Introduction to Generative Artificial Intelligence

AWS Educate

Machine learning's new path

Over the last several years, machine learning, or ML, has rapidly expanded the capabilities of artificial intelligence in the world of IT. Today, that expansion has entered the domain of generative artificial intelligence, or generative AI. Generative AI is making it easier to innovate faster and reduce the number of hours needed for development. This provides you with more time to grow your business.

Course objectives

In this course, you will be introduced to the domain of generative AI. By the end of this course you will be able to do the following:

- Define generative artificial intelligence
- Describe foundation models
- Discuss use cases for generative AI
- Identify the features and benefits of AWS generative AI services.

What is generative AI?

Generative AI is a type of artificial intelligence that can create new content and ideas, including conversations, stories, images, videos, and music. AI generators are powered by machine learning foundation models. These models are capable of producing content, so you don't have to. This AI-generated content can be edited, so you can make the necessary modifications that meet your needs.

How generative AI differs from traditional AI

Generative AI is a subset of machine learning. To help you understand the difference between traditional machine learning and generative AI, let's review some key differences. Traditional machine learning models perform tasks based on data that you provide. They can make predictions such as ranking, sentiment analysis, image classification, and more. However, each model can perform only one task. And to successfully do it, models need to be carefully trained on the data. As models are trained, they analyze the data and look for patterns. Then, these models make a prediction based on these patterns.

With generative AI, models are pre-trained on massive amounts of general domain data, beyond the data that you provide. These models can perform multiple tasks. Based on prompts, such as code comments, the model predicts what the outcome should be and then actually generates that content. The AI-generated content comes from learning patterns and relationships that allow the model to predict the desired outcome.

Foundation models

At the core of every AI generator is a foundation model. Foundation models are a class of powerful machine learning models that are differentiated by their ability to be pre-trained on vast amounts of data in order to perform a wide range of downstream tasks. These tasks include text generation, data summarization, information extraction, question and answer responses, and chatbot interactions. In contrast, traditional ML models are trained to perform a specific task from a data set.

Comparisons of traditional and foundational models

Let’s review some key differences between traditional machine learning models and foundation models. Traditional ML models are typically siloed to perform specific tasks, like analyzing text for sentiment, classifying images, and forecasting trends. In order to achieve each task, customers need to gather labeled data, train a model, and deploy that model.

The size and general-purpose nature of foundation models make them very different from traditional ML models. With foundation models, instead of gathering labeled data and training multiple models, you use the same pre-trained foundation model to adapt several tasks. Foundation models can also be customized to perform domain-specific functions that are differentiating to their businesses. This uses only a small fraction of the data and compute required to train a model from scratch.

Introduction to Generative Artificial Intelligence

AWS Educate

Prompt engineering

Generative AI architectures use prompt engineering to initiate an action from a foundation model. Prompt engineering is the process of designing and refining the prompts or input stimuli for a language model to generate specific types of output. Prompt engineering involves selecting appropriate keywords, providing context, and shaping the input in a way that encourages the model to produce the desired response. It is a vital technique to actively shape the behavior and output of foundation models.

How prompt engineering works

Prompt engineering begins with a prompt. This is the text that you input into the model. The model then generates a response to the prompt. This is known as inference. The result of the inference is displayed as the output. In this example, your prompt is: Where is Japan located? The model's inference on the prompt returns an output that says, "Japan is located in the northwest Pacific Ocean." The model might not always give you the output that you need. So, you might need to alter the prompt or provide the model with examples of the tasks in the prompt.

Prompt engineering IT example

Now, let's look at an example that's helping IT professionals save time in development. In this example, you're using Amazon Q Developer to code. Amazon Q Developer uses prompt engineering with a foundation model to generate code.

First, in your integrated development environment, or IDE, you type a natural language prompt for the code that you want to generate. For this example, you wrote, "write a function to verify an email address." Your foundation model generates options for functions that you might use. You can use your right and left arrows keys to review the AI-generated options. Then, you can press Tab to choose the one that best suits your needs. In this example, you choose the first choice. When you press Enter, the foundation model generates a common option to build on to this function. In this example, an if else statement is generated. You can choose whether you want to use this additional code or not. In this example, you choose to use it. Any of the generated code that you use is fully editable. You can understand now from this example how generative AI is a time saver for development.

