aws startups

Generative AI for every startup

Easily build and scale 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, venture capitalists (VCs) are increasing their positions in generative AI—from $408 million in 2018 to $4.8 billion in 2021 and to $4.5 billion in 2022. Angel and seed deals have grown as well, with $358.3 million invested in 2022 compared with just $102.8 million in 2018.[1]

Those VC numbers aren't surprising, considering the proven benefits of generative AI. As an entrepreneur, you can use the technology to automate tasks, 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.[2]

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 organizations of every size are choosing Amazon Web Services (AWS) for their generative AI journeys. But first, let's consider the fundamentals of the technology.

[1] "**Vertical Snapshot: Generative AI**," PitchBook, March 2023
[2] Haan, K., "**24 Top AI Statistics and Trends In 2023**," Forbes, April 2023

Understanding generative AI

Generative AI is a type of AI that can create new content and ideas, including conversations, stories, images, videos, and music. It is powered by large models that are pretrained on vast amounts of data, commonly referred to as foundation models (FMs).

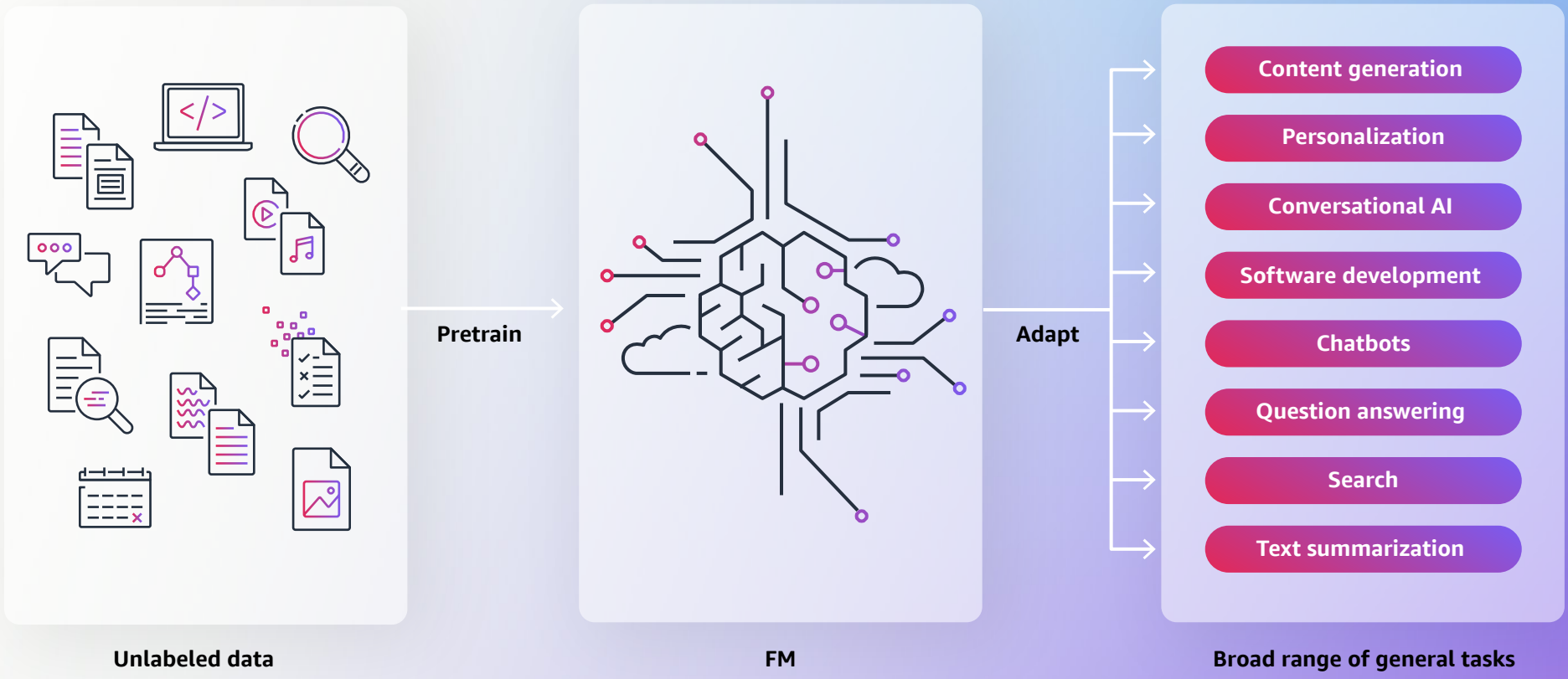
Recent advancements in ML—specifically the invention of the transformer-based neural network architecture—have led to the rise of models that contain billions of parameters or variables. To give a sense of the change in scale, the largest pretrained model in 2019 was 330M parameters. Now, in 2023, the largest models are more than 500B parameters—a 1,600 times increase in size in just a few years. Today's FMs, such as the large language models (LLMs) GPT3.5 or BLOOM and the text-to-image model Stable Diffusion, 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.

