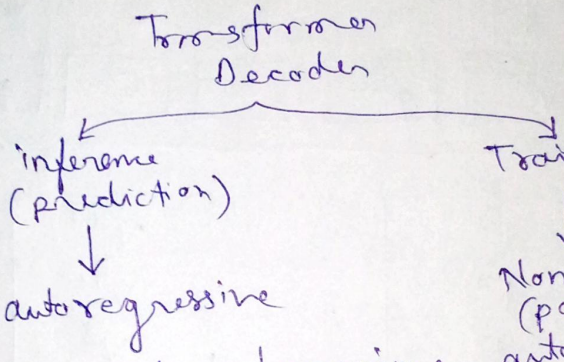
16 Mar 2025

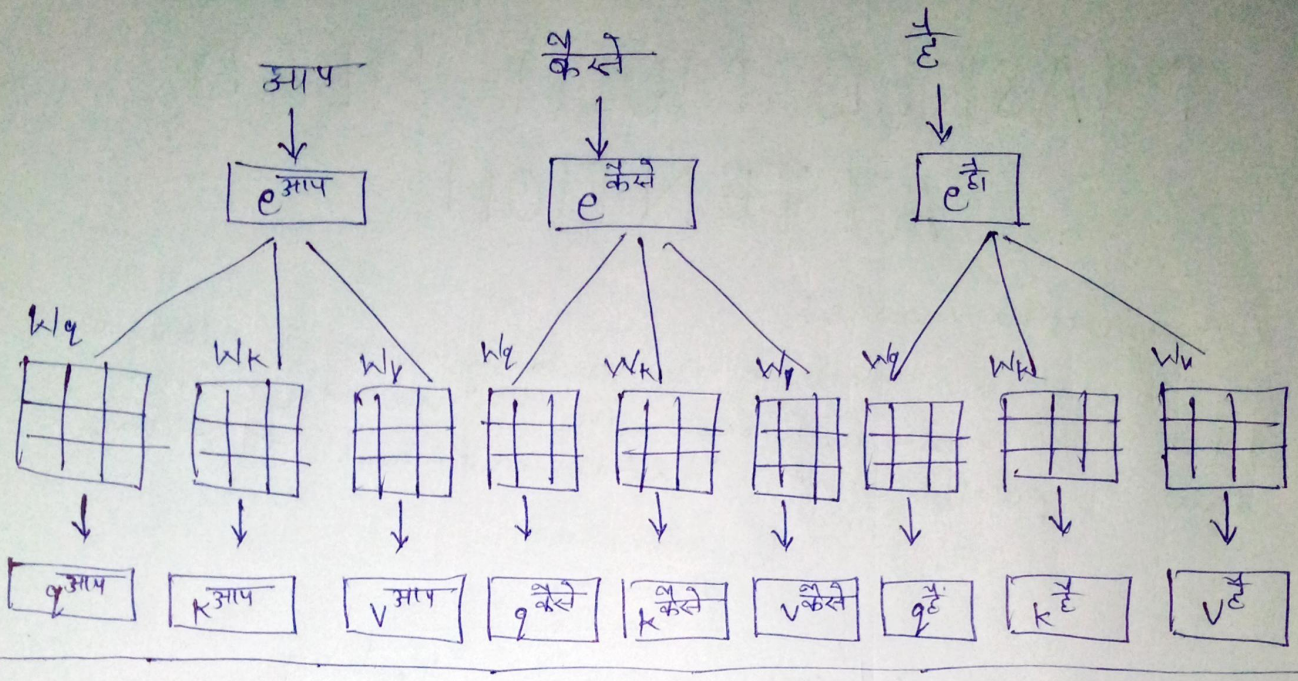
MASKED MULTI-HEAD ATTENTION

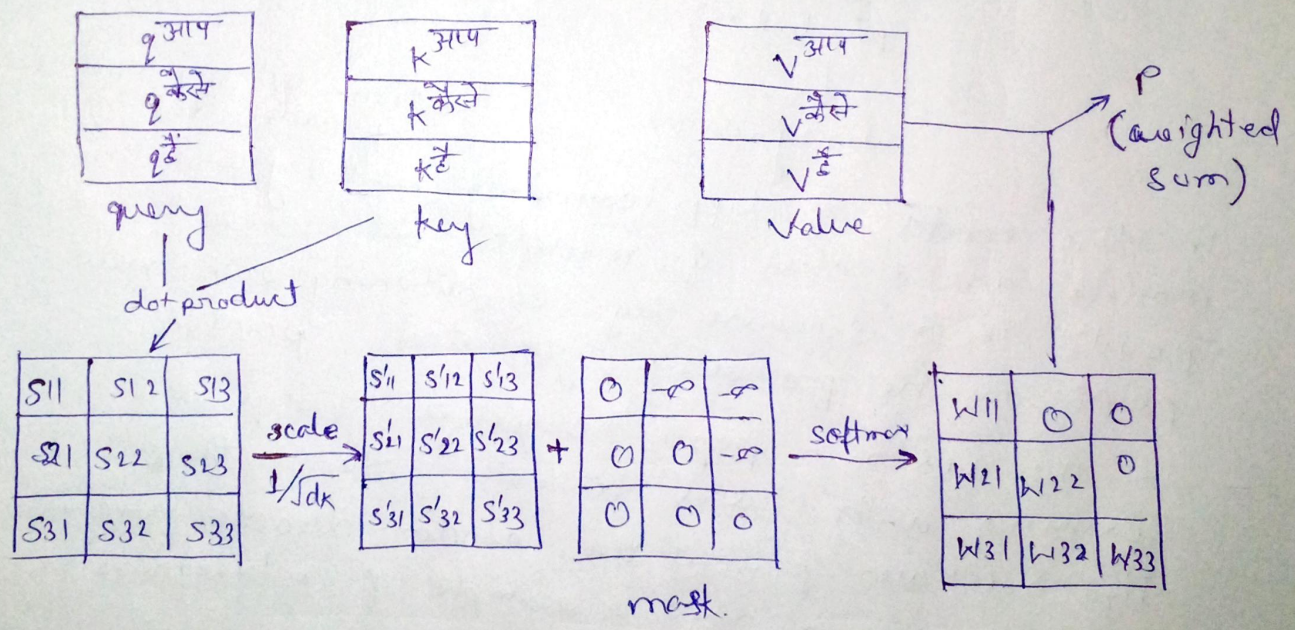
Autoregressive models:
The transformer decoder is autoregressive at inference time and non-autoregressive at training time. this is possible with masked multi-head attention.



- in the context of deep learning models are a class of models that generate data points in a sequence by conditioning each new point on the previously generated points.
- in the research paper "Sequence to Sequence Learning with Neural Networks" the author Ilya Sutskever present the ecoder-decoder ~~architecture~~ architecture is a also a example of autoregressive. model. (All seq2seq model).

masked multi-head Attention Architecture:





$$P = \begin{cases} \text{आप}_{ce} = W11 * v^{आप} + W12 * v^{वे\ कैसे} + W13 * v^{हैं} \\ \qquad = W11 * v^{आप} \\ \text{वे कैसे}_{ce} = W21 * v^{आप} + W22 * v^{वे\ कैसे} + W23 * v^{हैं} \\ \qquad = W21 * v^{आप} + W22 * v^{वे\ कैसे} \\ \text{हैं}_{ce} = W31 * v^{आप} + W32 * v^{वे\ कैसे} + W33 * v^{हैं} \end{cases}$$