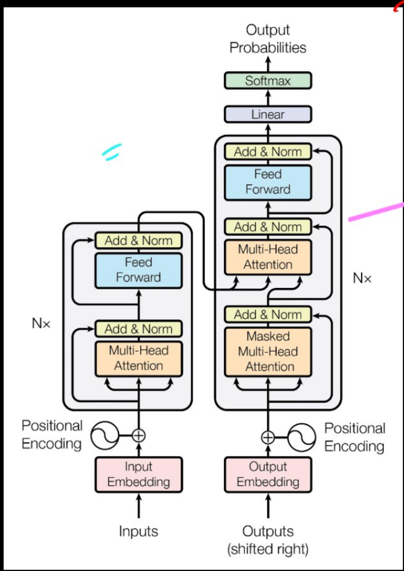
1. Encoder → Self attention → multihead self attention
2. Entire flow of encoder
3. Positional encoding
4. Residual connection
5. Importance of NN in encoder
6. masked self attention } decoder
7. cross attention
8. Entire flow of decoder → training
9. Entire flow of inferencing → Prediction

training
↓
(translation prob) O/P =
→ We are friend | (हम दोस्त हैं!)
masking



decoder
{non autoregressive training}
{Parallel training}
→ training

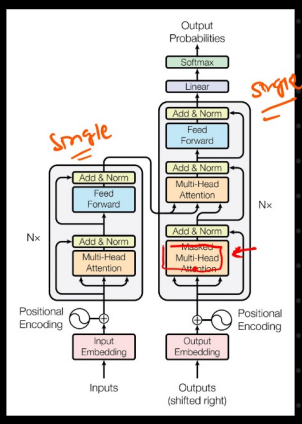
we are friend
↑
decoder
हम दोस्त हैं!

Infrencing (Prediction)
↳ Autoregressive (sequentially)
↳ sequentially

My name is Sunny → transformer → ? translate

Google (transformer) ←
Bert → encoder Part
NLM
(M)
1. text classification
2. Summarization
OpenAI → GPT → ... 3.5, 4
2. text generation

