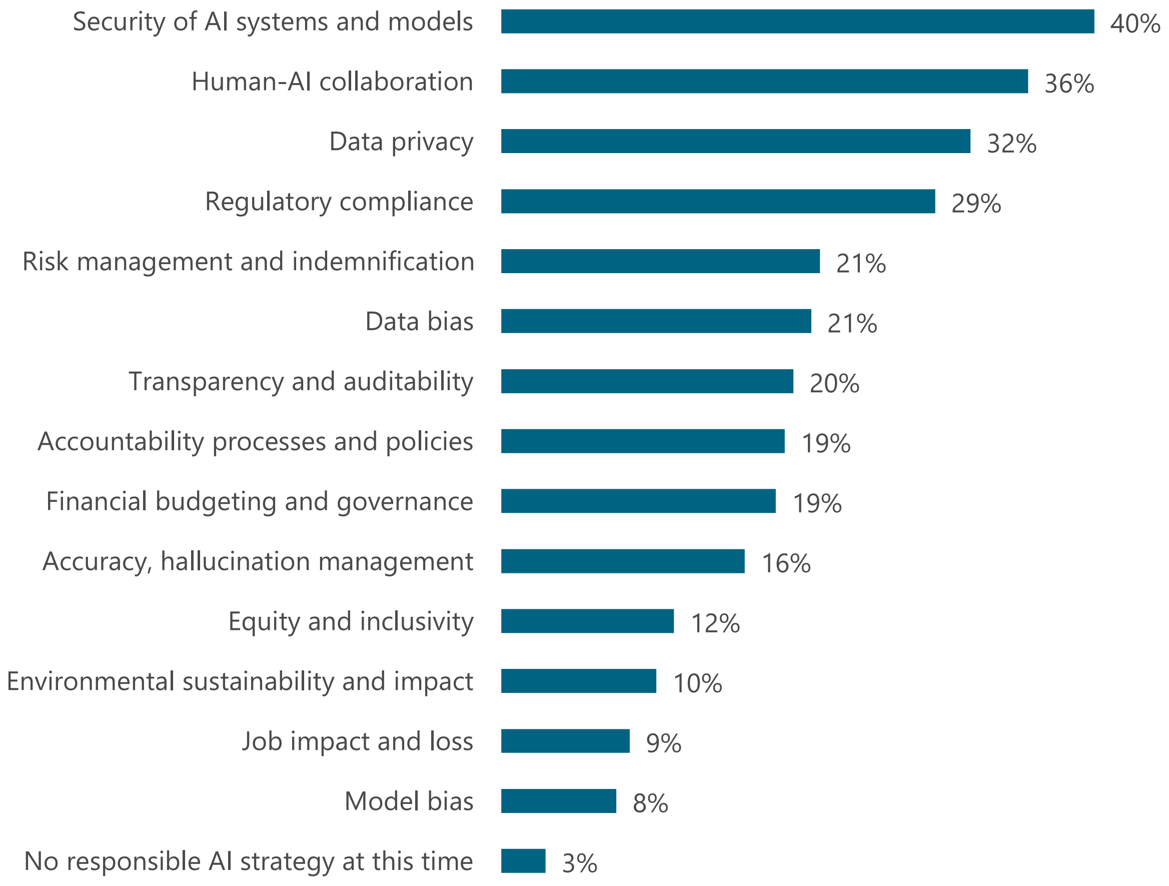

Figure 8: What components of Responsible AI is your organization prioritizing? (n=335)

This data clearly illustrates that companies are prioritizing the Security of AI Systems and Models and Data Privacy as the top Responsible AI strategies. High Readiness respondents had slightly higher scores for both of these. And, as noted above, this is the issue respondents identified as having the most mature planning efforts in this survey.

Previous BARC research on this topic, conducted in October 2023 with 539 participants, recorded that over 6% of respondents did not have a strategy for Responsible AI five months later. We see some incremental progress on the topic.

We expect risk management and indemnification to become more important over the coming months as customers adopt large language models trained on public data models. In 21% of our responses, this was selected as a priority.

High Readiness respondents were far less likely to cite Human-AI Collaboration: at 22%, it was not in the top 4 responses, with a gap of 17 percentage points compared to non-High Readiness respondents. It may be that they see the challenges of defining and tackling these issues more clearly.

Measuring Success and Failure

It's impossible to measure success for any project without clear metrics, and few are easily quantified. Organizations will be challenged to define how they will measure accuracy and alignment, although they are admirable goals. Only 1% of organizations say they have no plans to measure success, which is encouraging, though it may be either naive or disingenuous.

However, it is concerning to note that the speed of decisions, regulatory requirements and timely completion were at the bottom, all cited by fewer than 19% of respondents. All are relatively measurable and perhaps assumed to be well in hand. In the analysts' experience, they rarely are. High Readiness firms ranked them higher—with completion on time coming in fifth overall with 26%, exceeding the overall rate by 9 percentage points, and the non-High Readiness response by 11, the largest difference between the two groups. As elsewhere, High Readiness respondents were less concerned about data—in this case, protection and auditability (19%)—than the non-High Readiness cohort (28%). This was the second largest difference.

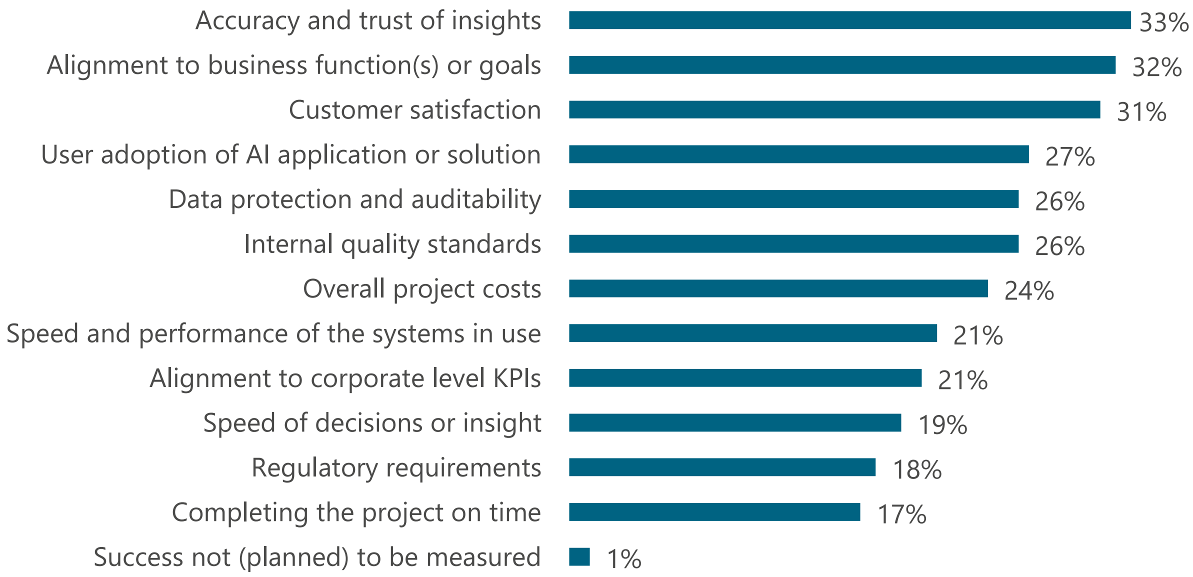

Figure 9: Select the top 3 ways your organization will measure success with its AI-driven projects (n=335)

AI Use Cases

General AI Use Cases

When it comes to leveraging AI and GenAI, many users ask, "Where should we use AI first?" Companies feel pressure to rush because of a fear of missing out (FOMO), so they often begin with general AI use cases that appear likely to help them expedite the project and experience an immediate return on investment (ROI) without taking on a lot of risks.