Foundation model types

When you build your application, it’s important to choose the right type of foundation model. Review each model type to learn more about different types of foundation models.

- Text-to-text: Text-to-text foundation models (FMs) are built for natural language processing tasks that are developed by the ML research community, startups, and established companies. Many organizations have had a chance to experiment with generative large language models (LLMs) that can do things like summarize text, extract information, respond to questions, and create content. These models take a user’s input text and extend it with new, generated text (for example, sentence auto-completion).

- Text-to-embeddings: Text-to-embeddings FMs are another type of LLM that can compare pieces of text (inputs), like what a user types into a search bar with indexed data, and makes a connection between the two. Amazon’s product search uses this type of LLM to compare a user’s ask with catalog data and present the user with more accurate and relevant results.

- Multimodal: Multimodal FMs can understand and generate different formats, such as text and images. These types of FMs can generate images based on natural language prompts. An example of this is Stable Diffusion.

Generative AI use cases

Generative AI can be used for a broad range of topics. Review each use case to learn more about generative AI use cases.

Enhance Customer Experiences:

- Chatbots and virtual assistants: Streamline customer self-service processes and reduce operational costs by automating responses for customer service queries through generative AI-powered chatbots, voice bots, and virtual assistants.
- Agent assist and call analytics: Concisely summarize customer conversations to reduce the time that agents and supervisors spend taking and reviewing notes or sharing context when transferring contacts. Analyze customer interactions to derive insights and monitor agent performance.
- Personalization: Deliver better personalized experiences and increase customer engagement with personalized offerings and communications.

Boost employee productivity

- Conversational search: Improve employee productivity by quickly and easily finding accurate information and summarizing content through a conversational interface.
- Code generation: Accelerate application development with code suggestions based on the developer’s comments and code.
- Automated report generation: Generative AI can be used to automatically generate financial reports, summaries, and projections, saving time and reducing errors.

Optimize processes

- Intelligent document processing: Improve business operations by automatically extracting and summarizing data from documents and insights through generative AI-powered question and answering.

- Data augmentation: Generate synthetic data to train ML models when the original dataset is small, imbalanced or sensitive.
- Supply chain optimization: Improve logistics and reduce costs by evaluating and optimizing different supply chain scenarios.

Enhance creativity and content creation

- AI-generated marketing content: Create engaging marketing content, such as blog posts, social media updates, or email newsletters, saving time and resources.
- AI-generated sales content, guidance, and enablement: Generate personalized emails and messages based on a prospect's profile and behavior, improving response rates. Generate sales scripts based on the customer's segment, industry, and the product or service.
- New product development: Generate multiple design prototypes based on certain inputs and constraints, speeding up the ideation phase, or optimize existing designs based on user feedback and specified constraints.

Generative AI industry use cases

Generative AI is being applied to almost every industry. Review each industry use case to learn how some of the top industries are using generative AI.

Health care:

- Ambient digital scribe: Automatically create transcripts, extract key details, and create summaries from clinician-patient interactions.
- Interpret medical images: Enhance, reconstruct, or even generate medical images like X-rays, MRIs, or CT scans, which can aid in better diagnosis.
- Personalized medicine: Based on a patient's genetics, lifestyle, and symptoms, generative AI can create personalized treatment plans.

- Intelligent health assist: Enable agent assist, call report summarization, and agent performance assessment for health care insurance providers.
- Automate medical coding: Automate medical coding of medical claims to reduce the timeframe for billing, errors, and administrative tasks and to meet regulatory and compliance requirements.

