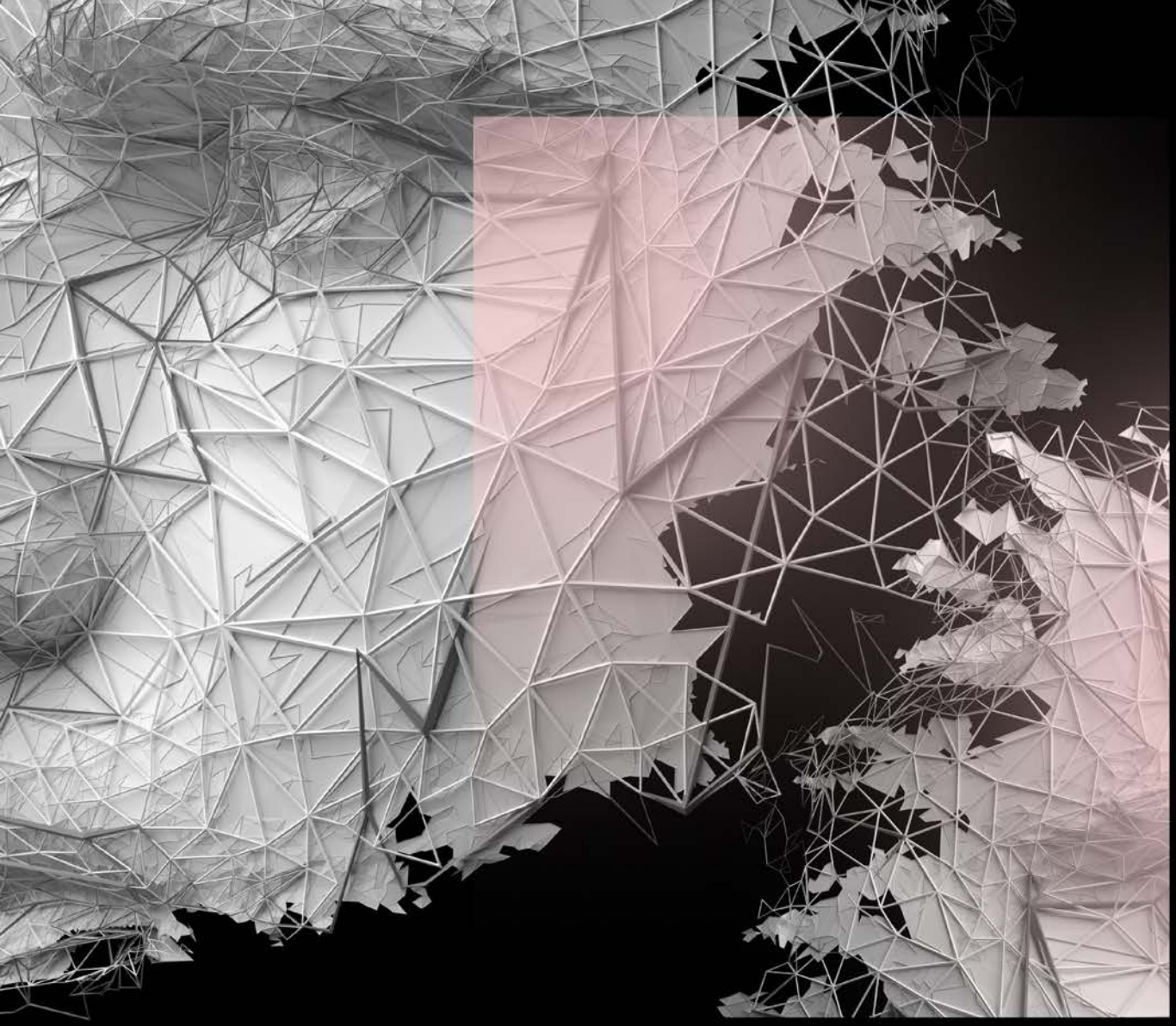


Sponsored by CLEARML

AI Infrastructure Alliance

Enterprise Generative AI Adoption

C-Level Key Considerations, Challenges, and Strategies for Unleashing AI at Scale

2023

When we started the AI Infrastructure Alliance in 2021, AI and machine learning was a niche topic, with a passionate and dedicated set of practitioners but it was mostly hidden from the general public. People used AI without thinking about it, like when they talked to their phone and it understood what they said or when they used Google Translate on their summer vacation.

All that changed over the last year with the release of **ChatGPT**, a revolutionary Large Language Model (LLM).

To say that ChatGPT was a titanic shift in public perception of AI is an understatement. It rocketed to over 100M users in just three months between February 2023 to April 2023. It hit its first million users in only 5 days. To put that in context, it had taken Netflix 3.5 years to hit 1M users.

Regular people everywhere were suddenly using AI directly, instead of hidden inside an app. The experience is raw, visceral and direct. Now everyone is aware of AI and has an opinion on it, whether it's governments rushing to pass landmark legislation, to artists either for or against AI, to big business, to the local cab driver at the airport.

Companies are racing to integrate generative AI and LLMs into their applications. But how are enterprises faring with weaving these models into their workflow and applications? Are they too unpredictable? Are companies having tremendous early success or really struggling? Or is the answer somewhere in the middle as enterprises learn how to work with these very new kinds of systems?

In this special report, the AIIA polled over 1000 businesses with over 1 billion USD in yearly revenue to see how they're using AI and whether they're successfully integrating LLMs and generative AI into their business and products.

Research and development is moving at a breakneck pace now, with new ideas, new research, new applications and models landing almost daily. Everywhere developers have raced to embrace the shift, building new kinds of apps on top of LLMs like GPT-4, the veteran LLaMA foundation model from Meta, Falcon, WizardLM, Starcoder and we've seen a flurry of state-of-the-art foundation models like SAM (Segment Anything Model), Stable Diffusion, Gen1 and Gen 2. The pace continues to speed up. But are enterprises able to keep up with the rate of change? Is anyone?

A few years ago, the general feeling in the MLOps industry was that everyone would have a huge team of data scientists and train advanced machine learning models from scratch. It looks like that future will never come to pass. That's because foundation models are hard to create, often costing millions, to tens of millions of dollars, to 100s of millions of dollars to train.
These models are also big. The largest transformer based foundation LLMs have memory requirements that scale quadratically with parameter count. Serving inference with LLMs often requires many datacenter level GPUs, hundreds of GBs of RAM, and backend tricks like weight streaming to make them work.

It can take hundreds or thousands of the most advanced GPUs running for months to train these models along with advanced supercomputing teams to manage them. Some analysts have noted that many of the top models lose money on inference as the world waits for economies of scale to kick in and drive down prices.

Because of all this, most companies are now turning to advanced foundational models created by a small subset of companies and researchers. They're looking to fine tune these base models to make them work for their personal use cases and needs and to get them integrated into their applications.

There's little doubt that since the release of ChatGPT, we've seen a Cambrian explosion of AI use cases and a massive uptick in public interest in AI.

But has that translated to Enterprise adoption and integration even as individual developers and small businesses adopt AI at a furious pace?

We set out to find out.

Demography

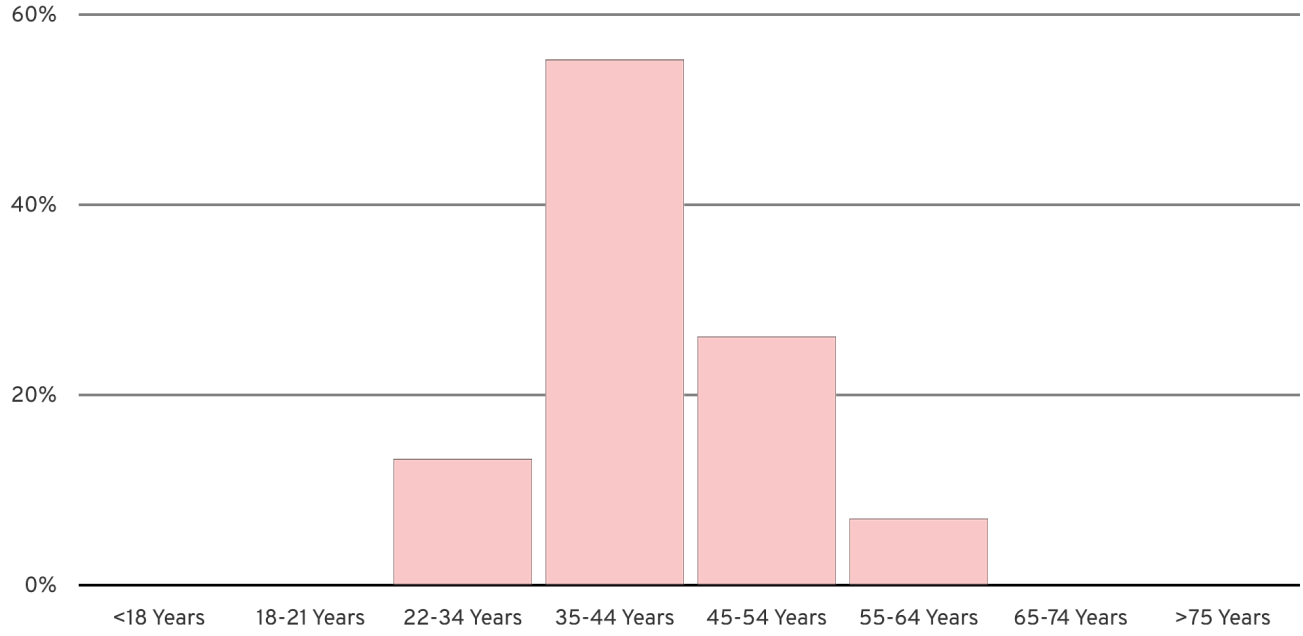
To start with, let's understand the demographics of who we talked with in our survey. Essentially, we talked to very big enterprises. There are no small or even medium businesses in our results. Every respondent had more than 10,000 employees and 50% had over 20,000.

We also primarily talked with leadership and the heads of teams, with job titles like CIO, CTO, Head of AI, VP of Data or VP of Artificial Intelligence. That means the results primarily represent the c suite and the team leads but not the engineers and their teams.

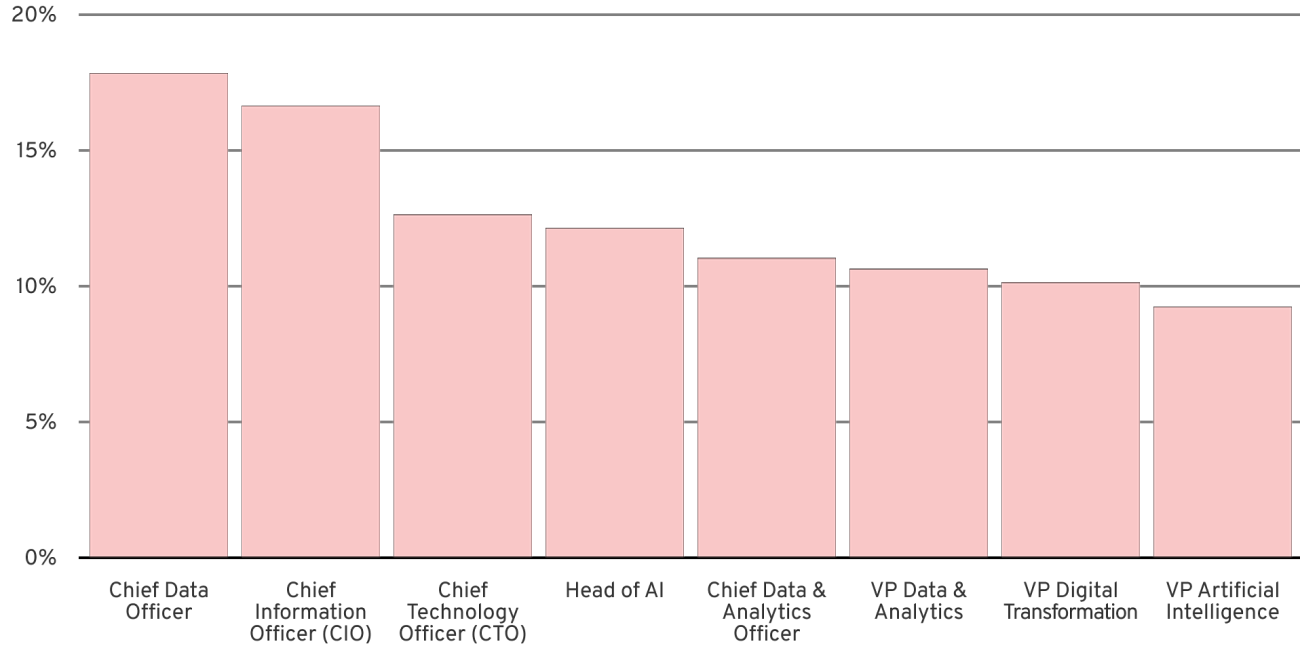
The survey primarily focused on companies in the US, UK and Canada, as well as across the EU. However, we did have a number of respondents from Japan and South Korea but we don't consider this survey a definitive representative of the Asian enterprise company perspective.

