## What is Generative AI?

Generative artificial intelligence (generative AI) is a type of AI that can create new content and ideas, including conversations, stories, images, videos, and music. It can learn human language, programming languages, art, chemistry, biology, or any complex subject matter. It reuses what it knows to solve new problems. For example, it can learn English vocabulary and create a poem from the words it processes. Your organization can use generative AI for various purposes, like chatbots, media creation, product development, and design.

## What is the difference between AI and Generative AI?

[Artificial intelligence](https://aws.amazon.com/what-is/artificial-intelligence/) is the broader concept of making machines more human-like. It includes everything from smart assistants like Alexa, chatbots, and image generators to robotic vacuum cleaners and self-driving cars. Generative AI is a subset that generates new content meaningfully and intelligently.

## When was generative AI created?

Generative AI emerged in the late 2010s with advancements in deep learning, particularly with models like Generative Adversarial Networks (GANs) and transformers. Advances in cloud computing have made generative AI commercially viable and available since 2022.

## What are foundation models in generative AI?

[Foundation models](https://aws.amazon.com/what-is/foundation-models/) are large generative AI models trained on a broad spectrum of text and image data. They are capable of performing a wide variety of general tasks like answering questions, writing essays, and captioning images.

## What are generative AI examples?

Generative AI has several use cases across industries

### **Financial services**

Financial services companies use generative AI tools to serve their customers better while reducing costs:

* Financial institutions use chatbots to generate product recommendations and respond to customer inquiries, which improves overall customer service
* Lending institutions speed up loan approvals for financially underserved markets, especially in developing nations
* Banks quickly detect fraud in claims, credit cards, and loans
* Investment firms use the power of generative AI to provide safe, personalized financial advice to their clients at a low cost

[Read more about generative AI in Financial Services on AWS](https://aws.amazon.com/financial-services/generative-ai/)

### **Healthcare and life sciences**

One of the most promising generative AI use cases is accelerating drug discovery and research. Generative AI can create novel protein sequences with specific properties for designing antibodies, enzymes, vaccines, and gene therapy.

Healthcare and life sciences companies use generative AI tools to design synthetic gene sequences for synthetic biology and metabolic engineering applications. For example, they can create new biosynthetic pathways or optimize gene expression for biomanufacturing purposes.

Generative AI tools also create synthetic patient and healthcare data. This data can be useful for training AI models, simulating clinical trials, or studying rare diseases without access to large real-world datasets.

[Read more about Generative AI in Healthcare & Life Sciences on AWS](https://aws.amazon.com/health/gen-ai/)

### **Automotive and manufacturing**

Automotive companies use generative AI technology for many purposes, from engineering to in-vehicle experiences and customer service. For instance, they optimize the design of mechanical parts to reduce drag in vehicle designs or adapt the design of personal assistants.

Auto companies use generative AI tools to deliver better customer service by providing quick responses to the most common customer questions. Generative AI creates new materials, chips, and part designs to optimize manufacturing processes and reduce costs.

Another generative AI use case is synthesizing data to test applications. This is especially helpful for data not often included in testing datasets (such as defects or edge cases).

[Read more about Generative AI for Automotive on AWS](https://aws.amazon.com/automotive/automotive-generative-ai/)

### **Media and entertainment**

From animations and scripts to full-length movies, generative AI models produce novel content at a fraction of the cost and time of traditional production.

Other generative AI use cases in the industry include:

* Artists can complement and enhance their albums with AI-generated music to create whole new experiences
* Media organizations use generative AI to improve their audience experiences by offering personalized content and ads to grow revenues.
* Gaming companies use generative AI to create new games and allow players to build avatars.

### **Telecommunication**

Generative AI usecases in telecommunication focus on reinventing the customer experience defined by the cumulative interactions of subscribers across all touchpoints of the customer journey.

For instance, telecommunication organizations apply generative AI to improve customer service with live human-like conversational agents. They reinvent customer relationships with personalized one-to-one sales assistants. They also optimize network performance by analyzing network data to recommend fixes.

[Read more about Generative AI for Telecom on AWS](https://aws.amazon.com/telecom/generative-ai/)

### **Energy**

Generative AI suits energy sector tasks involving complex raw data analysis, pattern recognition, forecasting, and optimization. Energy organizations improve customer service by analyzing enterprise data to identify usage patterns. With this information, they can develop targeted product offerings, energy efficiency programs, or demand-response initiatives.

Generative AI also helps with grid management, increases operational site safety, and optimizes energy production through reservoir simulation.

## What are the benefits of generative AI?

According to Goldman Sachs, generative AI could drive a 7 percent (or almost $7 trillion) increase in global gross domestic product (GDP) and lift productivity growth by 1.5 percentage points over ten years.

