# Overview

Generative AI is a subset of machine learning (ML) that is used for creating new content. Generative models use neural networks to generate data based on, but different from, the data they were trained on.

Generative AI burst onto the public scene on November 30, 2022 when OpenAI released a public demo of ChatGPT.

A Generative Pretrained Transformer (GPT) chatbot, like ChatGPT, works by taking a prompt (often a question) as input, and generating a response (an answer) based on statistical probabilities.

* It is *generative* because it generates some kind of output, like text.
* It is *pretrained* because it uses a data model that has been trained through machine learning.
* It is a *transformer* because it transforms a sequence of input text into a sequence of output text.

# Large Language Models

A *Large Language Model* (LLM) is a neural network build using *transformer architecture*. Transformer architecture focuses attention on words to weigh their importance in a sentence, allowing the transformer to "understand" context and relationships between the words. LLMs can perform a variety of tasks, such as answering questions, summarizing text, translating languages, and creative writing.

## Training an LLM

A large language model is trained using unsupervised learning with vast amounts of text data from diverse sources like books, articles, and websites. The model uses this data to learn patterns in language, such as grammar, context, and word associations. The training requires significant computational power and time, often utilizing specialized hardware like GPUs. Throughout training, the model's performance is evaluated and fine-tuned to improve accuracy and relevance. The result is a powerful tool capable of "understanding"[[1]](#footnote-1) and generating human-like text.

### Parameters

LLM parameters are the numerical values that define the behavior and performance of the model. These parameters include *weights* and *biases*, which are adjusted during training to minimize errors in predictions. The number of parameters in an LLM can range from millions to billions, contributing to the model's ability to generate human-like text. For example, the Llama 3.1 LLM from Meta has 405 billion parameters.

## Using an LLM (Inference)

A GPT chatbot uses the pretrained LLM to predict the most probable output for text input from a user.

The process of generating output is iterative. This is what a chatbot does:

1. Based on a prompt, what are the top few most likely candidates for the first word in the response. Pick one randomly.
2. Based on the prompt plus the words in the response so far, what are a few most likely candidates for the next word in the response, pick one randomly.  
   Repeat #2 until there is a low probability of there being a next word.

### Context Length

The *context length* of a LLM refers to the maximum number of tokens it can process at once. It's like the model's memory or attention span. For example, GPT-3 has a context length of 4,096 tokens, while GPT-4 can handle up to 8,192 tokens, and even up to 32,768 tokens in the extended version—which is used in Microsoft Copilot. This means the model can consider a larger amount of text when generating responses, which can improve coherence and relevance.

### Tokens

A *token* is a unit of text that the model processes. Tokens can be words, subwords, or even individual characters, depending on the tokenization method used. For example, the sentence "I love AI" might be tokenized into three tokens: "I", "love", and "AI". In more complex cases, words can be broken down into smaller subword tokens, especially for handling rare or compound words. The model uses these tokens to understand and generate text.

1. [↑](#footnote-ref-1)