Wallaroo.AI

Lessons from the Leading Edge:

Machine Learning Best Practices and Challenges from Chief Data Officers Already in Production

Table of Contents

Foreword

Nina Zumel Ph.D.
Vice President,
Data Science
Wallaroo.AI

The use of machine learning (ML) has become more widespread in industry over the past several years, and is expected to increase monumentally over the immediate future, as companies discover new ways to unlock the value of ML- and AI-driven processes. Medium to large companies across a variety of industries are currently spending millions and even billions of dollars a year on ML, and many expect to at least double (and some even to quadruple) their spend over the next three years. Common current use cases include marketing, personalization, and security; we expect that AI-driven document processing, analytics, and generation will also become more widespread in the near term.

But advances in these more sophisticated methods often come at the cost of increased complexity and its attendant challenges. Companies struggle to find talent with the necessary skills to develop these technologies and transition them effectively and efficiently into production. Other challenges include integrating ML tools and artifacts into production stacks, as well as managing ML tools, infrastructure, and costs.

In order to help companies meet these challenges and get the most value from their ML and AI efforts, Wallaroo.AI commissioned this survey to better understand the state of the industry, as seen by the executives responsible for these efforts: Chief Data Officers (CDO) and Chief Data Analysis Officers (CDAO). These insights will help us to build a platform that best addresses the most important challenges, and so accelerate AI adoption and unleash its attendant value.

CHAPTER 1

About the Research

Just like any bell curve, many organizations today that are interested in AI have not fully adopted the technology yet. They are still working on training datasets or setting up prototypes.

To help AI leaders in the next phase of their adoption, this report focuses on feedback from people who have already put ML models into production at scale, then monitored and optimized models to provide value for the enterprise quickly.

Diving Into Key Areas of ML Production

From team dynamics to production ML tactics, we sought to understand how CDOs and other ML leaders are generating value from their AI initiatives so we could share those insights with you.

We will review some highlights in the Executive Summary but now we'll do a deep dive into key areas. Topics covered are related to deployment, scaling, monitoring, and optimization of ML models in production.

The results should help CDOs and other ML leaders better understand the following areas:

Optimal AI Strategy

How leading-edge teams are assembled, the processes they employ to facilitate ML, the importance of automation in their processes, etc.

Deployment

How they run ML models, the time it takes their teams to build and deploy models, and the challenges therein.

Scaling

What they've found to be the most effective ways to scale their ML models, and how important scaling is to the effectiveness and future of their ML initiatives.

Monitoring

How they observe and monitor their models, and how they feel they can improve observability.

Optimizing

How well they respond to model drift and optimize models while in production.

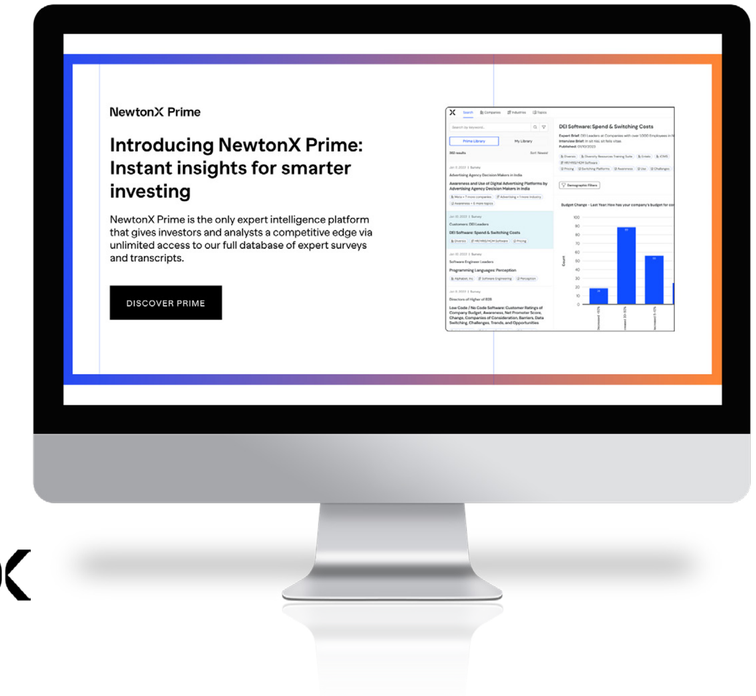
Growth

How these leading-edge enterprises are looking to grow AI/ML at their organizations, from expanding their teams to increasing the scope of ML in decision-making.

The Scientists Behind the Survey

The survey this report is based on was conducted by the research firm NewtonX, the same firm used by the New York Times and other leading information outlets.

NewtonX is the only B2B research company that solves the challenges of today's insights leaders by connecting them with verified business expertise.



To do this, they built the most sophisticated algorithm in the research industry, the NewtonX Graph. This AI-driven Graph custom-recruits the perfect audience for business questions from an open network of 1.1 billion professionals across 140 industries. Every professional is 100% verified, allowing NewtonX clients to make their next bold moves with confidence.

Decision-quality data is embedded in everything they deliver. They field large-scale quantitative surveys, facilitate qualitative or expert interviews, engage in long-term consultations, and create customized research plans. With their all-access platform NewtonX Prime, they deliver expert intelligence at scale, giving investors an edge via instant access to expert surveys and transcripts.

NewtonX partners with the Fortune 500, top consultancies, marketers, and investors. Together with their clients, they're ushering in a new standard of truth in B2B insights.

To learn more,

head to newtonx.com.

NewtonX

People at the Leading Edge of AI

Research Participants

In order to ensure the data reflects lessons from leading-edge organizations with ML already in production, NewtonX surveyed:

Individuals who are responsible for **business outcomes of AI** initiatives (no consultants)

Who have at least **one model** in production

With titles of **Chief Data Officer/Chief Data Analytics Officer, Vice President, or Director** (title in some of the largest enterprises)

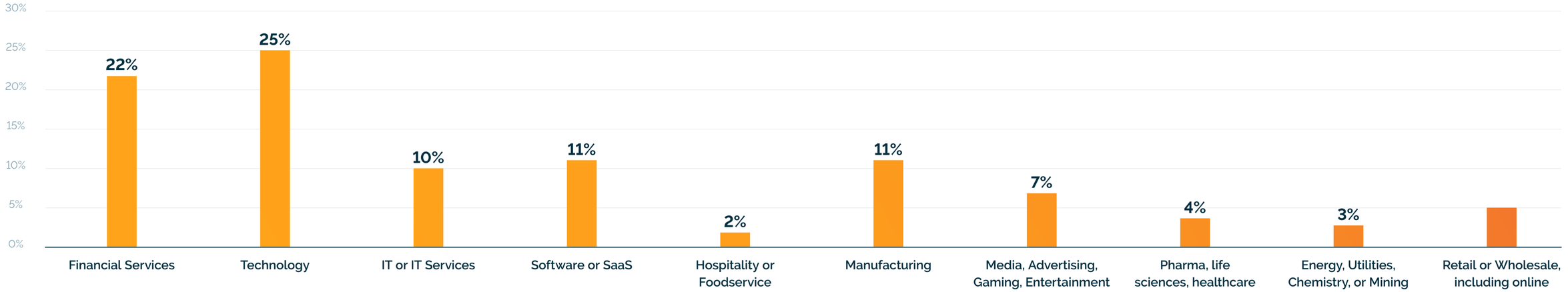
At U.S.-based **private companies** (no government entities)

With **more than 5,000 employees** (71% had than 100 people working in ML, 52% have more than 250)

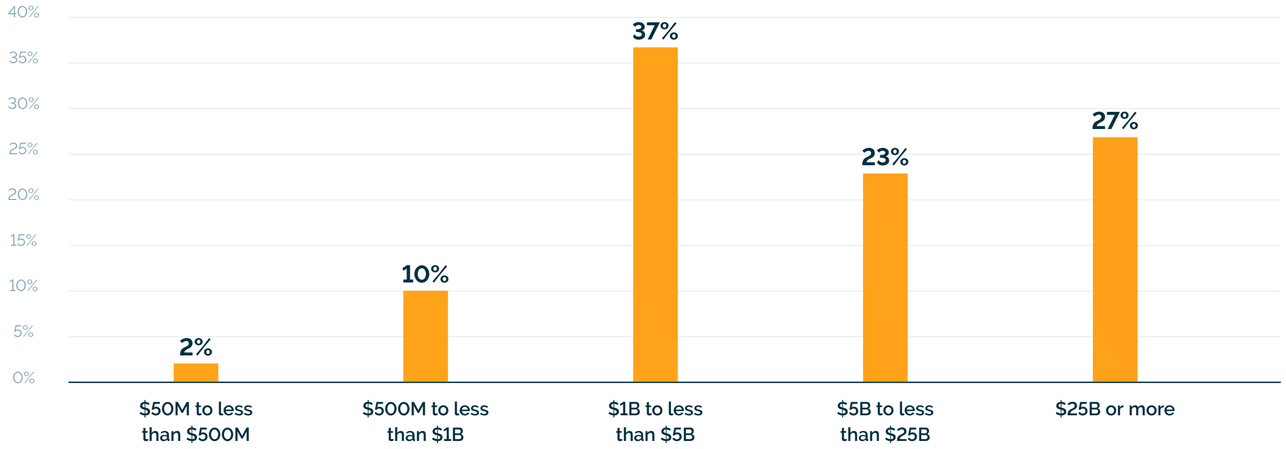
And/or a **revenue** greater than **$500 million** (U.S.)

Respondents came from a variety of industries, 87% working at companies with more than $1Bn in revenue, over half with more than 10.000 employees

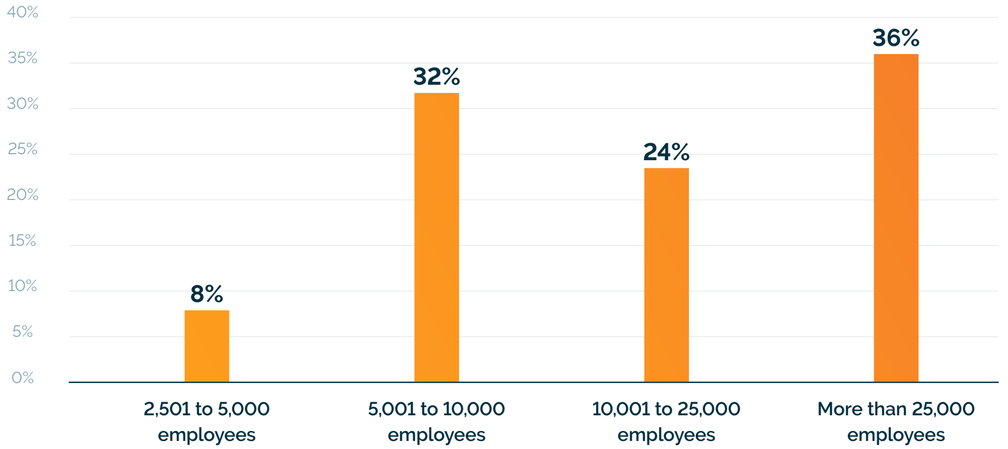
Industry



Revenue



Employees



BACK TO TABLE OF CONTENTS

CHAPTER 2

Executive Summary

For enterprises who are in the next wave of putting ML into production, it is critical to understand what CDOs/CDAOs at leading-edge enterprises are doing to succeed and what is still an obstacle for them.

Executive Summary

Discovering How to Get ROI from Machine Learning (ML)

With a $15.7 trillion (U.S.) potential contribution to the global economy by 2030 (PwC), demand for artificial intelligence (AI) has never been higher. But, artificial intelligence is still a project for early adopters in many cases, without the well-documented best practices and wealth of knowledge of a mature technology.

$15.7 trillion
potential contribution to the global economy by 2030

In fact, Gartner reports that 90% of AI initiatives fail to produce substantial return on investment (ROI), with about half never making past the prototype stage.

of AI initiatives **fail** to produce substantial return on investment (ROI)

But some AI initiatives are in production and generating value for the organizations conducting them.

What are these leading-edge firms doing that allows them to succeed when others do not? More importantly, how can others learn from their best practices and the things that did not work for them?

Wallaroo.AI set out to answer these questions.

NewtonX

We commissioned NewtonX, the world's leading B2B market research company with 100% verified research across 140 industries and 1.1 billion professionals, to reach out to ML leaders at leading-edge firms who have successfully put models in production to find out. This report summarized their findings.

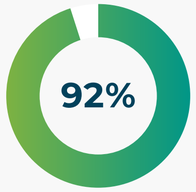
Executive Summary

ML in Production Today

There were several clear themes in the research, regarding the value of ML already in production:

The Leading Edge is Already Finding Value in Their Production ML Initiatives

- Early adopters do not always realize the promise of new technologies. However, despite all the challenges of being first, 92% of AI team leaders at leading-edge organizations felt they are already finding that their AI initiatives are generating value.

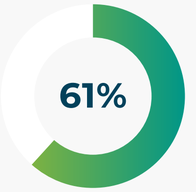


of AI team leaders at leading-edge organizations felt that their AI initiatives are **generating value**

- The value leading-edge enterprises are gaining comes in the areas of personalizing customer experience, fraud detection, optimizing sales and marketing, and improving real-time decision making, among other use cases.

