Generative AI

How Generative AI differs from discriminative models, how text/image generation actually works, a preview of Large Language Models, and the critical distinction between fluency and correctness.

Beginner 15 min

Last updated: August 22, 2026
Read online: https://www.learnforgex.com/en/topics/generative-ai

---

Generative AI

The "AI Fundamentals" category opened with a map of four nested circles: AI, Machine Learning, Deep Learning, and Generative AI. The first three lessons worked from the outside in -- what AI is, the Machine Learning mechanism behind it, and the Deep Learning technique (neural networks) that made today's most capable systems possible. This lesson covers the fourth and innermost circle: Generative AI. Unlike the other three, Generative AI isn't a distinct *technique* -- it's an *application* of the techniques you already know, aimed at a different kind of task. Understanding that distinction is the key to this whole lesson, and it sets up everything the rest of this course (Large Language Models, Tools & MCP, AI Agents) builds on top of.

What Is Generative AI?

Generative AI refers to AI systems trained not to classify or predict a single label, but to **create new content**: text, images, audio, code, video. The large majority of modern generative AI systems are deep-learning based -- but that doesn't mean "generative AI = deep learning," only that it's the most common way to build one today. Contrast this with the examples used throughout the previous three lessons: a spam filter predicts one of two labels ("spam" / "not spam"); an image classifier predicts one of many categories ("cat," "dog," "car"). Generative AI does something structurally different -- given a prompt, it produces an entirely new piece of content that didn't exist before, one word, pixel, or sound sample at a time.

"New content" spans several different types; roughly, it branches like this:

```
Generative AI
├── Text
│   └── LLM (see "Large Language Models: A Preview")
├── Image
├── Audio
├── Video
└── Code
```

The fourth and final lesson in this category, "Large Language Models," zooms in on just the top branch -- generative AI applied to text. The other branches are out of scope here, but they're the same underlying idea (models trained to create new content) applied to different kinds of data.

Discriminative vs. Generative: The Core Distinction

Most of the models discussed in "Machine Learning" and "Deep Learning" -- the spam filter, the cat classifier -- are what's called **discriminative** models: given an input, they discriminate between a fixed set of possible outputs (a label, a category, a number). A **generative** model, in contrast, learns the underlying patterns and structure of its training data well enough to produce brand-new examples that plausibly could have come from that same data. A discriminative model trained on millions of cat photos can tell you "yes, this is a cat." A generative model trained on the same photos can produce a picture of a cat that has never existed, by having learned what makes cat photos look the way they do.

> A simple gut-check for telling the two apart: if a system's output is a label, a category, or a number, it's almost certainly discriminative. If its output is a new sentence, image, sound clip, or block of code, it's generative. Keep in mind this is a simplified beginner-level mental model at this stage of the course -- some real systems sit at the boundary between the two, but the "discriminative vs. generative" check is still extremely useful for building the right intuition as a newcomer.

How Generative Models Produce New Content

Without getting into the underlying math (this course deliberately avoids that, as stated in "Deep Learning"), it helps to have an accurate mental model of the general idea. Most modern generative models -- especially text and image generators -- work by learning to predict "what comes next," one small piece at a time, based on everything that came before it:

- A text generator predicts the next most likely word (technically, a smaller unit called a *token*) given the words so far, then adds that word to the sequence and repeats -- building a full sentence or paragraph one token at a time.

- An image generator commonly works by starting from random noise and gradually refining it, step by step, into a coherent image that matches a text description -- a family of techniques called *diffusion models*.

In both cases, the model was trained (using the training loop from "Deep Learning" -- forward pass, loss, backpropagation) on enormous amounts of existing content, learning the statistical patterns of what tends to follow what. Generation at inference time is that same learned pattern, run forward to produce something new rather than to classify something given.

Large Language Models: A Preview

The text-generation idea above -- predicting the next token, one at a time, based on everything so far -- is exactly the mechanism behind **Large Language Models (LLMs)**, the systems behind tools like modern AI chat assistants and code-writing assistants. LLMs are generative AI applied to text, built on the transformer architecture introduced in "Deep Learning"'s "Common Types of Neural Networks" section, and trained on enormous amounts of text collected from books, websites, and code repositories. "Predicting the next token" is presented here as a deliberately simplified model -- how LLMs are actually trained (pretraining), the separate stage that teaches a model to follow instructions (post-training / instruction tuning), how "context" shapes a model's behavior at inference time, and inference itself, are each covered one at a time in this course's next category, "Large Language Models." That category is dedicated entirely to how these systems actually work, what a "prompt" really does, and their specific capabilities and limitations -- this lesson only needs you to see that LLMs are a specific, hugely important instance of the general idea of generative AI, not a separate concept.