Lastly, we spoke to people across a large range of verticals, everything from law firms, to manufacturing, to telecommunications, energy, food, healthcare and more. The largest representations came from Information Technology companies, but no vertical surveyed represented more than 8% of the total respondents, so we had a wide range of views across a wide range of companies.

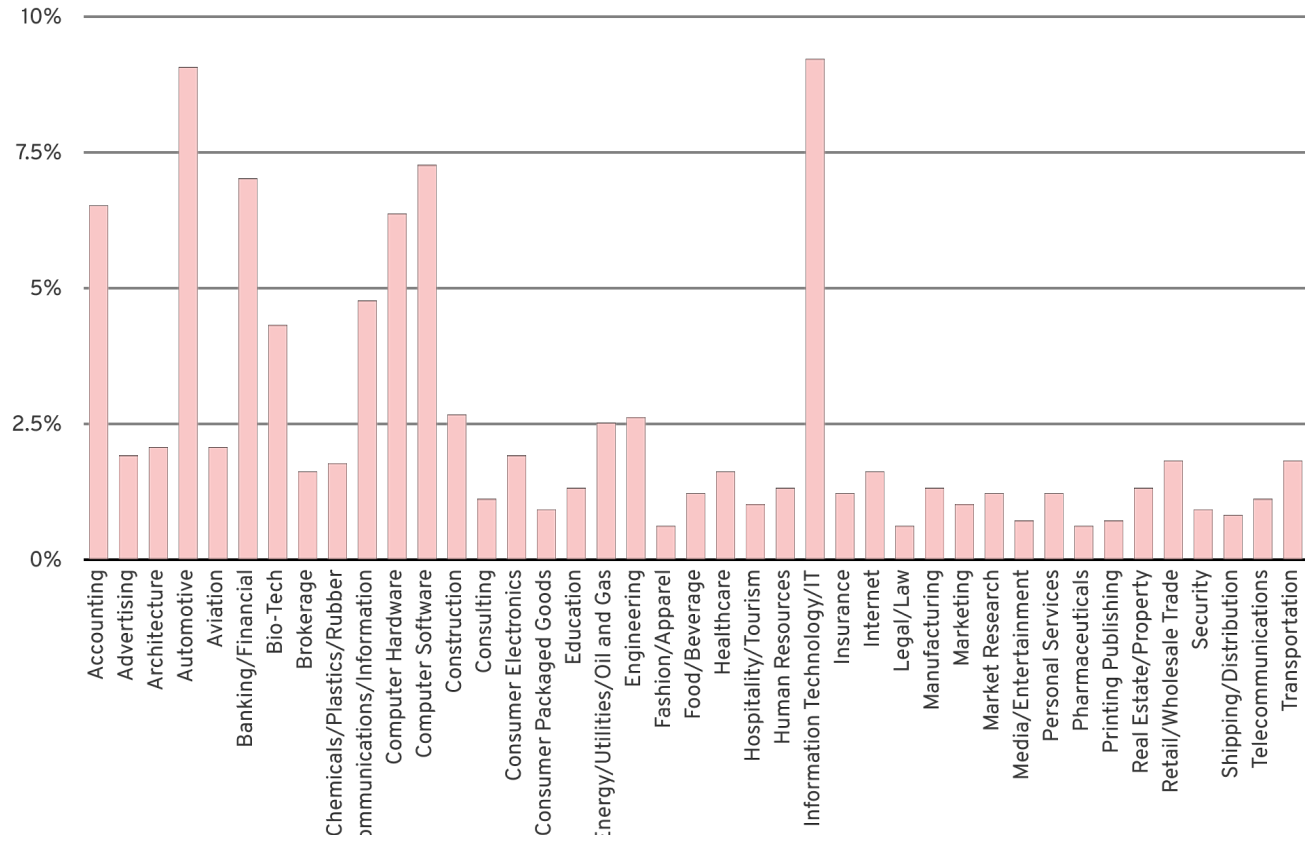
Age of Surveyed Respondents



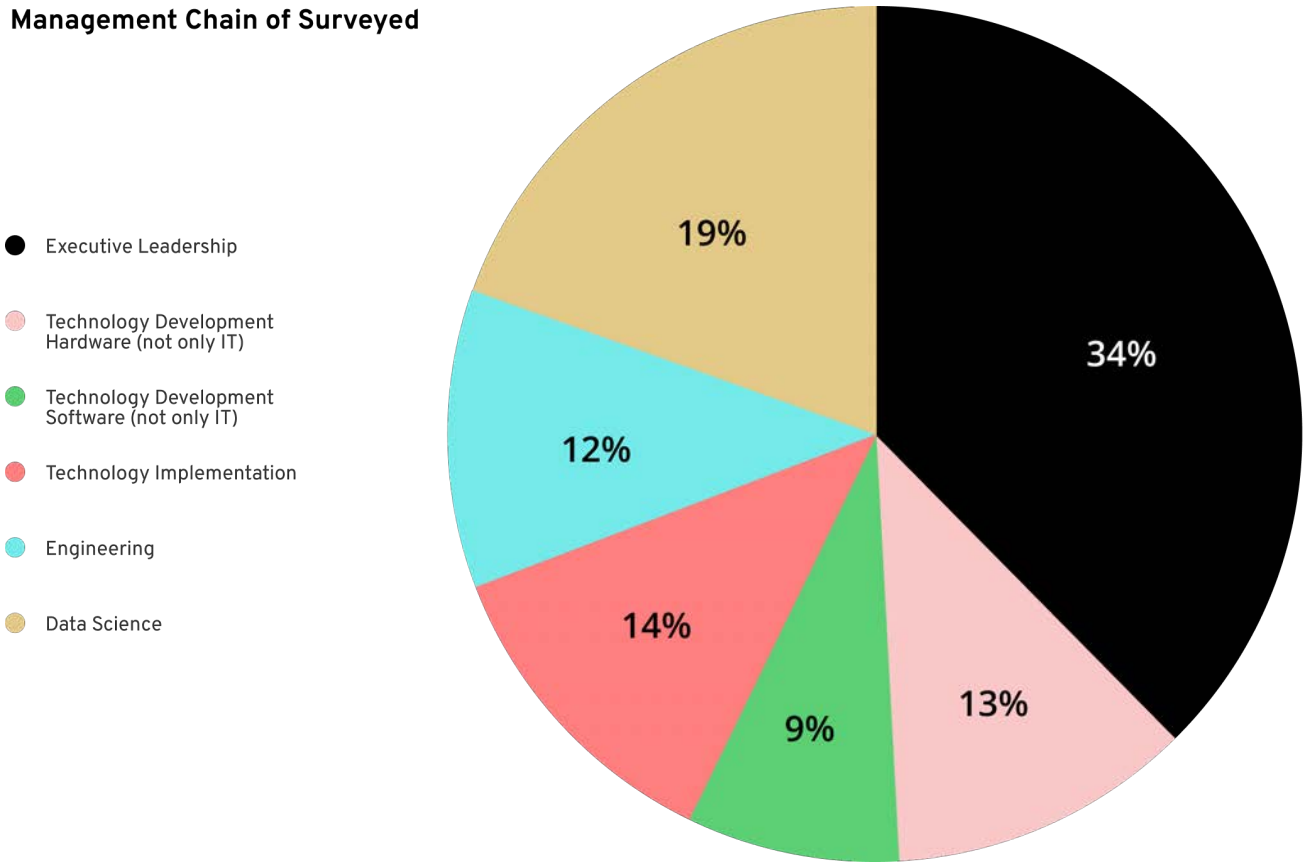
Title/Position of Surveyed Respondents



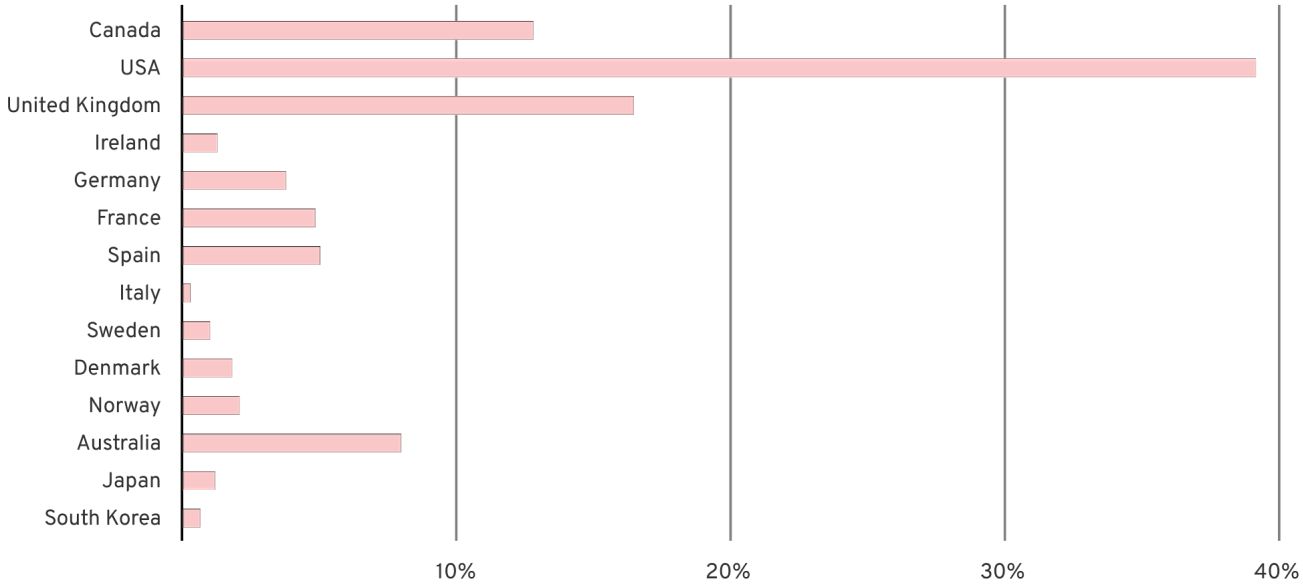
Industry Represented by Survey



Management Chain of Surveyed



Headquarters of Responding Businesses



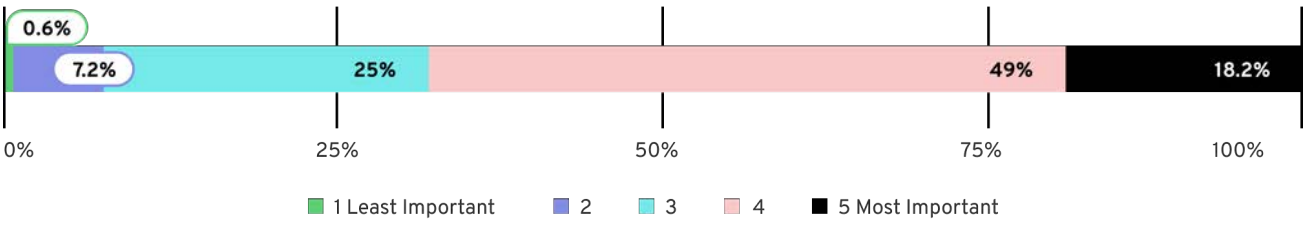
Questions

Now let's dive into the heart of the survey with the questions that matter most. Rather than include every question in order, we've picked out questions that offer the most insights into what we've uncovered with our outreach. We clustered them together with our analysis of what we found. Together the answers form a fascinating pattern that reveals the thinking of many of today's top companies when it comes to artificial intelligence and machine learning.

Let's start with the question on everyone's minds: LLMs.

Are big enterprises adopting LLMs like OpenAI's GPT 3.5/4, Anthropic's Claude or any open source alternatives right now?

