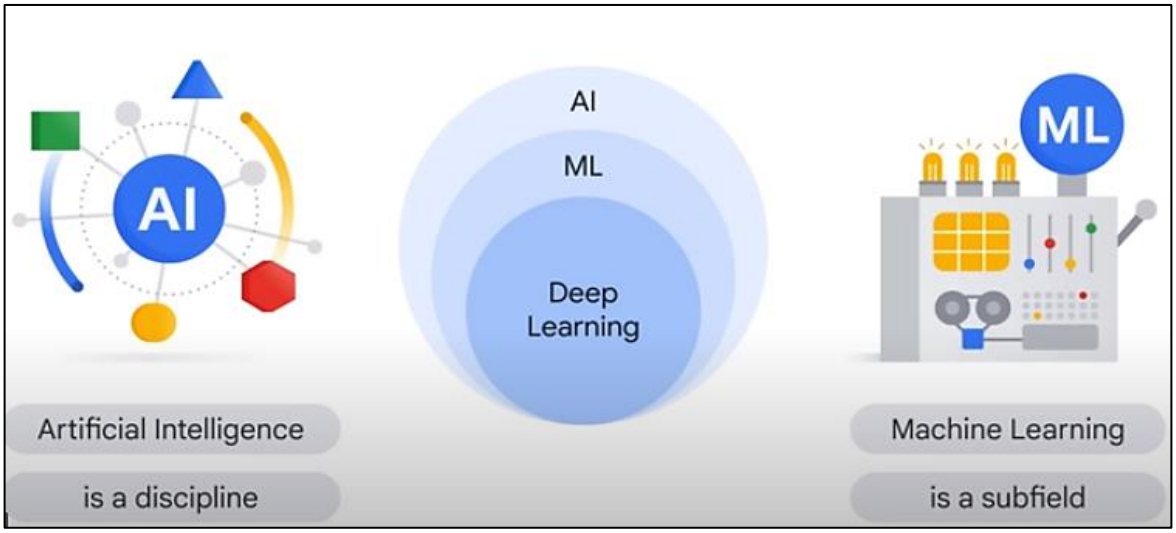
Lecture 1 Introduction To Generative AI

1.1 AI (Artificial Intelligence) vs Machine Learning

1.1.1 AI vs Machine Learning



AI, or Artificial Intelligence, is a field of computer science that creates intelligent agents that can reason, learn from experiences, and make independent decisions without human intervention. Think of AI as the brainpower that enables a computer to understand and make decisions. For example, imagine a self-driving car that uses AI to navigate through traffic.

Machine learning, on the other hand, is a subfield of AI. It involves teaching machines to recognize patterns and make predictions/decisions based on data. It trains a model from input data (part of the original data), which then can make useful predictions from new or unseen data (drawn from another part of the original data). Machine learning gives computers the ability to learn without explicit programming. For instance, think about a trained model to predict stock prices based on historical data.

1.1.2 Supervised Machine Learning vs Unsupervised Machine Learning

Supervised and Unsupervised are the two most common machine learning classes. The key difference is that the supervised learning method uses labeled data to learn and predict whereas the unsupervised learning method finds something interesting in the data as there is no labeled data for supervision.