Current Investment and Future Plans Will Fuel Growth, Create Challenges for Others

- Gains from current initiatives are at least partly the result of the strong investment made by leading-edge enterprises.
- Partly due to the perceived value they feel their AI initiatives are already achieving, leading-edge organizations plan to invest significantly in their AI initiatives in the future (61% will more than double investment). They also intend to scale significantly within the next three years.



of leading-edge organizations plan to **investment more than double** in their AI initiatives

- However, this investment (including cloud cycle costs, staffing costs, etc.) and scale will also impact other enterprises seeking the same human resources, vendors who support the ML ecosystem, and other elements of the AI ecosystem.

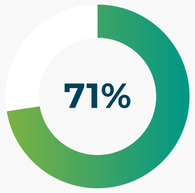
Executive Summary

Key Challenges

Leading-edge enterprises faced consistent challenges:

Addressing the ML Skills Gap

- Skilled ML experts are a critical factor to the success of leading-edge enterprises, according to CDO survey responses on several questions. In fact, 71% of respondents have more than 100 staffers working on ML. However, acquiring and retaining these ML experts represent some of the greatest challenges these organizations faced.



of respondents have more than **100 staffers** working on ML

The Platform is the Challenge

- Being at the leading edge means overcoming hurdles in whatever way you can. For enterprises trying to put ML into production and scale to reach ROI, that often means creating in-house ML frameworks that are later found to be suboptimal. In-house development teams then need to rework the software, which impacts staff learning curves, cost, and time-to-value. But data shows that these are all challenges that CDO survey respondents are already facing, even before they make any platform revisions. Fifty-one percent cited the effort required to integrate tools as a major challenge in getting ML into production, while another 51% said that the number of discrete tools needed throughout the process was a major challenge.

cited the **effort required to integrate tools** as a major challenge in getting ML into production

said that the **number of discrete tools** needed throughout the process was a major challenge

ML Production Best Practices and Challenges by Stage

- Where AI teams choose to deploy their models – most often in the cloud – has a major impact on the value they are able to generate for their organizations. However, leading-edge organizations encounter challenges across the entire ML production process that stem from the ML production framework they use and the intricacies of working in the cloud.

Executive Summary

Takeaways

- The research shows that leading-edge enterprises face many challenges that impact cost and productivity because of the dependency on hard to find and retain ML experts, the choice of cloud, and use of in-house developed ML production frameworks.

- All AI teams will be looking for and trying to retain ML experts but, fortunately, all leading-edge enterprises have strong ideas on how to do so that AI team leaders can use to acquire and retain ML experts that other organizations can use. There is also the option of working with a commercial vendor that has ML experts who can assist.

- To overcome the challenges faced by leading-edge enterprises, two viable options are available to the next wave of AI team leaders. They could adopt a purpose-built ML production platform or they could develop an in-house framework that includes the capabilities of commercial ML platforms. The challenge with an in-house framework that includes the latest innovations is that it would need to operate seamlessly across all phases of the ML production process without taking the significant time necessary to develop and test such a complex technology.

CHAPTER 3

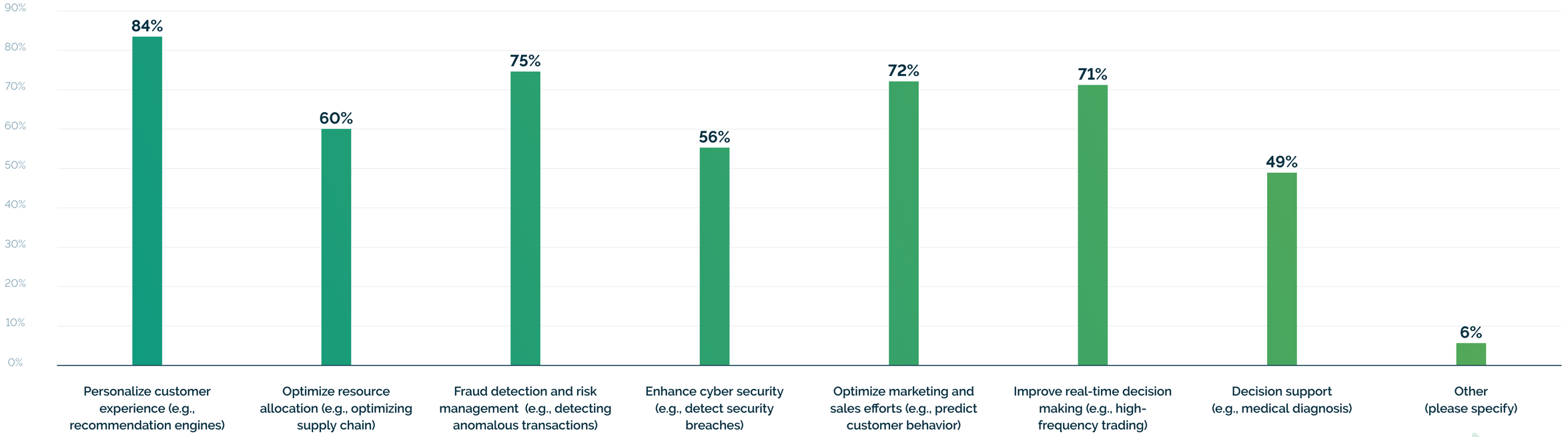
Finding Value in Enterprise ML Today

Early adopters do not always realize the promise of new technologies. However, the leading-edge organizations that responded to the survey showed surprisingly similar responses in this critical area.

Enterprises Use Machine Learning to Achieve Some Common Objectives

Enterprises sought to achieve a variety of goals through their ML initiatives, but several goals were common to over 70% of AI team leaders responding to the survey, including: personalizing customer experience, fraud detection, optimizing sales and marketing, and improving real-time decision making.

Goals of using ML



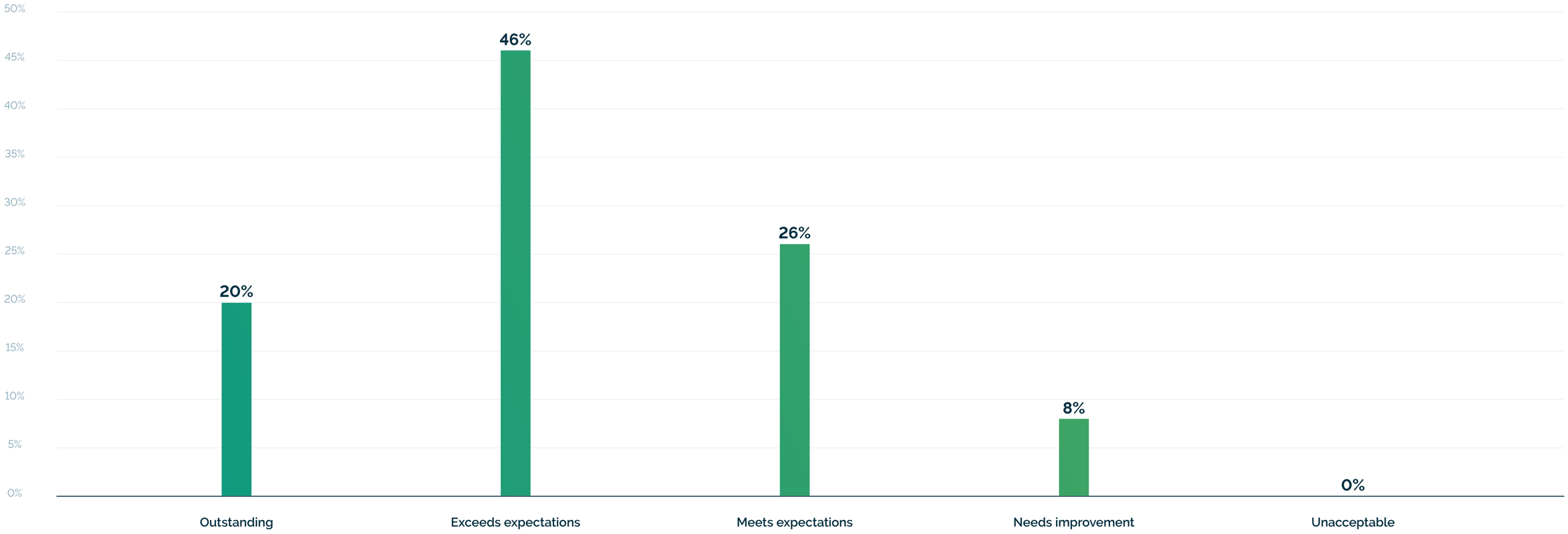
Other reasons specified

- Supply chain optimization
- Enabling better distribution
- Compliance problems
- HVAC, oil and gas, energy, MLO
- Support & predictive maintenance
- Document processing, conversational AI, text analytics

Machine Learning is Already Delivering Value

Though AI is considered to be in its infancy by many, 92% of CDO survey respondents feel that the ML models they already have in production deliver business value, and two thirds even feel the models go beyond expectations.

Are ML models delivering business value?



Takeaways

- One gratifying conclusion from this data is that even now, while AI is still in its early growth phase, there is enough value to enterprises to make AI a good investment, particularly for industries that can gain a competitive advantage from personalizing customer experience, fraud detection, optimizing sales and marketing, and improving real-time decision making.
- The challenge is that not all enterprises can match the size and ability to execute of the companies included in the survey, which may be factors in the ability to realize value from AI initiatives.
- The rest of this report focuses on how other enterprises can leverage the lessons learned by these CDOs in their own AI initiatives.

CHAPTER 4

Leading Edge Continues to Fuel AI Growth, Challenges for Others

The gains leading-edge organizations firms have seen are partly the result of the strong investments they have been making in AI historically.

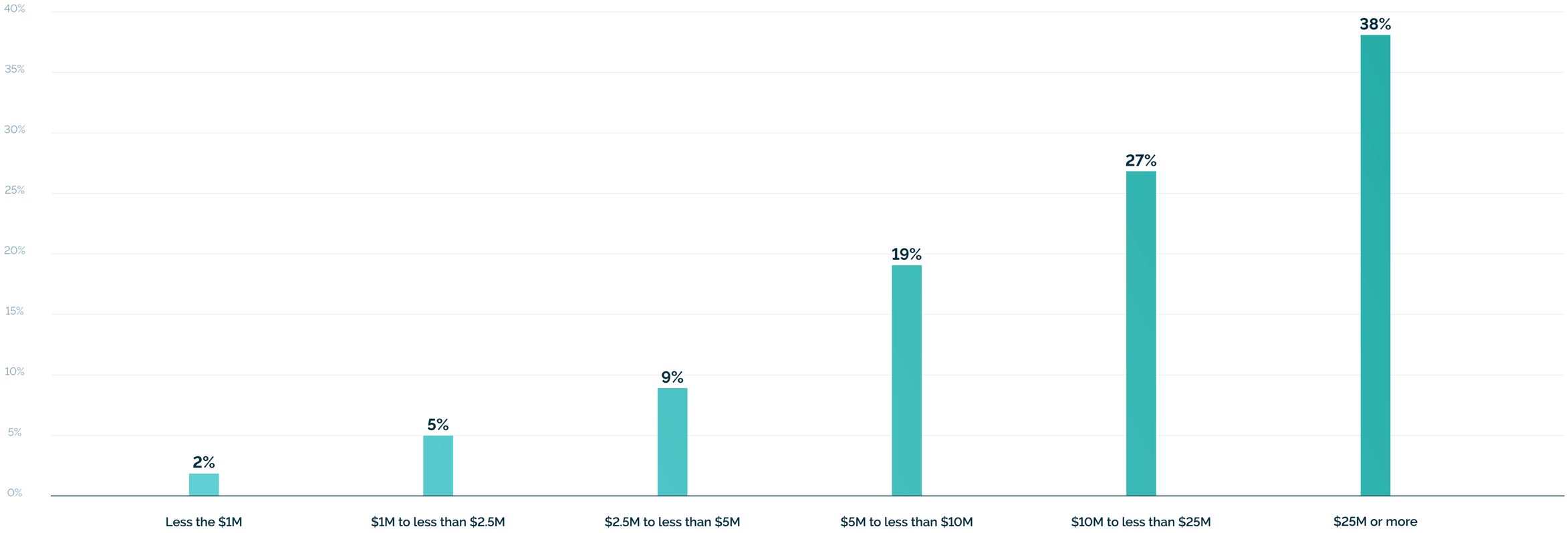
Based on the perceived potential value of AI - which was surprisingly high - they also plan a great deal of future investment in their AI initiatives.

This planned future spending and growth will impact not just the growth of their particular AI initiatives. It will also impact other enterprises seeking the same human resources, vendors who support the ML ecosystem, and many more elements of the AI ecosystem.

Significant Investments in AI at the Leading-Edge

Eighty-four percent of individuals responding to the survey say their firms are currently spending more than $5M (USD) per fiscal year. Two thirds are spending more than $10M (USD) a year on ML. Over a quarter are already spending more than $25M (USD), a significant amount even for a Fortune 500 firm. These levels reflect staffing, equipment, software, and other elements enterprises use to support their AI initiatives.