Importance of adopting xGPT/LLMs/generative AI as part of AI transformation during fiscal year 2023



Turns out, 67.2% saw it as a top priority to adopt LLMs and generative AI by the end of the year, with 49% rating it as 4 and 18.2% rating it as 5. That's a very unexpected result. While LLMs have proven incredibly useful already for individuals and are starting to work their way into small business workflows, we didn't expect them to rate as highly on enterprise radar, mostly because they are hard to control and don't fit the pattern of traditional deterministic IT applications. But it seems enterprises are hungry to unlock the value of generative text models and don't want to get left behind.

But while companies are eager to adopt LLMs and generative AI, it seems like there are a number of big blockers and challenges standing in their way.

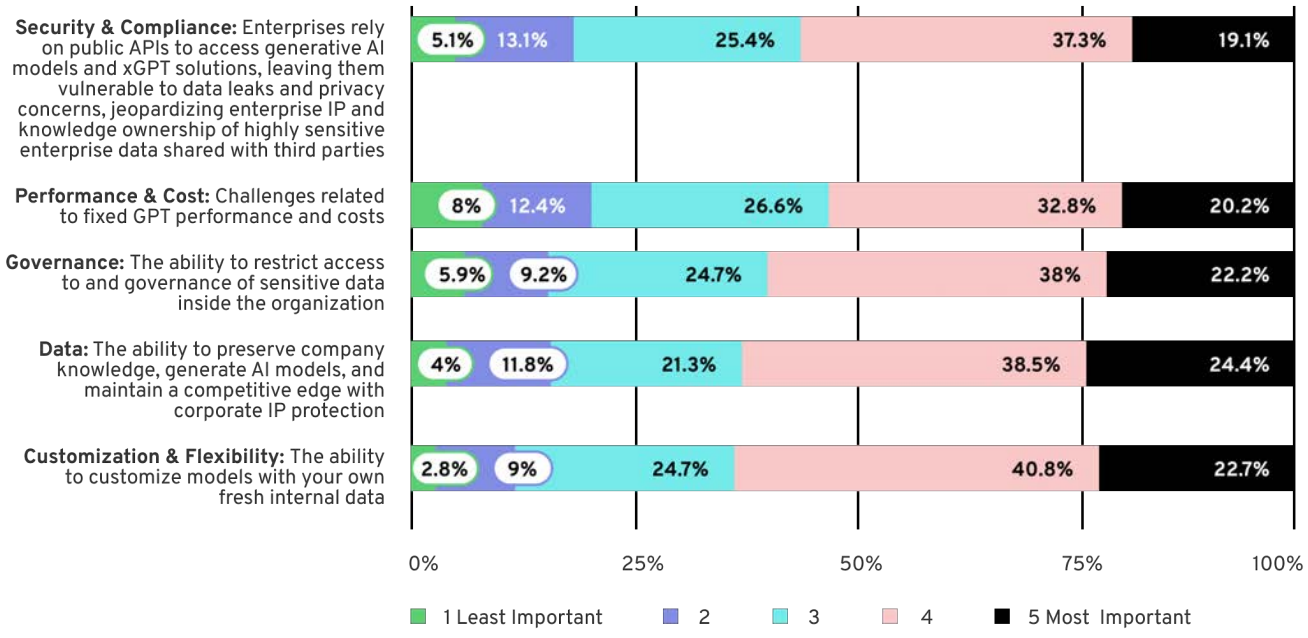
The biggest ones are customization and flexibility and the ability to preserve company knowledge and IP. Corporations have built up a storehouse of valuable assets and they want to protect them. They also want the ability to change the apps, models and frameworks as needed.

Down on the list are issues like compliance though they are not insignificant. Big business is facing a flurry of potential regulation, starting with the **EU AI Act** while also facing a mountain of current compliance challenges.

Performance and cost are also major challenges, with OpenAI API costs and performance hitting hard, a pattern we've seen with small businesses and now enterprises too. Though again, flexibility and customization and data sit at the top of corporate concerns.

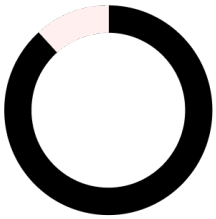
Privacy and security are linked to data concerns. Can companies legitimately use third party models without exposing company secrets?

Key challenges/blockers in adopting generative AI/LLMs/xGPT



Maybe that's why the overwhelming majority are implementing specific rules and policies around LLMs and generative AI, defining what developers and teams can and can't do with it.

Do you plan to implement policies specific to the adoption of Generative AI in your enterprise and/or implement internal policies around its usage?



88.3%

of companies polled plan to implement policies specific to the adoption of generative AI in enterprise and/or implement internal policies around its usage.

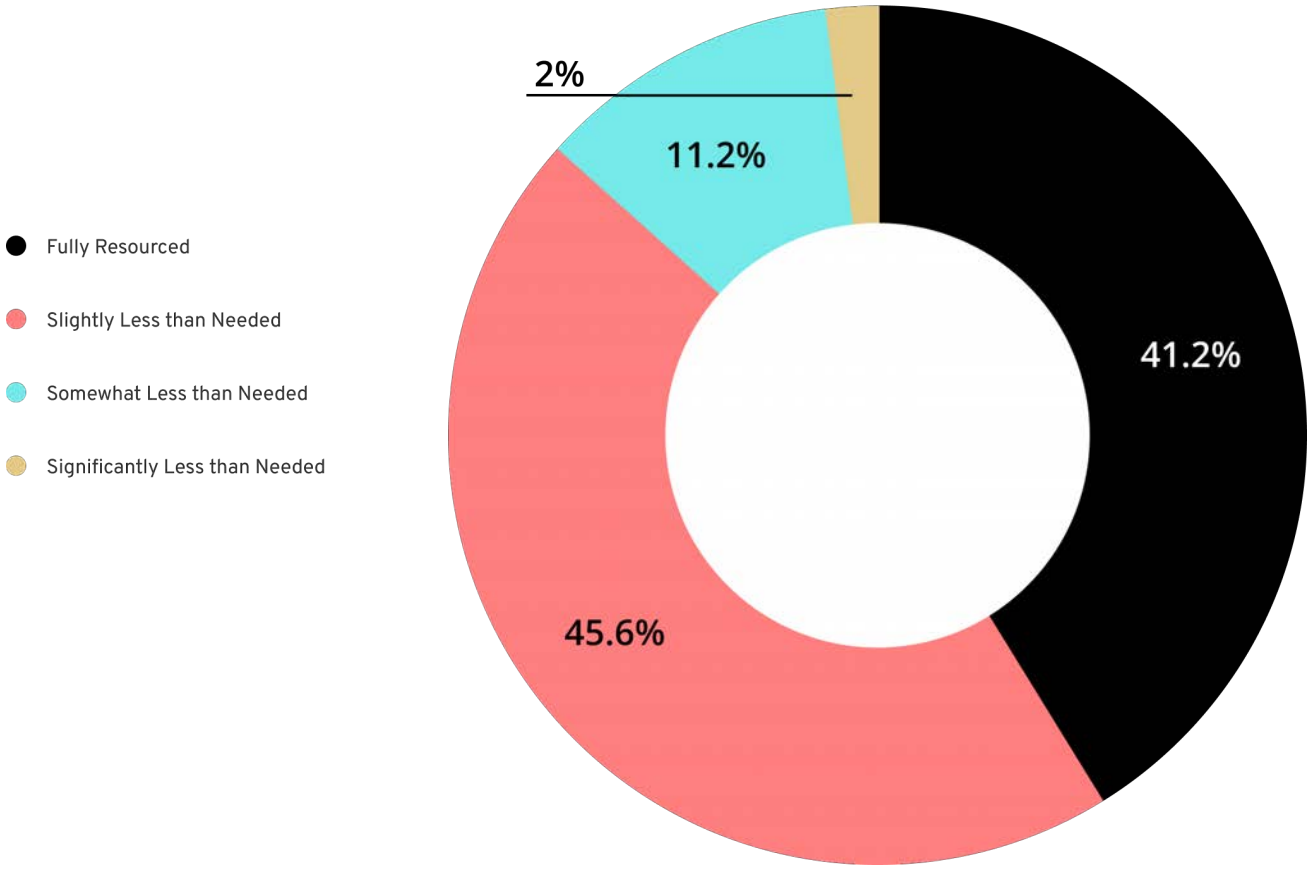
The next big blocker to successful deployment is resources. When we asked whether their teams had the right budget and people to make it happen, only 41.2% of the respondents felt they were fully staffed up and had the right budget to deliver on the promises of LLMs and generative AI.

For the rest of the people we talked to, 58.8% felt they were understaffed and didn't have the right funding to really drive transformation across the business, which is a tremendously worrying statistic.

While some companies likely need to rework their budgets for next year to deal with the surge of new possibilities, it's also a strong possibility that many companies simply can't get the people they need right now. Competition in the industry is fierce, with skilled people in data science, machine learning and MLOps hard to come by as multiple companies vye for their rare skillsets.

We expect to see hiring crunch across these fields continue for many years to come, as AI is only recently become a central focus of college and university curriculums and because the technologies are so new. It will take years to fill the gaps in the industry.

Budget/resources allocated for data science, ML/AI, and data engineering teams to create value and drive AI transformation/generative AI adoption across the enterprise



But if companies are planning to adopt LLMs what kinds of models are they looking to use? The vast majority are looking to use off the shelf models or models that call out to via API in the cloud.

That's a staggering turn around from only a few years ago where most organizations were set on training their own models from scratch and building data science teams to do it.

Today it seems that most teams prefer to use a model that can already do the heavy lifting. It seems many are not interested in using their own data to train their own models, though as we'll see later, many companies are interested in fine-tuning a foundational model or base model with their in-house data. It wasn't long ago that people thought of data as the new oil, but its turning out that many companies don't have overly unique data and that a well crafted base model will often serve a wide variety of use cases. Only a small subset of companies have very unique data they can leverage.

All of this points to a very clear trend. Most companies simply want things that work. They want the end result.

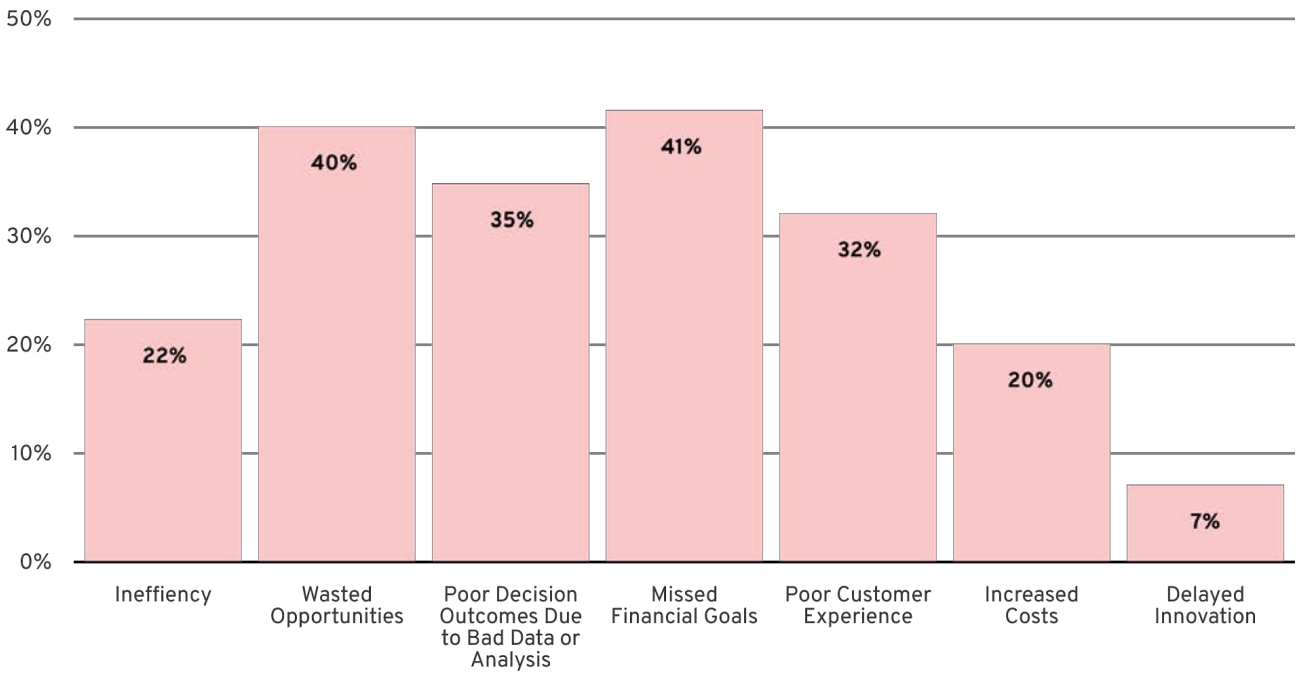
While many companies say they want to fine tune on their data, fine tuning is hard work. Fine tuning is updating the weights of a pre-trained model through more training on a smaller subset of data. A flurry of open source

work is making this increasingly accessible, particularly with Adaptors, such as **LoRAs**, which freezes the pre-trained model weights of an LLM and injects trainable rank decomposition matrices into each layer of the Transformer architecture, thereby greatly reducing the number of trainable parameters for downstream tasks. A number of these memory efficient fine tuning techniques, known as **Parameter Efficient Fine Tuning** (PEFT) have recently come storming out of labs across the world.