Google → T5 } → translation
Meta → BART



single

single

decoder

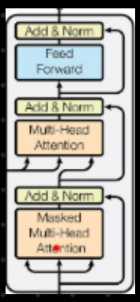
encoder

6 encoder

6 decoder

hyperparameter

2, 3, 4
10, 15
6, 8



Encoder

ANN

Self attention

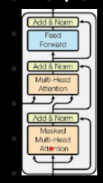
decoder

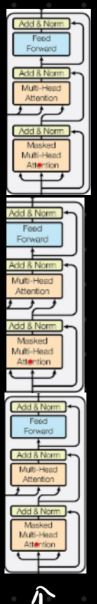
ANN

cross attention

mask self Attention

6decoder





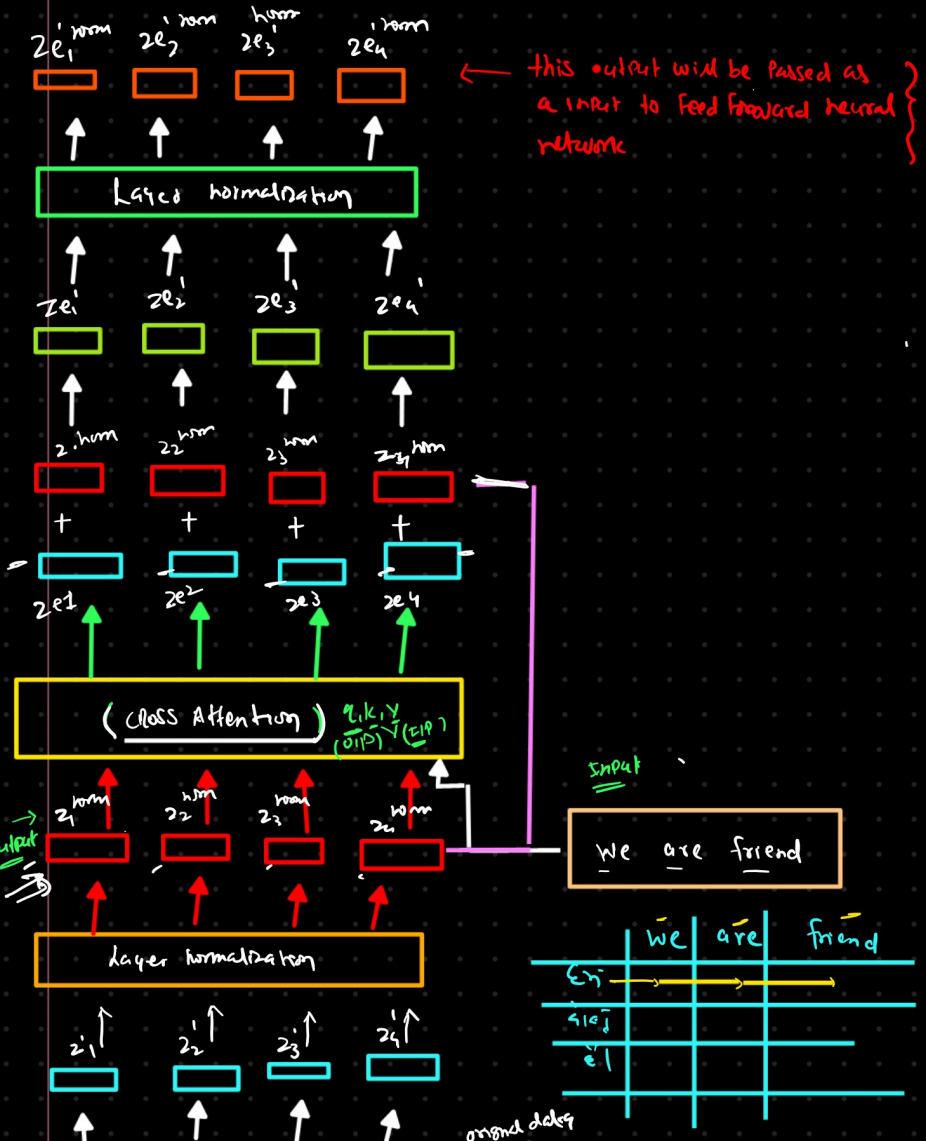
We are friend

512

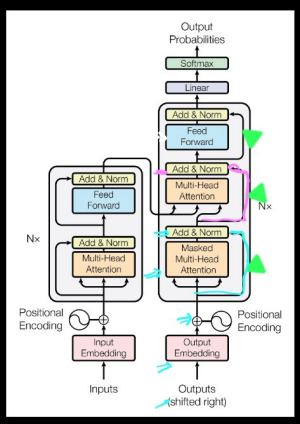
context

CE

We are friend | હમ દોસ્ત હૈ!



this output will be passed as a input to Feed forward neural network



હમ =

original data

{complete info}

Contextual vector

Skip connection

residual

← Masked self Attention (Masking) હમ દોસ્ત હૈ!

masking

input interaction matrix

હમ → $W_{11} \times$ હમ + W_{12} દોસ્ત + W_{13} હૈ

(Predict) (training)

(present token)

Future token X

← PE layer (seq of data) sine cosine

(512)

← word2vec (512)

