aporia

2024 AI & ML Report: Evolution of Models & Solutions

December 2023

Table of Contents

Introduction and Key Findings

Introduction & Methodology

Machine Learning (ML) is a field that has experienced rapid evolution over the past decade, is now seeing its production applications mature. Production ML workflows are becoming increasingly streamlined and the roles within this domain, which have significantly evolved and become more defined in recent years are crucial to the effective implementation of AI technologies.

Today, with a wide range of industries integrating AI into their processes and leveraging vast volumes of data to enhance product and workflow efficiency, optimizing production ML has become essential for maximizing value creation.

While some companies invest in large teams of data scientists and ML engineers to develop, deploy, and manage models in-house – often incurring high expenses due to the salaries these roles command many are recognizing the advantages of utilizing comprehensive ML platforms. These platforms not only automate various tasks but also enable highly qualified professionals to focus on value-generating activities, such as model development. Moreover, these platforms play a crucial role in managing aspects like AI hallucination, which is vital for maintaining business reputation, customer loyalty, and revenue

Our aim for this survey was to understand the day-to-day challenges of ML production leaders both from a technical and business perspective, how much time and money they spend dealing with these challenges, and what benchmarks are used when deploying ML models to production.

Methodology

To get more insight into the current state of machine learning models and solutions, we commissioned a survey of 1,000 ML professionals to shed light on their processes, priorities and most pressing challenges.This report was administered online by Global Surveyz Research, an independent global research firm. The survey is based on responses from ML engineers, MLOps engineers and ML Platform engineers, from companies ranging between 500-7,000 employees in the US (80%), Canada (10%) and UK (10%). Respondents hailed from all industries, except for Government and Education. The respondents were recruited through a global B2B research panel and invited via email to complete the survey, with all responses collected during September 2023. The average amount of time spent on the survey was 6 minutes and 6 seconds. The answers to most of the non-numerical questions were randomized to prevent order bias in the answers.

Key Findings

1

88% of ML practitioners agree that real-time observability is crucial for the success of ML models in production

There's a clear consensus among most ML engineers – regardless of their industry or team size – that real-time observability is a crucial requirement for the success of ML models in production (Figure 3), because without it, they are oblivious to any issues that may occur. 88% of ML practitioners said it was either somewhat crucial (61%) or very crucial (27%).

2

On average, companies spend four months building production monitoring tools and dashboards

96% of respondents reported they spent at least one month building production monitoring tools and dashboards for their last project, with the majority spending an average of four months building the tools (Figure 7). Given that the average annual salary for an ML engineer in the US is around $160k, the time versus cost consideration in building these tools in-house is therefore considerable, because it means that highly qualified talent is tied up for too long on building the tools themselves, when they can be working on more valuable tasks. Using an ML platform solution to build production monitoring tools and dashboards is a far more cost-effective alternative, because it frees up ML engineers to work on projects that actually create value, like deploying ML models.

3

93% of ML engineers encounter issues related to production models on a daily/weekly basis

Observability tools – which generally take time to build due to the complexity of AI models – are built by data scientists and then deployed into production by ML engineers. When issues arise with these models in production, ML engineers – who typically do not understand how the models work – rely on the data scientists to identify and triage the problems. The fact that 93% of the respondents encounter issues with their production models on a daily/weekly basis (Figure 8) highlights just how important it is to monitor and identify issues quickly, because a high volume of production issues can have financial implications for the business. Using an AI performance platform, however, can significantly reduce the time required to identify issues related to production models, freeing up data scientists to focus on more valuable tasks.

4

69% of respondents encounter infrastructure-related challenges when deploying new ML models to production

Given that machine learning is still a new and evolving field, this finding (Figure 2) is testament to how difficult it is to deploy an ML model to production and make sure it is running as it should, especially in smaller teams where resources are limited. Using 'off-the-shelf' ML platforms can compensate for the lack of resources in smaller teams, saving them significant time on building ML models from scratch and eliminating the need to deal with a variety of other infrastructure-related challenges.

5

89% of ML engineers say their GenAI models show signs of hallucination

The majority (89%) of ML engineers in companies that use LLMs and generative AI models (e.g., for chat bots, virtual assistants, etc.) say their GenAI models show signs of hallucinations to various degrees (Figure 11). The severity of hallucinations can range from factual errors to content that's biased and even dangerous, and given the immediacy with which hallucinations in ML models can occur, they must be dealt with as swiftly as possible. Failing to do so can not only result in bad customer experiences but also tarnish their trust in the brand/product and lead to lost revenue due to reputational damage. One way to do this effectively is with a platform that minimizes hallucinations by creating a deterministic layer of a company's ground truth. This helps the model to understand whether the AI-generated content reflects the ground truth or not, and prevents it from presenting incorrect content.

6

83% of the respondents agree that it's important to monitor AI bias in projects

With companies increasingly transitioning to AI models as a key business component, there is growing concern about AI models becoming self-efficient in making decisions, because by leveraging all possible data, including data that's biased, its output could also be biased or even discriminatory. This can have serious repercussions on the business, from disastrous PR to falling stock prices. Observability is therefore crucial for AI product developers to ensure that ML models meet regulatory standards by being fair, accurate and ethical. 83% of the respondents agree that it's crucially important to monitor AI bias in projects on some level (Figure 10), with 27% saying it is very crucial and 56% saying it is somewhat crucial. The main challenges they encounter in relation to AI bias (Figure 9) are difficulty in identifying biased data (19%), inadequate tools for monitoring bias (18%), and lack of understanding of bias implications (12%).