This doesn't mean that these use cases lack value, but they do provide a faster, less complex on-ramp to integrating AI into a company's culture. The most popular general use cases are AI chatbots and intelligent assistants, with 28% of respondents deploying them and another 24% in proof-of-concept (POC) testing.

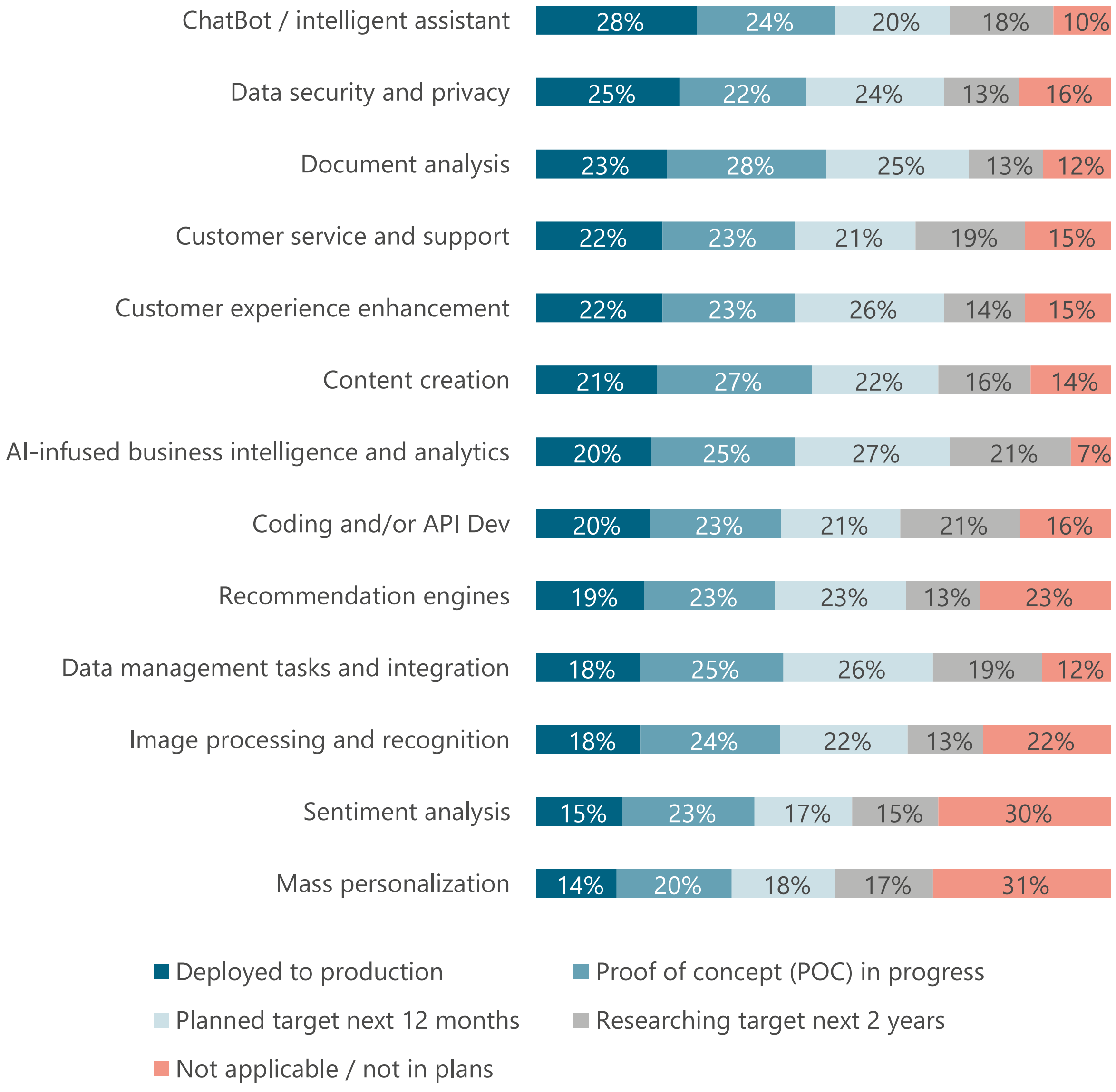

Figure 10: Which of the following general AI use cases has your organization deployed, planned or is researching? (n=334)

Use cases for AI-infused business intelligence and analytics, as well as AI-driven data management tasks and integration, are becoming critical for companies early on. When combining deployed, POC and planned projects in the next 12 months, the data shows that both are top-tier initiatives. Data management use cases fit this group with 69%, and business intelligence and analytics at 72%.

AI has been used to improve user experience via recommendation engines, automated interactions and personalized data. It's surprising that this use case came in lower, as early adopters have found positive ROI by infusing AI in it. 23% of respondents surveyed plan to adopt this use case in the next 12 months, and the same number are actively testing it.

Across all general use cases, mass personalization is the least popular, with 31% of respondents indicating that it is not part of their AI projects, and only 14% indicating that they have deployed that type of use case.

Specialized and Industry-Specific Use Cases

Specialized and industry-specific opportunities can be attractive because they may offer rapid competitive advantage and differentiation. They may also be technically “smaller” because they work with highly specific, already documented fact bases—a frequently cited reason for investing in retrieval-augmented generation (RAG, discussed below in “Model Management is a Critical Strategy for Success”).

The most frequently named use cases or industry-focused plans shown below in Figure 11 occur in Sales Forecasting / RFP Generation, with this use case deployed or underway with 80% of the respondents. Fraud Detection at 77%, Predictive Maintenance at 76% and Robotic Process Automation at 75% deployed or underway round out the four. Disease diagnosis was far down at the bottom—a huge surprise considering the early popularity of predictive healthcare stories in the market.

