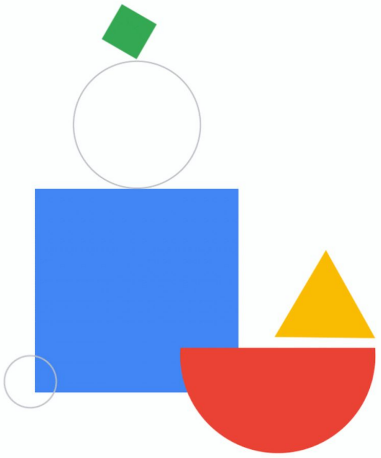
Developing Applications with Google Cloud: Foundations

Module 5: Adding Intelligence to Your Application



Welcome to Developing Applications with Google Cloud: Foundations, module 5: Adding Intelligence to Your Application.

Machine learning is about teaching machines to recognize patterns like humans do. Although even a two-year-old can easily distinguish between an apple and an orange, it's difficult to teach a computer to do the same thing. Google has developed pre-trained machine learning models and made them available to you as easy-to-use Google Cloud APIs. Now, with just a few lines of code, you can add artificial intelligence, or AI, to your own application.

Agenda

01	Using pre-trained machine learning models
02	Introduction to generative AI

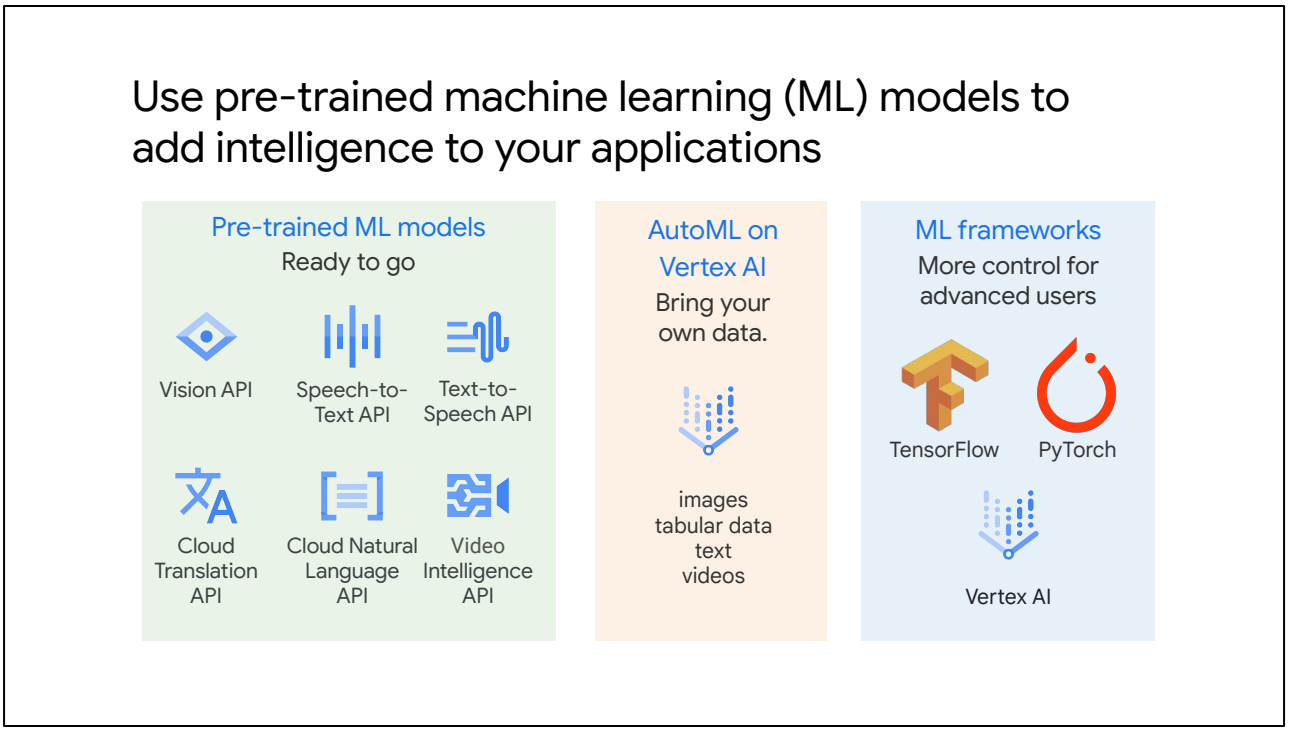
In this module, we explore Google's pre-trained machine learning APIs for vision, speech, video intelligence, and natural language processing.

