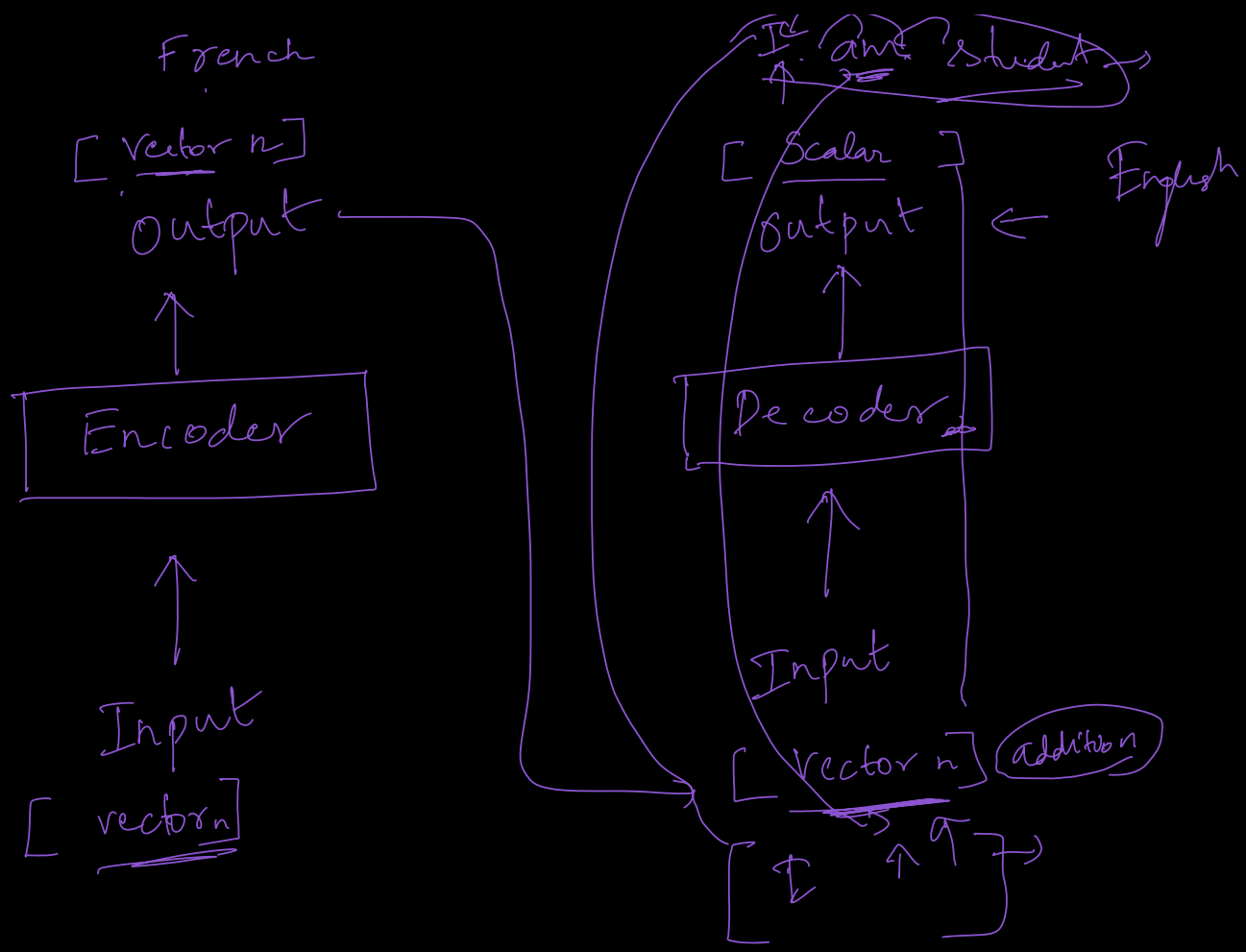
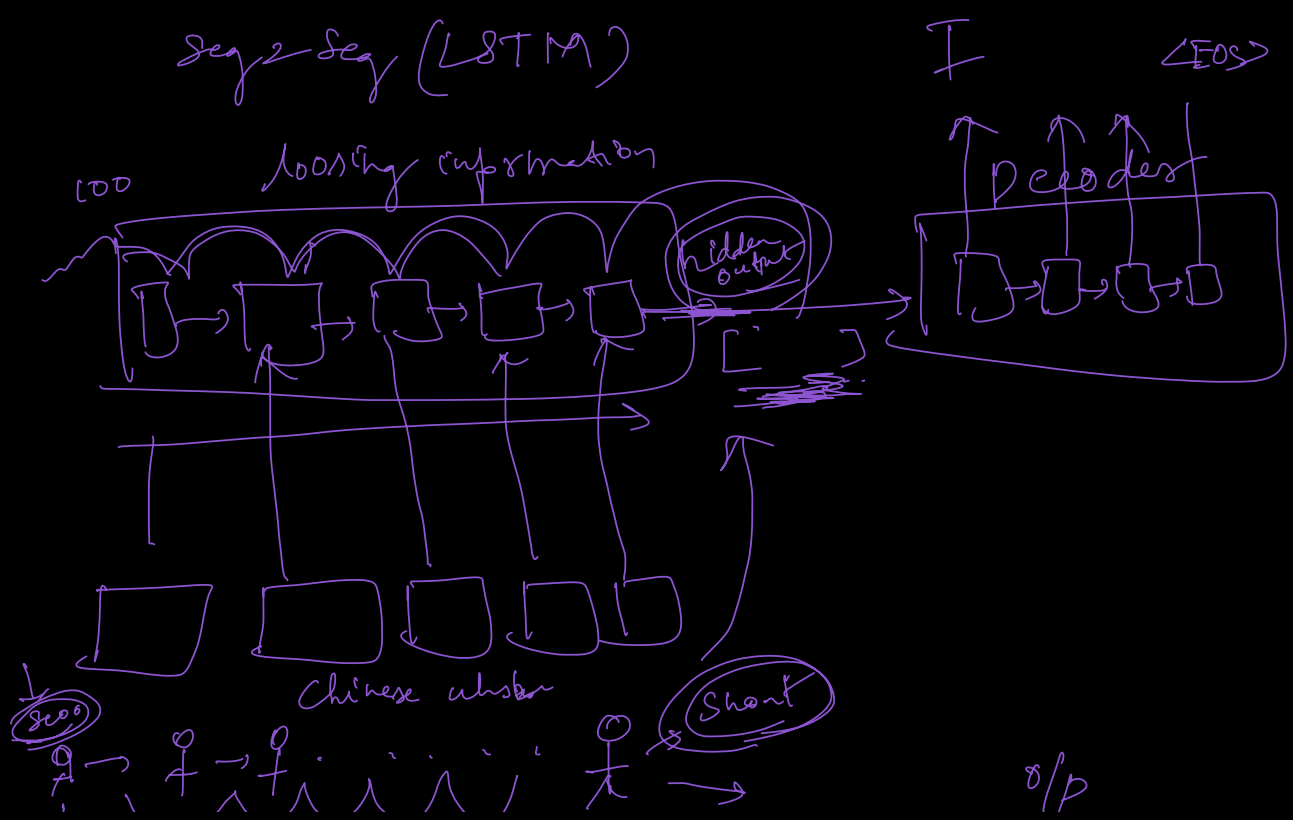
