Introduction to Large Language Models (LLM)

By Raghava
3rd Year,
M.Sc Artificial Intelligence and Machine Learning

Language Model (LM)

A language model is a computer program that is trained on a large amount of text data to generate language outputs that are coherent and natural-sounding. It predicts the next word or character in a sequence of text, based on the context and patterns learned from the training data.

Input: "The cat sat on the _____"

Language Model Output: "mat/floor/etc..."

Explanation: The language model predicts the next word "mat" based on the context of the sentence and the patterns learned from the training data.

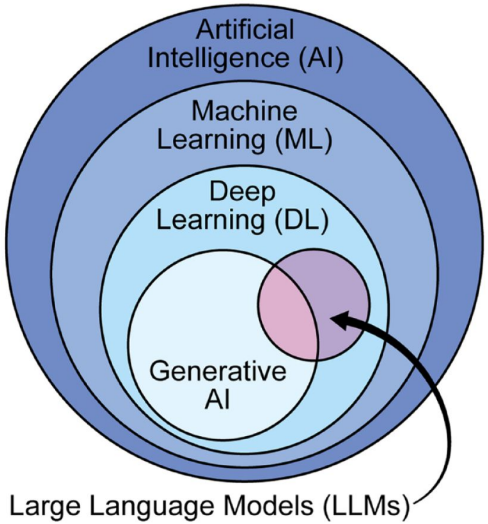
Some popular language models include BERT and Transformers.

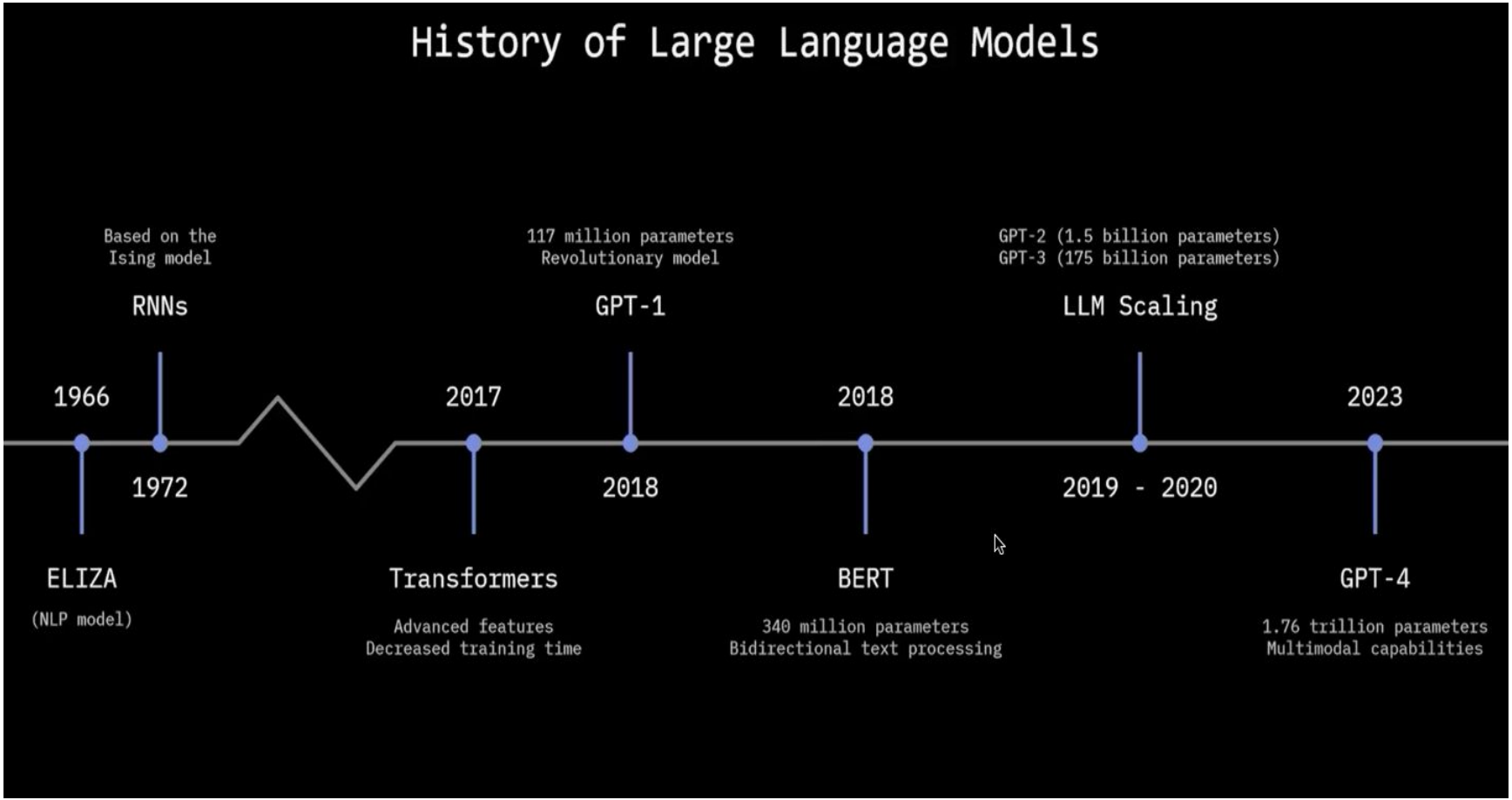
What is LLM?