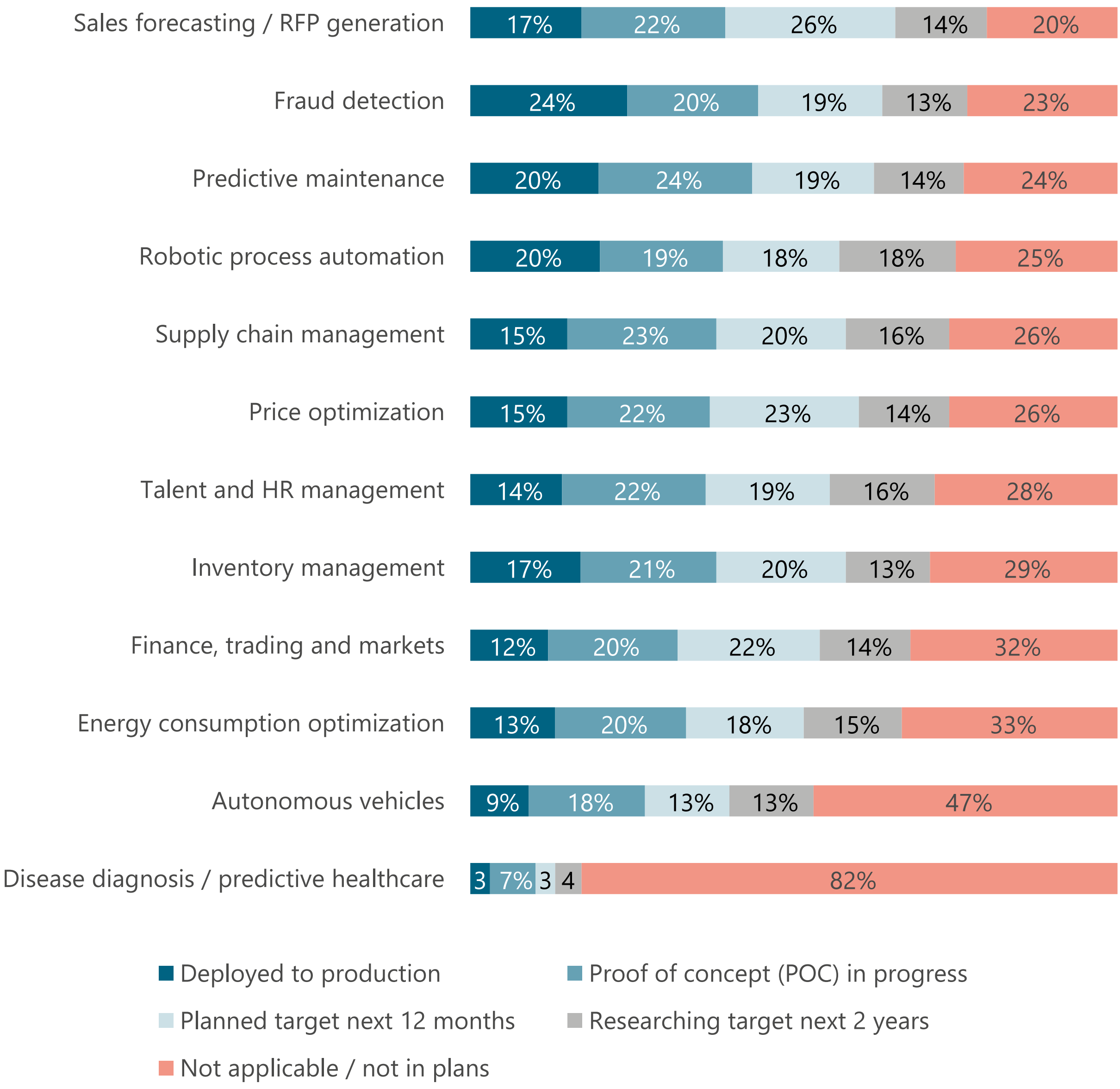

Figure 11: Which of the following specialized / industry-specific AI use cases has your organization deployed, planned, or is researching? (n=334)

Technology Choices

Transforming Your Architecture

Innovation often leads to disruption, especially when new technology is being incorporated. This can result in significant strategy shifts to ensure enterprise architectures can accommodate new use cases and workloads. The important question is whether your infrastructure is prepared for AI and GenAI, and what technology will be needed to meet new challenges.

When asked, *"How will AI innovation and projects transform your company's technology architecture?"* our respondents rejected the notion that previous architecture investments should be overhauled or scrapped to make way for AI. The data thus indicates that this is not necessary, and most respondents (61%) are simply integrating AI-specific technology into their existing architectures to fill gaps and add functionality.

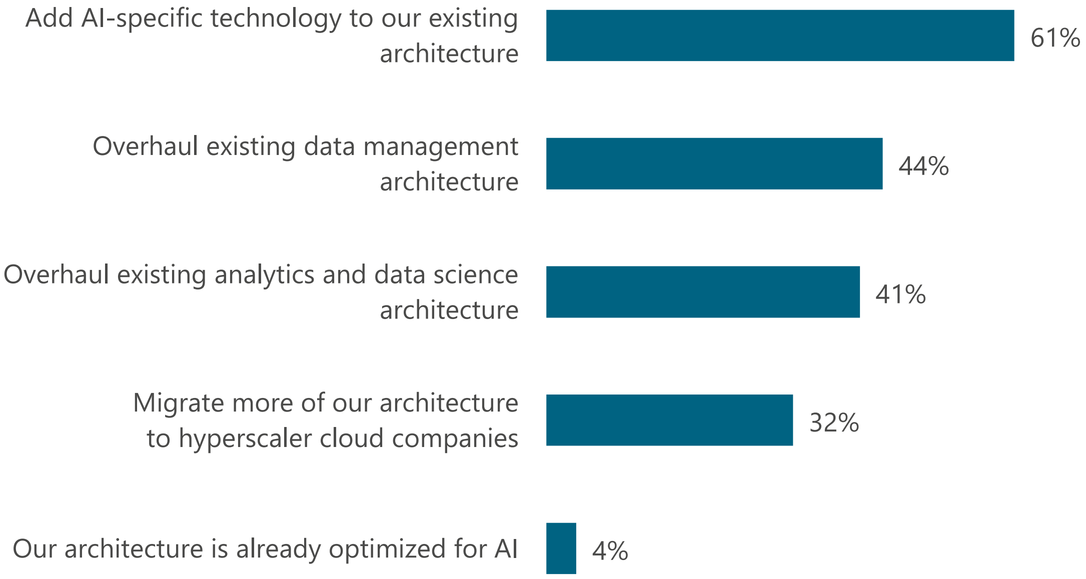

Figure 12: How will AI innovation and projects transform your company's technology architecture? (n=335)

Not at all surprising in the research is that only 4% of respondents state that their company architecture is already optimized for AI, further supporting the idea that while a sense of urgency exists around taking part in AI, you are likely in the 96% group that still has work to do to prepare for AI/GenAI innovation.

The survey data reveals that companies are focused on making changes to their data management, analytics and data science architectures, with consolidation around AI being a priority. This is not a surprise. Every year, data management software spending is the largest single category of IT infrastructure spending, and business intelligence/data science is not far behind. This has been the case for decades.

Hyperscaler cloud companies have made a compelling value proposition out of helping clients take advantage of their one-stop shops for easier AI implementation. Bringing AI projects on board will significantly accelerate cloud transformations for one-third of surveyed companies: 32% of respondents plan to migrate more of their architecture and workloads to cloud solutions.

However, we predict that cloud platform costs will play a role over the long term as customers face difficulty managing monthly cloud transactions. Those costs have not gone unnoticed, and platform providers are already taking concrete steps to demonstrate their commitment to help.

Augmenting Your Architecture for AI

A bewildering set of new technologies is vying for buyer attention. The confusion is evident in the similarity of responses on this question, and even more so in the size of the “not in our plans” or otherwise called “I don’t know” responses for each—typically in the 15% range. However, both vector databases (24%) and knowledge graphs (18%) came in higher. Foundation models and optimized hardware have moved the fastest, with combined testing and production rates at 58%, but the gap to other technologies is small.

Vector databases, AutoML environments and centralized feature stores are the laggards overall, but here again, the gap is small.

