Machine learning at scale

High-performance, low-cost machine learning to accelerate generative AI development

INTRODUCTION

Solve for machine learning scalability

Machine learning (ML) has emerged as a core technology ingredient for organizations that are focused on driving innovation. Today, more than 100,000 organizations leverage artificial intelligence (AI) solutions and services from Amazon Web Services (AWS) to improve business results. These businesses span virtually every industry, including financial services, healthcare, media, professional sports, retail, and the industrial sector.

The rapid emergence of generative AI is the most visible example of the impact ML innovations are having on the above industries. Generative AI applications have captured widespread attention because they can help reinvent customer experiences, create applications never seen before, and help users reach unprecedented levels of productivity. According to Goldman Sachs, generative AI could drive a 7 percent increase in global GDP over 10 years.[1]

Goldman Sachs also forecasted that AI investment could reach $200 billion by 2025 with the enormous economic potential from generative AI.[2] Like most AI, generative AI is powered by very large ML models that are pretrained on vast amounts of data. These models are commonly referred to as foundation models (FMs). Amidst this growth, obstacles to widespread ML adoption remain. Many organizations, enticed by ML's potential benefits, have grown frustrated by slow progress and a lack of return on their ML investments. For these organizations to reach their goals, they must find ways to move these large models into production faster and at a lower cost.

In this eBook, we will explore the major barriers to ML scalability and success. Then, we will demonstrate how AWS solutions and services can help virtually any organization overcome those challenges and leverage generative AI to drive meaningful innovation.

Top 5 barriers to achieving machine learning results at scale:

1 Resource constraints

2 Disconnected tools

3 Lack of responsible use of ML strategy

4 Expensive infrastructure

5 Lack of repeatable and reproducible ML workflows

[1] "Generative AI could raise global GDP by 7%," Goldman Sachs, April 2023

[2] "AI investment forecast to approach $200 billion globally by 2025," Goldman Sachs, August 2023

Table of contents

Examine the barriers to machine learning success

For many organizations, ML has proven difficult to scale, leading to a lack of progress and frustration with the technology.

With the right services, solutions, tools, and processes, any organization can achieve success with ML and scale it across the business. But determining what those solutions are—and how best to implement them—starts with examining and understanding the barriers that must be overcome.

In that spirit, let's take a look at the five greatest challenges to driving widespread adoption of and business results with ML:

1. **Resource constraints:** Today, many developers find it difficult to get started with ML. These developers want to get models up and running and integrate them into solutions to solve their business problems. However, preparing data and building, training, and deploying a model can take months for experienced ML practitioners and even longer for ML developers new to the technology.

   The size of FMs is very different from traditional ML models (the largest FMs are more than 500 billion parameters—a 1,600 times increase in size in just a few years), making them especially hard to train. Training FMs requires a high performance computing (HPC) cluster with thousands of Graphics Processing Units (GPUs) or **AWS Trainium** chips, along with software to efficiently utilize the cluster.

2. **Disconnected tools:** Due to the relative newness and rapidly changing nature of ML, most organizations have yet to codify standard processes for ML development. Most also lack a set of securely connected ML tools—such as an integrated development environment (IDE) with debuggers, profilers, and solutions for collaboration, workflows, and project management.

   Without access to tools and processes operationalized for ML development, teams are forced to rely on disparate, disconnected components. This makes it difficult to scale ML throughout the organization and involves non-technical teams in the process, as business analysts, developers, and data scientists will struggle to collaborate and deliver results at the speed of modern business.

3. **Lack of responsible use of ML strategy:** To build and maintain trust among customers, partners, and internal stakeholders, organizations need to prioritize responsible AI, including security and privacy. As generative AI continues to grow and evolve, adhering to responsible AI principles will become increasingly critical to building trust and balancing potential innovation with emerging risks.

   Encompassing a core set of concepts—fairness, explainability, robustness, security and privacy, transparency, and governance—responsible AI mitigates risks through the transparent use of data and models. It can enhance model performance, improve data protection, and establish bias detection and mitigation mechanisms in ML systems to improve fairness. As such, responsible AI is an integral, iterative part of the AI lifecycle. It extends from initial design, development, and secure infrastructure to deployment and, ultimately, use, testing, and ongoing auditing for potential bias and accuracy. Responsible AI demands a multidisciplinary effort by technology companies, policymakers, community groups, scientists, and others to tackle new challenges as they arise and work to share best practices and accelerate research.

