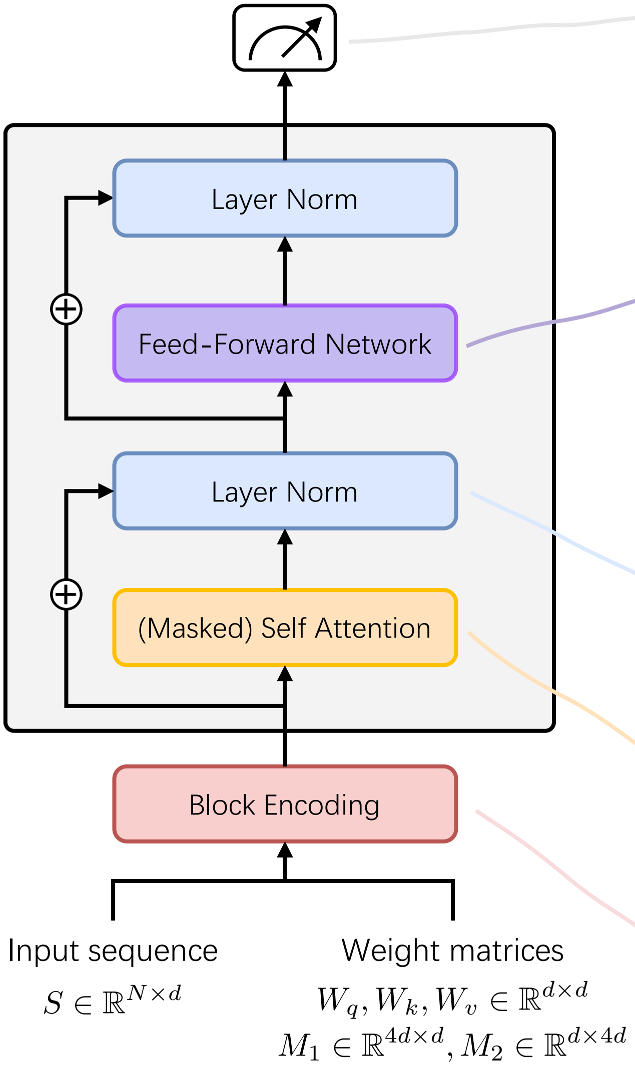
Layer Norm
Feed-Forward Network
Layer Norm
(Masked) Self Attention
Block Encoding
Input sequence
$S \in \mathbb{R}^{N \times d}$
Weight matrices
$W_q, W_k, W_v \in \mathbb{R}^{d \times d}$
$M_1 \in \mathbb{R}^{4d \times d}, M_2 \in \mathbb{R}^{d \times 4d}$

Quantum transformer for the j-th token

$$\mathrm{Transformer}(S, j) = \mathrm{LN}(\mathrm{FFN}(\mathrm{LN}(\mathrm{Attention}(S, j))))$$

Quantum feed-forward network with an activation function σ and an input vector ψ

$$\sum_{k=1}^{d} \left(M_2 \cdot \sigma(M_1 \cdot \psi)\right)_k |k\rangle$$

Quantum residual connection with layer normalization for the j-th token

$$\sum_{k=1}^{d} \mathrm{LN}(\mathrm{Atten}(S)_j + S_j)_k |k\rangle$$

Quantum self-attention matrix for the j-th token

$$\mathrm{Atten}(S)_j = \mathrm{softmax}(QK^T/\alpha_0)_j \cdot V$$

Block encoding of the input matrices:

$$\begin{bmatrix} S/\alpha & * \\ * & * \end{bmatrix}, \begin{bmatrix} Q/\alpha & * \\ * & * \end{bmatrix}, \begin{bmatrix} K/\alpha & * \\ * & * \end{bmatrix}, \begin{bmatrix} V/\alpha & * \\ * & * \end{bmatrix}$$