Survey Report Findings

Most Challenging Infrastructure Aspects when Deploying a New Model

Given that machine learning is still a new and evolving field, there are very few (if any) widely accepted best practices that can be used to eliminate infrastructure-related challenges when deploying new ML models. Predictably therefore, **99% of respondents indicated they encounter a very broad range of production ML challenges**, with setting up environments (15%), monitoring and observability (14%), and versioning and governance (14%) topping the list.

This shows that the process of setting up production ML environments and ensuring effective monitoring and observability is incredibly challenging in itself, before even reaching the stage of tackling other challenges relating to deploying new models.

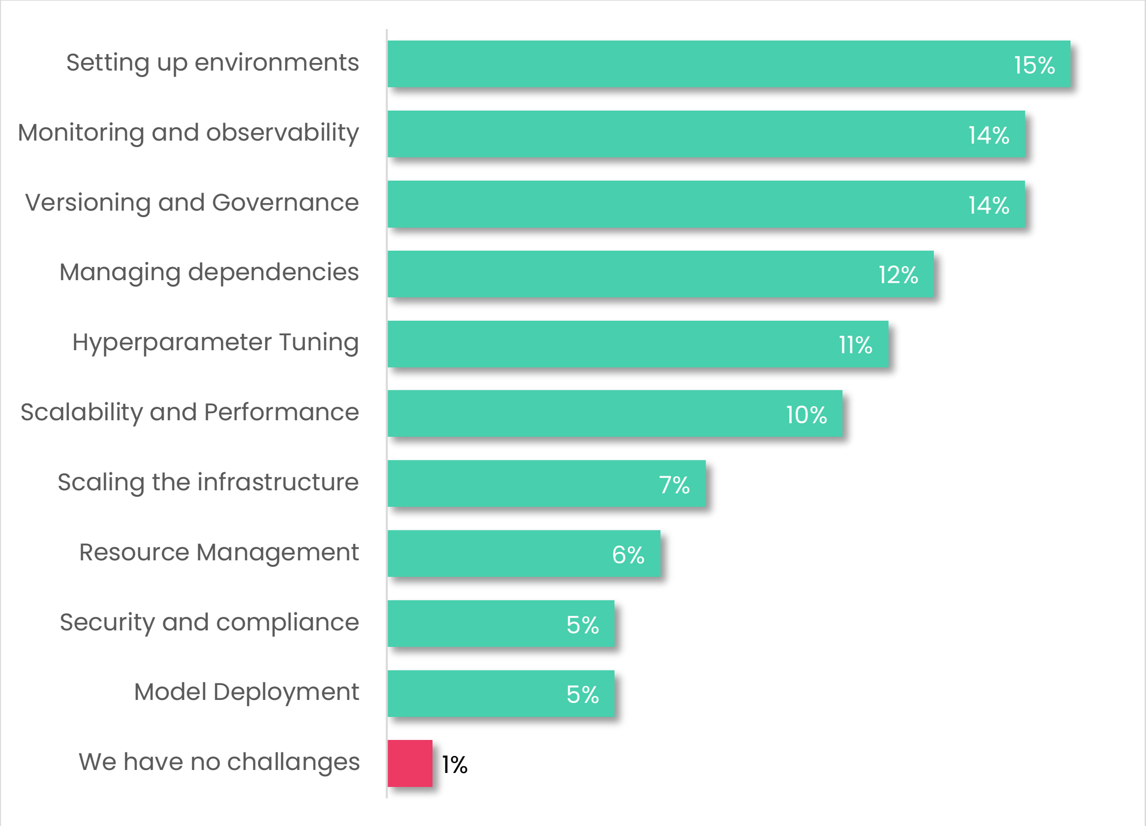

Figure 1: Most Challenging Infrastructure Aspects when Deploying a New Model

Frequency of Difficulties When Deploying New ML Models to Production Due to Infrastructure-Related Challenges

When asked how often they encounter difficulties when deploying new ML models to production due to infrastructure-related challenges, 69% of the respondents reported they experience difficulties either sometimes (46%) or often (23%). 24% say they rarely encounter difficulties, and 7% say they never do.

Deploying a new ML model to production is a resource-intensive process, so it's not surprising that small companies with <10 ML engineers encounter difficulties more than twice as often (28%) as large companies with 10+ ML engineers (12%).

Given that machine learning is still a new and evolving field, this finding is testament to how difficult it is to deploy an ML model to production and make sure it is running as it should, compared with DevOps deployments of software to production, for example, where there are already proven best practices and benchmarks for success.

ML platforms are a good 'off-the-shelf' solution because they can compensate for the lack of resources in smaller teams, saving them significant time on building ML models from scratch and eliminate the need to deal with a variety of other infrastructure-related challenges.

