DAY – 10

11 August 2025

Machine Learning (ML) Large Language Models (LLMs) are advanced AI systems using deep learning (Transformers) trained on massive text datasets to understand, generate, and predict human language, enabling tasks like translation, chatbots, and content creation, by learning complex patterns and context far beyond older models, though requiring huge resources. They function by predicting the next "token" (word/subword) in a sequence, using self-attention to weigh word importance, making them powerful tools for natural language understanding and generation.

How LLMs Work

- Deep Learning & Transformers: LLMs are built on transformer neural networks, processing entire text sequences in parallel (unlike older RNNs) for faster training on GPUs.
- Tokens & Context: They break text into tokens (words/subwords) and use self-attention to understand relationships and context, assigning weights to how much each token influences others.
- Training: Trained on vast amounts of text (books, web pages), they learn grammar, facts, and reasoning by predicting missing words or sequences.
- Parameters: They have billions or trillions of parameters (variables), allowing them to capture intricate language nuances, making them far more complex than traditional models.

Machine Learning & LLMs

- LLMs are a specialized, advanced form of machine learning, specifically deep learning.
- They represent a leap over older NLP models due to scale (data & parameters) and the transformer architecture, enabling emergent abilities (complex reasoning, coding).

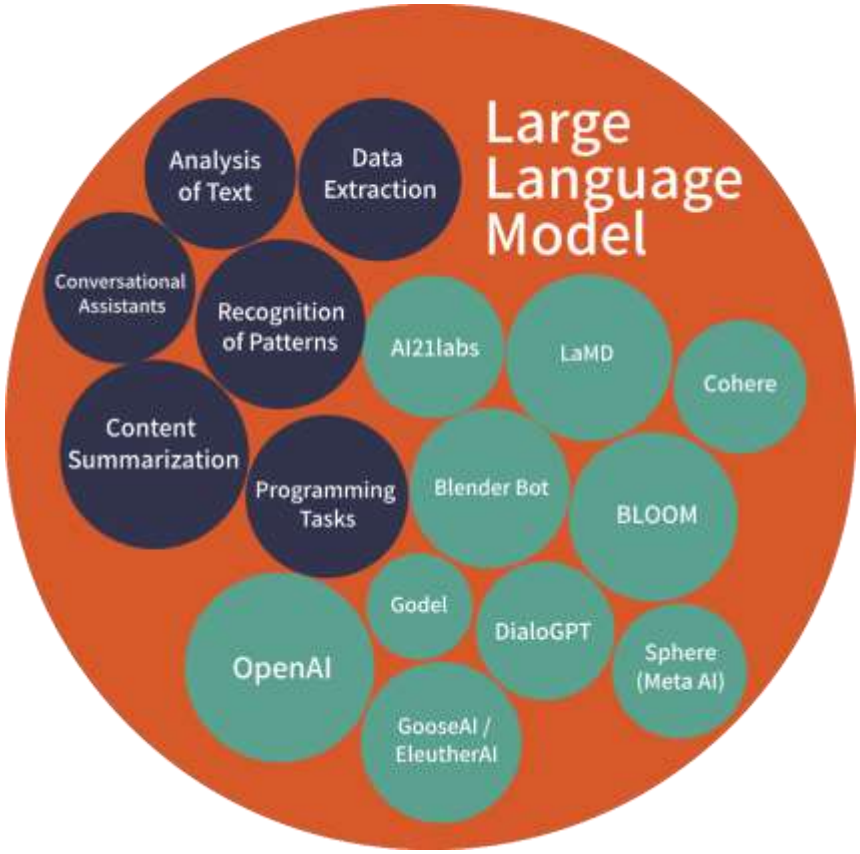
❖ **Popular LLMs**

Some of the most widely used LLMs include:

- GPT-4 and GPT-4o (OpenAI): Advanced multimodal reasoning and dialogue capabilities.
- Gemini 1.5 (Google DeepMind): Long-context reasoning, capable of handling 1M+ tokens.
- Claude 3 (Anthropic): Safety-focused, strong at reasoning and summarization.
- LLaMA 3 (Meta): Open-weight model, popular in research and startups.
- Mistral 7B / Mixtral (Mistral AI): Efficient open-source alternatives for developers.
- BERT and RoBERTa (Google/Facebook): Strong embedding models for NLP tasks.
- mBERT and XLM-R: Early multilingual LLMs.
- BLOOM: Large open-source multilingual model, collaboratively developed.

❖ **Use Cases**

- Code Generation: LLMs can generate accurate code based on user instructions for specific tasks.
- Debugging and Documentation: They assist in identifying code errors, suggesting fixes and even automating project documentation.
- Question Answering: Users can ask both casual and complex questions, receiving detailed, context-aware responses.
- Language Translation and Correction: LLMs can translate text between over 50 languages and correct grammatical errors.
- Prompt-Based Versatility: By crafting creative prompts, users can unlock endless possibilities, as LLMs excel in one-shot and zero-shot learning scenarios.



How do large language models work?

- **Machine learning and deep learning**

At a basic level, LLMs are built on machine learning. Machine learning is a subset of AI, and it refers to the practice of feeding program large amounts of data in order to train the program how to identify features of that data without human intervention.

LLMs use a type of machine learning called deep learning. Deep learning models can essentially train themselves to recognize distinctions without human intervention, although some human fine-tuning is typically necessary.

Deep learning uses probability in order to "learn." For instance, in the sentence "The quick brown fox jumped over the lazy dog," the letters "e" and "o" are the most common, appearing four times each. From this, a deep learning model could conclude (correctly) that these characters are among the most likely to appear in English-language text.

Realistically, a deep learning model cannot actually conclude anything from a single sentence. But after analyzing trillions of sentences, it could learn enough to predict how to logically finish an incomplete sentence, or even generate its own sentences.

- **LLM neural networks**

In order to enable this type of deep learning, LLMs are built on neural networks. Just as the human brain is constructed of neurons that connect and send signals to each other, an artificial neural network (typically shortened to "neural network") is constructed of network nodes that connect with each other. They are composed of several "layers": an input layer, an output layer, and one or more layers in between. The layers only pass information to each other if their own outputs cross a certain threshold.

- **LLM transformer models**

The specific kind of neural networks used for LLMs are called transformer models. Transformer models are able to learn context — especially important for human language, which is highly context-dependent. Transformer models use a mathematical technique called self-attention to detect subtle ways that elements in a sequence relate to each other. This makes them better at understanding context than other types of machine learning. It enables them to understand, for instance, how the end of a sentence connects to the beginning, and how the sentences in a paragraph relate to each other.

This enables LLMs to interpret human language, even when that language is vague or poorly defined, arranged in combinations they have not encountered before, or contextualized in new ways. On some level they "understand" semantics in that they can associate words and concepts by their meaning, having seen them grouped together in that way millions or billions of times.