



AWS FOR EVERY APPLICATION

**Unleash the power
of generative AI:
build better
applications, faster**



Table of contents

Generate incredible outcomes	3
Are you ready for the generative AI revolution?	4
Choose the most cost-effective cloud for generative AI	5
Empower developers to new levels of productivity	7
Build, train, and deploy machine learning (ML) models for any use case	9
Train and run models with more power—and reduced costs	11
Ready to put that incredible power to good use?	12
Why build with AWS?	14

Generate incredible outcomes

The adoption of artificial intelligence (AI) is growing thanks to the availability of scalable compute capacity, exponential growth of data, and the rapid advancements in machine learning (ML) technologies.

And now generative AI is changing the data landscape even faster.

Generative AI represents a new frontier in AI, enabling the creation of fresh content and innovative ideas across multiple domains. The possibilities are limitless, from generating conversations, stories, images, music, video, and even code.

The power of foundation models—and why you don't need to create your own

At the heart of generative AI are the foundation models (FMs), large-scale models with billions of parameters that have been pre-trained on

vast amounts of data. FMs possess an extraordinary range of capabilities. Moreover, these highly capable FMs can be tailored to perform domain-specific tasks, making them invaluable for businesses seeking to differentiate themselves in competitive markets. By leveraging a fraction of the data and compute resources required to build, train, and deploy a model from scratch, organizations can customize FMs that embody their unique voice, style, and services. This level of customization empowers companies across diverse consumer industries to provide personalized and exceptional customer experiences - without having to invest vast sums developing their own FMs. For instance, a financial firm that wants to auto-generate daily activity reports can train a pre-existing FM on past reports and other relevant proprietary data, to create a fine-tuned model capable of generating new reports from scratch (at considerably less cost than creating a new FM.)





As tools using advances in natural language processing work their way into businesses and society, they could drive a 7% (or almost \$7 trillion) increase in global GDP and lift productivity growth by 1.5 percentage points over a 10-year period.

Goldman Sachs²

Are you ready for the generative AI revolution?

Amazon Web Services (AWS) is helping customers rapidly adopt generative AI applications to drive new levels of productivity and transform their offerings—while preparing for what's next.

AWS is helping organizations overcome the hurdles associated with this complex and demanding technology. AWS offers developers the easiest place to build with FMs, the most price-performant infrastructure for ML, flexibility to work with open source FMs or build your own, and use FM powered tools like Amazon CodeWhisperer to rapidly improve developer productivity.

By embracing AWS's cloud-based solutions, developers can leverage innovative generative AI products that meet the ever-growing demands of customers. In this document, we will take a deep dive into each of these dedicated tools and services, exploring their unique capabilities and demonstrating how they can elevate your business in the new world of generative AI.

A report from McKinsey found that 90% of commercial leaders expect to use generative AI products regularly in the next two years.¹

Choose the most cost-effective cloud for generative AI

For more than 25 years, AWS has harnessed AI to deliver remarkable customer experiences, from the [Amazon Prime Air](#) drone delivery system to [Amazon Go](#) stores and [Alexa](#). AWS has helped more than 100,000 customers of all sizes and industries harness this transformative technology—including Intuit, Thomson Reuters, AstraZeneca, Ferrari, Bundesliga, 3M, and BMW, as well as thousands of startups, Small and Medium Enterprises (SMEs), and government agencies around the world.

AWS extends this democratizing approach to generative AI: taking the tech out of the realm of research and development, and extending its availability beyond a handful of startups and large, well-funded tech companies. With generative AI on AWS, you can reinvent your applications, create entirely new customer experiences, drive unprecedented levels of productivity, and transform your business.

AWS provides an ever-evolving suite of products and services to enable organizations to harness the full power of generative AI. 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.



The easiest place to build applications with foundation models



[Amazon Bedrock](#) is the easiest way for customers to build and scale generative AI-based applications using FMs, democratizing access for all builders. Amazon Bedrock makes FMs from Amazon and leading AI startups including AI21 Labs, Anthropic, Cohere, and Stability AI accessible via an API.

Democratizing access to high-performing FMs



Customers often seek high-performing FMs that deliver exceptional results tailored to their specific needs. Amazon Bedrock eliminates the complexities of managing large infrastructure clusters and incurring substantial costs. By providing seamless integration with applications, Amazon Bedrock empowers users to effortlessly build and scale generative AI applications without compromising on performance.

Securely building differentiated applications with proprietary data



With Amazon Bedrock, customers can easily build customized applications using their own data while keeping their data secure and private. Amazon Bedrock allows users to securely build upon base FMs, ensuring their applications stand out in the market. By securely leveraging proprietary data, users can create differentiated applications that meet their unique requirements.