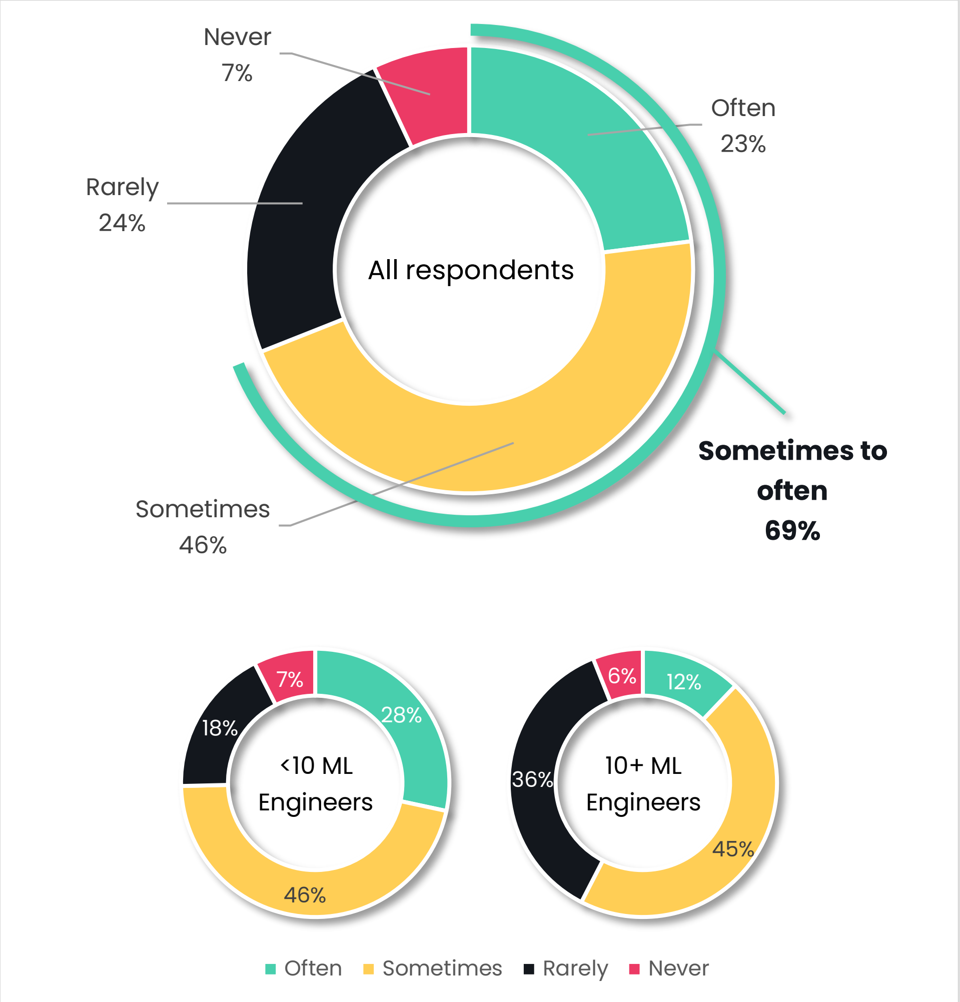

Figure 2: Frequency of Difficulties When Deploying New ML Models to Production Due to Infrastructure-Related Challenges

The Need for Real-Time Observability in ML Models

When asked how crucial they think real-time observability is for the success of ML models in production, 88% of the respondents say it is crucial on some level, with 61% indicating it is somewhat crucial and 27% indicating it is very crucial. Only 12% of the respondents say it isn't crucial at all.

This shows that **most practitioners in the market understand that real-time observability is an indisputable requirement for the success of ML models in production**, because without it, they are oblivious to any issues that may occur.

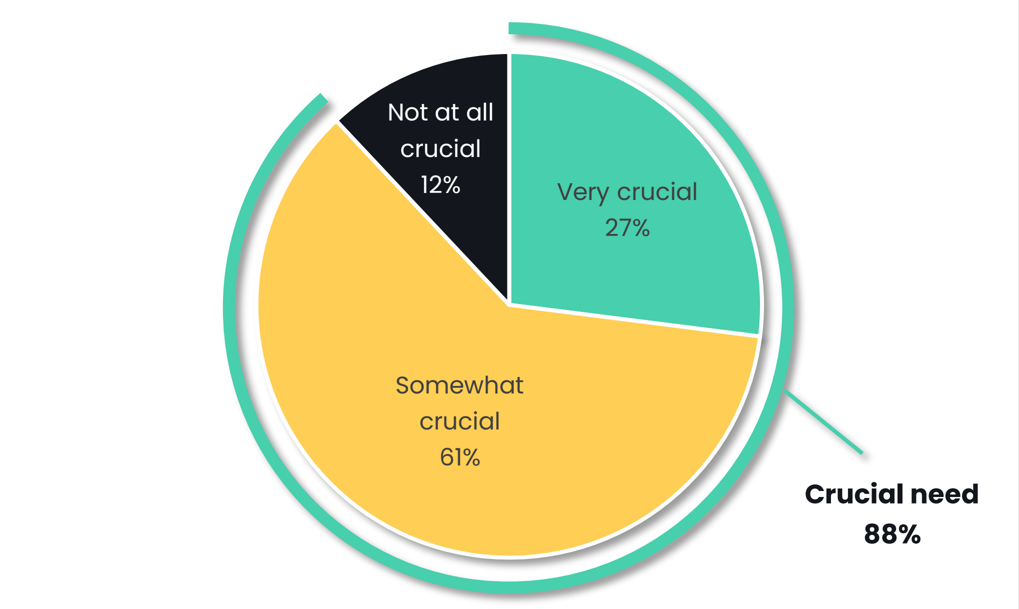

Figure 3: Need for Real-Time Observability for the Success of ML Models in Production

Challenges in Real-Time Observability in ML Models

We asked respondents, "What do you believe is the main challenge in implementing real-time observability in ML models?" to get an idea of the issues they encounter most regularly, which can potentially set them back once they are in production.

Once again, the range of challenges is extremely broad. **The top responses are complexity in integrating observability (19%), scaling and distributed systems (12%), and feature engineering (10%).**

As with the infrastructure-related challenges, integrating observability is the main challenge also when it comes to in-house observability tools. Once the tools have been built in-house, integrating them with the existing environment and extracting value can be incredibly difficult.

