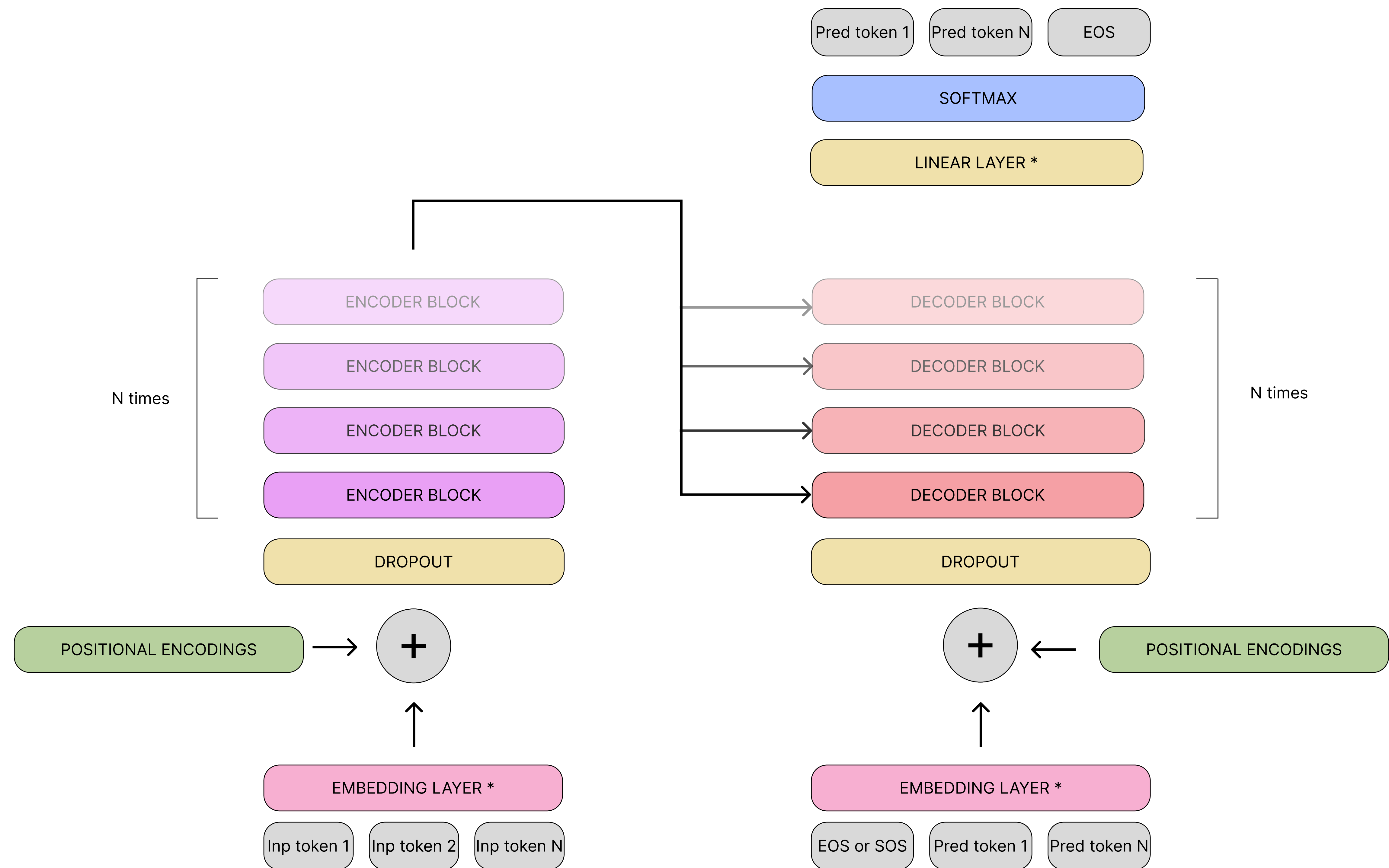
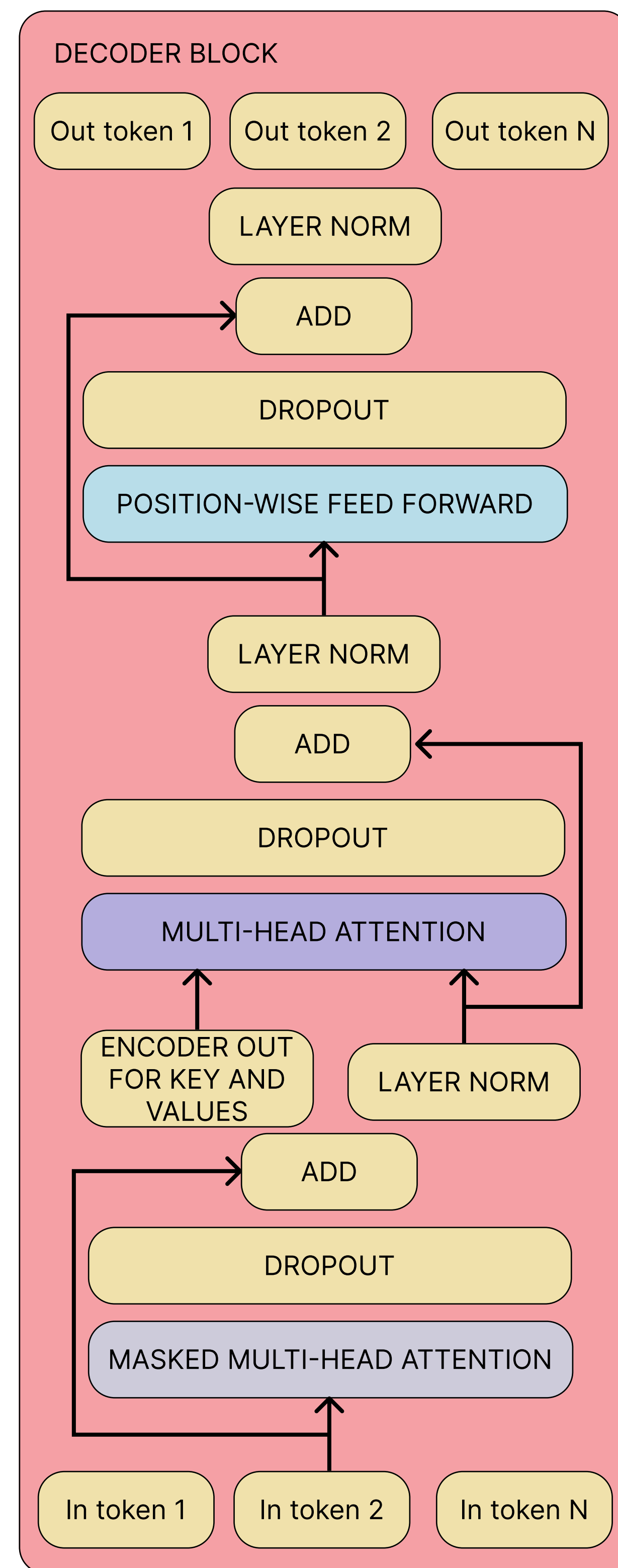