Company spend on ML latest fiscal (USD)

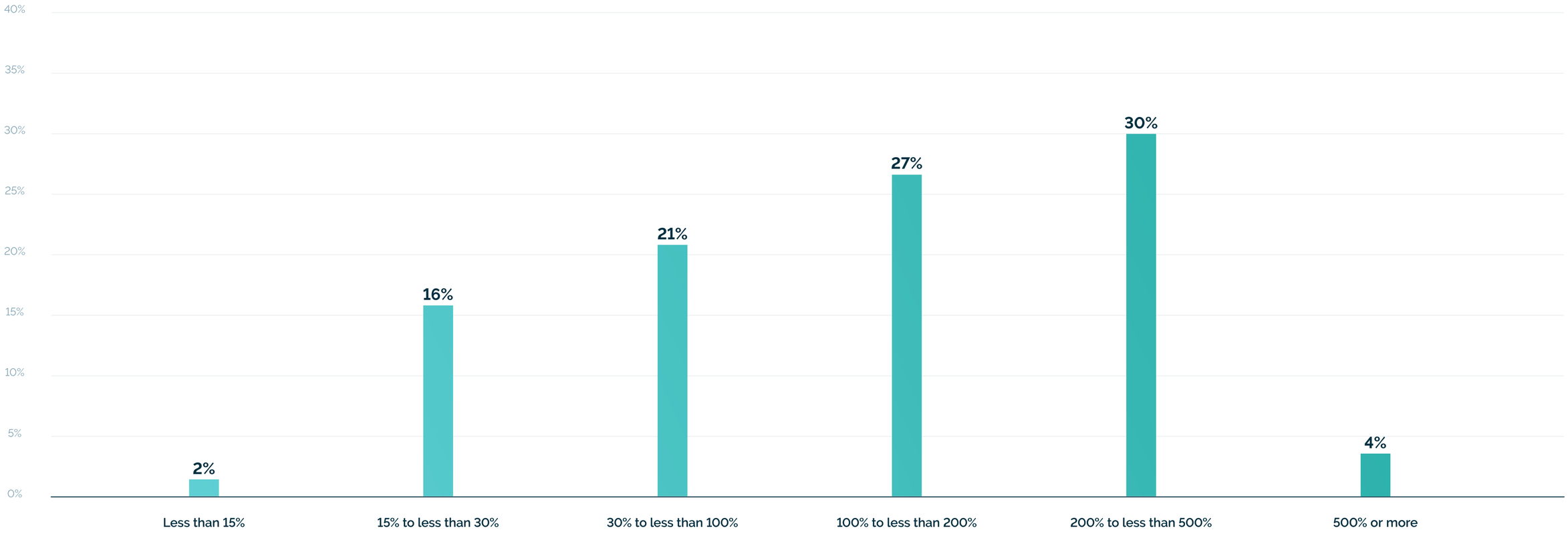


BACK TO TABLE OF CONTENTS

CDOs Plan Even Greater ML Investment

Despite news coverage of a potential economic recession, 61% of CDO survey respondents anticipate their firms will be at least doubling their spend on ML in the next three years, while 34% expect to at least quadruple their spend. This investment is likely related to the high existing value CDOs already feel their AI initiatives are already contributing to their enterprises, as well as the increased potential they see from scaling and optimization.

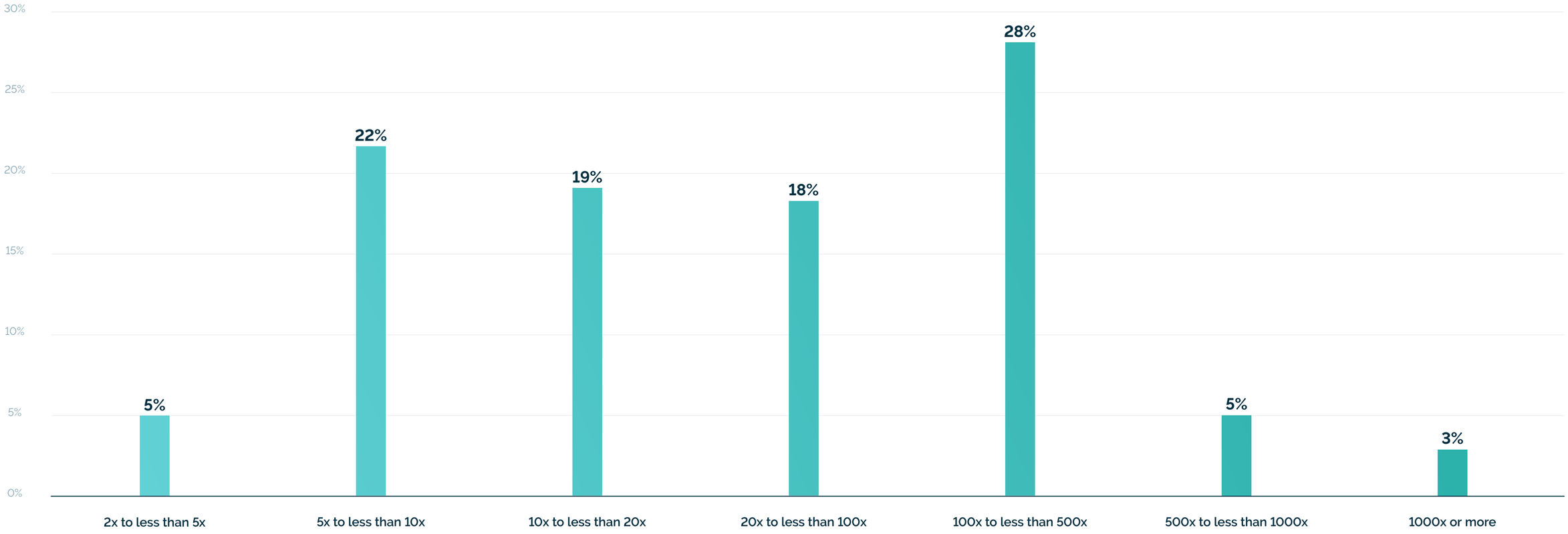
Expected growth in spend on ML next three years



Scaling Growth

The value of AI is so strong already that leading-edge enterprises plan to scale significantly over the next three years. Almost all companies expect to scale their ML models more than five times in the next three years and 36% plan to scale over 100x. This planned growth helps explain the planned spending. However, it also suggests that leading-edge firms believe the ML production challenges they face are issues they can overcome by making some changes.

Scaling goal next three years



BACK TO TABLE OF CONTENTS

Takeaways

- Beyond the day-to-day costs of cloud usage, payroll for ML engineers and data scientists, and other recurring costs that are a universal part of the ML production process, many large enterprises on the leading-edge have been forced to build their own production ML frameworks.
- Thus, the cost to hire developers, potentially buy point solution software (and specialized consultants familiar with that software) for key elements of the ML production process, and other expenses are also included in the total spend. Similarly, costs for expanding AI initiatives (including additional development, integration, training, etc.) are included in the forecasts.
- Fortunately, the market is maturing quickly and other firms looking to put their models into production can eliminate some of these platform-related costs by implementing a turn-key framework such as that offered by Wallaroo.AI. They can also use these learnings to plan out and develop a comprehensive, integrated ML framework.
- Beyond the platform-related costs, however, clients who have built their own ML production platforms, noted that there were several important – and challenging – downstream consequences to "do-it-yourself" ML production frameworks that carried costs of their own. These are highlighted in subsequent sections of this report.

CHAPTER 5

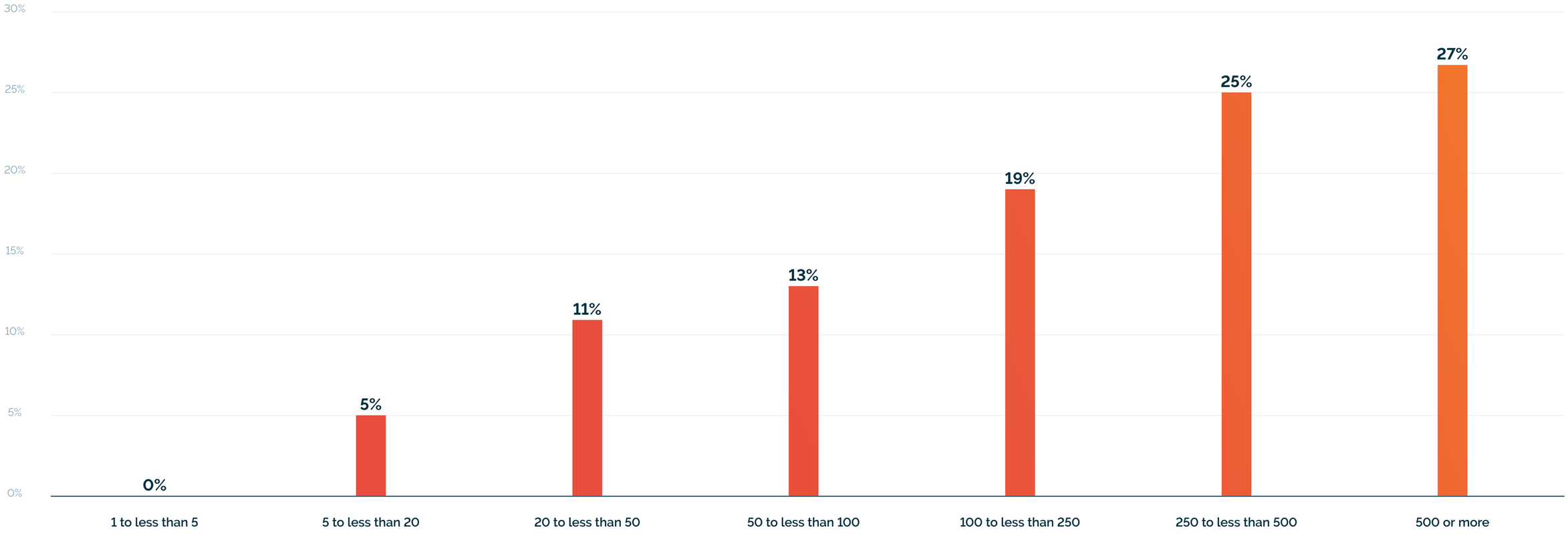
The Unique Challenge of ML Experts

Skilled ML experts were such a critical factor to ML success, according to CDO survey responses on several questions, that we wanted to call out the topic for a deeper discussion. (We'll do a deeper dive into other challenges in subsequent sections.)

Current Production ML Needs Significant Staff

ChatGPT and the subsequent rush to adopt generative AI (or build a competitor) has restarted the AI stampede. However, ML production is still in its infancy for most organizations, even for some large enterprises. Nevertheless, 71% of respondents from leading-edge enterprises say their companies have more than 100 people working in ML – and 52% have more than 250 people working on ML. These high numbers may be due in part to the high effort required to build ML platforms in-house (which many firms do today) and then manually optimize processes and the models.

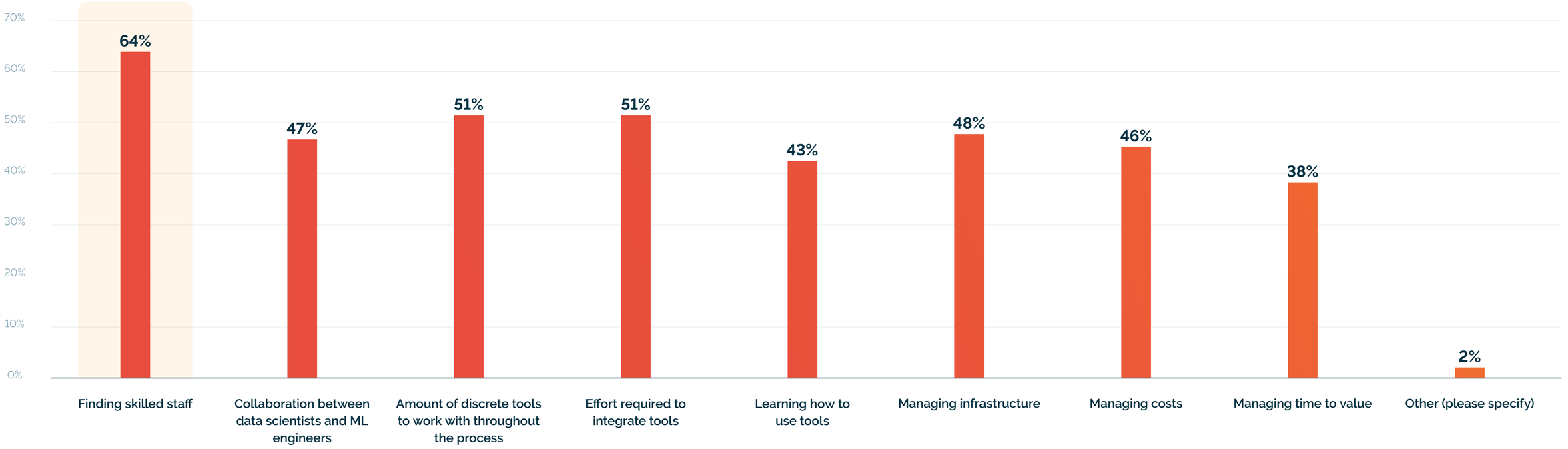
People working in ML



ML Experts: The Greatest Challenge to Production ML

The single most frequently cited challenge to running production ML, according to survey respondents, was finding skilled staff. Not only did two thirds of respondents say finding ML experts was the greatest challenge, they provided details on how they are dealing with this issue, including retaining existing employees (ranked issue number 1). If large, leading-edge enterprises are having such trouble finding the experts necessary to execute production ML, its logical to assume that other organizations will also face this challenge. It may be even worse for smaller firms or those that wait to adopt AI. Some of the other challenges to production ML also have a potential impact on staff retention, which was the number 1 way leading-edge enterprises are addressing the staffing challenge.