Effortless integration and deployment



Amazon Bedrock simplifies the model selection process with a serverless experience that provides scalability and agility while eliminating infrastructure management tasks. Users can easily find the perfect model for their needs and quickly get started. Amazon Bedrock enables customers to privately customize FMs with their own data and seamlessly integrate and deploy them into their applications using familiar AWS tools and capabilities. Users can leverage [Amazon SageMaker](#) ML features like Experiments for testing different models and Pipelines for managing FMs at scale, all without the need to manage any infrastructure.

Cutting-edge FMs at your fingertips



Amazon Bedrock offers access to some of the most cutting-edge FMs available today. For instance, customers can utilize the [Jurassic-2](#) family of multilingual large language models (LLMs) from AI21 Labs, capable of generating text in Spanish, French, German, Portuguese, Italian, and Dutch-based on natural language instructions. Additionally, Anthropic's latest LLM, [Claude 2](#), excels in thoughtful dialogue, content creation, complex reasoning, creativity, and coding, based on Constitutional AI and harmlessness training. Cohere's [Command](#) text generation model is trained for practical business applications and its Embed model is trained for search, clustering, or classification tasks across 100+ languages. Stability AI's text-to-image foundation models, including the highly popular [Stable Diffusion](#), enable the generation of unique, realistic, high-quality images, art, logos, and designs.

Customization made simple



One of Amazon Bedrock's most remarkable features is its ease of customization. None of the customer's data is used to train the original base models. Customers can configure their [Amazon Virtual Private Cloud \(Amazon VPC\)](#) settings to access Amazon Bedrock APIs and provide model fine-tuning data in a secure manner and all data is encrypted. Customer data is always encrypted in transit (TLS1.2) and at rest through service managed keys.

Early adopters praise Amazon Bedrock

Amazon Bedrock has already garnered significant interest from customers such as Coda. Shishir Mehrotra, Co-founder and CEO of Coda, praises Amazon Bedrock for delivering quality, scalability, and performance to their AI solutions while maintaining data security. Amazon Bedrock's accessibility has driven enthusiasm among enterprises of all sizes, including prominent partners like Accenture, Deloitte, Infosys, and Slalom, who are building practices to accelerate generative AI adoption. Independent Software Vendors (ISVs) like C3 AI and Pega are also leveraging Amazon Bedrock, appreciating its vast selection of FMs and AWS's commitment to security, privacy, and reliability.

Empower developers to new levels of productivity

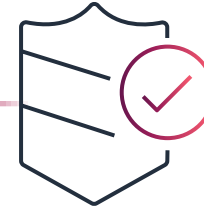


There are many ways AWS helps you take advantage of the advances in generative AI. One of those ways is by infusing generative AI under the hood of services that help accelerate and simplify daily tasks. For example, Amazon CodeWhisperer is a game changing AI coding companion that is revolutionizing developer productivity.



Transforming the developer experience with generative AI

Before [Amazon CodeWhisperer](#), software developers spent a significant amount of time writing code. With CodeWhisperer, developers can now rely on AI-generated code suggestions in their Integrated Development Environment (IDE). By turning over more of the writing of code to CodeWhisperer, developers can work faster and unleash their creativity. The positive response from CodeWhisperer customers demonstrates its potential as one of the most powerful applications of generative AI.



The AI coding companion with built-in security scanning

Developers' productivity isn't truly enhanced if the code suggested by generative AI tools contains hidden security vulnerabilities or fails to handle open source responsibly. CodeWhisperer also has built-in security scanning, powered by automated reasoning. It identifies and suggests remediations for hard-to-detect vulnerabilities, including those listed in the top ten Open Worldwide Application Security Project (OWASP) and crypto library best practices. CodeWhisperer also filters out biased or unfair code suggestions and can flag code resembling open-source code for customers' reference or licensing.



Best coding companion for working with AWS

In addition to learning from the billions of lines of publicly available code, CodeWhisperer has been trained on Amazon code. We believe CodeWhisperer is now the most accurate, fastest, and secure way to generate code for AWS services, including [Amazon Elastic Compute Cloud \(Amazon EC2\)](#), [AWS Lambda](#), and [Amazon Simple Storage Service \(Amazon S3\)](#).

CodeWhisperer seamlessly integrates with popular IDEs like VS Code, IntelliJ IDEA, AWS Cloud9, and more through the AWS Toolkit IDE extensions.

CodeWhisperer is available in 15 languages including Python, Java, JavaScript and many more.