LLM stands for Large Language Model. A Large Language Model is a type of artificial intelligence (AI) model that is trained on a vast amount of text data to generate language outputs that are coherent and natural-sounding. LLMs are designed to process and understand human language, and they have numerous applications, possibly every domain.

Popular LLMs:

- ChatGPT by OpenAI
- Gemini & Gemma by Google
- llama by Meta




History of Large Language Models
1966
ELIZA
(NLP model)
1972
RNNs
Based on the
Ising model
2017
Transformers
Advanced features
Decreased training time
2018
GPT-1
117 million parameters
Revolutionary model
2018
BERT
340 million parameters
Bidirectional text processing
2019 - 2020
LLM Scaling
GPT-2 (1.5 billion parameters)
GPT-3 (175 billion parameters)
2023
GPT-4
1.76 trillion parameters
Multimodal capabilities

Foundational models

Foundational models in large language models (LLMs) refer to the initial models that serve as the basis or foundation for further research, development, and fine-tuning within the field of natural language processing (NLP). These models are typically pre-trained on vast amounts of text data using unsupervised learning techniques, such as self-supervised learning or autoencoding, to learn the underlying patterns and structures of language.

GPT (Generative Pre-trained Transformer) developed by OpenAI

BERT (Bidirectional Encoder Representations from Transformers) developed by Google

