Transformers continued

- Attention: The sequence of vectors from the embedding step are passed through an operation called a attention block

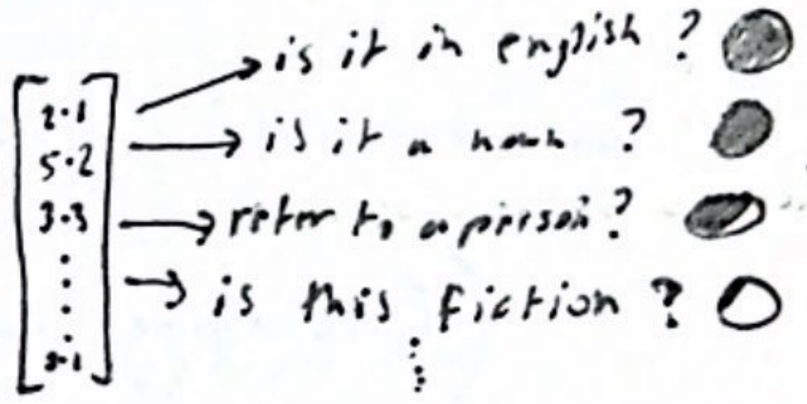
This allows the vectors to talk to each other and pass information back and forth to update there value, giving context to words.

Ex a Machine learning Model vs A fashion Model

The attention block figures out which words in context (concept) are relevant for updating the meanings of which other words "" and how exactly those meanings should be updated (the numbers in vector)

- MLP: a Multi layer preceptron AkA Feed forward layer.

The vectors after the attention operation pass through a MLP operation. and here the vectors dont communicate they all go through the same operation in parralel and is like asking a long list of Questions &
about each vector and updating them based on the answer, this can help the LLM store facts.



- both the attention and MLP are a block of Matrix multiplications
- there are also normalization steps in between and a final unembedding

Then you go through repetitions of attention and MLP on and on
the hope is that all the meaning of the prompt has been baked into the very last vector in the sequence. we then do a certain operation on the last step which produces the probability distribution of all possible next words (tokens). then feed it prompt + prediction and repeat until stop word or limit