KAMARAJ IAS ACADEMY
Only IAS Academy by Grandson of "Perunthalaivar Kamarajar"

Large Language Model

Published On: 27-02-2024

Why is in news? What is an LLM, the backbone of AI chatbots like ChatGPT, Gemini?

Ever since the launch of OpenAI's sensational chatbot ChatGPT, conversations about artificial intelligence have become common from living rooms to boardrooms.

When computers were invented, they were machines that executed instructions given by programmers. Now, computers have now gained the ability to learn, think and hold conversations.

Not only that, they can perform several creative and intellectual tasks once only limited to humans. This is what we call generative AI.

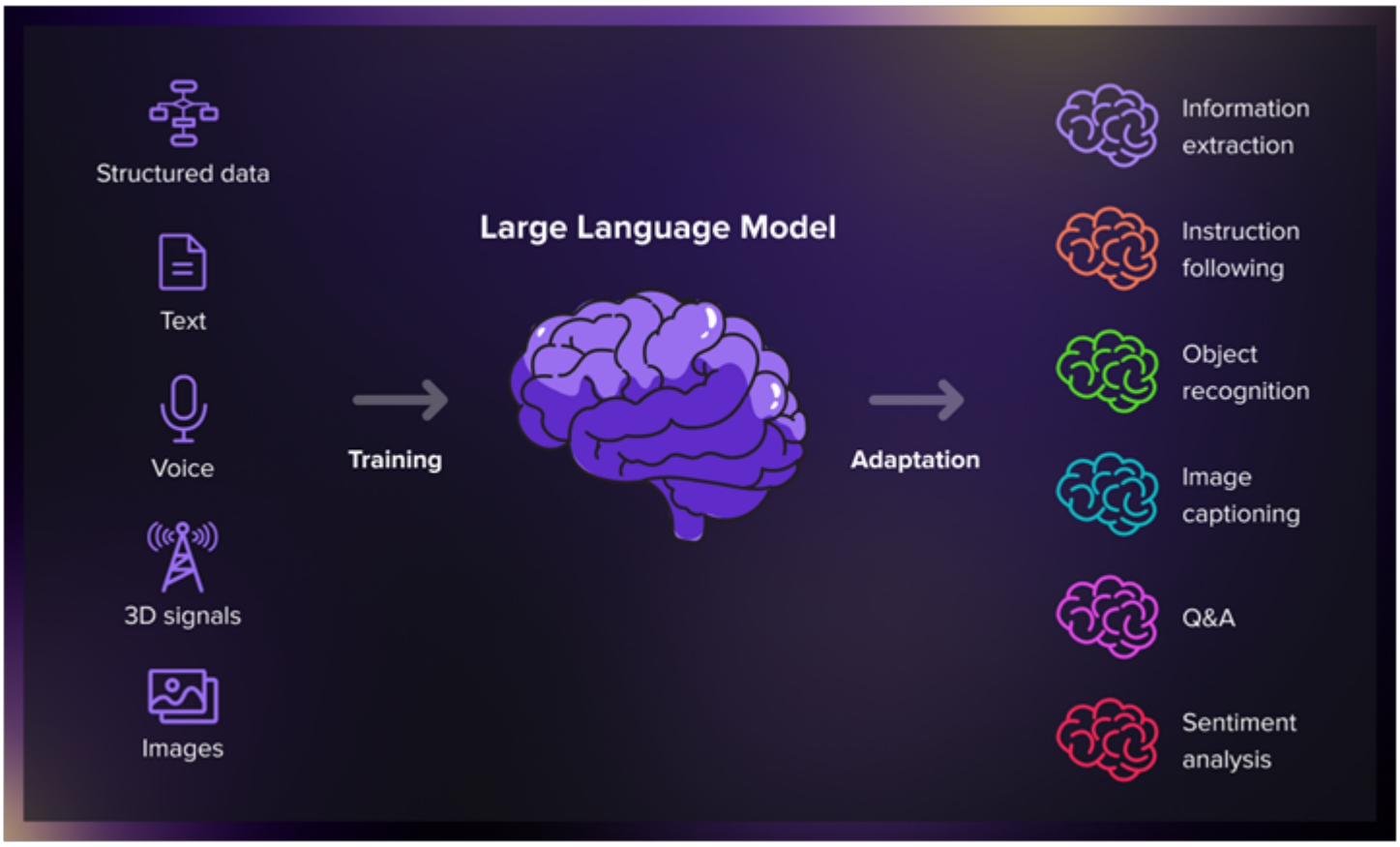
The **ability of Generative AI models** to "converse" with humans and predict the next word or sentence is due to something known as the **Large Language Model, or LLM**.

