Token Alignment

- token alignment is the process of matching tokens from one sequence to their corresponding tokens in another sequence, often used in translation, attention, embedding comparisons to know which parts of the inputs relate to which parts of the output

EX: english: i am a student → fr am — suis
french: je suis a etudiant

This mapping is token alignment: i → je, am → suis, a → un, student → etudiant
no direct match

- to eval use Precision, Recall, F1 to see accuracy of token alignment to true vals

Multi query attention

Multi query attention is a attention mechanism where all heads share the same key and value vectors but have seperate querie vectors unlike multi-head mention where each head has its own key query and value vectors. Compared to multi-head attention its more memory efficient and computationally efficient since key and values are shared, while still allowing multiple query prespectives

Pooling (CNN)

- pooling is a dimentionality reduction method in ML used in CNN (convolutional neural networks), its known as a downsampling operation used to reduce the spacial size (height and width of the 2D dimentions of the feature map) of feature maps (output of convolution layer showing where certion features like edges or textures appere in the input), this helps decrese computation control overfitting and make features more invariant to small shifts

Ex: $\begin{bmatrix} 1 & 3 & 2 & 4 \\ 5 & 6 & 7 & 8 \\ 3 & 2 & 1 & 0 \\ 4 & 5 & 2 & 3 \end{bmatrix}$ ⇒ 2x2 max pooling with stride of 2 on this 4x4 matrix ⇒ we devide it into blocks of 2x2 and take the max value from each block stride=2 so skip every other pixel ⇒ $\begin{bmatrix} 6 & 8 \\ 5 & 3 \end{bmatrix}$

Stride is the num of pixels the Filter moves or slides across at each step during convolution or pooling

Ex Result: the output is smaller (2x2) but keeps the strongest features (max values)

- Note: there are other pooling types too like avg pooling, which takes the avg instead of the max (which our max pooling ex did)