General challenges with production ML



ML Experts: The Greatest Challenge to Production ML (cont.)

Actions to deal with challenge of finding skilled staff (ranked)

1.

Focus on retention of own employees

2.

Focus on being employer of choice

3.

Active recruitment straight out of college/university

4.

Training technical or operational skills

5.

Offering internships for college students

6.

Offer more financial incentives

‘Other’ responses

- Explaining value to business
- Passing risk gates

In Their Own Words:

Beyond responses to specific questions, we asked survey respondents to share anything else they felt was relevant.

> *Hiring skilled ML professionals and retaining them can be expensive, especially given the demand for expertise in the field*
>
> — Healthcare Technology Software Firm

Takeaways

- The recent change in employee-employer dynamics and the high wages being offered by top firms seeking to cash in on the excitement around generative AI may be contributing to the challenge CDO survey respondents are facing in finding skilled ML production staff. But, however challenging the issue is for large, leading-edge firms that are already well-known among practitioners, the staffing challenge is even greater for other organizations.

- Automation is certainly one way to reduce the amount of staff hours that must be found (and paid for) to support AI initiatives, as fewer staff hours are required to support mundane efforts. In fact, removing time devoted to routine, repeatable tasks via automation freed up 40% of ML engineer and data scientist time for Wallaroo.AI clients. Automation also reduces the potential for human error.

- In addition, automation also creates an opportunity for practitioners to spend more time focusing on the more skilled portions of the ML production process (such as model optimization or scaling) which generate value for the company.

- This focus on activities that more directly impact the enterprise's bottom line may contribute to greater employee satisfaction, which could in turn help employee retention efforts (which survey respondents said was their #1 way to combat the need for skilled ML experts) given that keeping AI talent fully engaged is a critical factor in retaining these unique employees, according to Boston Consulting Group[1].

- One item that was notably absent from the list of methods leading-edge organizations are using to deal with the challenges of finding skilled staff is outsourcing. This is likely due to the complex nature of in-house development and the very small set of skilled AI consulting organizations.

- However, for organizations that are in the early stages of ML production or are willing to overlay a framework on top of existing in-house work, the opportunity to take advantage of a turn-key ML production platform (especially one that addresses all the stages from deployment to scaling to monitoring to optimization), there are some vendors such as Wallaroo.AI who also have ML experts available to support the client's ML production process.

[1] "How to Attract, Develop, and Retain AI Talent", Boston Consulting Group, 2023

CHAPTER 6

The Platform is the Challenge

Being at the leading edge means overcoming hurdles in whatever way you can. For enterprises trying to put ML into production and scale to reach ROI, that often means creating in-house ML frameworks that are later found to be suboptimal.

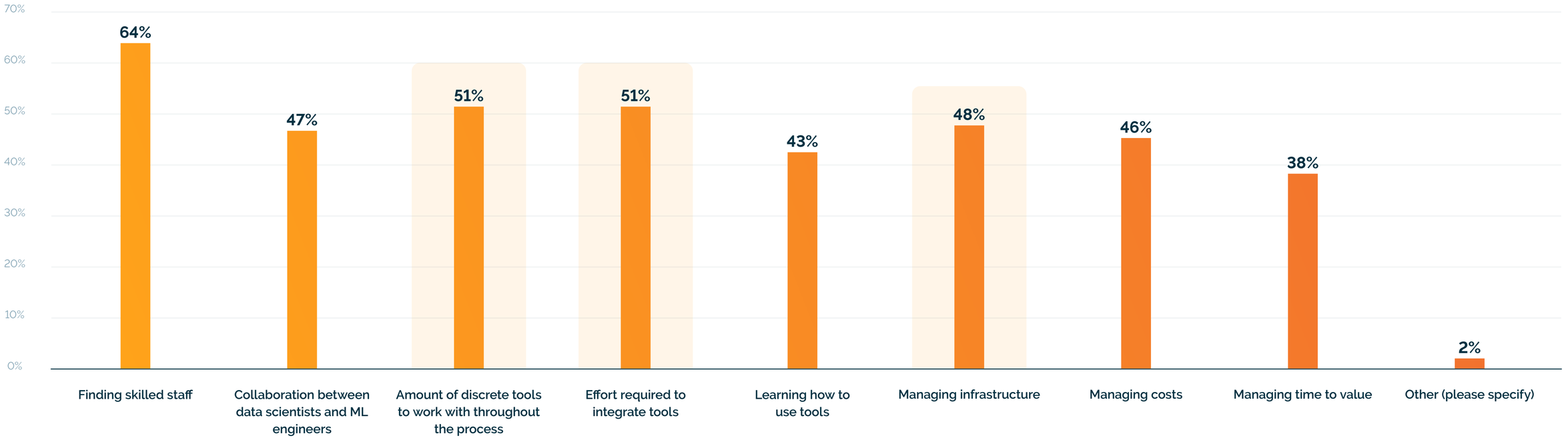
At that point, in-house development teams often need to rework the software, which impacts staff learning curves, cost, and time-to-value.

But data shows that these are all challenges that CDO survey respondents are already facing, even before they make any platform revisions.

Infrastructure Implications of DIY Platforms

More than half of CDO survey respondents said that the number of discrete tools they must work with throughout the process and the effort required to integrate those tools are challenges. Slightly less than half said that managing their infrastructure is a challenge. All of these challenges are likely due to the fact that most leading-edge firms build their own ML production frameworks, either as a collocation of point solutions that have been bolted together or built entirely from scratch within the enterprise, often by corporate IT departments with little previous experience in ML.

General challenges with production ML

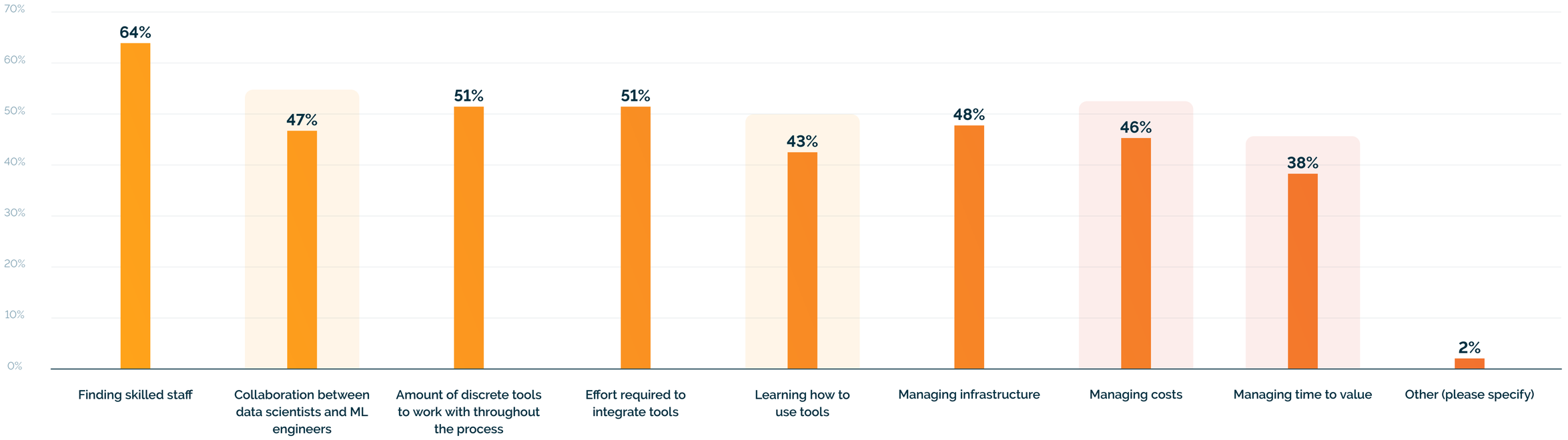


Connection Between Platform and Productivity

Almost half of CDO survey respondents listed collaboration issues and learning how to use tools as major challenges to putting ML into production, suggesting that platform choices are negatively impacting the ability of staff to produce value for the enterprise.

A close second set of challenges are the time to value and the cost of putting ML into production. These challenges can be caused by the complexity of models and other enterprise-specific issues but they can also be impacted by infrastructure decisions.

General challenges with production ML



In Their Own Words: Challenges

Beyond responses to specific questions, we asked survey respondents to share anything else they felt was relevant. Some of the responses are below:

> *As ML systems grow in complexity and size, scaling the infrastructure to handle increased data and computational demands can be costly.*

— Worldwide Publishing Firm

> *The lack of proper data management and quality can negatively impact TCO in ML.*

— Food Technologies Company

> *Training complex ML models on large datasets is computationally taxing, requiring powerful hardware or cloud resources.*

— Data Technologies Organization

> *Autonomy/lack of a central ML team that has the ability to provide guardrails and common consistent MLOps experiences to our data scientists and ML engineers.*

— Managed Information Security Organization

> *Experimenting with multiple ML models to find the best fit for the problem at hand can increase costs*

— Data Technologies Organization

> *The time which it takes us to deploy models to production and achieve results is the biggest inhibitor to success for us.*

— Top Financial Services Institution

> *Model developer-tester-monitor coherence and communication between team members needs to be improved.*

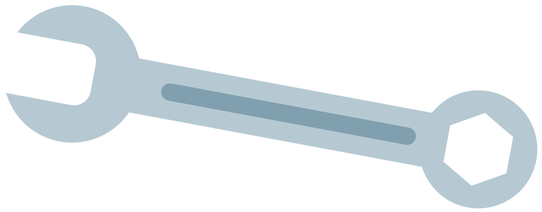
— Worldwide Chip Manufacturer

> *Broken ownership between various components of the machine learning model development and deployment process.*

— Worldwide Computer/Laptop/Phone Innovator

Takeaways

- The data reinforces the concept that early adopters don't have the luxury of purpose-built platforms and must forge their own way technologically. However, do-it-yourself ML production technologies have consequences, including ML component research time and acquisition cost, then development and integration time and costs, as well as long-term infrastructure management.
- Fortunately, as the AI market rapidly matures, there are more and more point solutions for different parts of the ML production process (deployment, scaling, monitoring, and optimization). And now, AI teams can even take advantage of comprehensive ML platforms that encompass the full spectrum of ML production (like the Wallaroo.AI Enterprise Edition) and provide ongoing, expert support.
- Commercial ML platforms must be responsive to the market, which benefits enterprises as commercial vendors develop product upgrades and feature enhancements on a regular basis. Some of the new features available from commercial producers today, for example, include ways to better take advantage of existing technologies, such as CPUs. This is a critical strategy as the GPUs typically used in ML acceleration (or cycles at a cloud provider who has GPUs) are challenging to acquire in the wake of the COVID-19 supply chain disruptions.
- Beyond the cost and time involved in building and maintaining do-it-yourself ML frameworks, home-grown technology is often less than ideal for AI teams to use, which impacts their output and how successful the enterprise can be.



Takeaways

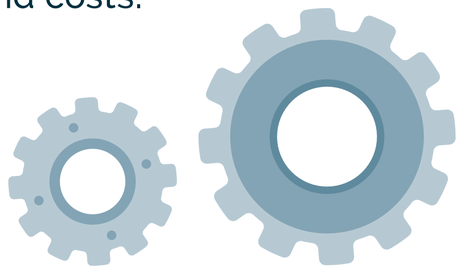
- Most in-house ML frameworks force ML engineers and data scientists to work in silos, using technologies they likely have never worked with before, are not easy to use, or have never been set up in the unique way required by this particular enterprise. The data shows that these two issues, while understandable, were significant enough to be cited as major challenges by almost half of all respondents.

- This lack of collaboration and familiarity with the ML framework impacts everything from initial ML production deployment (since the two groups often speak in different "languages" and don't see each other's work in a common way) to monitoring and optimization efforts (which unnecessarily extends the length of the feedback loop).

- While there are inevitable delays in the ML production process that impact time-to-value, lack of collaboration and excessive learning time are avoidable. Unfortunately, beyond realization of ROI, this delay can also impact competitiveness and even the ability to respond to market conditions in a timely fashion.

- Ability to manage costs was also cited as a major challenge of putting ML into production by about half of the respondents. As with time, some cost overages in ML production are unavoidable. However, costs incurred by human factors such as errors in repeatable processes, poor communication, slow understanding of tools, and so on, can be corrected.

