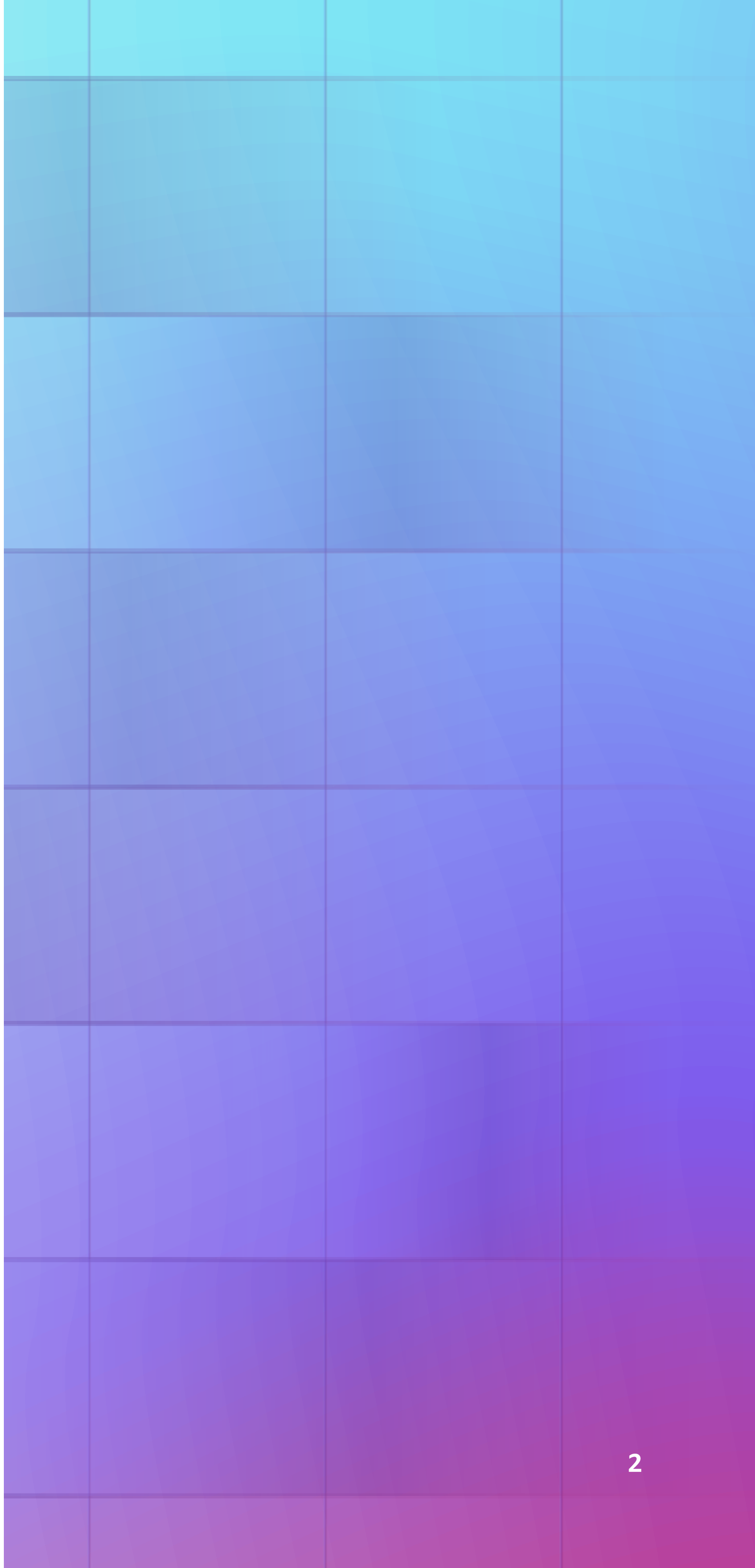
aws startups

Generative AI for every startup

Easily build and scale with generative AI on AWS

Table of contents



INTRODUCTION

The power and promise of generative AI for startups

The seeds of a machine learning (ML) paradigm shift have existed for decades, but with the availability of scalable compute capacity, the proliferation of data, and the rapid advancement of ML technologies, customers across industries are transforming their businesses. Generative AI tools like OpenAI's ChatGPT and Google's Bard have captured widespread attention, and the appetite for investment keeps growing.

