Democratized, operationalized, trusted: the 3 keys to successful AI outcomes

Unlocking the potential of machine learning in the generative AI era

Table of contents

PART I - THE MARKET

The rapid development of generative AI is fueling ever-growing AI investments and innovations

INTRODUCTION

Why artificial intelligence matters more than ever

Artificial intelligence (AI) has existed for decades, but with the ready availability of scalable compute capacity, a massive proliferation of data, and the rapid advancement of machine learning (ML) technologies, organizations across industries are transforming their businesses.

With consumer-facing applications like ChatGPT delivering impressive results and demonstrating the power and sophistication of today's ML models in ways that are exciting and plainly evident, **generative AI** has captured widespread attention and imagination. We are truly at an inflection point, and we believe most customer experiences and applications will be reinvented with generative AI.

Although specific estimates vary on the growth of generative AI, there is no debate about the impact of this technology on the global economy. The potential effects of these numbers are staggering—both for the organizations that make the investments and for those that do not. According to Goldman Sachs, generative AI could drive a 7 percent (or almost $7 trillion) increase in global GDP and lift productivity growth by 1.5 percentage points over a 10-year period.[1] And Bloomberg expects the spend on generative AI to reach over $109 billion by 2030, a compound annual growth rate (CAGR) of 36.5 percent from 2024 to 2030.[2]

Whether you are already expanding your AI initiatives with generative tools or just getting started with AI and ML, a clear reference guide can help you develop your strategy and help ensure successful outcomes. This eBook outlines the three strategic pillars to success and provides practical recommendations that you can apply to your organization.

[1] "Generative AI could raise global GDP by 7%," Goldman Sachs, April 2023
[2] "Generative AI Market Size, Share & Trends Analysis Report By Component (Software, Services), By Technology, By End-use, By Application, By Model, By Region, And Segment Forecasts, 2024 - 2030," Research and Markets, February 2024

OVERVIEW

What is generative AI?

Generative AI is a type of AI that can create new content and ideas, including conversations, stories, images, videos, and music. Like all AI, generative AI is powered by ML models—very large models that are pretrained on vast amounts of data and commonly referred to as foundation models (FMs). Recent advancements in ML have led to the rise of models that contain billions of parameters or variables.

To give a sense of the change in scale, the largest pretrained model in 2019 was 330 million parameters. Now, the largest models are more than 500 billion parameters—a 1,600 times increase in size in just a few years. Today's FMs can perform a wide range of tasks that span multiple domains, like writing blog posts, generating images, solving math problems, engaging in dialogue, and answering questions based on a document. The size and general-purpose nature of FMs make them different from traditional ML models, which typically perform specific tasks like analyzing text for sentiment, classifying images, and forecasting trends.

FMs can perform so many more tasks because they contain such a large number of parameters that make them capable of learning complex concepts. And through their pretraining exposure to internet-scale data in all its various forms and myriad of patterns, FMs learn to apply their knowledge within a wide range of contexts. While the capabilities and resulting possibilities of pretrained FMs are amazing, organizations place an even higher value on the fact that these generally capable models can also be customized to perform domain-specific functions that are differentiating to their businesses yet use only a small fraction of the data and compute required to train a model from scratch.

Customized FMs can create a unique customer experience, embodying the company's voice, style, and services across a wide variety of consumer industries, like banking, travel, and healthcare. For instance, a financial firm that needs to automatically generate a daily activity report for internal circulation using all the relevant transactions can customize the model with proprietary data, which will include past reports, so that the FM learns how these reports should read and what data was used to generate them.

PART II – THE KEYS TO SUCCESS

3 keys that can help organizations bring AI to more parts of their business

KEY #1

Democratize

To keep pace with the rapid innovations in AI, it's important to make generative AI applications easy and practical for all. This is known as *democratization*, and for many Amazon Web Services (AWS) customers, it's the first step to unlocking the potential of these powerful technologies. Organizations look to AWS for straightforward ways of finding and accessing high-performing FMs that deliver outstanding results and are best suited for different business tasks. The second step focuses on ensuring the seamless integration of FMs into applications without having to manage huge clusters of infrastructure or incur significant costs. The third and final step simplifies the process of building differentiated apps on the base FM using each organization's data (a little data or a lot) while also keeping that data secure.

Democratized AI automates more of the applications we use to live, work, and play, allowing more time to focus on high-value activities.

Customers like Intuit, Thomson Reuters, AstraZeneca, Ferrari, Bundesliga, 3M, and BMW, as well as thousands of startups and government agencies around the world, are transforming themselves, their industries, and their missions with AI.