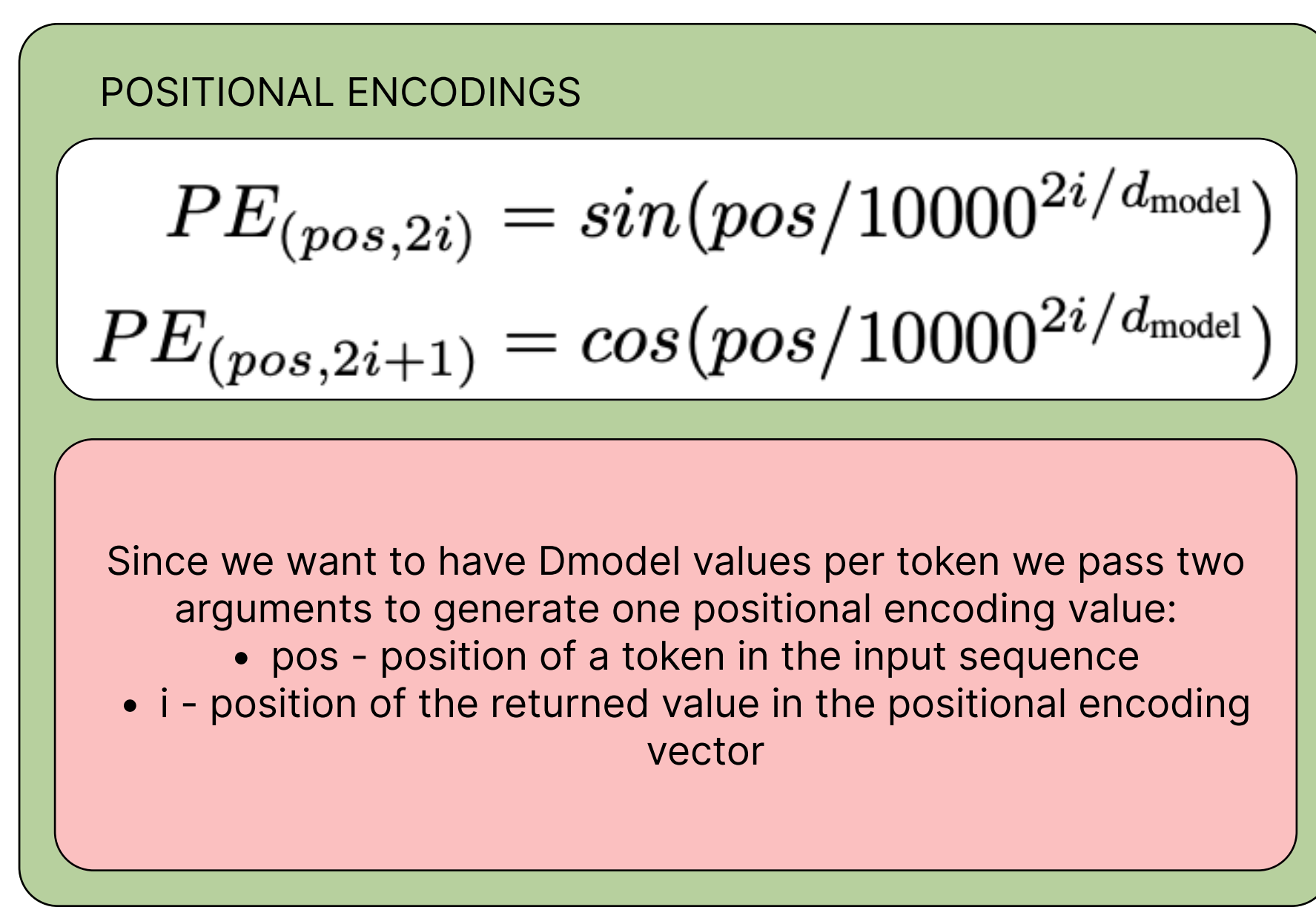
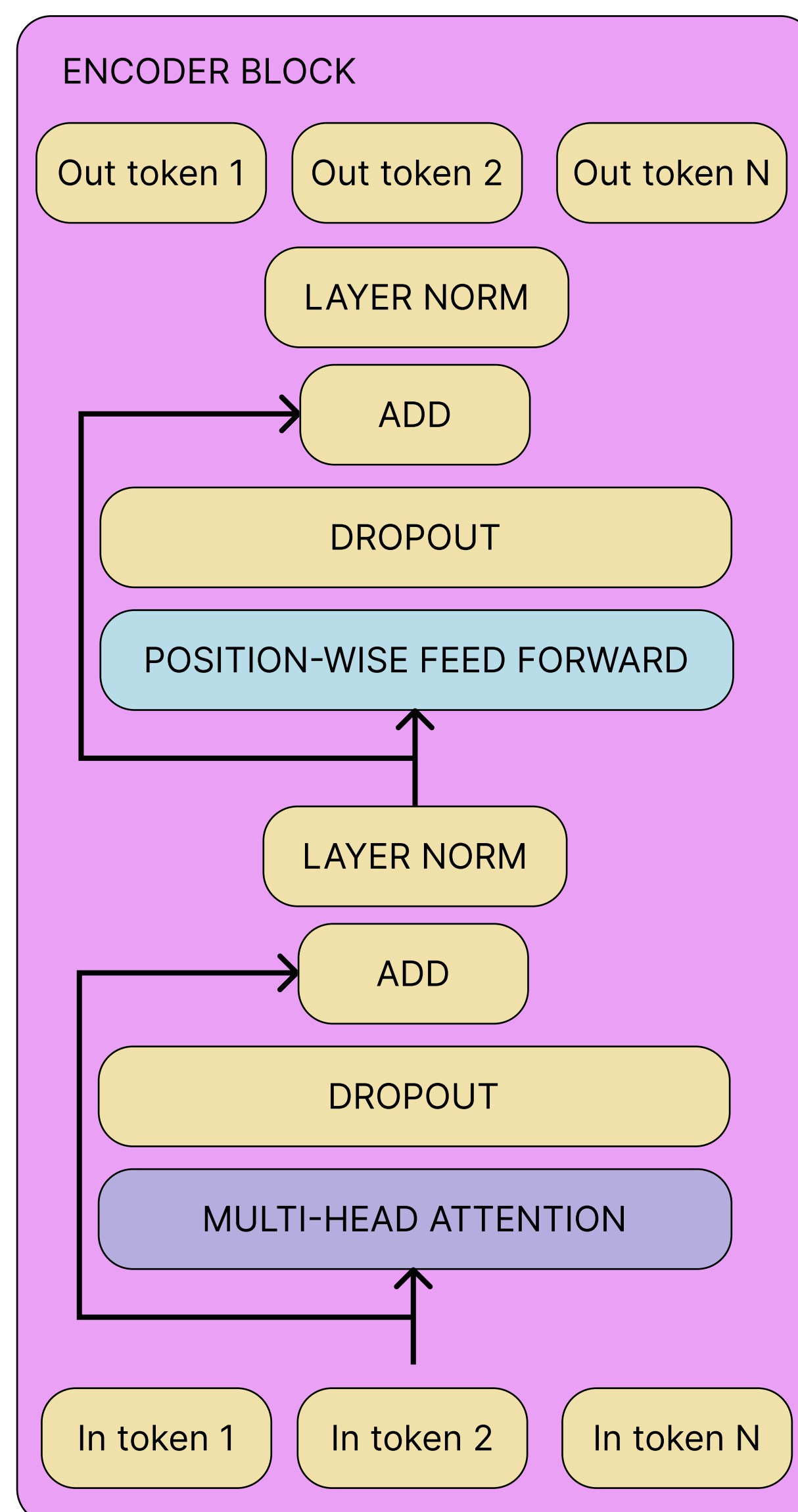
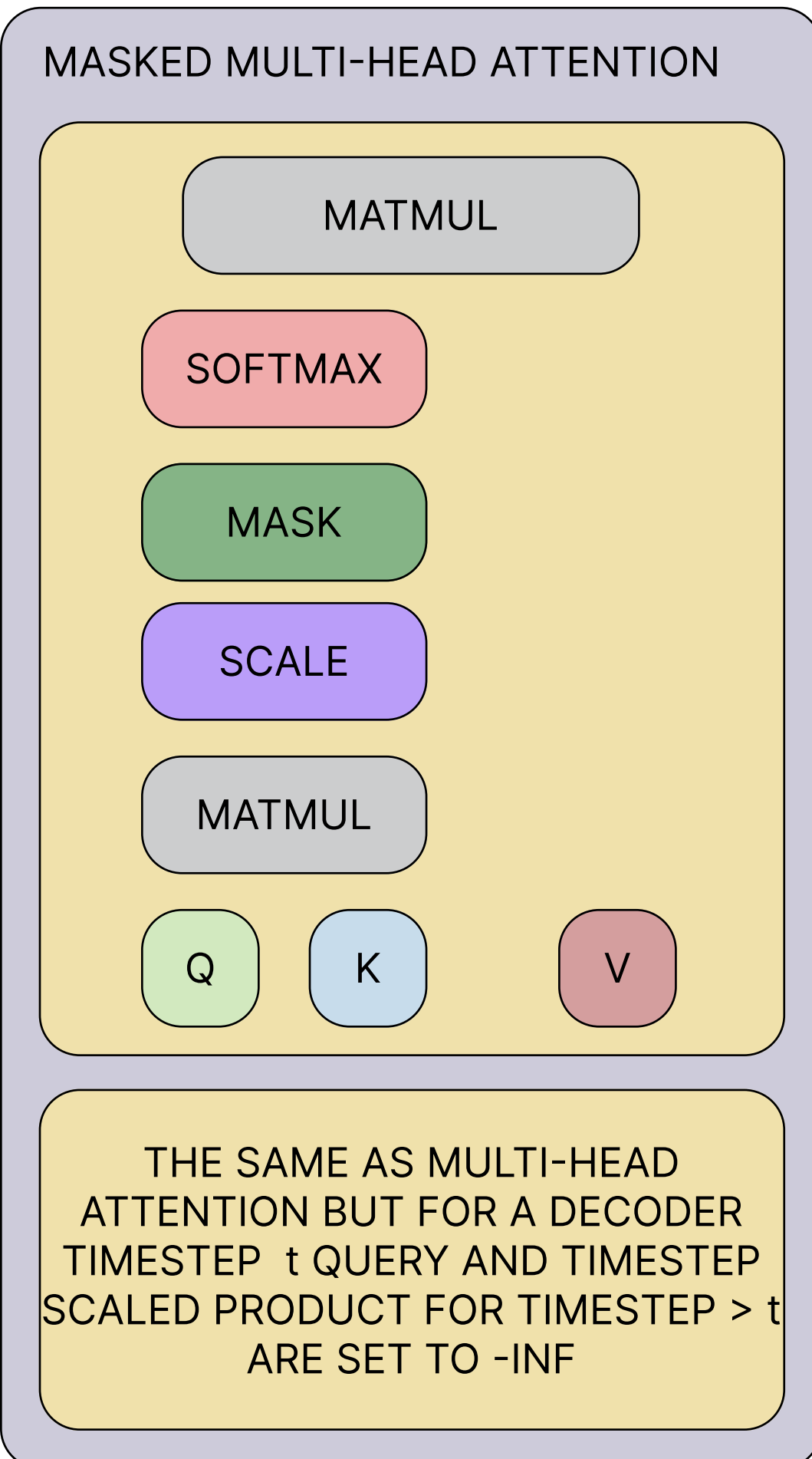