Through their pretraining exposure to internet-scale data in all its various forms and patterns, FMs learn to apply their knowledge within a wide range of contexts. While the capabilities of a pretrained FM are amazing, the exciting thing is that these models can also be customized to perform domain-specific functions that use only a small fraction of the data and compute required to train a model from scratch.

Upleveling the customer experience with customized FMs

Customized FMs can create unique customer experiences that embody a company's voice, style, and services across a wide variety of industries. For instance, a fintech startup that needs to auto-generate a daily activity report using all the relevant transactions can customize the model with proprietary data. This data includes past reports, enabling the FM to learn how reports should read and what data was used to generate them.

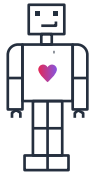
Generative AI on AWS enables you to reinvent your applications, create entirely new customer experiences, drive unprecedented levels of productivity, and transform your startup. You can choose from a range of popular FMs or use AWS services that have generative AI built in, all running on the most cost-effective cloud infrastructure for generative AI.



Top ways to apply generative AI to your startup

Content generation

Assist with tasks such as writing essays, reports, emails, concept art, and designs or generating unique pieces of content that may not have been possible with human effort alone.



Chatbots

Create natural language-based conversational interfaces, enhancing the user experience by providing more human-like interactions.

Personalization

Create a more personalized experience for your customers with highly relevant content and product recommendations for your site and communications.

Question answering

Find and synthesize information and quickly answer customer questions using natural language prompts from a large body of data, such as the internet.

Conversational AI

Create natural language-based conversational interfaces, such as chatbots and virtual assistants, and leverage speech-to-text and translation capabilities.

Search

Find content and information in documents and other assets to improve search accuracy, generate search results faster, and unlock insights to make data-driven business decisions.

Software development

Generate code snippets, comments, and documentation based on natural language inputs to improve the efficiency and accuracy of software development tasks.

Text summarization

Generate a shorter version of an article, document, or webpage; a concise overview of a large amount of text; or the key points of a block of text.

Why AWS for generative AI?

AI and ML have been a focus for Amazon for over 20 years, and many of the capabilities customers use with Amazon are driven by ML. Our ecommerce recommendations engine, the paths that optimize robotic picking routes in our fulfillment centers, and our supply chain, forecasting, and capacity planning are all informed and driven by ML.

Prime Air (our drones) and the computer vision (CV) technology in Amazon Go (our physical retail experience that lets consumers select items off a shelf and leave the store without having to formally check out) use deep learning (DL). Alexa, powered by more than 30 different ML systems, helps customers billions of times each week to manage smart homes, shop, get information and entertainment, and more. We have thousands of engineers at Amazon committed to ML, and it's a big part of our heritage, current ethos, and future.

AI and ML have been a focus for Amazon for over 20 years

Democratizing AI for startups of any size

Our approach is to democratize 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. There are a few key reasons customers use AWS for generative AI applications.

1. **The most cost-effective infrastructure:** To achieve your goals with generative AI, you need the most performant, cost-effective infrastructure that is purpose-built for ML. Over the last five years, AWS has been investing in our own silicon to push the envelope on performance and price performance for demanding workloads like ML training and inference. And our **AWS Trainium** and **AWS Inferentia** chips offer the lowest cost for training models and running inference in the cloud. With ML infrastructure powered by NVIDIA GPUs and AWS-designed ML chips, customers get the flexibility to choose the optimal infrastructure that will maximize performance while controlling costs.
2. **Flexibility:** Choose from the broadest choice of models from leading AI startups and Amazon to suit your unique business requirements and choose from a wide selection of FMs from AI21 Labs, Anthropic, Stability AI, and Amazon to find the right model for your use case. No other vendor offers the same breadth and depth of choice.
3. **Secure customization:** Customize FMs for your business with just a few labeled examples. Because all data is encrypted and does not leave your **Amazon Virtual Private Cloud** (Amazon VPC), you can trust that your data will remain private and confidential. AWS offers 300 security, compliance, and governance services and features, offering the flexibility and customization needed for customers to build an end-to-end security strategy that is right for them.
4. **The easiest way to build with FMs:** Quickly integrate and deploy FMs into your applications and workloads running on AWS. Use familiar controls and integrations with our breadth and depth of capabilities and services like **Amazon SageMaker** and **Amazon Simple Storage Service** (Amazon S3).
5. **Generative AI-powered solutions:** With generative AI built in, services such as **Amazon CodeWhisperer**, an AI coding companion, can help you improve productivity. In addition, you can deploy common generative AI use cases like call summarization and question answering using AWS sample solutions that combine AWS AI services with leading FMs.