T5 (Text-To-Text Transfer Transformer) developed by Google Brain

LLM Modalities

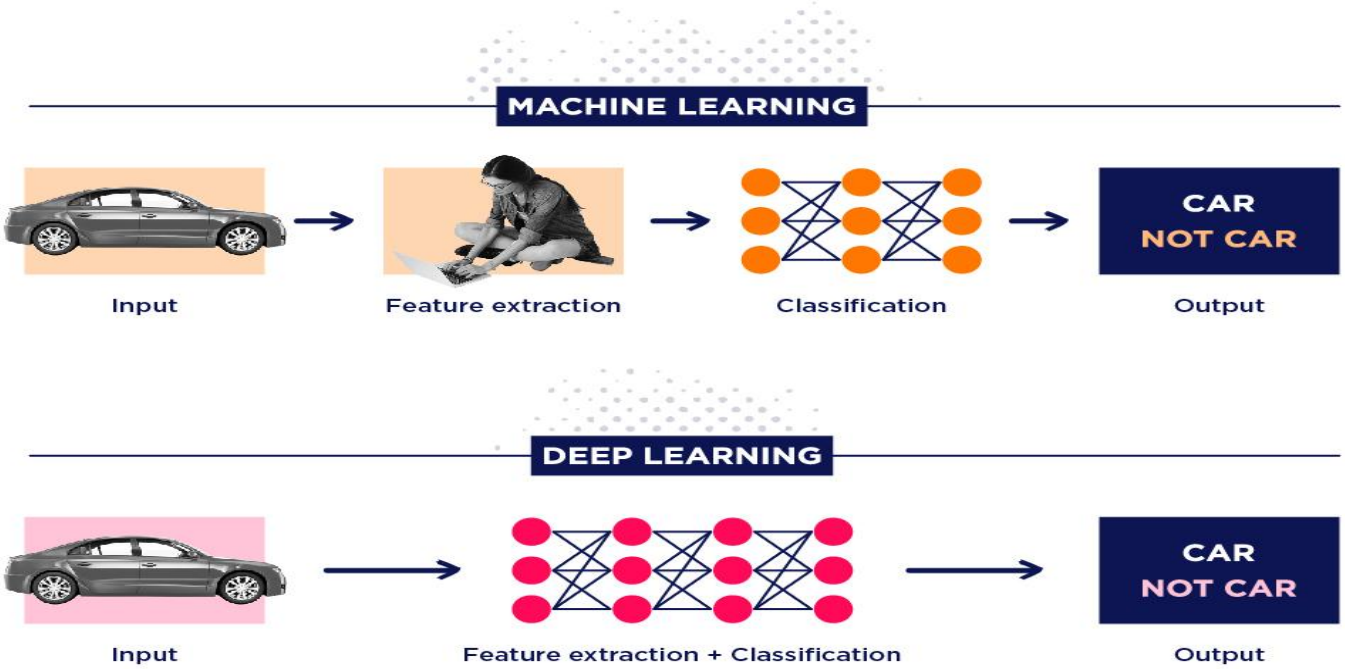
TEXT MODELS: These models focus primarily on processing and generating text data. They are trained on large text corpora and excel at tasks such as language modeling, text generation. *GPT (Generative Pre-trained Transformer) models developed by OpenAI and BERT (Bidirectional Encoder Representations from Transformers) developed by Google.*

IMAGE-TEXT MODELS: LLMs that integrate both text and image modalities can understand and generate content based on both textual and visual inputs. *Gemini developed by Google.*

TEXT-VIDEO MODELS: Given a text prompt as a input, these models can generate video based on the context from the prompt and make a new real life video. *Sora by OpenAI*

MULTIMODAL MODELS: These models combine information from multiple modalities (text, image, audio) to generate coherent outputs that leverage the strengths of each modality. *Google AI's PaLM-E, OpenAI's GPT-4V*

