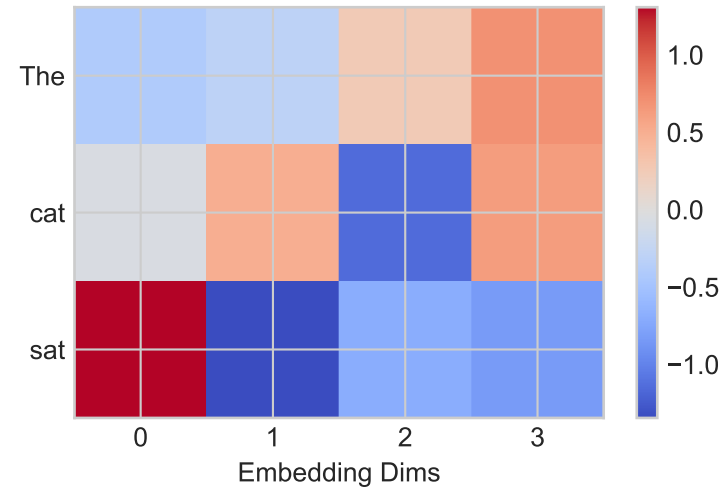
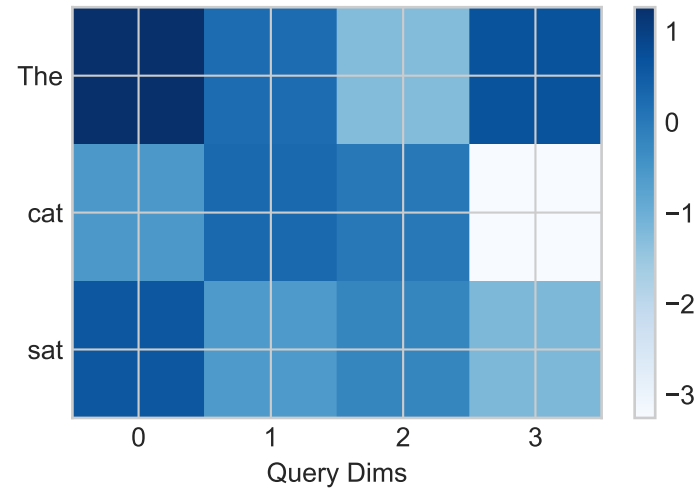


Self-Attention Mechanism: Step-by-Step Computation

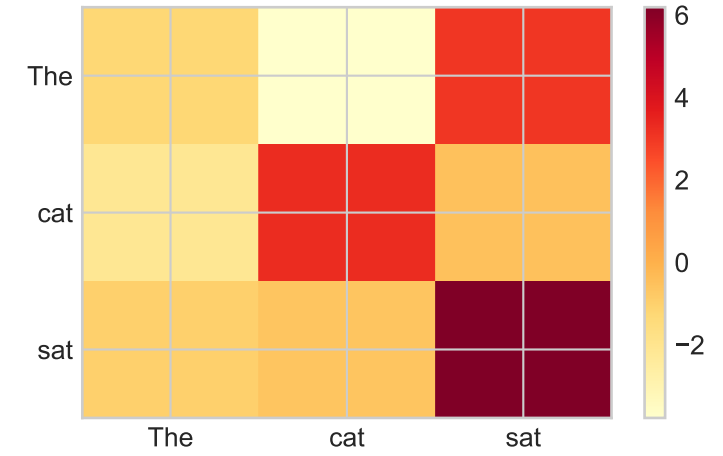
Step 1: Input Embeddings



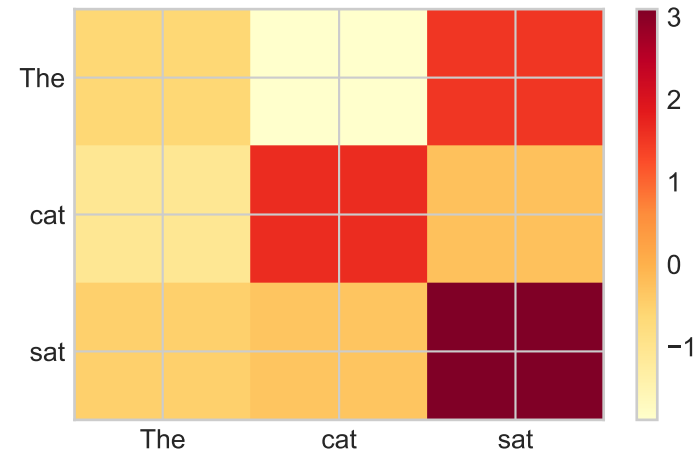
Step 2: Query Matrix (Q)



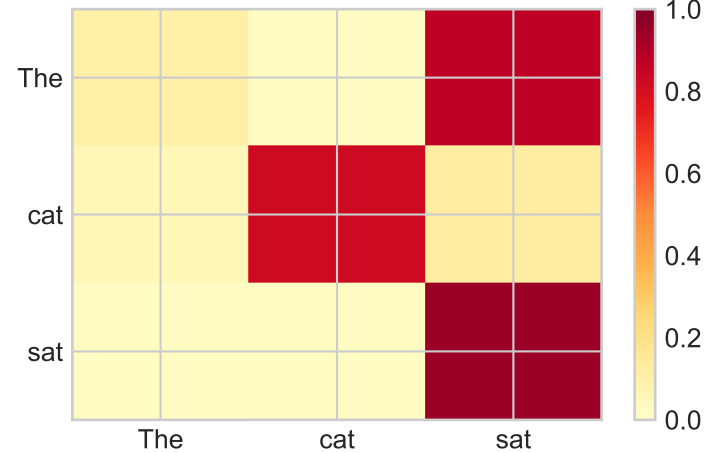
Step 3: QK^T (Raw Scores)



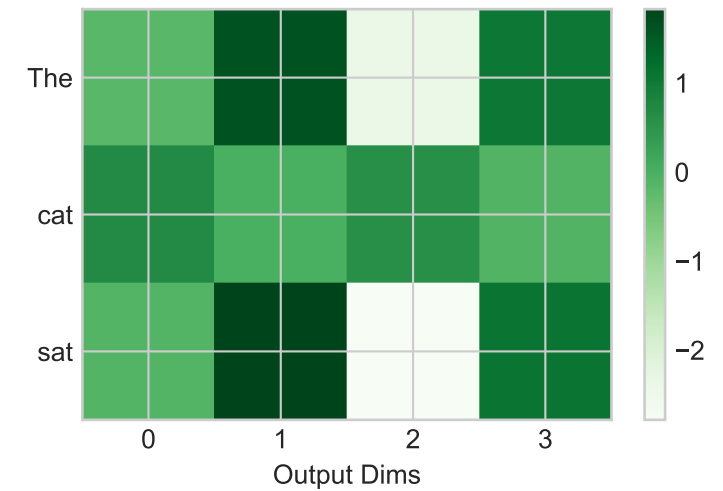
Step 4: Scale by $\sqrt{4}$



Step 5: Softmax (Probabilities)



Step 6: Attention \times Values



$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where each word decides what to attend to based on Query-Key similarity