But the fact remains, even with these new methods, a range of new MLOps tools focused almost exclusively on fine tuning, many teams admit quietly that fine-tuning is much harder than it sounds and break a model without warning or skew it in strange ways.

Our sense is that as better and better out of the box models become available more companies will adopt them rather than struggle with the hassle of fine tuning unless they truly have a treasure trove of unique data. Most companies will opt for out of the box models and we expect this will increase with time.

Percentage of organizations that experienced the following impacts due to poor AI/ML operationalization or commercialization

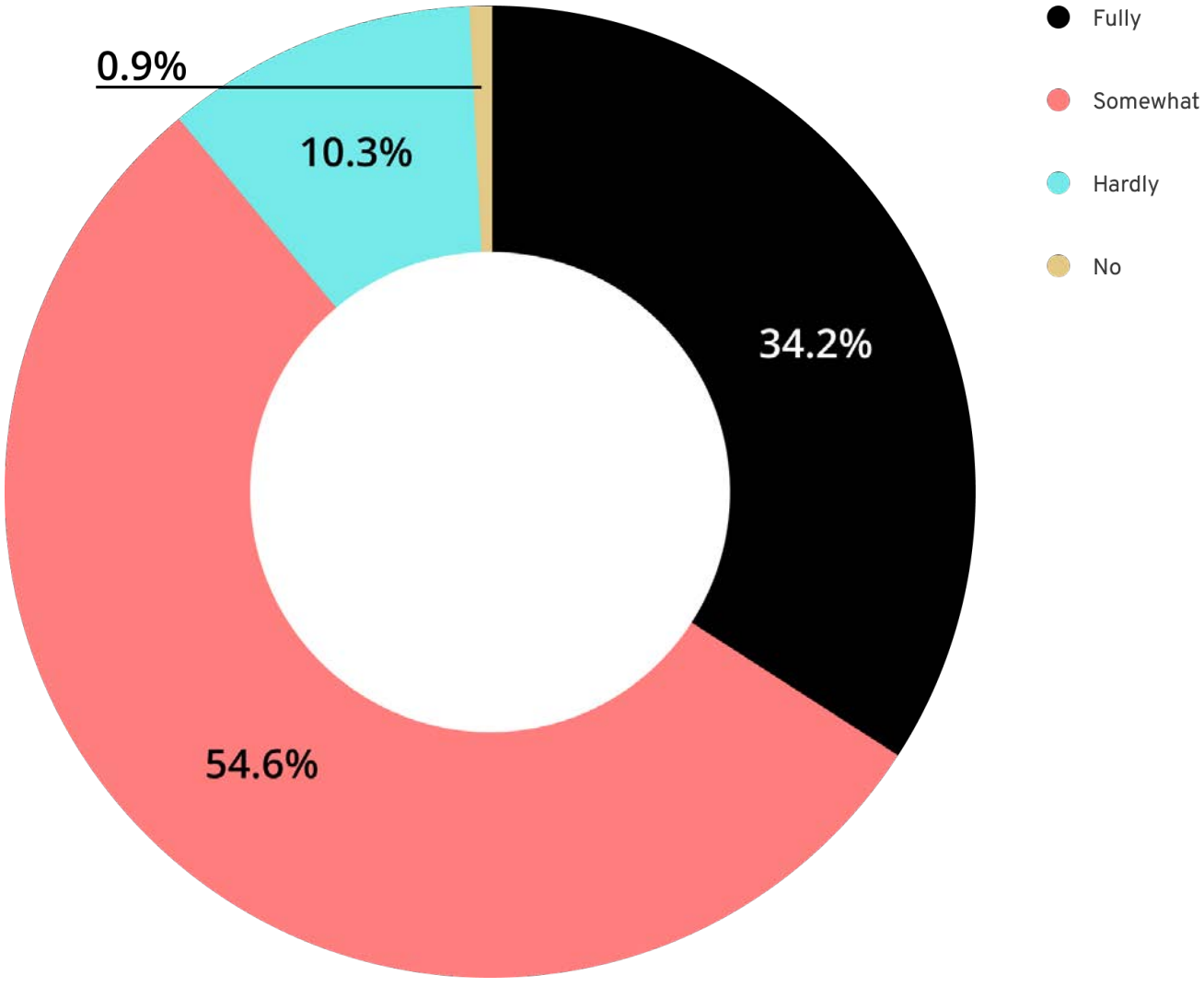


There are other big roadblocks to adoption in the enterprise. Many of them boil down to simple logistics and politics. Others revolved around showing ROI.

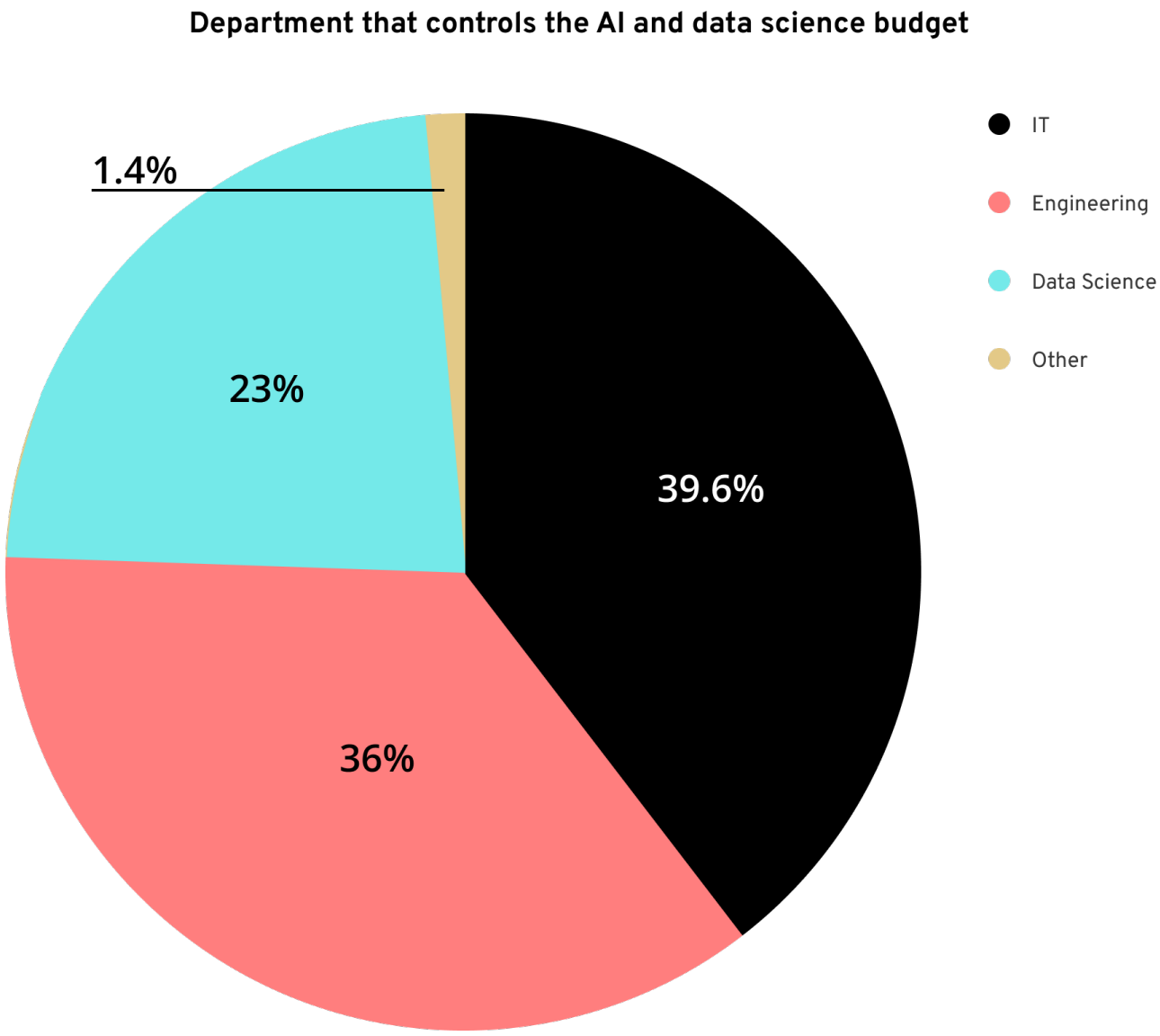
The following questions tell the tale.

The biggest shocker here is that most companies don't have a way to really show ROI for what they're spending on AI/ML. Only 34.2% felt they could fully show the return on investment. That makes it harder to justify budgets for the next year.

Ability to measure the impact and ROI of AI/ML projects on the bottom line

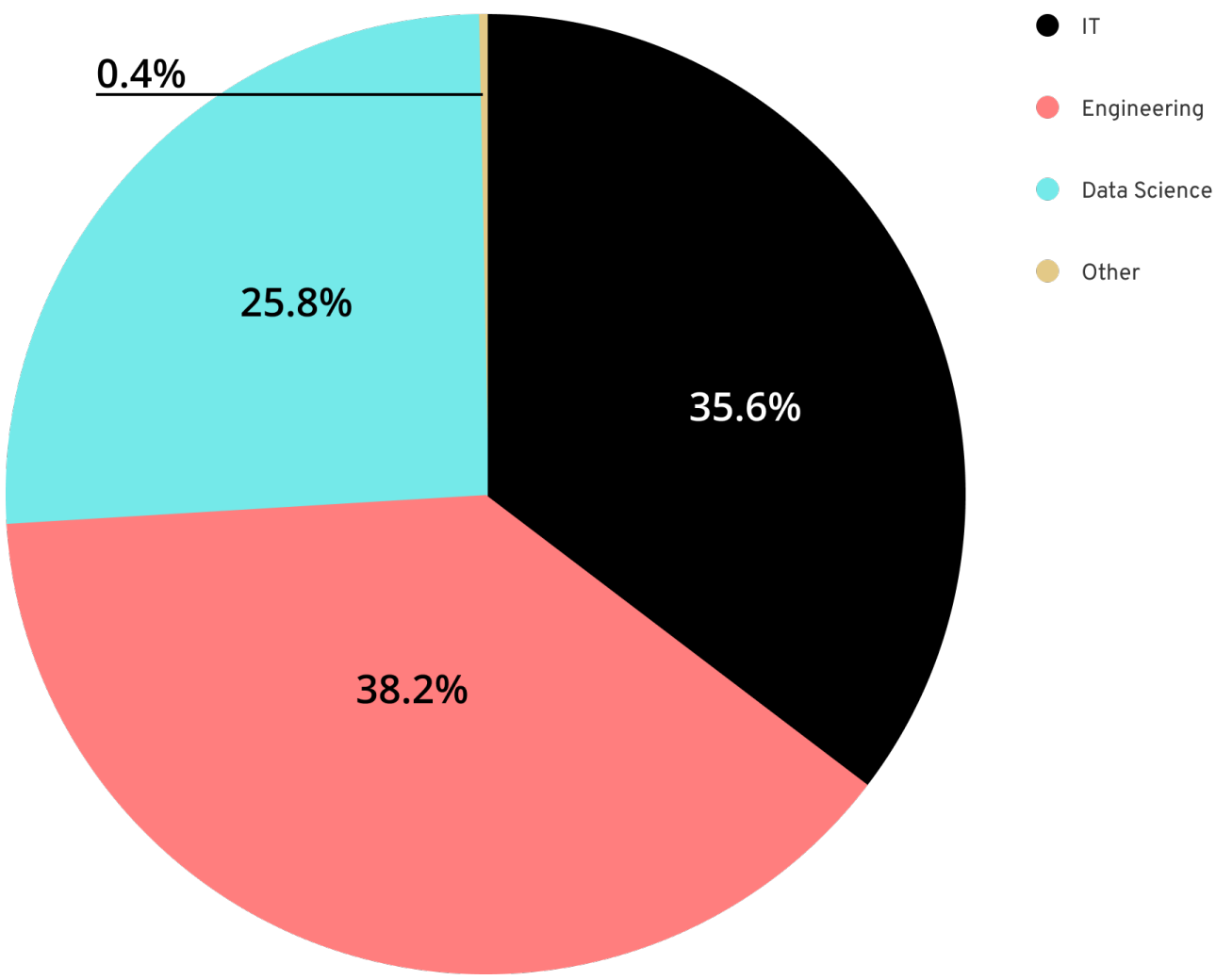


The other challenge is scattered teams who control AI transformation across the company, along with the budget to get things done. We see a near even split across data science teams, engineering and IT.