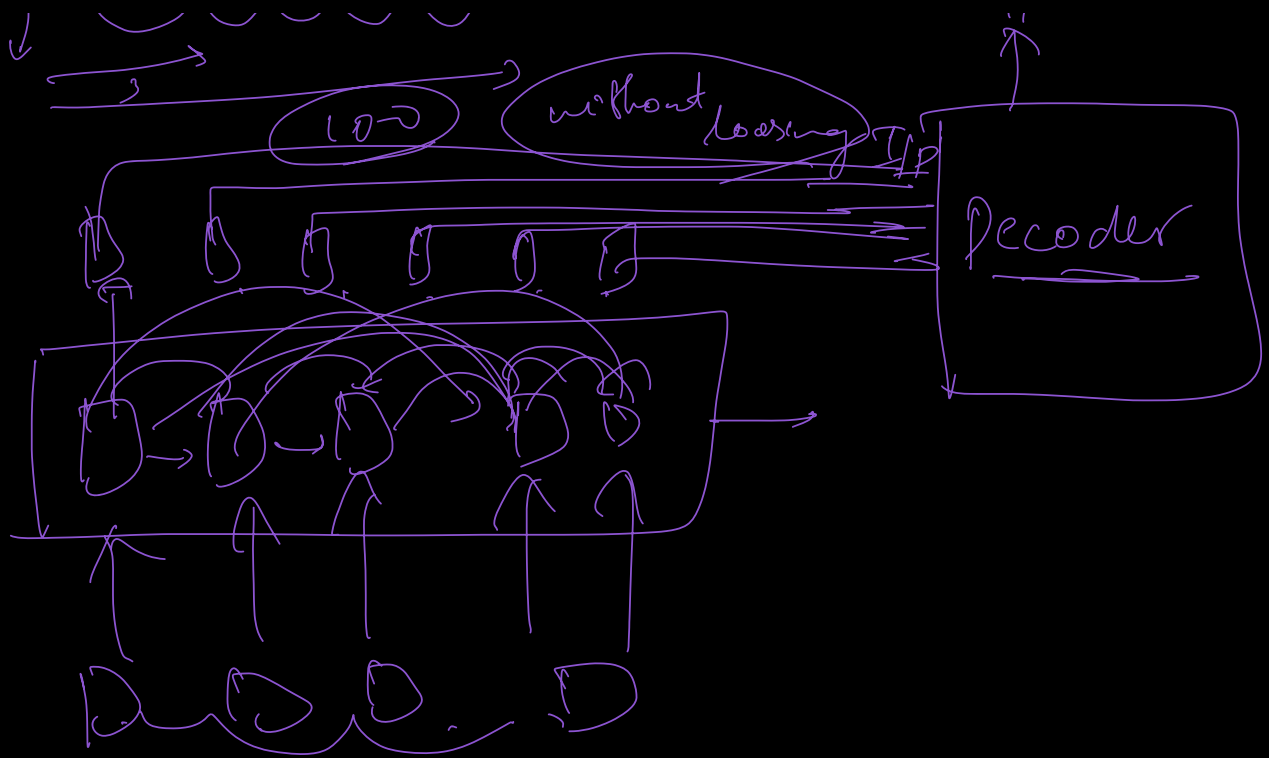
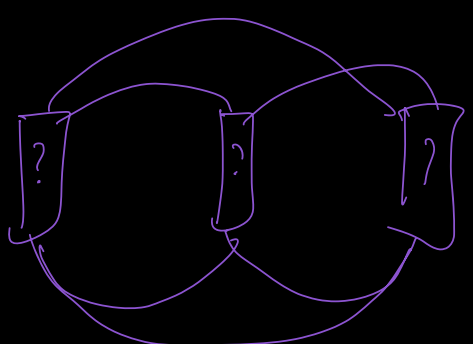

