Attention in transformers (in LLM)

- allows each word (token) to interact with one another in the input. They pass information with one another allowing the model to understand context about the words and the positions they appear in.

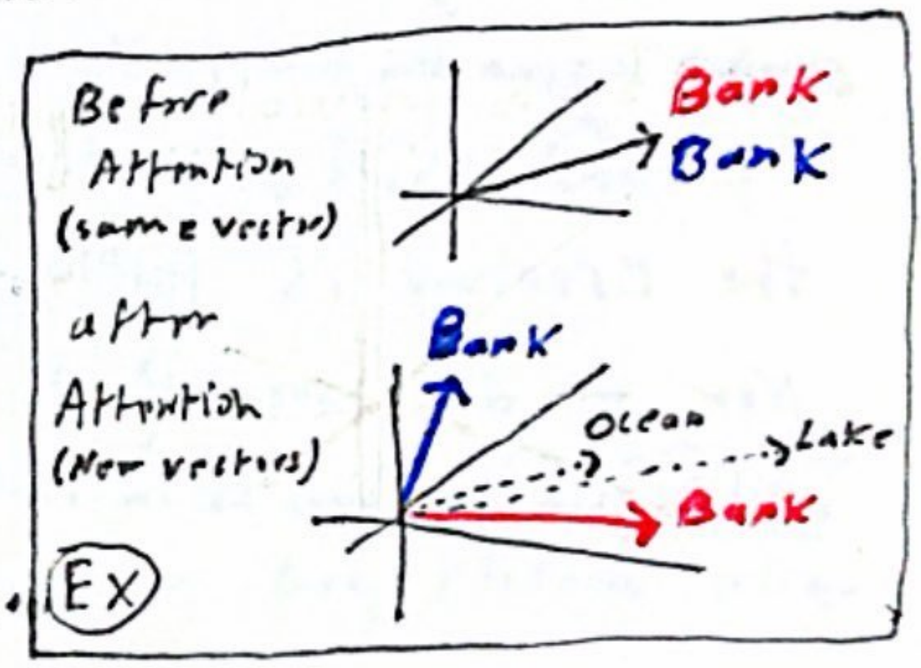
Ex attention is responsible for figuring out that

by the River Bank

is Different from

Deposit at the Bank

since bank has the same position and is same word it will have the same embedding but both words have different meaning and the only way the word "Bank" can find its meaning is by the words passing information back and forth this allows the Attention model to update the vectors (words) to closely align with there true meaning allowing the model to understand the true meaning of the prompt and not just the individual words in isolation.



Ex: see how after attention Bank aligns with "bodies of water"

Example:

- input: a fluffy blue creature roamed the verdant forest

• Now imagine the only 2 late we care about is having the adjectives adjust the meaning of their corresponding noun (Too add context to words)

This is called A Single Head of Attention. in reality an Attention block consists of many different heads. called Multi Headed Attention. and all these heads are ran in parralle/ and update words based on many different words so that all words have a chance to update all other relevant words. not just River attending to Bank or Deposit 54 attending to bank (single attention)