The AWS approach to responsible AI

AWS builds FMs with responsible AI in mind at each stage of its comprehensive development process. Throughout design, development, deployment, and operations, we consider a range of factors, including:

Accuracy

Evaluating factual correctness or how closely a summary reflects the underlying document

Fairness

How a system impacts different subpopulations of users (e.g., by gender, ethnicity)

Intellectual property (IP) and copyright considerations

Appropriate usage

Filtering out user requests for legal advice, medical diagnoses, or illegal activities

Toxicity

Restricting hate speech, profanity, violence, and offensive and inappropriate language

Privacy

Protecting personal information, and customer prompts

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 of our FMs, we invest actively to improve features and learn from customers as they experiment with new use cases. At AWS, we know that generative AI technology and its uses will continue to evolve, posing new challenges that require additional attention and mitigation. Together with academic, industry, and government partners, we are committed to the continued development of generative AI in a responsible way.

Learn more about the AWS approach to responsible AI and ML ›

To learn more about these challenges and emerging solutions read this Amazon Science blog ›

Tools for building with generative AI on AWS

1. Amazon Bedrock

The easiest way to build and scale generative AI applications with FMs

Amazon Bedrock is a fully managed service that makes FMs from leading AI startups and Amazon available via an API, so you can choose from a wide range of FMs to find the model that is best suited for your startup's use case. With the Bedrock serverless experience, you can get started quickly, privately customize FMs with your own data, and easily integrate and deploy FMs into your applications using the AWS tools and capabilities you are familiar with. These include integrations with **SageMaker**, features like SageMaker Experiments to test different models and SageMaker Pipelines to manage your FMs at scale without having to manage any infrastructure.

With Bedrock, you can build and scale generative AI applications that can generate text and images in response to prompts. Your team will gain access to FMs from top AI startups, including AI21, Anthropic, and Stability AI and exclusive access to the **Amazon Titan** family of FMs developed by AWS.

Learn more about Amazon Bedrock ›

stability.ai

Stability AI is the open-source, generative AI company behind the popular image-generation FM Stable Diffusion. "I'm excited to strengthen our ongoing partnership with AWS by making our Stable suite of open models available to AWS customers through Amazon Bedrock. This collaboration with AWS is a testament to our commitment to deliver cutting-edge open artificial intelligence solutions that can help businesses make more informed decisions and achieve greater stability in an ever-changing world. We believe that this partnership will unlock significant value for AWS customers, and we look forward to working closely together to bring these powerful capabilities to a wider audience."

Emad Mostaque, CEO, Stability AI

2. Amazon Titan models

Innovate responsibly with high FMs from Amazon

Amazon Titan currently consists of two FMs. The first is Titan Text, a generative LLM for tasks such as summarization, text generation (creating a blog post, for example), classification, open-ended Q&A, and information extraction. The second is Titan Embeddings, an LLM that translates text inputs (words, phrases, or large units of text) into numerical representations (known as embeddings) that contain the semantic meaning of the text. While this LLM does not generate text, it is useful for applications like personalization and search because by comparing embeddings, the model will produce more relevant and contextual responses than word matching. In fact, the Amazon.com product-search capability uses a similar embedding model to help customers find the products they're looking for. To continue supporting best practices in the responsible use of AI, Amazon Titan FMs are built to detect and remove harmful content in the data, reject inappropriate content in the user input, and filter outputs that contain inappropriate content, such as hate speech, profanity, and violence.

Learn more about Amazon Titan models ›

3. Trainium and Inferentia

High-performance, cost-effective infrastructure for generative AI

Amazon Elastic Compute Cloud (Amazon EC2) Trn1 instances, powered by **Trainium** accelerators, are purpose-built for high-performance DL training of generative AI models, including LLMs and latent diffusion models. Trn1 instances offer up to 50 percent cost-to-train savings over other comparable Amazon EC2 instances. You can use Trn1 instances to train DL models across a broad set of applications, such as text summarization, code generation, question answering, image and video generation, recommendation, and fraud detection.

