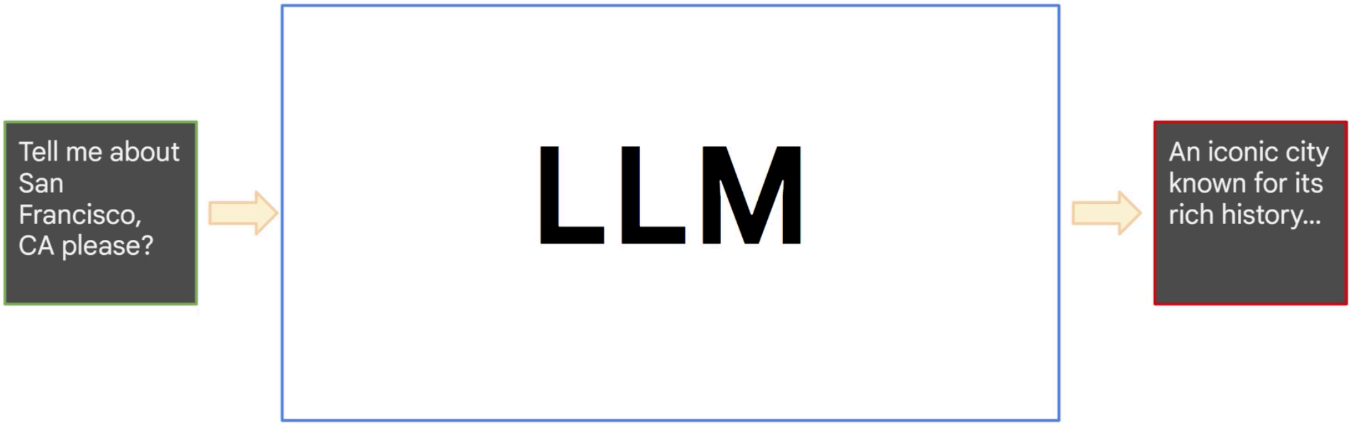
Concept

LLMs (Large Language Model)


Tell me about San Francisco, CA please?
LLM
An iconic city known for its rich history...

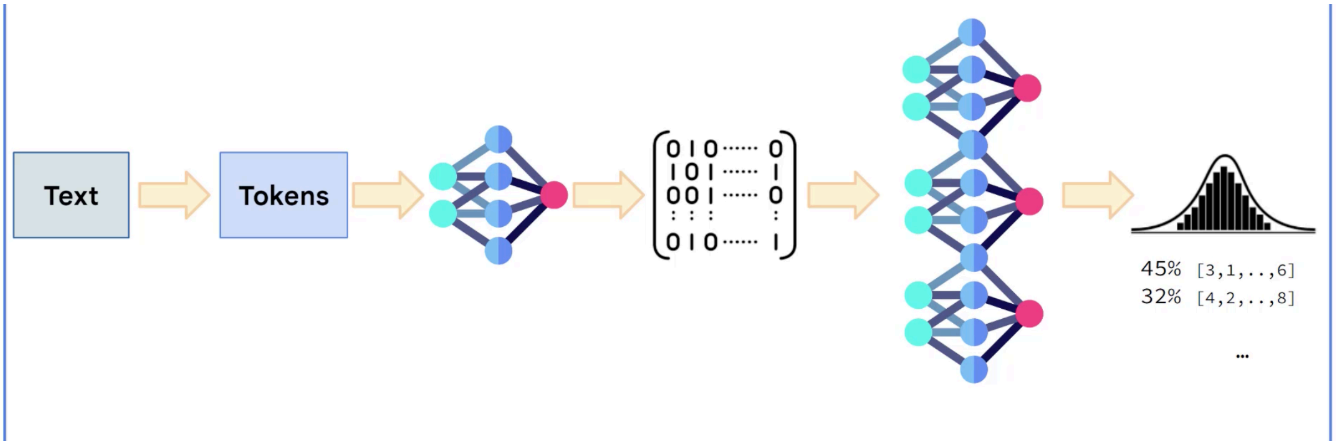
Transformer uses self-attention*. Instead of trying to process every part of an input equally, the attention mechanism allows the model to focus on the most crucial parts of the input.
PROS: less memory needed.

Full transformers consist of an encoder and decoder.

- Encoder: converts input text into a intermediate representation (high-dimensional numerics that capture the semantic nuances of text)
- Decoder: converts that intermediate representation into a useful text.
- Self-attention: weigh the importance of different words in a sequence when processing each word

LLMs is a pre-trained transformer with lots of data capable of understand the probabilistic relationship between words.

It transform:



Words -> Tokens -> Id (number) to each token -> vector representation-> prob. of vectors showing up.