4. **Expensive infrastructure:** With the increased use of ML comes more requirements for compute, storage, and networking. This can lead to demands on time, cost, and resources—especially for organizations that choose to house and manage their ML infrastructure on premises. As organizations push the boundaries of ML complexity—creating models that use billions of parameters to make thousands of predictions—these problems can escalate exponentially if left unchecked.

   Costs can be controlled by procuring only the amount of infrastructure necessary for an organization's ML workloads. But this can prove difficult as infrastructure requirements evolve throughout the ML lifecycle. For example, moving ML workloads to production can account for up to 90 percent of the overall operational budget.

5. **Lack of repeatable and reproducible ML workflows:** Without repeatable ML workflows, the ML development process can slow, and the quality of models can deteriorate. By adopting **ML operations** (MLOps) processes and standardizing ML development, organizations can move faster and more efficiently toward achieving success with ML at scale. MLOps integrates ML workloads into release management, continuous integration and continuous delivery (CI/CD), and operations. MLOps require the integration of software development, operations, data engineering, and data science. Organizations need purpose-built MLOps tools to automate and standardize processes in order to accelerate model development and time to production.

5 AWS machine learning solutions

1 **Access to hundreds of publicly available FMs**

2 **Integrated ML tools**

3 **Responsible ML**

4 **Flexible infrastructure**

5 **Purpose-built MLOps tools**

Achieve machine learning success with AWS

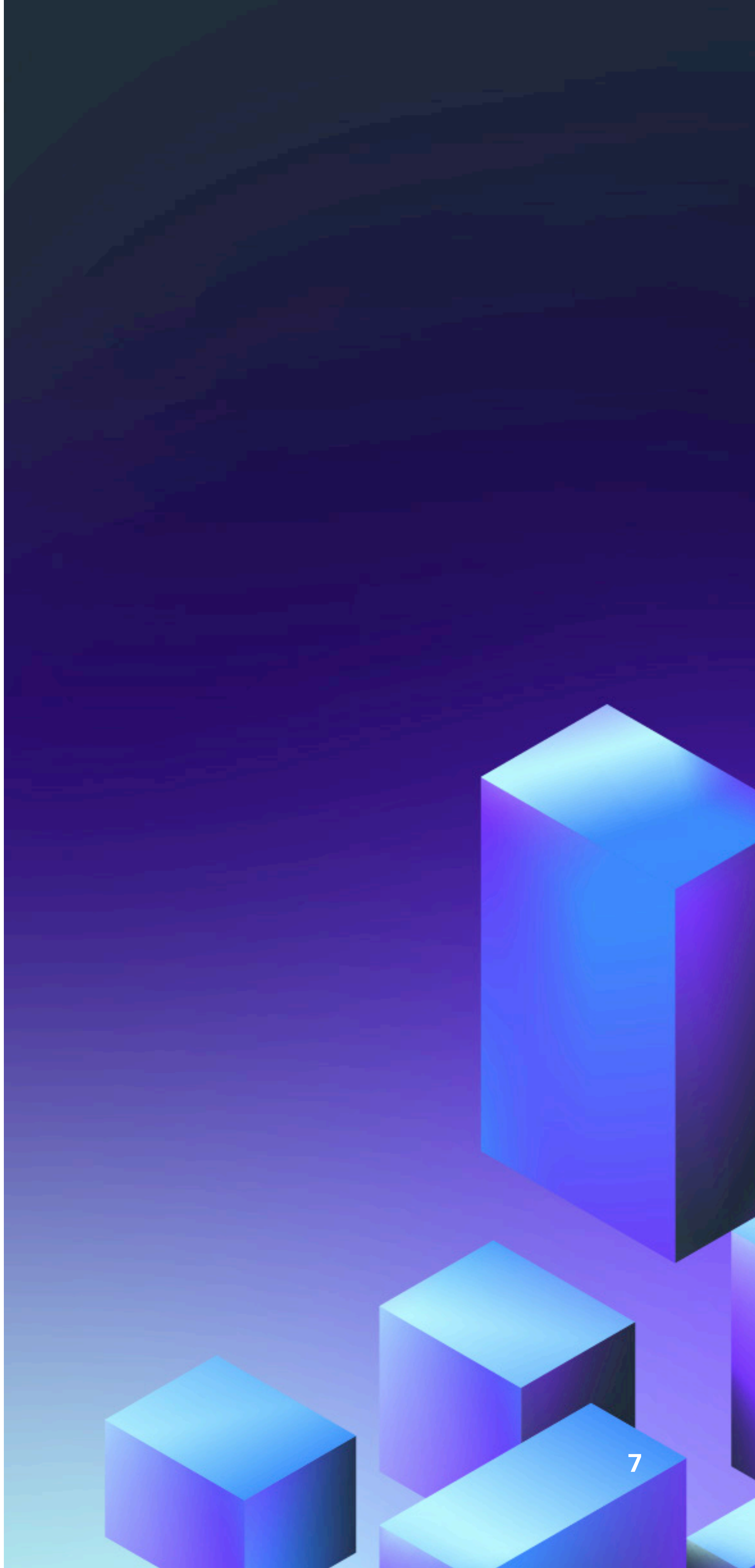
Now you can overcome ML challenges, accelerate your ML journey, and reach your business goals faster using cloud services designed specifically for ML.

