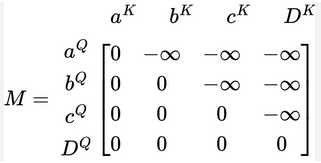
**Masked Self Attention**

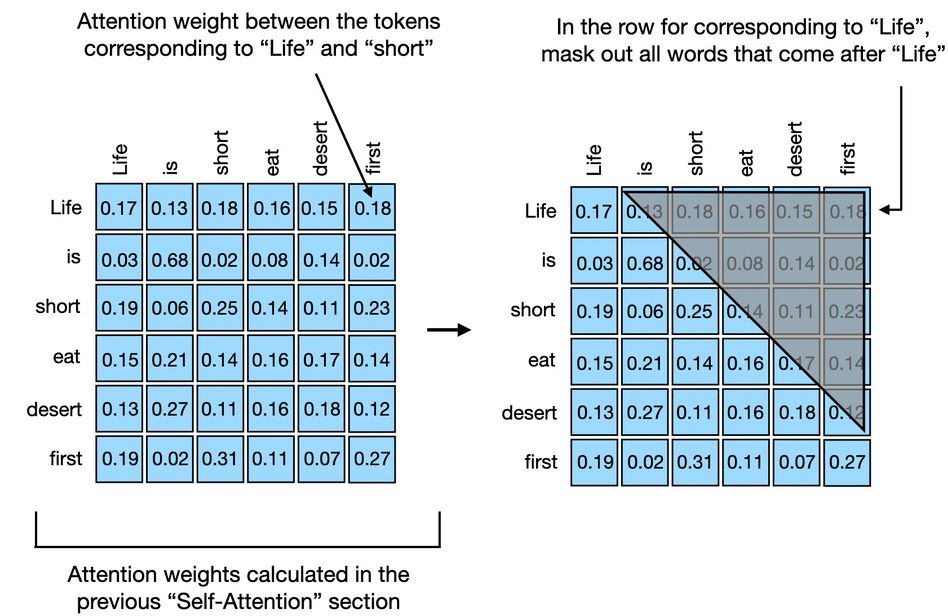
======================================================================

Masked Self-Attention is a variant of self-attention used in Transformers, specifically in decoder blocks (like in GPT models) where the model generates text one token at a time.

**Why is it needed?**

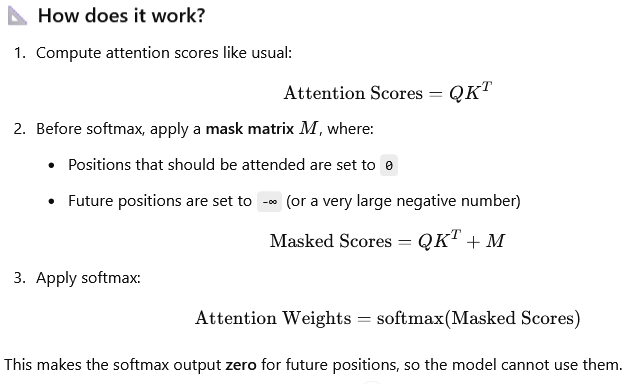
In autoregressive tasks (e.g., text generation), the model should not "peek" at future tokens when predicting the current token.  
To enforce this, masked self-attention is used to block access to future positions in the sequence during training.





**What does it do?**

It applies a mask to the attention scores to prevent the model from attending to tokens ahead of the current one.



--------------------------------------------------------------------------------------------------------------

**Example (Sequence length = 4):**

For token at position 2, we want it to only see tokens [1, 2].  
The mask might look like this (upper triangular):