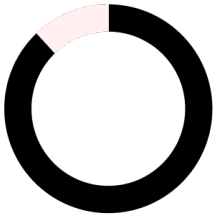
Even worse, who controls the priorities is split evenly across teams as well, with companies letting IT lead, others letting engineering take center stage and the rest letting data science teams lead the way. Each of these groups has different priorities and perspectives and sees problems in different ways so this often leads to inconsistency in how organizations rollout AI and see the benefit from it.

Department that determines the organization's AI, ML, and data science priorities



Perhaps that's why well over 80% of the people surveyed wanted to unify platforms across their teams, in an attempt to bring order and consistency to something that is proving anything but consistent. Unfortunately, the likelihood of this is incredibly low, as the tools that aid model training and deployment are tremendously different from the emerging stack of guardrails and orchestration tools needed to keep foundation models and the code surrounding them running right.

Is your organization seeking to standardize on a single AI/ML platform across departments (versus using different point solutions for different teams)?

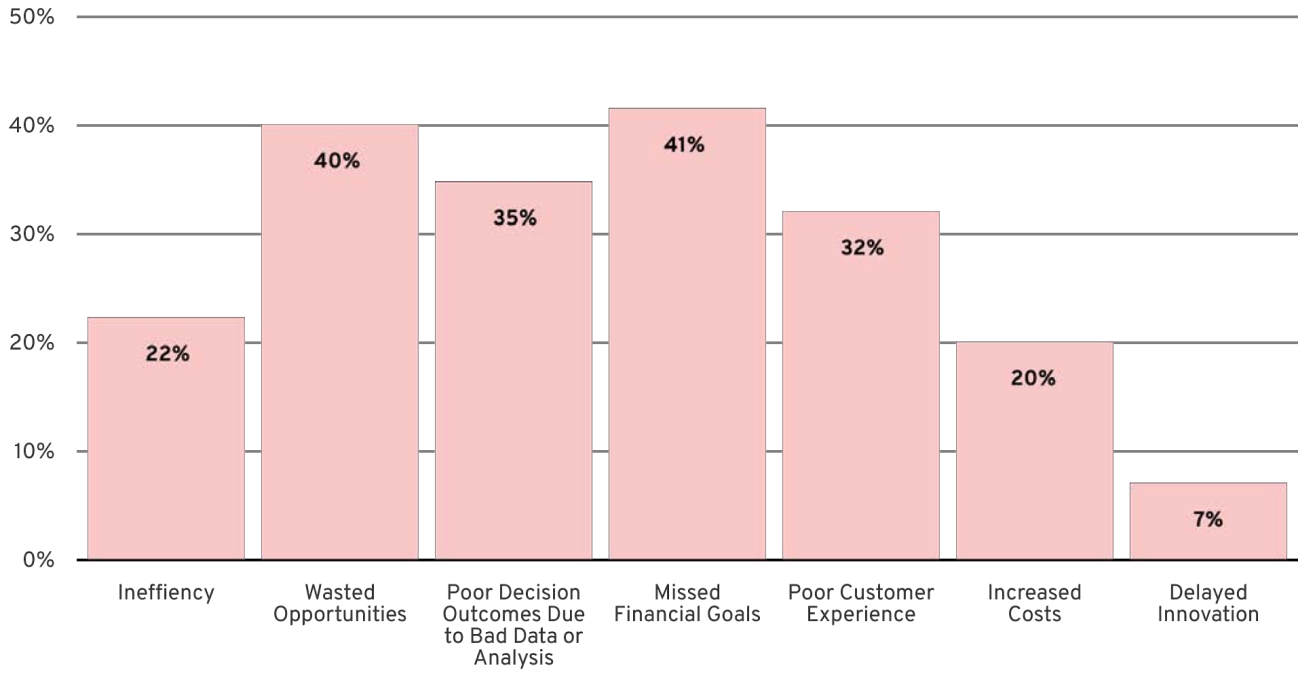


87.7%

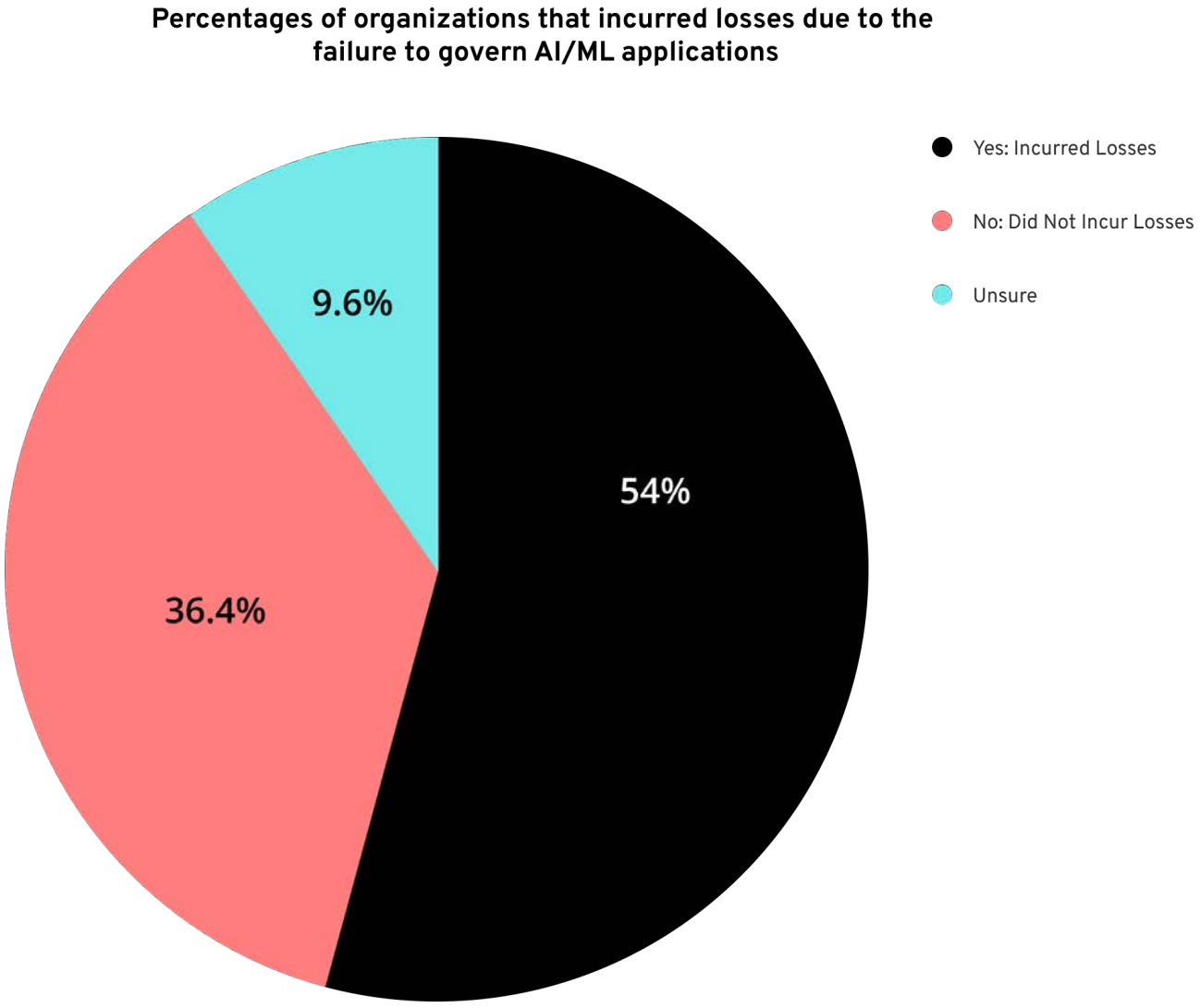
of organizations polled are seeking to standardize on a single AI/ML platform across departments (versus using different point solutions for different teams)

Companies have also faced big problems trying to drive value from their AI investments. Many organizations faced issues from poor customer experience to missed financial goals because of their AI investments, or just wasted opportunities and bad decisions.

Percentage of organizations that experienced the following impacts due to poor AI/ML operationalization or commercialization



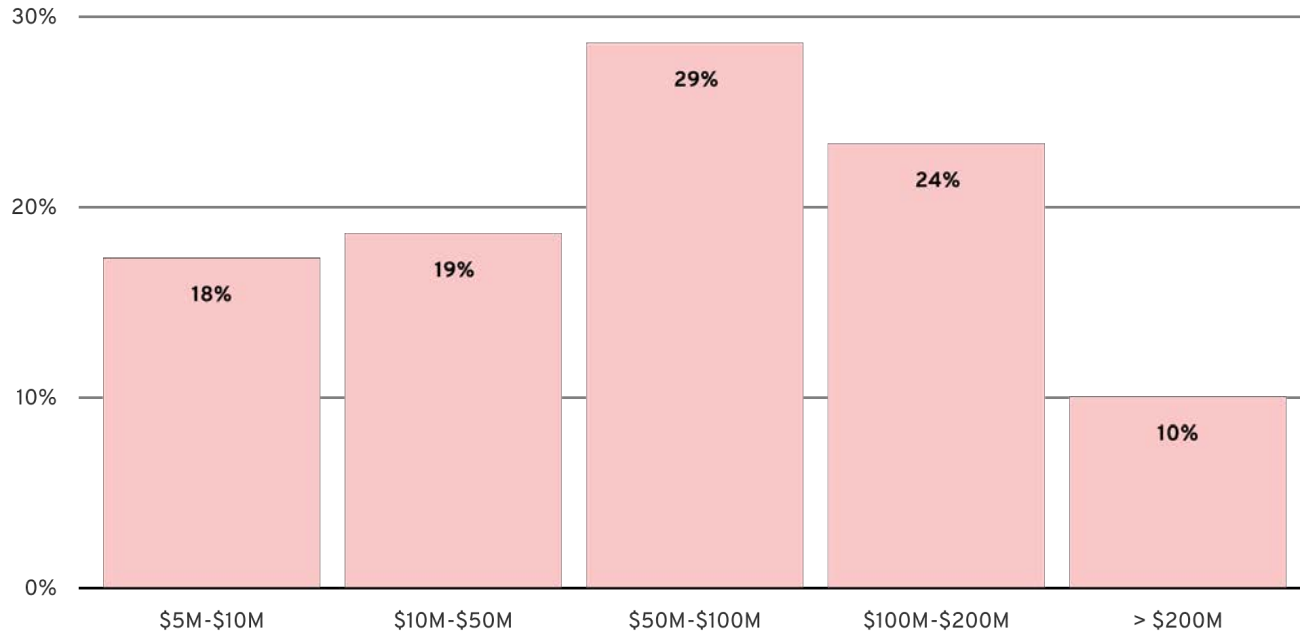
Even worse, struggles to manage and govern these applications have led to major financial losses across more than half of the companies we surveyed.



Worst of all, the size of those losses is staggering. 20% saw losses of 50M to 100M. 24% was losses of 100M to 200M and 10% saw losses of 200M or more from failures to govern the models and applications right.

Many organizations we've spoken to admit privately that governance is incredibly challenging and that they lack the tools they need to really understand why models and applications are making decisions that are severely lacking.

