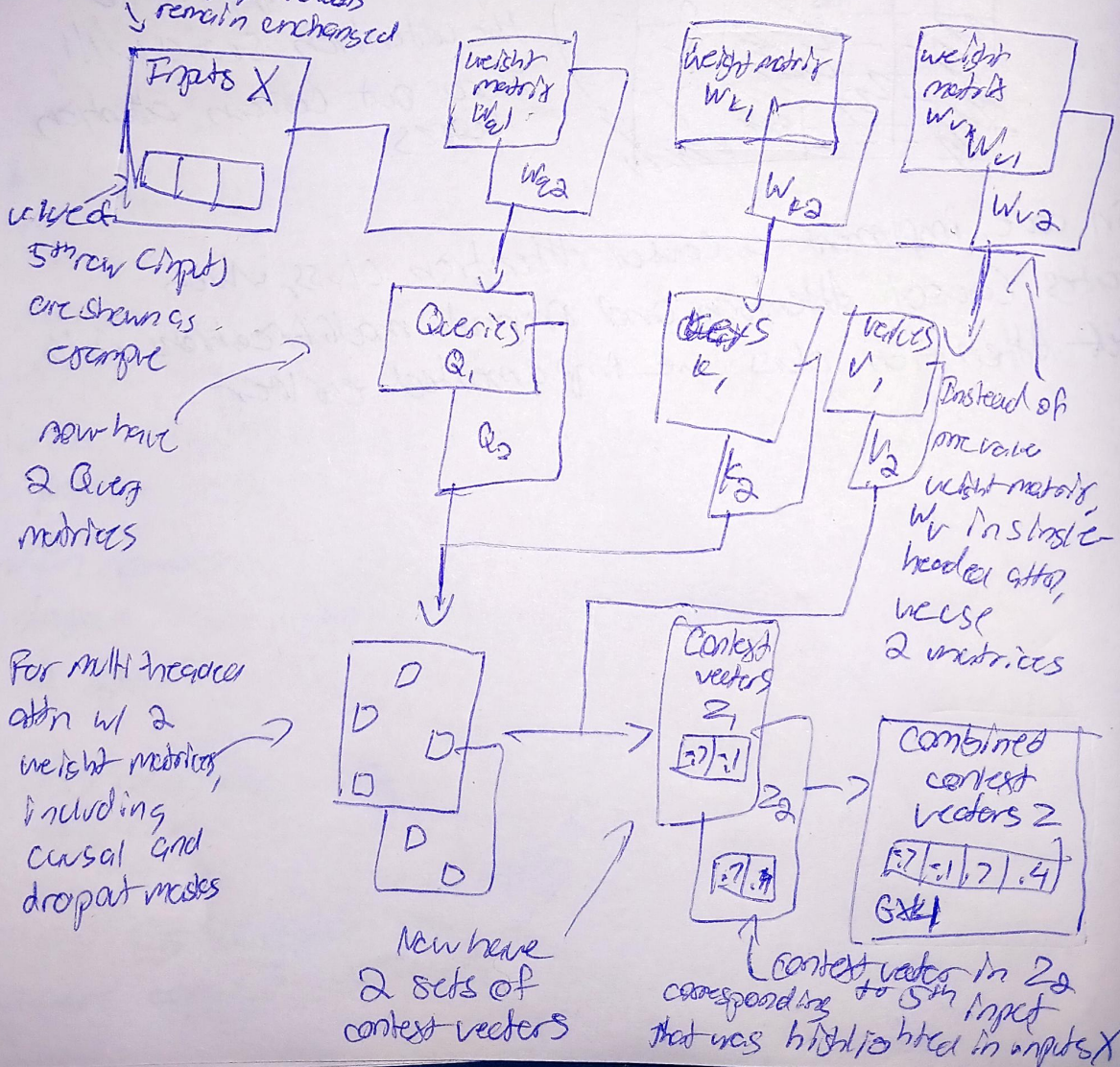
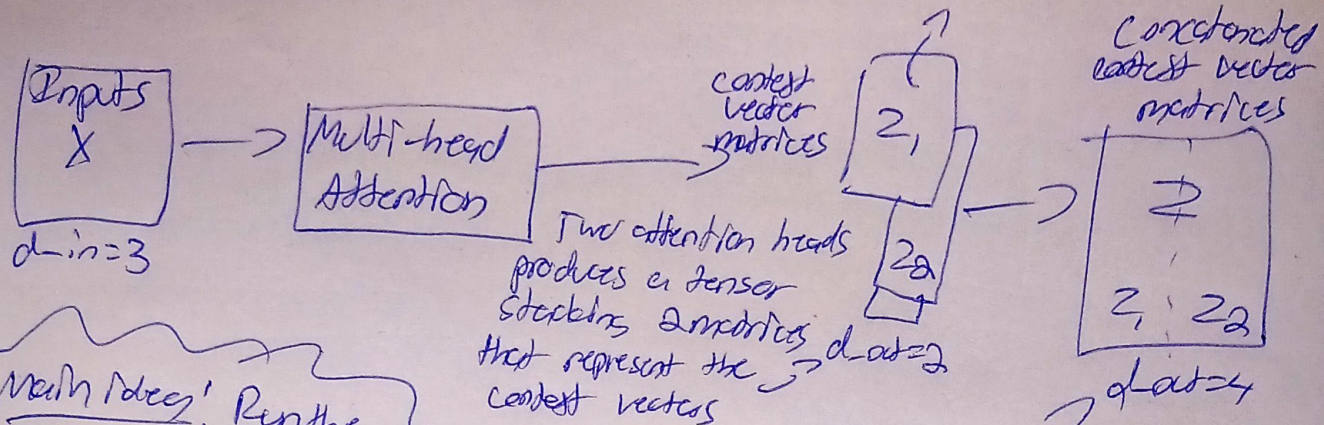
Multi-head Attention Mechanism

- Dividing attention mechanism into multiple heads
  - Each operating independently

Can stacking multiple single head attention layers

1. Implementing multi-head attention involves creating multiple instances of the self-attention mechanism, each w/ it's own weights, and then combining their outputs
2. This can be computationally intensive, but it makes LLMs powerful at complex pattern recognition tasks





Multi-head: Run the attention mechanism multiple times (in parallel) w/ different, learned linear projections; the results of multiplying the input data (like query, key, value vectors) by a weight matrix

Choosing an embedding dimension of 2 (d_out=2) for context vectors results in a final embedding dimension of 4 (d_out x num_heads)