A clear path to democratized artificial intelligence

AWS has played a key role in democratizing AI and making it accessible to anyone who wants to use it, including more than 100,000 customers of all sizes and industries. We have the broadest and deepest portfolio of AI and ML services. AWS has invested and innovated to offer the most performant, scalable infrastructure for cost-effective ML training and inference; developed **Amazon SageMaker**, the easiest way for all developers to build, train, and deploy models; and launched a wide range of **AI services** that allow customers to add AI capabilities like image recognition, forecasting, or intelligent search to applications with a simple API call.

We take the same democratizing approach to generative AI: We work to take these technologies out of the realm of research and experiments and extend their availability far beyond a handful of startups and large, well-funded tech companies.

Amazon Bedrock is the easiest way to build and scale generative AI-based applications using FMs, democratizing access for all builders. Bedrock provides access to a range of powerful FMs for text and images—including **Amazon Titan** FMs—through a scalable, reliable, and secure AWS managed service. With the Bedrock serverless experience, AWS customers can easily find the right model for what they're trying to get done, get started quickly, privately customize FMs with their own data, and easily integrate and deploy them into their applications using the AWS tools and capabilities they are familiar with, without having to manage any infrastructure.

Bedrock customers can choose from some of the most cutting-edge FMs available today. This includes the **Jurassic-2** family of multilingual large language models (LLMs) from AI21 Labs, which follows natural language instructions to generate text in Spanish, French, German, Portuguese, Italian, and Dutch. Anthropic **Claude**, is a family of FMs that can be used in a variety of applications for conversations, image analysis, document processing, and workflow automation based on research into training honest and responsible AI systems. Cohere's **Command** text generation model is trained to follow user commands and be useful instantly in practical business applications such as summarization, copywriting, dialogue, extraction, and question answering. Cohere's text understanding model, Embed, can be used for search, clustering, or classification tasks across over 100 languages, allowing organizations to easily search by meaning or categorize text. **Meta Llama 2** pretrains and fine-tunes LLMs for natural language tasks like question and answering and reading comprehension. Bedrock also makes it easy to access Stability AI's suite of text-to-image FMs, including **Stable Diffusion** (the most popular of its kind), which is capable of generating unique, realistic, high-quality images, art, logos, and designs.

One of the most important capabilities of Bedrock is how easy it is to customize a model. Customers simply point Bedrock at a few labeled examples in **Amazon Simple Storage Service** (Amazon S3), and the service can fine-tune the model for a particular task without having to annotate large volumes of data (as few as 20 examples is enough).

While Bedrock democratizes access to FMs, generative AI can also be used to democratize software development. Case in point: **Amazon CodeWhisperer**, an AI coding companion that uses an FM under the hood to radically improve developer productivity by generating code suggestions in real time based on developers' natural language comments and prior code in their integrated development environment (IDE). During the preview of CodeWhisperer (now part of **Amazon Q Developer**), we ran a productivity challenge, and participants using CodeWhisperer completed tasks 57 percent faster on average and were 27 percent more likely to complete them successfully than those who didn't use CodeWhisperer. This is a giant leap forward in developer productivity, and we believe it's only the beginning. Plus, CodeWhisperer is free to individual developers, so getting started is easy.

KEY #2

Operationalize

With AI adoption growing rapidly, business and technical teams are challenged to build more. In this rush to leverage the technologies, organizations rarely stop to set standard tools and processes for ML development. As a result, different teams with different skills and requirements often use completely different and disconnected tools, making collaboration impractical, if not impossible.

For example, an R&D team might be working on a computer vision (CV) application with state-of-the-art algorithms and frameworks, while sales and marketing teams are building a linear regression model to forecast customer demand on a locally stored spreadsheet. Or developers could be coding a mobile purchasing app and want to add a recommendation engine to make the customer experience more personal.

Relatively few organizations utilize operational ML tools and practices—such as infrastructure, IDEs, debuggers, profilers, collaboration tools, workflows, and project management tools—that can be connected securely. This reality complicates management across teams of business analysts, developers, and data scientists and in coordination with existing software tools and processes. In these common scenarios, scaling up or down becomes exceedingly difficult.

The good news is that there's a proven way to minimize the risks and complications of ML while providing straightforward, repeatable practices for teams—by operationalizing ML.