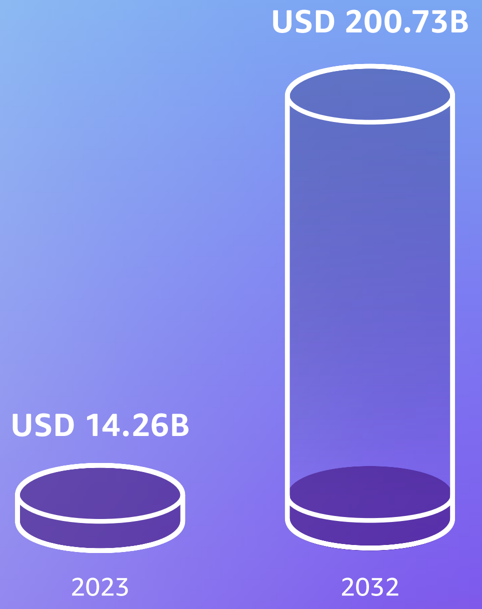
According to a March 2023 PitchBook report, many venture capitalists (VCs) have cooled on investment in other tech innovations such as cryptocurrency and blockchain. VC funding in general is down 53 percent year over year in Q1 2023–but generative AI is a bright spot. In 2022 alone, investment in the sector reached US$2.1 billion–a 425 percent increase compared to 2020.[1] Through 2023, AI21 Labs, a large language model startup specializing in text-based generation, raised a $155 million Series C. Databricks, the developer of a data analytics platform with its own Lakehouse generative AI program, raised more than $500 million from investors including Nvidia and T. Rowe Price. And Corti, the creator of a generative AI copilot for use in healthcare, raised a $60 million Series B in the second half of 2023.[2]

Those VC numbers aren't surprising, considering the proven benefits of generative AI. As a founder, you can use the technology to go to market faster, boost innovation, personalize customer experiences, and optimize costs. It's time to join the over 60 percent of business owners who already believe that artificial intelligence (AI) will increase their productivity.[3]

This eBook is a guide for startup leaders interested in integrating generative AI solutions into their businesses. It includes examples of startups that have leveraged generative AI, and it illustrates why startups at any stage are choosing Amazon Web Services (AWS) for their generative AI journeys. But first, let's consider the fundamentals of the technology.

Global generative AI market

Market forecast to grow at a CAGR of 34.2%[4]



[1] "Vertical Snapshot: Generative AI," PitchBook, March 2023
[2] "Robbins, J., "Generating less momentum? Generative AI deal count dips in Q3," PitchBook, October 2023
[3] Haan, K., "24 Top AI Statistics and Trends In 2023," Forbes, April 2023
[4] "Generative AI Market," Polaris Market Research, January 2023

The strategic imperative

Startups across industries and around the world are looking to leverage generative AI to innovate faster, boost employee productivity, ignite creativity, and optimize business processes.

For most startups, however, the path to achieving these benefits remains unclear.

Most founders acknowledge the need to invest in generative AI—and to do it quickly, lest their competitors gain the advantage. However, few have succeeded in developing strategies for how they will adopt the technology, where they will put it to use, or how they will achieve and measure their results.