Next, we give some more benefits of generative AI.

### **Accelerates research**

Generative AI algorithms can explore and analyze complex data in new ways, allowing researchers to discover new trends and patterns that may not be otherwise apparent. These algorithms can summarize content, outline multiple solution paths, brainstorm ideas, and create detailed documentation from research notes. This is why generative AI drastically enhances research and innovation.

For example, generative AI systems are being used in the pharma industry to generate and optimize protein sequences and significantly accelerate drug discovery.

### **Enhances customer experience**

Generative AI can respond naturally to human conversation and serve as a tool for customer service and personalization of customer workflows.

For example, you can use AI-powered chatbots, voice bots, and virtual assistants that respond more accurately to customers for first-contact resolution. They can increase customer engagement by presenting curated offers and communication in a personalized way.

### **Optimizes business processes**

With generative AI, your business can optimize business processes utilizing machine learning (ML) and AI applications across all lines of business. You can apply the technology across all lines of business, including engineering, marketing, customer service, finance, and sales.

For example, here's what generative AI can do for optimization:

* Extract and summarize data from any source for knowledge search functions
* Evaluate and optimize different scenarios for cost reduction in areas like marketing, advertising, finance, and logistics.
* Generate synthetic data to create labeled data for supervised learning and other ML processes.

### **Boosts employee productivity**

Generative AI models can augment employee workflows and act as efficient assistants for everyone in your organization. They can do everything from searching to creation in a human-like way.

Generative AI can boost productivity for different kinds of workers:

* Support creative tasks by generating multiple prototypes based on certain inputs and constraints. It can also optimize existing designs based on human feedback and specified constraints.
* Generate new software [code suggestions](https://aws.amazon.com/what-is/ai-coding/) for application development tasks.
* Support management by generating reports, summaries, and projections.
* Generate new sales scripts, email content, and blogs for marketing teams

You can save time, reduce costs, and enhance efficiency across your organization.

## How did generative AI technology evolve?

Primitive generative models have been used for decades in statistics to aid in numerical data analysis. Neural networks and deep learning were recent precursors for modern generative AI. Variational autoencoders, developed in 2013, were the first deep generative models that could generate realistic images and speech.

### **VAEs**

VAEs introduced the capability to create novel variations of multiple data types. This led to the rapid emergence of other generative AI models like generative adversarial networks and diffusion models. These innovations were focused on generating data that increasingly resembled real data despite being artificially created.

### **Transformers**

In 2017, a further shift in AI research occurred with the introduction of transformers. Transformers seamlessly integrated the encoder-and-decoder architecture with an attention mechanism. They streamlined the training process of language models with exceptional efficiency and versatility. Notable models like GPT emerged as foundational models capable of pretraining on extensive corpora of raw text and fine-tuning for diverse tasks.

Transformers changed what was possible for natural language processing. They empowered generative capabilities for tasks ranging from translation and summarization to answering questions.

### **The future**

Many generative AI models continue to make significant strides and have found cross-industry applications. Recent innovations focus on refining models to work with proprietary data. Researchers also want to create text, images, videos, and speech that are more and more human-like.

## How does generative AI work?

Like all artificial intelligence, generative AI works by using machine learning models—very large models that are pre-trained on vast amounts of data.

### **Foundation models**

[Foundation models](https://aws.amazon.com/what-is/foundation-models/) (FMs) are ML models trained on a broad spectrum of generalized and unlabeled data. They are capable of performing a wide variety of general tasks.

FMs are the result of the latest advancements in a technology that has been evolving for decades. In general, an FM uses learned patterns and relationships to predict the next item in a sequence.

For example, with image generation, the model analyzes the image and creates a sharper, more clearly defined version of the image. Similarly, with text, the model predicts the next word in a string of text based on the previous words and their context. It then selects the next word using probability distribution techniques.

### **Large language models**

[Large language models](https://aws.amazon.com/what-is/large-language-model/) (LLMs) are one class of FMs. For example, OpenAI's generative pre-trained transformer (GPT) models are LLMs. LLMs are specifically focused on language-based tasks such as such as summarization, text generation, classification, open-ended conversation, and information extraction.

[Read about GPT](https://aws.amazon.com/what-is/gpt/)

What makes LLMs special is their ability to perform multiple tasks. They can do this because they contain many parameters that make them capable of learning advanced concepts.

An LLM like GPT-3 can consider billions of parameters and has the ability to generate content from very little input. Through their pretraining exposure to internet-scale data in all its various forms and myriad patterns, LLMs learn to apply their knowledge in a wide range of contexts.