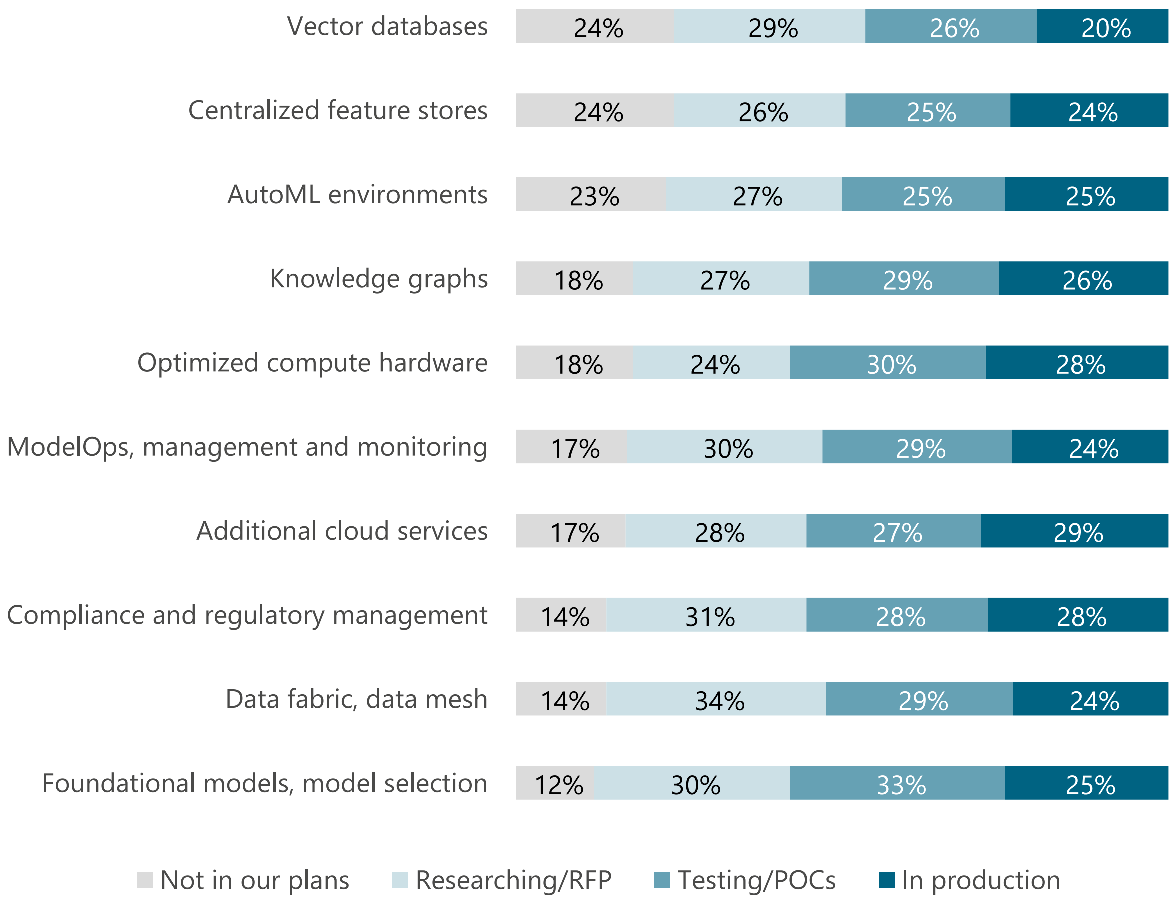

Figure 13: What is the status of the following AI technology in your existing environment? (n=298)

Perhaps the most interesting data comes from comparing the responses of the High Readiness cohort to the rest of the survey sample. It comes as no surprise that respondents who are identified as highly ready would be ahead of the other panels. The following chart, Figure 14, reflects the experience and priorities of the two groups and can be used as guidance for anyone who has yet to qualify as a High Readiness user, as we defined in the first section of this research (How Mature is Your Approach?).

From a priority standpoint, optimized compute hardware has risen to the top for the High Readiness group, with 83% already in production or testing/POC. From the analyst's viewpoint, this is likely due to their focus on cost optimization and budgeting for AI projects.

Choosing foundational models early in their journey has also been a priority, with 77% already in production or testing/POC, representing a gap of 29 percentage points between their experience and that of all other respondents.

Knowledge graphs round out the top three with High Readiness respondents at 77% in production or testing/POC, representing a 28 percentage point gap in their experience versus all others. The same level of attention has not been paid to the inclusion of vector databases. Our analyst view is that often, this technology is integrated with other components of the stack, especially DBMS, and while it delivers significant value in an AI technology stack, it's not always upfront in the solutions being used but powering them instead.

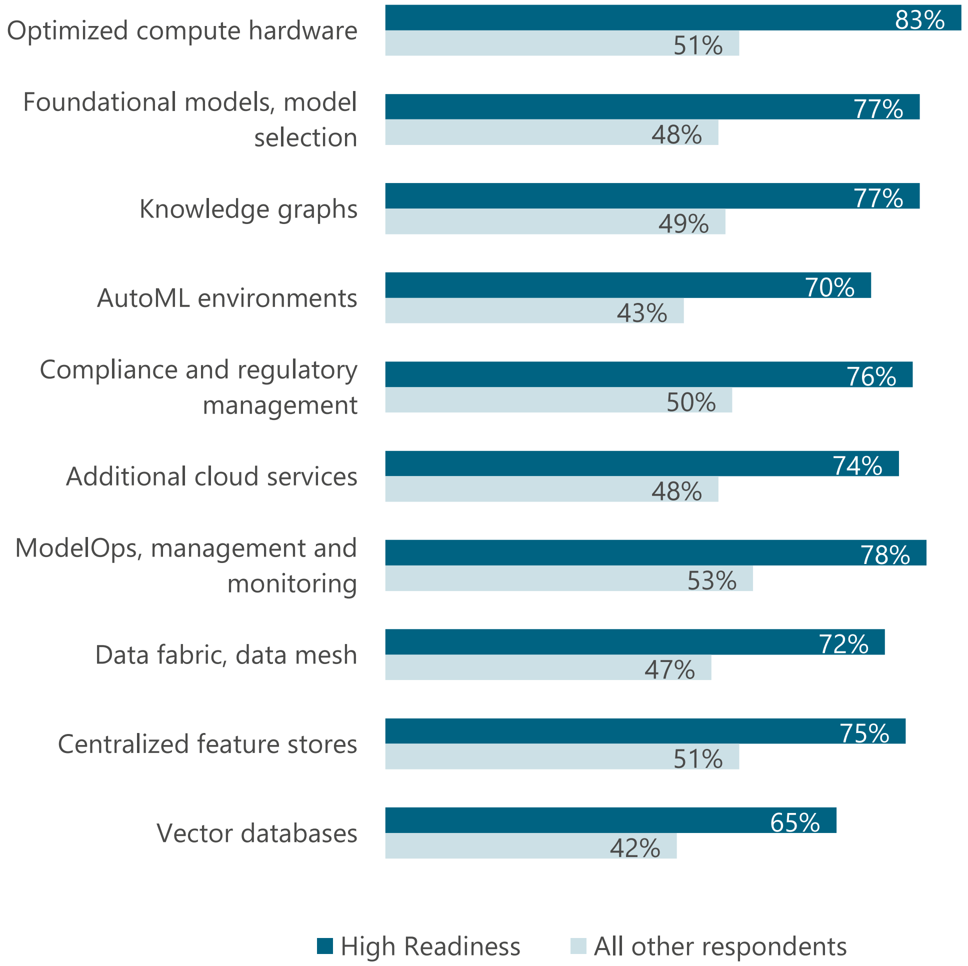

Figure 14: AI technologies with status "Testing/POCs" or "In Production" by Readiness (n=296)

AI Cloud Strategies

Companies utilize cloud platforms for a variety of mission-critical workloads. This research investigated AI use and compared and contrasted how companies are using the cloud today for AI, adding cloud to support AI workloads, and whether they are already using the cloud for data analytics.

The most utilized cloud platforms among respondents are Microsoft Azure at 76%, Amazon Web Services (AWS) at 69% and Google Cloud Platform (GCP) at 61%. Leadership by the "Big Three," as they are often called when referring to Hyperscalers, is not a surprise in these results. General platform utilization is consistent with High Readiness customers, without significant differences. The least utilized for any workloads are Alibaba Cloud at 30%, Baidu Cloud at 24% and Tencent Cloud at 23%. These findings match nearly all global cloud utilization research and our global audience sample trends.

There is a similar trend when examining how respondents rank their use for data and analytics workloads. 38% of respondents use Azure for data and analytics, 30% utilize AWS and 22% are using GCP. There are some differences when examining how High Readiness users adopt these platforms for data and analytics, though their rank is the same. For Azure, it's 41%, AWS is 34%, and GCP is used by 26% of these respondents.