Access to hundreds of publicly available FMs

Amazon Bedrock offers third-party models from leading AI startups and Amazon, such as AI21 Labs, Anthropic, Claude 2, Cohere Command, Jurassic 2, **Meta Llama 2**, and Stability AI SDXL 1.0. Amazon Bedrock is a fully managed service that makes these FMs available through an API. With the serverless experience, you can easily experiment with FMs, privately customize FMs with your own data, and seamlessly integrate and deploy them into your applications using AWS tools and capabilities. In addition, **Amazon SageMaker JumpStart** is an ML hub that offers hundreds of publicly available FMs from various model providers, including a growing list of the highest-performing open-source FMs, such as Falcon 40B, Stable Diffusion, OpenLLaMA, and Flan-T5/UL2.

Integrated ML tools

Amazon SageMaker offers the most comprehensive tools for each step of the ML lifecycle, from preparing data at petabyte scale to training and debugging models to managing experiments, tuning, deploying and monitoring models, and managing pipelines—all in **Amazon SageMaker Studio**, an IDE for ML. Using SageMaker tools, customers can easily train, test, troubleshoot, deploy, and manage FMs at scale and boost the productivity of developers while maintaining model performance in production.



Responsible ML

SageMaker provides explainability, security, and governance to support the responsible use of ML. It also detects potential bias during data preparation, after model training, and in deployed models, while feature importance graphs help explain model predictions and reporting for stakeholders. Customized FMs are encrypted using **AWS Key Management Service** (AWS KMS) keys and stored. **Amazon Bedrock** offers comprehensive monitoring and logging capabilities and tools for deep governance and audit requirements.

Flexible infrastructure

SageMaker offers the ideal combination of high performance with the most cost-effective, and energy-efficient infrastructure for generative AI available in a fully managed service. **AWS Trainium** and **AWS Inferentia** are designed by AWS to deliver the best performance at the lowest cost for deep learning training and inference. Customers can use Trainium and Inferentia to train and deploy natural language processing (NLP), computer vision (CV), and recommender models across a broad set of applications, such as speech recognition, recommendation, fraud detection, and image and video classification. With **Amazon Bedrock**, your business can fine-tune and deploy FMs without creating instances, implementing pipelines, or setting up storage.

Purpose-built MLOps tools

SageMaker offers integrated capabilities for **MLOps**, which help your teams improve productivity. Purpose-built tools for MLOps help you automate and standardize processes across the ML lifecycle, so you can easily train, test, troubleshoot, deploy, and govern ML models at scale to produce models faster while maintaining performance in production.

Simplify machine learning at scale with Amazon SageMaker

To maintain focus on your core business objectives, avoid the struggle of building your own ML solution. Instead, offload the heavy lifting to **SageMaker**, which provides high-performance, cost-effective, and scalable ML capabilities to implement an ML environment across your entire business. Regardless of your organization's level of ML skills and experience, your teams can use SageMaker to prepare data and build, train, and deploy ML models for virtually any use case. With SageMaker, your organization can access a broad set of purpose-built ML capabilities under one unified visual user interface.

How does Amazon deliver packages so quickly?

