**Key Value Caching in Transformers**

Key and Value are the states of transformers used to calculate the attention that is helpful in understanding the context of the input sequence.

Key Value caching occurs in decoder only models or encoder – decoder model, where there are token generation steps. In decoder, we calculate attention of tokens, which depends on the previous tokens and at each step we repeat this step, but in reality, we only need to calculate attention for the new token. Hence by caching previous Key and Value states, we can focus on calculating attention of the new token.

This way, we can reduce our matrix multiplication operation, hence improving computational efficiency.