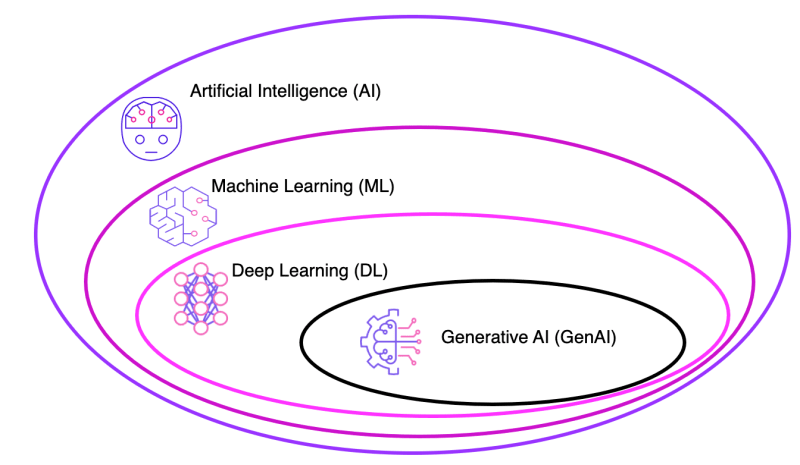
**Large Language Model (LLM) & Generative AI (Gen AI)**



# **What are Foundation Models?**

* Foundation model as an “AI model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks”.
* It’s been trained on a huge amount of unsupervised data
* By referring to all the dataset, it can predict and generate the next word based on previous words that it’s seen before.
* Foundation Models is a part of Generative AI.
* Can use prompt engineering. For example, prompting a model to perform classification tasks like giving a sentence and asking if it is positive sentiment or negative.
* Once foundation models combine with LLM, it will become more powerful because it has more dataset or parameters.

# **What are LLM?**

LLM can be considered a subset of generative AI. It is an instance of foundation model which is trained on massive datasets.

LLMs are a type of generative AI that are trained on text and produce textual content.

LLM refers to general purpose language models being **pre-trained** and then **fine-tuned** for specific purposes.

**Pre-trained?**

* Trained to solve common language problems, like:
  + Text classification (E.g. spam mail, sentiment of review, categorize news article topics)
  + Question answering
  + Document summarization
  + Text generation

**Fine-tuned?**

* Then be tailored to solve specific problems in different fields, like:
  + Retail

**[Example]**

**Prompt:**

“Fine-tune the model to generate creative and engaging product descriptions for an online retail platform.”

**Process:**

Train the model with promote tone, optimize description with informative and persuasive.

**Application:**

Automatically generate product descriptions for new items in the online retail catalog.

* + Finance
  + Entertainment
* Trained with a relatively small size of field datasets.

## **3 Major Features of LLM**

* Large
  + Large training datasets - Sometimes at the petabyte scale (1 petabyte = 1000 TB)
  + Large number of parameters – the memories and knowledge that the machine learned from model training.
* General purpose
  + Sufficient to solve common problems
  + Commonality of human languages
  + Resource restriction – only certain organizations have the capability to train such large language models with huge data sets and a tremendous number of parameters.
* Pre-trained and fine-tuned

## **How LLM work?**

* LLM is based on a transformer model and works by receiving an input, encoding it, and then decoding it to produce an output prediction.

## **What are Transformers?**

* It is a type of neural network architecture
* For example, GPT uses Recurrent Neural Network (RNN) and long short-term memory (LSTM), but it can’t handle large sequences of text and it will forget what happened in the beginning when it comes to the end of paragraph. Then Transformer fixed this problem.
* Transformers have it self-attention, it allows neural network to understand the context and the surrounding words.
* Also have positional encoding, neural network will learn the word order from the data.

**Step 1: Training**

* LLM are pre-trained using large textual datasets from sites like Wikipedia, GitHub, or others.
* Unsupervised learning, meaning it processes the datasets fed to it without specific instructions.
* LLM's AI algorithm can learn the meaning of words, and of the relationships between words.

**Step 2: Fine-Tuning**

* For a large language model to perform a specific task, such as translation, it must be fine-tuned to that activity. Fine-tuning optimizes the performance of specific tasks.

**Step 3: Prompt-Tuning (Prompt Engineering)**

* Used to guide Generative AI (Gen AI) to generate desired output.
* It requires detailed instructions to create high-quality and relevant output.
* **Zero-shot Learning**
  + Using single command to get an LLM to take on a behavior
  + Implies that the model can recognize things that have not explicitly been taught in the training.
* **Few-shot Learning**
  + Adding example or instruction can be helpful to show the model

## **Benefits of LLM**

* **Large set of applications**

Can be used for language translation, sentence completion, sentiment analysis, question answering, mathematical equations, and more.

* **Always improving**

LLM performance is continually improving because it grows when more data and parameters are added. Once an LLM has been pretrained, few-shot prompting enables the model to learn from the prompt without any additional parameters.

* **Learn fast**

They do not require additional weight, resources, and parameters for training.

## **Challenges and limitation of LLM**

* **Security**

LLM present important security risks when not managed or surveilled properly

* **Scaling**

It can be difficult in time and resource-consuming to scale and maintain large language models.

* **Deployment**

Deploying large language models requires deep learning, a transformer model, distributed software and hardware, and overall technical expertise.

## **Examples of popular large language models**

* **GPT (Generative pre-trained transformers)**

Developed by OpenAI. Currently, it is the largest or generalized LLM in terms of natural language generation.

* **PaLM (Pathways Language Model)**

Developed by Google. It is a transformer language model capable of common-sense, joke explanation, code generation, and translation.

* **BERT (The Bidirectional Encoder Representations from Transformers)**

Can understand natural language and answer questions.

* **XLNet**
* **Bard (Updated to Gemini)**

A conversational AI tool by Google.

# **Generative AI (Gen AI)**

* Part of the foundation models
* Type of artificial intelligence that can produce new content, including text, images, audio, and synthetic data.
* Powered by foundation models (large AI models) that can multi-task and perform out-of-the-box tasks, including summarization, Q&A, classification, and more.

## **How does Gen AI work?**

* Neural networks identify the patterns and structures within existing data to generate new and original content.
* Gen AI models are the ability to leverage different learning approaches, including unsupervised or semi-supervised learning for training.

## **Challenges of Gen AI**

* **Scale of compute infrastructure:**

Diffusion models could require millions or billions of images to train. Moreover, to train such large datasets, massive compute power is needed, and AI practitioners must be able to procure and leverage hundreds of GPUs to train their models.

* **Lack of high-quality data:**

Troves of data are being generated globally every day, not all data can be used to train AI models. Because models require high-quality data like high-resolution image to train it, low resolution image will impact to low accuracy of AI models.

# **Relationship**

## **Generative AI and LLMs**

* + Many large language models, including GPT-3, are generative AI models. They can generate human-like text based on the input they receive.

## **Foundation Models and LLMs**

* + In the context of GPT-3, for example, it serves as both a large language model and a foundation model. It is pre-trained on diverse data and can be fine-tuned for specific tasks, making it a foundational building block for various applications.