Transformers:

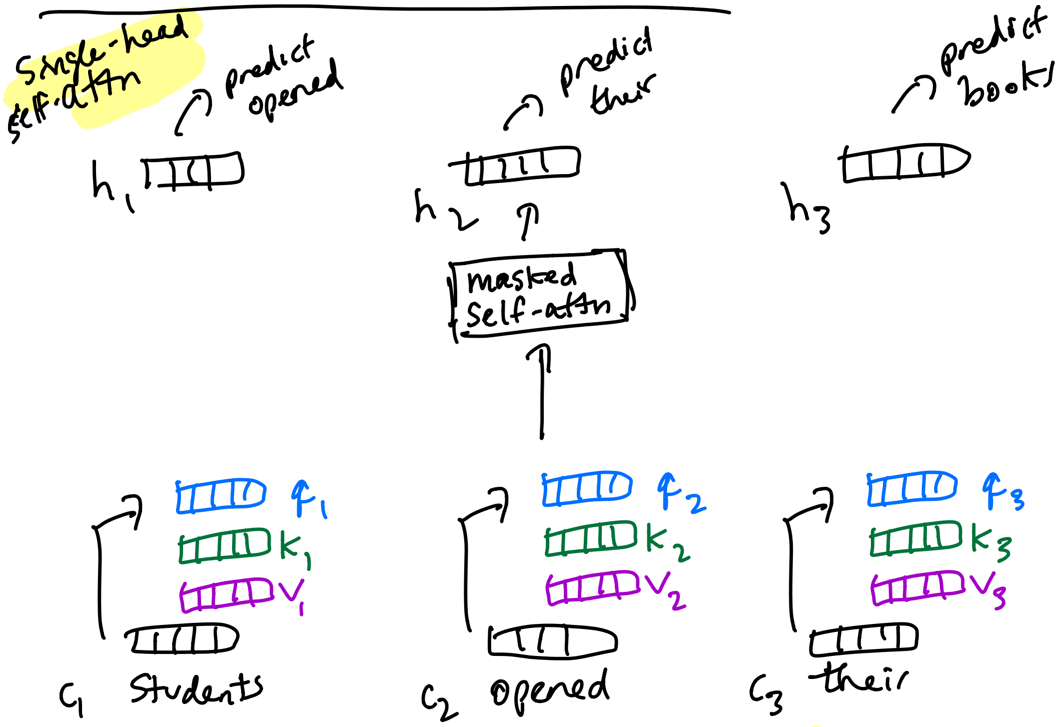
↳ a neural LM whose component is multi-head self-attention

multi-head self-attn:

↳ instead of just one set of query/key/value vectors, let's have many sets (heads)

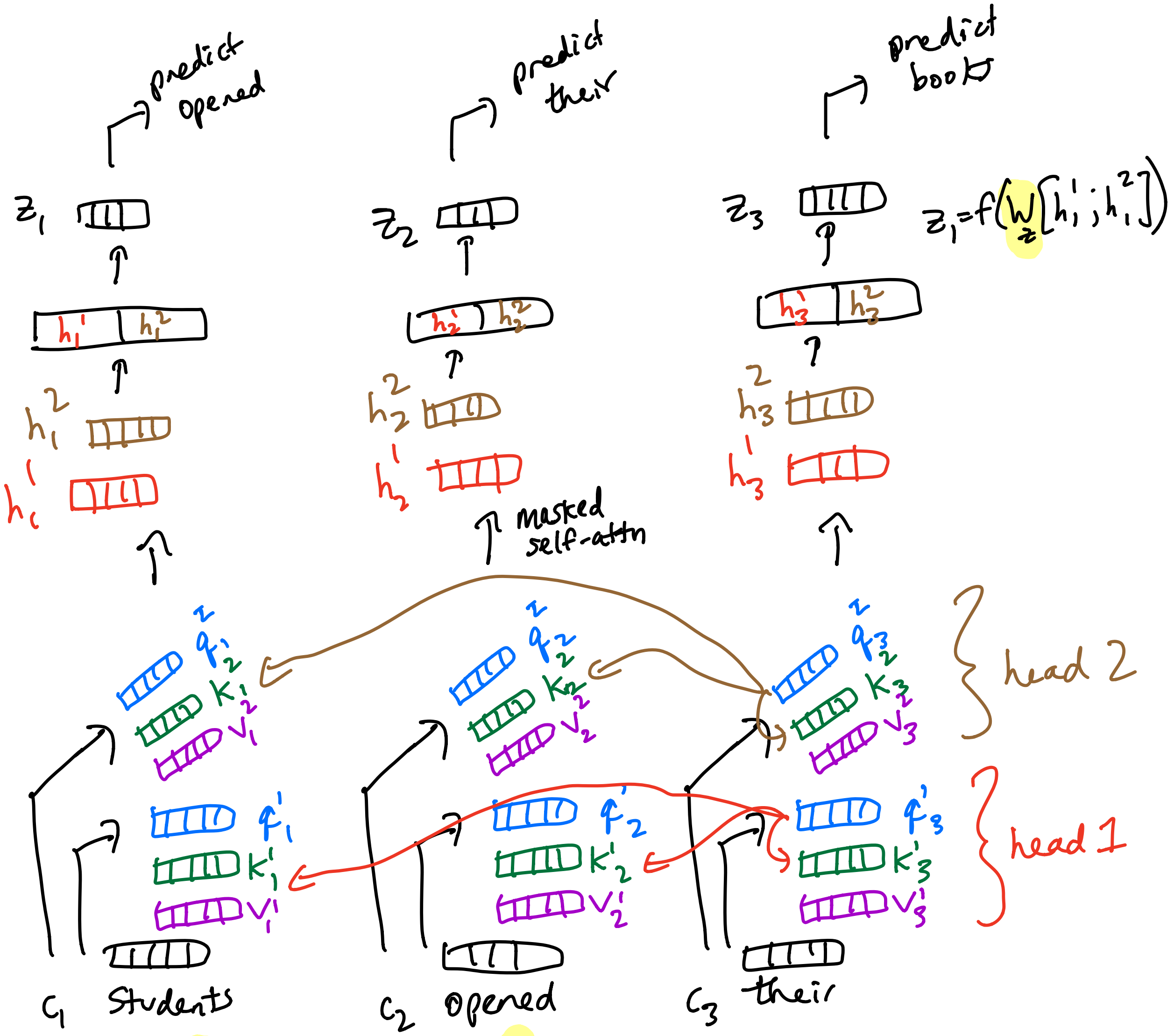
↳ intuition: having multiple sets of Q, K, V can allow each head to attend to different linguistic properties of the prefix