French
[vector n]
output
Encoder
Input
[vector n]
I am student
[Scalar]
output
English
Decoder
Input
[vector n]
addition


Seq 2 Seq (LSTM)
I
<EOS>
100
loosing information
hidden output
Decoder
Chinese whisker
Short
o/p

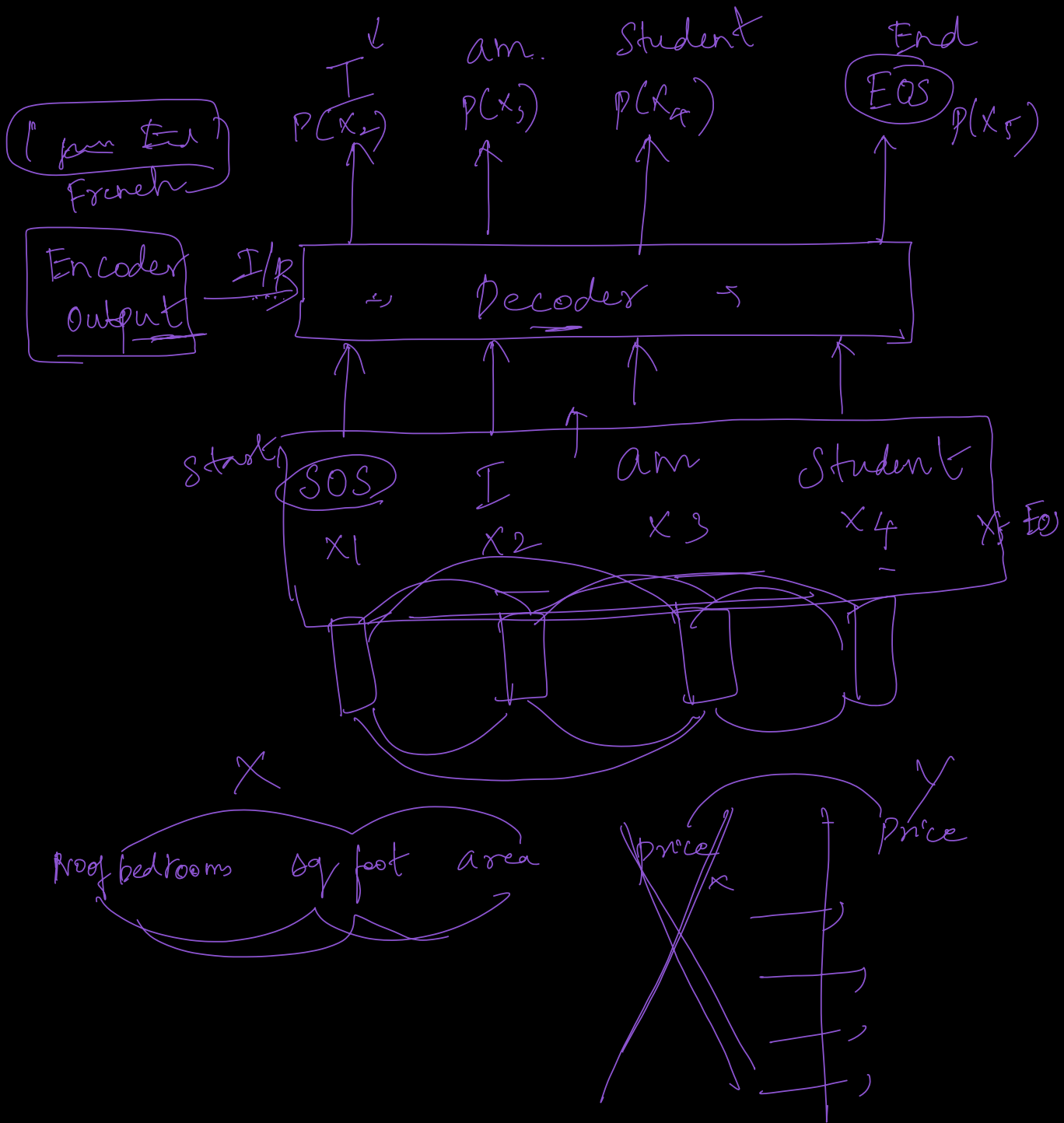


Masked Self Attention

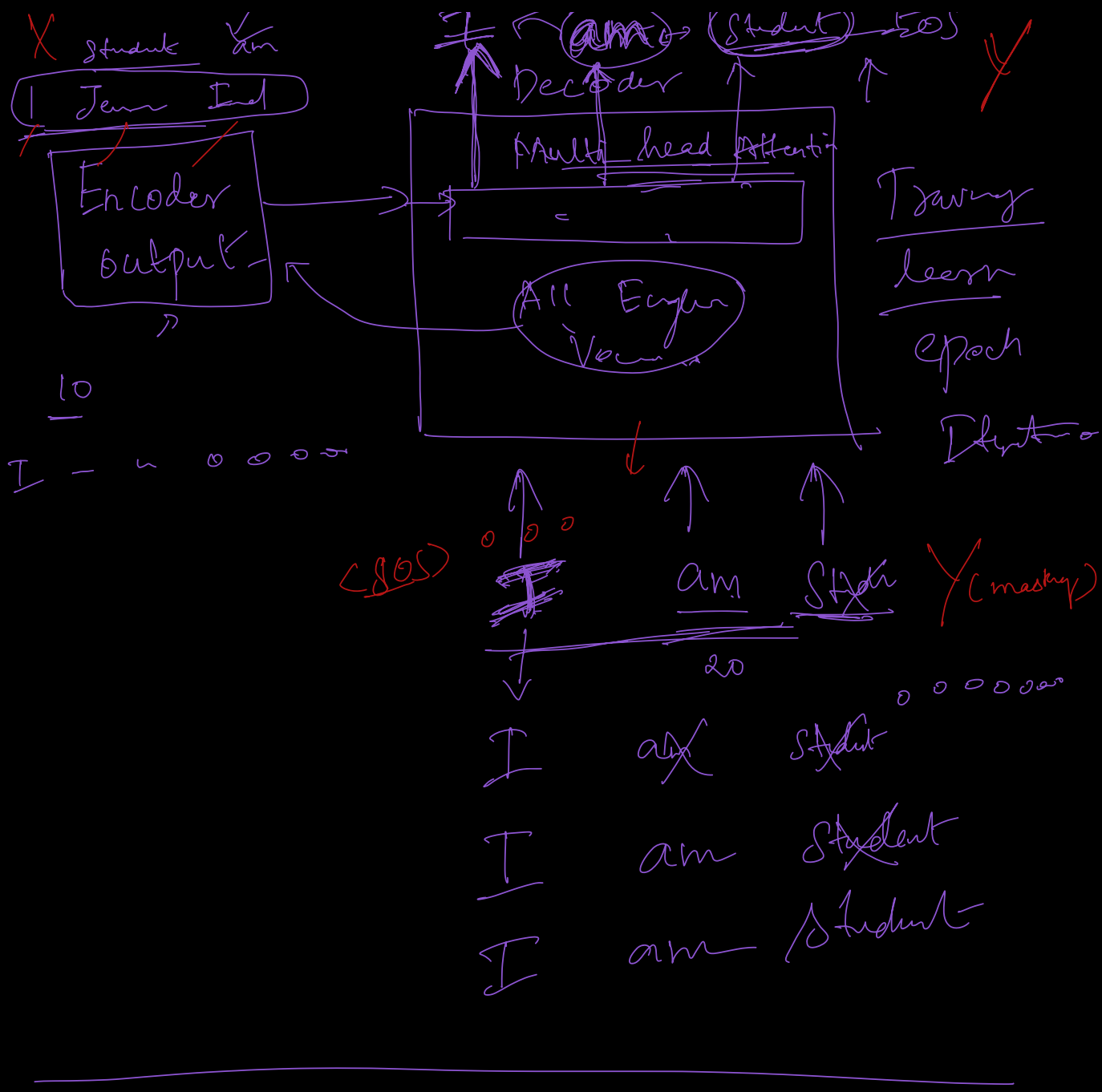


Decoder

Encoder



I
am.
Student
End
EOS
P(X2)
P(X3)
P(X4)
P(X5)
French
Encoder output
I/P
Decoder
start
SOS
I
am
Student
X1
X2
X3
X4
X5 EOS
X
Y
No of bedrooms
sq foot
area
price
Price


X student am
1 Jan Ind
Encoder output
10
I am (student) EOS
Y
Decoder
Multi head Attention
All Engine Vocab
Training
learn
epoch
Iteration
(SOS)
I am Student
Y (masking)
20
I am Student
I am Student
I am Student

Training

Jusis

Encoder output + SOS → I am

Ecoder output + SOS + I → am

Encoder output + SOS + I + am → Student

Encoder output + SOS + I + am + Student → EOS

1,0000000

Training

English embedder

From

German

Keras embedder

No Pretrained

Vocabulary

10,000 → random

255, 512

Learnerable weights

Keras embedding

Englisht

1 2 3 4

Hi How are you ?

