

h_{i-1} h_i Residual
Stream h_{i+1}

Feedforward

Layer Norm

MultiHead
Attention

Layer Norm

 x_{i-1} x_i x_{i+1}

...

...