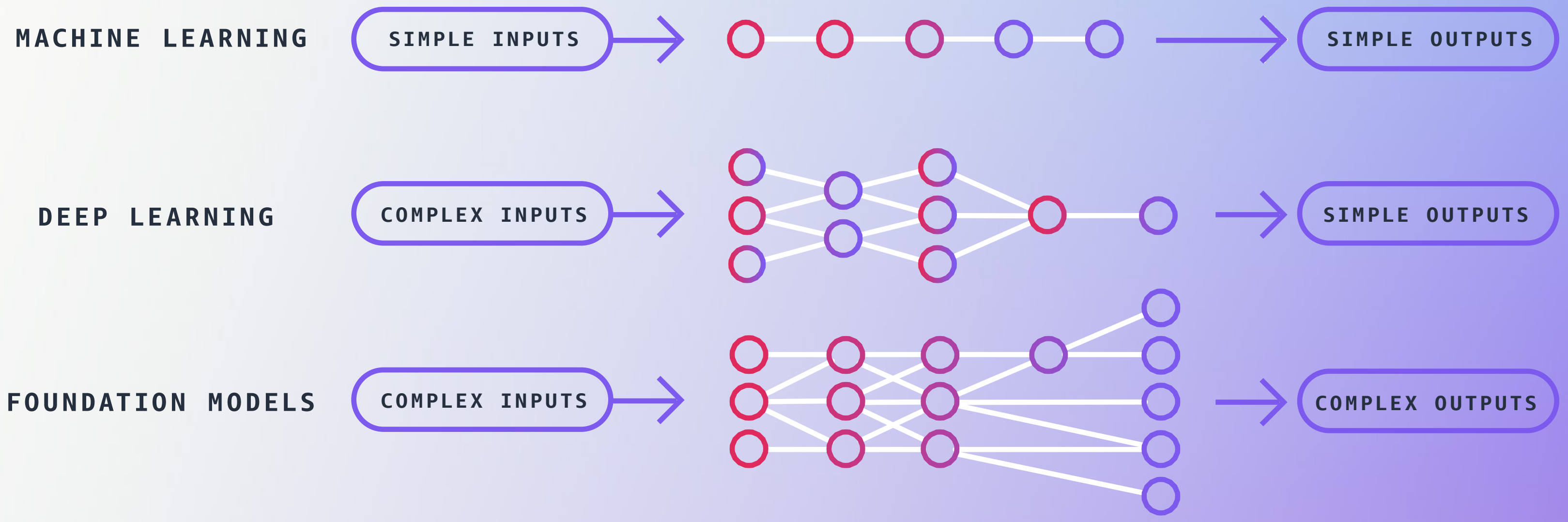
Read on to learn how your startup can begin realizing the business value of generative AI today—so you can keep pace with the market and leapfrog your competition.

Understanding generative AI

Before your startup can fully unlock the business value of generative AI, it's important to have a fundamental understanding of how the technology works. Generative AI is a term used to describe algorithms that can create new content and ideas, including conversations, stories, images, videos, and music. Generative AI is powered by extremely large ML models that are pretrained on vast amounts of data. These are commonly known as **foundation models (FMs)**.

Traditional forms of ML allowed us to take simple inputs, like numeric values, and map them to simple outputs, like predicted values. With the advent of deep learning, we can take complicated inputs, like videos or images, and map them to relatively simple outputs, for example, if the image contains a cat or not. With generative AI, we can leverage massive amounts of complex data to capture and present knowledge in more advanced ways—mapping complicated inputs to complicated outputs, like summarizing a long document and extracting the key insights.

Text-based generative AI systems use a specific type of FM called a **large language model (LLM)**. LLMs can perform a wide range of tasks that span multiple domains, like writing code, solving math problems, engaging in dialogue, and analyzing the information in documents to answer questions.

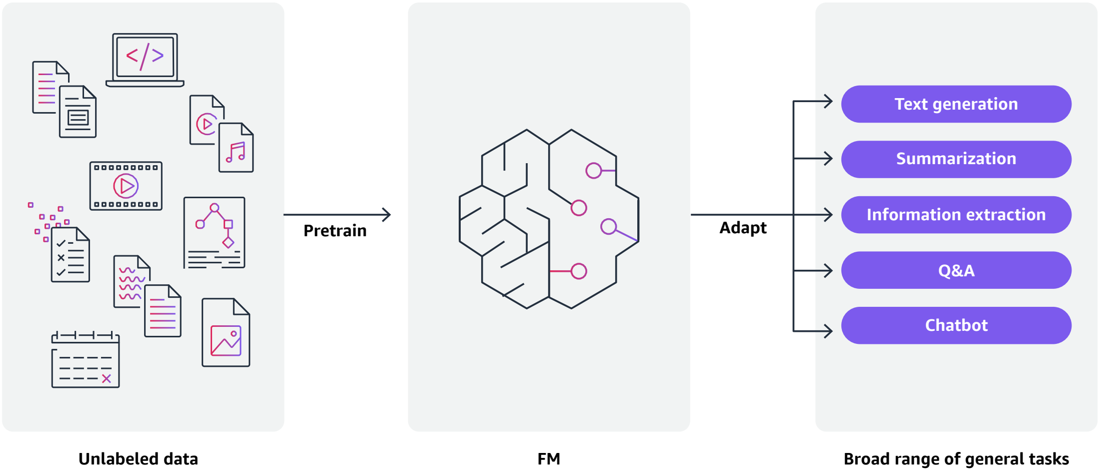


Make data your differentiator

When you want to build generative AI applications that are unique to your business needs, your startup's data is your strategic asset. FMs can be customized and fine-tuned with your startup's proprietary data to deliver a more differentiated experience compared to an "out-of-the-box FM." For example, a grocery chain that tracks shopper preferences can customize an FM to produce a better recommendation engine that is highly differentiated from competitors' offerings.