5 6 7

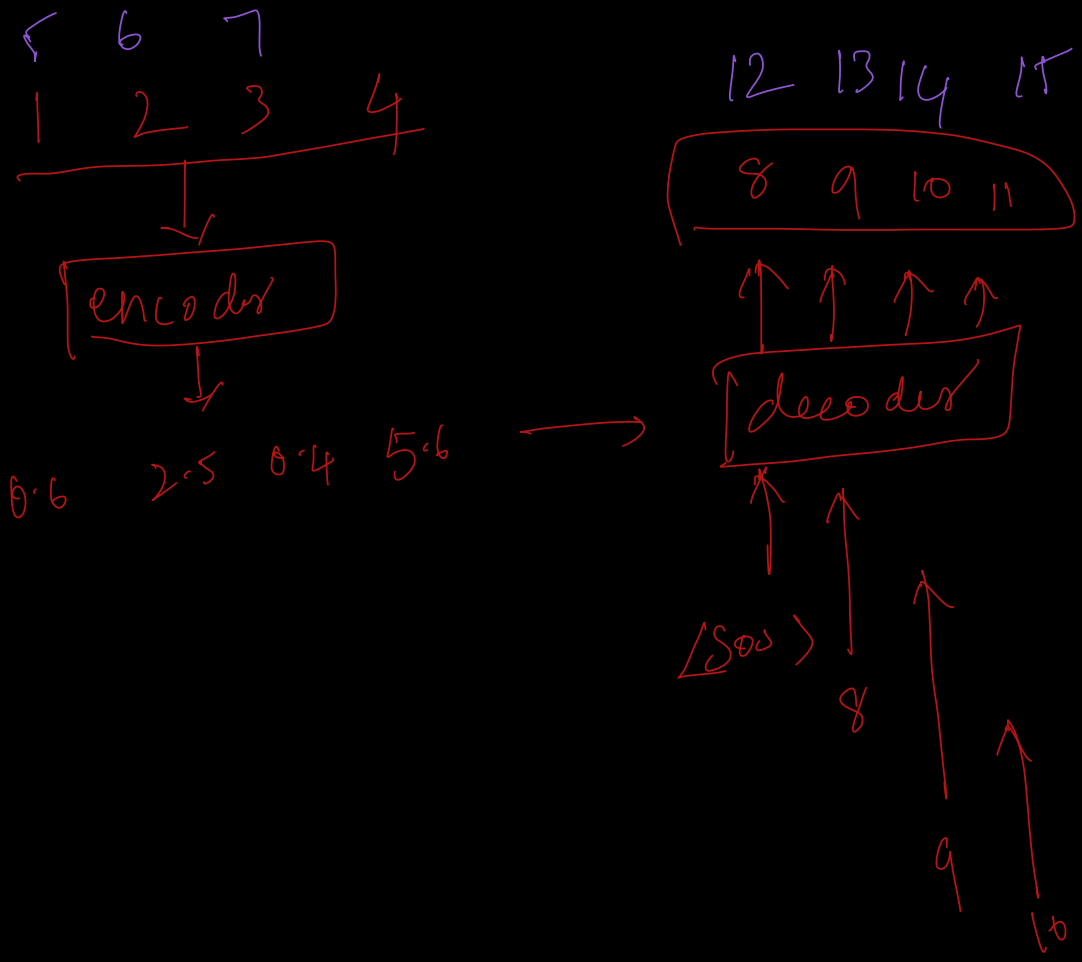
I am good ?

8 9 10 11

I o are g ?

12 13 14 15

h ŷ âr ŷôu ?



How Masked Attention ?

X_1 X_2

token 1 token 2

Embedding 1 Embedding 2

Similarity

$X_1 \cdot X_2$ ⟶ Self Attention

linear tranformation to achieve Multi head

Query $= W_Q \cdot X_1$

Key $= W_K \cdot X_2$

$[W_Q \cdot X_1] \cdot [W_K X_2^T]$ ⇒ Similarly with Linear Transformation

$$Z = \frac{[W_Q X_1][W_K X_2^T]}{\sqrt{d}}$$

Softmax (Z) ⇒ probability $\frac{e^{x_1} \cdot e^{x_2}}{e^{x_1} + e^{x_2}}$

Value = Matrix (weights)

