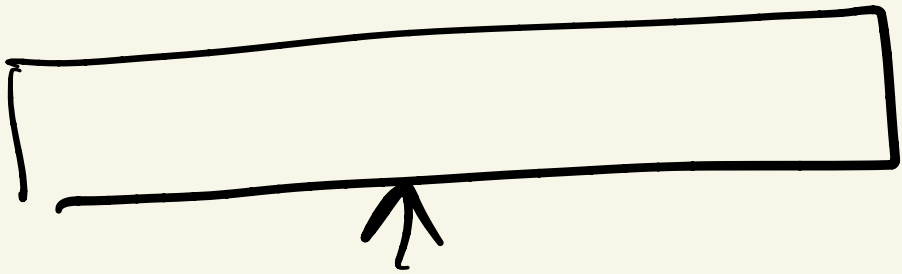
Suppose we have a pre-trained decoder-only model.

We have some data of the form (news art., summ)

How do we fine-tune the model on this

(x, y)



$x_1 \cdots x_N$ $y_1 \cdots y_M$

News Article

Sumary

$$\{(x, y)\}_{i=1}^{K}$$

$$\max_{\theta} \prod_{i=1}^{K} P(y \mid x)$$

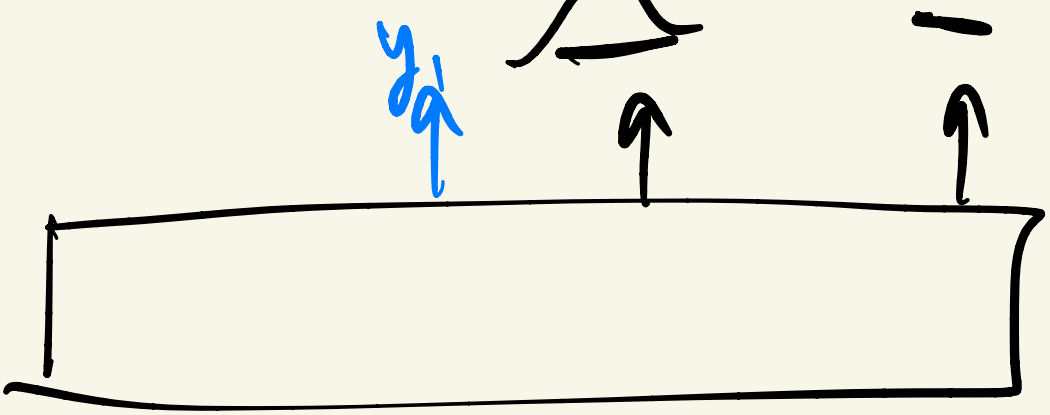
$$\max_{\theta} \sum_{i=1}^{K} \log P(y \mid x)$$

$$\min -\sum \log P(y \mid x)$$

loss

$$\sum_{i=1}^{N} \log P(y_i \mid x \; y_{i-1})$$

$P(y_2 \mid x, y_1)$



$$Loss = -\sum_{j=1}^{N} P(y_j \mid x \; y_1 \cdots y_{j-1})$$

Pre-training

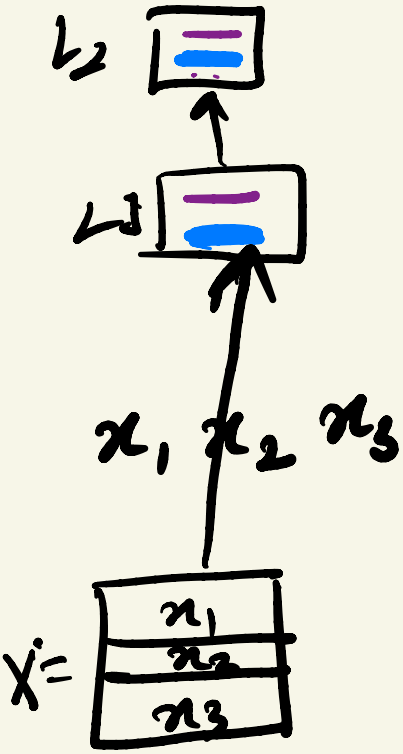
$$\max_{\theta} p(x)$$

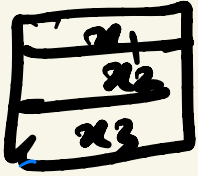
$$= \max_{\theta} \prod p(x_i \mid \ldots x_{i-1})$$