Training ML vs DL



ITERATORS
Source: www.iteratorshq.com

Types of LLMs

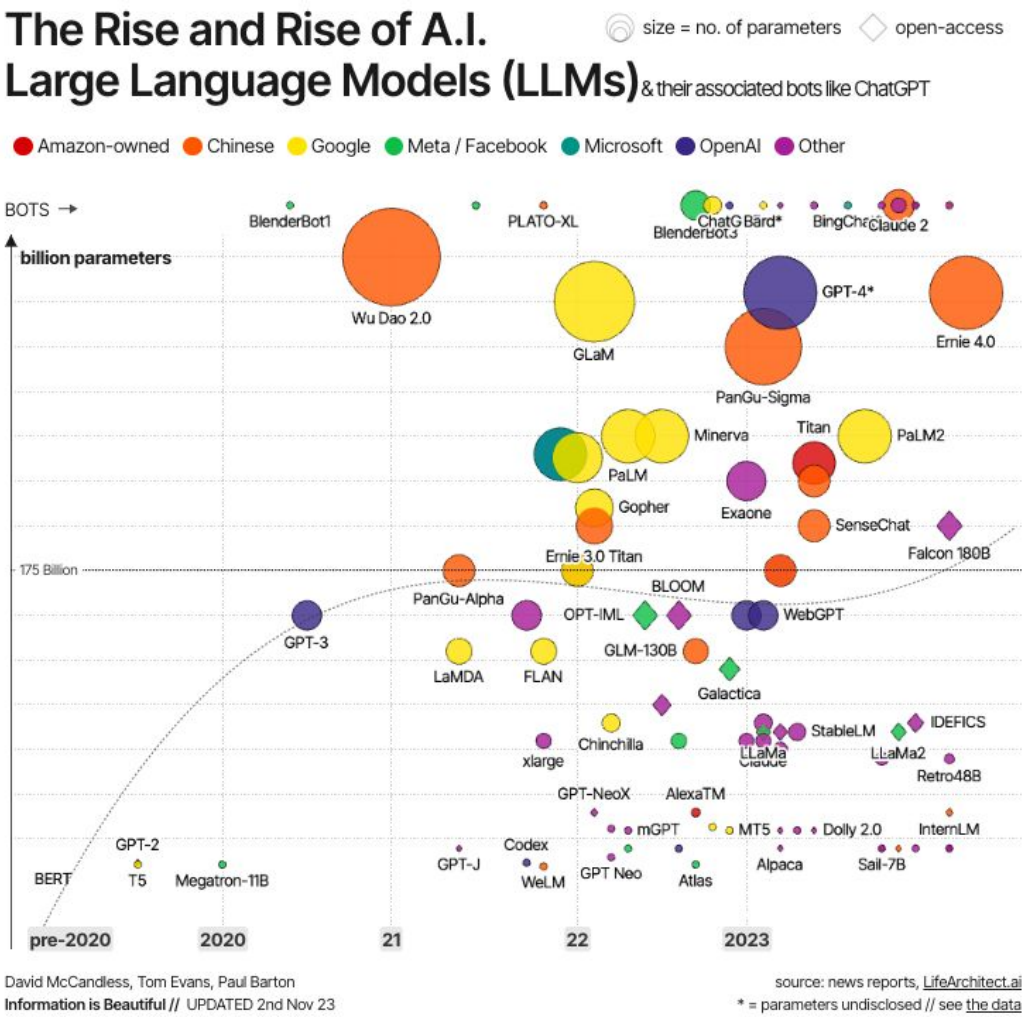
AUTOREGRESSIVE LLMS: Generate text one token at a time, predicting the next token based on previous context. Examples: GPT-3, GPT-4.

AUTOENCODER LLMS: Compress input text into a smaller representation and then reconstruct the original text. Examples: BERT, RoBERTa.

DECODER-ONLY LLMS: Use only the decoder part of the transformer architecture, generating text without encoding input text. Examples: GPT-3, GPT-4.

ENCODER-DECODER LLMS: Use both encoder and decoder parts of the transformer architecture, encoding input text and generating output text. Examples: BERT, RoBERTa.

More...



Source: Datacamp

Buzz words in the world of LLMs

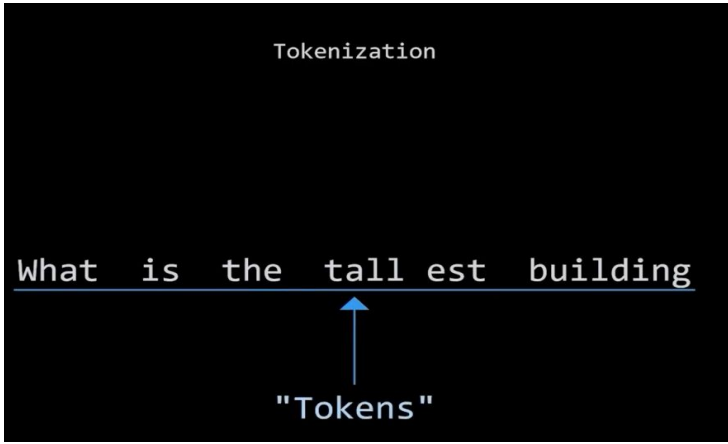
PARAMETERS:

The purpose of having a large number of parameters in an LLM is to increase its capacity to learn complex patterns and relationships in language. More parameters allow the model to capture finer nuances of language, including syntax, semantics, context, and domain-specific knowledge. larger models with more parameters also require more computational resources for training and inference.
Example: Gemma 2B, 7B

- A 7B (7 billion) parameter model has 7,000,000,000 learnable weights and biases.
- A 2B (2 billion) parameter model has 2,000,000,000 learnable weights and biases.

