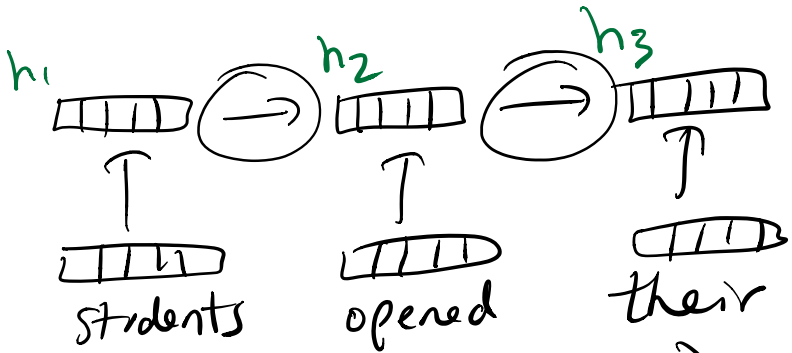
Parallelization of self-attention

at training time

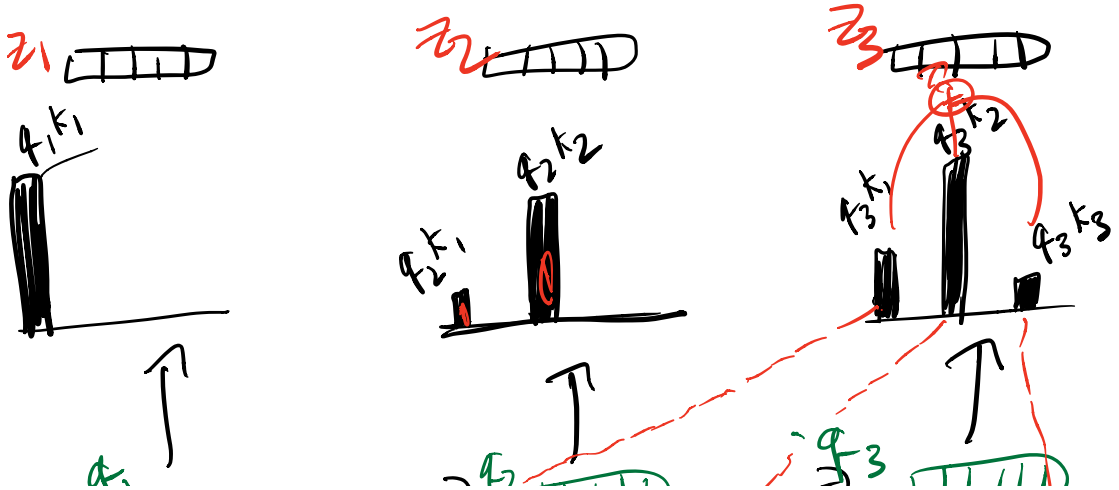
RNNs:

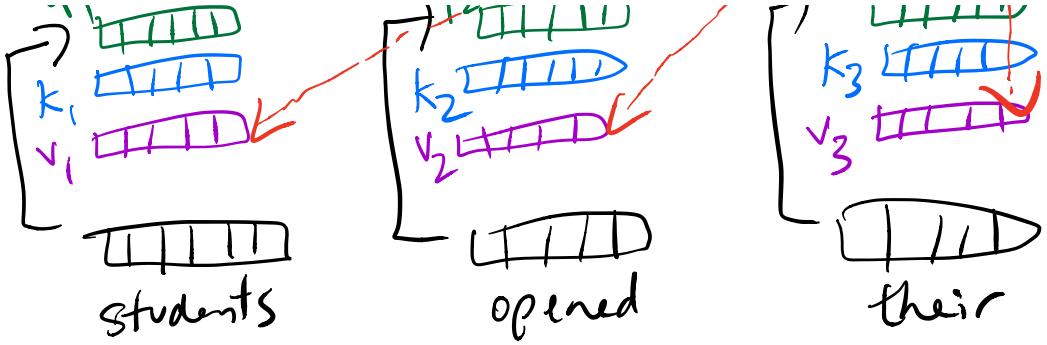


$$h_n = f(W_h h_{n-1} + W_e c_n)$$

↳ each hidden state is a direct fn of the previous hidden state

Sequential computation in RNNs allows us to model word order





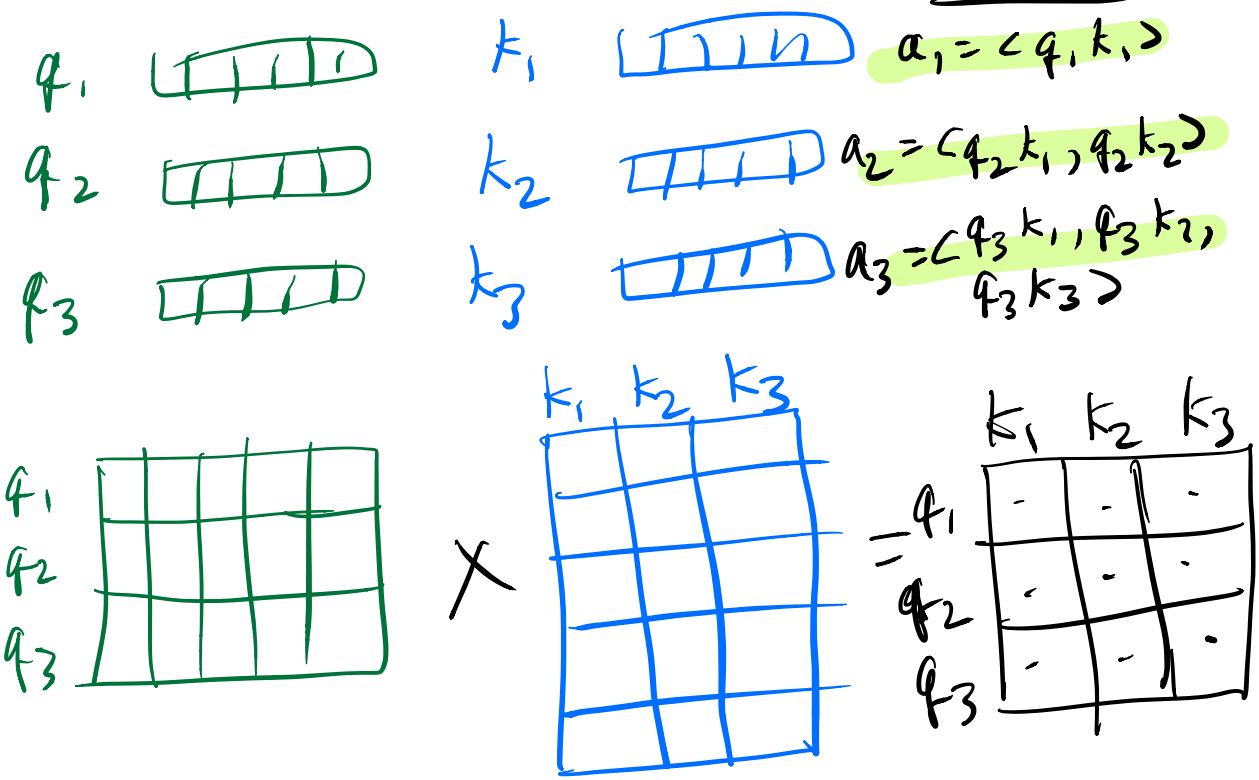
there is no dependency between z_n and z_{n-1}!

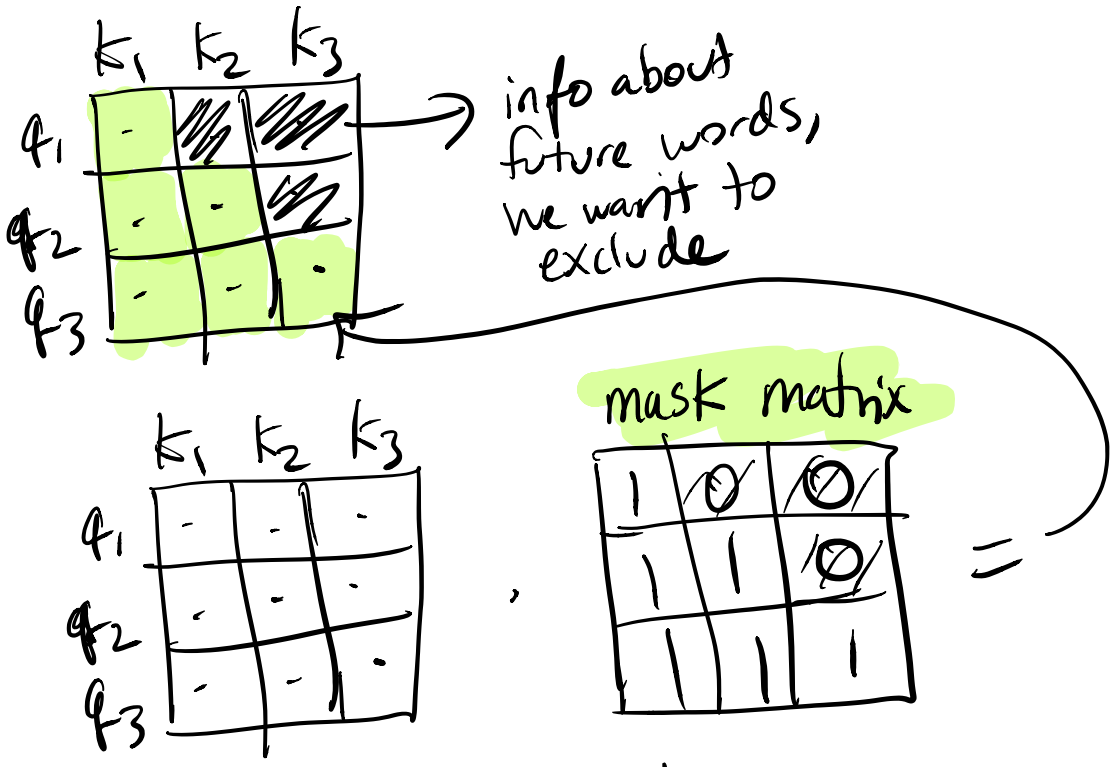
we no longer have to compute the z_s one at a time.

---

how do we parallelize the attn computation?

attn scores





after masking, we apply softmax and then we get the valid attn distributions without any "cheating"