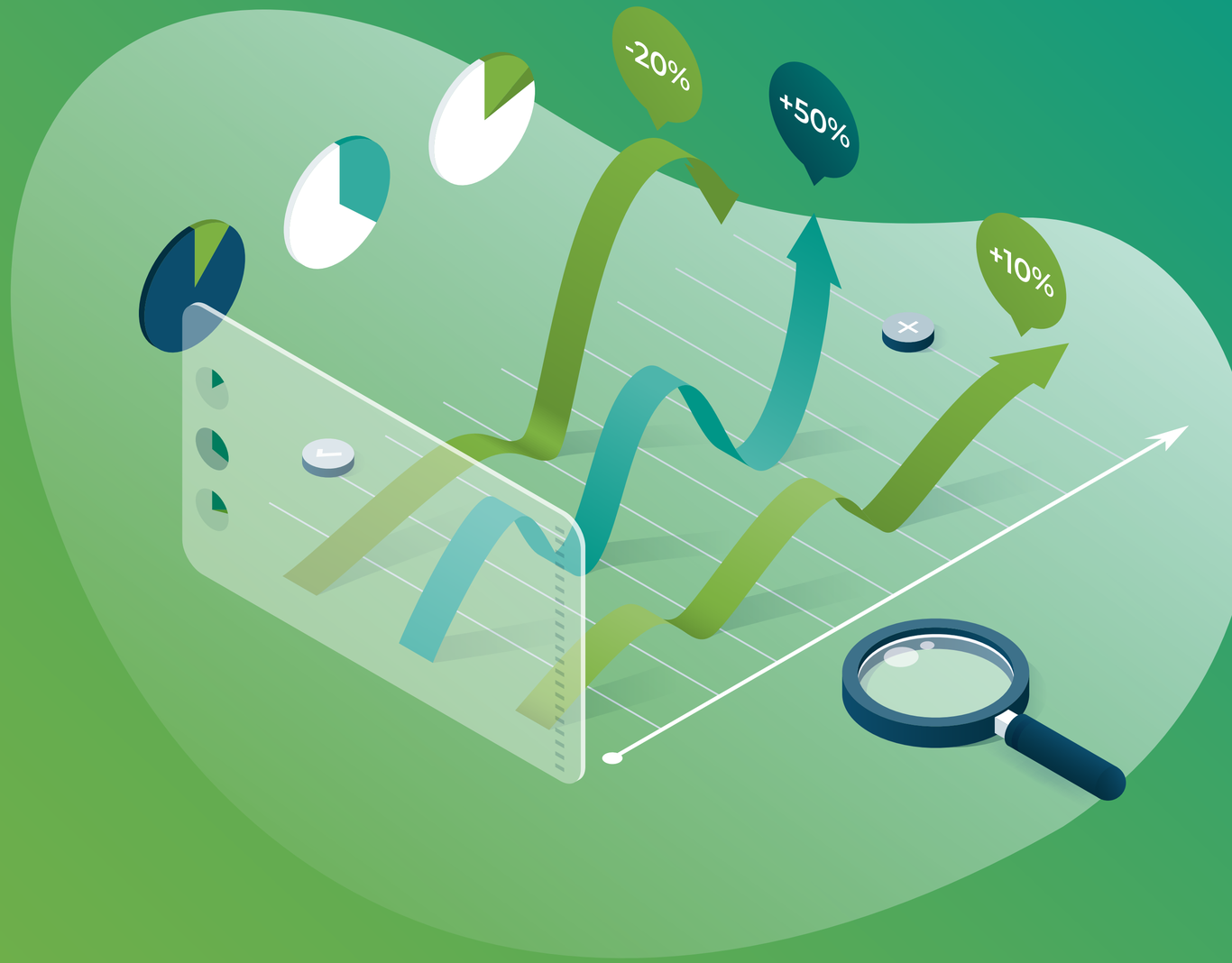
- Enterprises can address these human-related issues by creating intensive and comprehensive training by someone who understands the technology and ideal ML production processes from the perspective of the various users. Alternatively, CDOs can use purpose-built ML frameworks with built-in collaboration capacity, have user feedback-optimized interfaces, monitoring baselines, alarms, and other monitoring and optimization features, as well the ability to easily share data with popular cloud platforms for faster and more predictable timelines and costs.



CHAPTER 7

ML Production Best Practices and Challenges by Stage

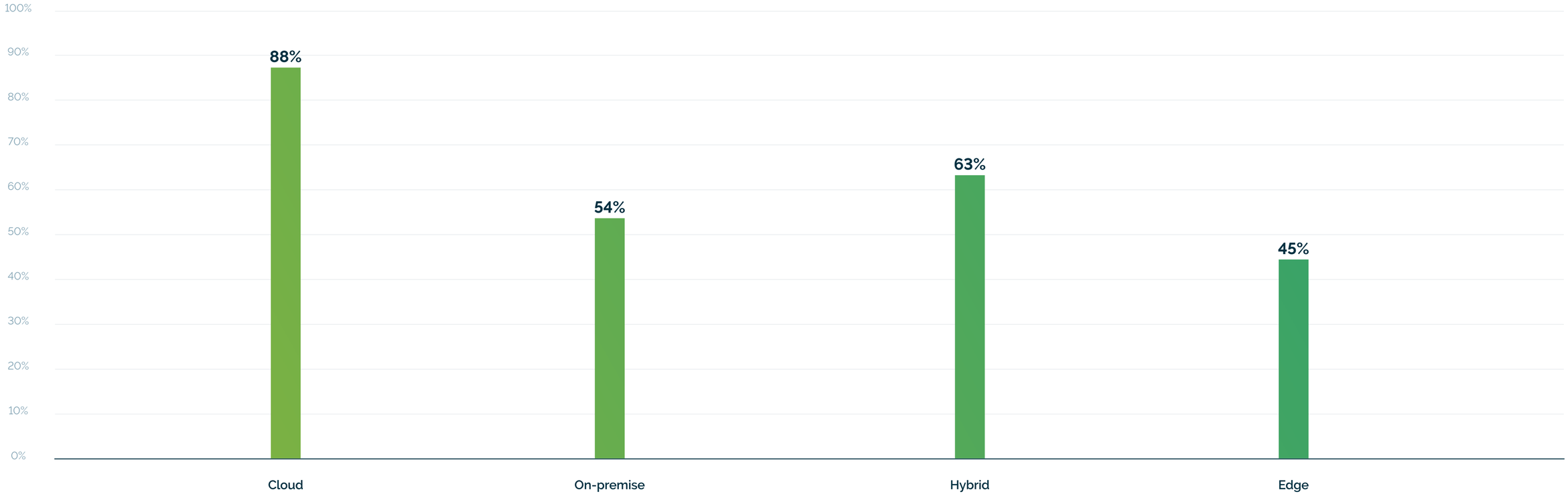
Where AI teams choose to deploy their models has a major impact on the value they are able to generate for their organizations. But, at each stage of the ML production process, leading-edge organizations face challenges because of ML production frameworks they must work through, though there are some clear ideas of how they can overcome these challenges.



Cloud is the Deployment Environment of Choice

Almost half of the survey respondents indicated they were deploying on the edge, which has historically been seen as the most challenging place to deploy AI. However, the vast majority (88%) of companies responding to the survey are deploying their ML models in the cloud while another 63% are deploying in a hybrid manner. Working with outside cloud providers or even in-house cloud can speed up the deployment and cost less than dedicated infrastructure. But, given the variety of challenges and the frequency in which they were cited by respondents, even cloud deployment has substantial issues.

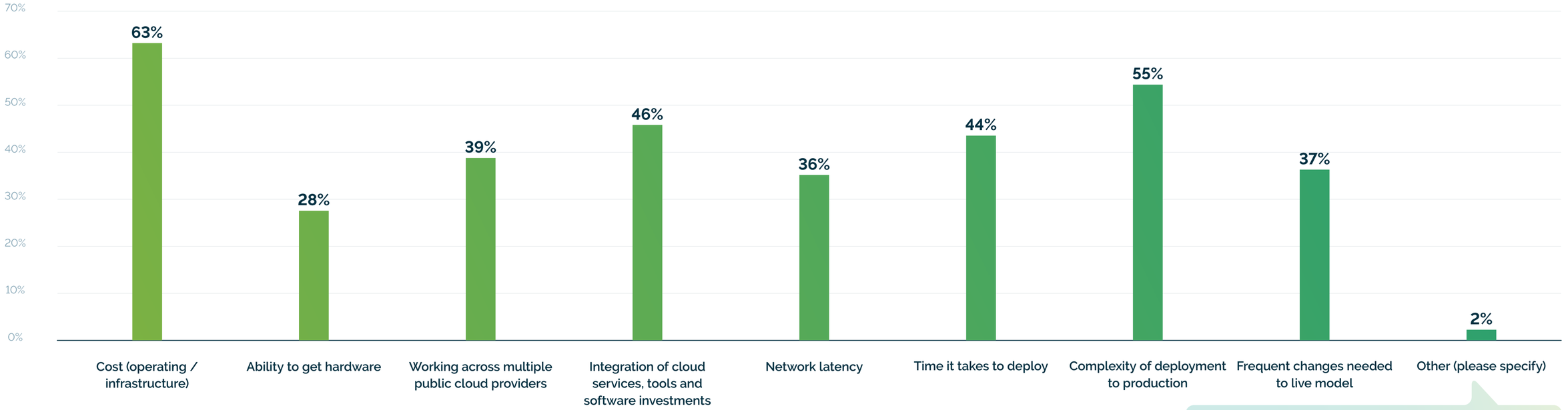
Deployment model



Cloud-Related Challenges Common During Deployment

Platform-related issues were common challenges during the deployment phase of ML production, but a surprising number were directly connected to the choice of cloud as an ML deployment method. Fifty-five percent of respondents noted that complexity was a challenge, for example, while a combined 85% of respondents cited cloud infrastructure-related issues as a challenge. Likely because of the prevalence of cloud as the deployment platform of choice (per previous chart), latency was cited by over a third of respondents. Potentially because of these issues, 63% of respondents cited cost as a challenge during the deployment phase of ML production.

Challenges deploying to production



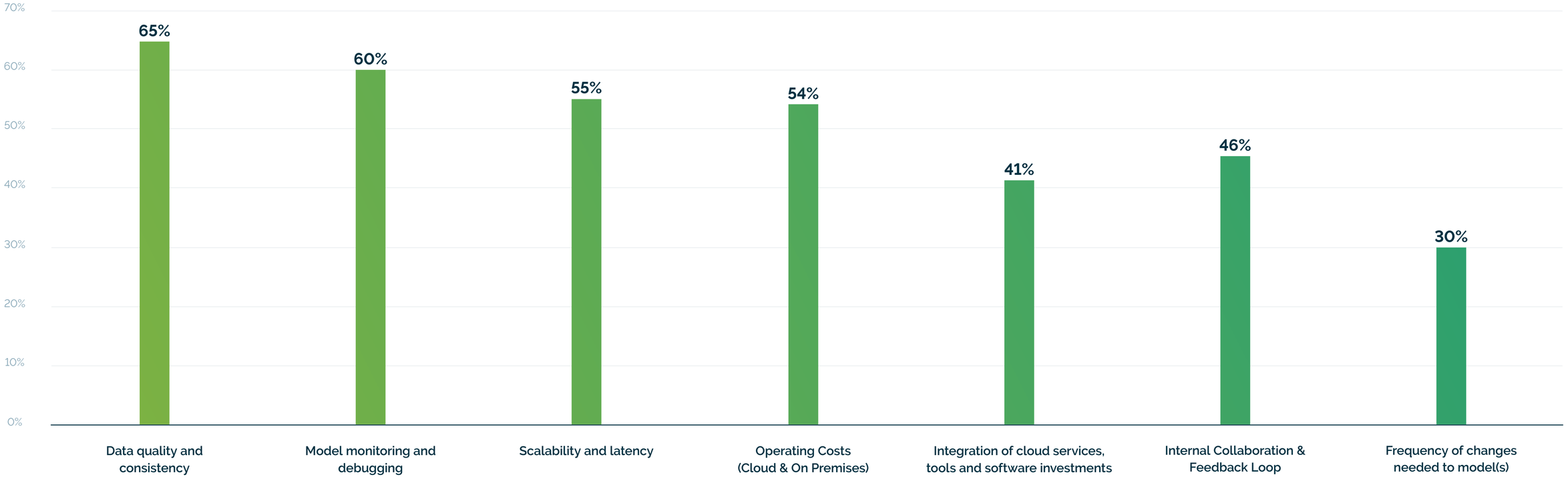
Other reasons specified

- Security reviews
- Getting IT staff to allow deployment and management to accept it

Platform Choices Impact Model Serving

Cloud-related issues (including costs, integration of cloud services and tools, latency) were also challenges for serving models. However, data quality and consistency was the single most frequently cited challenge at this stage (65%). Model monitoring and debugging was the next most frequently cited challenge (60%), which is not surprising given most ML production frameworks used by the leading edge are created in-house, potentially by people unfamiliar with the team interactions, model data issues, and other AI production processes. In fact, team collaboration and feedback loops were specifically called out as challenges by almost half of respondents, underscoring the impact of platform decisions.