It is to be noted that while **not all generative AI tools are built on LLMs, all LLMs are forms of Generative AI** which in itself is a **broad and ever-expanding category or type of AI**

LLM:

According to Google, LLMs are **large general-purpose language models** that can be pre-trained and then fine-tuned for specific purposes.

In simple words, these models are trained to solve common language problems such as text classification, question answering, text generation across industries, document summarisation, etc



The LLMs can also be **tailored to solve specific problems** in a variety of domains such as finance, retail, entertainment, etc., using perhaps a relatively small size of field datasets.

Three primary features of LLM:

The '**Large**' indicates two meanings — the enormous size of training data; and the parameter count. In Machine Learning, parameters, also known as **hyperparameters**, are essentially the memories and knowledge that a machine learned during its model training. Parameters define the skill of the model in solving a specific problem.

The second most important thing to understand about LLM is the **General Purpose**. This means the model is sufficient to solve general problems that are based on the commonality of human language regardless of specific tasks, and resource restrictions.

In essence, an LLM is like a **super smart computer program** that can comprehend and create human-like text. It is trained on massive data sets which are essentially patterns, structures, and relationships with languages. An LLM can also be seen as a tool that helps computers understand and produce human language.

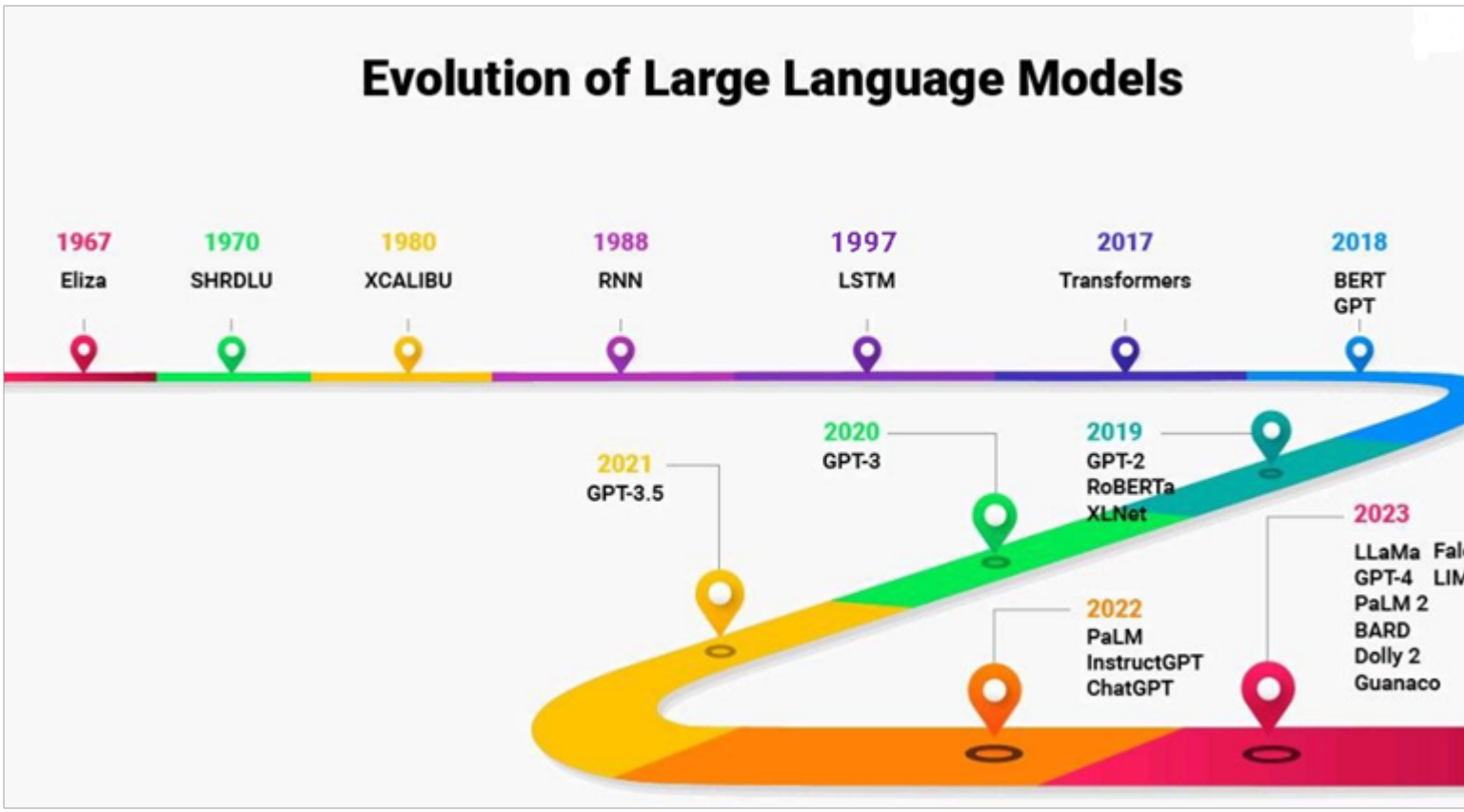
Types of LLMs:

There are various ways to categorise LLMS. It is to be noted that the type depends on the specific aspect of tasks they are meant to do.

Basis of architecture: **Autoregressive, transformer-based, and encoder-decoder**.

GPT-3 is an example of an **autoregressive model** as they predict the next word in a sequence based on previous words

Similarly, **LaMDA or Gemini** (formerly Bard) are **transformer-based** as they use a specific type of neural network architecture for language processing



Then there are the **encoder-decoder** models that encode input text into a representation and then decode it into another language or format.

Based on **training data**: **Pre-trained and fine-tuned, multilingual** or models that can understand and generate text in multiple languages, and **domain-specific** or models that are trained on data related to specific domains such as legal, finance or healthcare.

Based on their size: Large models usually require more computational resources. However, they offer better performance.

Based on availability: Categorised as open-source and closed-source as some are freely available while some are proprietary.

LLaMA2, BlOOM, Google BERT, Falcon 180B, OPT-175 B are some open-source LLMs, while **Claude 2, Bard, GPT-4,** are some proprietary LLMs.

Working of LLMs:

At the core of it is a technique known as "**deep learning**".

It involves the **training of artificial neural networks**, which are mathematical models which are believed to be inspired by the structure and functions of the human brain.

For LLMs, this neural network learns to predict the probability of a word or sequence of words given the previous words in a sentence.

As mentioned earlier, this is done by analysing the patterns and relationships between words in the data set used for training.

Once trained, an LLM can predict the most likely next word or sequence of words based on inputs also known as **prompts.**

An LLM's learning ability can be best described as similar to how a baby learns to speak. You don't give a baby an instruction manual, he/she learns to understand language by listening to people speak.

Generative Pre-trained Transformers (GPTs):

GPTs are a **type of large language model (LLM)** that use transformer neural networks to generate human-like text.

GPTs are **trained on large amounts of unlabelled text data** from the internet, enabling them to understand and generate coherent and contextually relevant text.

They can be **fine-tuned for specific tasks** like: Language generation, Sentiment analysis, Language modelling, Machine translation, Text classification.

GPTs **use self-attention mechanisms** to focus on different parts of the input text during each processing step.

This allows GPT models to capture more context and improve performance on **natural language processing (NLP)** tasks.

NLP is the ability of a computer program to understand human language as it is spoken and written -- referred to as natural language.

Applications of LLMs:

LLMs come with an array of applications across domains.

They **generate text** and are **capable of producing human-like content** for purposes ranging from stories to articles to poetry and songs. They can strike up a conversation or function as virtual assistants.

In conversational settings, LLMs engage with users, providing information, answering questions, and maintaining context over multiple exchanges.