## How do generative AI models work?

Traditional machine learning models were discriminative or focused on classifying data points. They attempted to determine the relationship between known and unknown factors. For example, they look at images—known data like pixel arrangement, line, color, and shape—and map them to words—the unknown factor. Mathematically, the models worked by identifying equations that could numerically map unknown and known factors as x and y variables.

Generative models take this one step further. Instead of predicting a label given some features, they try to predict features given a certain label. Mathematically, generative modeling calculates the probability of x and y occurring together. It learns the distribution of different data features and their relationships.

For example, generative models analyze animal images to record variables like different ear shapes, eye shapes, tail features, and skin patterns. They learn features and their relations to understand what different animals look like in general. They can then recreate new animal images that were not in the training set.

Next, we give some broad categories of generative AI models.

### **Diffusion models**

Diffusion models create new data by iteratively making controlled random changes to an initial data sample. They start with the original data and add subtle changes (noise), progressively making it less similar to the original. This noise is carefully controlled to ensure the generated data remains coherent and realistic.

After adding noise over several iterations, the diffusion model reverses the process. Reverse denoising gradually removes the noise to produce a new data sample that resembles the original.

### **Generative adversarial networks**

The [generative adversarial network](https://aws.amazon.com/what-is/gan/) (GAN) is another generative AI model that builds upon the diffusion model’s concept.

GANs work by training two neural networks in a competitive manner. The first network, known as the generator, generates fake data samples by adding random noise. The second network called the discriminator, tries to distinguish between real data and the fake data produced by the generator.

During training, the generator continually improves its ability to create realistic data while the discriminator becomes better at telling real from fake. This adversarial process continues until the generator produces data that is so convincing that the discriminator can't differentiate it from real data.

GANs are widely used in generating realistic images, style transfer, and data augmentation tasks.

### **Variational autoencoders**

Variational autoencoders (VAEs) learn a compact representation of data called latent space. The latent space is a mathematical representation of the data. You can think of it as a unique code representing the data based on all its attributes. For example, if studying faces, the latent space contains numbers representing eye shape, nose shape, cheekbones, and ears.

VAEs use two neural networks—the encoder and the decoder. The encoder neural network maps the input data to a mean and variance for each dimension of the latent space. It generates a random sample from a Gaussian (normal) distribution. This sample is a point in the latent space and represents a compressed, simplified version of the input data.

The decoder neural network takes this sampled point from the latent space and reconstructs it back into data that resembles the original input. Mathematical functions are used to measure how well the reconstructed data matches the original data.

### **Transformer-based models**

The transformer-based generative AI model builds upon the encoder and decoder concepts of VAEs. Transformer-based models add more layers to the encoder to improve performance on text-based tasks like comprehension, translation, and creative writing.

Transformer-based models use a self-attention mechanism. They weigh the importance of different parts of an input sequence when processing each element in the sequence.

Another key feature is that these AI models implement contextual embeddings. The encoding of a sequence element depends not only on the element itself but also on its context within the sequence.

#### **How transformer-based models work**

To understand how transformer-based models work, imagine a sentence as a sequence of words.

Self-attention helps the model focus on the relevant words as it processes each word. The transformer-based generative model employs multiple encoder layers called attention heads to capture different types of relationships between words. Each head learns to attend to different parts of the input sequence, allowing the model to simultaneously consider various aspects of the data.

Each layer also refines the contextual embeddings, making them more informative and capturing everything from grammar syntax to complex semantic meanings.

## What are the limitations of generative AI?

Despite their advancements, generative AI systems can sometimes produce inaccurate or misleading information. They rely on patterns and data they were trained on and can reflect biases or inaccuracies inherent in that data. Other concerns related to training data include

### **Security**

Data privacy and security concerns arise if proprietary data is used to customize generative AI models. Efforts must be made to ensure that the generative AI tools generate responses that limit unauthorized access to proprietary data. Security concerns also arise if there is a lack of accountability and transparency in how AI models make decisions.

[Learn about the secure approach to generative AI using AWS](https://aws.amazon.com/ai/generative-ai/security/)

### **Creativity**

While generative AI can produce creative content, it often lacks true originality. The creativity of AI is bounded by the data it has been trained on, leading to outputs that may feel repetitive or derivative. Human creativity, which involves a deeper understanding and emotional resonance, remains challenging for AI to replicate fully.

### **Cost**

Training and running generative AI models require substantial computational resources. Cloud-based generative AI models are more accessible and affordable than trying to build new models from scratch.

### **Explainability**

Due to their complex and opaque nature, generative AI models are often considered black boxes. Understanding how these models arrive at specific outputs is challenging. Improving interpretability and transparency is essential to increase trust and adoption.

