RNNs & The Attention Mechanism

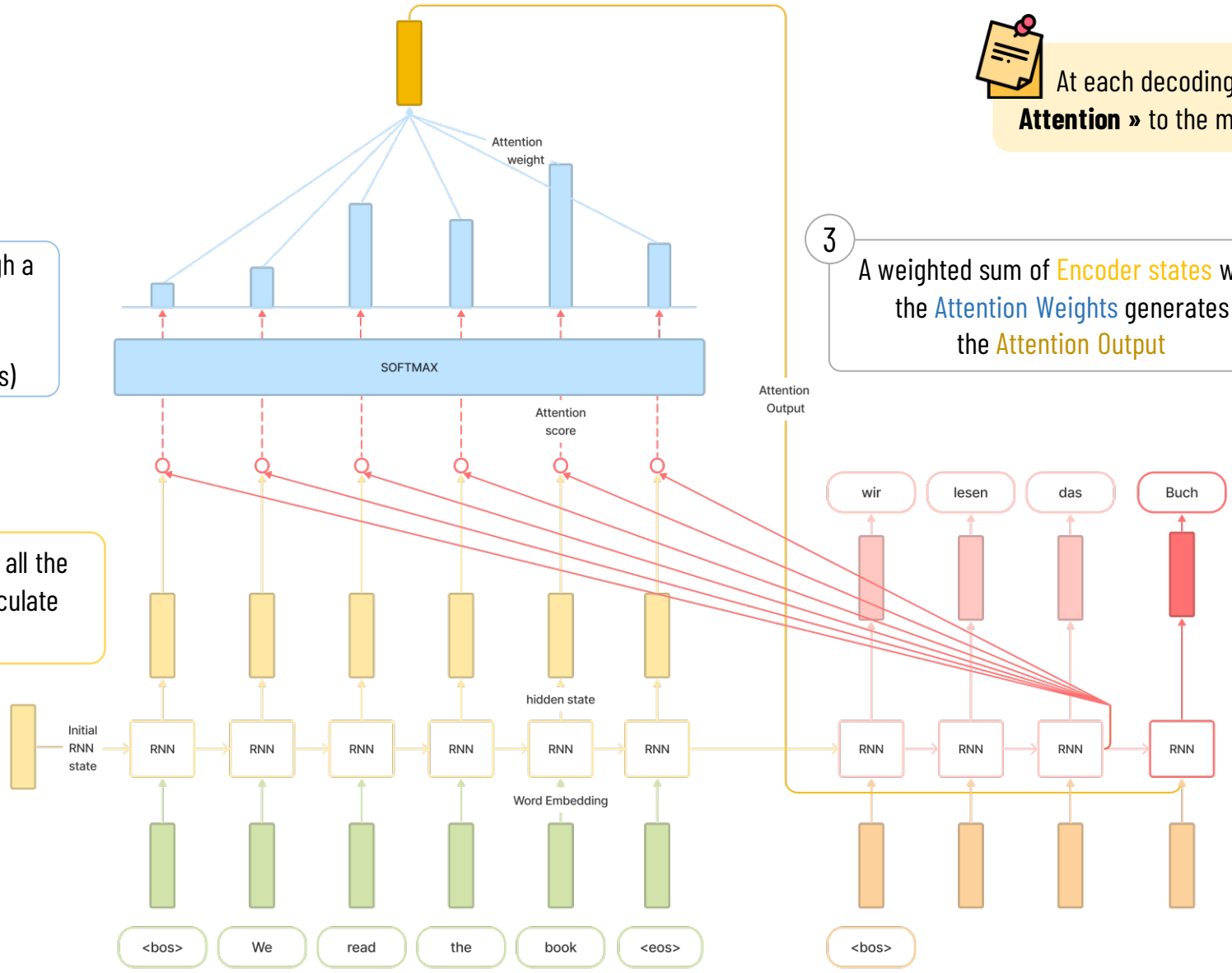
At each decoding step, the model « **pays Attention »** to the most relevant input token

2 The Attention Scores are passed through a softmax function to compute the Attention Weights (The distribution over the input tokens)

3 A weighted sum of Encoder states with the Attention Weights generates the Attention Output

1 The Decoder's Hidden state and all the Encoder states are used to calculate attention scores

4 The Attention output is passed back to the Decoder, to generate the next token



Mouad ATERHI
Abderrazzak BAJJOU