Generative AI takes artificial intelligence a step further. Generative AI is a type of artificial intelligence that creates new content based on what it has learned from existing content. Generative AI can make applications more powerful and the development experience more efficient. This module will introduce you to generative AI, and explain why developers should care about it.

Agenda

01	Using pre-trained machine learning models
02	Introduction to generative AI

Pre-trained learning models can help you add powerful features to your applications.



Google Cloud offers several pre-trained machine-learning (ML) models that you can use to add intelligence to your application.

- The Vision API lets you perform complex image detection.
- The Speech-to-Text and Text-to-Speech APIs enable developers to convert audio to text and text to audio.
- The Cloud Translation API lets you translate an arbitrary string into any supported language. The Cloud Translation API is highly responsive. Websites and applications can use the Cloud Translation API for fast, dynamic translation of text from a source language to a target language.
- The Cloud Natural Language API lets you extract information about entities that are mentioned in text documents, news articles, or blog posts. You can use the API to understand sentiment about your product on social media, or parse intent from customer conversations.
- The Video Intelligence API lets you search video files to extract and label entities at the shot, frame, or video level. The API annotates videos stored in Cloud Storage and helps you identify key entities in your video and when they occur within the video.

AutoML on Vertex AI enables users with limited ML expertise to train high-quality models specific to their business needs. AutoML on Vertex AI lets you train models on images, tabular data, text, or videos without writing any code.

You can also use your own data to build and train your own custom ML models by

using frameworks like TensorFlow, PyTorch, and Vertex AI.

Invoke REST APIs to use machine learning APIs; no machine learning knowledge is required

It's really easy to invoke the REST APIs to implement machine learning in your application, no ML knowledge is required. In this example, we are using the Vision API to process an image that's stored in Cloud Storage. We invoke the REST API and send it a JSON request, and we receive a JSON response with attributes that describe the image. Let's take a look at a few examples now.

Analyze images

Label detection

Optical character recognition (OCR)

Landmark detection

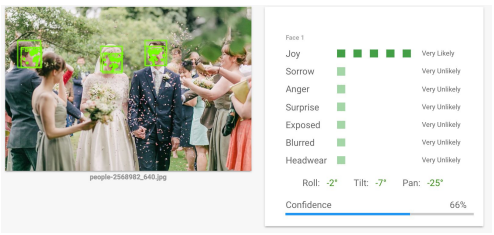
Logo detection

Face detection

Explicit content detection

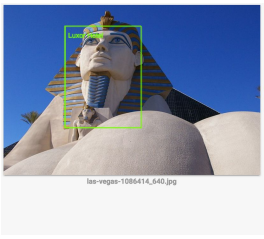
The Vision API can categorize objects under labels and perform optical character recognition, or OCR. The Vision API can detect landmarks, logos, faces, and explicit content.