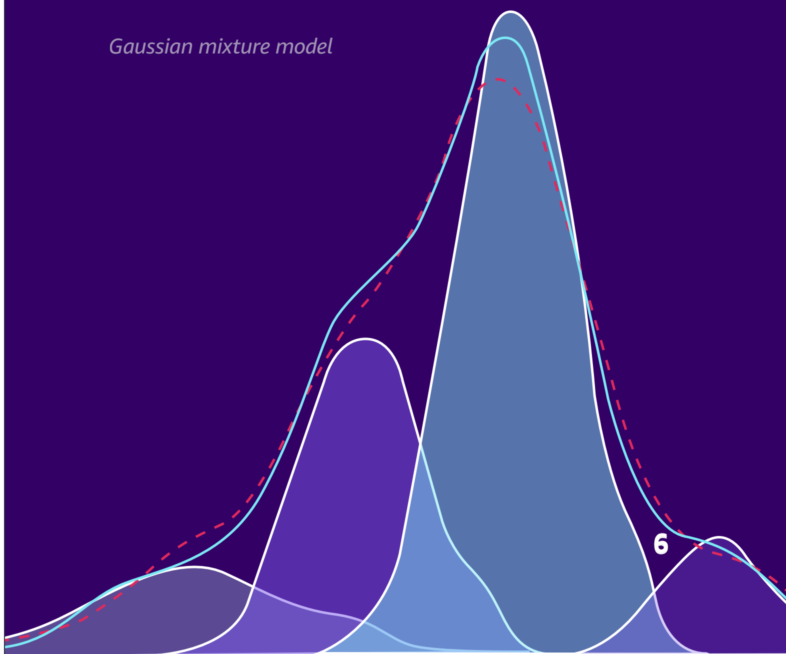
Startups can also use customized FMs to easily create unique content that embodies their brand's tone and style. For instance, a fintech startup that needs to auto-generate a daily activity report for internal circulation can customize an FM with proprietary data, including past reports. The FM could then learn how these reports should read and what data was use to generate them—to deliver a report that better reflects the needs of the business.

Now that you have a general understanding of how the technology works, let's begin exploring how you can put generative AI to work for your startup.



Moments in generative AI history:

Today's FMs that are used to create generative AI applications are built atop a long history of AI innovation. Two of the earliest models with generative AI capabilities are the hidden Markov model (HMM) and the Gaussian mixture model (GMM), both developed in the 1950s. HMMs use known data to make educated guesses about unknown data (for example, predicting whether a card player is cheating based on their results). GMMs can examine a group of data (such as a music playlist) and subgroups within that data (for example, genres) to infer unknown information (such as "this is a rap song"). Both are still used today.



Business capabilities of generative AI

Startups at all stages are using generative AI to enhance productivity and create business value in many ways, including:

Innovation and creativity

Get suggestions, generate prototypes, and explore innovative concepts

Code generation

Improve developer productivity by 57% with AI coding companion **Amazon CodeWhisperer**[5]

Contact center analytics

Summarize and extract insights from customer calls

Personalization

Improve personalized recommendations and generate tailored content

Virtual assistants

Enhance customer experience with human-like responses

Conversational search

Extract insights from all your corporate information

Content generation

Create text, images, videos, and music

Moments in generative AI history:

Another early example of generative AI is ELIZA, a chatbot (or "chatterbot," as they were previously known) developed by an MIT professor from 1964 to 1966. Like its namesake, Eliza Doolittle of *Pygmalion* and *My Fair Lady*, the program grew more sophisticated by "learning" from human interactions. ELIZA was most famously used to mimic the behavior of a therapist conducting an initial psychiatric interview, with the user playing the role of the patient.

[5] Data collected from a "productivity challenge" conducted by Amazon during the Amazon CodeWhisperer preview

Business considerations for generative AI

As you work to identify the capabilities of generative AI that are most useful to your startup—and develop a strategy for implementing them into your business processes—you will need to determine which FMs to use in creation of generative AI applications to suit different needs.