The **AWS Neuron** SDK helps developers train models on AWS Trainium and deploy models on **Inferentia** accelerators. It integrates natively with frameworks, such as PyTorch and TensorFlow, so you can continue using your existing code and workflows to train models on Trn1 instances.

Learn more about Amazon EC2 Trn1 instances ›

Money Forward

"We launched a large-scale AI chatbot service on the Amazon EC2 Inf1 instances and reduced our inference latency by 97% over comparable GPU-based instances while reducing costs. As we keep fine-tuning tailored NLP models periodically, reducing model training times and costs is also important. Based on our experience from successful migration of inference workload on Inf1 instances and our initial work on AWS Trainium-based EC2 Trn1 instances, we expect Trn1 instances will provide additional value in improving end-to-end ML performance and cost."

Takuya Nakade, CTO, Money Forward, Inc.

Amazon EC2 Inf2 instances powered by AWS Inferentia2

Amazon EC2 Inf2 instances are purpose-built for DL inference. They deliver high-performance inference at the lowest cost in Amazon EC2 for generative AI models, including LLMs and vision transformers. You can use Inf2 instances to run your inference applications for text summarization, code generation, video and image generation, speech recognition, personalization, fraud detection, and more.

Inf2 instances are powered by AWS Inferentia2, the second-generation Inferentia accelerator. Inf2 instances raise the performance of Inf1 by delivering up to four times higher throughput and up to 10 times lower latency. Inf2 instances are the first inference-optimized instances in Amazon EC2 to support scale-out distributed inference with ultra-high-speed connectivity between accelerators. Your teams can now efficiently and cost-effectively deploy models with hundreds of billions of parameters across multiple accelerators on Inf2 instances. Inf2 instances deliver up to 40 percent better price performance than comparable Amazon EC2 instances.

Learn more about Amazon EC2 Inf2 ›

Finch Computing reduces inference costs by 80% using AWS Inferentia for language translation

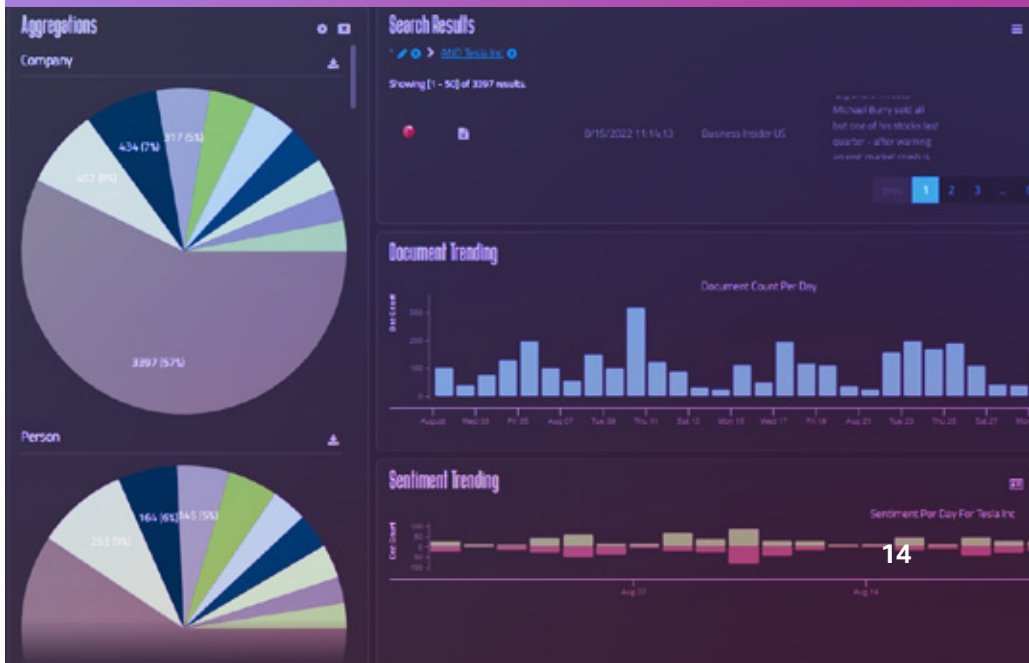
Finch Computing, which develops natural language processing (NLP) technology that uncovers insights from huge volumes of text data, wanted to fulfill customers' requests to support additional languages. Finch had built its own neural translation models using DL algorithms with a heavy compute requirement that depended on GPUs. Now the company needed a scalable solution to support global data feeds and enable it to iterate new language models quickly without incurring prohibitive costs. It created a compute infrastructure centered around the use of AWS Inferentia. As a result, Finch has accelerated its time to market, expanded its NLP to support three additional languages, and attracted new customers.