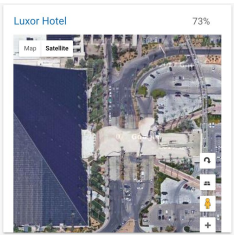
Get insight from images



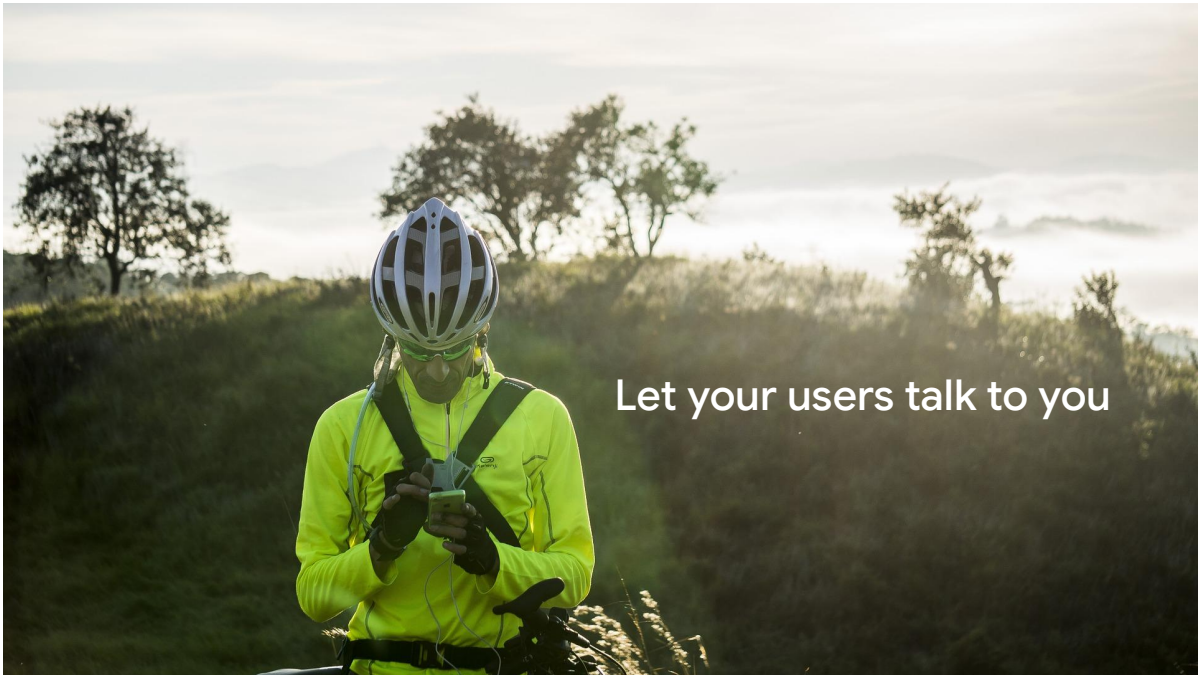
For example, the Vision API can analyze faces and return information about emotions and head wear. In the wedding picture, the API accurately returns the emotional expressions on the faces in the picture.

Get insight from images

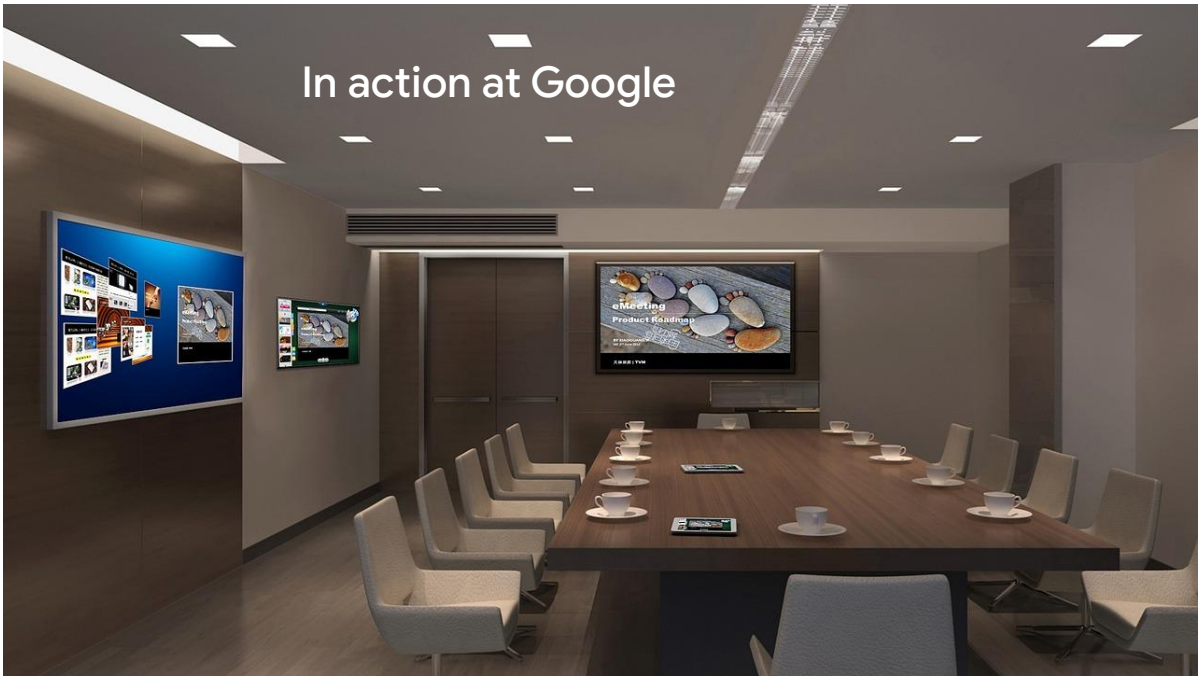




In the picture of the Sphinx, Vision API correctly detects that the image is from the Sphinx in Las Vegas and not the Sphinx in Egypt.