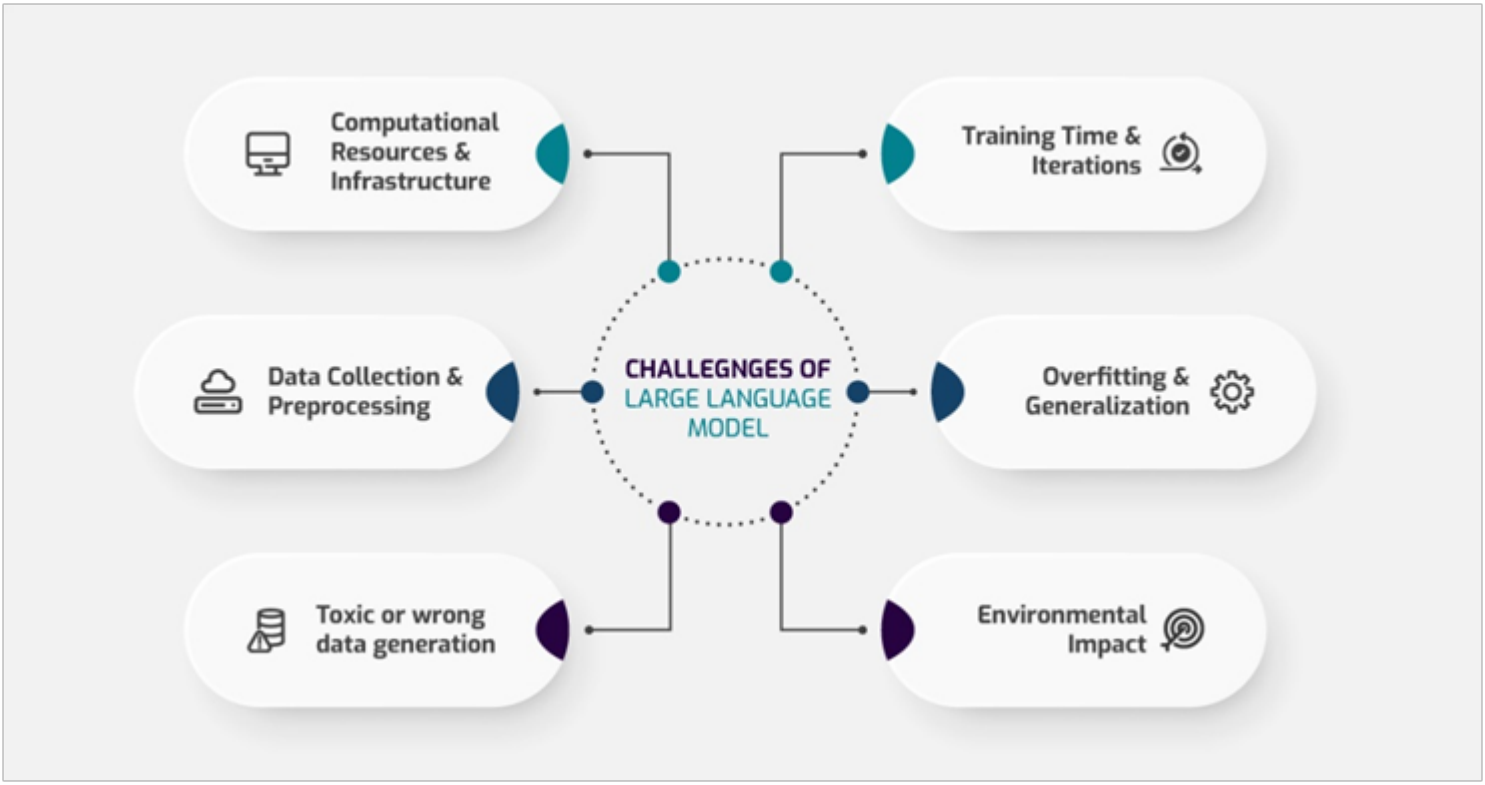
Additionally, they play **a crucial role in content creation and personalisation**, aiding in marketing strategies, offering personalised product recommendations, and tailoring content to specific target audiences.

Advantages of LLMs:

Perhaps, the biggest advantage of LLMs is their **versatility**.

A single model can be used for a wide variety of tasks. Since they are trained on large data sets, they are capable of generalising patterns which can be later applied to different problems or tasks

When it comes to data, LLMs can reportedly perform well even with limited amounts of domain or industry-specific data. This is possible because LLMs can leverage the knowledge they learned from general language training data.



Another important aspect is their **ability to continuously improve** their performance. As more data and parameters are infused into LLMs, their performance improves.

LLMs are **continuously developing and proliferating into new dimensions**.

Possible future for LLM developments:

Continued improvements in the accuracy and capabilities of large language models, allowing them to understand and generate more complex and nuanced language.

Expansion of the applications of large language models beyond text-based tasks, such as speech recognition and natural language understanding in virtual assistants and chatbots.

Integration of large language models with other technologies, such as computer vision and knowledge graphs, to create more powerful and versatile AI systems.

Development of more efficient and sustainable training methods, such as using smaller models or unsupervised learning techniques, to reduce the computational and energy costs of training large language models.