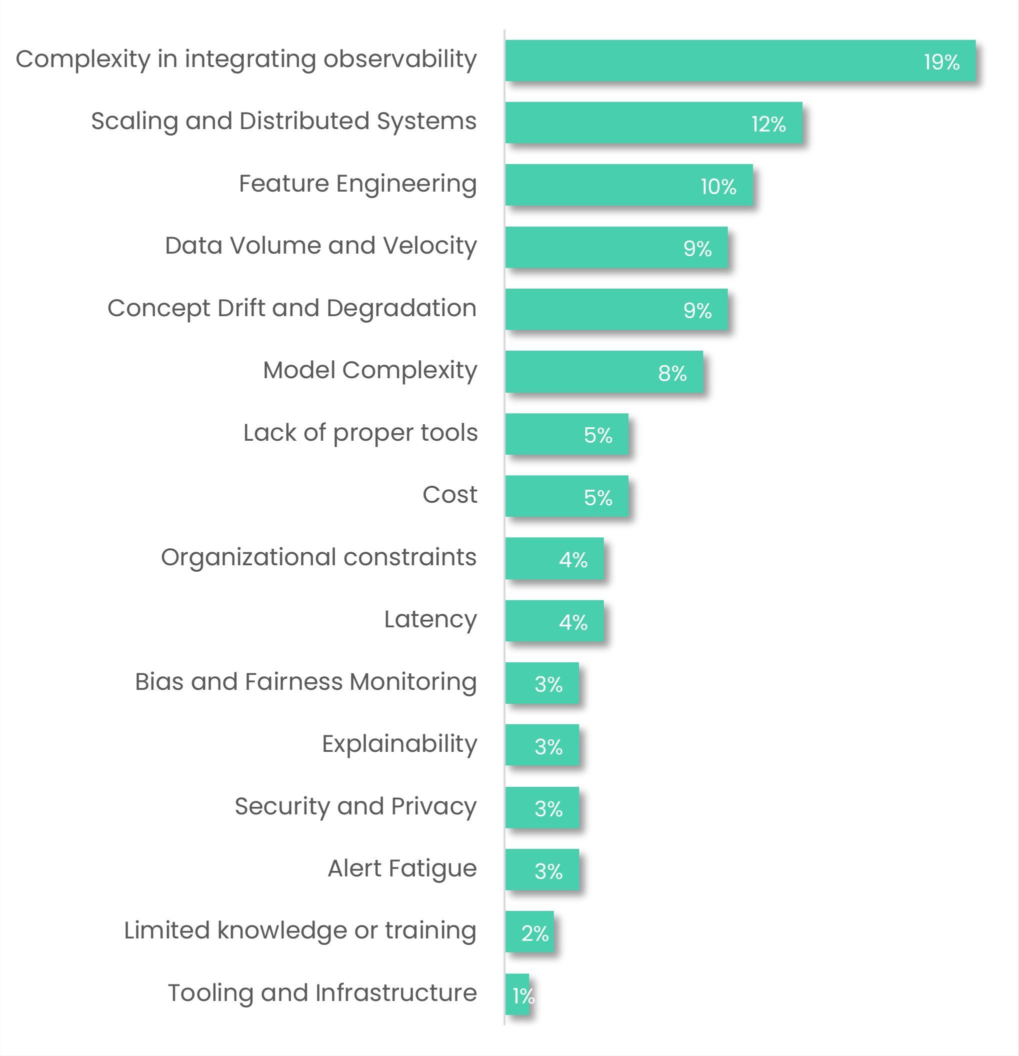

Figure 4: Main Challenges in Implementing Real-Time Observability in ML Models

Top Challenges with Current ML Monitoring System

The top challenges respondents face with their current ML monitoring system are difficulty in setup and maintenance (32%), integration with existing tools (30%), and limitations in providing value quickly (29%), which likely due to the extensive time involved in building and integrating these monitors, and understanding what needs to be tracked and monitored.

When further investigating this finding according to team size (Figure 6), **36% of those with <10 ML engineers say they feel limited, compared to only 15% of companies with 10+ ML engineers**. Given the complexity of building monitoring in-house, small teams tend to be less confident in their ability to deliver value quickly since there are fewer of them. But given the high salaries they command, companies that feel they need a larger ML team to deliver value quickly may not necessarily be better off, because they need to spend more to grow their team. For smaller companies, a more economical solution is to invest in an ML platform that allows their engineers to work more efficiently and provide value as quickly as larger teams do, but with smaller overheads.

*Question allowed more than one answer and as a result, percentages will add up to more than 100%