Free tier for individual developers

To ensure CodeWhisperer is accessible to as many developers as possible, AWS offers a free tier for individual users. There are no qualifications or time limits for generating code; anyone can sign up with just an email account. No AWS account is required. For business users, we offer the CodeWhisperer Professional Tier, which includes administration features like single sign-on (SSO) with [AWS Identity and Access Management \(IAM\)](#) integration and higher limits on security scanning.

CodeWhisperer increases Accenture's developer productivity by up to 30%

"Accenture is using Amazon CodeWhisperer to accelerate coding as part of our software engineering best practices initiative in our Velocity platform [...] After searching for multiple options, we came across Amazon CodeWhisperer to **reduce our development efforts by up to 30%** and we are now focusing more on improving security, quality, and performance."

Balakrishnan Viswanathan, Senior Manager, Tech Architecture at Accenture

[Read full story >](#)

Build, train, and deploy machine learning (ML) models for any use case



Amazon SageMaker is built on AWS's two decades of experience developing real-world ML applications, including product recommendations, personalization, intelligent shopping, robotics, and voice-assisted devices. The unique service offers a broad set of ML capabilities used by tens of thousands of customers to access and analyze data, and build, train, and deploy high-quality ML models.

With SageMaker, you can:

- Enable more people to innovate with ML through a choice of tools—integrated development environment (IDE) for data scientists and no-code interface for business analysts.
- Access, label, and process large amounts of structured data (tabular data) and unstructured data (photo, video, geospatial, and audio) for ML.
- Reduce training time from hours to minutes with optimized infrastructure. Boost team productivity up to 10 times with purpose-built tools.
- Automate and standardize MLOps practices and governance across your organization to support transparency and auditability.
- [Amazon SageMaker JumpStart](#) is a 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 foundation models](#). ML practitioners can deploy FMs to dedicated SageMaker instances from a network isolated environment and customize models using SageMaker for model training and deployment.

10x

increase in team productivity

1 trillion+

predictions per month

54%

lower TCO

Up to 50%

faster training through more efficient use of GPUs

40%

reduction in labeling costs

22

compliance programs (PCI, HIPAA, SOC 1/2/3, FedRAMP, ISO, and more)

<10ms

inference overhead latency

SageMaker for business analysts

- Easily prepare data, train models, and generate predictions using a point-and-click interface
- Improve collaboration by sharing models and datasets with data scientists
- AutoML integrated into common BI tools such as Domo, Snowflake, and [Amazon Redshift](#)

SageMaker for data scientists

- Access data from structured and unstructured data sources
- Improve productivity with purpose-built tools
- Use fully managed Jupyter Notebooks with just a few clicks

SageMaker for MLOps

- 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

Watch: [Use Amazon SageMaker to Build Generative AI Applications - AWS Virtual Workshop](#)

Technology Innovation Institute trains its 40 billion parameter LLM on SageMaker

United Arab Emirate's Technology Innovation Institute (TII) is a leading research center working on transformative technologies. Using Amazon SageMaker, TII was able to pre-process and train data for its open-source Falcon models. Both models (Falcon-40B and Falcon-7B) are available through SageMaker JumpStart.

[Read full story ›](#)

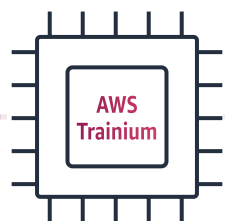
Forethought lowers cost by 66% for its AI models using SageMaker

Helping support teams cut costs with its generative AI software, Forethought turned to AWS and Amazon SageMaker for more efficient utilization of hardware resources like GPUs. With SageMaker, Forethought was able to use multi-model endpoints (MMEs) to run multiple AI models on a single inference endpoint and scale, without the high cost usually associated with training at this scale.

[Read full story ›](#)

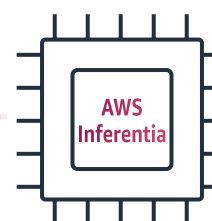
Train and run models with more power—and reduced costs

Generative AI requires extensive processing power. AWS has invested significantly in developing its own silicon to push the boundaries of performance and price performance in demanding workloads. AWS Trainium and AWS Inferentia chips provide the most cost-effective solutions for training models and running inference in the cloud.



Access unparalleled training performance with Trn1 instances

[Amazon EC2 Trn1](#) instances, powered by [AWS Trainium](#), introduce key innovations in ML training. These instances can deliver up to 50 percent cost savings compared to other EC2 instances, offering an exceptional value proposition. With optimization for distributed training across multiple servers connected via second-generation [Elastic Fabric Adapter \(EFA\)](#) networking at 800 Gbps, Trn1 instances enable accelerated training workflows. Customers such as Helixon, Money Forward, and the Amazon Search team have significantly reduced training times for large-scale deep learning models, turning months of training into weeks or even days. Building upon this success, AWS introduces the general availability of Trn1n instances, which feature 1600 Gbps of network bandwidth, enabling 20 percent higher performance for network-intensive models.