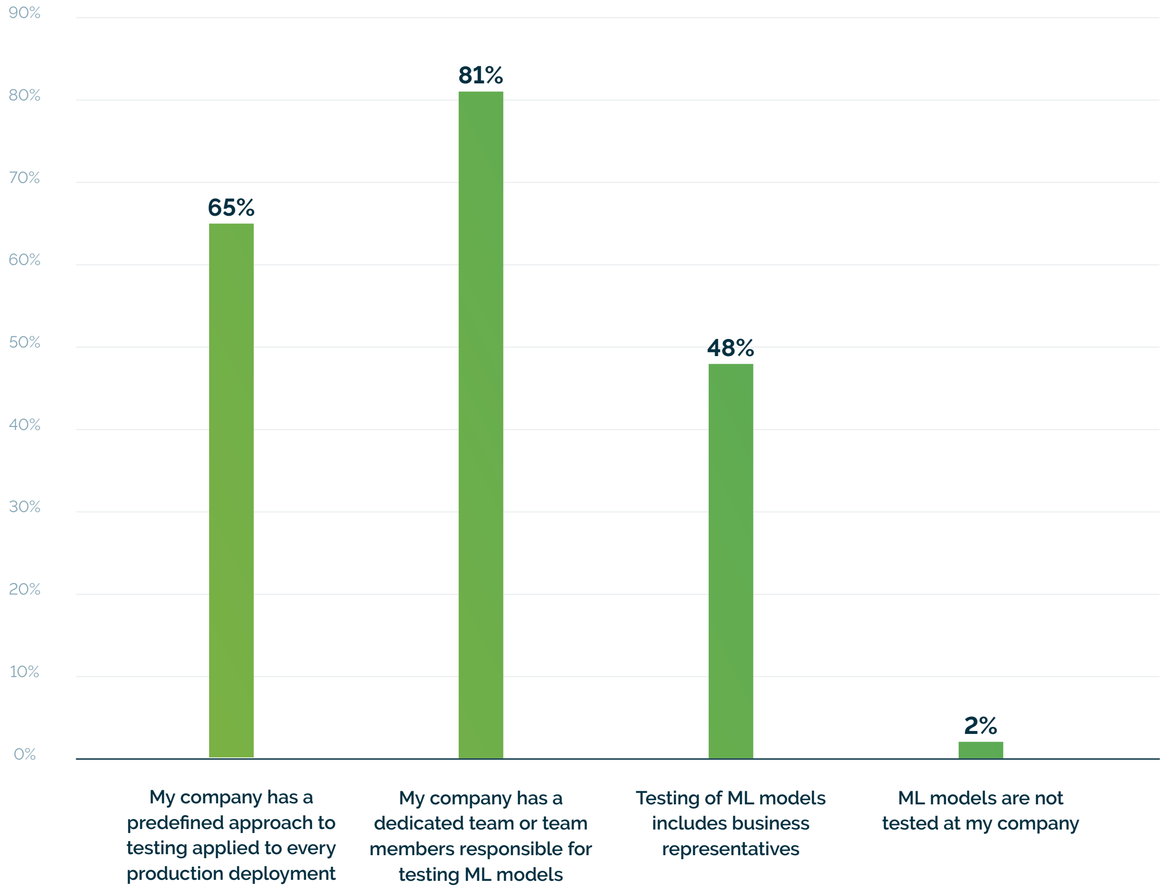
Challenges serving ML models



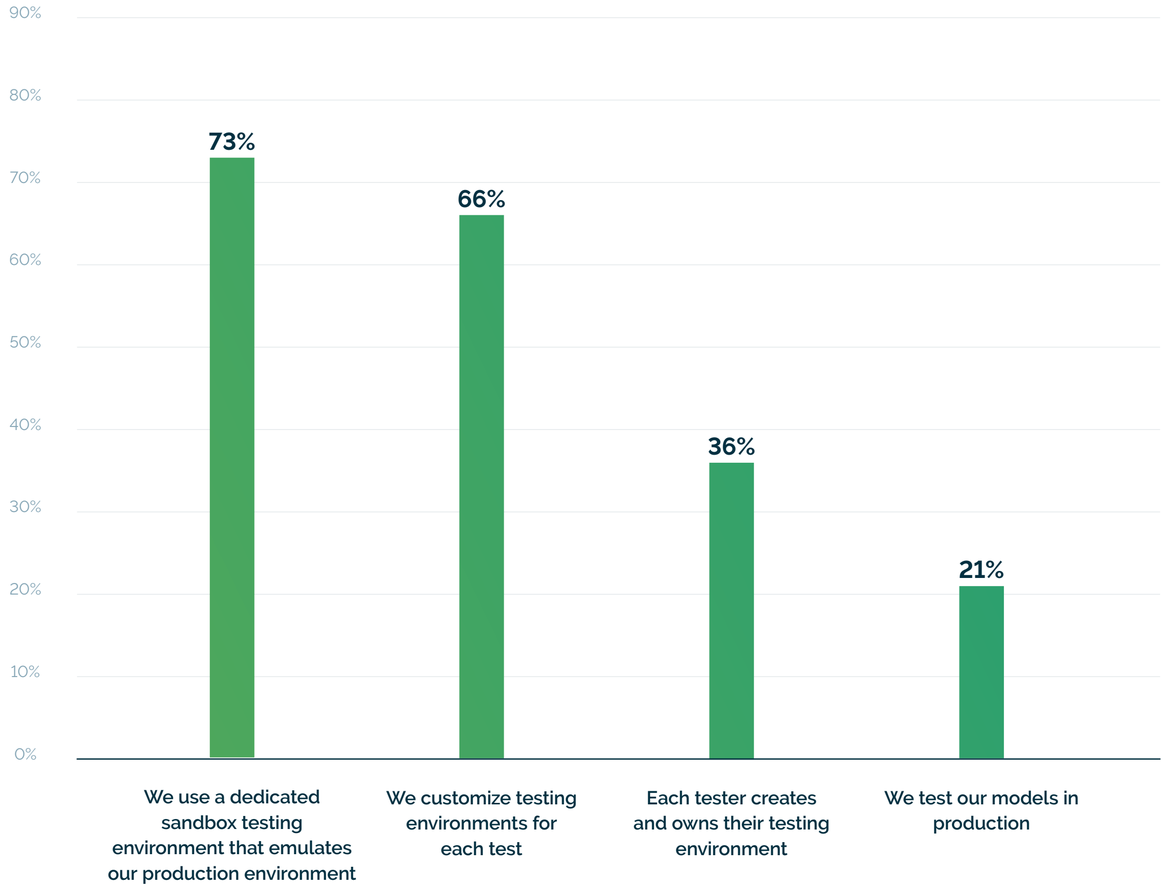
Testing ML Models

The majority of companies have a dedicated testing team, defined testing approach and use sandbox environments for testing ML models.

Testing process



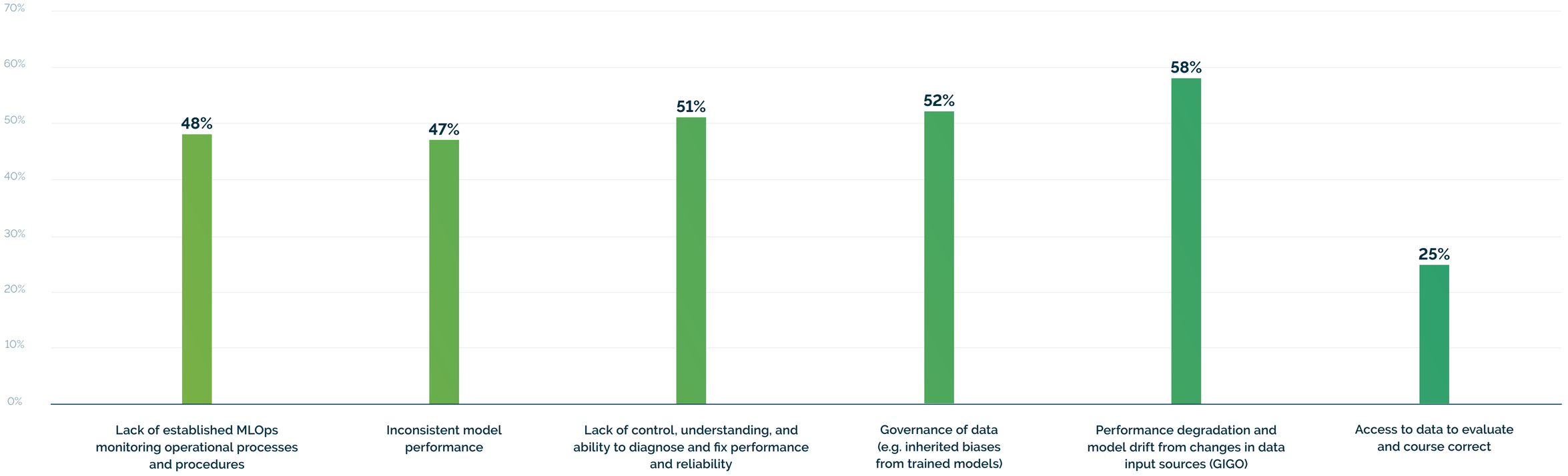
Testing approach



Monitoring Challenges Impact ROI

About half of survey respondents endured challenges related to the data and/or model performance/degradation, which might mean training and/or model development stage problems have carried over to the production process. However, half of survey respondents indicated that the lack of established MLOps monitoring processes was a challenge during model monitoring. Even more surprising, 51% of respondents indicated they had a lack of control, understanding, or ability to diagnose and fix performance and reliability. Given the critical importance of monitoring for drift, test and optimization results, impacts of scaling and other changes, suboptimal monitoring is likely reducing the ROI of models, despite the confidence expressed in earlier charts that AI initiatives were generating value for the organization.

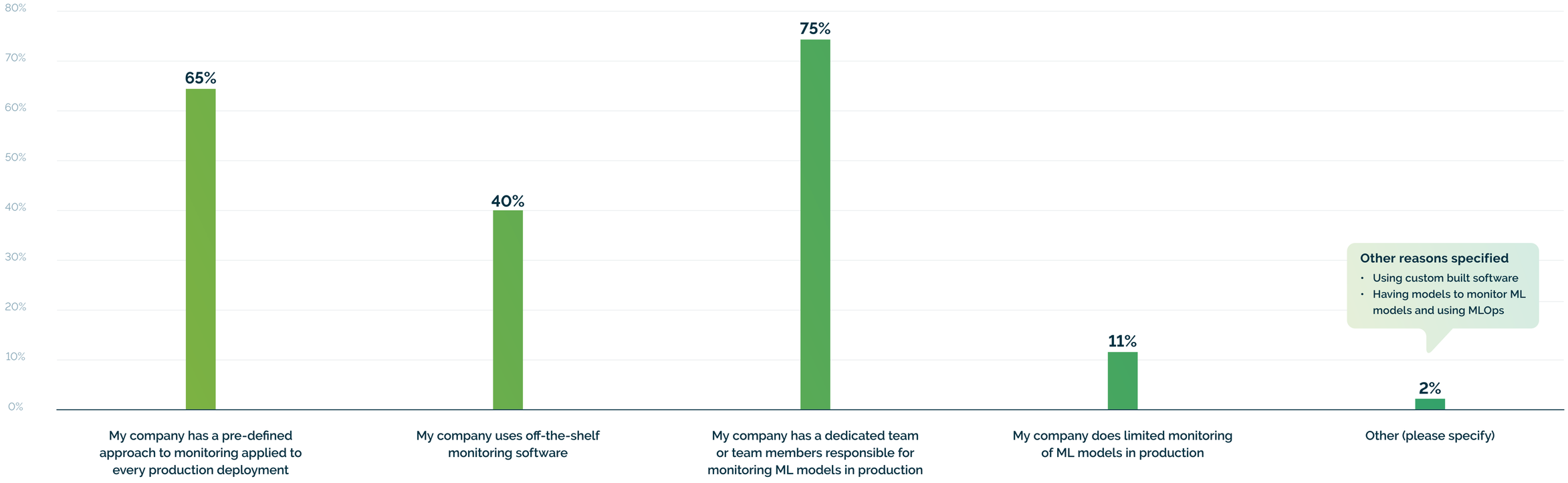
Challenges monitoring models in production



Leading-Edge ML Model Monitoring Processes Vary

While 65% of responding companies do have a pre-defined approach to monitoring models, the methods vary widely. This is logical given that there is likely a wide variety in the ML production platforms created by leading-edge firms. The efficacy of different approaches also varies, given that 46% of respondents (see previous chart) cited internal collaboration and feedback loop (both of which are impacted by monitoring practices and technology) as a challenge in serving models. But, surprisingly, 11% of leading-edge organizations have only limited monitoring of ML, which will certainly have an impact on the ROI of AI initiatives as you cannot optimize or respond to market conditions unless you monitor models.

