$$Z = \text{softmax}\left(\frac{1}{\sqrt{d}} Q K^T\right) V$$

$L \times d'$ $\quad$ $L \times L$ $\quad$ $L \times d'$ $\quad$ $d' \times L$

$$Q = W^Q X \quad K = W^K X \quad V = W^V X$$

$L \times d'$ $\quad$ $d' \times L$ $\quad$ $L \times d$

$M \times d$ $\quad$ $K' = W^K y$ $\quad$ $V = W^V y$

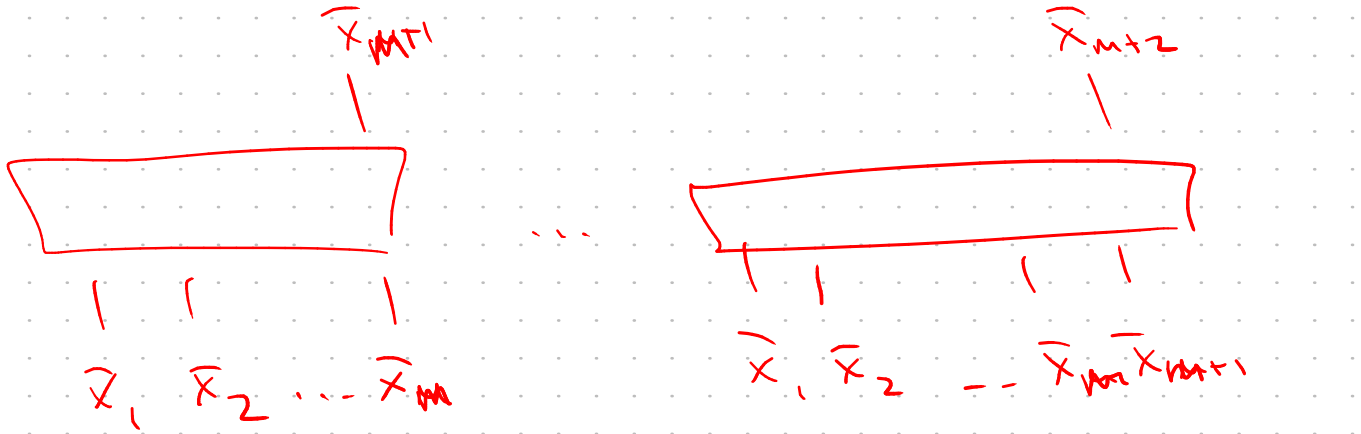
$L \times d'$ $\quad$ $d' \times M$ $\quad$ $M \times d'$ $\quad$ $L \times M$

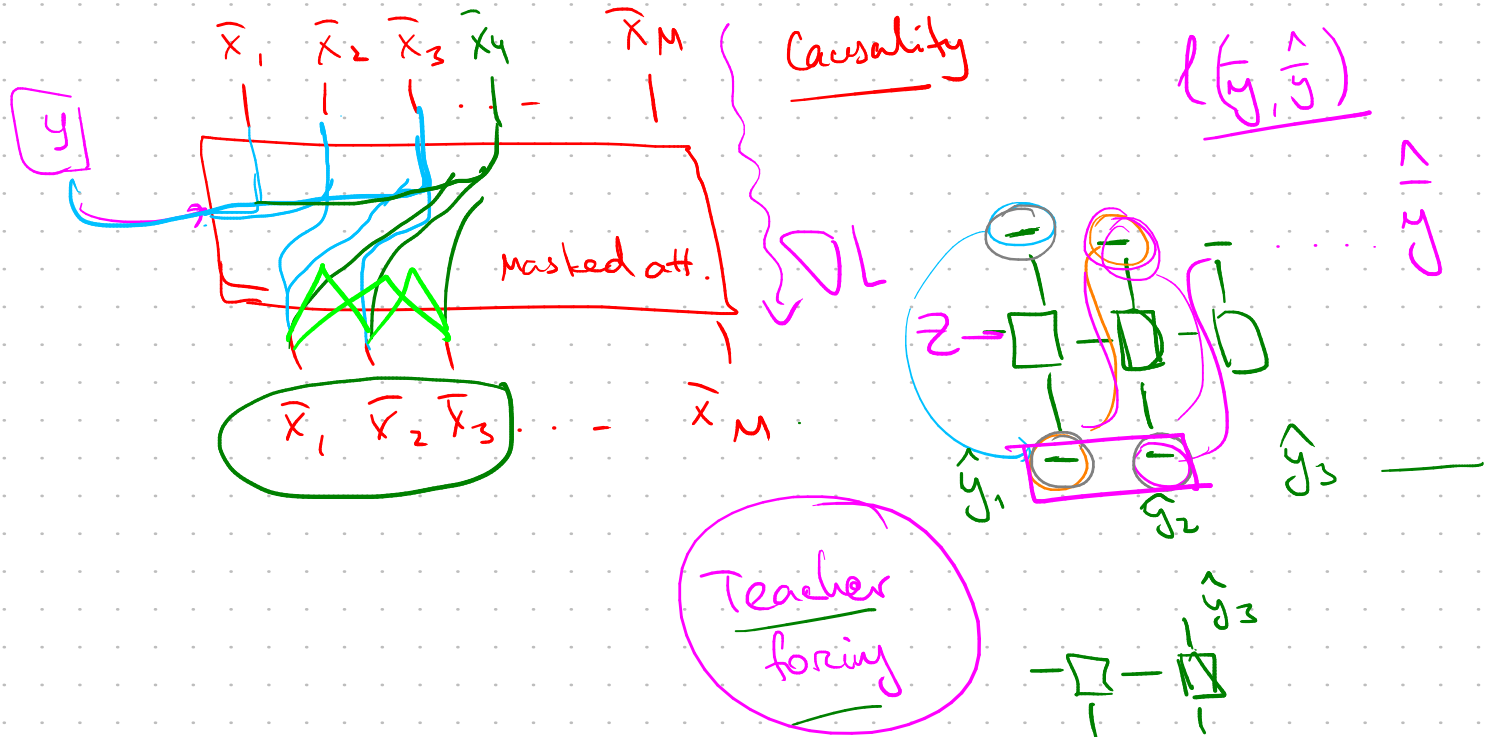
encoder output

GPT-3: $d = 12k$ $\quad$ ≈ 100 layers

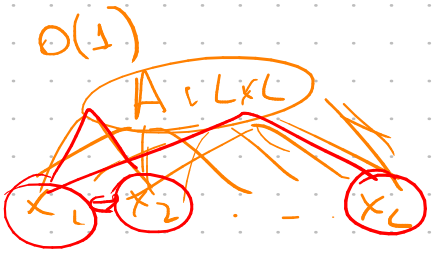
$$(4d^2 + 8d^2) \cdot 100 = 12 \cdot 100 \cdot 12k^2 = 1200 \cdot 144M \approx$$

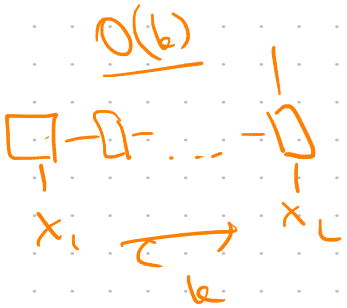
$\approx 170B$





Teacher forcing


O(1)
A : L×L
x1
x2
xL


O(b)
x1
xL
b