Part 1 : Encode Layer

$X = [t_1, t_2, t_3 \dots]$

$E^1 = W_{emb}(x_i)$

$E^1 = E^1 \sqrt{d}$

$X^0 = E^1 + PE$

$PE = \left[\cos\left[\frac{pos}{10^{8i/d}}\right], \sin\theta, \dots\right]$

$$PE_{pos,2i} = \sin\left(\frac{pos}{10^{2i/d}}\right) \quad \text{(even positions)}$$

$$PE_{pos,2i+1} = \cos\left(\frac{pos}{10^{2i/d}}\right) \quad \text{(odd positions)}$$

$$i \in \left[0, \dots, \frac{d}{2} - 1\right]$$

$Q = X^0 W^Q \qquad K = X^0 W^K \qquad V = X^0 W^V$

Single head

$score = \frac{Q_i K_j^T}{\sqrt{d_K}}$ mask to $-\infty$

$head_j = Attention(Q, K, V) = Softmax(scores) \cdot V$

multi-head

$\Psi = concat(head_1, \dots, head_h)$

$MHA(H) = \Psi \cdot W^O$

Add initial X^0 and normalize :

$residue(x) = X_0 + MHA(H)$

$H' = \gamma \odot \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta$ } layernorm

$\mu = \frac{1}{d}\sum x_i \quad , \quad \sigma^2 = \frac{1}{d}\sum (x_i - \mu)^2$

$\gamma, \beta \Rightarrow$ learnable params

$FFN(x) = \phi(x W_1 + B_1) W_2 + B_2$

$\phi = x \cdot \Phi(x)$

$\Phi(x) = \frac{1}{2}\left[1 + erf\left(\frac{x}{\sqrt{2}}\right)\right]$ GELU

$H'' = Layernorm(H' + FFN(H'))$

SO FAR : Encode = Embed(X) + PE → Q, K, V → score → Single head → H_1, H_2, H_3, H_4 → concat → MHA → add & layernorm → FFN → A&L

$H^l = Encode(H^{l-1}) \quad , \quad H^0 = X^0$

Part 2 : Decode Layer (Analogous to Encoder Layer)

$T_0 = Embed(Y)\sqrt{d} + PE$

$T_1 = LayerNorm(T_0 + MHA_{masked}(T_0))$

Cross attention : T_L from iter 2 ; T_2 from iter 2

$T_2 = Layernorm(T_1 + MHA_{cross}(Q = T_1, K \Rightarrow H^l, V \Rightarrow H^l))$

$T^3 = Layernorm(T_2 + FFN(T_2))$

$T^{l+1} = T_3 \Longrightarrow$ Stack decoder layer $\Rightarrow T^l = Decode(T^{l-1})$

Part 3 : Prediction

$Logits = T^l W_{out} + B_{out}$

$P(next\ token\ |\ context) = softmax(logits)$