The Speech-to-Text API enables developers to convert audio to text. It handles 110 languages and variants to support your global user base. You can transcribe the text of users dictating to an application’s microphone, enable command-and-control through voice, transcribe audio files, and more.



Here's an example of how Google uses machine learning. Google’s conference room systems perform occupancy detection by using motion detection with the VC camera. Every 30 seconds, the system sends a Pub/Sub notification indicating whether motion was detected or not. It also sends a Pub/Sub notification when a call starts or ends.

If motion is detected between 6 and 8 minutes after the meeting start time, the room counts as occupied. Otherwise, it's empty, and is available for someone else to reserve.

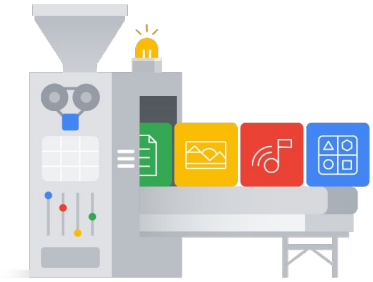
Agenda

01	Using pre-trained machine learning models
02	Intro to generative AI

One type of artificial intelligence that you can use for your applications is generative artificial intelligence, or generative AI for short.

What is generative AI?

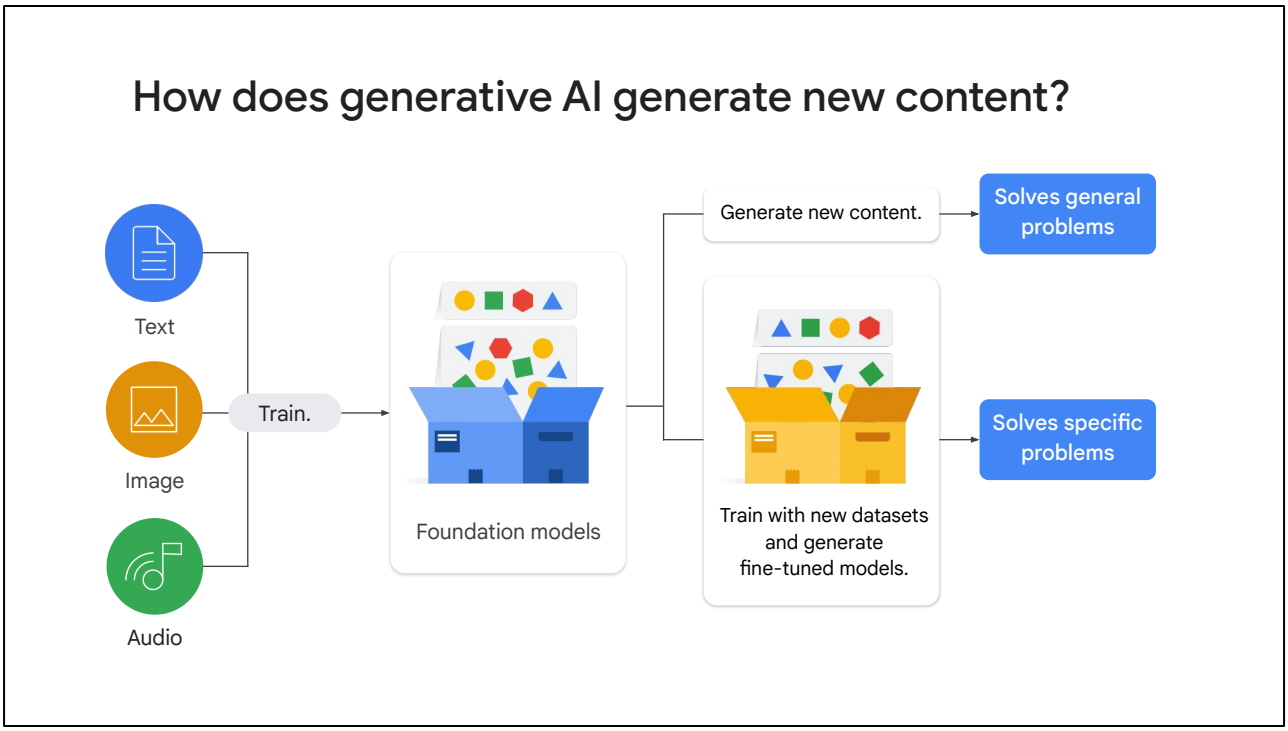
- Generative AI creates new content based on what it has learned from existing content.
- Learning from existing content ("training") results in the creation of a statistical model.
- When provided a prompt, this model can be used to predict an expected response and generate new content.