TOKENIZATION:

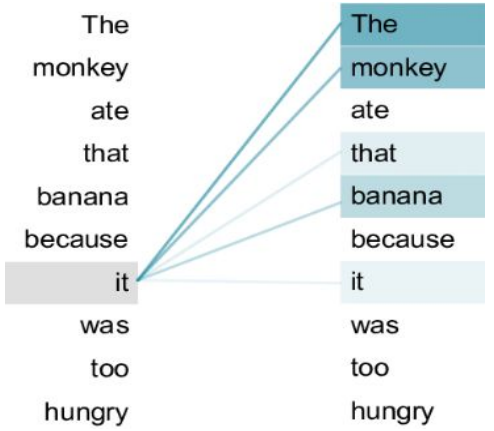
It is the process of breaking down a sequence of text into smaller units called tokens. These tokens can be words, subwords, characters, or any other meaningful units, depending on the specific tokenization strategy used. Tokenization is a fundamental step in natural language processing (NLP) tasks, as it allows machines to understand and process human language.



Buzz words in the world of LLMs

SELF ATTENTION:

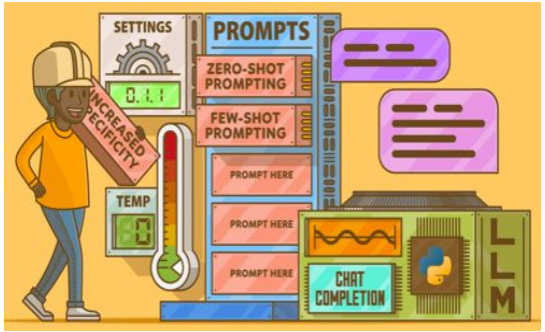
Self-attention is a mechanism that computes the importance of each token in a sequence by considering its relationships with all other tokens in the same sequence. It allows the model to assign different levels of importance or attention to each token based on its relevance to other tokens in the sequence.



PROMPT ENGINEERING:

Prompt engineering refers to the process of designing or crafting natural language prompts or instructions to control the behavior of large language models (LLMs) and elicit desired responses. It involves carefully constructing input text that guides the LLM to produce outputs aligned with specific objectives or tasks. *It includes Instructions, Examples, Constraints, Contextual Information.*

Example: Give me 300 words assignment content on methods of classification in machine learning with examples include KNN, K-Means etc.



Buzz words in the world of LLMs

HALLUCINATION

It refers to a phenomenon where the model generates text or outputs that are not grounded in the input data or are inconsistent with reality. Hallucinations occur when the model generates content that is either unrelated to the input prompt or contains factual inaccuracies, contradictions, or nonsensical information.

Confidently incorrect: The model produces answers that are wrong, but with high confidence, as if it believes they are correct.

Not supported by training data: The responses are not based on any actual training data or evidence, but rather the model's own internal workings.

Often plausible-sounding: Hallucinated responses can be grammatically correct, coherent, and even persuasive, making them difficult to detect.

Example:
What is crane?

