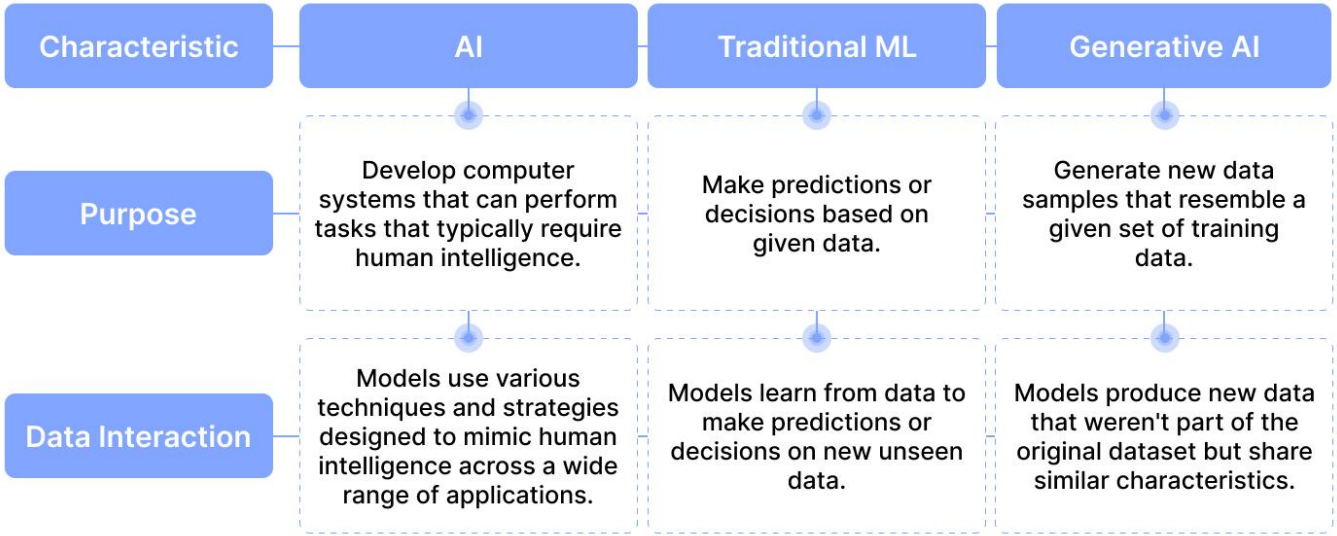
# Generative AI

01

# Gen AI

- Generative AI (genAI) is a broad label describing any type of artificial intelligence (AI) that can produce new text, images, video, or audio clips. Technically, this type of AI learns patterns from training data and generates new, unique outputs with the same statistical properties.
- Generative AI models use prompts to guide content generation and use transfer learning to become more proficient. Early genAI models were built with specific data types and applications in mind.

# Artificial Intelligence vs. Traditional Machine Learning, Generative AI



# How GenAI work?

- Generative AI, at its core, is about predicting the next piece of data in a sequence, whether that's the next word in a sentence or the next pixel in an image.
- GenAI is actually built on "Large Language Models" (LLM not AI), and these LLMs are based on the "Generative Pre-trained Transformer" architecture (or GPT—as invented by Google). These models learn patterns from a massive amount of text data (ChatGPT 3.5 was trained on 175B parameters, whilst ChatGPT-4 is trained on 1 Trillion parameters.)
- LLMs use these patterns (and not logic) to generate responses (they are basically *very* powerful autocorrect). Unlike other computer systems that are particularly good at math, LLMs are subject to "hallucinations," where they may generate seemingly meaningful responses, that are not otherwise correct.

# Type of GenAI

1. Generative Adversarial Networks(GANs)
2. variational autoencoders (VAEs)
3. Autoregressive models
4. Transformer-based models
5. Reinforcement learning for generative tasks

# Types of GenAI models

- Text to Text
- Text to Image
- Image to text
- Image to 3D
- Image or video to 3D
- Text to Audio
- Text to code
- Image to Science
- Text to video
- Audio to text

# How GenAI models are evaluated?

Generative AI models are evaluated in a more concise, point-by-point format:

1. **Objective Evaluation:**
   - Use a separate validation or test dataset not seen during training.
   - Measure quantitative metrics like accuracy, perplexity, or F1 score.
   - Assess the model's ability to generate outputs that match ground truth data.
2. **Subjective Evaluation:**
   - Human evaluators assess the relevance, coherence, and quality of generated outputs.
   - Conduct user studies or surveys to gather feedback on the usefulness of the model's outputs.