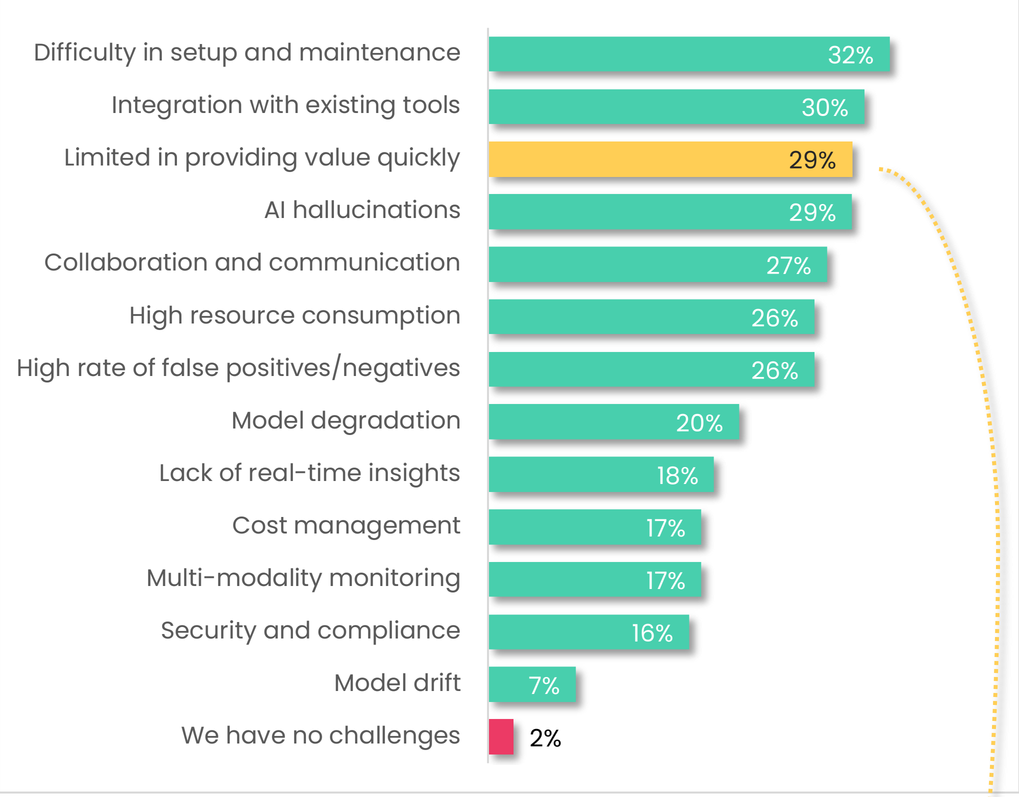

Figure 5: Top Challenges with Current ML Monitoring System

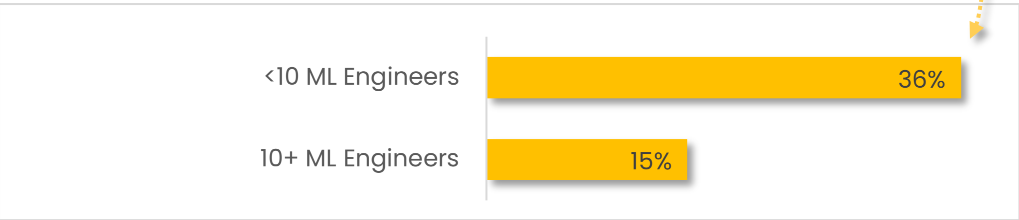

Figure 6: “Limited in providing value quickly”, by Number of ML Engineers

Time Spent on Building Production Monitoring Tools and Dashboards

When asked how much time they spent on building production monitoring tools and dashboards for their last project, 96% of respondents reported they spent at least one month, with the majority spending an average of four months just on building the tools.

If you consider that out of a team of, say, five ML engineers at an average annual salary of $160k (many ML engineers command even higher salaries), four of whom spend four months coding the monitoring without knowing if their efforts will succeed – that's an investment by the company of $212k for the four months (around $53k per ML engineer), which is a lot of money to spend on something whose value is not guaranteed. By comparison, an ML platform solution costs around $60k-$70k annually, making it an extremely cost-effective alternative for companies that use more than one ML engineer to build their production monitoring tools and dashboards.

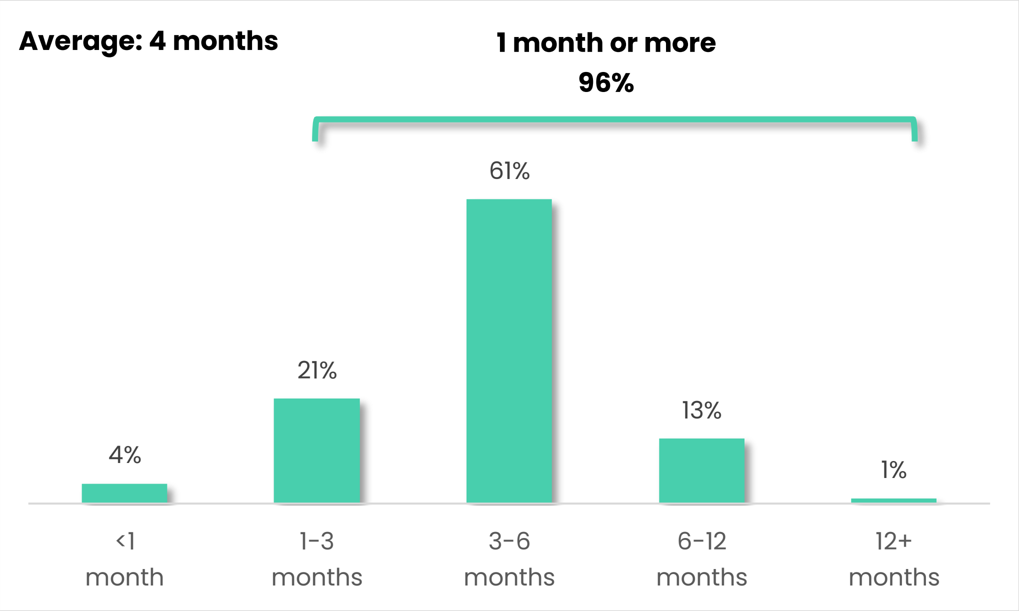

Figure 7: Time Spent on Building Production Monitoring Tools and Dashboards

Frequency of Issues Related to Production Models

When asked how often they encounter issues related to production models, a whopping 93% of the respondents reported they encounter issues on a daily/weekly basis: 5% encounter them 2-3 times a day, 14% once a day, 14% 4-6 times a week, 43% 2-3 times a week, and 17% once a week.

5% of respondents experience issues less than once a week but more than once a month, and 2% experience issues less than once a month.

Given that it's typically difficult to identify problems related to monitoring and observability quickly, it can take time to discover that ML models are spewing out bad predictions, which can be exacerbated by ongoing Issues related to production models.

This finding is therefore significant from a financial perspective because when issues occur frequently, they can be detrimental to the business running smoothly, and this can manifest in a poor customer experience that can also impact loyalty and revenue. Minimizing the frequency of issues is therefore important from a business perspective.

