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 input relate to which parts of the output

Ex: english: i am a student — this mapping is token alignment: i → je, am → suis, a → un, student → etudiant
french: je suis un etudiant — no direct word 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