Size of loss for organizations that incurred losses due to failure in governing AI/ML applications



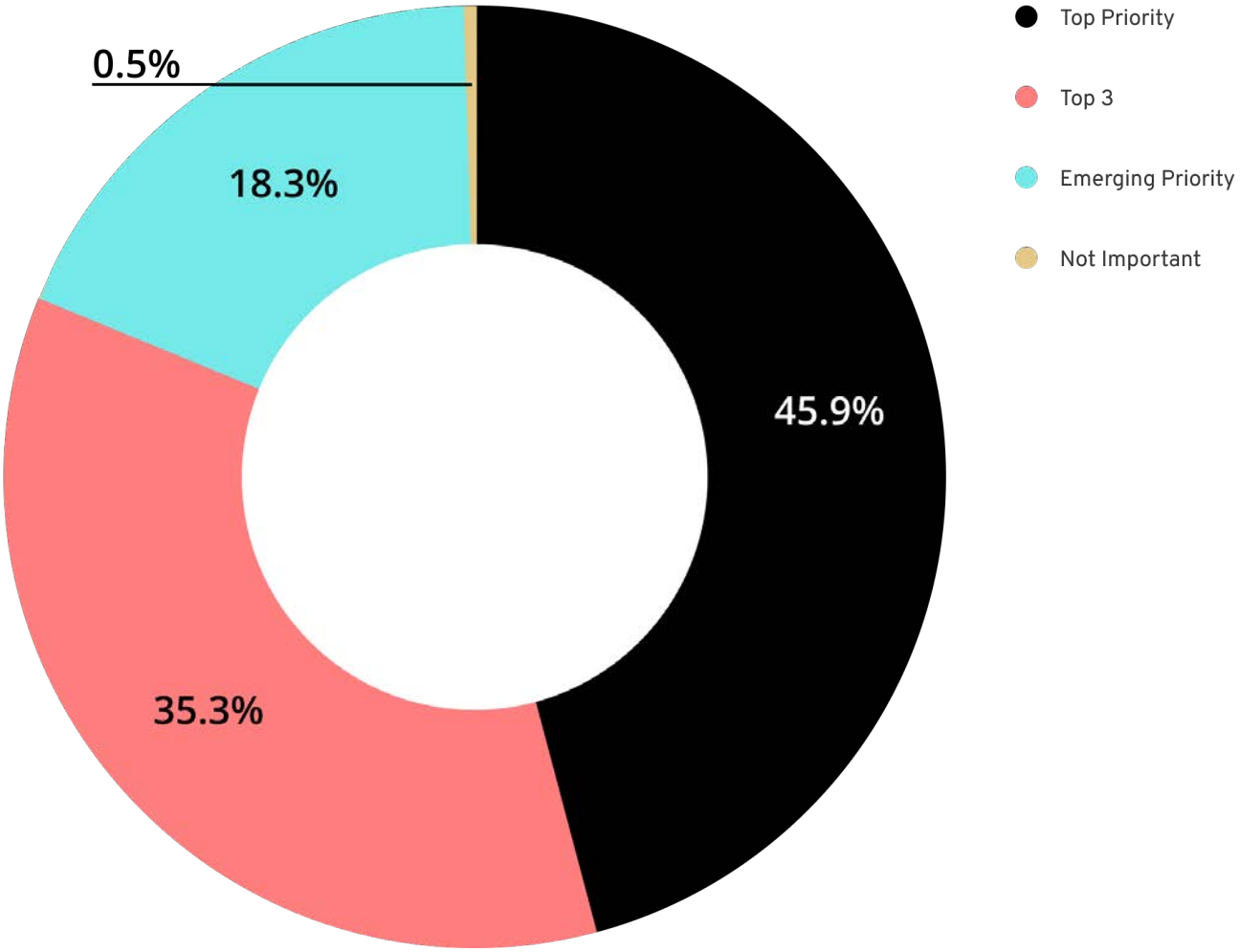
Some even confessed that the tools might not exist yet which could make explainability rules from the EU a pipe dream. ML models are still often a mysterious black box that makes it hard to understand what's going on under the hood.

These challenges will only get worse with generative AI and LLMs which are by their very nature non-deterministic and harder to control.

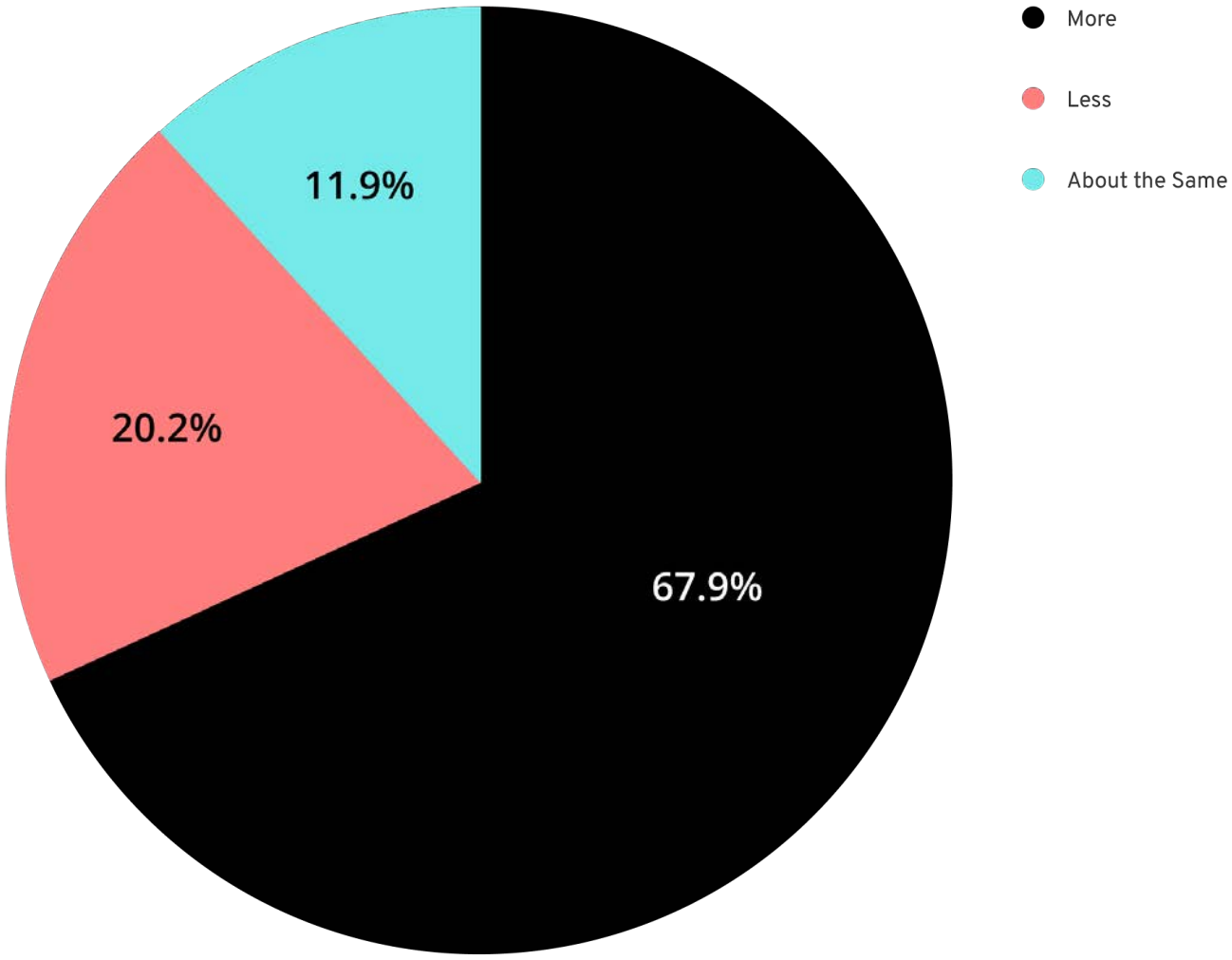
These are big numbers that should give every organization pause as they think about how to deliver value from their artificial intelligence investments.

Maybe that's why most of the organizations see it as a top priority to drive value from the AI spending this year, versus in the past where it might have fallen into more of an R&D budget with no expected value in the short term.

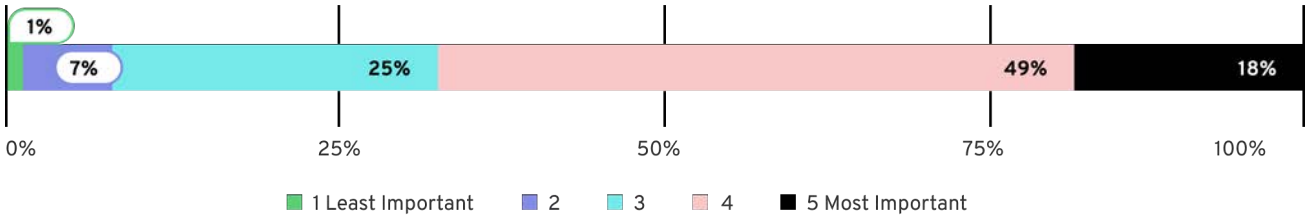
Priority of scaling AI and ML use cases to create business value for enterprise's data strategy over the next 36 months



Importance of creating value from AI investments compared to last year
(given the latest advancements and release of generative AI and LLM platforms, such as ChatGPT)

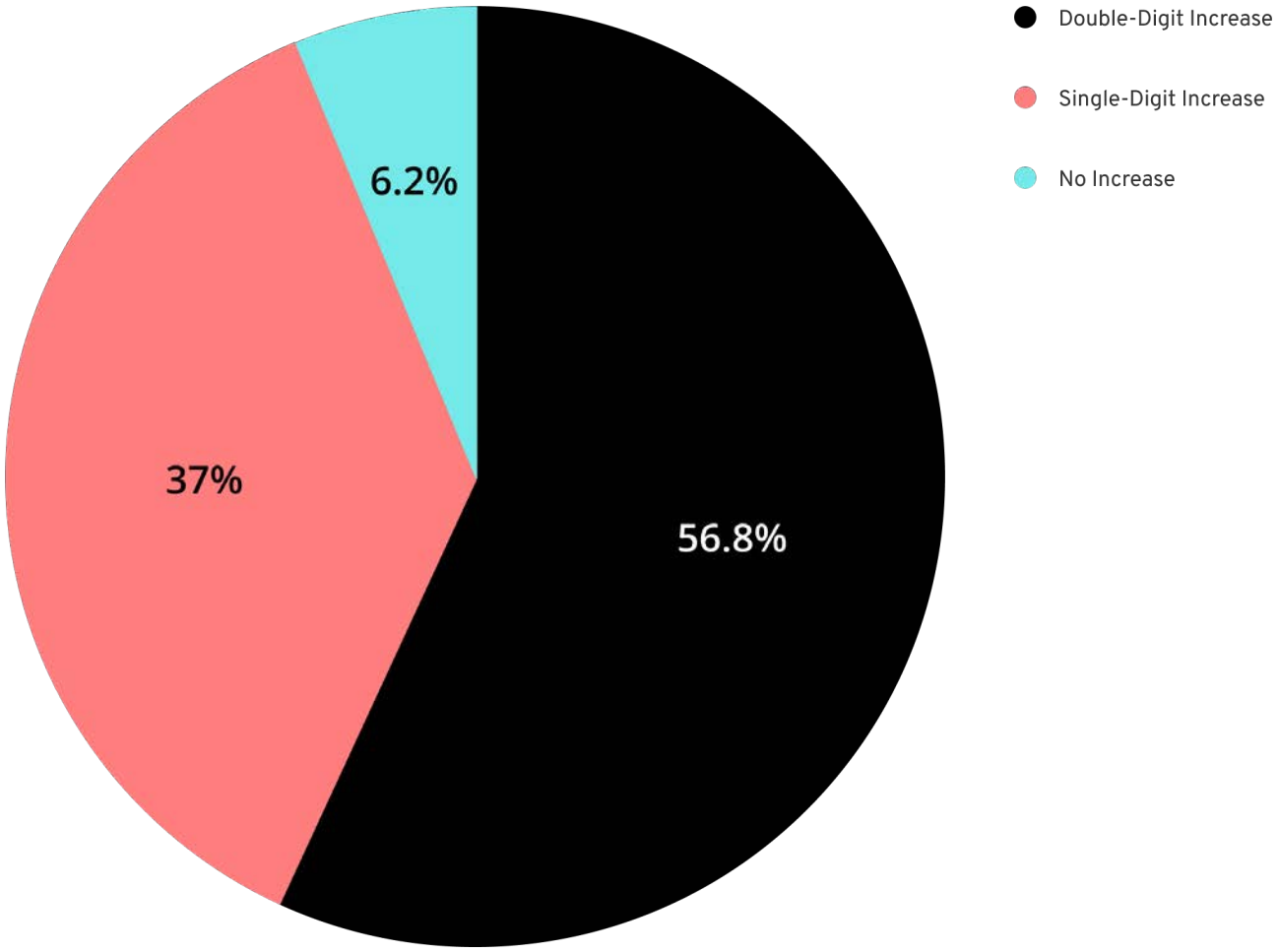


Importance for organization to create value from your AI/ML investments



The pressure to get value from these apps is increasing at all levels of these companies, going right up to the board of directors.