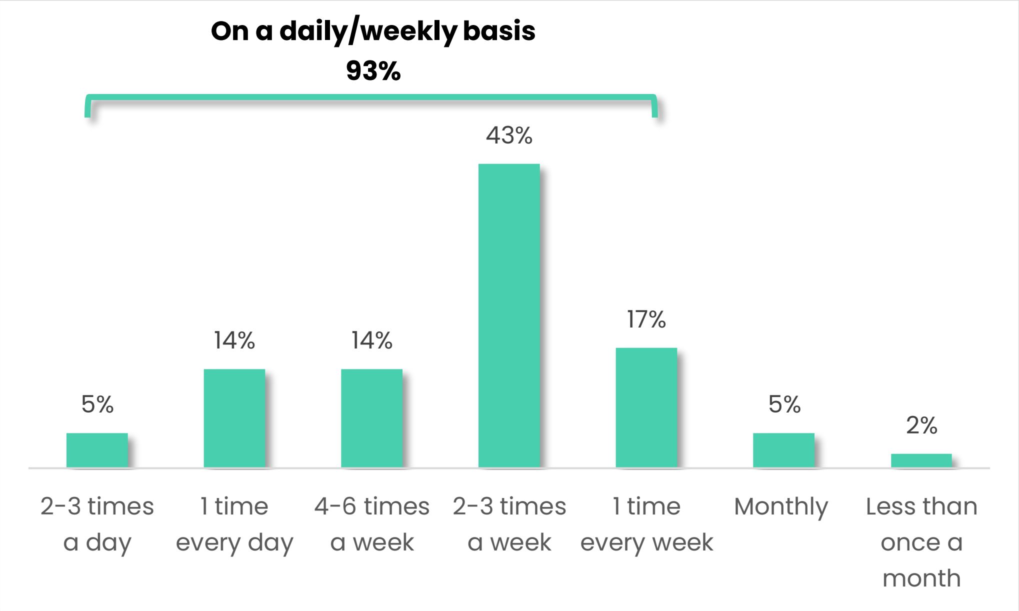

Figure 8: Frequency of Encountering Issues Related to Production Models

AI Bias in Current Projects – Challenges and Importance

Observability isn't just a tool for monitoring ML production issues, it also helps companies understand the effectiveness and impact of ML models from a use-case perspective.

As companies increasingly transition to AI models as a key business component, people and regulators around the world are weary of AI models becoming self-efficient in making decisions, because by leveraging all possible data, including data that's biased or inaccurate, its output could also be biased or even discriminatory. **Observability is therefore crucial for AI product developers to ensure that ML models meet regulatory standards by being fair, accurate and ethical.**

The main challenges respondents encounter in relation to AI bias are difficulty in identifying biased data (19%), inadequate tools for monitoring bias (18%), and lack of understanding the bias implications (12%), seen in Figure 9.

83% of the respondents say that monitoring AI bias in projects is crucial on some level: 27% say it is very crucial, 56% say it is somewhat crucial, and 17% say it isn't crucial at all (Figure 10).

