Attention Layer N

W_o

W_v

W_q

W_k

x_{l+1}

MLP

x'_l

x_l

x_0^1 x_0^2 x_0^3 x_0^4

residual stream

Embedding Layer

Input tokens

t_1 t_2 t_3 t_4