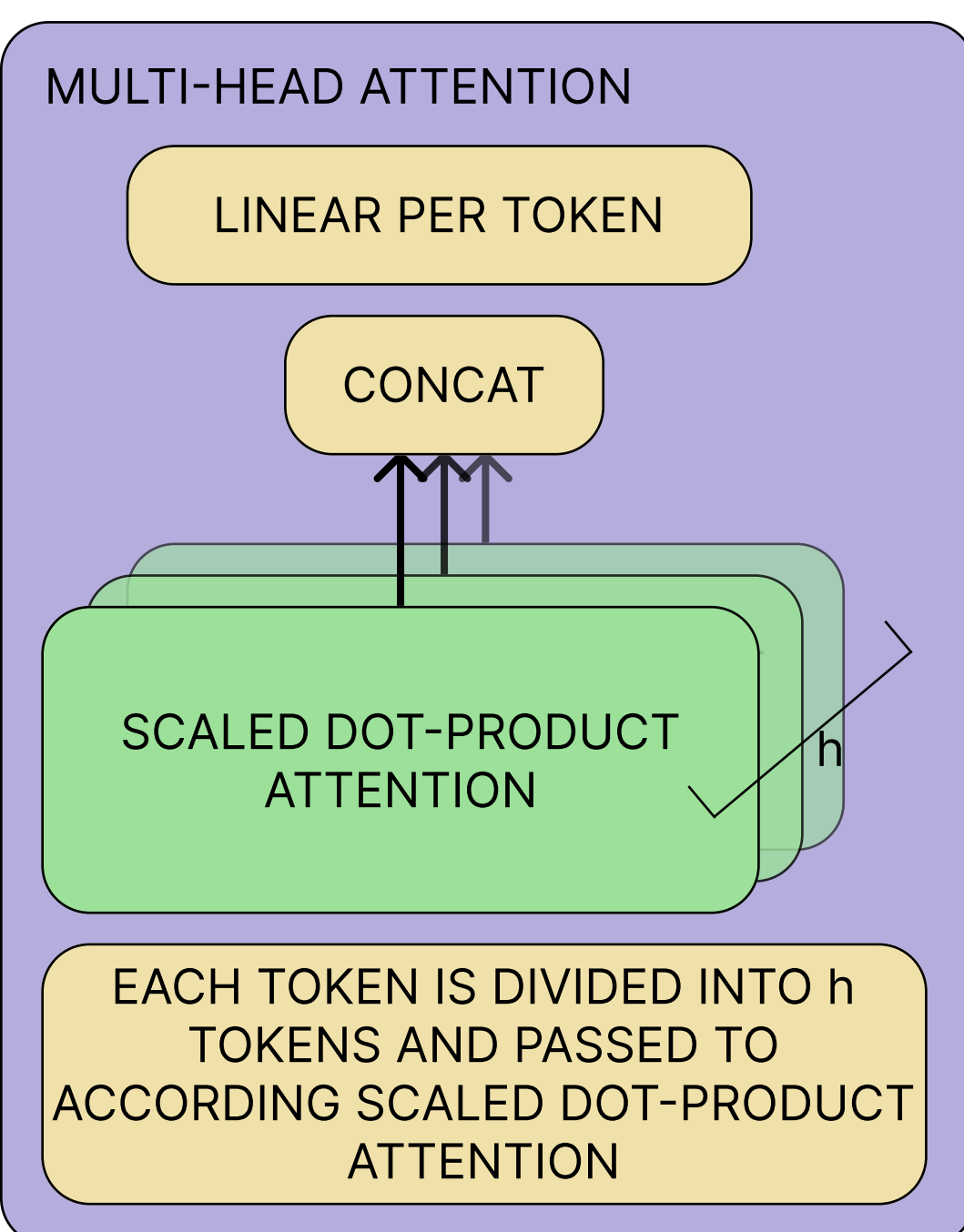
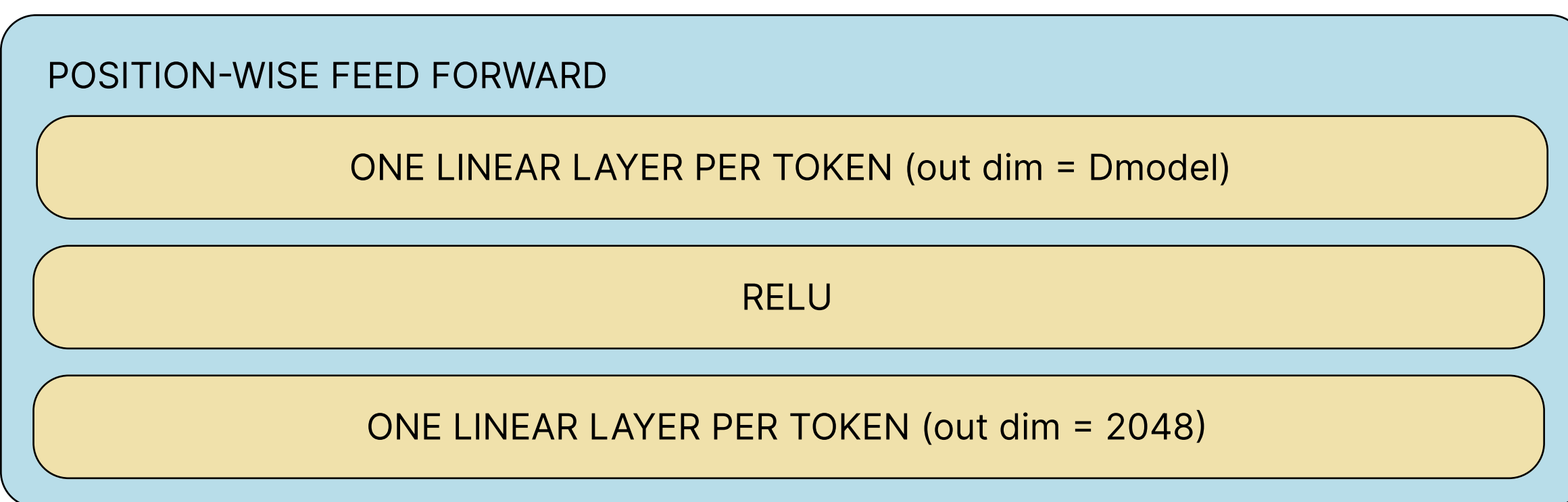