Take a **virtual tour** of an Amazon Fulfillment Center to find out. Discover how Amazon uses a "symphony of machine learning" to help fulfill, sort, and deliver packages in record time.

The top 5 benefits you can achieve with Amazon SageMaker:

1. Accelerate generative AI development with access to hundreds of publicly available FMs that can be customized easily
2. Increase team productivity with an integrated set of tools for ML workflows all in one place
3. Optimize performance and cost with purpose-built, energy-efficient ML infrastructure
4. Automate and standardize MLOps practices across your organization to build, train, deploy, and manage models at scale
5. Ensure model governance and compliance with built-in governance tools

LG AI Research

LG AI Research develops FM using Amazon SageMaker

LG AI Research, the AI research hub of South Korean conglomerate LG Group, was founded to promote AI as part of its digital transformation strategy to drive future growth. The research institute developed its FM EXAONE engine within 1 year of using **SageMaker**.

Built on AWS, the FM mimics humans as it thinks, learns, and takes actions on its own through large-scale data training. The multipurpose FM can be employed in various industries to carry out a range of tasks.

Read the case study ›

Accelerate generative AI development with access to hundreds of publicly available FMs

SageMaker JumpStart is an ML hub that offers models, algorithms, and pre-built ML solutions. Developers can access hundreds of publicly available FMs from various model providers, including a growing list of best-performing open-source models.

FMs, such as Falcon 40B, Stable Diffusion, OpenLLaMA, and Flan-T5/UL2, are large-scale ML models that contain billions of parameters and are pretrained on terabytes of text and image data. They inform a wide range of tasks, such as article summarization and text, image, and video generation. Because these FMs are pretrained, they can help reduce training and infrastructure costs while supporting customization for your use case.

FMs available in SageMaker JumpStart:

- Llama 2 from Meta is an auto-regressive language model that uses an optimized transformer architecture. It comes in a range of parameter sizes—7 billion, 13 billion, and 70 billion—and pretrained and fine-tuned variations.
- Falcon 40B Instruct BF 16 is a 40 billion parameter causal decoder-only model built by TII based on Falcon 40B and fine-tuned on a mixture of Baize. It is a publicly available chat/instruct model based on Falcon 40B.
- Stable Diffusion XL 1.0 is the largest open-source image model from Stability AI, designed for creative professionals who want to generate the highest-quality images without sacrificing performance.
- Cohere Command provides businesses and enterprises with the best quality, performance, and accuracy in all generative tasks. With its intuitive SDK, this FM can unlock the full potential of large language models (LLMs) for your applications.
- AI21 Jurassic-2 Ultra is optimized to follow natural language instructions and context, so there is no need to provide it with any examples.

Forethought

Forethought saves over 66% in costs for generative AI models using Amazon SageMaker

Forethought is a leading generative AI suite for customer service. At the core of its suite is the innovative SupportGPT technology, which uses ML to transform the customer support lifecycle—increasing deflection, improving CSAT, and boosting agent productivity. SupportGPT leverages state-of-the-art information retrieval (IR) systems and LLMs to power over 30 million customer interactions annually.

One of the significant challenges it has is the efficient utilization of hardware resources such as GPUs. Because the hyper-personalization of models requires unique models to be trained and deployed, the number of models scales linearly with the number of clients, which can become costly. To tackle this challenge, Forethought uses SageMaker multi-model endpoints (MMEs) to run multiple AI models on a single inference endpoint and scale. SageMaker MMEs enable Forethought to deliver high-performance, scalable, and cost-effective solutions with sub-second latency, addressing multiple customer support scenarios at scale.

Read the success story ›

Improve cost-efficiency with purpose-built integrated machine learning tools

As your use of ML grows, so will your infrastructure requirements. To prevent your costs from becoming prohibitive, you will need tools and processes that allow you to dynamically match your spend to your specific compute, storage, and networking needs throughout the ML lifecycle. You will also need to maximize productivity and efficiency, enabling your developers to avoid wasted time and duplicative efforts and put models into production quickly.

By using services and tools that are purpose-built for ML, you can achieve speed, scale, and cost-efficiency that go far beyond general-purpose and on-premises solutions.