↳ n-gram windows, subject/obj of sentence, discourse (global) context, entities, verbs, . . .



$q_1 = f(W_q c_1)$ $k_1 = f(W_k c_1)$ $v_1 = f(W_v c_1)$

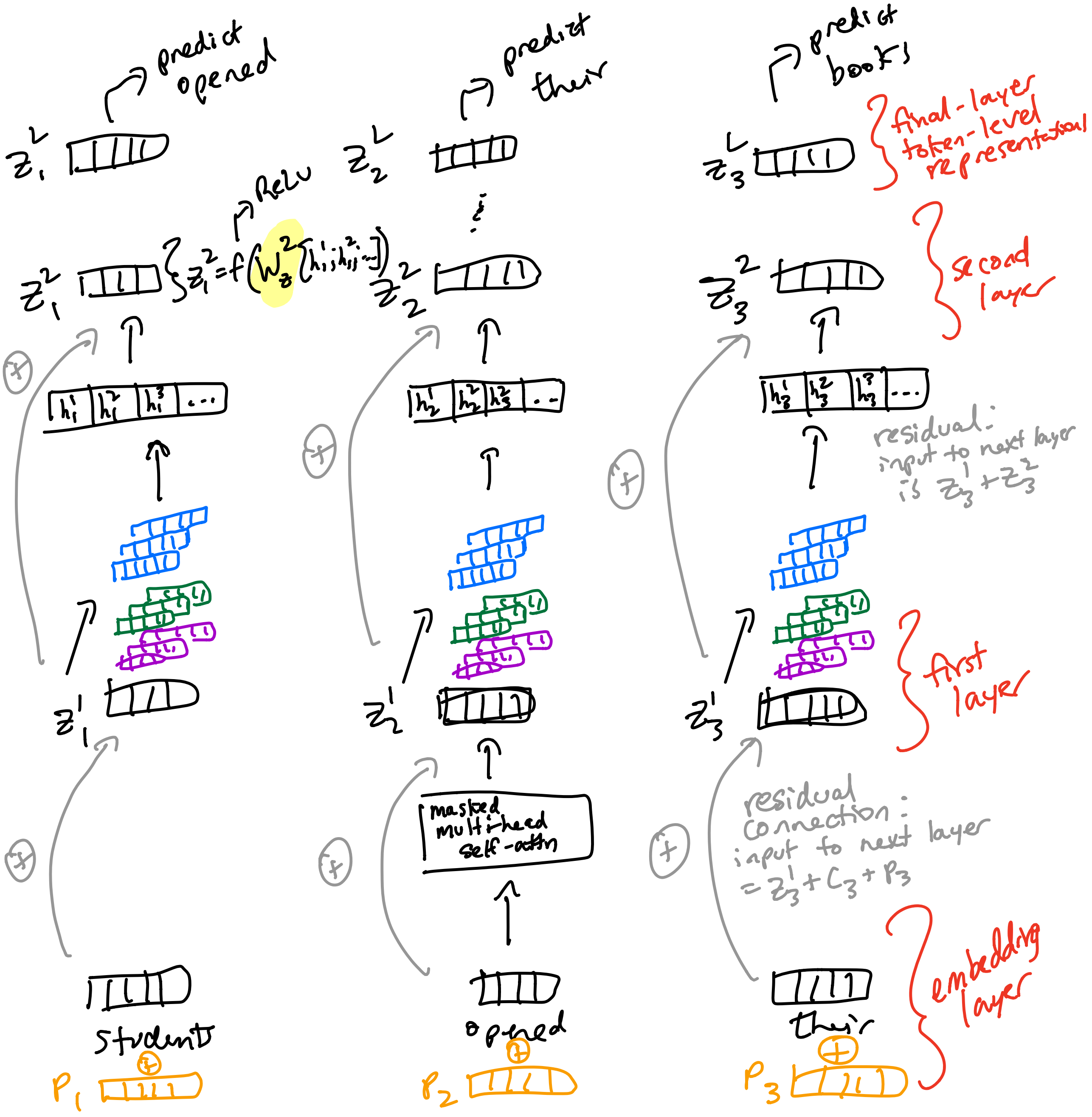
multi-head self-attn



$q_1^1 = f(W_q^1 c_1)$ $q_1^2 = f(W_q^2 c_1)$

$q_2^1 = f(W_q^1 c_2)$

Adding depth

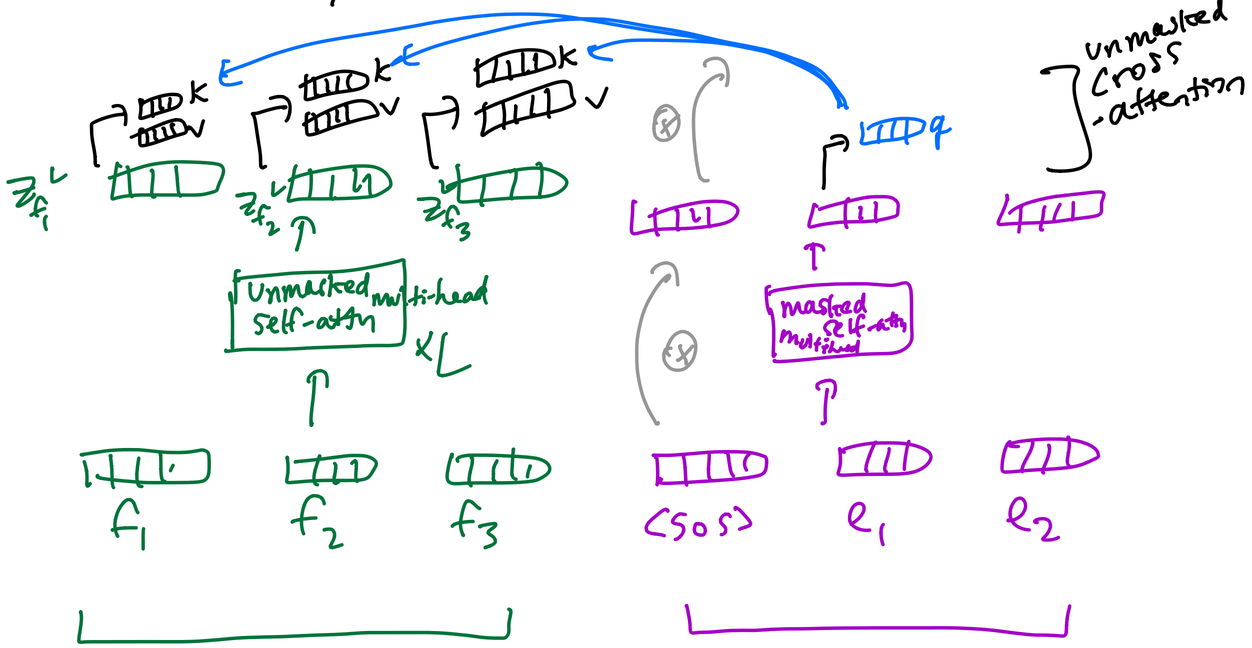


predict opened
predict their
predict books
z_1^L
z_2^L
z_3^L
final-layer token-level representational
z_1^2
$z_1^2 = f(W_z^2 [h_1^1; h_1^2; \dots])$
ReLU
z_2^2
z_3^2
second layer
h_1^1 h_1^2 h_1^3 ...
h_2^1 h_2^2 h_2^3 ...
h_3^1 h_3^2 h_3^3 ...
residual: input to next layer is $z_3^1 + z_3^2$
z_1^1
z_2^1
z_3^1
first layer
masked multihead self-attn
residual connection: input to next layer $= z_3^1 + c_3 + p_3$
embedding layer
students
opened
their
p_1
p_2
p_3

seq2seq

what if we want to give the model some input and have it generate a completion

↳ let's say we're translating from French to English



encoder
(no need to predict the next word)

decoder:
responsible for generating text

$$p(e_n \mid e_{1,\dots,n-1}, f)$$

conditional LM