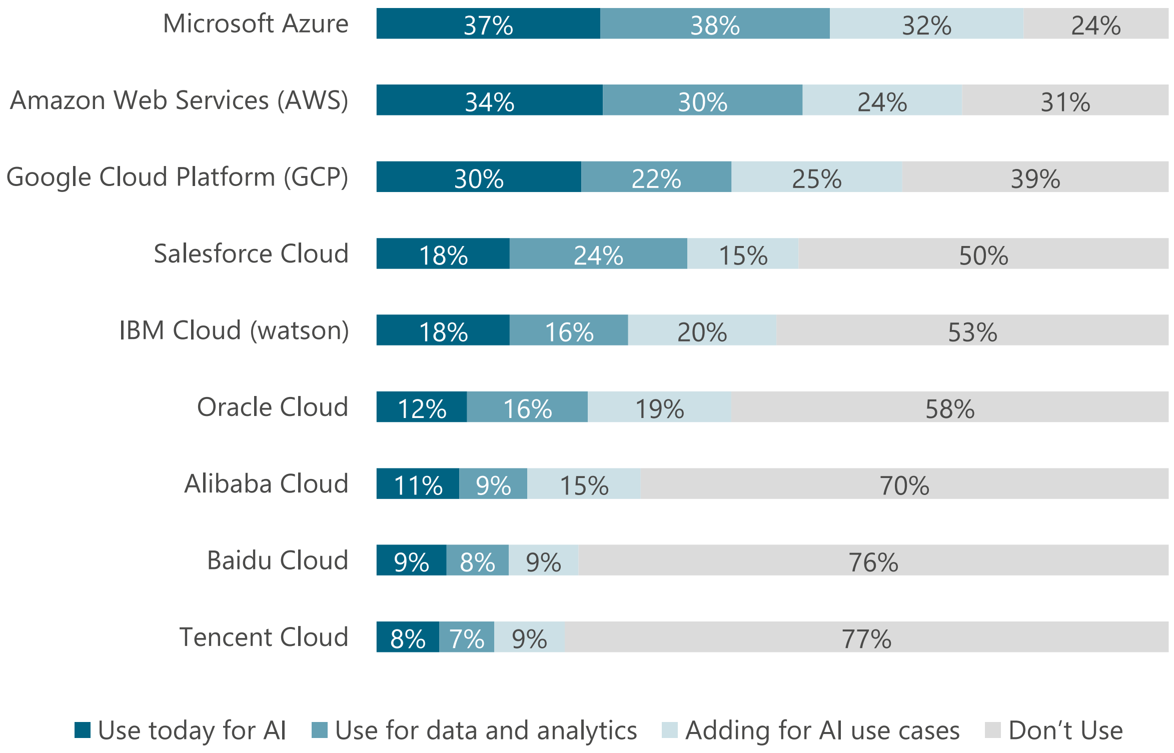

Figure 15: Which of the following cloud platforms does your company utilize for general data, analytics and AI use cases? (n=352)

The most interesting data comes from analyzing how these companies are competing for AI use workloads. Hyperscalers, in general, are leading the race to build new and competing models to power AI. They have the resources, depth of skill sets and money to make this investment. From the user perspective, it is an enticing message to shift AI workloads and often other data and analytics to these platforms, all offering varying degrees of an integrated approach to simplifying AI.

37% of respondents are already using Azure for AI today, and AWS is close behind with 34%. There is a significant gap between Azure and third place with GCP coming in at 30%. High Readiness respondents are closely aligned with these results except for GCP, where we see higher usage at 37.5%.

The final measurement of success focuses on where companies are adding new AI use cases. Azure remains in the lead, with 32% of respondents growing their AI footprint on this platform. AWS and GCP are falling behind, with a significant gap of up to 8 percentage points. This gap is not helped by High Readiness respondents as their planned growth is mainly focused on Azure. Taking the overall total addressable market (TAM) for cloud AI into consideration, estimated to be well over 62 billion dollars, this gap should be concerning for the market laggards.

Oracle, IBM and Salesforce are in the middle of the pack, and our conversations with end users and vendors indicate they remain focused on enabling their existing clients with greater value through AI across their platforms.

Data and the Dreaded Model Choices

Data is King

Structured data (61%) leads the data type rankings. It is followed closely by real-time data (56%, with time series at 43% and sensor data at 29%)—although streaming data comes in near the bottom with 27%. The surprise here is that geospatial data was cited less frequently than any other response. Anyone who has seen vendor demos in use case stories—they have been hard to avoid at conferences—will recall how frequently this data is at the heart of solutions in many categories.

It is also worth noting that despite years of hype, commercial product development and marketing, and widespread sales efforts to address it, unstructured (and/or semi-structured—they are often conflated) data is only cited by a third of respondents. The story for many organizations will begin with their structured data—and that will actually be likely to accelerate progress because of the metadata, quality management functions and governance capabilities provided by the database management software used for it.

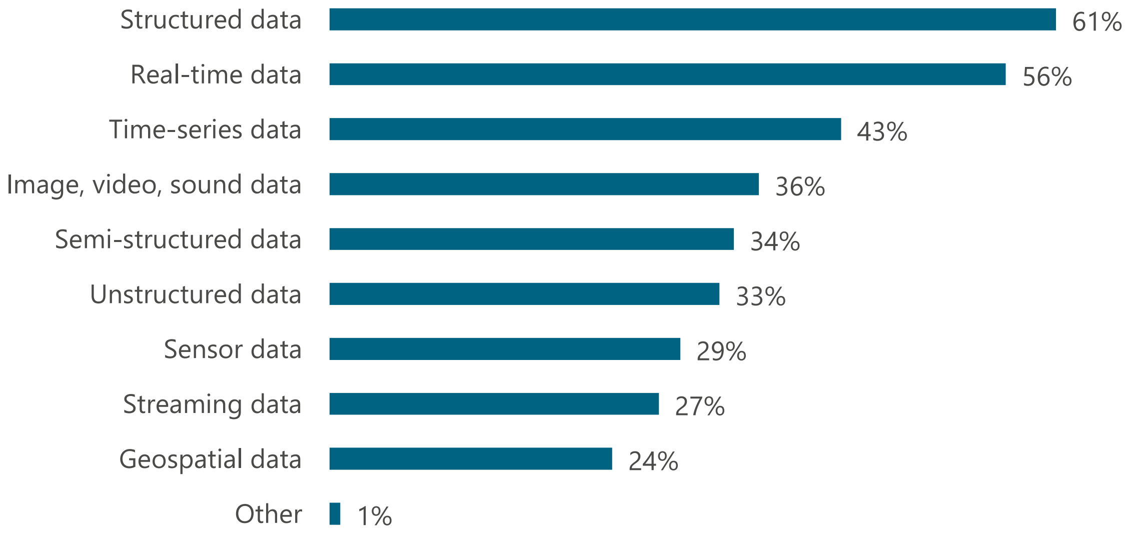

Figure 16: What data types are critical to your AI innovation? (n=335)

Data's role in these projects is central for obvious reasons, not least of which is how it informs, enriches and validates models. But models themselves are data, and the processes associated with them mirror those in broad use for data today. Organizations can leverage this knowledge and skill set to their advantage as they initiate a new layer of architecture.

Biggest Doesn't Always Equal Best with Models

We are witnessing a global arms race initiated by OpenAI with ChatGPT in the LLM and foundational model sector. The number of models available has grown exponentially—new models premier daily on Hugging Face and other centralized distribution points. The Hugging Face community hosts over 400,000 models to browse. In a short number of months, it has become clear that there will not be "one model to rule them all."

Our respondents identified Google Gemini and BARD (renamed to Gemini in February 2024) at a combined 57% as the top choice for LLMs, followed by GPT 3.5 plus 4.0 at 53%. Coding models continue to find a specialized home in many environments, serving development and data science teams at 19%. High Readiness respondents submitted similar responses but showed a significantly higher level of interest in GPT 3.5 and 4.0, with 70% choosing this model, 17 percentage points higher than the rest of the respondents. Early adopters are very likely to have this pioneering model in their environment. Only 3% of High Readiness users were still researching what models they plan to use, in comparison to the 12% of standard respondents.