You should also carefully consider the infrastructure you will be using to support your FMs. Your models will benefit from a cost-efficient infrastructure that meets your requirements for performance.

When evaluating FMs used to create generative AI applications, look for models that offer:

1. Easy ways to build and scale generative AI applications with security and privacy built in
2. Performant, low-cost infrastructure to train your own models and run inference at scale
3. Generative AI-powered applications to transform how work gets done
4. Data as your differentiator

Responsible AI, security, and privacy

With their vast size and open-ended nature, FMs raise new issues in defining, measuring, and mitigating responsible AI concerns across the development cycle, such as accuracy, fairness, intellectual property (IP) considerations, hallucinations, toxicity, and privacy. For example, looking at the issue of fairness, can we ask an LLM to assign male and female pronouns at the same rate in reference to a doctor? Does that still apply if the prompt describes the doctor as having a beard? And should we do the same for other professions? What about the Women's National Basketball Association (WNBA)? You can see that simply defining fairness in the context of an LLM is challenging and requires new approaches and solutions.

Generative AI technology and how it is used will continue to evolve, posing new challenges that will require additional attention and mitigation. To tackle these challenges and foster innovation, **academic, industry, and government partners** are working together to explore new solutions and concepts to ensure that generative AI continues to evolve in a responsible, private, and secure way.

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

Read the blog post *Responsible AI in the Generative Era*
Learn more ›

Moments in generative AI history:

In 2014, the development of the first generative adversarial network (GAN) marked one of the biggest breakthroughs in generative AI. In a GAN, two models (a "generator" and a "discriminator") compete in a zero-sum game. The generator manufactures content that appears increasingly "genuine," while the discriminator analyzes its opponent's techniques to better identify fakes. This novel approach of using AI to train other AI proved revelatory, while GANs themselves unlocked a new era for digital imagery.

How AWS can help you succeed with generative AI

You can unlock the full business value of generative AI for your startup with AWS. Reinvent your applications, create entirely new customer experiences, drive unprecedented levels of productivity, and ultimately transform your startup.

Experience and expertise

One of the key advantages of AWS lies in a rich AI heritage built over two decades of focused investment. In fact, more than 100,000 customers currently use AWS for AI and ML.

Amazon, the driving force behind AWS, harnesses ML capabilities to power its ecommerce recommendations engine, optimize robotic picking routes in fulfillment centers, and much more. Further, ML informs Amazon's supply chain, forecasting, and capacity planning.

Deep learning is also employed in the Amazon Prime Air drone delivery system and the computer vision (CV) technology behind Amazon Go, the innovative retail experience that allows customers to select items and leave the store without traditional checkouts. And Alexa, which is supported by more than 30 different ML systems, helps customers with a wide array of tasks billions of times each week.

With thousands of dedicated ML engineers, AI and ML are deeply ingrained in the heritage of Amazon and AWS—continuing to shape the future.

More than

100,000

customers currently use AWS for AI and ML

Why startups build with AWS

Startups at any stage choose to build generative AI and other AI applications on AWS for many reasons. Here are some of the top advantages of building on AWS, according to our customers:

The easiest way to build and scale generative AI applications with security and privacy built in

Amazon Bedrock is the easiest way for customers to build and scale generative AI–based applications using FMs. Bedrock makes **Amazon Titan** FMs and models from leading AI companies such as AI21 Labs, Anthropic, Cohere, Stability AI, and Meta accessible via an API. Customers using Bedrock can leverage the benefits of AWS, which is architected to be the most flexible and secure cloud computing environment available today. **Agents for Amazon Bedrock** is a fully managed capability that makes it easier for developers to create generative AI applications that can deliver up-to-date answers based on proprietary knowledge sources and complete tasks for a wide range of use cases.

The most performant, low-cost infrastructure for generative AI

For years, AWS has invested in developing silicon that delivers the highest levels of performance and cost optimization for ML workloads. The results—**AWS Trainium** and **AWS Inferentia**—deliver the lowest costs for training models and running inference in the cloud. AWS has also developed **Amazon Elastic Compute Cloud** (Amazon EC2) instances to help you take advantage of these capabilities. For example, **Amazon EC2 Trn1** instances powered by Trainium save you up to 50 percent on training costs,[6] while **Amazon EC2 Inf2** instances powered by AWS Inferentia2 deliver up to 40 percent lower cost per inference.[7]

Data as your differentiator

With AWS, it's easy to use your startup's data as a strategic asset to customize FMs and build more differentiated experiences. Data is the difference between a general generative AI application and one that truly knows your business and your customer. And with the most comprehensive set of data and AI services, you can securely customize an FM on AWS with your data and build a model that is an expert on your business, your data, and your customers.