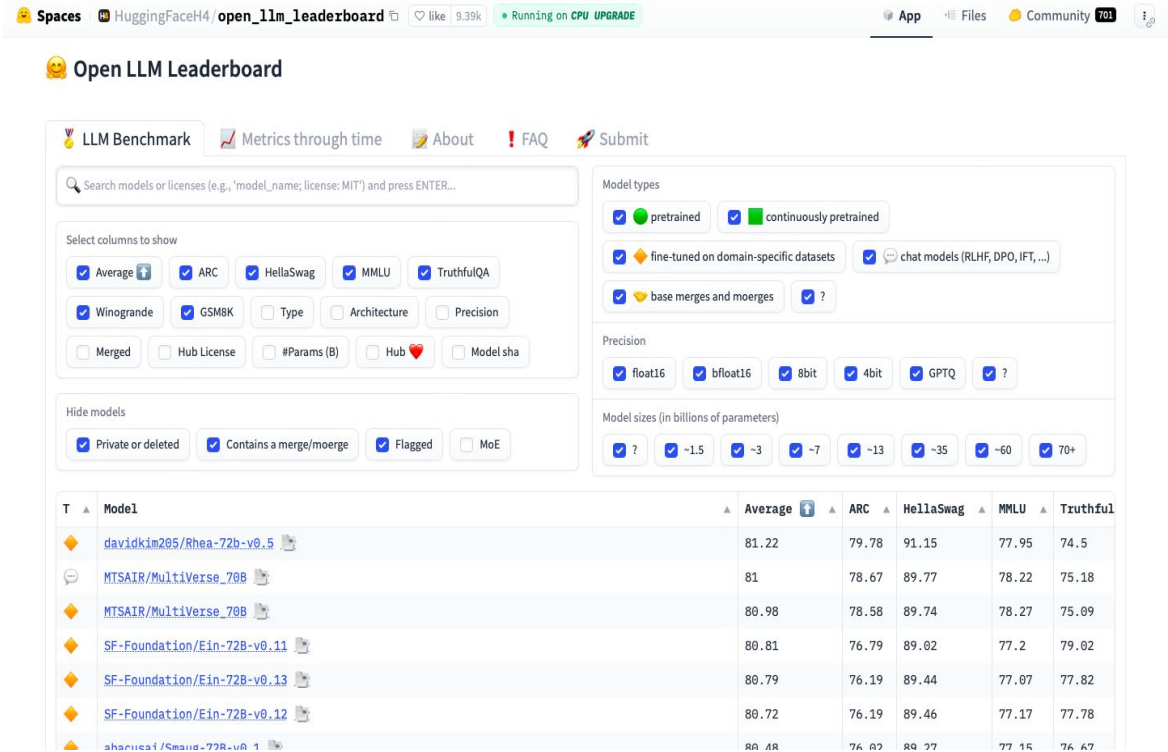
LLM Leaderboard

Tracks open-source LLMs and chatbots: It focuses on models freely available for everyone to use and improve.

Ranks models based on benchmarks: It evaluates models on various tasks like question answering and summarization.

Provides detailed results: You can explore how each model performs on individual benchmarks.

Collection of top models: It features a curated list of high-performing LLMs.



https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Evaluation Metrics for LLMs

ARC (question answering): Evaluates a model's ability to answer questions based on a given text.

HellaSwag (natural language understanding): Assesses a model's ability to understand natural language by identifying the correct ending to a story.

TruthfulQA (factual accuracy): Evaluates a model's ability to provide accurate and truthful answers to questions.

MMLU (massive multitask learning): Tests a model's ability to perform well across a wide range of tasks simultaneously.

ROUGE scores (summarization quality): Measures the quality of a model's summarization capabilities by comparing it to human-written summaries.

BLEU scores (machine translation quality): Evaluates the quality of a model's machine translation outputs by comparing them to human-translated references.

Applications of LLM

Chatbots and Virtual Assistants: Conversational AI, enabling chatbots to understand and respond to user queries.

Text Summarization: Summarize long texts, extracting key points and saving time.

Content Generation: Create content, such as articles, stories, and even entire books.