## What are the best practices in generative AI adoption?

If your organization wants to implement generative AI solutions, consider the following best practices to enhance your efforts.

### **Begin with internal applications**

It’s best to start generative AI adoption with internal application development, focusing on process optimization and employee productivity. You get a more controlled environment to test outcomes while building skills and understanding of the technology. You can test the models extensively and even customize them on internal knowledge sources.

This way, your customers have a much better experience when you eventually use the models for external applications.

### **Enhance transparency**

Clearly communicate about all generative AI applications and outputs so your users know they are interacting with AI and not humans. For instance, the AI can introduce itself as AI, or AI-based search results can be marked and highlighted.

That way, your users can use their own discretion when they engage with the content. They may also be more proactive in dealing with any inaccuracies or hidden biases the underlying models may have because of their training data limitations.

### **Implement security**

Implement guardrails so your generative AI applications don't allow inadvertent unauthorized access to sensitive data. Involve security teams from the start so all aspects can be considered from the beginning. For example, you may have to mask data and remove personally identifiable information (PII) before you train any models on internal data.

### **Test extensively**

Develop automated and manual testing processes to validate results and test all types of scenarios the generative AI system may experience. Have different groups of beta testers who try out the applications in different ways and document results. The model will also improve continuously through testing, and you get more control over expected outcomes and responses.

## What is generative AI training for beginners?

Generative AI training begins with understanding foundational machine learning concepts. Learners also have to explore neural networks and AI architecture. Practical experience with Python libraries such as TensorFlow or PyTorch is essential for implementing and experimenting with different models. You also have to learn model evaluation, fine tuning and prompt engineering skills.

A degree in artificial intelligence or machine learning provides in-depth training. Consider online short courses and certifications for professional development. [Generative AI training on AWS](https://aws.amazon.com/training/learn-about/generative-ai/) includes certifications by AWS experts on topics like:

* [Introduction to generative AI](https://explore.skillbuilder.aws/learn/course/external/view/elearning/17176/introduction-to-generative-ai-art-of-the-possible)
* [Generative AI for executives](https://explore.skillbuilder.aws/learn/course/external/view/elearning/16666/generative-ai-for-executives)
* [Generative AI essentials for business](https://training.resources.awscloud.com/partner-ai-ml-training/aws-partner-generative-ai-essentials-business)

## How can AWS help Generative AI?

Amazon Web Services (AWS) makes it easy to build and scale generative AI applications for your data, use cases, and customers. With [generative AI on AWS](https://aws.amazon.com/ai/generative-ai/), you get enterprise-grade security and privacy, access to industry-leading FMs, generative AI-powered applications, and a data-first approach.

Choose from a range of [generative AI services](https://aws.amazon.com/ai/generative-ai/services/) that support all types of organizations in every stage of generative AI adoption and maturity:

* Code generation is one of the most promising applications for generative AI. With [Amazon Q Developer](https://aws.amazon.com/q/developer/), an generative AI- powered assistant for software developement, you can get great results in developer productivity.
* [Amazon Bedrock](https://aws.amazon.com/bedrock/) is another fully managed service that offers a choice of high-performing FMs and a broad set of capabilities. You can easily experiment with various top FMs, privately customize them with your data, and create managed agents that execute complex business tasks.
* You can also use [Amazon SageMaker JumpStart](https://aws.amazon.com/sagemaker/jumpstart/) to discover, explore, and deploy open source FMs—or even create your own. SageMaker JumpStart provides managed infrastructure and tools to accelerate scalable, reliable, and secure model building, training, and deployment.
* [AWS HealthScribe](https://aws.amazon.com/healthscribe/) is a HIPAA-eligible service empowering healthcare software vendors to build clinical applications that automatically generate clinical notes by analyzing patient-clinician conversations. AWS HealthScribe combines speech recognition and generative artificial intelligence (AI) to reduce the burden of clinical documentation by transcribing patient-clinician conversations and generating easier-to-review clinical notes.
* [Amazon Q in QuickSight](https://aws.amazon.com/quicksight/q/) helps business analysts easily create and customize visuals using natural-language commands. The new [Generative BI authoring capabilities](https://aws.amazon.com/blogs/business-intelligence/generative-bi-dashboard-authoring-capabilities-now-available-in-preview-for-amazon-quicksight-q-customers/) extend the natural-language querying of QuickSight Q beyond answering well-structured questions (such as “What are the top 10 products sold in California?”) to help analysts quickly create customizable visuals from question fragments (such as “Top 10 products”), clarify the intent of a query by asking follow-up questions, refine visualizations, and complete complex calculations.