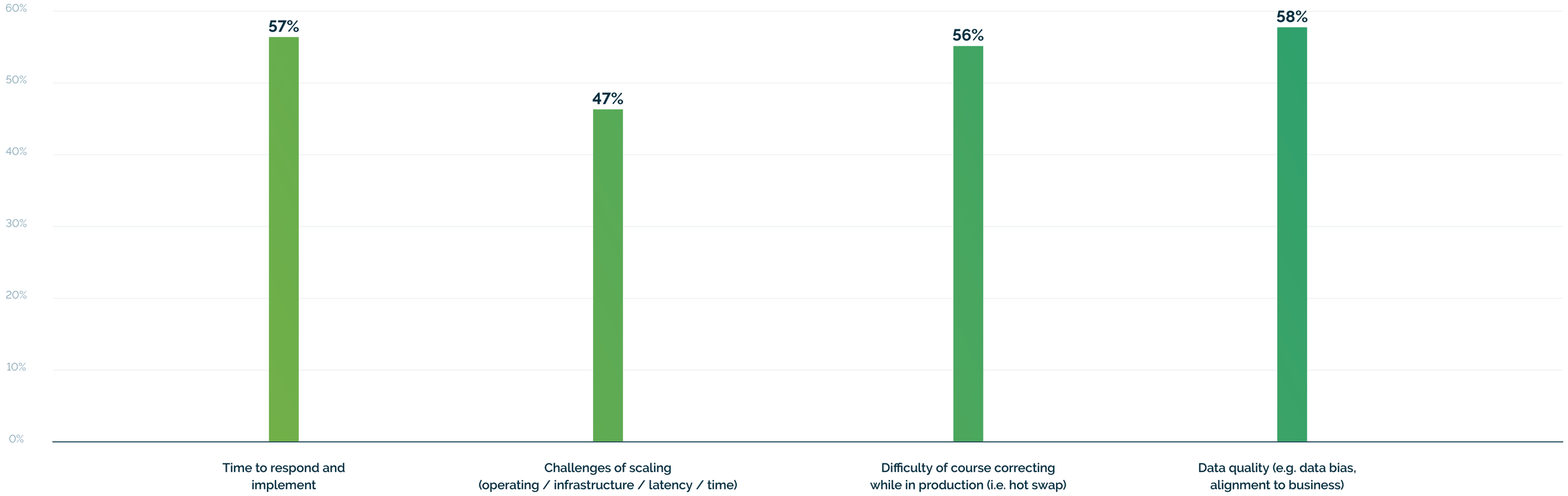
How companies monitor ML models in production



Challenges Optimizing ML Models in Production

Leading-edge organizations were remarkably consistent in the challenges they say they face, with almost the same percentage saying the time to respond, difficulty in ability to respond, and quality of data (e.g., data bias, alignment ot business, etc.) were their greatest challenges. Human-related issues may contribute to some of these problems. However, optimizing models at scale was also a significant issue, meaning that, once again, the platform is impacting the ability of the leading edge to generate value for the enterprise.

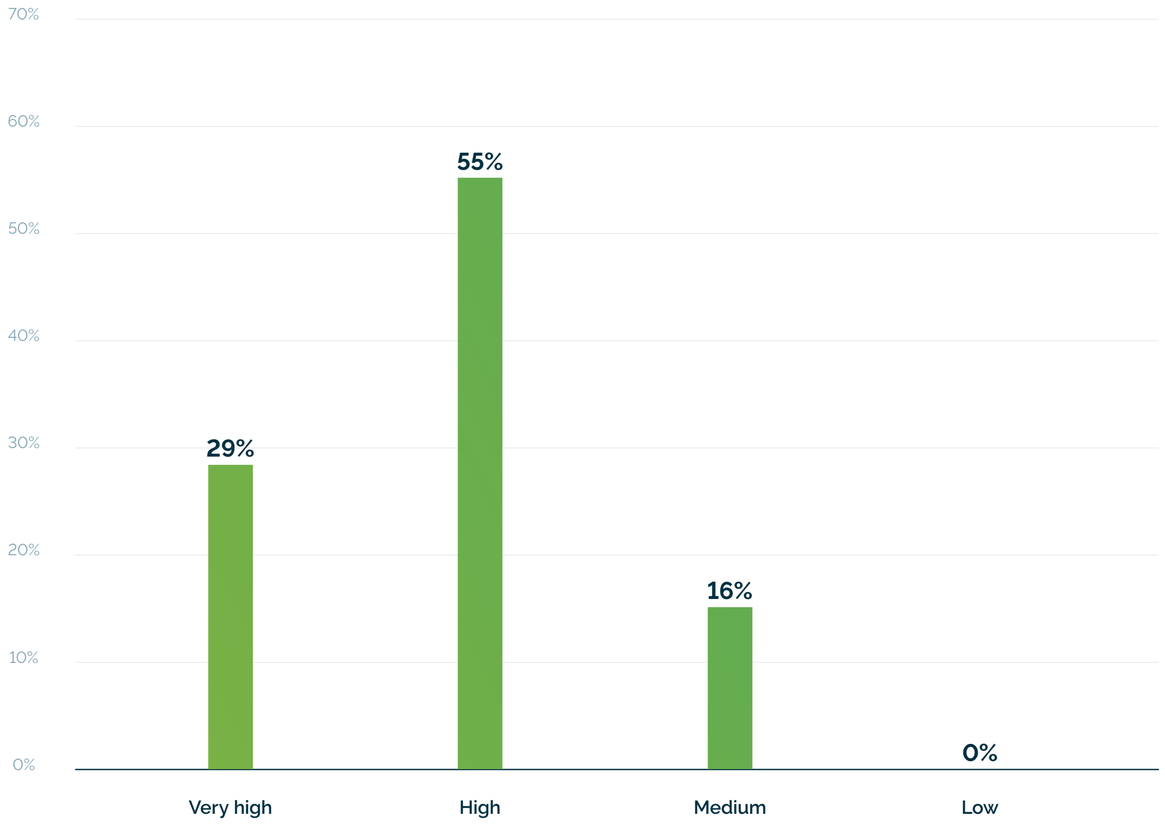
Challenges optimizing models in production



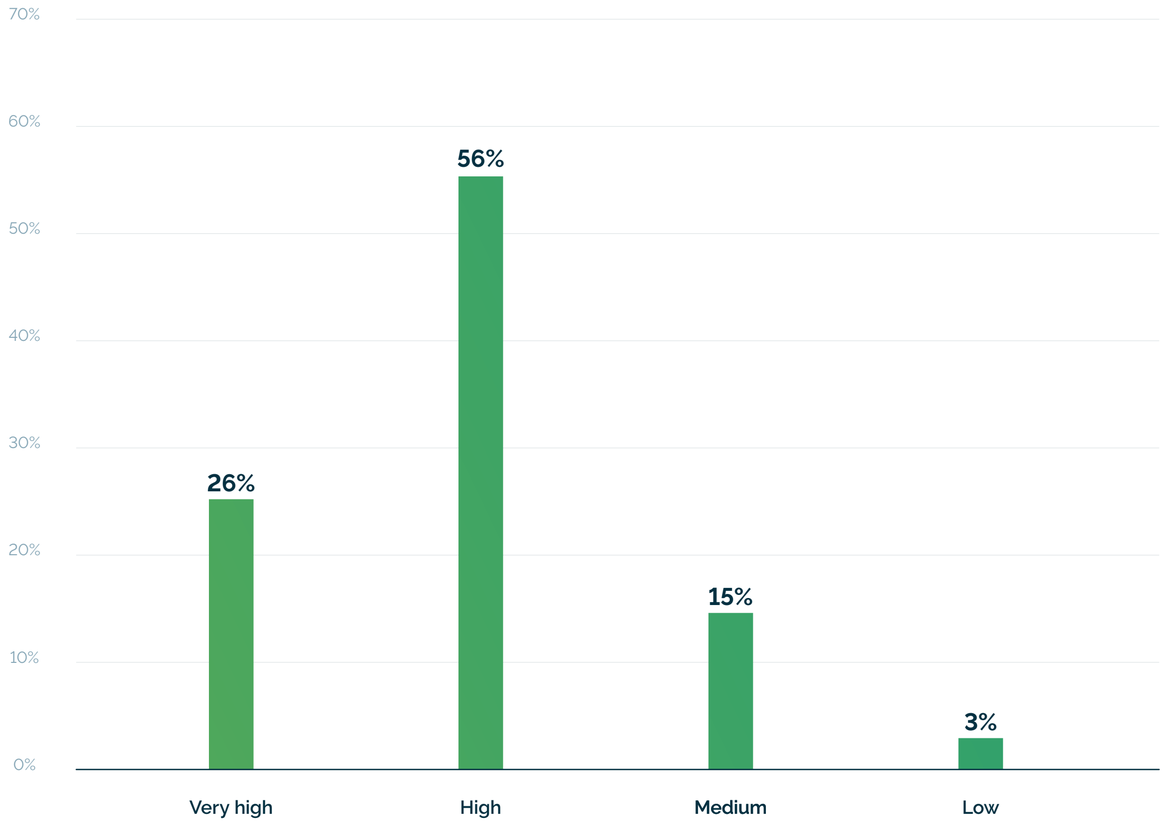
Areas of Improvement: Automation is the Key

Over 80% of companies expect automation to present a high or very high improvement to both the testing and monitoring processes. Benefits of automation include freeing up practitioner time from routine tasks so they can focus on tasks that bring value to the enterprise. Automation can also reduce human error and speed up processes.

Expected improvement from testing automation



Expected improvement from monitoring automation



In Their Own Words:

Beyond responses to specific questions, we asked survey respondents to share anything else they felt was relevant. Some of the responses are below:

> *[What sets us apart?] Implementing a highly automated deployment process for ML models, incorporating automated testing, continuous integration, and deployment pipelines to minimize manual effort and errors.*

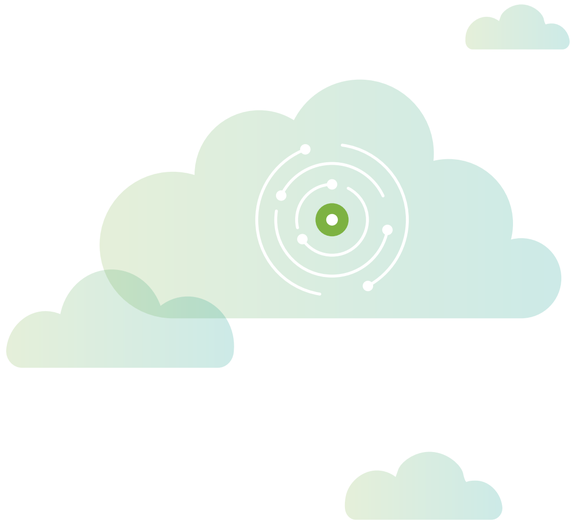
— Home Furnishings Manufacturer

> *We currently have pre-defined protocols that can monitor for common issue. But there is always a human intervention required for most of the cases till now. Automation is required for observing the model in production.*

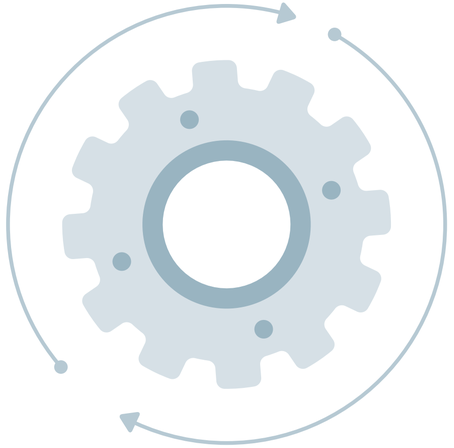
— Global Chip Manufacturer

Takeaways

- Cloud is the deployment location of choice for leading-edge enterprises. However, using the platform for ML comes with its own challenges, from latency to costs to integration of cloud infrastructure.



- Many of these issues are related, as a more performant infrastructure that is pre-integrated and cloud-optimized can remove integration and latency, which reduces cloud cycles, which reduces costs. Some modern ML production platforms (including Wallaroo.AI's Enterprise Edition) also offer a way to better take advantage of CPUs, which is another way to reduce costs related to models (such as large language models, which are very popular today) that require scarce and expensive GPU-powered hardware.
- One clear way CDO survey respondents identified as a way to overcome some challenges relates to the use of automation, which did not appear to have been included in most in-house developed ML platforms, given how many organizations cited it as a way to improve.



- However, due to the rapidly maturing ML operations market there are commercial vendors that have ML production platforms that have workload orchestration and other automation elements already included in their software. Some, such as Wallaroo.AI, have not only automation that supports all stages of the production process but also integration with Microsoft AzureML and Azure Databricks Jupyter notebooks, to reduce complexity associated with one of the most commonly used enterprise cloud platforms.

In Their Own Words:

We asked respondents what set their processes apart. The responses, in their own words, are below:

General Best Practices That Set Leading-Edge Firms Apart

DEPLOYMENT PROCESS

Designing the deployment process to handle large-scale ML models, accommodating high data volumes, and ensuring efficient resource utilization to meet growing demands.

— Financial Services Firm A

AUTOMATION

Implementing a highly automated deployment process for ML models, incorporating automated testing, continuous integration, and deployment pipelines to minimize manual effort and errors.

— Home Furnishings Manufacturer

AVAILABILITY / SETUP OF INFRASTRUCTURE

Adopting infrastructure as code practices to define and provision the required infrastructure for ML model deployment consistently, enabling scalability and reproducibility.

— Healthcare Software Firm A

MONITORING

Implementing robust monitoring systems that provide real-time insights into model performance, data drift, and anomalies, enabling proactive interventions and rapid issue resolution.

— Data Quality and Integration Software Firm

OPTIMIZATION OF MODELS

Creating feedback loops and mechanisms to collect insights from production ML models, enabling continuous learning, model iteration, and performance enhancement.