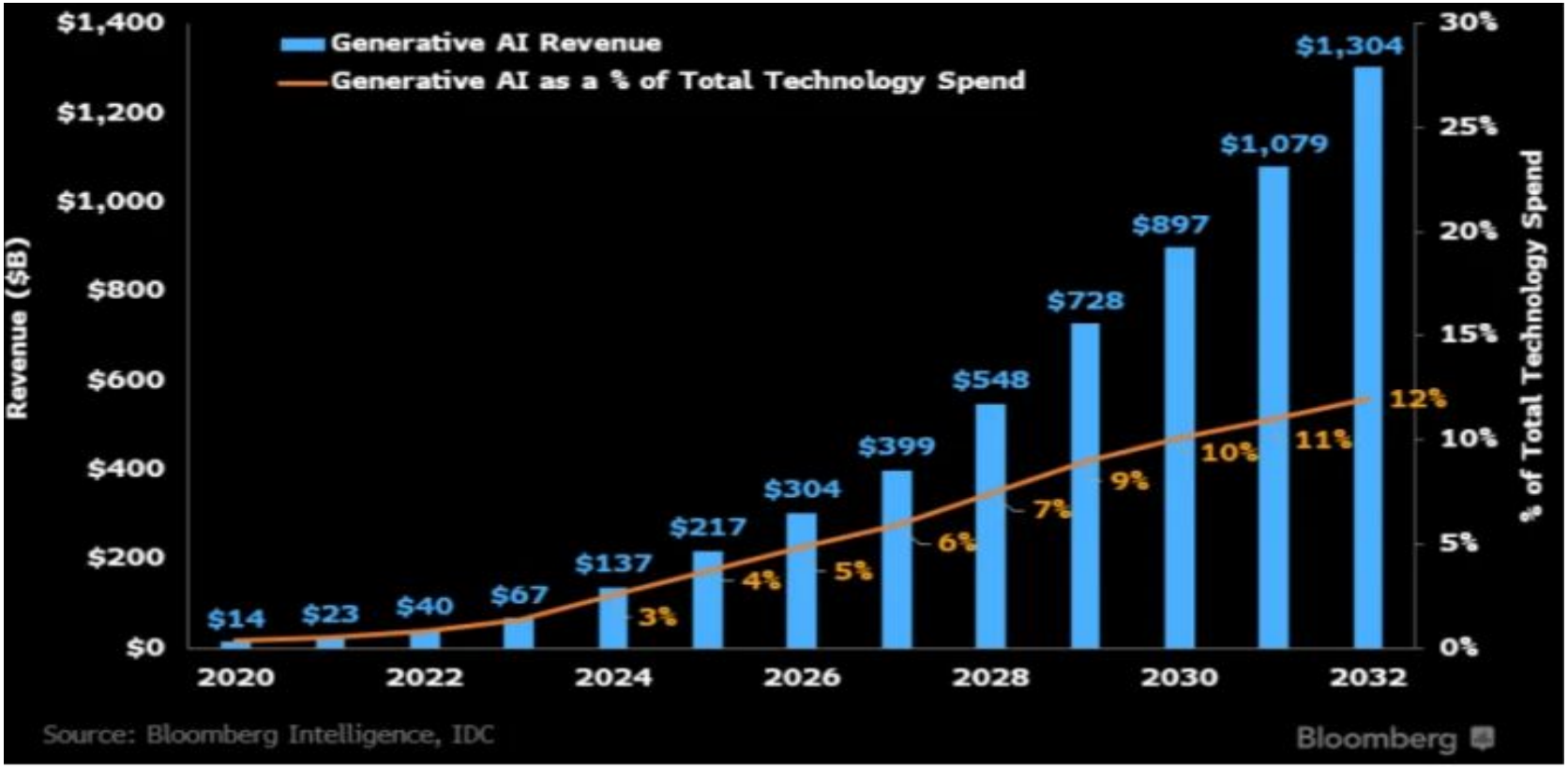
Question Answering: Answer questions, providing accurate and relevant information based on context.

Creative Writing: Assist writers with ideas, suggestions, and even co-authoring content.

Customer Service: Power automated customer support, providing quick and accurate solutions.

WHAT NOT ?! Much more …

Gen AI in business



New Job Roles

Prompt Engineer

AI Trainer

AI Ethics Officer

Data Annotation Specialist

NLP Engineer

AI Researcher

Urban Planner

AI Personality Designer

These a just a few…

Quiz

Since the last quarter of year 2022, (ie. The launch & rise of LLMs, Generative AI such as ChatGPT, Gemini etc…).

Which company had a significant rise in their market share and net profit?

NVIDIA

To train and run models effectively. Nvidia's GPUs (Graphics Processing Units) are particularly well-suited for this task, due to their parallel processing architecture.

Nvidia A100 GPUs. Nvidia is ramping up its next-generation of GPUs, called H100, which delivers six times more throughput than the A100.

Resources to learn...

Youtube Channels:

Krish Naik: https://www.youtube.com/@krishnaik06 *(For beginners)*

Deep Learning AI by Prof. Andrew NG : https://www.youtube.com/@Deeplearningai

CodeBasics: https://www.youtube.com/@codebasics

Assembly AI: https://www.youtube.com/@AssemblyAI

Varun Mayya: https://www.youtube.com/@VarunMayya *(Know new inventions)*

Nicholas Renotte: https://www.youtube.com/@NicholasRenotte

Resources to learn...

