What is GPT?

Generative Pre-trained Transformers, commonly known as GPT, are a family of neural network models that uses the transformer architecture and is a key advancement in artificial intelligence (AI) powering generative AI applications such as ChatGPT. GPT models give applications the ability to create human-like text and content (images, music, and more), and answer questions in a conversational manner. Organizations across industries are using GPT models and generative AI for Q&A bots, text summarization, content generation, and search.

## What is GPT?

Generative Pre-trained Transformers, commonly known as GPT, are a family of neural network models that uses the transformer architecture and is a key advancement in artificial intelligence (AI) powering generative AI applications such as ChatGPT. GPT models give applications the ability to create human-like text and content (images, music, and more), and answer questions in a conversational manner. Organizations across industries are using GPT models and generative AI for Q&A bots, text summarization, content generation, and search.

## Why is GPT important?

The GPT models, and in particular, the transformer architecture that they use, represent a significant AI research breakthrough. The rise of GPT models is an inflection point in the widespread adoption of ML because the technology can be used now to automate and improve a wide set of tasks ranging from language translation and document summarization to writing blog posts, building websites, designing visuals, making animations, writing code, ]researching complex topics, and even composing poems. The value of these models lies in their speed and the scale at which they can operate. For example, where you might need several hours to research, write, and edit an article on nuclear physics, a GPT model can produce one in seconds. GPT models have sparked the research in AI towards achieving artificial general intelligence, which means machines can help organizations reach new levels of productivity and reinvent their applications and customer experiences.

## What are the use cases of GPT?

The GPT models are general-purpose language models that can perform a broad range of tasks from creating original content to write code, summarizing text, and extracting data from documents.

Here are some ways you can use the GPT models:

### Create social media content

Digital marketers, assisted by artificial intelligence (AI), can create content for their social media campaigns. For example, marketers can prompt a GPT model to produce an explainer video script. GPT-powered image processing software can create memes, videos, marketing copy, and other content from text instructions.

### Convert text to different styles

GPT models generate text in casual, humorous, professional, and other styles. The models allow business professionals to rewrite a particular text in a different form. For example, lawyers can use a GPT model to turn legal copies into simple explanatory notes.

### Write and learn code

As language models, the GPT models can understand and write computer code in different programming languages. The models can help learners by explaining computer programs to them in everyday language. Also, experienced developers can use GPT tools to autosuggest relevant code snippets.

### Analyze data

The GPT model can help business analysts efficiently compile large volumes of data. The language models search for the required data and calculate and display the results in a data table or spreadsheet. Some applications can plot the results on a chart or create comprehensive reports.

### Produce learning materials

Educators can use GPT-based software to generate learning materials such as quizzes and tutorials. Similarly, they can use GPT models to evaluate the answers.

### Build interactive voice assistants

The GPT models allow you to build intelligent interactive voice assistants. While many chatbots only respond to basic verbal prompts, the GPT models can produce chatbots with conversational AI capabilities. In addition, these chatbots can converse verbally like humans when paired with other AI technologies.

## How does GPT work?

Though it’s accurate to describe the GPT models as artificial intelligence (AI), this is a broad description. More specifically, the GPT models are neural network-based language prediction models built on the Transformer architecture. They analyze natural language queries, known as prompts, and predict the best possible response based on their understanding of language.

To do that, the GPT models rely on the knowledge they gain after they’re trained with hundreds of billions of parameters on massive language datasets. They can take input context into account and dynamically attend to different parts of the input, making them capable of generating long responses, not just the next word in a sequence. For example, when asked to generate a piece of Shakespeare-inspired content, a GPT model does so by remembering and reconstructing new phrases and entire sentences with a similar literary style.

There are different types of neural networks, like recurrent and convolutional. The GPT models are transformer neural networks. The transformer neural network architecture uses self-attention mechanisms to focus on different parts of the input text during each processing step. A transformer model captures more context and improves performance on natural language processing (NLP) tasks. It has two main modules, which we explain next.