Softmax $(Z) *$ Value

↳ Update embedding

$$Q = W_Q X_1$$

$$K = W_K X_2$$

$$\text{Values} = V = W_V(X_1, X_2)$$

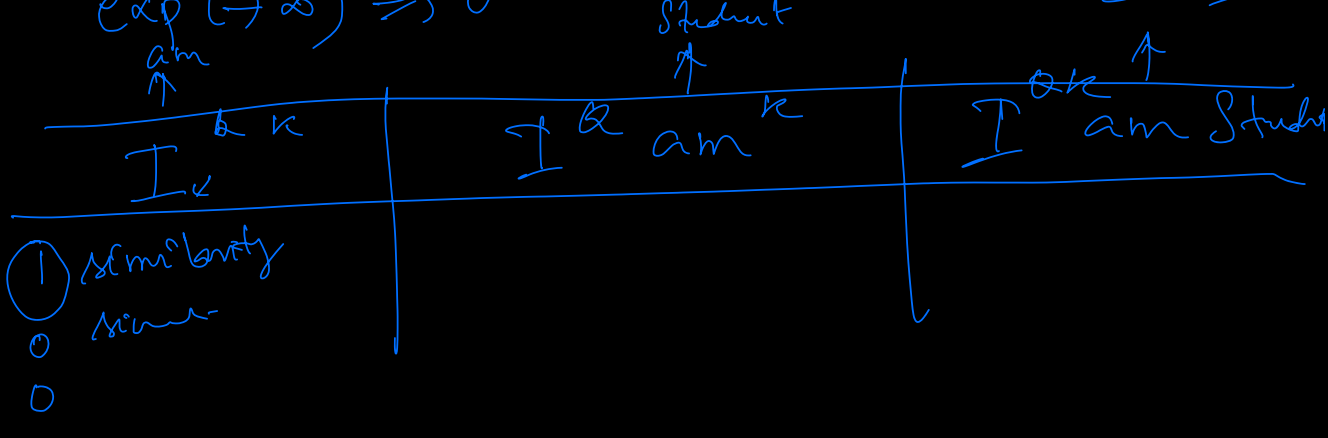
$$Z = \frac{K^T Q}{\sqrt{d}}$$

$$W = \begin{bmatrix} \exp(z_{11}) & \exp(z_{12}) - & \exp(z_{13}) - \\ 0 & \exp(z_{22}) - & \exp(z_{23}) - \\ 0 & 0 & \exp(z_{33}) - \end{bmatrix}$$

