* 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 a 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. In order to enable this type of deep learning, LLMs are built on neural networks.

**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.
* The specific kind of neural networks used for LLMs are called transformer models.

**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

**Building own LLMS**

* To build LLM applications, developers need easy access to multiple data sets, and they need places for those data sets to live. Both cloud storage and on-premises storage for these purposes may involve infrastructure investments outside the reach of developers' budgets. Additionally, training data sets are typically stored in multiple places, but moving that data to a central location may result in massive egress fees.
* Data egress fees are charges from cloud providers for moving or transferring data from the cloud storage where it was uploaded. These fees, also known as bandwidth or data transfer fees, are separate from the fees organizations pay for cloud storage and computing.
* Fortunately, Cloudflare offers several services to allow developers to quickly start spinning up LLM applications, and other types of AI.
* Vectorize is a globally distributed vector database for querying data stored in no-egress-fee object storage (R2) or documents stored in Workers Key Value.
* Combined with the development platform Cloudflare Workers AI, developers can use Cloudflare to quickly start experimenting with their own LLMs.