Throughout the ML lifecycle—including labeling, data preparation, feature engineering, training, hosting, monitoring, and workflows—your team can use a single visual inference in **SageMaker Studio**. This provides you with greater control over your infrastructure spend. Furthermore, it can improve your data science team's productivity by up to 10 times and enable them to develop models in weeks instead of months.[3]

AWS customers are achieving massive scale, productivity, and cost-efficiency with purpose-built tools from AWS:

Vanguard	**NerdWallet**	**Mueller Water Products**
AstraZeneca	**Zendesk**	

Learn more about accelerating training and development of ML models ›

AI21labs

AI21 Labs accelerates generative AI model adoption using Amazon SageMaker

AI21 Labs (AI21), a leader in generative AI and LLMs, wanted to empower businesses with state-of-the-art LLMs and AI applications to build generative AI solutions. Initially, AI21 released 2 models: 1 with 7 billion parameters and another with 178 billion parameters. However, the company saw an opportunity to offer customers a midsize model of 17 billion parameters that bridged the gap between the existing sizes. The new pretrained language model would preserve the quality of text generation, making it nearly the same as the largest-size model at a much lower inference cost to AI21 and its customers. To build that model efficiently, AI21 looked to AWS and trained the FM in under 20 days using **SageMaker**.

Read the case study ›

[3] "Lowering Total Cost of Ownership for Machine Learning and Increasing Productivity with Amazon SageMaker," AWS Machine Learning Blog, February 2020

Foster responsible use of machine learning

Responsible use of ML is key to achieving tangible benefits that scale across the business. AWS is committed to developing fair and accurate AI services and helping organizations transform responsible AI from theory into practice with purpose-built tools and guidance.

To use ML in a responsible manner, ML models need to be built with transparency, fairness, and security in mind. **Amazon SageMaker Clarify** provides bias detection across the ML workflow and includes feature importance graphs. These explain model predictions and produce reports to support internal presentations while also identifying issues with models to enable course correction.

To help your organization meet security criteria applicable to ML workloads, SageMaker includes solutions for encryption, private network connectivity, authorization, authentication, monitoring, and auditability.

Achieve responsible and secure ML with Amazon SageMaker Clarify:

- Gain greater visibility into data and models to identify and limit bias
- Detect potential bias throughout the entire workflow
- Explore feature importance graphs to help explain model predictions

TII Technology Innovation Institute

Technology Innovation Institute trains Falcon LLM 40B FM on Amazon SageMaker

The **Technology Innovation Institute** (TII) is a leading global research center dedicated to pushing the frontiers of knowledge. TII's team of scientists, researchers, and engineers focuses on delivering discovery science and transformative technologies. Its work focuses on breakthroughs to future-proof our society. Trained on 1 trillion tokens, **TII Falcon LLM** delivers high performance while remaining incredibly cost-effective. Falcon 40B matches the performance of other high-performing LLMs and is the top-ranked open-source model on the public **Hugging Face Open LLM leaderboard**. It's available as open source in two different sizes—Falcon 40B and Falcon 7B—and was created from scratch using data pre-processing and model training jobs built on **SageMaker**.

Read the success story ›

Scale machine learning across your business with MLOps

MLOps practices help you streamline the ML lifecycle by automating and standardizing ML workflows. With standardized MLOps processes in place, your teams can get models into production faster and collaborate more effectively. Over time, MLOps can help you reach your ultimate goal—scaling ML adoption and using ML to improve results across the entire organization.

SageMaker delivers the capabilities, automation, standardization, and centralization you need to make MLOps a reality for your organization.