Generative AI is a type of artificial intelligence that creates new content based on what it has learned from existing content.

We call this type of learning "training." A statistical model is created by using the existing content.

You can provide an input, called a prompt, to the model, which can predict an expected response. New content can be generated based on the expected response.

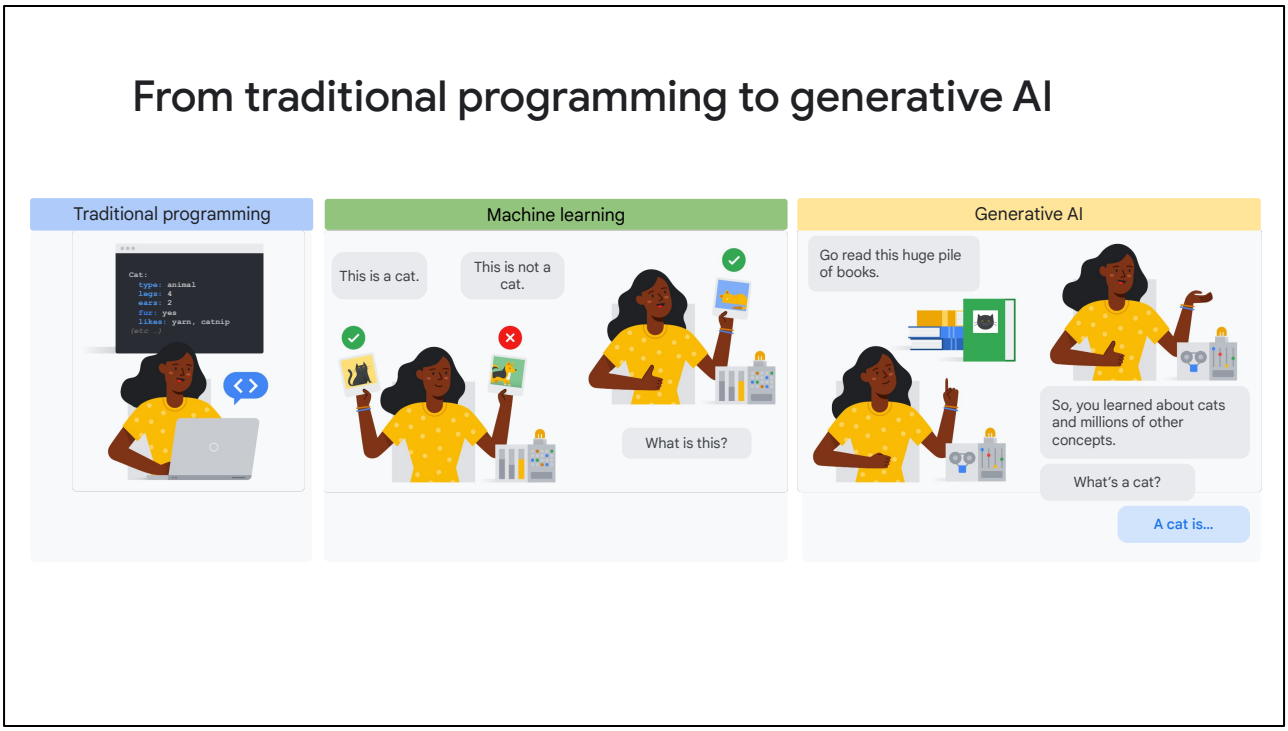


How does AI generate new content? It learns from a massive amount of existing content such as text, images, and audio.

Training results in the creation of a “foundation model.” The most popular type of foundation model is a large language model, or LLM. LLMs are trained on text data only, but other types of foundation models may be trained on other types of data, like images or programming code.