Websites:

Hugging Face: https://huggingface.co/learn/nlp-course/chapter0/1

Cohere: https://docs.cohere.com/docs/llmu

Datacamp: https://www.datacamp.com/courses/large-language-models-llms-concepts

Deep Learning AI: https://www.deeplearning.ai/courses/

Full Stack DL: https://fullstackdeeplearning.com/llm-bootcamp/

KDNuggets: https://www.kdnuggets.com/a-comprehensive-list-of-resources-to-master-large-language-models

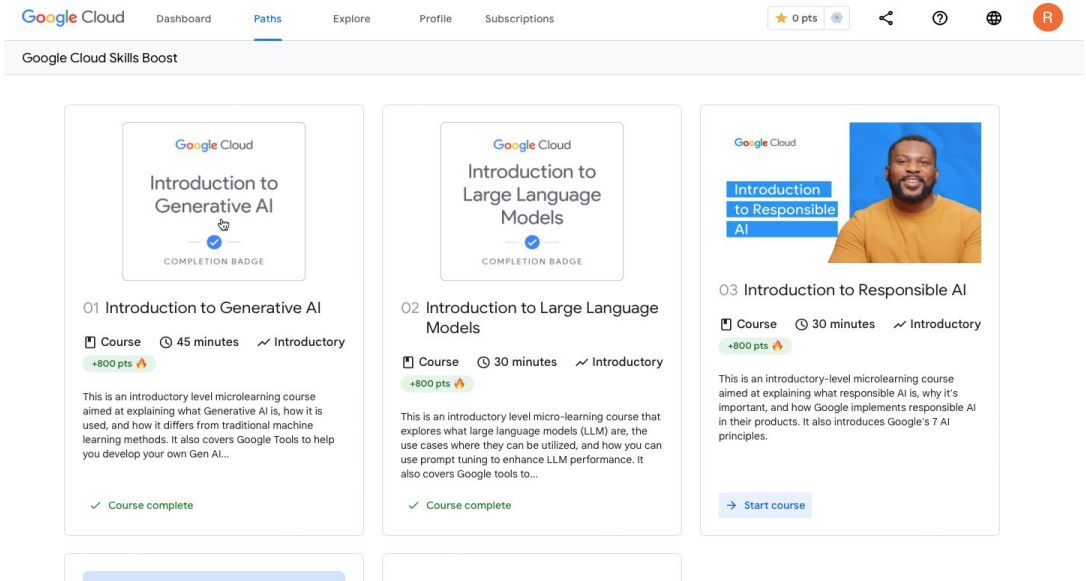
Resources (BONUS)

You Can get free access to learn LLM & Generative AI in Google.

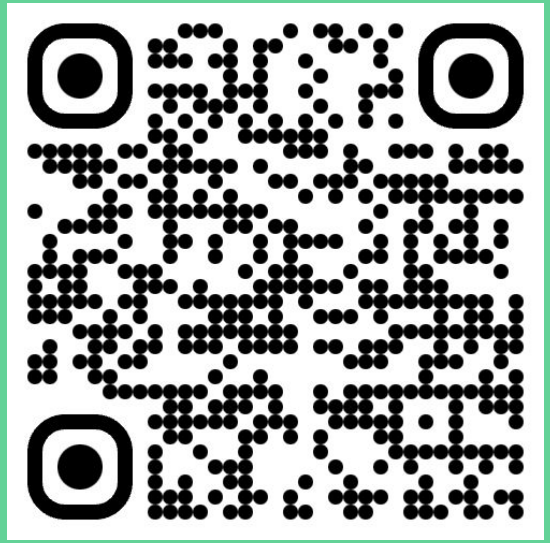
Google offers FREE BADGE Course for beginners and students

Checkout:

https://www.cloudskillsboost.google/catalog



Thankyou



Connect with ME!
@raghavtwenty