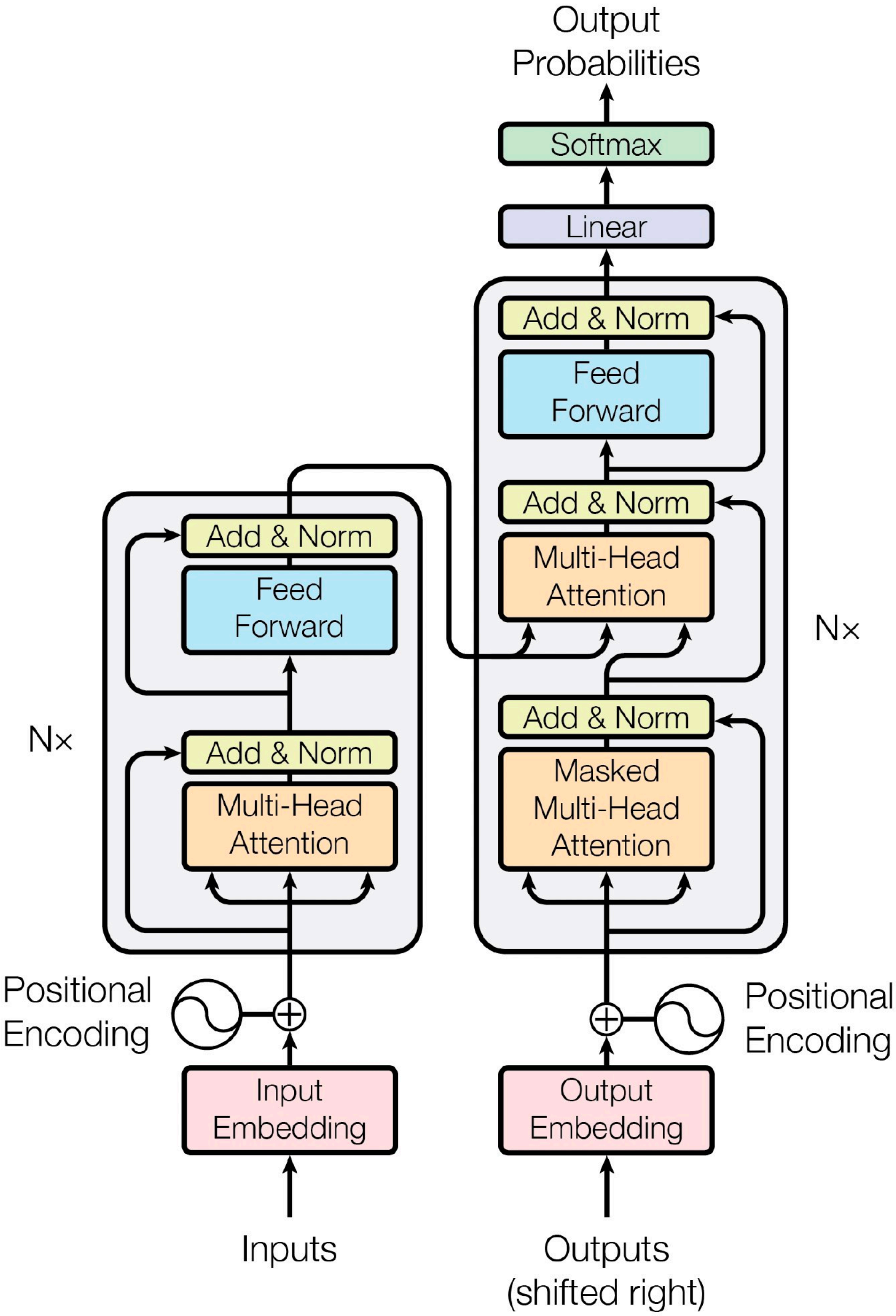
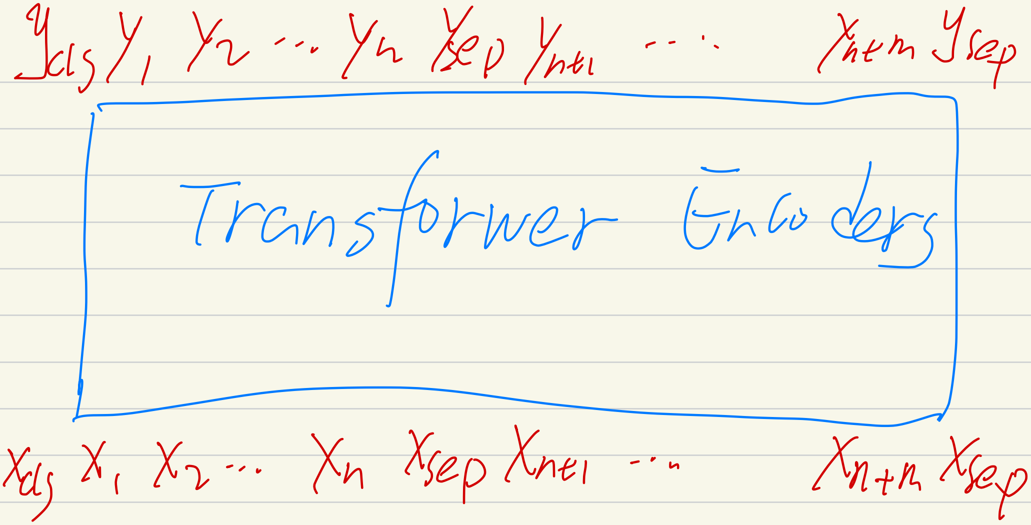


Figure 1: The Transformer - model architecture.


y_{cls} y_1 y_2 ... y_n y_{sep} y_{n+1} ... x_{n+m} y_{sep}
Transformer Encoders
x_{cls} x_1 x_2 ... x_n x_{sep} x_{n+1} ... x_{n+m} x_{sep}

NSP: $\sigma(A_{NSP} \cdot y_{cls} + b_{NSP})$

$\Downarrow$

CE - Loss

MLM: For all [mask] / [replace] / [nochange]

$$z = \text{Softmax}(\underset{\mathbb{R}^{|V| \times d_h}}{\underset{\uparrow}{A_{MLM}}} y + b_{MLM})$$

$-\log(z_w)$