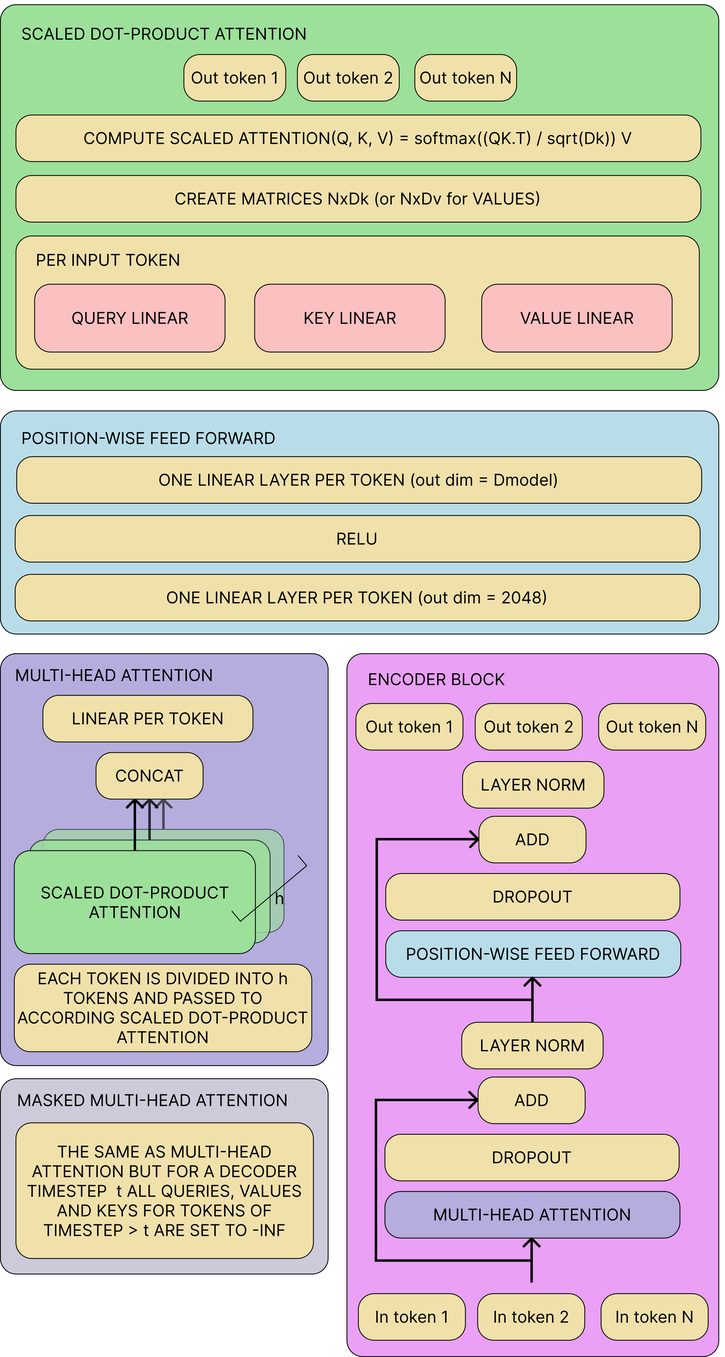
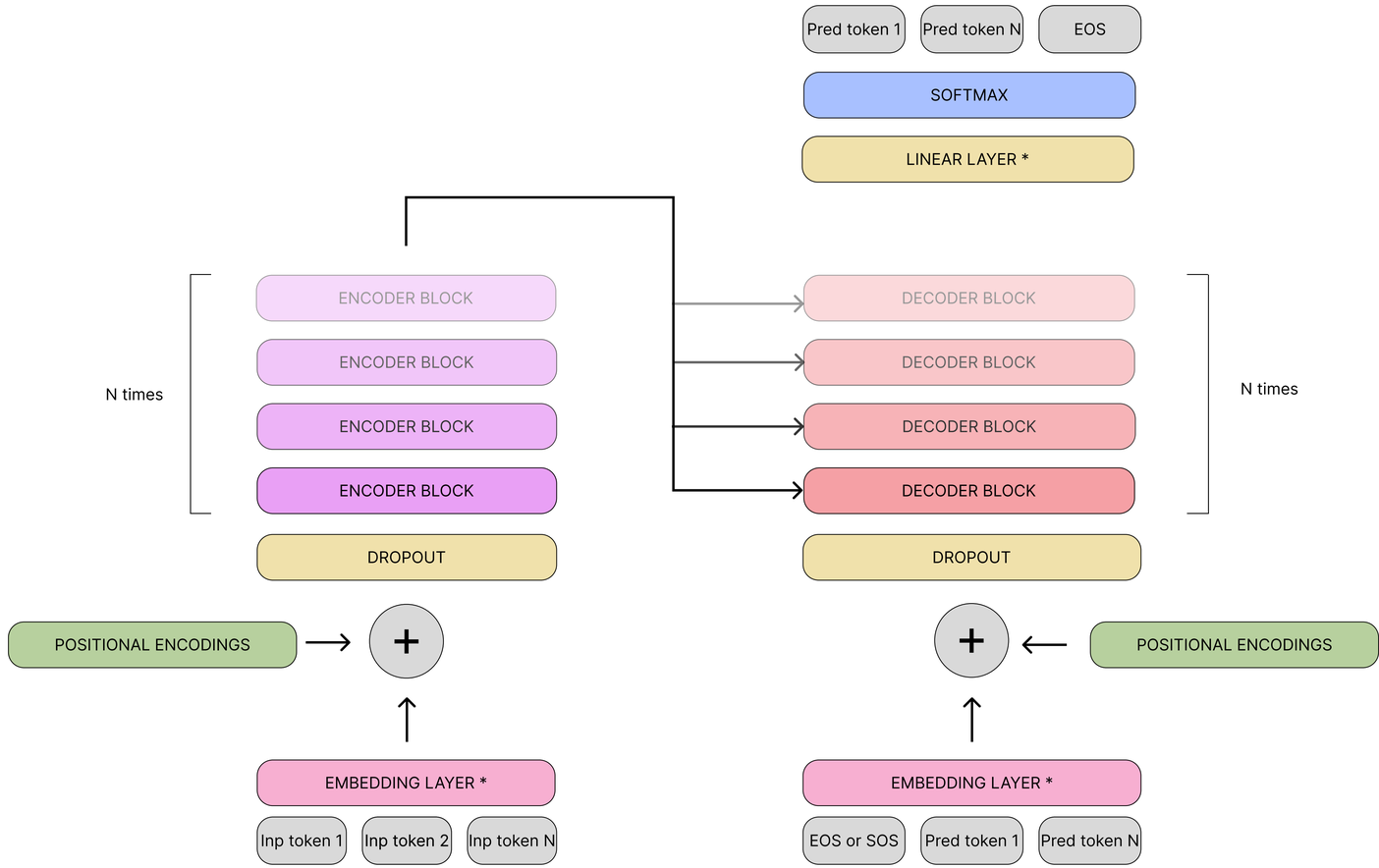
TRANSFORMER DIAGRAM

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

Dmodel = 512 (e.g. embeddings size)
h = 8 (no of attention heads)
Dk = Dv = Dmodel / h (k for keys and queries, v for values)





layers with * share the same weights but for embedding layers those weights are scaled by sqrt(Dmodel)