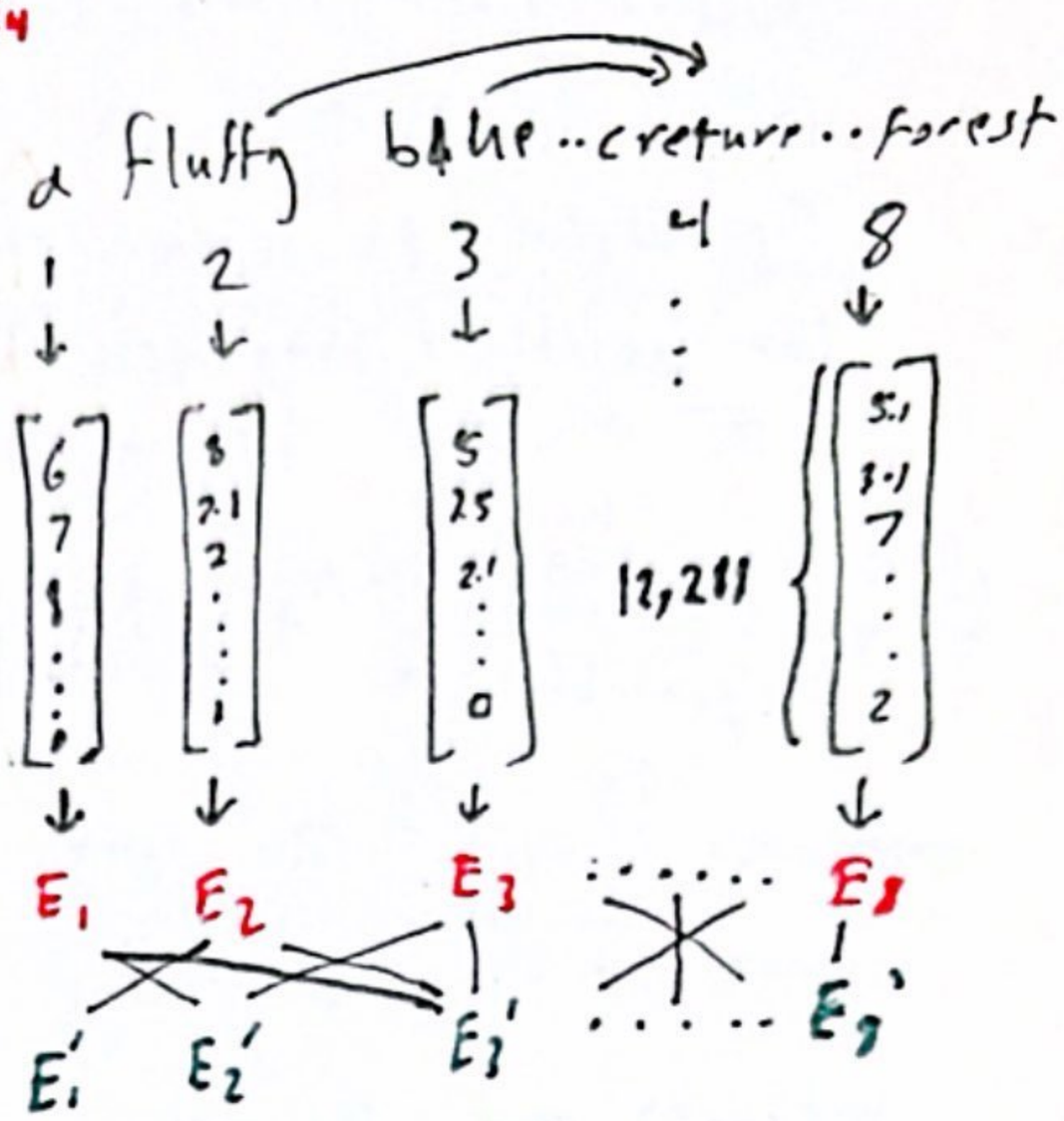
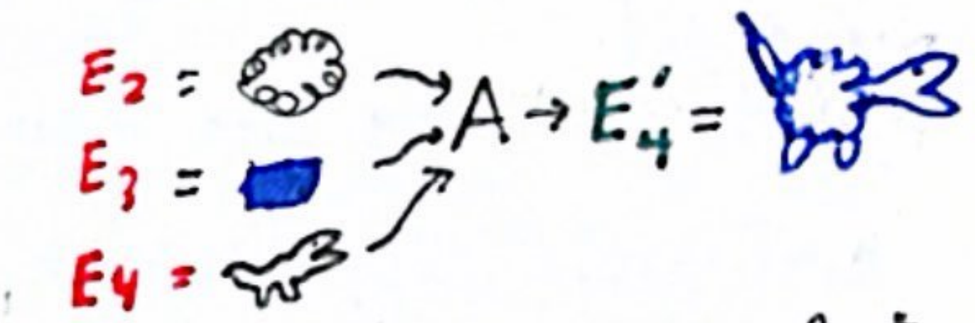
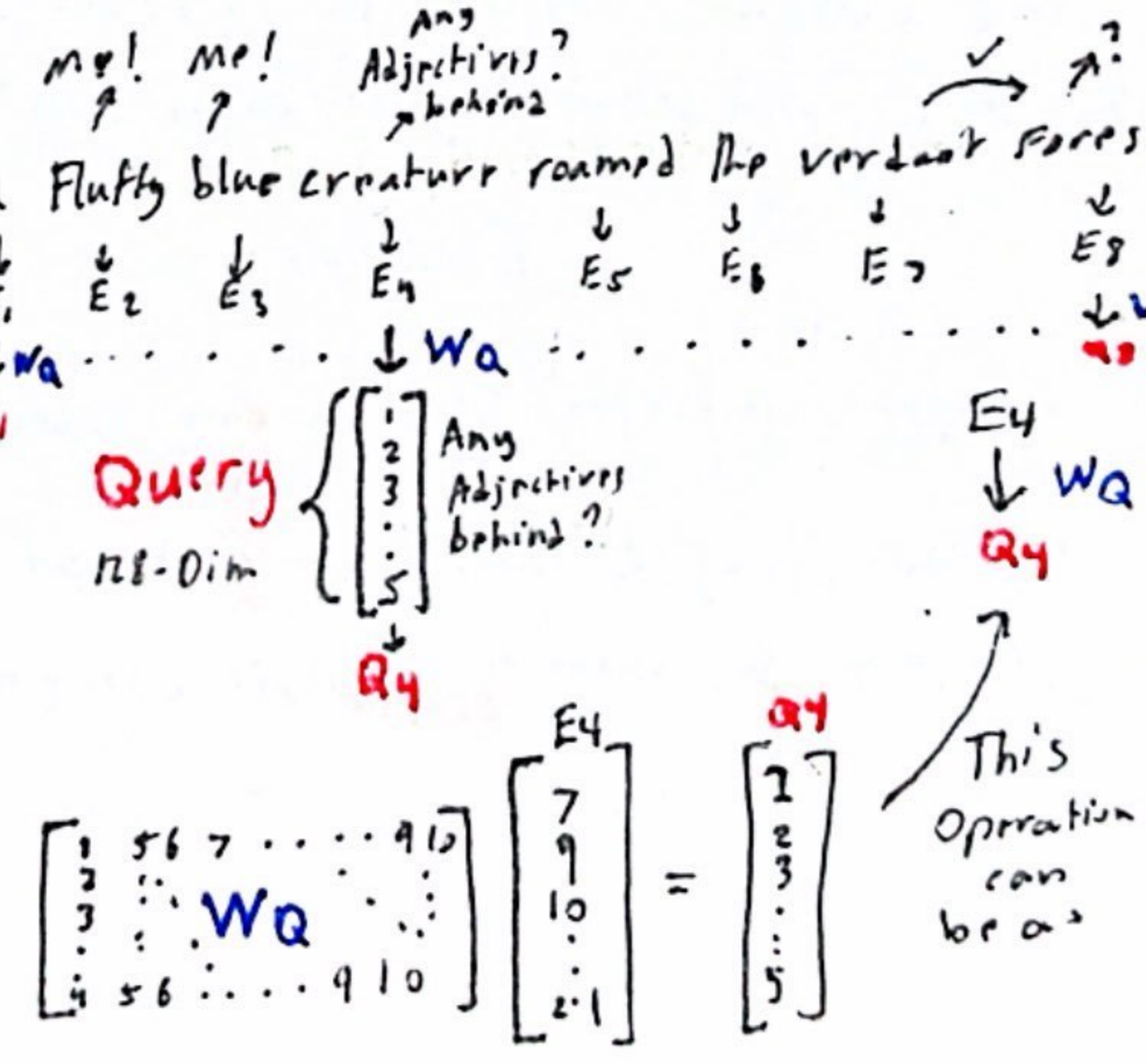
Attention (P-2) (LLM notes)

- Ex from last Page: a fluffy blue creature roamed the verdant forest



- The initial embedding for each word is a high dimentional vector that only encodes the meaning and position of that word with no context. We Denote the Embeddings as $\vec{E}_n$. The goal is to have a series of computations that produce a new refined set of embeddings where for Ex those corresponding to the nouns have ingested the meaning from there corresponding adjectives. for ex here creature is noun and blue, fluffy are the adjectives the noun "blue" and after Attention we want the LLM to know that the creature is fluffy and blue. This will add relevant context as Now we dont have just a creature we h.. a fluffy blue creature. This is just one head Ex in reality true Multi head Attention is learned throug training





- for this ex imagine each noun asking if there is a adjective behind of me and fluffy / blue to be able to say yes im an adjective and im in that position. that Question is encoded as another vector called the Query for this word. its a smaller dimension than the embedding. 128 for ex. Computing (answering) this Query Looks Like taking a Query. matrix W_Q and multiplying it by the embedding. here you multiply all embedding by W_Q and get $Q_1 \rightarrow Q_8$. the W_Q is 55 learned in traing and true behaviour is