The foundation model can then be used directly to generate content and solve general problems, such as content extraction or document summarization.

The model can also be trained further with new datasets in your field to solve specific problems, such as financial model generation or healthcare consulting. This training results in the creation of a new model that is tailored to your specific needs.



How is generative AI different from traditional programming and other types of machine learning?
In traditional programming, you have to specify the rules, then the machine will act on them and return the answers.

For example, using traditional programming, you might specify these attributes of a cat:

- type: animal
- legs: 4
- ears: 2
- fur: yes
- likes: yarn, catnip

However, writing these algorithms is difficult because it's impossible to implement all possible rules.

So you need a new method: machine learning with neural networks.

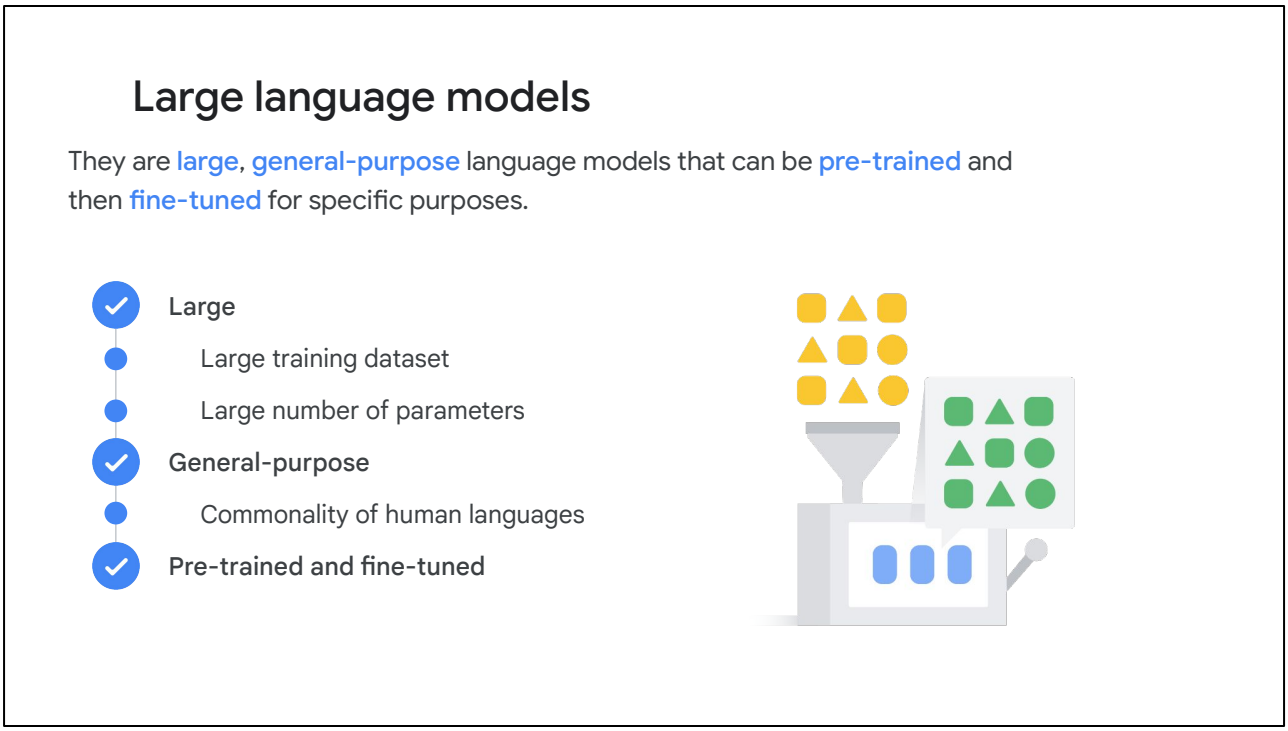
With machine learning, you feed the machine data and answers and let it discover the rules itself. For instance, you train the machine on many pictures of cats and other animals. The machine learns the pattern and predicts whether a new picture is a cat.

However, this type of learning is typically in a **narrow** field to solve a **specific** task. What if you want a machine to develop some **fundamental intelligence** to solve **general problems**?

Generative AI aims to solve this problem.

With generative AI, you feed a machine a huge amount of multimodal data. The machine learns a seemingly endless number of concepts and develops **foundation** models like an LLM.