Operationalization of ML provides the tools, infrastructure, and operations support to scale. Operationalizing ML starts with the data acquisition and modeling activities of the data science team being informed by a clear understanding of the business objectives for the ML application and of all governance and compliance issues. MLOps ensures that the data science, production, and operations teams work seamlessly together across a series of ML workflows that are as automated as possible. Human intervention is incorporated as needed, ensuring smooth deployments, data monitoring, and model performance tracking.

How AWS helps customers operationalize machine learning

Amazon SageMaker, which we already know is a powerful service to help democratize ML, is equally suited for operationalization. It automates and standardizes every step of the MLOps workflow to help projects scale without limits. Thanks to SageMaker, AWS customers are running millions of models with billions of parameters and generating hundreds of billions of predictions.

SageMaker also offers an end-to-end ML service for data labeling, data preparation, feature engineering, training, hosting, monitoring, and workflows that can be accessed using a single visual interface in **Amazon SageMaker Studio**. In comparison to self-managed ML environments, the productivity of data science teams can improve by up to 10 times, and model development time is reduced from months to weeks. And all SageMaker capabilities are offered on fully managed, low-cost, high-performance infrastructure in the cloud.

AWS customers are realizing massive scale (and savings) with SageMaker tools:

- **Vanguard** has fully automated the setup of its ML environments and now deploys ML models 20 times faster
- **AstraZeneca** can deploy new ML environments in five minutes versus one month to generate insights that improve R&D and accelerate the commercialization of new therapeutics
- **NerdWallet** reduced training costs by close to 75 percent, even while increasing the number of models trained
- **Zendesk** reduced ML inference costs by 90 percent by deploying thousands of models per endpoint using SageMaker multi-model endpoints
- **Mueller Water Products** cut the number of false alerts in half and maximized the potential to identify true leak events

Operationalization also means that we need to deliver the breadth and depth of AI use cases, including intelligent contact centers, intelligent document processing (IDP), content moderation, personalization, intelligent search, fraud prevention, identity verification, predictive maintenance, AI for DevOps, health AI, and ML-powered business intelligence (BI). AWS offers services for all these use cases and more.

MUELLER

"Being on AWS enables a faster development process for us."

Kenji Takeuchi, SVP of Technology Solutions, Mueller Water Products, 2021

Read more ›

KEY #3

Build trust

It's essential for organizations to build trust with their customers, partners, and internal stakeholders regarding their use of generative AI. To create and maintain this trust, organizations need to make investments and considerations across responsible AI, security, and privacy.

Responsible AI

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 be used to enhance model performance, improve data protection, and establish bias detection and mitigation mechanisms in ML systems to improve fairness.

"94% (of companies) are struggling to operationalize across all key elements of responsible AI."

Accenture, 2022

Responsible AI is an integral part of the complete AI lifecycle, extending from initial design, development, and secure infrastructure to deployment and, ultimately, ongoing use. It is an iterative process that requires ongoing testing and auditing for potential bias and accuracy. While most companies have begun their journey to responsible AI, the majority (94 percent) are struggling to operationalize across all key elements of responsible AI.[3]

So, how do organizations transform responsible AI from theory into practice? They begin by educating the next generation of ML leaders to elevate fairness and mitigate bias by bringing more diverse perspectives to the table, providing resources to promote education and training, and ensuring data protection and privacy. Responsible AI also requires a multidisciplinary effort by technology companies, policymakers, community groups, scientists, and more to tackle new challenges as they arise and work to share best practices and accelerate research.

[3] "From AI compliance to competitive advantage," Accenture, June 2022

Build more responsible, secure, and private AI with AWS

Security and privacy

Data security and privacy are also critical to scaling generative AI responsibly. When it comes time to customize and fine-tune a model, organizations need to know where and how their data is being used. They need to be confident their protected data or intellectual property (IP) is not being used to train a public model and that customer data remains private. Organizations need security, scalability, and privacy to be baked in from the start to be viable for their business applications.

Gain purpose-built protections with Amazon SageMaker

As organizations scale their use of AI technologies, they can leverage AWS resources to help implement responsible AI across the entire ML lifecycle.

Organizations can mitigate bias and improve explainability with AWS purpose-built services. **Amazon SageMaker Clarify** helps mitigate bias across the ML lifecycle by detecting potential bias during data preparation, after model training, and in the deployed model by examining specific attributes.

Monitoring is also important to maintaining high-quality ML models and ensuring accurate predictions. **Amazon SageMaker Model Monitor** automatically detects and issues alerts when models deployed in production generate inaccurate predictions.