# Cont.

1. **Relevance and Quality Assessment:**
   - Evaluate how well the model's outputs align with the input prompt or task requirements.
   - Assess the diversity and creativity of generated outputs to avoid repetitive or biased results.
2. **Fine-Tuning and Retraining:**
   - Based on evaluation results, fine-tune the model's parameters to improve performance.
   - Consider retraining the model with additional data if it struggles with specific tasks or domains.
3. **Architecture Revision:**
   - Revisit the model's architecture if evaluation reveals fundamental issues or limitations.
   - Experiment with different architectures, such as larger networks or different attention mechanisms.

# Cont.

Popular metrics for assessing generative AI model performance include quantitative and/or qualitative scores for the following criteria:

- **Inception (IS) Score** assesses the quality and diversity of generated images.
- **Fréchet Inception Distance (FID) Score** assesses the similarity between the feature representations of real and generated data.
- **Precision and Recall Scores** assess how well-generated data samples match real data distribution.
- **Kernel Density Estimation (KDE)** estimates the distribution of generated data and compares it to real data distribution.

# Cont.

- **Structural Similarity Index (SSIM)** computes feature-based distances between real and generated images.
- **BLEU (Bilingual Evaluation Understudy) Scores** quantify the similarity between the machine-generated translation and one or more reference translations provided by human translators.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Scores** measure the similarity between a machine-generated summary and one or more reference summaries provided by human annotators.
- **Perplexity Scores** measure how well the model predicts a given sequence of words.
- **Intrinsic Evaluation** assesses the model's performance on intermediate sub-tasks within a broader application.

# Cont.

- **Extrinsic Evaluation** assesses the model's performance on the overall task it is designed for.
- **Few-Shot or Zero-Shot Learning** assesses the model's ability to perform tasks with very limited or no training examples.
- **Out-of-Distribution Detection** assesses the model's ability to detect out-of-distribution or anomalous data points.
- **Reconstruction Loss Scores** measure how well the model can reconstruct input data from the learned latent space.

# Large Language Models(LLM)

# LLM

- Large language models (LLMs) are a category of foundation models trained on immense amounts of data making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks.
- In a nutshell, LLMs are designed to understand and generate text like a human, in addition to other forms of content, based on the vast amount of data used to train them. They have the ability to infer from context, generate coherent and contextually relevant responses, translate to languages other than English, summarize text, answer questions (general conversation and FAQs) and even assist in creative writing or code generation tasks.

# Types of LLM

1. **Autoregressive Models:** These models generate text one token at a time based on the previously generated tokens. Examples include OpenAI's GPT series and Google's BERT.
2. **Conditional Generative Models:** These models generate text conditioned on some input, such as a prompt or context. They are often used in applications like text completion and text generation with specific attributes or styles.

# Difference between LLM & GenAI

**Generative AI** is like a big playground with lots of different toys for making new things. It can create poems, music, pictures, even invent new stuff!

**Large Language Models** are like the best word builders in that playground. They're really good at using words to make stories, translate languages, answer questions, and even write code!

So, generative AI is the whole playground, and LLMs are the language experts in that playground.

# Open LLMs

# T5

- Release on Oct 2019
- T5 (Text-to-Text Transfer Transformer) is a large language model (LLM). It was developed by Google Research and is based on the Transformer architecture. T5 is designed for various text-related tasks, where both inputs and outputs are in the form of text.
- T5 differs from some other LLMs in its approach to handling tasks. Instead of designing specific architectures for different tasks (e.g., question answering, text generation, translation), T5 uses a unified framework where tasks are represented as text-to-text transformations. This approach allows T5 to handle a wide range of tasks by framing them as text transformation tasks, making it a versatile and powerful LLM.

# UL2

- Release on Oct 2022
- UL2 stands for **Unifying Language Learning Paradigms**. It's a framework for creating pre-trained language models that can be applied to a wide range of tasks. Here's a breakdown of how it works:
- **Core Idea:** Traditionally, different pre-training methods are used for different tasks. UL2 proposes a single, unified approach that works well across various tasks.
- **Mixture-of-Denoisers (MoD):** This is UL2's key concept. It combines multiple pre-training techniques into one objective function. This allows the model to learn from various data types and improve its overall performance.
- **Mode Switching:** UL2 can be fine-tuned for specific tasks by selecting the most appropriate pre-training scheme within MoD. This helps the model adapt to the specific demands of the task at hand.

Overall, UL2 aims to create versatile language models that can be effectively used for various purposes.

# Dolly

- Release on April 2023
- Dolly  refers to **Databricks Dolly**, an LLM created by Databricks. Here's what makes Dolly unique:
- **Instruction-Following:** Unlike some LLMs that focus on general understanding, Dolly excels at following specific instructions. You can provide clear instructions, and Dolly will try its best to complete the task as instructed.
- **Open-Source and Commercially Available:** Databricks offers Dolly in two versions:
    - **Open-source version (databricks/dolly-v2-12b):** This freely available version allows anyone to experiment with Dolly and its capabilities.
    - **Commercially licensed version (databricks-dolly-15k):** This paid version offers additional features and training data, making it suitable for enterprise use.
- **Focus on Enterprise Needs:** Databricks designed Dolly specifically for enterprise applications. It's particularly useful for tasks like data analysis, report generation, and code generation within a company setting.