Generative AI-powered applications to transform how work gets done

AWS is building powerful new applications that transform how our customers get work done with generative AI. Boost productivity with purpose-built conversational agents that streamline coding with **Amazon CodeWhisperer**, simplify business intelligence with **Amazon QuickSight Generative BI**, and improve clinical efficiency for healthtec startups with **AWS HealthScribe**. With security, privacy, and responsible AI at the forefront, easy customization, and integration into your existing data sources and applications, startups can quickly take advantage of generative AI without the heavy lifting.

Further reading on responsible AI for startups:

Additional generative AI resources ›

Additional ML resources ›

[6] Over other comparable Amazon EC2 instances

[7] Compared to prior generation AWS Inferentia–based instances

AWS generative AI services

Facilitate your generative AI applications with a range of AWS technologies, including:

Amazon Bedrock ›

Build and scale generative AI applications with FMs. Bedrock supports a variety of FMs, including:

- **Amazon Titan**: For text summarization, generation, classification, open-ended Q&A, information extraction, embeddings, and search
- **AI21 Labs Jurassic-2 Multilingual LLMs**: For text generation in various languages
- **Anthropic Claude 2**: LLM for conversations, question answering, and workflow automation based on research into training honest and responsible AI systems
- **Stability AI Stable Diffusion**: Generates unique, realistic, high-quality images, art, logos, and designs
- **Cohere Command + Embed:** Text generation model for business applications and embeddings model for search, clustering, or classification in over 100 languages
- **Meta Llama 2**: Pretrained and fine-tuned LLMs for natural language tasks like question and answering and reading comprehension

AWS Trainium: Train models faster with up to 50% cost savings[8] using this ML model accelerator

AWS Inferentia2: Run high-performance FM inference with up to 40% lower cost per inference using this accelerator[9]

Amazon CodeWhisperer: Enjoy 57% faster application development[10] while helping to ensure security with this AI coding companion, which is at no cost for individual use

Amazon QuickSight Generative BI: Transform traditional multistep business Intelligence (BI) tasks into intuitive and powerful natural language experiences with Generative BI capabilities in Amazon QuickSight

Amazon SageMaker: Build your own FMs with managed infrastructure and tools to accelerate scalable, reliable, and secure model building, training, and deployment

Amazon SageMaker JumpStart: ML hub that provides access to algorithms, models, and ML solutions so you can quickly get started with ML. With SageMaker JumpStart, ML practitioners can choose from a broad selection of **publicly available FMs**, deploy FMs to dedicated SageMaker instances from a network-isolated environment, and customize models using SageMaker for model training and deployment.

[8] AWS Trainium delivers up to 50 percent cost-to-train savings over comparable Amazon EC2 instances

[9] AWS Inferentia delivers up to 40 percent cost per inference over comparable Amazon EC2 instances

[10] Data collected from a "productivity challenge" conducted by Amazon during the Amazon CodeWhisperer preview

CUSTOMER STORIES

Startups are proving what's possible with generative AI

Featured customer stories

Startups of all sizes are integrating generative AI into their businesses to innovate faster and build a competitive advantage over competitors. Here's how AWS is helping four startups take advantage of this revolutionary technology.

CUSTOMER STORIES

InsightFinder kick-starts success with AWS solutions

The startup **InsightFinder**, an AI-driven predictive observability platform, faced a problem of scale as the number of students and teachers using the platform grew quickly. The company lacked the internal infrastructure to filter the alerts sent its way. By connecting the InsightFinder engine with data from **Amazon CloudWatch**, the company was able to receive essential insights quickly and easily.