Transformer — decoder-only

2 Layers

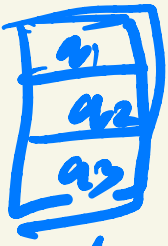
— → Attention

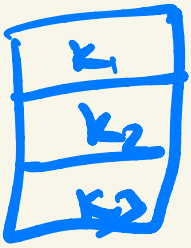


L1 → X = 

→ Attention

W^Q W^V W^K

$q =$ 

$k =$ 

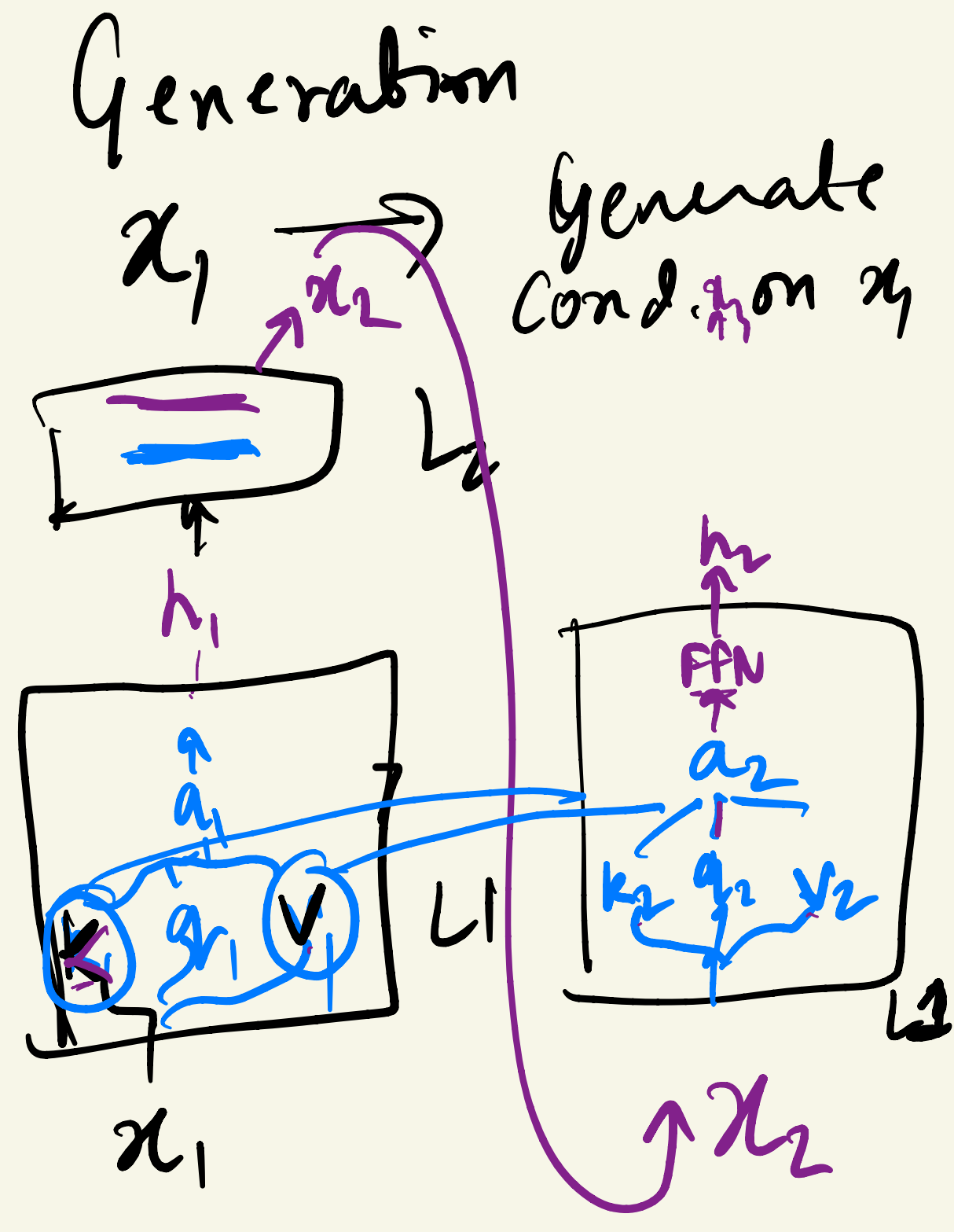
$v =$ 

$$O = \text{Softmax}\left(\frac{q k^T}{\sqrt{d}} + \text{mask}\right)$$

$$a = O \cdot V$$

→ Output of Attention

→ LN + Residuck + FFN


Generation
Generate conditioned on x_1
x_1
x_2
L_2
h_1
h_2
FFN
a_1
a_2
K
q
V
k_2
q_2
v_2
L1
L1
x_1
x_2