I am Student

$-\infty$ $-\infty$ $-\infty$

$\exp(-\infty) \Rightarrow 0$



Softmax

I	am $(-\infty)$	Student $(-\infty)$
0.5	0.2	0.3
1	0	0

$\exp(\cdot)$

$\exp(-\infty) = 0$

I	am	Student
0.8	0.2	0

I	am	Student
0.7	0.1	0.2

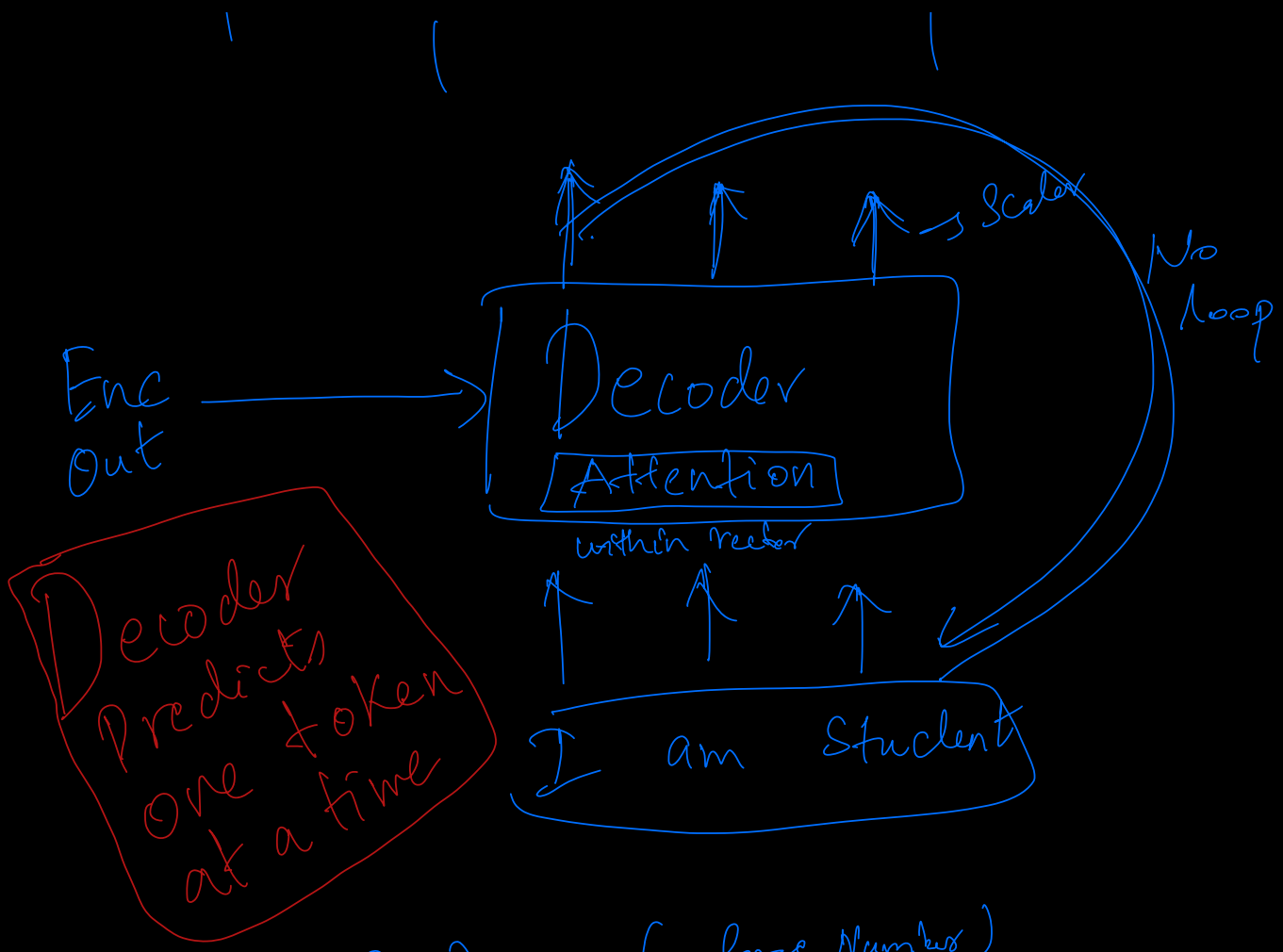
Softmax

am Student

$\langle F, O \rangle$

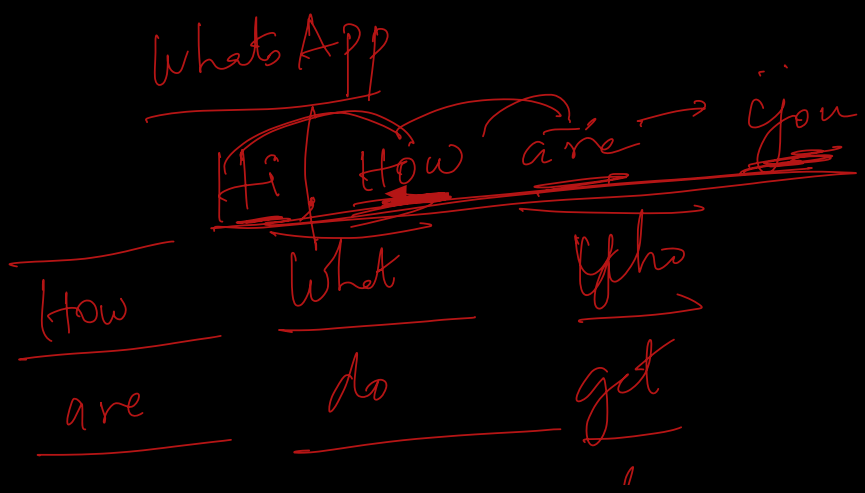
I am Stud