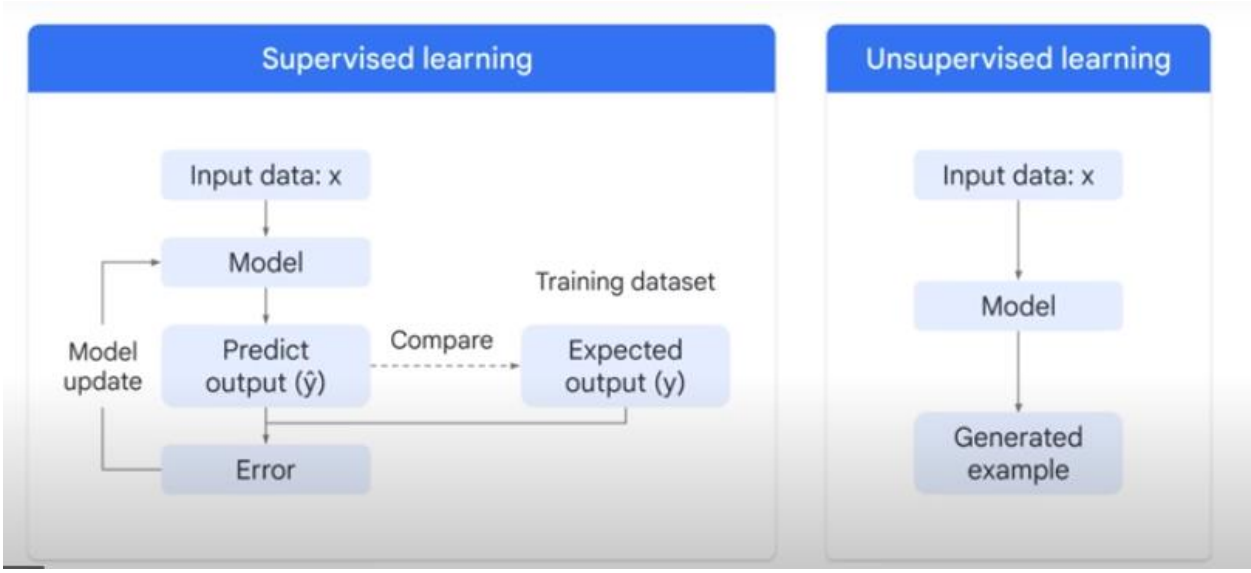
Labeled data comes with a tag like a name, a type, or a number.

For one stock, imagine we have the past five years of stock prices used as training data. We label the data using some trend reversal (bullish or bearish) trading signals. We want to predict how precise these trading signals are when capturing the sudden change in stock price movements. We need to use labels: reversal '1', and non-reversal '0'. We can use this labeled data to train a model to see how well those trend reversal trading signals can capture the change in direction of stock price movements (e.g. using metrics such as recall, precision, accuracy, etc.).

Correspondingly, for a portfolio of 1000 stocks, we do not know which stock will perform well in the next three months and which stocks will fail. We want to conduct a trading strategy by longing those 'winners' but short-selling those 'losers.' We prepared 200 features using firm fundamental data for each of these 1000 stocks (DataFrame: 1000 x 200). Using unsupervised learning, we group these stocks into two

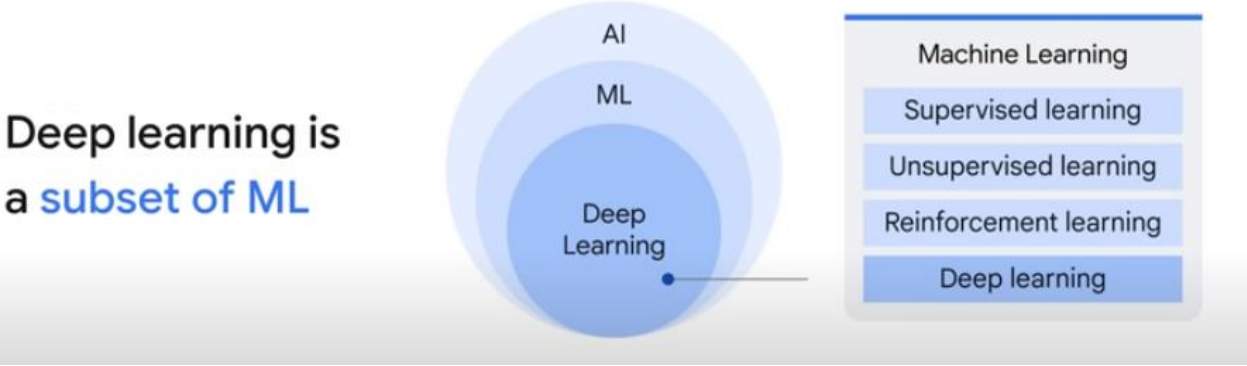
categories: winners and losers. We short-sell those losers and use the cash we received to buy those winners (the idea of ‘cash-neutral’).