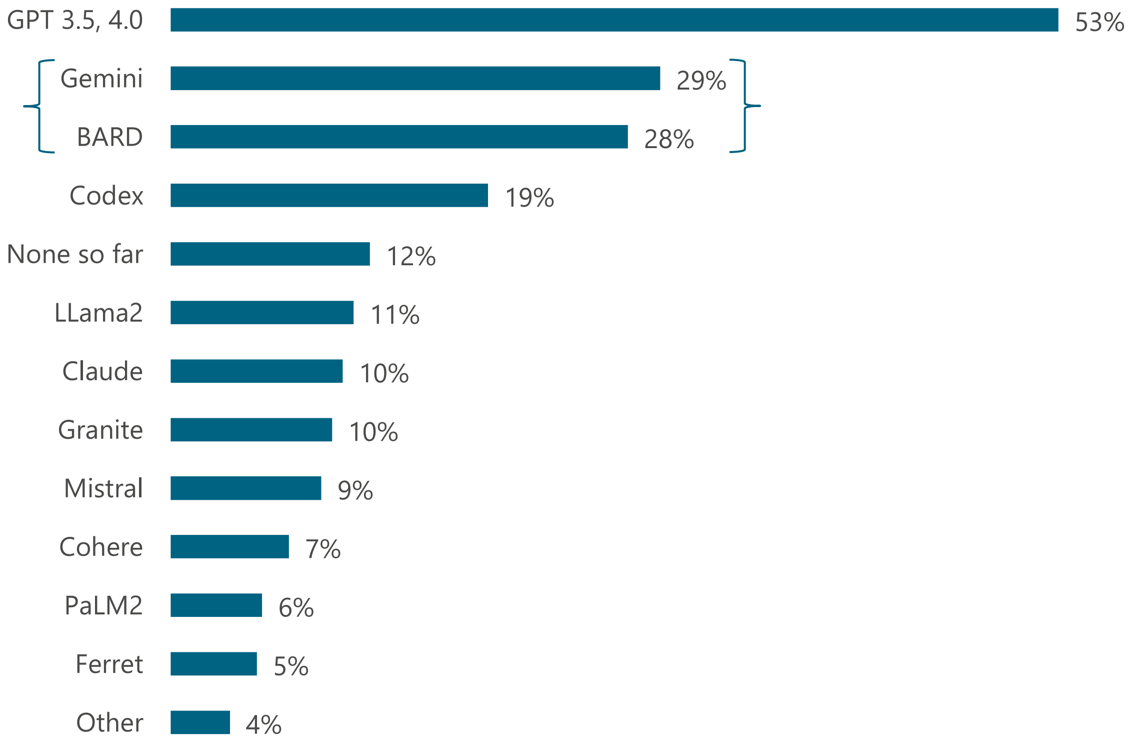

Figure 17: Which of the following large language models (LLMs) and foundational models are part of your company's AI strategy? (n=309)

Early adopters are already evolving toward multi-model environments to align the work with the model capabilities, and it is also clear that there will be a variety of model sizes in most sophisticated environments. Companies are aligning narrow models to specific work and medium and larger-performing models to cover wider-scoped projects.

Specialty Models in Your Ecosystem

21% of respondents indicated that domain-specific models are not a priority at this time. However, that number drops when analyzing High Readiness respondents. Only 16% of these users are not planning to leverage domain-specific models. Finance models at 40% and code models to assist developers at 39% have generated significant interest with standard respondents and are even more popular with the High Readiness group, where finance models are at 55% and code models are at 49%. Again, the surprise was that medical (13%) and biomedical (10%) came in lowest for the standard group and nearly the same for the High Readiness cohort.

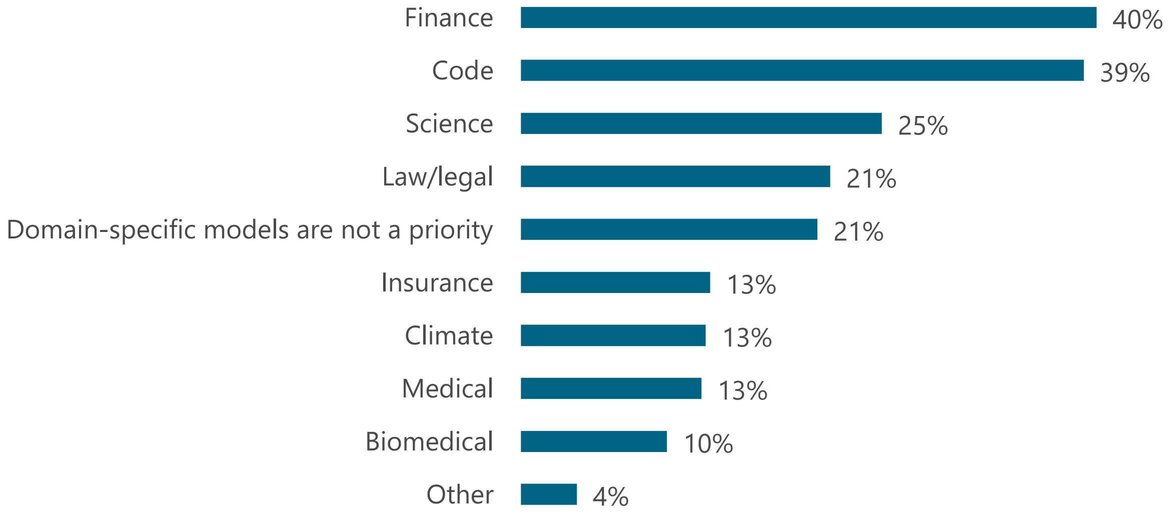

Figure 18: Which industry or domain-specific models are included in your AI model strategy? (n=335)

Model Management is a Critical Strategy for Success

Effectively managing AI models allows companies to better manage quality and accuracy, leading to stronger decision-making and customer experiences. In a world of significant regulation and compliance pressures, managing and understanding the model and the information within a model is a requirement for growing and innovating with AI.

Monitoring and maintenance of models is a foundational capability. For overall model management, survey respondents were equally split on executing that work in-house (33%), on third-party platforms (32%) or a hybrid of both (31%). Respondents defined as High Readiness are managing monitoring and maintenance primarily from a hybrid mix at a rate of 42%, 15 percentage points more than standard respondents. This is similar to the results for model benchmarking, where 30% are executing this in-house, 32% are relying on third-party platforms and 30% cite hybrid strategies. Meanwhile, High Readiness respondents are 8 percentage points more likely to leverage a hybrid strategy.

Prompt engineering, the skill of creating and refining input prompts to achieve optimal model outputs, is a growing competency. Highly proficient prompts can drive quality, accuracy and overall output relevance. For companies that include prompt engineering in their plans, the majority (35%) are doing this work in-house. High Readiness respondents leverage an in-house strategy at a rate of 38%.

RAG or retrieval-augmented generation is the technique of retrieving information from often proprietary enterprise sources to enhance the “knowledge” of large language models (LLMs) and decrease “hallucinations.” Coupling this highly valuable and nuanced corporate information with the model allows it to respond in a more relevant and diverse way. Survey respondents (35%) are primarily relying on third-party solutions to accomplish this task. 14% of respondents have yet to deploy this sophisticated approach to model management, and that is likely driven by the use cases and adoption rates of these respondent companies. High Readiness respondents are only slightly ahead in RAG adoption but are more likely to execute it in-house or in a hybrid configuration.

Fine-tuning allows companies to adapt pre-trained AI models to their specific needs and contexts. This process is important because it enables businesses to tailor the model's responses, style and focus to align with their brand, objectives and customer expectations. By fine-tuning models on proprietary data or specific use cases, companies can significantly improve the performance, relevance and effectiveness of AI applications, leading to better outcomes and higher customer satisfaction. Most survey respondents, including High Readiness respondents, do this continuously or sporadically as required. The largest segment (34%) rely on third-party platforms for this type of work.

