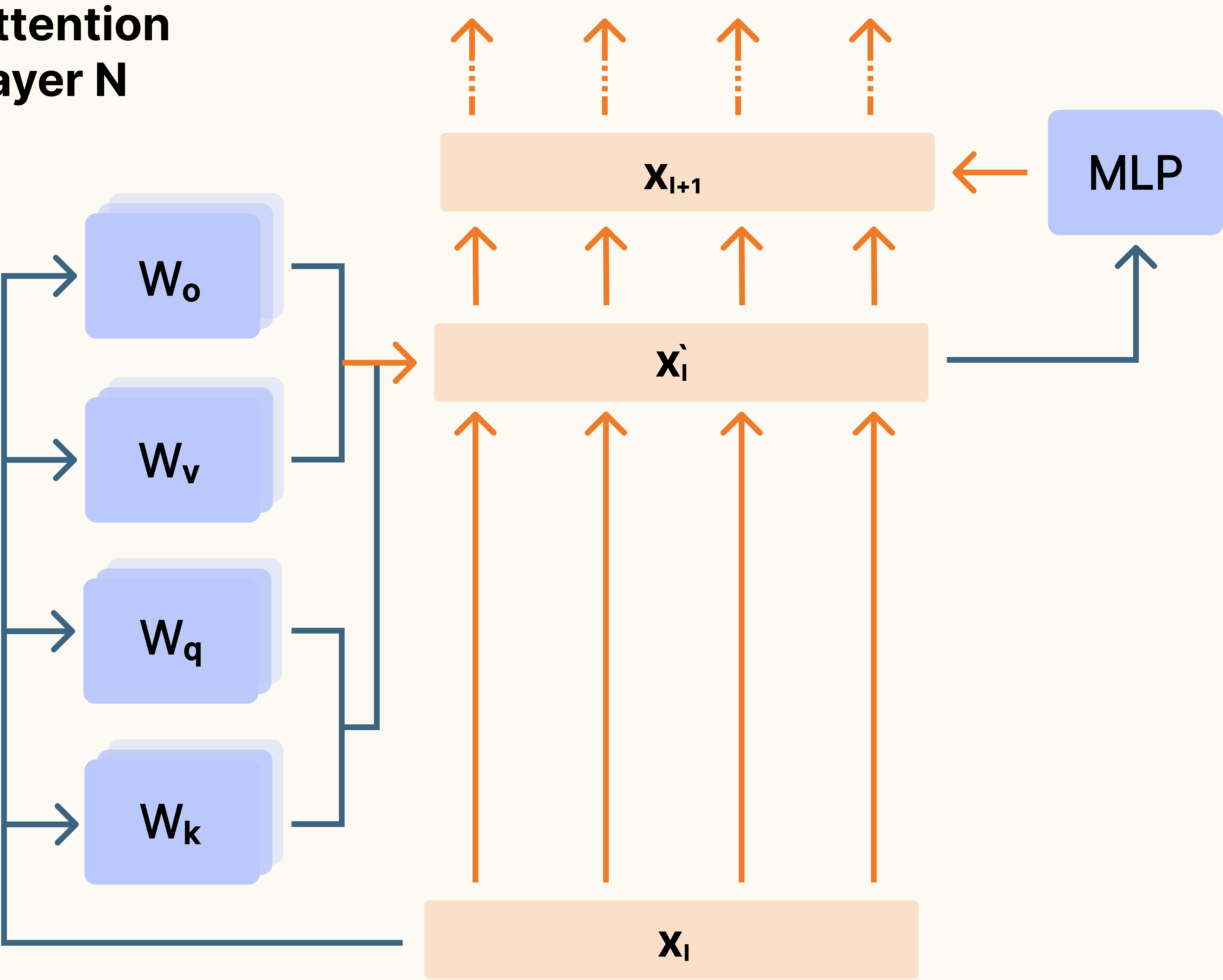
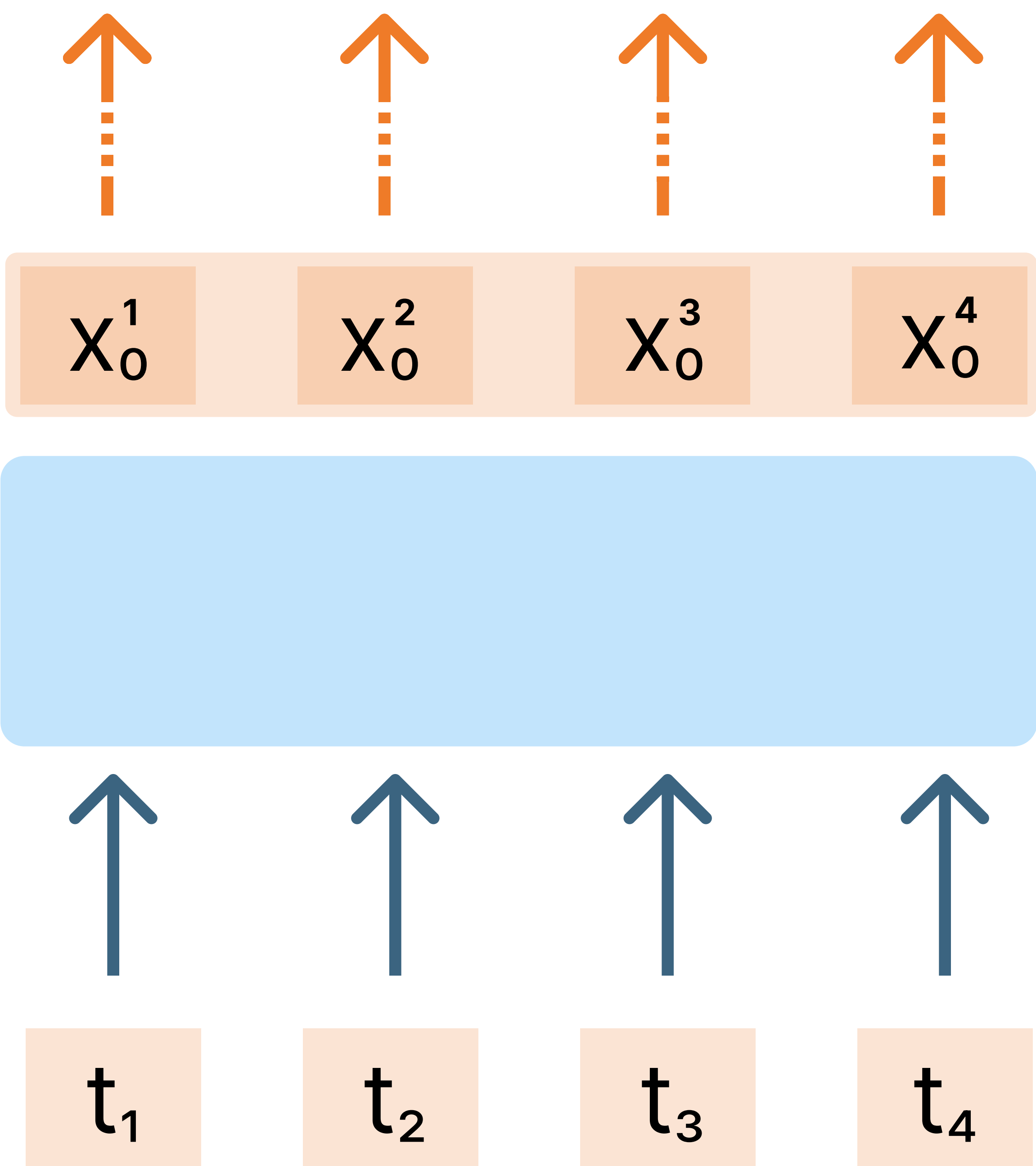


Attention
Layer N



Embedding
Layer

Input tokens



residual stream