Below is the diagram to get a better understanding of both approaches.

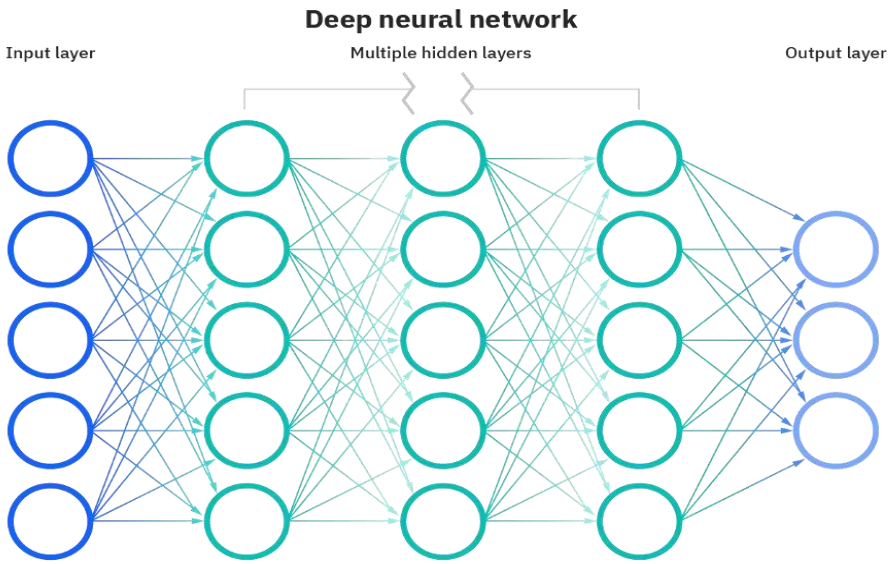


1.2 Deep Learning

1.2.1 Introduction of Deep Learning



Deep Learning is a subset of machine learning. Deep Learning uses artificial neural networks, which allows them to process more complex patterns than traditional machine learning. Neural Networks can use labeled or unlabeled data. It is also called semi-supervised learning. In semi-supervised learning, the neural networks are trained on a small amount of labeled data and a large amount of unlabeled data. The labeled data helps the neural networks to learn the basic concept of the task. The unlabeled data helps the neural networks to generalize to new examples.