KV - cache

k_1 v_1

k_2 v_2

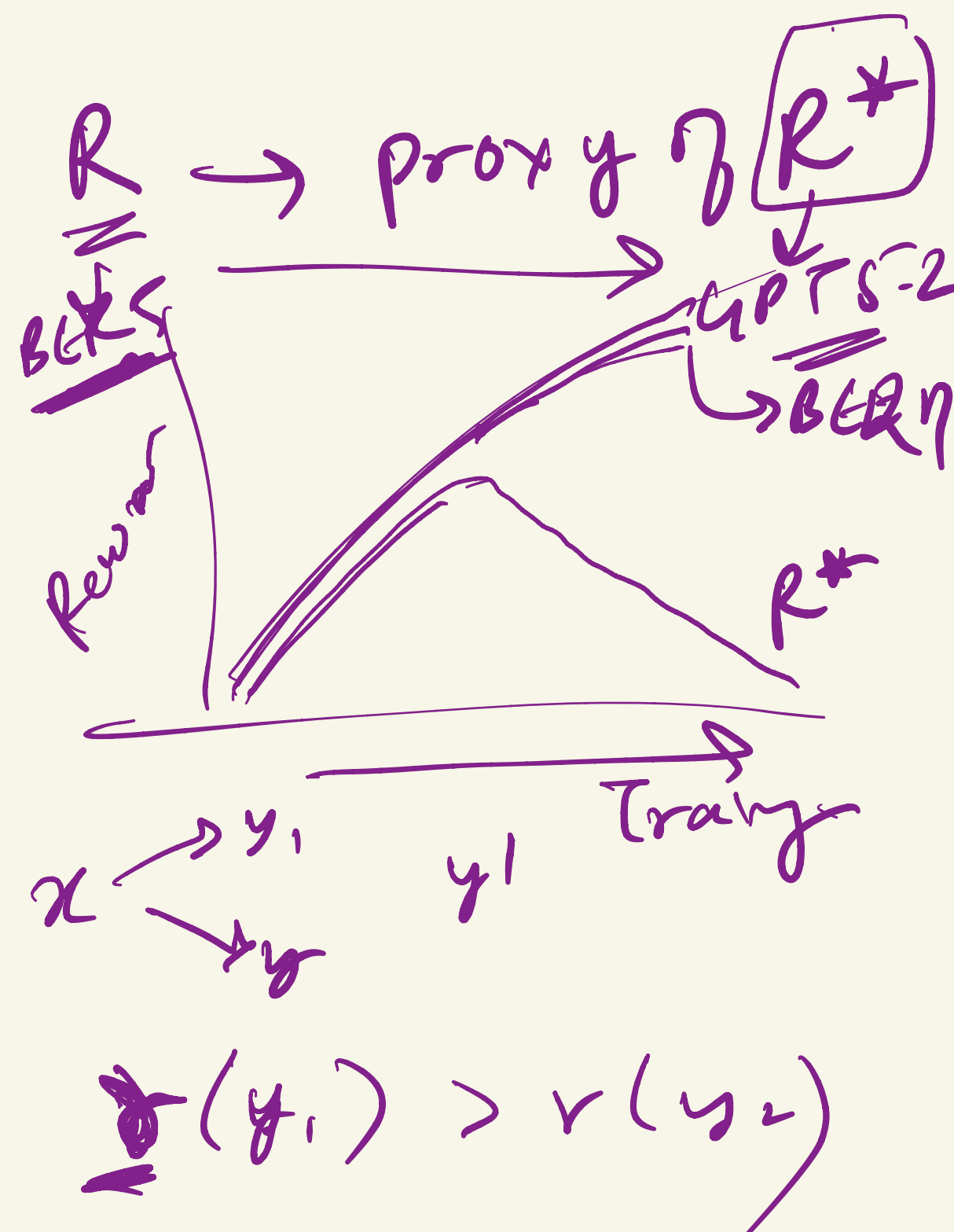
.

.

.

for each layer

for each prev. timestep



$r(y_1) > r(y_2)$