Loss = Actual - Prediction

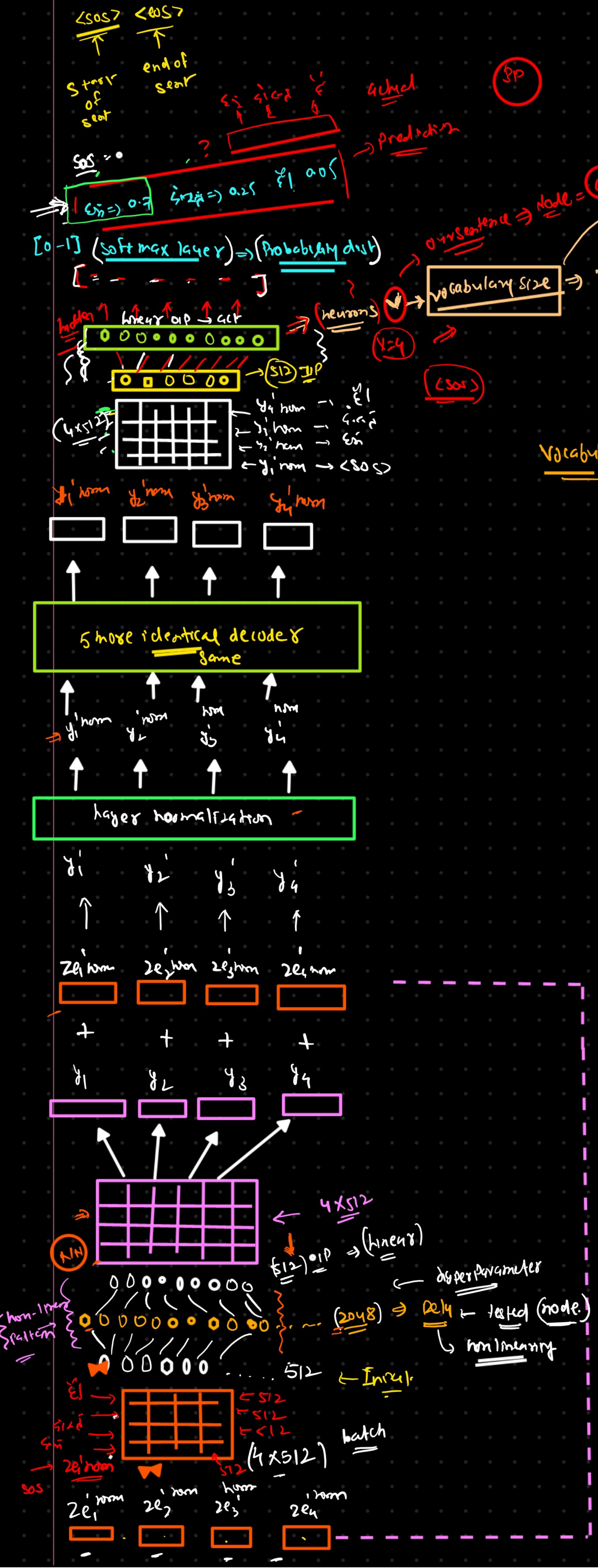
Loss ↓ training → bp

Pred.

we are enemy :) ?

↳ generate

(seq)



o/p sentence => node = 4

vocabulary size => vocabulary of output data

↳ o/p colums

yes

we are friends

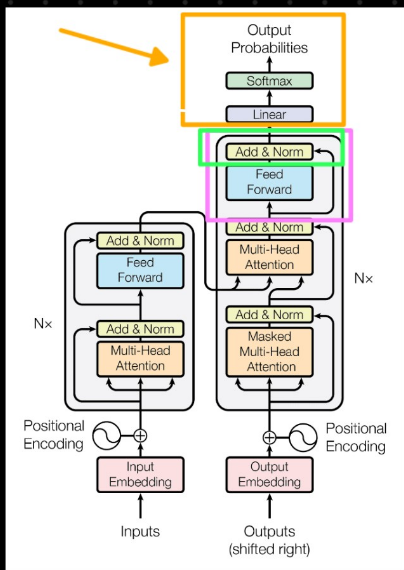
Vocabulary => mean => {unique words in the given data} ?

I/P <=> hidden (matrix multiplication)

4x512 512x2048

=> (4x2048)

ANN matrix



→ word2vec

En [- - - - - -]

[- - - - - -] Actual

○ = loss ← ↓↓↓↓

training

[- - - - - -] Prediction

(Softmax layer)

[- - - - - -]

= linear operation → ANN

⇒ fix → voc of output column

[illegible] dynamic

(V=4)

Eng hindi

<SOS> En [illegible]

V = 4

⇒ (batch)

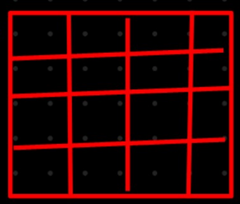
(Parallel)

↓

matrix

. . . . ⇒ 512

512 →



(trained)

I/P	O/P
how are you	[illegible]
what you doing	[illegible]
what your name	[illegible]

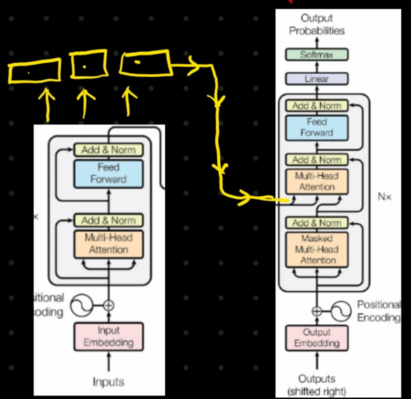
⇒ transformer

New sentence is coming

(We are friends) → OP ?