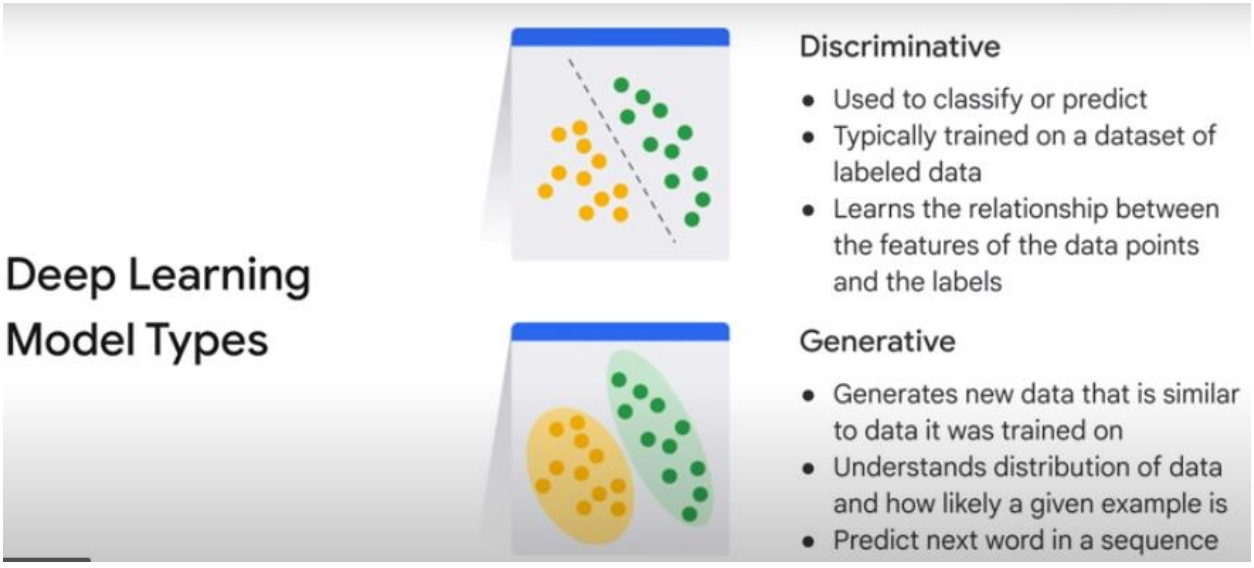


Note: The “deep” in deep learning refers to the depth of layers in a neural network.

Deep learning models typically have many layers of neurons which allows them to learn more complex patterns than traditional machine learning methods. They typically use larger datasets for training. Deep learning models can be used for a variety of tasks, including classification, regression, object detection, natural language processing, and more. The depth of the neural network allows it to capture complex patterns and relationships in the data. Some of the applications of deep learning are - Convolutional Neural Networks (CNNs) for image-related tasks, Recurrent Neural Networks (RNNs) for sequential data, speech recognition, recommendation systems (used in social media or shopping websites), and autonomous vehicles, etc.

1.2.2 Types of Deep Learning Models

Deep learning models can be broadly categorized into two types: generative models, and discriminative models.



Discriminative models

As the name suggests, discriminative models discriminate between different classes or categories based on provided data. Discriminative Model once trained, can predict labels of the new data points. It focuses on learning:

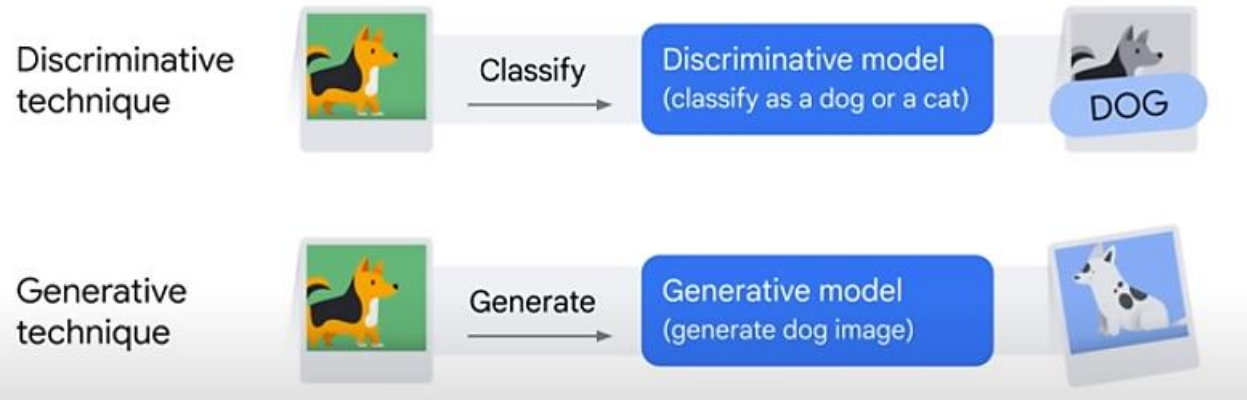
- the relationship between the features of the data points and the labels,
- the boundary or decision boundary that separates different classes or categories within the data,

It then classifies or predicts labels for the given data. discriminative models are typically trained on a dataset of labeled data (classifications).

Generative models

Generative models generate new data instances that are like the data it was trained on. Its learning is based on the probability distribution of existing data and predicts new words in a sequence.