There is a significant cost involved that needs to be considered in ongoing AI strategies when utilizing RAG and fine-tuning. Compute resources, data preparation, storage, integration and API calls are all part of this scenario. Utilizing third-party platforms, i.e., cloud environments, can result in surprising bills. The return on investment (ROI) of models that produce fewer hallucinations and highly relevant responses is clear. This needs to be balanced with the costs of achieving it.

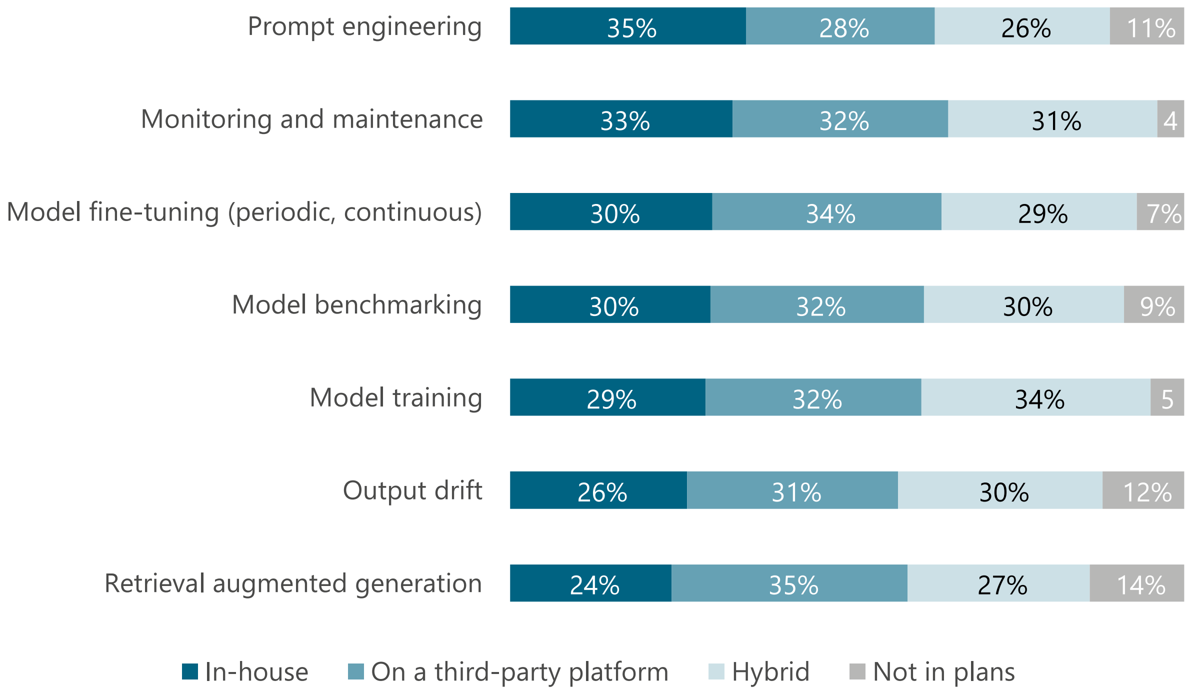

Figure 19: How will your organization manage LLMs? (n=289)

Methodology and Demographics

This research was designed to explore the impact of AI and GenAI on enterprise companies. Its goal is to understand better how prepared these firms are to take advantage of this disruptive technology, with a specific focus on how users are optimizing their technology architectures to leverage AI.

The study employs a quantitative approach utilizing survey responses from a qualified panel of respondents. The design of the survey instrument allows for a detailed exploration of participant subject matter expertise. It focuses on AI maturity, challenges to success, funding, strategy, technology gaps and strategies, and other critical components of leveraging this type of technology.

The respondent population is global with significant European and North American audiences, and a total sample size of 335 completed surveys. Random sampling was used to ensure representation across different job roles, company sizes and industries. Participants were included based on their knowledge of their company's AI strategy and projects. Those with no project experience were excluded.

Respondent Panel

Survey results were drawn from business and IT functional professionals whose job titles included CXO, VP/EVP/SVP, Senior Director, Director, Manager, Engineer and Analyst.

Respondents represent a wide variety of industries, including information technology, manufacturing, retail, financial services, healthcare and others.

Respondents represent companies of all sizes based on company worldwide employee size and annual revenue.

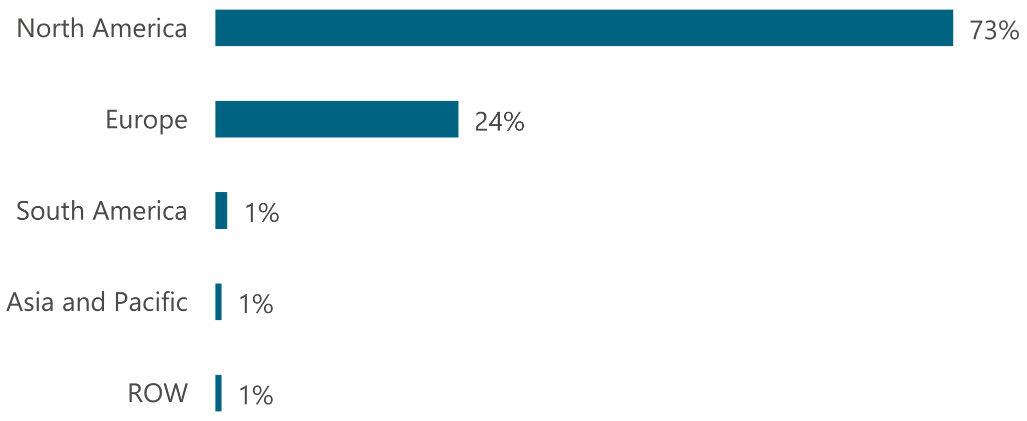

Figure 20: In which region are you located? (n=335)

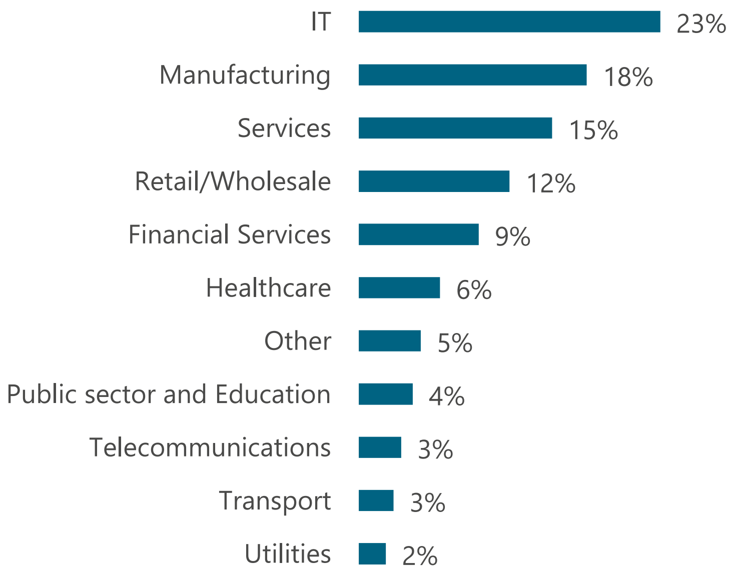

Figure 21: Select the primary sector served by your organization (n=335)