To improve governance, SageMaker provides purpose-built **tools**—including SageMaker Role Manager, SageMaker Model Cards, and SageMaker Model Dashboard—that deliver tighter control and deeper visibility over ML models. AWS customers can set up users with least-privilege permissions in minutes; easily capture, retrieve, and share essential model information; and stay informed on model behavior, like bias, all in one place.

Enhance security and privacy with Amazon Bedrock

When customers use Amazon Bedrock to customize a model, Bedrock can fine-tune the model for a particular task without having to annotate large volumes of data. Then, Bedrock makes a separate copy of the base FM that is accessible only to the customer and trains this private copy of the model.

AWS customers can also configure their Amazon Virtual Private Cloud settings to access Bedrock APIs and provide model fine-tuning data in a secure manner. Customer data is always encrypted in transit (TLS1.2) and at rest through service managed keys.

AWS keeps responsible AI in mind at each stage of its comprehensive FM development process. Throughout FM design, development, deployment, and operation, we consider:

Accuracy

Fairness

IP and copyrights

Appropriate usage

Toxicity

Privacy

Resources

Hear from an Amazon Scholar on the emerging challenges and solutions to build generative AI responsibly ›

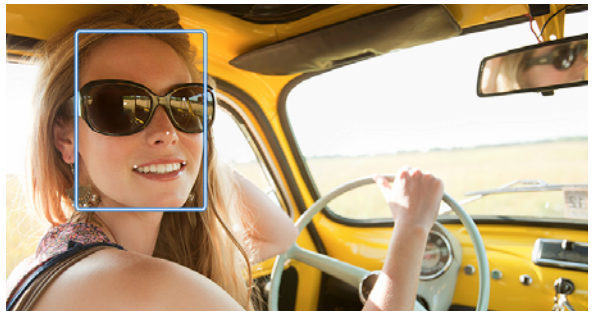
Learn more about the new commitments from the White House, technology organizations, and the AI community to advance responsible and secure use of AI ›

To address these issues, we build solutions into our processes for acquiring training data into the FMs themselves, and into the technology that we use to pre-process user prompts and post-process outputs. For all our FMs, we actively invest to improve our features and to learn from customers as they experiment with new use cases.

Check out three essential resources to enable more responsible AI:

1. **The Responsible Use of Machine Learning Guide** provides considerations and recommendations for the responsible use and development of ML systems across three major phases of their lifecycles: 1) design and development, 2) deployment, and 3) ongoing use. **Learn more ›**

2. **Continuous education** on the latest developments in ML is an important part of responsible use. AWS offers the latest in ML education across the learning journey through programs like the **AWS Machine Learning University (MLU) Bias and Fairness Course**, **Training and Certification program**, and **AI and ML Scholarship program**.



3. **AWS AI Service Cards** provide transparency and document the intended use cases and fairness considerations for our AWS AI services. Explore the AI Service Cards: **Amazon Rekognition face matching**, **Amazon Textract AnalyzeID**, and **Amazon Transcribe – Batch (English-US)**.

AWS is committed to the continued development of AI and ML in a responsible way.

Learn more ›

Successful outcomes start here

More than 100,000 customers have chosen AWS for AI to create new customer experiences, optimize their businesses, augment their employees' ingenuity, help improve the quality of their products, and so much more. That's because AWS supports you no matter where you are on your ML journey—with the solutions you need to scale without limits.

Connect with the experts from AWS

- **AWS Professional Services** is a global team of experts that can help you realize your desired business outcomes for the AWS Cloud
- The **AWS Generative AI Innovation Center** connects you with AWS AI and ML experts to help you envision, design, and launch new generative AI products, services, and processes

Collaborate with an official AWS Partner

- **AWS Partners** are uniquely positioned to help AWS customers accelerate the journey to the AWS Cloud

Construct it yourself with proven solutions

- **AI Use Case Explorer** helps you discover the top AI use cases, customer stories, and implementation paths based on your business objectives
- **AWS Solutions Library** offers solutions built by AWS and AWS Partners for a broad range of use cases
- **AWS AI services** let you easily add intelligence to applications—no ML skills required
- **Amazon SageMaker** empowers users to build, train, and deploy ML models for any use case with fully managed infrastructure, tools, and workflows
- **Amazon Bedrock** makes FMs from Amazon and leading AI companies such as AI21 Labs, Anthropic, Cohere, Stability AI, and Meta accessible via an API

From the world's largest enterprises to emerging startups, more AI is built on AWS than anywhere else.

Learn more about how your business can deliver successful AI and ML outcomes ›