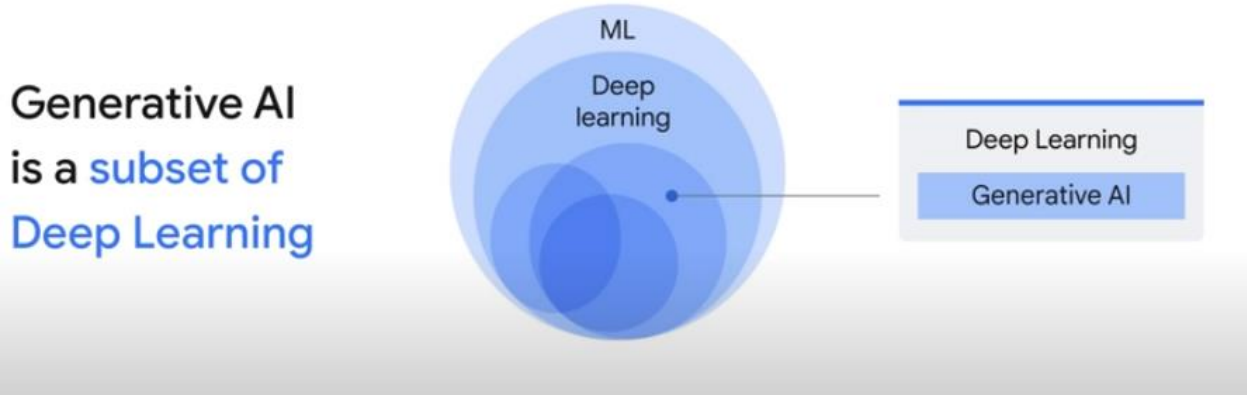
Let's take an example, assume you have a dataset containing images of dogs and cats. The discriminative model learns the probability of y (output), given x (input), that this is a dog (this is also called conditional probability) and classifies it as a dog or a cat. On the other hand, the generative model learns the probability of x and y together (also called joint probability), predicts the conditional probability that this is a dog, and then generates a picture of a dog.



1.3 Generative AI

1.3.1 Introduction of Generative AI

Generative AI is a subset of deep learning, but they refer to different aspects of artificial intelligence.



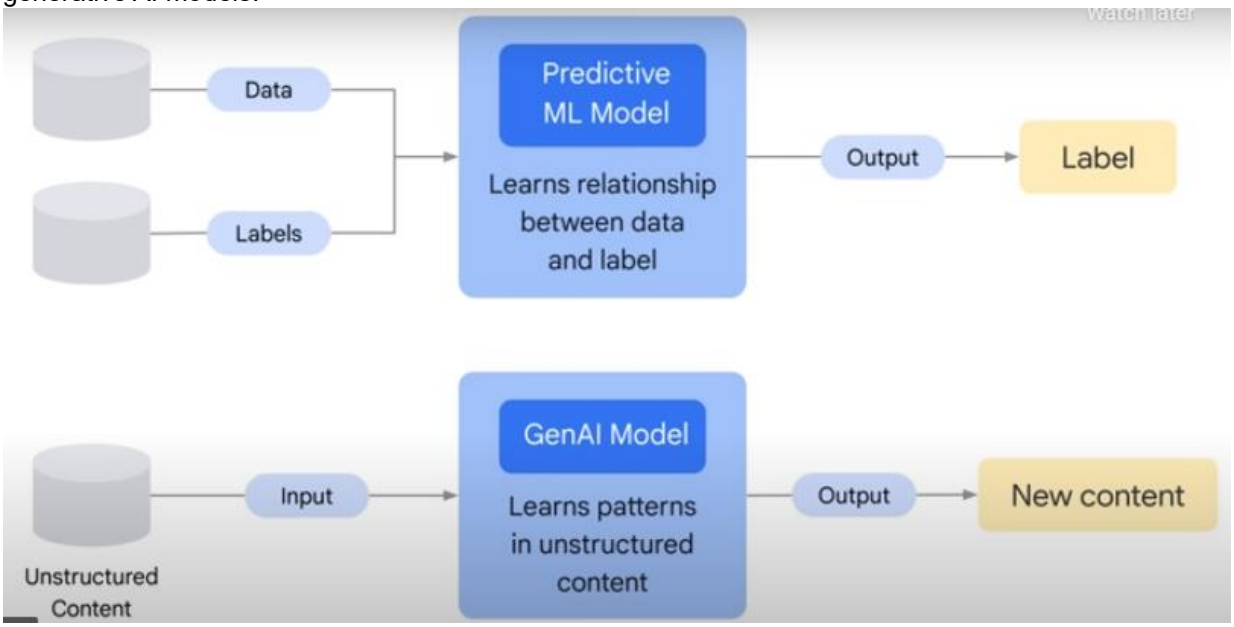
Generative AI refers to a class of algorithms and models designed to generate new, original content. This content could include images, text, audio, video, and even human-like conversations. For example, Large Language models (LLMs) can engage in natural language conversation. It is widely used in creating a virtual assistant who can chat with you about a wide range of topics, answer your questions, and even tell you jokes.