What Generative AI Is Not

Because generative AI is the newest and most visible piece of the map from "What Is Artificial Intelligence?", it's worth being explicit about a few things it does not automatically mean:

- **Generative AI is not the same as general intelligence (AGI).** Recall "Narrow AI and General AI" from the first lesson -- a generative model, however fluent its output looks, is still a narrow AI system trained on a specific kind of task (predicting the next token, or the next refinement step), not a general-purpose reasoner.
- **Generating fluent output is not the same as generating correct output.** A language model can produce a grammatically perfect, confident-sounding sentence that is factually wrong -- a behavior often called "hallucination." Fluency is a property of the generation *mechanism* (predicting plausible next tokens); correctness is a separate property that has to be checked, not assumed.
- **Generative AI is not automatically "agentic."** A system that generates a paragraph of text or an image in response to one prompt is not, by itself, making autonomous

decisions or taking multi-step actions in the world -- that's the subject of this course's later "AI Agents" category, which is built *on top of* generative models but adds distinct additional machinery.

Common Uses of Generative AI Today

Tying back to "Where Is AI Used Today?" from the first lesson, generative AI specifically shows up in:

- **Text generation:** drafting emails, summarizing documents, writing and explaining code, answering questions in conversation.
- **Image generation:** producing illustrations, product mockups, or art from a text description.
- **Audio and speech generation:** synthesizing realistic speech from text, or generating music.
- **Code generation:** the specific case of text generation applied to programming languages -- autocompleting a function, generating a whole file from a description, or explaining what an unfamiliar piece of code does.

Best Practices

- Always distinguish *fluency* from *correctness* when evaluating generative AI output (see "What Generative AI Is Not") -- a confident, well-written response is not evidence that it's accurate.
- When you see an impressive generative AI demo, ask which of the four circles from "What Is Artificial Intelligence?" it's really showing you -- a specific, narrow, trained capability, not general intelligence.
- Match the model type to the task: use a discriminative model when you need a classification or a number, and a generative model when you actually need new content produced -- using a generative model to solve a simple classification problem is usually unnecessary overhead.

Common Mistakes

- **Assuming every AI system is "generative AI."** As "Discriminative vs. Generative: The Core Distinction" showed, huge, useful categories of AI (spam filters, fraud detectors, recommendation systems) are discriminative, not generative -- the terms are not synonyms for "AI" in general.
- **Trusting generated text as automatically factual.** As covered in "What Generative AI Is Not," a fluent answer and a correct answer are two different things -- this course's later "AI Evaluation" material goes much deeper into how to actually check for this.

- **Treating "generative AI" and "AI agent" as the same thing.** A chatbot that generates a single reply to a single prompt is generative AI; a system that plans, takes multiple actions, and uses tools to accomplish a goal is an agent built on top of generative AI -- the two terms describe different layers, not the same capability (see "What Generative AI Is Not").

Summary, Cheat Sheet, and Glossary

Summary

- Generative AI is the innermost circle on the map from "What Is Artificial Intelligence?" -- not a separate technique from deep learning, but an *application* of it aimed at producing new content instead of predicting a label.
- **Discriminative** models choose between fixed outputs (a label, a category); **generative** models produce entirely new content.
- Most modern generative models work by predicting "what comes next" -- the next token for text, the next refinement step for images -- based on everything learned during training.
- **Large Language Models** are generative AI applied to text, built on the transformer architecture from "Deep Learning" -- the subject of this course's next category.
- Fluent output is not automatically correct output, and generative AI is not automatically "agentic" -- both are common points of confusion worth keeping straight.

Cheat Sheet

- Discriminative = chooses a label/category. Generative = creates new content.
- LLM = generative AI applied to text, via the transformer architecture.
- Diffusion model = a common approach for generative image models.
- Fluency ≠ correctness. Generation ≠ agentic behavior.

Glossary

- **Generative AI:** AI systems trained to produce new content (text, images, audio, code) rather than classify or predict a fixed label.
- **Discriminative model:** a model that chooses among a fixed set of possible outputs, such as a label or category.
- **Token:** the small unit (often close to a word or word-piece) a text generation model predicts one at a time.
- **Diffusion model:** a family of generative image techniques that gradually refine random noise into a coherent image.
- **Large Language Model (LLM):** a generative AI system, built on the transformer architecture, trained to predict and generate text.

- **Hallucination:** when a generative AI system produces fluent, confident output that is factually incorrect.