[Read about neural networks »](https://aws.amazon.com/what-is/neural-network/)

[Read about natural language processing (NLP) »](https://aws.amazon.com/what-is/nlp/)

### Encoder

Transformers pre-process text inputs as embeddings, which are mathematical representations of a word. When encoded in vector space, words that are closer together are expected to be closer in meaning. These embeddings are processed through an encoder component that captures contextual information  from an input sequence. When it receives input, the transformer network’s encoder block separates words into embeddings and assigns weight to each. Weights are parameters to indicate the relevance of words in a sentence.

Additionally, position encoders allow GPT models to prevent ambiguous meanings when a word is used in other parts of a sentence. For example, position encoding allows the transformer model to differentiate the semantic differences between these sentences:

* A dog chases a cat
* A cat chases a dog

So, the encoder processes the input sentence and generates a fixed-length vector representation, known as an embedding. This representation is used by the decoder module.

### Decoder

The decoder uses the vector representation to predict the requested output. It has built-in self-attention mechanisms to focus on different parts of the input and guess the matching output. Complex mathematical techniques help the decoder to estimate several different outputs and predict the most accurate one.

Compared to its predecessors, like recurrent neural nets, transformers are more parallelizable because they do not process words sequentially one at a time, but instead, process the entire input all at once during the learning cycle. Due to this and the thousands of hours engineers spent fine-tuning and training the GPT models, they’re able to give fluent answers to almost any input you provide.

## How was GPT-3 trained?

In a published research paper, researchers described generative pretraining as the ability to train language models with unlabeled data and achieve accurate prediction. The first GPT model, GPT-1, was developed in 2018. GPT-4 was introduced in March 2023 as a successor to GPT-3.

GPT-3 was trained with over 175 billion parameters or weights. Engineers trained it on over 45 terabytes of data from sources like web texts, Common Crawl, books, and Wikipedia. Prior to training, the average quality of the datasets was improved as the model matured from version 1 to version 3.

GPT-3 trained in a semi-supervised mode. First, machine learning engineers fed the deep learning model with the unlabeled training data. GPT-3 would understand the sentences, break them down, and reconstruct them into new sentences. In unsupervised training, GPT-3 attempted to produce accurate and realistic results by itself. Then, machine learning engineers would fine-tune the results in supervised training, a process known as reinforcement learning with human feedback (RLHF).

You can use the GPT models without any further training, or you can customize them with a few examples for a particular task.

## What are examples of some applications that use GPT?

Since its launch, the GPT models have brought artificial intelligence (AI) to numerous applications in various industries. Here are some examples:

* GPT models can be used to analyze customer feedback and summarize it in easily understandable text. First, you can collect customer sentiment data from sources like surveys, reviews, and live chats, then you can ask a GPT model to summarize the data.
* GPT models can be used to enable virtual characters to converse naturally with human players in virtual reality.
* GPT models can be used to provide a better search experience for help desk personnel. They can query the product knowledge base with conversational language to retrieve relevant product information.

## How can AWS help you run large language models like GPT-3?

[Amazon Bedrock](https://aws.amazon.com/bedrock/) is the easiest way to build and scale [generative AI](https://aws.amazon.com/generative-ai/) applications with large language models, also known as foundation models (FMs), similar to GPT-3. Amazon Bedrock gives you access via an API to foundation models from leading AI startups, including AI21 Labs, Anthropic, and Stability AI—along with Amazon’s newest foundation model family, [Amazon Titan](https://aws.amazon.com/bedrock/titan/) FMs. With Bedrock’s serverless experience, you can get started quickly, privately customize FMs with your own data, and easily integrate and deploy them into your applications using the AWS tools and capabilities you are familiar with (including integrations with Amazon SageMaker ML features like Experiments to test different models and Pipelines to manage your FMs at scale) without having to manage any infrastructure. Learn more about [building with foundation models on Amazon Bedrock](https://aws.amazon.com/bedrock/).