— Publishing Firm

In Their Own Words:

We asked respondents for tips on deploying ML models. The responses, in their own words, are below:

Best Practices for Deployment

> *"Our company provides a strong MLOps and Feedback loop process which allows our model to be robust and tuned real-time so customers have the best experience."*

— Computer Software Firm A

> *"Integrating ML model deployment with DevOps practices, fostering collaboration between data science and operations teams, and streamlining the end-to-end development and deployment lifecycle."*

— Data Quality and Integration Software Firm

> *"Establishing clear governance policies and processes to ensure compliance, ethical use, and responsible handling of ML models in production, addressing privacy, fairness, and security concerns."*

— Education Firm

> *"Establishing mechanisms to manage and track different versions of ML models in production, allowing for easy rollback in case of issues or performance degradation."*

— Computer Software Firm B

> *"Implementing an experimentation framework that allows for efficient testing, comparison, and iteration of ML models, facilitating quick deployment of the best-performing models."*

— Financial Services Firm B

In Their Own Words:

We asked respondents for tips on scaling ML models. The responses, in their own words, are below:

Best Practices for Scaling

CLOUD INFRASTRUCTURE

Deploying the Ml model in cloud and on a platform that can easily scale its resources depending upon the workload is the most cost effective way to go (e.g. using EMR, ECS clusters etc underneath).

— Financial Services Firm A

AUTOMATION

Utilize pre-trained models as a starting point and fine-tune them on specific tasks or domains, reducing the overall training time and resource requirements.

— Financial Services Firm C

CONTAINERIZATION

Use of kubernetes clusters is helpful while scaling up to more users.

— Chip Manufacturer

DEPLOYMENT PROCESS

Make User Interface extendable such that Business Analysts or Data Scientists can deploy algorithm without the help of developers.

— Healthcare Technology Firm B

DISTRIBUTED COMPUTING

Split the model across multiple devices or machines, with each handling a portion of the model's computation.

— Financial Services Firm B

Conclusion

After understanding the stumbling blocks and hearing the advice from CDOs at leading-edge enterprises, certain key points stand out. Along with these critical topics, there are also some clear learnings that the next wave of ML production adopters can take advantage of.

Summary

The AI team leaders at the leading-edge of ML production have proven surprisingly resourceful in their ability to pioneer a new technology while simultaneously being able to bring value back to the enterprise.

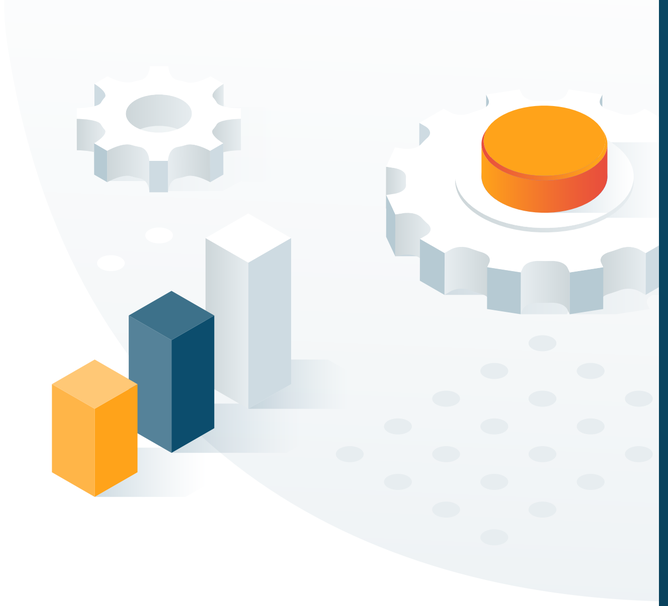
However, the research shows that leading-edge enterprises face many challenges that impact cost and productivity because of the dependency on hard to find and retain ML experts, the choice of cloud, and use of in-house developed ML production frameworks.

All AI teams will be looking for and trying to retain ML experts but, fortunately, all leading-edge enterprises have strong ideas on how to do so that AI team leaders can use to acquire and retain ML experts that other organizations can use. There is also the option of working with a commercial vendor that has ML experts who can assist.

But for organizations seeking to emulate these CDOs, the greatest challenge is clearly the right operations platform to execute ML production as effectively and efficiently as possible. This single decision impacts every aspect of the process, from deployment to scaling to monitoring and optimization.

A purpose-built ML production platform or taking the time to develop an in-house framework that includes the capabilities of commercial ML production software and operates in a seamless fashion (versus some of the tools infrastructure issues cited in the report) would solve many of the platform or cloud-related challenges leading-edge enterprises face.

Costs and time-to-value are closely connected to the choices made during the development of the platform. Even the challenge of staffing can be reduced – or at least freeing up existing resources to work on more valuable areas – with changes to the ML production operations platform used.



BACK TO TABLE OF CONTENTS

What's Next

Even if next wave enterprises decide to build their own operations platform for ML production, it's critical they understand the state of the art of commercially-available ML production platforms to understand what is possible and how it benefits practitioners and enterprises.

We've noted that several of the challenges leading-edge CDOs face can be solved with Wallaroo.AI's Enterprise Edition, which covers the deployment, scaling, monitoring, and optimization stages of the ML production process. Wallaroo.AI also has a team of ML experts available to help AI teams across all those steps in the process.

You can get a no-cost consultation with Wallaroo.AI's ML experts and see how you can apply the best practices and mitigate the challenges described by leading-edge AI team leaders.

Click here to specify your use case, model type, industry, and other factors so that we can identify the correct associate to assist you.

About Wallaroo.AI

Scaling AI to ROI

> **90%** of AI initiatives fail to produce ROI

Wallaroo.AI is changing that with our self-service ML production platform and expert ML team.

Client Results:

Realized value
4-6X faster

Analyzed data up
to **12.5X** faster

Scaled the
number of
AI initiatives
5-10X

Freed up practitioner
time by **40%**

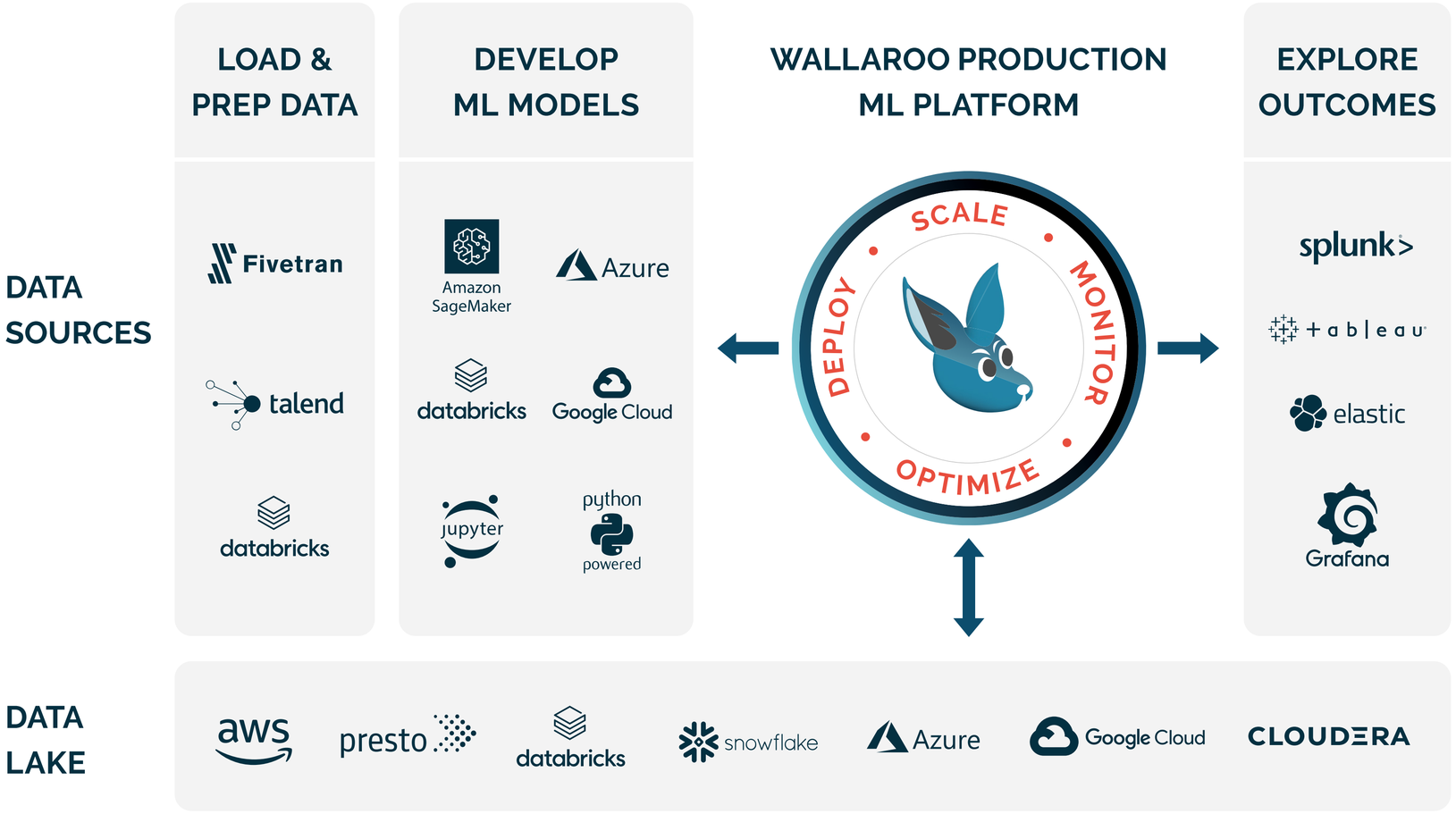
Reduced ML production
cost by **50-80%**

BACK TO TABLE OF CONTENTS

Leading-Edge Software

Wallaroo.AI Enterprise Edition ML Platform

Works with Your Tools. Your Data. Your Ecosystem



ML Experts

- Wallaroo.AI was founded in 2017 by an AI team leader at a financial services firm who was fed up with having to build the optimal ML production framework while trying to generate value for his enterprise. **Because of this, our mission is to empower enterprise AI teams to operationalize ML to drive positive outcomes.**
- **We're a completely remote firm** so we can hire the best, regardless of location. And we're all focused on being customer-first, from engineers who turn client feedback into leading-edge product features to customer success team members **who help clients with everything from initial implementation to expansion to optimization.**

BACK TO TABLE OF CONTENTS

Why Wallaroo.AI

DEPLOY IN SECONDS

- Self-service toolkit to deploy and scale ML
- Easy-to-use SDK, UI, and API for fast, repeatable, low-code/no-code ML operations

SCALABILITY + PERFORMANCE

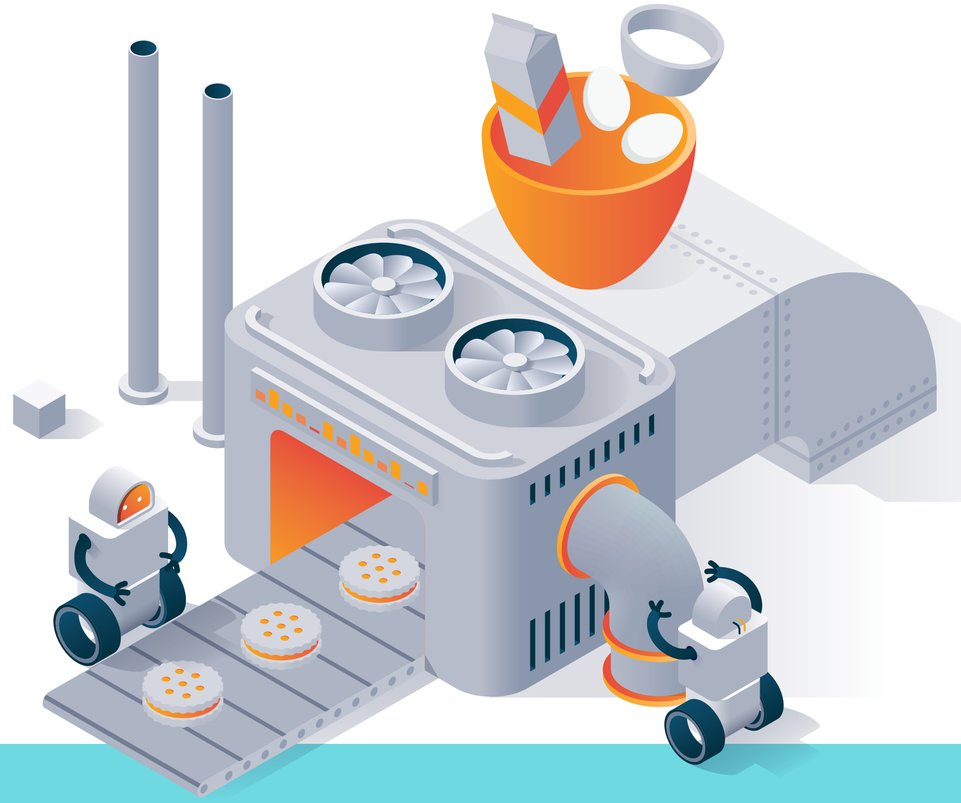
- Blazingly fast inference server
- Distributed computing core written in Rust-Lang uses up to 80% less infrastructure

CONTINUOUS OPTIMIZATION

- Advanced observability
- Comprehensive audit logs, advanced model insights, full A/B testing

EXPERT ML TEAM

- Dedicated Customer Success team
- Experts to support client as needed, from deployment to scaling to optimization



Featured Partners AMPERE vmware Azure

Investors boldstart CONTOUR VENTURE PARTNERS ENIAC VENTURES RESOLUTEVENTURES GREYCROFT FiftySix RIDGELINE RRE VENTURES M12

Wallaroo.AI info@Wallaroo.AI www.wallaroo.ai