AI

Generative artificial intelligence refers to models that predict and generate various types of outputs (such as text, images, or audio) based on what’s statistically likely, pulling from patterns they’ve learned from their training data. For example:

- Given a photo, a generative model can generate a caption.
- Given an audio file, a generative model can generate a transcription.
- Given a text description, a generative model can generate an image.

A **large language model (LLM)** is a subset of generative models focused primarily on **text**. An LLM takes a sequence of words as input and aims to predict the most likely sequence to follow. It assigns probabilities to potential next sequences and then selects one. The model continues to generate sequences until it meets a specified stopping criterion.

LLMs learn by training on massive collections of written text, which means they will be better suited to some use cases than others. For example, a model trained on GitHub data would understand the probabilities of sequences in source code particularly well.

However, it's crucial to understand LLMs' limitations. When asked about less known or absent information, like the birthday of a personal relative, LLMs might "hallucinate" or make up information. It's essential to consider how well-represented the information you need is in the model.

An **embedding model** is used to convert complex data (like words or images) into a dense vector (a list of numbers) representation, known as an embedding. Unlike generative models, embedding models do not generate new text or data. Instead, they provide representations of semantic and synactic relationships between entities that can be used as input for other models or other natural language processing tasks.

In the next section, you will learn about the difference between models providers and models, and which ones are available in the AI SDK.