So when you ask the machine “what’s a cat,” it can give you everything it learned about a cat.



So, what are large language models? Large language models refer to **large**, **general-purpose** language models that can be **pre-trained and then fine-tuned** for specific purposes.

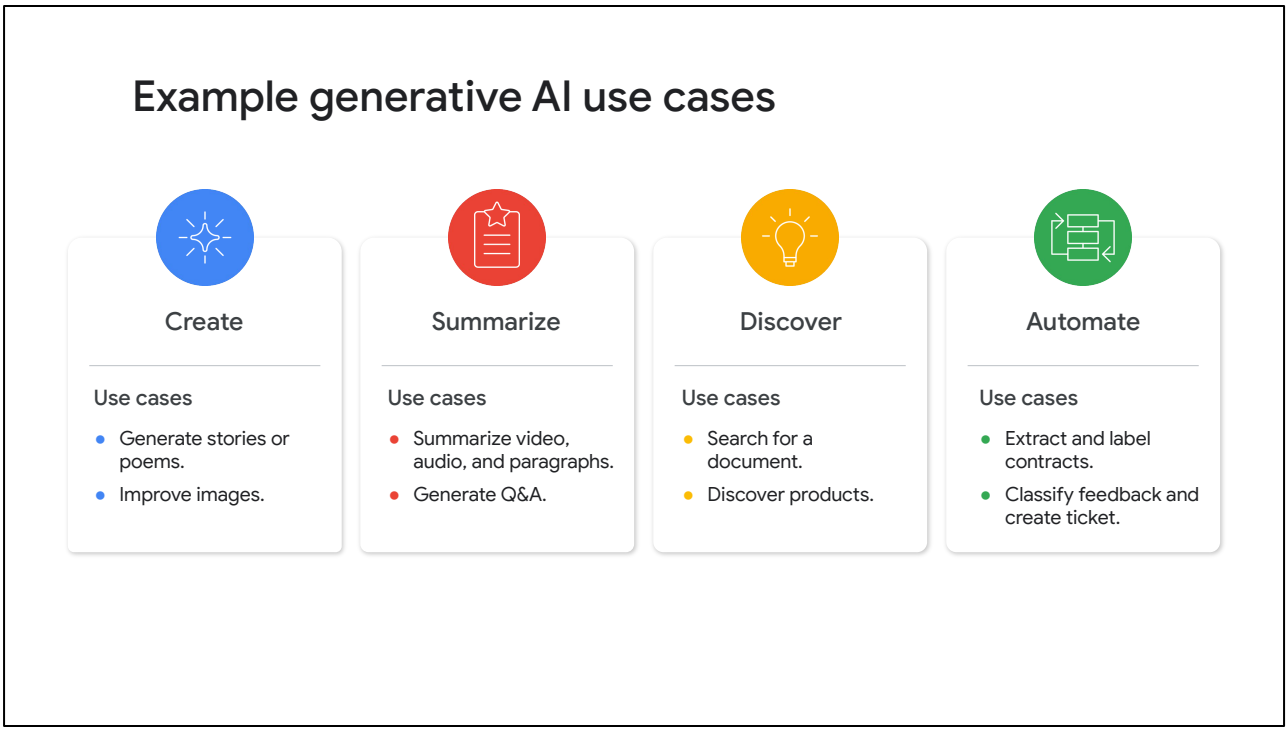
What does large mean? **Large** has two meanings.

First is the enormous size of the **training dataset**, sometimes at the petabyte scale.

Second it refers to the number of **parameters**, which now reaches billions and even trillions. Parameters are essentially the memories and knowledge that the machine has learned during model training. Parameters determine the ability of a model to solve a problem, such as predicting text.

General-purpose means that the models are sufficient to solve common problems. The models work due to the commonality found in a human language, regardless of the specific tasks you are trying to do.

This leads to the last point: **pre-trained and fine-tuned**. A large language model can be pre-trained for general purpose use with a large dataset. Later, it can be fine-tuned for a specific purpose by using a much smaller dataset.



What are the potential use cases of generative AI?

Generative AI can create content and bring your thoughts and visions to life. It can do things like:
Generate stories of poems based on prompts that you provide, or improve images based on instructions.
Creation of content is a key benefit of generative AI, but generative AI can do much more.