Generative AI is capable of creating data that was not present in the training set. It involves learning the underlying patterns and structures of the data to produce new instances that share similarities with the training data. Generative AI can use supervised, unsupervised, and semi-supervised methods. Generative AI models include Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and autoregressive models like Transformers.

1.3.2 Generative AI vs Machine Learning

Model Input

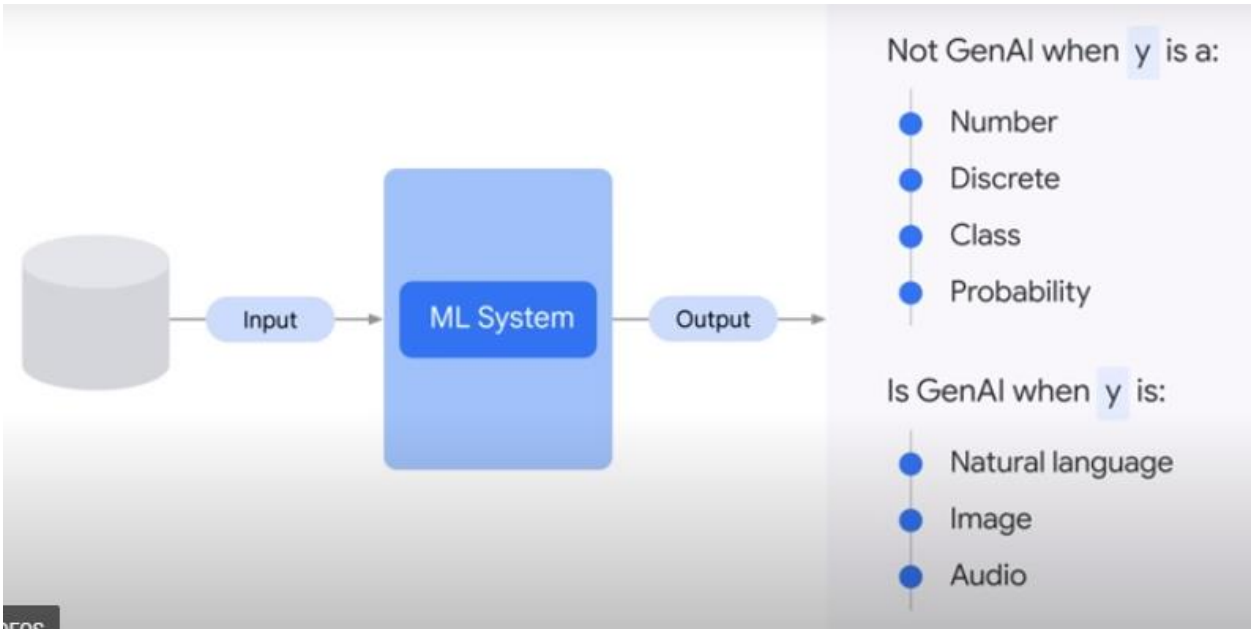
The following diagram illustrates the difference between the predictive machine learning models and the generative AI models.



The top flow diagram is the basic architecture of a traditional machine-learning model. These models learn the relationship between data and labels, and then predict the label (this process can be a simple prediction, classification, or clustering). The bottom flow diagram shows how generative AI models learn patterns on the unstructured content so as to generate new content, like text, image, music or even the code.