Life sciences

- Clinical development: Analyze large data sets to identify potential adverse drug reactions for both clinical and in-market drugs.
- Drug discovery: Use generative AI tools for protein folding, protein sequence design, and docking and molecule design to accelerate drug discovery and the design process while reducing costs.
- Enhance clinical trials: Augment and accelerate clinical trials by rapidly synthesizing vast amounts of combinatorial trial data, simulating patient populations, and optimizing protocol design.
- Automated research reporting: Generate documents and narratives based on drug discovery research dataset, such as proprietary scientific reports. For example, collate phase 1 clinical trials for a candidate therapeutic.
- Optimized trial enrollment: Use generative AI to match patients to clinical trials based on inclusion and exclusion criteria. For example, determine whether the patient is eligible based on co-morbidities.

Financial services

- AI-managed portfolios: Employ generative AI to create highly tailored investment strategies and portfolios aligned to specific financial goals and risk profiles.

- Increase the business value of unstructured content: Create on-demand structured data products (competitor maps, supply chain relationships, or product and service catalogs) from large unstructured data sources such as emails, document repositories, and filings.
- Drive product innovation and automate business processes: Use generative AI to develop new tools for end-users, like stock screening using natural language search. Examples include wealth management and brokerage clients and advisors and institutional investment analysis.
- Intelligent advisory: With chatbots and call center assist, firms can automatically translate complex questions from internal users and external customers into their semantic meaning, analyze for context, and then generate highly accurate and conversational responses.
- Transform financial documentation: Quickly draft investment research, loan documentation, insurance policies, regulatory communications, request for information, business correspondence, and more.

Manufacturing

- Product design optimization: Generative AI can quickly generate and assess countless design options, helping manufactures find the most optimized, efficient, and cost-effective solutions.
- Operational efficiency: Generative AI can simulate production to identify improvements, find hidden insights, validate models with synthetic data, and boost predictive accuracy, all without disrupting operations.
- Real-time equipment diagnostics: By ingesting historical data, generative AI can diagnose equipment failures in real time and recommend maintenance actions like input adjustments, repairs, or likely spare parts.
- Supply chain traceability: Gain end-to-end traceability of component parts through multi-tier supply chains and identify anomalies or gaps in the supply chain data.

- AI-powered maintenance assistants: Generative conversational agents can be trained on product manuals, troubleshooting guides, and maintenance notes to deliver swift technical support to workers, reducing downtime.

Retail

- Product design optimization: Generative AI can quickly generate and assess countless design options, helping manufactures find the most optimized, efficient, and cost-effective solutions.
- Operational efficiency: Generative AI can simulate production to identify improvements, find hidden insights, validate models with synthetic data, and boost predictive accuracy, all without disrupting operations.
- Real-time equipment diagnostics: By ingesting historical data, generative AI can diagnose equipment failures in real time and recommend maintenance actions like input adjustments, repairs, or likely spare parts.
- Supply chain traceability: Gain end-to-end traceability of component parts through multi-tier supply chains and identify anomalies or gaps in the supply chain data.
- AI-powered maintenance assistants: Generative conversational agents can be trained on product manuals, troubleshooting guides, and maintenance notes to deliver swift technical support to workers, reducing downtime.

Media and entertainment

- Produce high-quality content at scale: Generate characters, animations, and visual effects tailored to specific themes, genres, or formats.
- Optimize subscriber experiences: Create effective, personalized content that adapts in real time based on user engagement and preference.
- Enrich broadcast content: Enhance live broadcast content through automated graphics, speech, and video generation tailored to each program.

- Automated highlight generation: For sports, generative AI can detect highlights and automatically generate polished packages and promos.
- Automatic content tagging: Use generative AI to auto-tag and index massive media libraries for easier search and recommendations.

Benefits of using AWS for your generative AI solutions

As the leader in cloud computing, AWS is also leading the way in providing the best platform for all your generative AI needs. Review each topic to learn more about the challenges generative AI builders face and the solutions AWS offers.

Flexibility

Challenge: Customers need a straightforward way to find and access high-performing FMs that are best suited for their business.

AWS solution: AWS offers a wide selection of FMs built by Amazon and top AI startups, including AI21 Labs, Anthropic, and Stability AI.

Secure customization

Challenge: Customers want to easily take the base FM and build differentiated applications using their own data. They need their data to stay completely protected, secure, and private.