Read the story ›

InsightFinder

> **"A lot of AI tech companies think you need to invest heavily in hardware resources. [With AWS,] we can actually build a high-performance engine, and with reasonable cost."**
>
> Helen Gu, Founder, InsightFinder

CUSTOMER STORIES

Fraud.net builds a modern anti-fraud app using AWS machine learning solutions

Fraud.net, a fraud and compliance platform, was founded to solve the multi-percent fraud rates that harm many lenders, banks, payments processors, and digital commerce companies, as well as their customers. It realized that a lack of transparency into data was the biggest impediment to this goal. Fraud.net built a rapidly deployable, scalable, and secure platform on which to unify fraud data and create actionable insights. The startup leveraged an event-driven architecture on AWS, giving it the ability to scale up and down according to the number of events. It used AWS solutions, including Amazon EC2 and Lambda for compute and Amazon S3 for highly scalable object storage. These solutions helped it to unify and analyze three levels of data: customer-level, institution-level, and cross-institution-level data.

Read the story ›

Fraud.net

> "AWS helps us process thousands of transactions per second, at a scale that was virtually impossible three or four years ago."

Whitney Anderson, Co-Founder & CEO, Fraud.net

CUSTOMER STORIES

Mantium achieves low-latency GPT-J inference with DeepSpeed on SageMaker

Mantium, a global cloud-platform provider for building and managing AI applications, enables businesses of all sizes to build AI applications and automation faster and easier than what has been traditionally possible. But Mantium faced a challenge: Open-source models are rarely designed for production-grade performance. Response latency is a core obstacle for the generative pretrained transformers, such as GPT-J, that power modern text generation. This can make production deployment impractical and even infeasible. Mantium leveraged DeepSpeed's inference engine to inject optimized CUDA kernels into the Hugging Face Transformers GPT-J implementation, dramatically increasing text generation speeds with GPT-J.

Read the story ›

MANTIUM

> "DeepSpeed's inference engine is simple to integrate into a SageMaker inference endpoint. SageMaker makes it very easy to deploy custom inference endpoints, and integrating DeepSpeed was as simple as including the dependency and writing a few lines of code."

Joe Hoover, Senior Applied Scientist, R&D, Mantium

Stability AI gains resiliency, performance, and cost savings with SageMaker

FMs—large models that are adaptable to a variety of downstream tasks in domains such as language, image, audio, and video—are hard to train because they require a high-performance compute cluster with thousands of GPU or Trainium chips, along with software to efficiently utilize the cluster. **Stability AI**, a community-driven, open-source AI company developing breakthrough technologies, selected AWS as its preferred cloud provider to provision one of the largest-ever clusters of GPUs in the public cloud. Using SageMaker-managed infrastructure and optimization libraries, Stability AI's model training has become more resilient, performant, and cost-efficient: It has cut training time and costs by over half.

Read the story ›

stability.ai

> **"AWS has been an integral partner in scaling our open-source foundation models across modalities, and we are delighted to bring these to SageMaker to enable tens of thousands of developers and millions of users to take advantage of them."**
>
> Emad Mostaque, Founder & CEO, Stability AI

CUSTOMER STORIES

Runway scales in-house research infrastructure with AWS

Runway partnered with AWS to scale its high performance computing (HPC) cluster and leverage our research infrastructure to bring best-in-class user experiences across its Generative Suite. Runway's Gen-2 system, trained on AWS, can generate novel videos with text, images, or video clips. Gen-2 improves on Runway's multimodal generative models and represents a major advancement in state-of-the-art AI systems for video generation.

Read the story ›

runway

> **"AWS was instrumental in the development and training of this groundbreaking video generation model. We look forward to continuing to pioneer what's possible with generative AI together."**
>
> Cristóbal Valenzuela, Co-Founder & CEO, Runway

NEXT STEPS

Get started with generative AI

Now that you have a better understanding of generative AI, what it can do, and its potential business benefits, the next step is to clearly define your objectives and identify use cases for leveraging it. It's best to start with smaller experiments and simple, precise goals. Once you've achieved some quick wins, you can begin scaling your efforts upward and outward.

Collaboration with experts is highly recommended to ensure you consider factors such as data availability, data quality, and ethical implications related to generative AI. Furthermore, infrastructure considerations should not be an afterthought, as they can significantly impact costs, scalability, and energy consumption. Engaging with AWS experts can provide valuable guidance throughout the decision making process and stages of implementation.

Get started today!

Generative AI promises to be one of the most disruptive technologies in generations— one that can enhance human creativity, push the limits of innovation, and maximize output. AWS is at the forefront, committed to developing fair and accurate AI services and providing your startup with the tools and guidance needed to build AI applications responsibly. It's time for your startup to get started.

Learn more about generative AI for startups with AWS ›