Generative AI can summarize knowledge. Such as:
Automatically summarizing video, audio, and paragraphs, or
generating questions and answers based on the content.

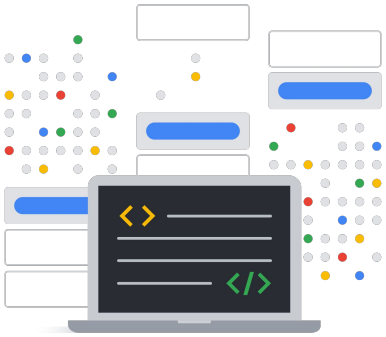
Generative AI can do search and discover for you. For example, it can:
Search for a document, or
Discover products based on desired features.

Generative AI can also automate workflows. For instance, it can:
Extract and label contracts, or
classify feedback and create support tickets.

You can use generative AI to create powerful and compelling applications.

Generative AI use cases for application development

- Code generation
- Documentation
- Code explanation



- Fixing code
- Code completion
- Code translation

Vertex AI Codey APIs and Gemini

Generative AI will revolutionize how applications are developed.

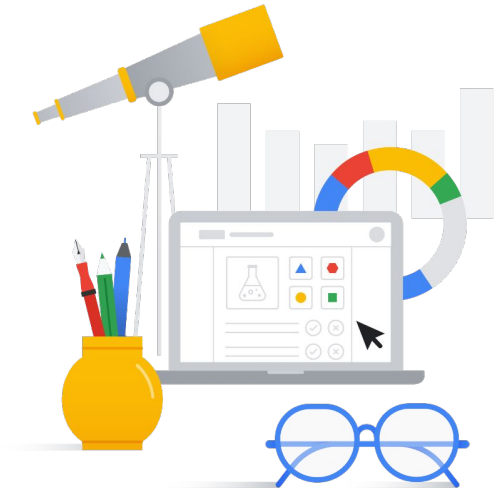
You will code your apps with help from your own generative AI-powered coding assistant.

Features will include:

- Code generation. Generate code based on a natural language description of the desired code. Automatically generate unit tests for a piece of code or ask your assistant to optimize code.
- Documentation. The assistant can add comments to your code or generate release notes based on the changes.
- Code explanation. Ask your assistant to explain what the code does, and how it does it.
- Fixing code. Your AI assistant will be able to find bugs in your code, and then fix them.
- Code completion. As you type your code, the context of your code will be used to finish the line of code you're writing. Or your code editor might suggest code for the entire function.
- Code translation. Take code written in one coding language, and have your assistant translate it into another, while adhering to coding conventions of the new language.

The Vertex AI Codey APIs, powered by the Codey foundation model, can assist with code generation, code chat, and code completion to provide these features. Gemini will provide you this assistant to help you write code faster and more efficiently.

[Code models overview:
https://cloud.google.com/vertex-ai/docs/generative-ai/code/code-models-overview
The next chapter of our Gemini era:
https://blog.google/technology/ai/google-gemini-update-sundar-pichai-2024/]



Hands-on lab:

Adding User Authentication and Intelligence to Your Application

In this lab, "Adding User Authentication and Intelligence to Your Application," you enhance the bookshelf application by using OAuth to authenticate users. You use Secret Manager to store sensitive application data. Finally, you use the Cloud Translation API to translate the description of the book to the preferred language of the logged in user.

In this module, you learned ...

- Google Cloud provides pretrained models for vision, speech, video intelligence, and natural language processing.
- You can use these pretrained models or train your own models without being a machine learning expert.
- Tools like **Vertex AI** make it easy to train and deploy custom machine learning models on Google Cloud.

We began this discussion by looking at the machine learning services, tools, and pretrained models that help you add intelligence to your applications.

Google Cloud provides pretrained models for vision, speech, video intelligence, and natural language processing.

You can use these pretrained models or train your own models without being a machine learning expert. You can then use these models to build exciting features in your applications.

Google Cloud also provides tools like Vertex AI that can help you use your own data to build customized models. Using these models in your apps is as simple as making API calls.

In this module, you learned ...

- Generative AI differs from traditional programming.
- You learned about large language models.
- We discussed Generative AI use cases for a variety of applications and application development tasks.

We also discussed generative AI, and how it differs from traditional programming.

You learned about large language models, and we discussed generative AI use cases for applications and application development.