An LLM typically has at least one billion or more parameters. *Parameters* are a [machine learning](https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML) term for the variables present in the model on which it was trained that can be used to infer new content. These models are trained on **large datasets**, which contain hundreds of millions to billions of words.

Modern LLMs emerged in 2017 and use [transformer neural networks](https://www.techtarget.com/searchenterpriseai/feature/Transformer-neural-networks-are-shaking-up-AI), commonly referred to as *transformers*. The key component of the transformer architecture is the self-attention mechanism, which enables the model to attend to different parts of the input sequence to compute a representation for each position. This ability to selectively focus on specific parts of the input is crucial for capturing long-range dependencies and understanding the nuances of natural language. The transformer consists of two main components: the encoder network and the decoder network. The encoder network takes an input sequence and produces a sequence of hidden states, while the decoder network takes a target sequence and uses the encoder’s output to generate a sequence of predictions. Both the encoder and decoder are composed of multiple layers of self-attention and feedforward neural networks

* With a large number of parameters and the transformer model, LLMs are able to understand and generate accurate responses rapidly, which makes the AI technology broadly applicable across many different domains.
* While traditional NLP algorithms typically only look at the immediate context of words, LLMs consider large swaths of text in order to better understand the context.

**How do large language models work?**

1. Training on very large datasets
2. Fine Tuning (optional)
3. D.L. using Transformer NN – uses something called as Self Attention Mechanism. The transformer architecture enables the LLM to understand and recognize the relationships and connections between words and concepts using a self-attention mechanism.

TYPES:

1. **Autoregressive Language Models (e.g., GPT)**
2. **Autoencoding Language Models (e.g., BERT)**