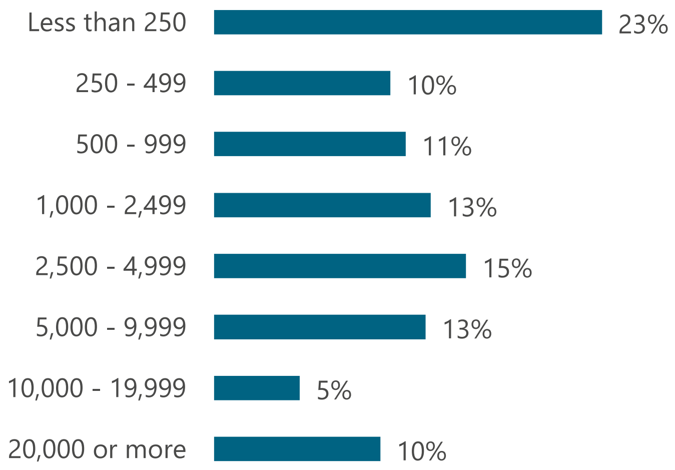

Figure 22: How many employees does your organization have worldwide? (n=335)

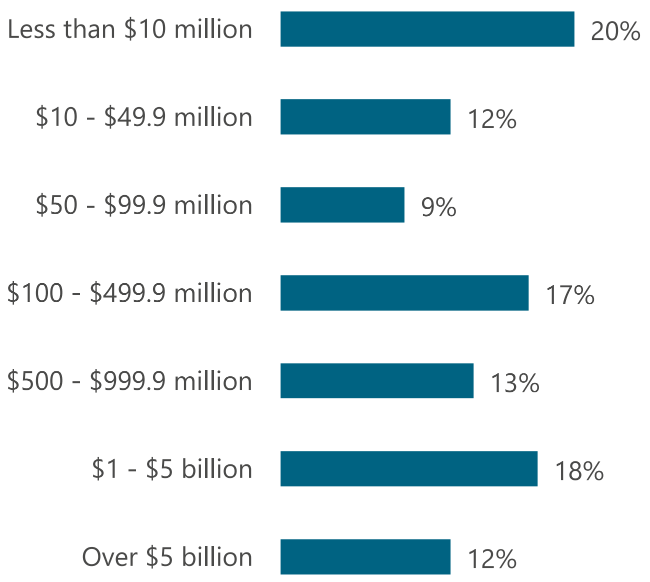

Figure 23: What is the annual revenue range that best fits your organization? (n=317)

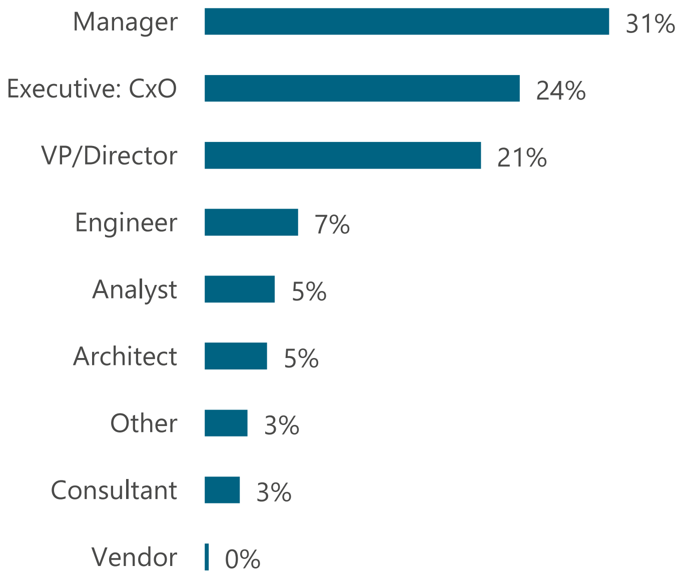

Figure 24: What best describes your job level? (n=335)

The survey was administered online, and data collection occurred over a three-week period starting in February and ending in March 2024. The survey was also quality-tested in the field, with a test phase to refine clarity and ensure response quality.

The survey instrument was designed by Shawn Rogers and Merv Adrian and deployed by BARC GmbH.

Authors

Shawn Rogers

CEO BARC, US, and BARC Fellow

For over 28 years, Shawn Rogers has been an internationally respected industry analyst, thought leader, speaker, author and instructor on data, business intelligence, analytics, AI/ML and cloud technologies.

His former executive strategy roles with Dell, Statistica, Quest Software and TIBCO Software give him a unique perspective on the enterprise software industry. As a CEO of BARC US and BARC Fellow, he is responsible for analyst coverage of these technologies in the North American market.

Merv Adrian

Founder and Principal, IT Market Strategy

Merv Adrian, founder and principal at IT Market Strategy, is a trusted advisor to leading software firms and investors, a frequent conference speaker, and a prolific researcher, editor and writer on IT issues. He is a former Research Vice President of Gartner, Senior Vice President of Forrester Research and Research Manager at Giga Information Group, specializing in data management.

Mr. Adrian began his career at the Federal Reserve Bank and was an independent programmer and consultant before moving into the software industry as a technical journal editor and strategic marketer with Information Builders and Sybase before becoming an analyst in 1997.

About BARC

BARC (Business Application Research Center) is one of Europe's leading analyst firms for business software, focusing on the areas of data, business intelligence (BI) and analytics, enterprise content management (ECM), customer relationship management (CRM) and enterprise resource planning (ERP). Our passion is to help organizations become digital companies of tomorrow. We do this by using technology to rethink the world, trusting data-based decisions and optimizing and digitalizing processes. It's about finding the right tools and using them in a way that gives your company the best possible advantage. This unique blend of knowledge, exchange of information and independence distinguishes our services in the areas of research, events and consulting.

Research

BARC studies are based on internal market research, software tests and analyst comments, giving you the security to make the right decisions. Our independent research brings market developments into clear focus, puts software and vendors through their paces and gives users a place to express their opinions.

Events

Decision-makers and IT industry leaders come together at BARC events. BARC seminars in small groups, online webinars and conferences with more than 1,000 participants annually all offer inspiration and interactivity. Through exchange with peers and an overview of current trends and market developments, you will receive new impetus to drive your business forward.

Consulting

In confidential expert workshops, coaching and in-house consultations, we transform the needs of your company into future-proof decisions. We provide you with successful, holistic concepts that enable you to use the right information correctly. Our project support covers all stages of the successful use of software.

BARC

About Aerospike

Contact info

Aerospike
2440 W. El Camino Real, Suite 100
Mountain View, CA 94040

(+1) 408-462-AERO (2376)

info@aerospike.com

aerospike.com

Aerospike is the real-time, multi-model database built for infinite scale, speed, and savings. Blazing fast and reliable, Aerospike performs on gigabytes to petabytes of data with sub-millisecond latency, so enterprises can make better decisions faster with new and constantly changing data. Built for efficiency and sustainability from the start, Aerospike enables organizations to operate on a fraction of the infrastructure required from legacy databases, helping them reduce their server footprint by 80 percent to drive lower all-around costs and significantly reduce carbon emissions.

Aerospike customers are ready for what's next with the lowest latency and the highest throughput data platform and no re-platforming required even as their data grows. Aerospike serves organizations as a catalyst for progress, believing that there are no limits to how businesses can innovate with their data.

Cloud and AI-forward, Aerospike empowers businesses like Adobe, Airtel, Criteo, DBS Bank, Experian, PayPal, Snap, and Sony Interactive Entertainment to optimize for the cloud and expand their use of transformative AI applications. Aerospike is headquartered in Mountain View, California, with offices in London, Bangalore, and Tel Aviv. aerospike.com.

BARC

Data Decisions. Built on BARC.

www.barc.com

Germany
BARC GmbH
Berliner Platz 7
D-97080 Würzburg
+49 931 880651-0

Austria
BARC GmbH
Hirschstettner Straße 19 / I / IS314
A-1220 Wien
+43 660 6366870

Switzerland
BARC Schweiz GmbH
Täfernstr. 22a
CH-5405 Baden-Dättwil
+41 56 470 94 34

United States
BARC US
13463 Falls Drive
Broomfield, CO. 80020
USA
+01 720-381-4988

IT MARKET STRATEGY

www.itmarketstrategy.com

United States
39654 Gladiolus Ln
Palm Desert, CA 92211