# DLite

- Release on May 2023
- DLite refers to a specific model called **D-Lite** developed by AI Squared. Here's what makes DLite interesting:
- **Lightweight and Efficient:** DLite is a lightweight LLM compared to other models like GPT-3. This means it requires less computational power to run, making it suitable for deployment on devices with limited resources like laptops or even powerful smartphones.

- **ChatGPT-like Interaction:** Despite its smaller size, DLite exhibits impressive capabilities for conversation and interaction. It can follow instructions and engage in open-ended dialogue similar to ChatGPT.

- **Fine-Tuning and Customization:** DLite is designed to be fine-tuned on specific datasets. This allows users to tailor the model to their particular needs and domains.

- **Open-Source and Accessible:** DLite is available as an open-source model, making it accessible to anyone who wants to experiment with it.

# Bloom

- Release on Nov 2022
- Bloom is a 176-billion parameter open-source multilingual LLM created by BigScience.

Here are some key features of Bloom:

- **Multilingual Capabilities:** Bloom can process and generate text in 46 natural languages and 13 programming languages. This makes it a valuable tool for tasks involving multiple languages.
- **Open-Source and Accessible:** Bloom is publicly available, allowing anyone to access and experiment with the model. You can find it on platforms like Hugging Face.
- **Focus on Research:** The model was designed to facilitate research on LLMs. Researchers can use Bloom as a starting point for fine-tuning models for specific tasks or exploring LLM capabilities further.

# OpenLLaMA

- Release on May 2023
- OpenLLaMA is an open-source project that offers a complete training pipeline for building large language models (LLMs). It's a non-gated version of another LLM called LLaMA, developed by Meta AI. Here's a breakdown of OpenLLaMA:
  - **Open-Source and Permissively Licensed:** Anyone can access, study, and modify OpenLLaMA, making it a valuable resource for researchers and developers.
  - **Complete Training Pipeline:** OpenLLaMA provides all the necessary steps for training LLMs, from data preparation and tokenization to pre-training and fine-tuning. This simplifies the process for those who want to build their own LLMs.
  - **Multiple Model Sizes:** The project offers different model sizes, ranging from 3 billion to 13 billion parameters, allowing users to choose a model that best suits their computational resources and needs .
  - **Focus on Research and Commercial Applications:** OpenLLaMA's open-source nature makes it valuable for research on LLMs, while its permissive licensing allows for commercial use as well .

# Falcon

- Release on May 2023

Falcon LLM is a powerful and open-source large language model (LLM) developed by the Technology Innovation Institute (TII) based in Abu Dhabi. Here are some key aspects of Falcon LLM:

- **Open-Source and Accessible**
- **Multiple Model Sizes:** Falcon LLM comes in various sizes, with the most prominent ones being Falcon-40B (40 billion parameters) and Falcon-7B (7 billion parameters). This allows users to choose a model that best suits their computational resources and task requirements. A wider range of sizes including 180B, 1.3B, and 7.5B models is also offered.
- **Strong Text Generation:** Falcon LLM excels at generating different creative text formats, like poems, code, scripts, musical pieces, and more.
- **Focus on Future-Proofing Applications:** The developers designed Falcon LLM to contribute to advancements in various applications and use cases. This could involve tasks like machine translation, chatbot development, content creation, and more.

# Gemma

- Release on Feb 2024

Gemma is a family of open-source large language models (LLMs) developed by Google AI as part of the Gemini initiative. Here are some key features that make Gemma stand out:

- **Open-Source and Accessible**
- **Multiple Model Sizes:** Gemma comes in two sizes: 2 billion and 7 billion parameters. This offers flexibility for users depending on their computational resources. The 2 billion parameter model can even run on a personal computer, while the 7 billion parameter model is suitable for more powerful setups.
- **Strong Generalist Capabilities:** Despite their smaller size compared to some LLMs, Gemma models exhibit impressive performance in various tasks like question answering, code generation, and creative text formats.
- **Instruction Tuning:** Gemma offers both base and instruction-tuned variants. Base models are pre-trained on a massive dataset, while instruction-tuned models are further fine-tuned on specific tasks or domains, potentially improving their performance for those particular uses.

# Reference


- https://www.techopedia.com/definition/34633/generative-ai
- https://yellow.ai/blog/types-of-generative-ai/
- https://www.analyticsvidhya.com/blog/2023/03/an-introduction-to-large-language-models-llms/
- https://www.databricks.com/resources/webinar/build-your-own-large-language-model-dolly