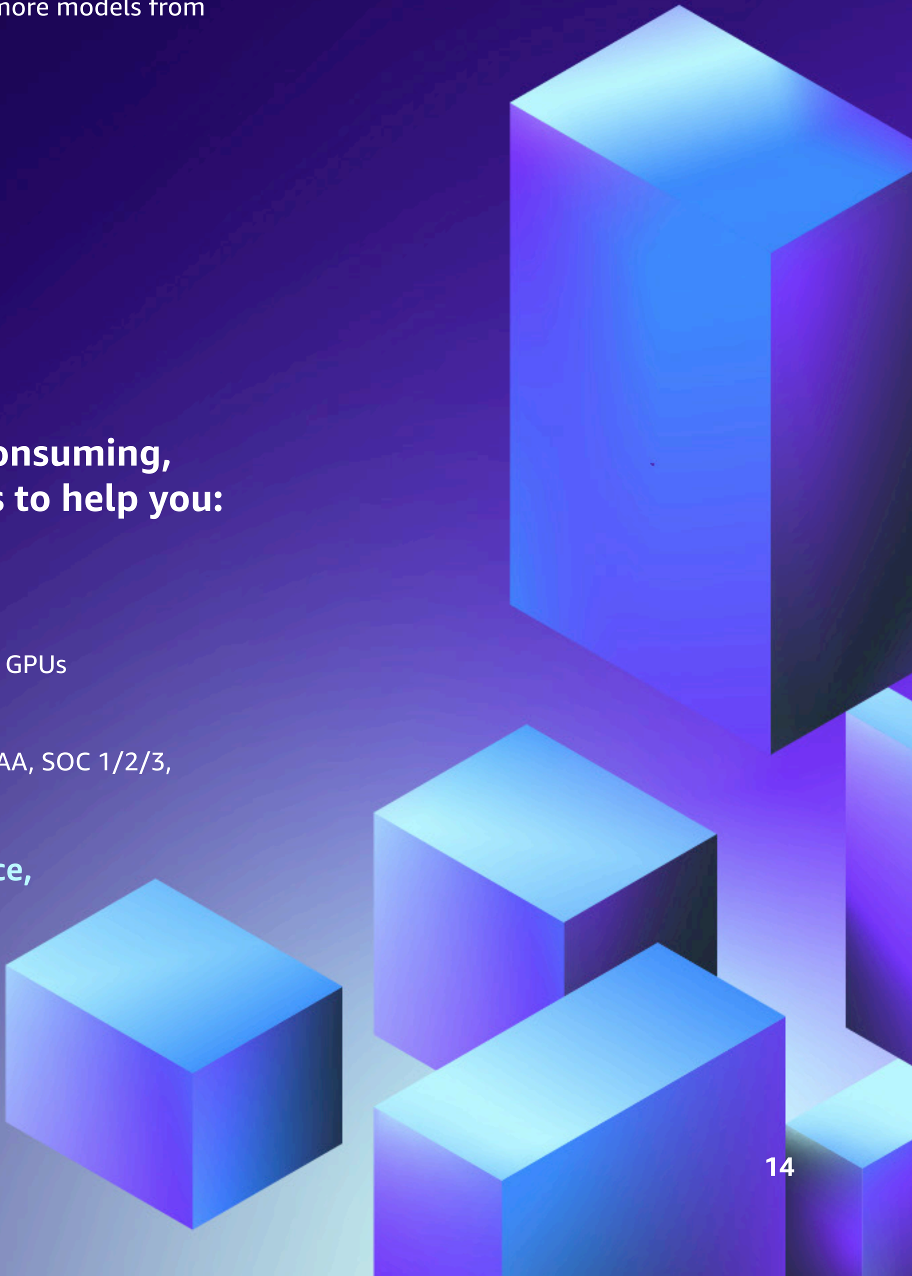
With the purpose-built MLOps tools provided by SageMaker, you can:

- Create repeatable training workflows to accelerate model development
- Catalog ML artifacts centrally for model reproducibility and governance
- Integrate ML workflows with CI/CD pipelines for faster time to production
- Continuously monitor data and models in production to maintain quality

Learn more about Amazon SageMaker for MLOps ›

It’s time to embrace machine learning

By using publicly available FMs, purpose-built development tools, MLOps, fully managed infrastructure, and solutions focused on the responsible use of data and models on AWS, you can propel many more models from concept to production in a repeatable way for less cost.



Amazon Bedrock makes it easier than ever to build and scale generative AI application development with access to FMs through an API.

Discover how Amazon Bedrock accelerates the development of generative AI applications ›

Amazon SageMaker eliminates the need for time-consuming, difficult, and expensive self-managed ML platforms to help you:

- Perform more than one trillion predictions per month
- Cut data labeling costs by 40 percent
- Accelerate model training by up to 50 percent through more efficient use of GPUs
- Reduce overhead latency to less than 10 milliseconds inference
- Support security standards and 22 compliance programs (including PCI, HIPAA, SOC 1/2/3, FedRAMP, and ISO)

Learn more about Amazon SageMaker for high-performance, low-cost ML development at scale ›