Model Output

The best way to differentiate between traditional machine-learning algorithms and generative AI is by distinguishing them based on the output.



A model is NOT GenAI when the output (y or label) is (may be deep learning)

- a number – 3.1415926
- a discrate – a cate or dog
- a class – Yes or No
- a probability – 0.05

A model is GenAI when the output (y or label) is

- natural language
- image
- audio

Let's see it mathematically, below is our model equation which represents that model output is the function of all the inputs.

$$y = f(x, \beta)$$

Here, y is the model output, x represents the model input, and $f(\cdot)$ embodies the model methodology.

If y is a number, like the predicted stock price, the model is not generative AI. If y is a sentence that discuss about the stock performance, the model is generative AI, because the model could elicit a text response based on the massive data input the model was trained on.

In generative AI, users can generate their content, whether it is text, audio, video, or code.
For example, models like Sparrow, LaMDA (Language Model for Dialogue Applications), or PaLM (Pathways Language Model) ingest very large data from multiple sources and build the foundation language models to help you generate content, that we can simply use by asking a question, either by typing or speaking. Like, you can ask "*What is the cat?"* or "*How do you say "hello" in different languages?",* and it will give you everything it has learned about languages or cats.

1.4 Types of Generative AI Models

In particular, Generative AI renders statistical models to train prompts (data input you feed into GenAI), predicts potential responses, and thereby generates new content. It essentially grasps the underlying structure of the data and uses this knowledge to produce entirely new information that closely resembles the training data.

1.4.1 Transformer Models

Transformers have gained significant popularity in recent years, especially in natural language processing tasks. Models like GPT (Generative Pre-trained Transformer) are pre-trained on large datasets and fine-tuned for specific generative tasks.

The Transformer is a type of neural network architecture introduced in the paper "Attention is All You Need" by Vaswani et al. It's designed to process sequential data using attention mechanisms, allowing the model to consider different parts of the input sequence when making predictions. Key Features of Transformer model include but are not confined to self-attention mechanism, multi-head attention, positional encoding.

Transformers consist of an encoder and a decoder. The encoder encodes the input sequence and passes it to decode the representation for a relevant task. Then the decoded text is passed as input to the pre-trained generative transformer model. Note that in pre-training, the model is trained on enormous amounts of data and billions of parameters.

Hallucination

In generative AI, “hallucination” means that the AI can make up things that are not true. It hallucinates. It might invent stories, mix up facts, or guess things that it does not know.
Technically, hallucinations are words, or phrases generated by the model that are often nonsensical or grammatically incorrect.
Developers work to make sure that AI will not do it, but it can be a challenge like –

- Not enough training data
- Noisy or dirty data
- Not enough context is given to the model, or
- Not enough constraints are given to the model.

Hallucinations can be a problem for transformers because they can make the output text difficult to understand. They can also make the model more likely to generate incorrect or misleading information. Suppose you ask an AI chatbot, "Tell me about a famous historical event in 1800." The AI might respond with something like, "In 1800, astronauts landed on the moon." This answer is a hallucination because it's not true. The moon landing did not happen until 1969, not in 1800. The AI made up a fictional event, which is not true and misleading.

1.4.2 Large Language Models (LLM)

Large Language Model (LLM) is a specific instance of a neural network, often based on the transformer architecture. It is trained on a massive amount of textual data to understand and generate human-like language. Large Language Model (LLM) is pre-training on a diverse dataset, fine-tuning for specific tasks. Large language models are used for a range of natural language processing tasks, such as text completion, question answering, summarization, and language translation.