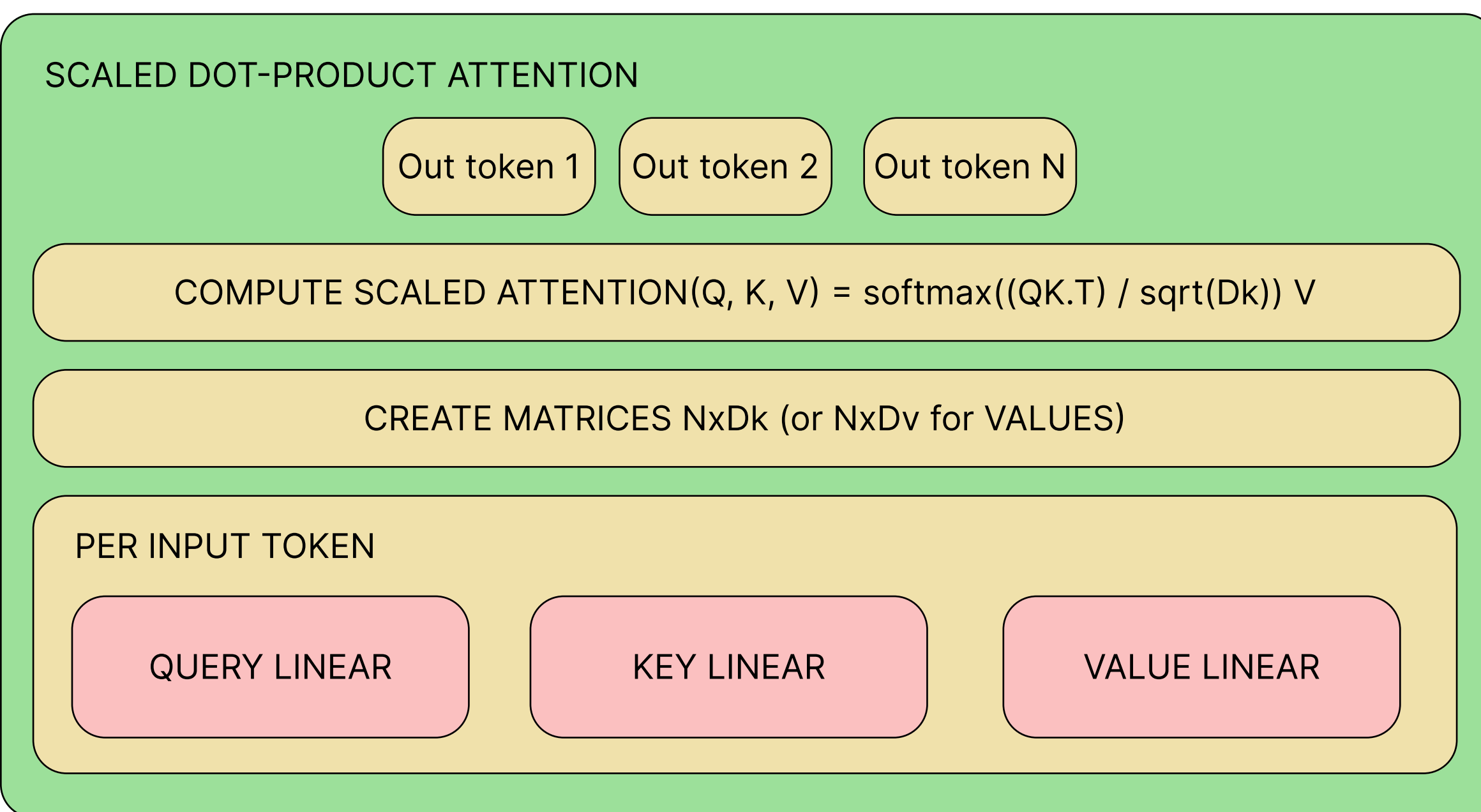


TRANSFORMER DIAGRAM

<https://arxiv.org/pdf/1706.03762.pdf>

$D_{\text{model}} = 512$ (e.g. embeddings size)
 $h = 8$ (no of attention heads)
 $D_k = D_v = D_{\text{model}} / h$ (k for keys and queries, v for values)
dropout rate is 0.1



layers with * share the same weights but for embedding layers those weights are scaled by $\sqrt{D_{\text{model}}}$