	I am	I am Stud
1	$Sm_1(Z_{12}, Z_{22})$,	$Sm_1(Z_{13}, Z_{23}, Z_{33})$
0	$Sm_1(Z_{12}, Z_{22})$,	$Sm_1(Z_{13}, Z_{23}, Z_{33})$
0	0	$Sm_1(Z_{13}, Z_{23}, Z_{33})$



$e^{(-\infty)} \rightarrow e^{(-\text{large Number})}$

0.00001

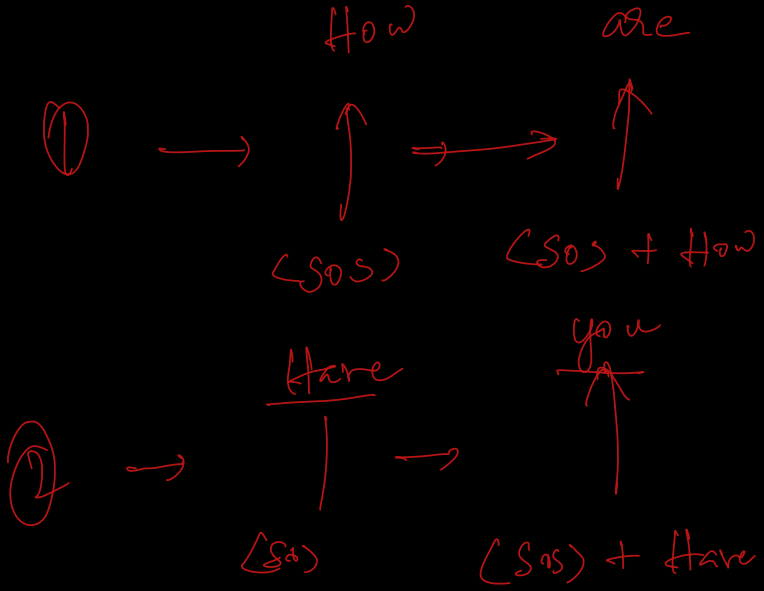


you what do

1 Billions

How are you?

Have you eaten?

Hey good morning

Whats app ?