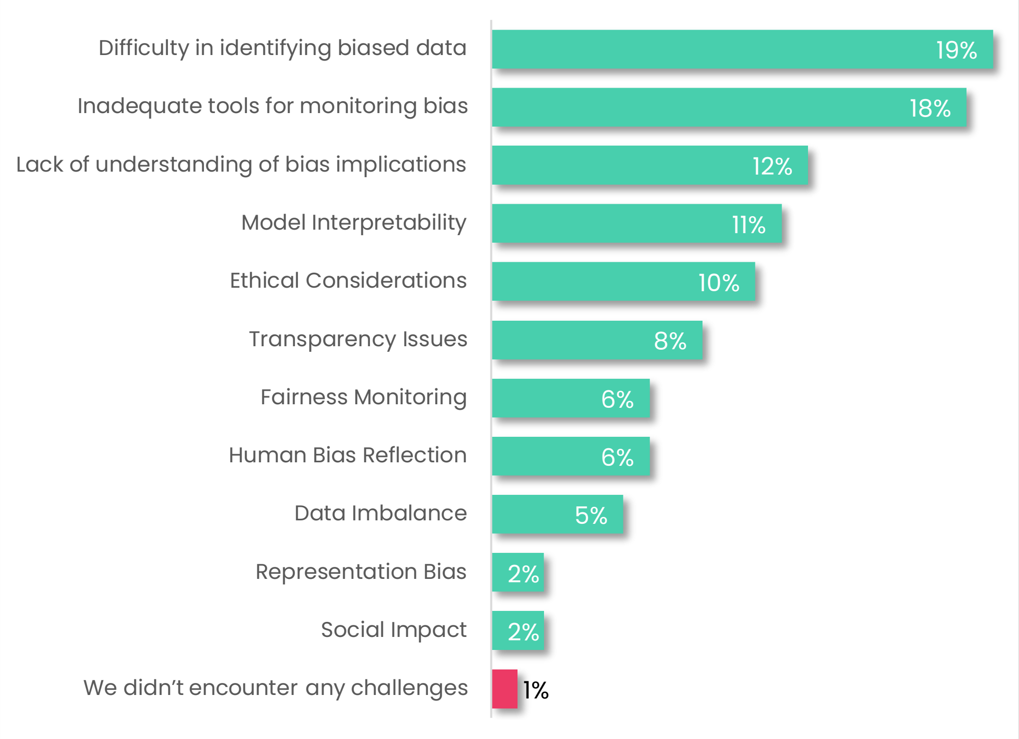

Figure 9: Main Challenges Encountered in Relation to AI Bias

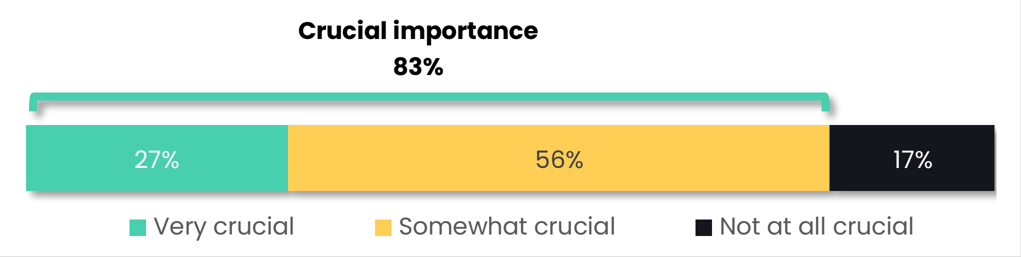

Figure 10: Importance of Monitoring of AI Bias in Projects

Percentage of ML Models Showing Signs of Hallucination

Most of the respondents (89%) say their ML models show signs of hallucinations (when AI generates false information, or "hallucinative content" in real time) to some degree: for 32% of respondents, 1%-25% of their ML models show signs of hallucination, for 44%, 26%-50% show signs of hallucination, and for 13%, 51%-75% show signs of hallucination. 11% say that none of their models are showing signs of hallucination, which is impossible, and likely means they're not working with generative AI.

There is no significant difference in the volume of ML models showing signs of hallucination between companies with <10 engineers and companies with 10+ ML engineers, confirming that **hallucinations occur regardless of team size in companies that use LLMs and generative AI models**.

The severity of hallucinations can range from factual errors to content that's blatantly biased and even dangerous, so hallucinations in ML models must be dealt with as quickly as possible to prevent reputational and even financial damage. One way to do this effectively is with a platform that minimizes hallucinations by creating a deterministic layer of a company's ground truth. This helps the model to understand whether the AI-generated content reflects the ground truth or not, and prevents it from presenting incorrect content.