Power large-scale generative AI applications with Inf2 instances

[Amazon EC2 Inf2](#) instances powered by [AWS Inferentia2](#) are designed specifically for large-scale generative AI applications, featuring models with hundreds of billions of parameters. They offer up to 4 times higher throughput and up to 10 times lower latency compared to the previous generation. With ultra-high-speed connectivity between accelerators, Inf2 instances facilitate distributed inference at scale. These advancements result in up to 40 percent better inference price performance compared to other EC2 instances, offering the lowest cost for inference in the cloud. Customers like Runway are already experiencing up to 2 times higher throughput with Inf2 instances, enabling them to introduce more features and complex models while delivering a superior experience to millions of creators.

Ready to put that incredible power to good use?

Exploring generative AI sector use cases

The potential of generative AI is vast, with AWS tools and services at your disposal. Here are some of the sectors and key use cases:



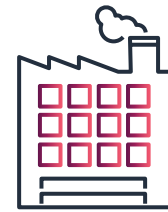
Financial Services

Fraud detection

Identify anomalies, unusual transaction patterns, and potential risks to help financial institutions combat fraud.

Portfolio optimization

Analyze large volumes of financial data to help asset managers make informed decisions based on risk profiles, return expectations, and market conditions.



Manufacturing

Product design optimization

Generate optimized product designs, improve the efficiency and performance of manufactured goods.

Process optimization

Analyze production data to identify bottlenecks, inefficiencies, and opportunities to maximize productivity.



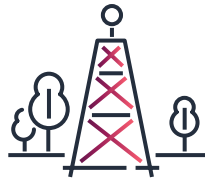
Health and Life Sciences

Drug discovery

Generate novel molecular structures, predict their properties, optimize drug candidates, and accelerate the drug development process.

Medical image analysis

Analyze images, such as X-rays, CT scans, or MRI scans, to detect abnormalities, tumors, or other medical conditions and improve diagnoses.



Telecoms

Marketing preferences

Provide instant responses to customer queries that are both accurate and highly personalized.

Network performance

Identify potential network issues, suggest troubleshooting steps, and automate optimization and maintenance tasks.



Government

Public safety and security

Analyze surveillance footage or sensor data to detect unusual activities, identify potential threats, or monitor public safety.

Policy simulation and planning

Simulate various scenarios and predict the outcomes of policy decisions to help governments make informed decisions.



General


Chatbots and virtual assistants

Provide human-like automated customer support, answering common questions and handling basic tasks.

Language translation and summarization

Translate between different languages or summarize lengthy documents for global communication, content localization, and information extraction.

Why build with AWS?



Our suite of dedicated generative AI tools and services offers the most comprehensive, performant, and scalable infrastructure for cost-effective ML training and inference. Customers who choose AWS for generative AI benefit from:



The easiest place to build with FMs

Quickly integrate and deploy FMs into your applications and workloads running on AWS using familiar controls and integrations with the depth and breadth of AWS capabilities and services such as SageMaker and Amazon S3.



The most price performance infrastructure for ML

Get the best price performance for generative AI with infrastructure powered by AWS-designed ML chips and NVIDIA GPUs. Cost-effectively scale infrastructure to train and run FMs containing hundreds of billions of parameters.



Flexibility to build from scratch

Choose from a wide selection of FMs from AI21 Labs, Anthropic, Cohere, Stability AI to find the right model for your use case.



Generative AI-powered solutions

With generative AI built in, services such as CodeWhisperer, an AI coding companion, can help you improve productivity. In addition, you can deploy common generative AI use cases such as call summarization and question answering using AWS sample solutions that combine AWS AI services with leading FMs.



Secure customization

Customize FMs for your business with just a few labeled examples. Since all data is encrypted and does not leave your Amazon VPC, you can trust that your data will remain private and confidential.

Build tomorrow better with AWS

AWS provides an integrated suite of services to build and deploy generative AI applications leveraging LLMs and FMs. Amazon SageMaker helps developers train, tune, and optimize LLMs, while Amazon Bedrock offers a control plane to manage and monitor models. With Amazon CodeWhisperer, coders can bypass time-consuming coding tasks and accelerate building with unfamiliar APIs

With its end-to-end capabilities, AWS enables rapid development of accurate, robust generative AI applications for any organization.

Learn more about generative AI on AWS ›

¹ ["AI-powered marketing and sales reach new heights with generative AI," McKinsey, May 2023](#)

² ["Generative AI could raise global GDP by 7%," Goldman Sachs, April 2023](#)