Expected revenue boosts from AI/ML investments and enterprise AI transformation in the upcoming fiscal year

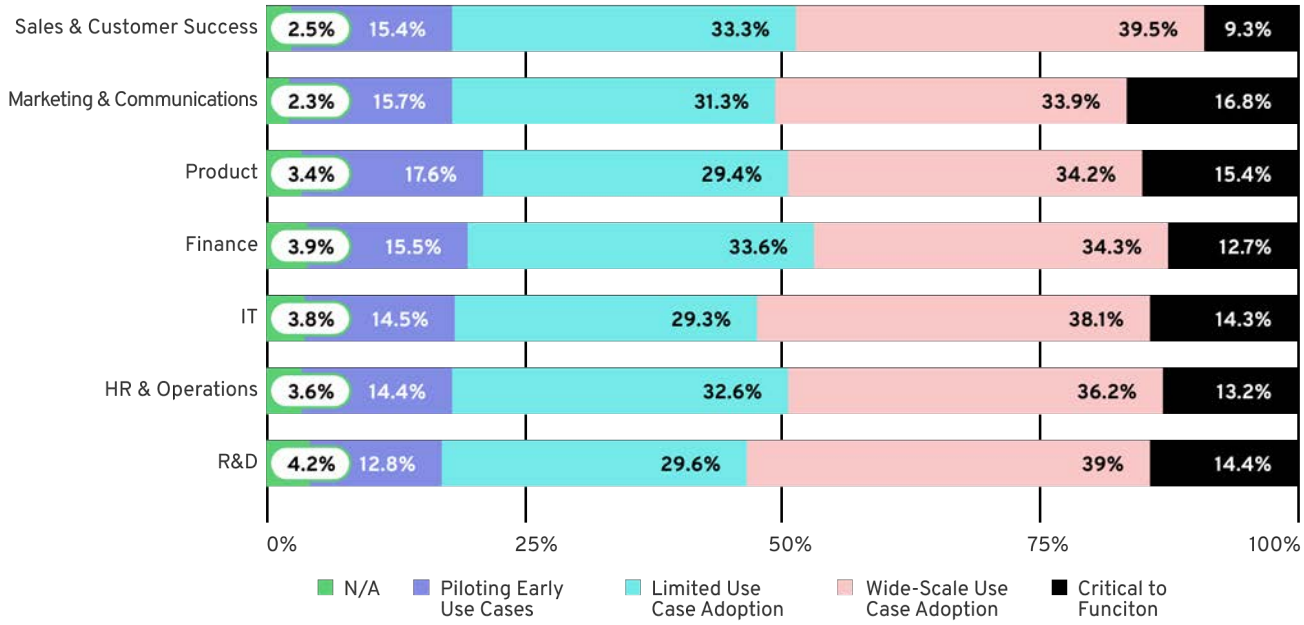


Interestingly enough though, despite all the challenges with driving ROI, AI has already become crucial to many businesses at the top level of the economy.

It's become essential in areas like marketing, sales and even the products these companies produce.

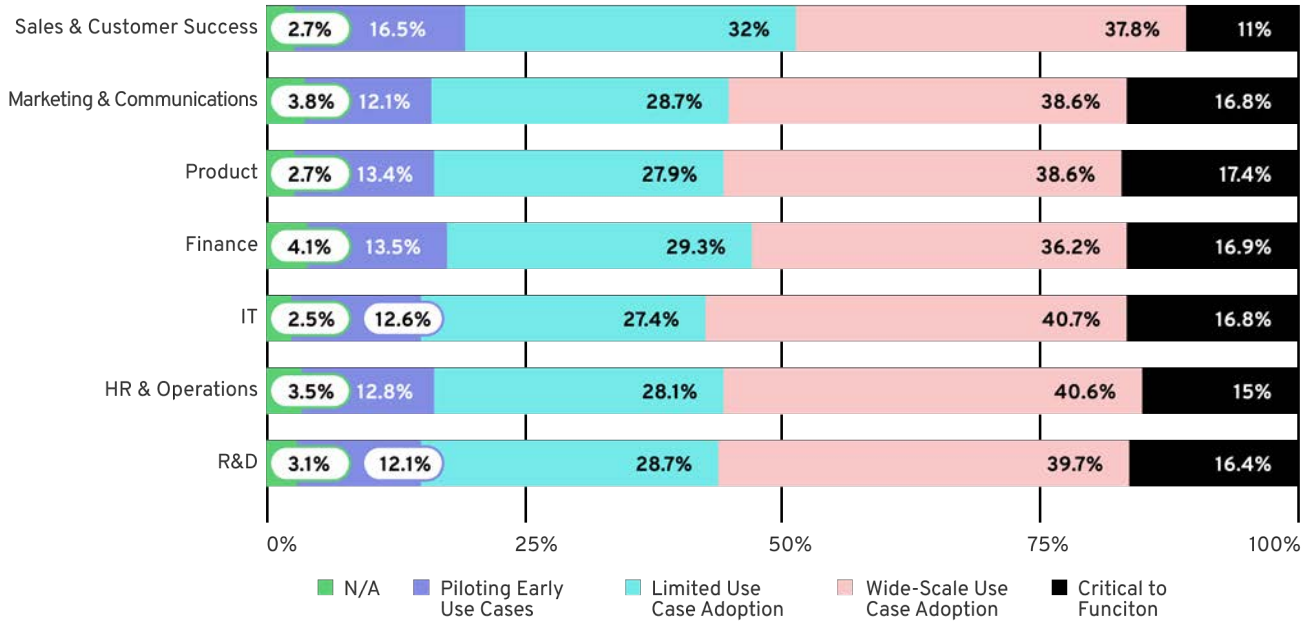
Where it's not critical we've already seen widespread adoption of Artificial Intelligence across a range of functions at these companies.

Extent of AI use in the core functions of the business today



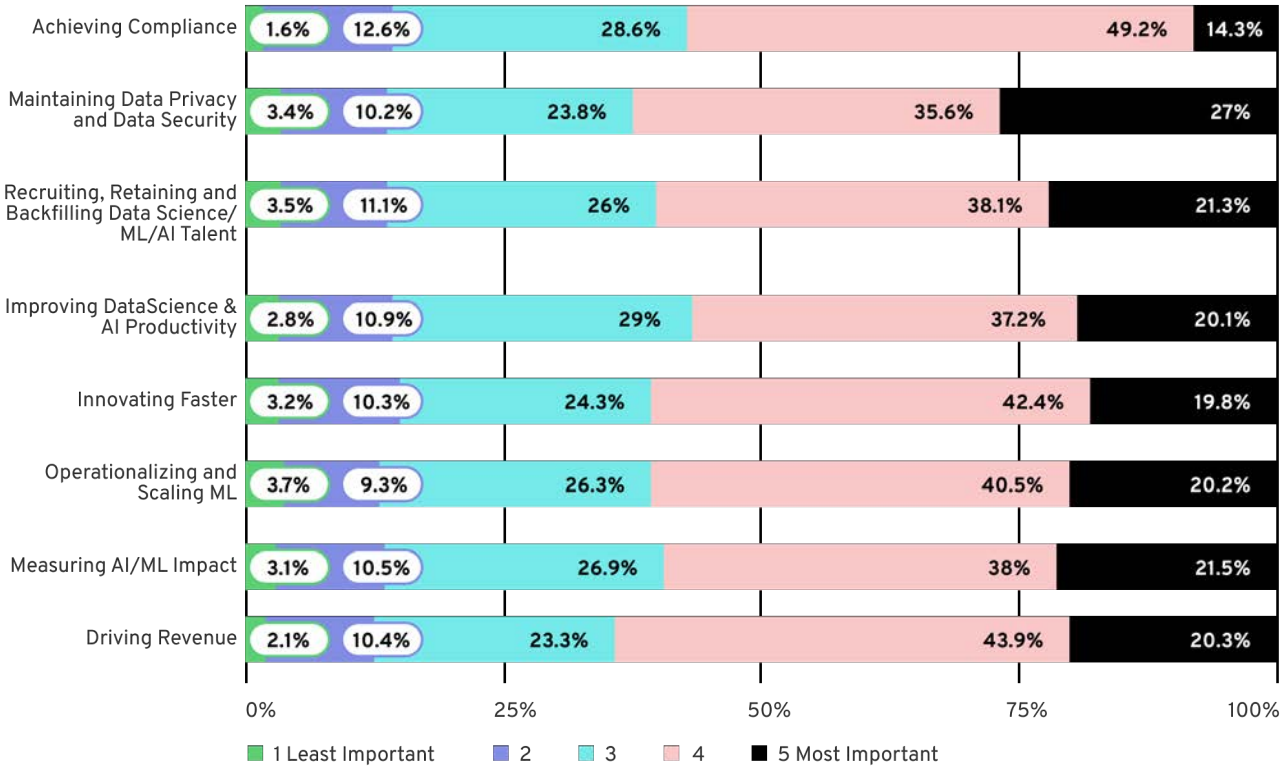
Executives expect that use to widen and become ever more critical over the next two two years. It's clear that AI is not slowing down even as we see slowdowns in other parts of technology.

Expected extent of AI use in core functions of the business in 2025



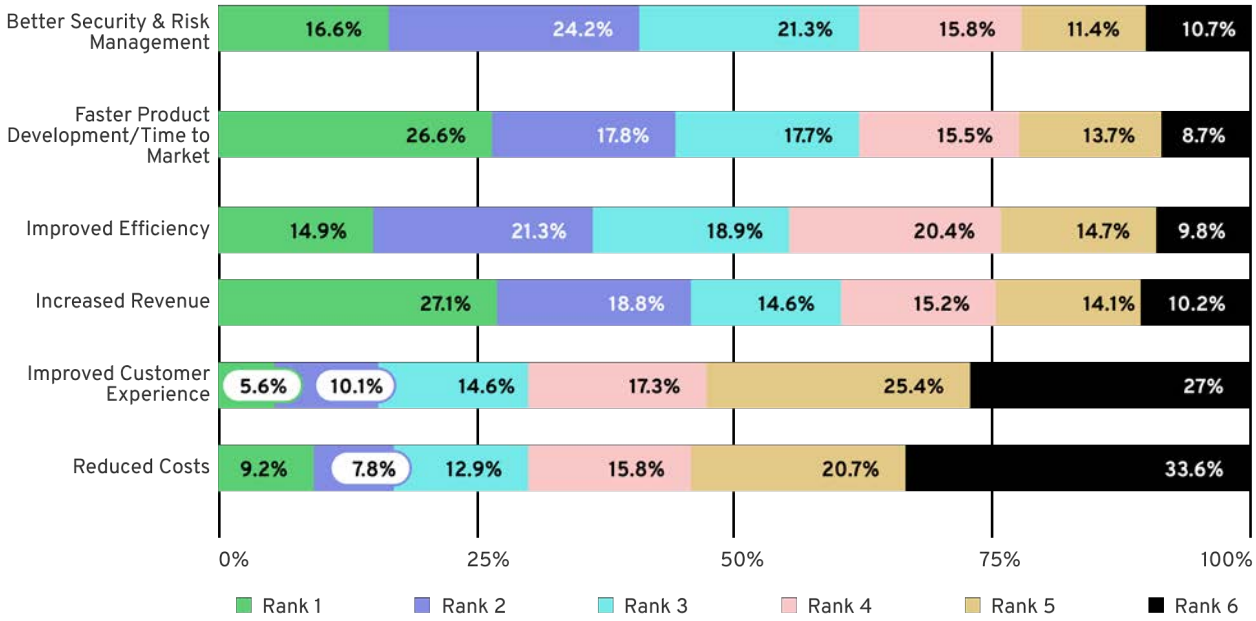
But despite all of these challenges the main thought on everyone's mind seems to be revenue. We expect these next few years to be the year of applications, not training and R&D. We're moving into the age of industrialized AI, where it gets more widespread and refined and engineered to be easier and easier to use. This is no longer the experimental phase of AI. It's now time for applications that matter and increase the bottom line.