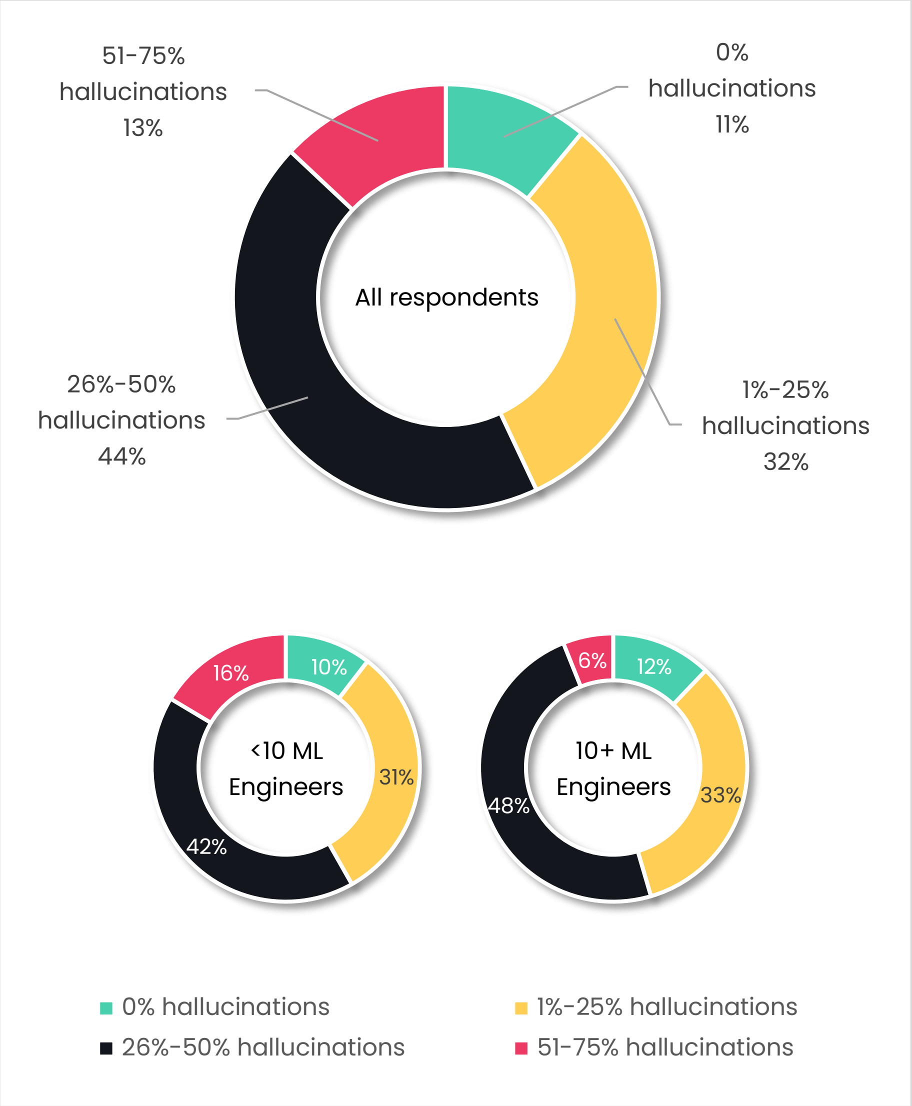

Figure 11: Percentage of ML Models Showing Signs of Hallucination

Demographics

Country, Industry, Company Size, and More

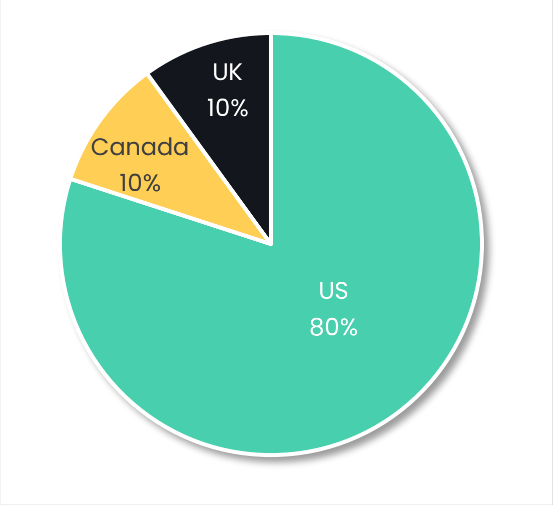

Figure 12: Country

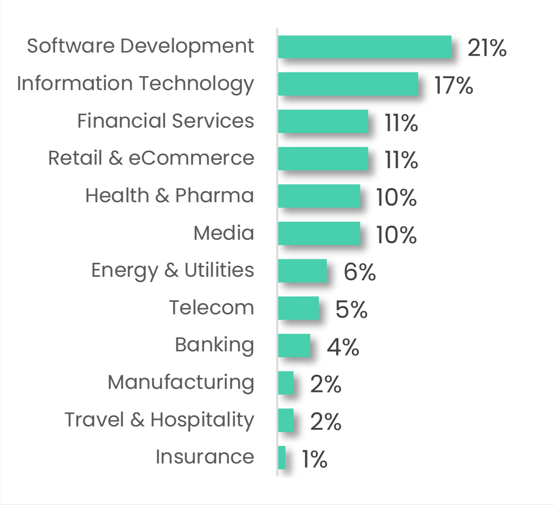

Figure 13: Industry

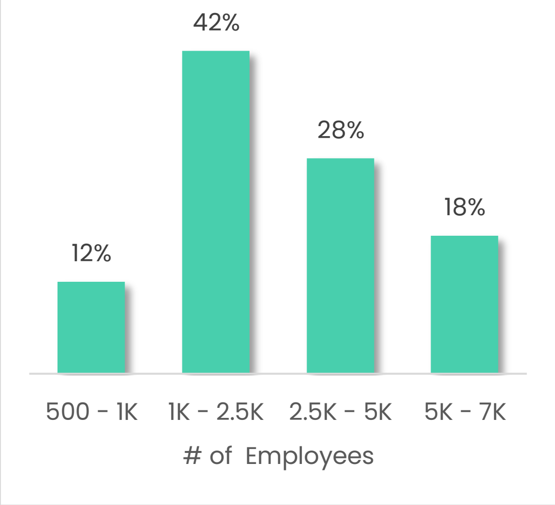

Figure 14: Company Size

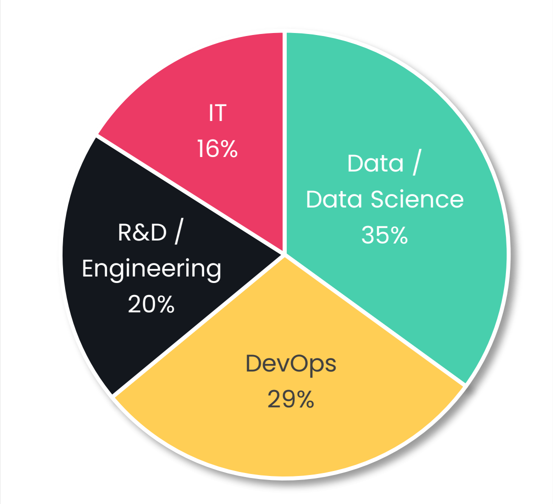

Figure 15: Department

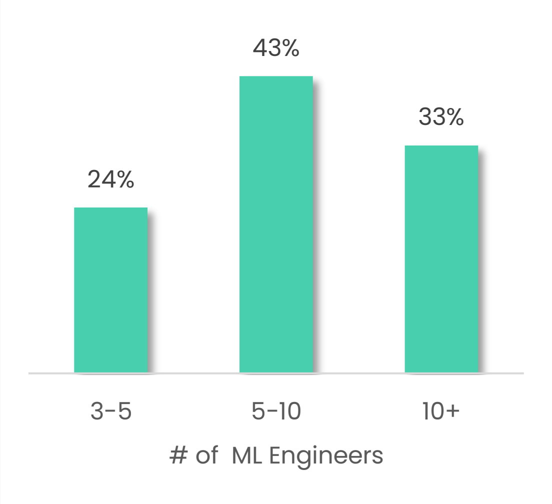

Figure 16: # of ML Engineers

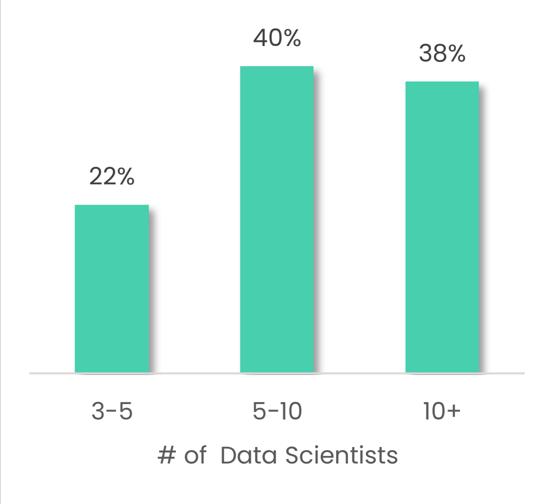

Figure 17: # of Data Scientists

About Aporia

Aporia is the ML Observability platform, trusted by Fortune 500 companies and industry leaders to monitor, manage, and improve ML models in production. Data science teams use Aporia to maximize model performance by detecting data drift, model degradation, bias, and data integrity issues.

With a centralized view of all production models, users gain deep model visibility and can easily track key metrics and performance. Additionally, Aporia's Production IR is an innovative feature that facilitates collaborative root cause analysis of production issues, enabling teams to analyze and optimize model performance swiftly. As a one-stop solution for ML observability, Aporia is pivotal for businesses seeking to harness the true potential of machine learning models.

Try us out (Free trial)

For more information, please visit us:

Email: office@aporia.com

Thank you!

aporia