Prompt Design

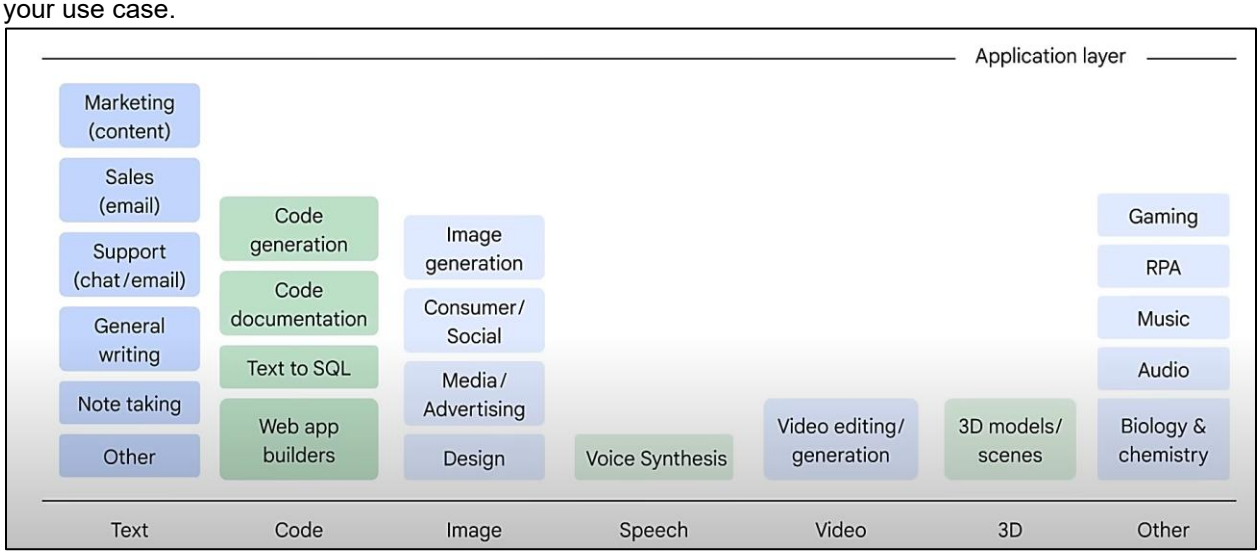
A prompt is a short piece of text that is given to the large language model as input and can be used to control the output of the model in many ways. The prompt can be a question, a statement, or context you give to AI to get the desired outputs.

Prompt design is the process of creating a prompt that will generate the desired output from a large language model. It is an important aspect of working with generative AI and can significantly impact the quality and relevance of the responses.
Gen AI heavily depends on the training data that you have fed into it. It analyzes the patterns and structures within the input data to learn. However, with the availability of browser-based prompts, you, as the user, can create your content.

In summary, the transformer is a type of neural network architecture, while a large language model is a specific implementation of that architecture, usually with a vast amount of pre-training on textual data. Large language models, like OpenAI's GPT (Generative Pre-trained Transformer) series, utilize the transformer architecture to achieve impressive results in understanding and generating human-like language. The transformer architecture itself, however, can be applied to various types of data beyond language, showcasing its versatility in handling sequential information.

1.5 Applications of Generative AI Models

Generative AI has a wide variety of applications from generating text, to generating code depending on your use case.



Text-to-text: Text-to-text models take a natural language input and produce text output. These models are trained to learn the mapping between a pair of texts (e.g., translation from one language to another). Applications – Generation, Classification, summarization, etc.

Text-to-image: Text-to-image models are relatively new and are trained on a large set of images, each captioned with a short text description. Diffusion is one method used to achieve this. Applications include - Image generation, Image editing, etc.

Text-to-video/Text-to-3D: Text-to-video models aim to generate a video representation from text input. The input text can be anything from a single sentence to a full script, and the output is a video that corresponds to the input text. Similarly, text-to-3D models generate three-dimensional objects that correspond to a user's text description (for use in games or other 3D worlds). Applications include - Video generation, Video editing, Game assets, etc.

Text-to-task: Text-to-task models are trained to perform a specific task or action based on text input. This task can be a wide range of actions such as answering questions, performing a search, making predictions, or taking some sort of action. For example, a text-to-task model could be trained to navigate web UI (User

Interface) or make changes to a doc through the GUI. Applications include – Software agents, virtual assistants, automation, etc.