Learn more ›

FinchAI

"Migrating many production workloads to Inf1 instances, we've achieved an 80% reduction in cost over GPUs. Now we're developing larger, more complex models that enable deeper, more insightful meaning from written text. The performance on Inf2 instances will help us deliver lower latency and higher throughput over Inf1 instances. All told, we are improving our cost-efficiency, elevating the real-time customer experience, and helping our customers glean new insights from their data."

Franz Weckesser, Chief Architect, Finch Computing



aws startups

4. Amazon CodeWhisperer

Build applications faster and more securely

Amazon CodeWhisperer helps developers write code quickly and securely without needing to leave their integrated development environment (IDE) to research something. CodeWhisperer understands comments written in natural language (in English) and can generate multiple code suggestions in real time to improve developer productivity. The service suggests entire functions and logical blocks of code (often consisting of up to 10–15 lines of code) directly in the IDE code editor. CodeWhisperer includes the following benefits:

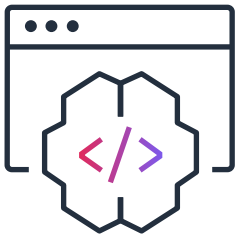
Optimized for use with AWS services
CodeWhisperer makes it more efficient for developers to use AWS services by providing code suggestions that are optimized for AWS APIs, including Amazon EC2, **AWS Lambda**, and Amazon S3. As you write code in your IDE, CodeWhisperer automatically analyzes your code and comments.

Built-in security scans
With CodeWhisperer, you can scan Java, JavaScript, and Python projects to detect hard-to-find vulnerabilities, such as those in the top 10 Open Worldwide Application Security Project (OWASP) or those that don't meet crypto library best practices and other similar security best practices. The service analyzes existing code in the IDE (whether generated by CodeWhisperer or written by you), identifies problematic code with high accuracy, and provides intelligent suggestions on how to remediate it.

Code responsibly: reference tracker for open-source code
CodeWhisperer provides a built-in reference tracker that detects whether a code suggestion might resemble open-source training data and can flag such suggestions. These suggestions are annotated with the open-source project's repository URL, file reference, and license information so that you can review before deciding whether to incorporate the suggested code.

Code responsibly: bias avoidance
Responsible use of AI and ML technologies is key to fostering continued innovation. CodeWhisperer helps developers avoid bias by filtering out code suggestions that might be considered biased or unfair.

Learn more about CodeWhisperer ›



In a productivity challenge, participants who used Amazon CodeWhisperer were 27% more likely to complete tasks successfully than non-users and completed the tasks 57% faster.