Areas of importance from an AI/ML process standpoint



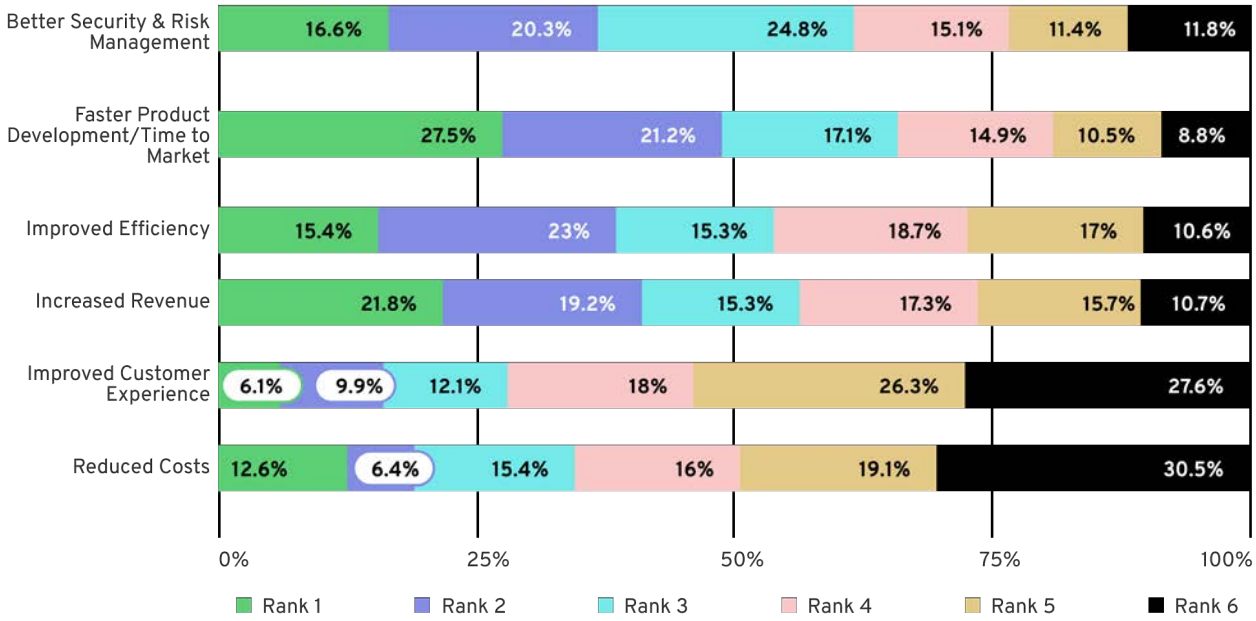
Even with challenges and losses, many companies are beginning to see real value from their investments in many areas, even if it's not consistent across the board. What accounts for the discrepancy? The most likely reason is that a company may see a very successful implementation of AI in one area of their business while struggling in another. We have not yet reached parity of AI applications across domains. For instance, a finance system might be very good at detecting fraud with ML but have challenges deploying a customer question answering system that proves effective and doesn't enrage end users.

Benefits of AI for enterprises today



Businesses expect even more value from AI in the next 18 months, especially from faster times to market and faster development cycles. They also expect it to improve the customer experience and reduce costs.

Expected benefits of AI for enterprises in 18 months



The Big Finish

Despite some big losses from previous AI deployments and big challenges to getting AI working and running effectively in large organizations, nearly every company we surveyed was very enthusiastic about weaving LLMs and generative AI into their products and workflow. Not only that, they expect it to drive more revenue or reduce costs or both.

Maybe it seems strange that enterprises would remain upbeat and enthusiastic about AI after taking some early hits but it's likely a result of where we are in the development of AI. In the past few years, we were just starting to see where it could fit and what it could do at the application layer. We had a lot of data science and research but not as many applications. All of that is changing. LLMs are very general purpose tools and when you combine them with the power of other models for doing specific tasks or with external knowledge bases they promise to deliver the long hoped for value in AI. It's sometimes dangerous to say "this time is different" but it does seem that this time is different and that's why enterprises are lining up to bring these advanced capabilities in house as fast as possible.

But there's the real challenge. Speed. Even with all that enthusiasm, large enterprises face a much larger headwind of regulation and compliance that can be a headache for even the most well run teams. With the EU focusing on explainability and interpretability in models and applications, especially for applications deemed high risk, companies may find their enthusiasm only gets them so far as much of interpretability and explainability remains in the early stages of research. There are already advanced explainability systems for traditional models but as of yet there are no ways to effectively detect truly bad answers or logical flaws from LLMs. It's likely to take a lot more research to get there. At the AIIA, we expect regulators to take a gentler hand at the dawn of new legislation, while software catches up to the need for better understanding of what these systems are doing and how they make sense.

Yet even with all of these challenges, AI is weaving its way into more and more enterprise domains and companies remain optimistic that it can drive more revenue and slash costs. All of this points to one thing: We're entering an age of industrialized AI.

Enterprises will not be left out and will drive much of the shift. Industrialization is where we see the rapid acceleration of research into real world applications, supported by ever more advanced AI chips, better software, and ideas imported from other domains that cross pollinate with traditional machine learning. It marks a time when companies around the world pour huge amounts of money, people and time into the pursuit of ever smarter software and machines. Now regular people and traditional coders and business people are getting their hands on super powerful models and taking them in bold new directions. They'll bring with them their know-how from decades of traditional computing and make the models, stacks and pipelines safer, more resilient and more predictable.

AI has already busted out of the walls of Big Tech R&D labs. Now even companies like Google are back on their heels as newer, faster and more agile smaller companies sense a once in a generation chance to reset the order of things and build new tech powerhouses. This was unthinkable even a year ago but it's the reality now. AI is poised to disrupt the old business model of the web in major ways. While the incumbents have the early advantage, they also have innovator's dilemma. They need to protect their old business model which is centered on advertising. But what happens when someone no longer needs to go to that ugly recipe site filled with ads after every paragraph

because AI can just tell them how to make herb chicken and rice for dinner? That's when the game really starts to change and a new business model will have to emerge to replace the old one.

As enterprises and small businesses find more and more use cases for AI and drive more revenue, it will turbocharge development of AI, rapidly advancing and refining the ideas of the research labs. As more and more product people, technical people, and traditional coders work with super charged models they'll take us in directions none of the researchers could have possibly imagined. They'll make the models smaller and faster, and find ways to weave them together with other traditionally coded apps, all while figuring out better ways to put guardrails on them so they deliver more and more consistent results.

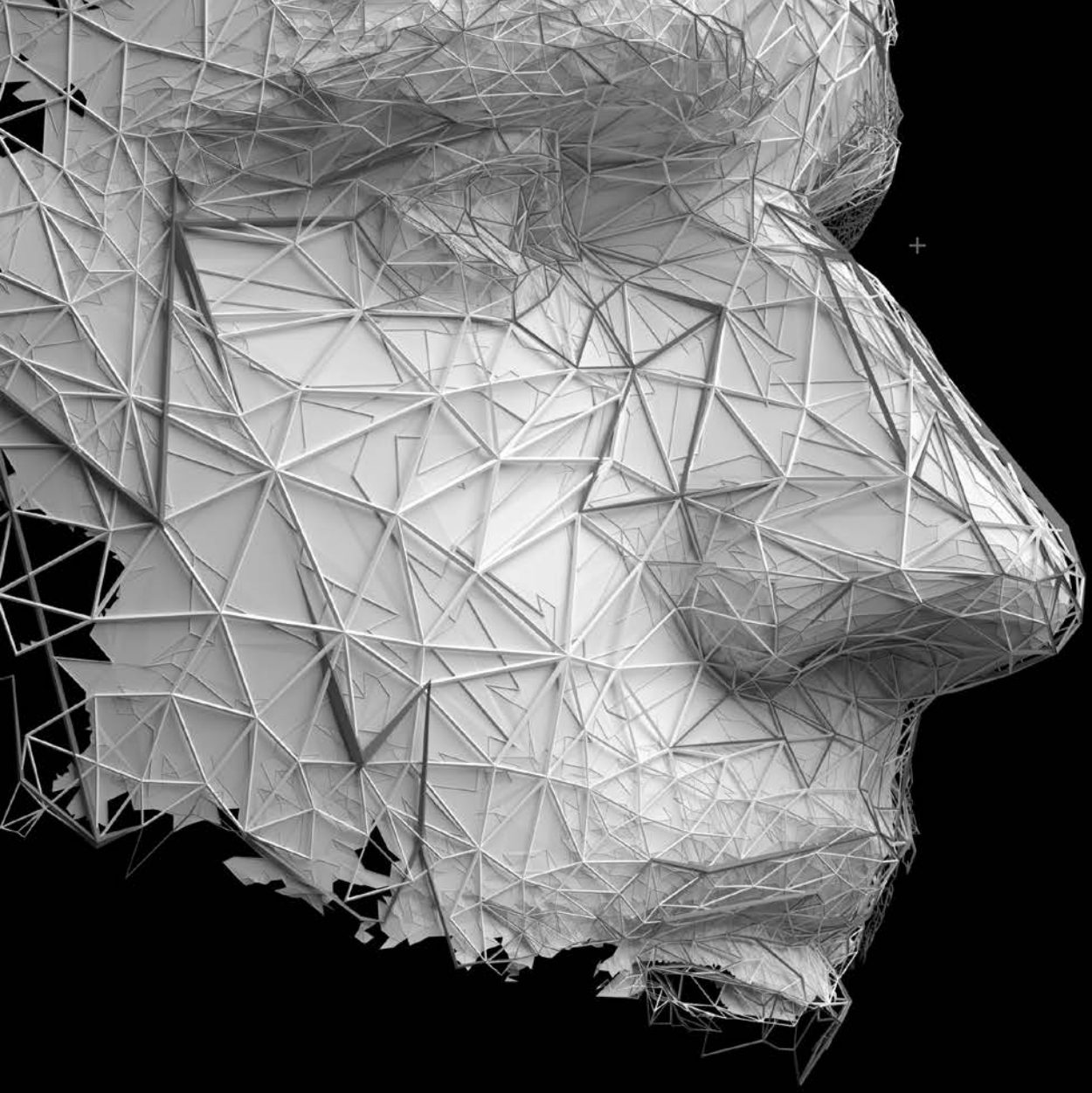
It's also industrialization that will push these systems to get safer, smarter and more aligned with what we want them to do. That's because companies and regulators and the general public will demand they get more reliable so they're more consistent, manageable and predictable. Not perfectly consistent because these are not and never will be deterministic systems, but they will get much more consistent, much more often.

We're also entering a time when many companies will never train their own models. They may fine tune one, but we expect even that to fall by the wayside as more and more models get stronger right out of the box. It's just too complicated and too expensive for the average company to build out a supercomputer, hire an ML team, a data team, an MLOps team and build a top of the line model. Expect more companies to look for finished products rather than train their own.

In the coming years, the vast majority of companies won't even interact with AI at a low level. That's like writing in Assembler, essential for a small subset of tasks, but way too complicated for most projects. Companies just don't have the time, money and people power to ingest billions of files, label them, do experiments, train a model, deploy it, optimize it and scale it.

But no matter what, big enterprises everywhere are already undergoing a remarkable transformation, weaving AI into every aspect of their business. We expect that to speed up in the coming years and to change many businesses dramatically as they drive untapped value from intelligent machines.

Enterprises are ready, willing and able to bring these systems in house and their efforts will make these systems better for everyone all up and down the economic chain. AI is poised to remake every aspect of the economy from logistics, to manufacturing, to healthcare and more. It's not a matter of if companies will overcome their challenges, it's a matter of when. It may not move as fast as everyone expects but make no mistake, big changes are coming to enterprises everywhere as AI weaves its way into every aspect of our lives.



AI Infrastructure Alliance

Website
ai-infrastructure.org

LinkedIn
linkedin.com/company/ai-infrastructure-alliance

Twitter
twitter.com/AiInfra

CLEAR|ML

Website
clear.ml

LinkedIn
linkedin.com/company/clearml

Twitter
twitter.com/clearmlapp