AWS solution: Using Amazon Bedrock, customers can fine tune models for a particular task without having to annotate large volumes of data (as few as 20 examples is enough). No customer data is used to train the underlying models. All data is encrypted and stays within the customer's virtual private cloud (VPC).

Cost-effective infrastructure

Challenge: To fully use FMs, customers need the most high-performing, cost-effective infrastructure purpose-built for ML.

AWS solution: AWS offers the best price performance for generative AI with infrastructure powered by AWS-designed ML chips and NVIDIA GPUs. With AWS, customers can cost-effectively scale infrastructure to train and run FMs containing hundreds of billions of parameters.

Builder friendly

Challenge: Customers want ease of use. They want to quickly integrate and deploy FMs into their applications and workloads running on AWS.

AWS solution: Using AWS generative AI capabilities, customers don't need to send their data to the model. Instead, they can bring the model to their data using familiar controls and integrations and services such as Amazon SageMaker and Amazon Simple Storage Service (Amazon S3).

AWS services

Challenge: Customers are looking for generative AI solutions to help improve productivity while seamlessly interacting with applications and systems.

AWS solution: With generative AI built in, services such as Amazon Q Developer can help customers improve productivity. They can also deploy generative AI solutions, such as call summarization and question answering, that combine AWS AI services with leading FMs.

AWS generative AI services

Review each option to learn more about the generative AI services that AWS offers.

Amazon Bedrock: Amazon Bedrock is a fully managed service that makes FMs from Amazon and leading AI startups available through an API. This means you can choose from various FMs to find the model that's best suited for your use case. Amazon Bedrock makes it easier for developers to create generative AI applications that can deliver up-to-date answers based on proprietary knowledge sources. They can also complete tasks for a wide range of use cases.

Amazon Q Developer: Amazon Q Developer is an AI coding companion that generates real-time, single-line or full-function code suggestions in your IDE to help you quickly build software. With Amazon Q Developer, you can write a comment in natural language that outlines a specific task in English, such as “Upload a file with server-side encryption.” Based on this information, Amazon Q Developer recommends one or more code snippets directly in the IDE that can accomplish the task. You can quickly and easily accept the top suggestion (Tab key), view more suggestions (arrow keys), or continue writing your own code. You should always review a code suggestion before accepting them, and you might need to edit it to ensure it does exactly what you intended.

AWS Inferentia: AWS Inferentia is a custom machine learning chip designed by AWS that you can use for high-performance inference predictions. In order to use the chip, set up an Amazon Elastic Compute Cloud instance and use the AWS Neuron software development kit (SDK) to invoke the Inferentia chip. To provide customers with the best Inferentia experience, Neuron has been built into the AWS Deep Learning AMIs (DLAMI). AWS Inferentia accelerators are designed by AWS to deliver high performance at the lowest cost for your deep learning (DL) inference applications.

AWS Trainium: AWS Trainium is an AWS-designed DL training accelerator that delivers high performance and cost-effective DL training on AWS. Amazon EC2 Trn1 instances, powered by AWS Trainium, deliver the highest performance on DL training of popular natural language processing (NLP) models on AWS. Trn1 instances offer up to 50% cost-to-train savings over comparable Amazon EC2 instances. Trainium has been optimized for training NLP, computer vision, and recommended models used in a broad set of applications. These applications include text summarization, code generation, question answering, image and video generation, recommendation, and fraud detection.

Amazon SageMaker JumpStart: SageMaker JumpStart helps you quickly and easily get started with ML. SageMaker JumpStart provides a set of solutions for the most common use cases that can be deployed readily in just a few steps. The solutions are fully customizable and showcase the use of AWS CloudFormation templates and reference architectures so you can accelerate your ML journey. SageMaker JumpStart also provides foundation models and supports one-step deployment and fine-tuning of more than 150 popular open-source models, such as transformer, object detection, and image classification models.

Summary

You have completed this section of the course. Next, you test what you learned with a final assessment.

Additional resources

Choose each link to learn more about using AWS services for generative AI solutions.

- What is generative AI?
- Generative AI for every business
- Announcing New Tools for Building with Generative AI on AWS
- Generative AI Innovation Center