5. The power of 2: AWS Partner Hugging Face with SageMaker

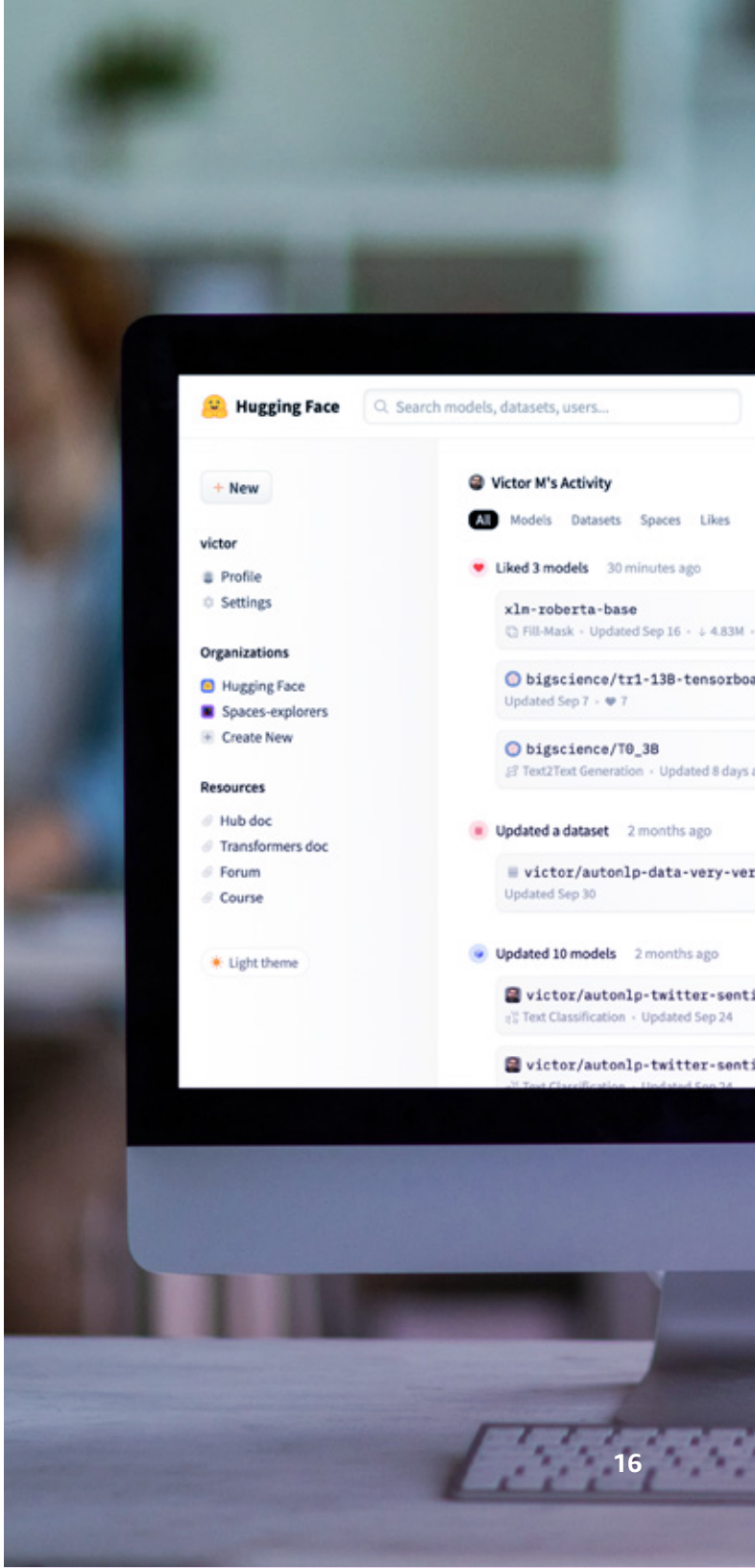
Making open-source models more accessible and cost-efficient

Hugging Face is a large open-source community focused on ML. AWS has a strong partnership with Hugging Face to accelerate the training, fine-tuning, and deployment of large language and vision models used to create generative AI applications. You can start using Hugging Face models on AWS in three ways: through **Amazon SageMaker JumpStart**, **Hugging Face AWS Deep Learning Containers** (DLCs), or the **tutorials** to deploy your models to Trainium or Inferentia. The Hugging Face DLC is packed with optimized transformers, datasets, and tokenizer libraries that enable you to fine-tune and deploy generative AI applications at scale in hours instead of weeks with minimal code changes.

Learn more about Hugging Face on SageMaker ›

"The future of AI is here, but it's not evenly distributed. Accessibility and transparency are the keys to sharing progress and creating tools to use these new capabilities wisely and responsibly. Amazon SageMaker and AWS-designed chips will enable our team and the larger machine learning community to convert the latest research into openly reproducible models that anyone can build on."

Clement Delangue, CEO, Hugging Face



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 ›

aws startups

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 ›

aws startups

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 ›

aws startups

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

CUSTOMER STORIES

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 ›

aws startups

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

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 and ML services and providing your startup with the tools and guidance needed to build AI and ML applications responsibly.

It's time for your startup to get started. Once you've familiarized yourself and your team with generative AI's potential and initial concepts, you can start by clearly defining your objectives. Identifying specific real-world use cases can help keep initial experiments on a smaller scale with more clearly defined goals. Collaboration with experts is encouraged as you consider your data availability and quality, select FMs that best fit your application, and develop your implementation plan. Generative AI can have ethical implications that should be discussed or addressed in your use case.

Considering the scale and growth of generative AI within your organization, infrastructure should not be an afterthought; it can have profound implications on your cost, scale, and energy consumption. Working with experts from AWS can give you a head start across all steps and decisions.

Learn more about generative AI for startups with AWS ›