⇒ Prediction

En



We are friend

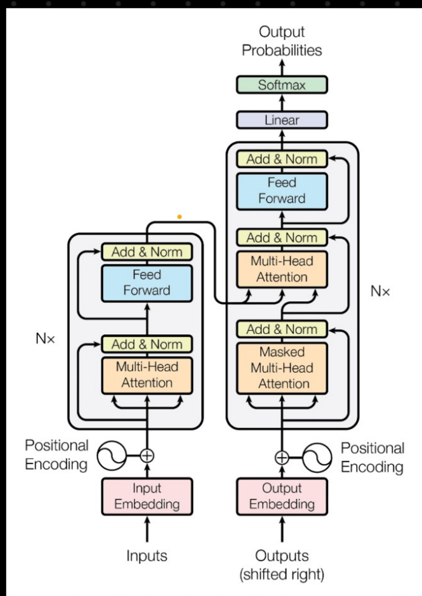
<SOS>

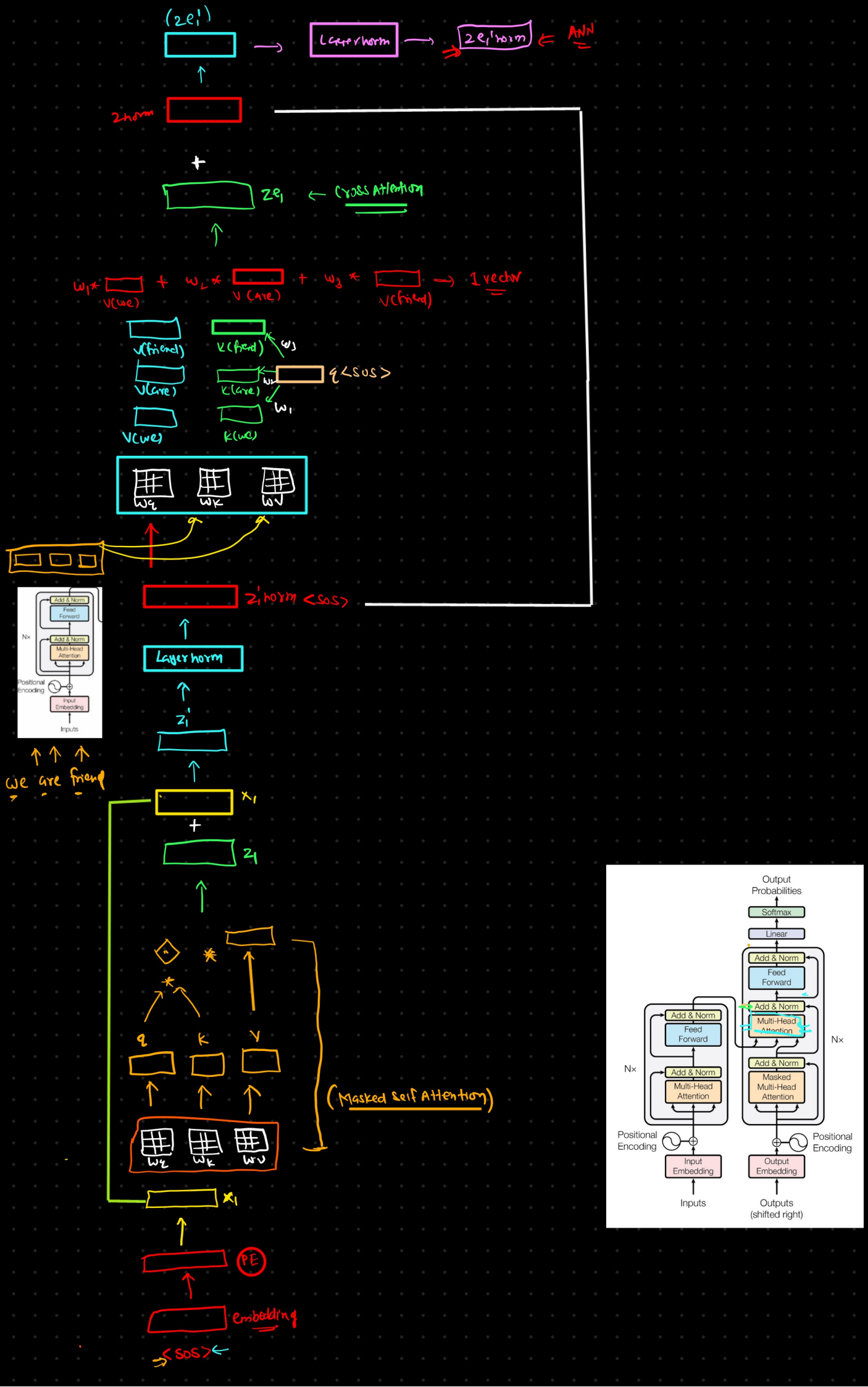
→ training

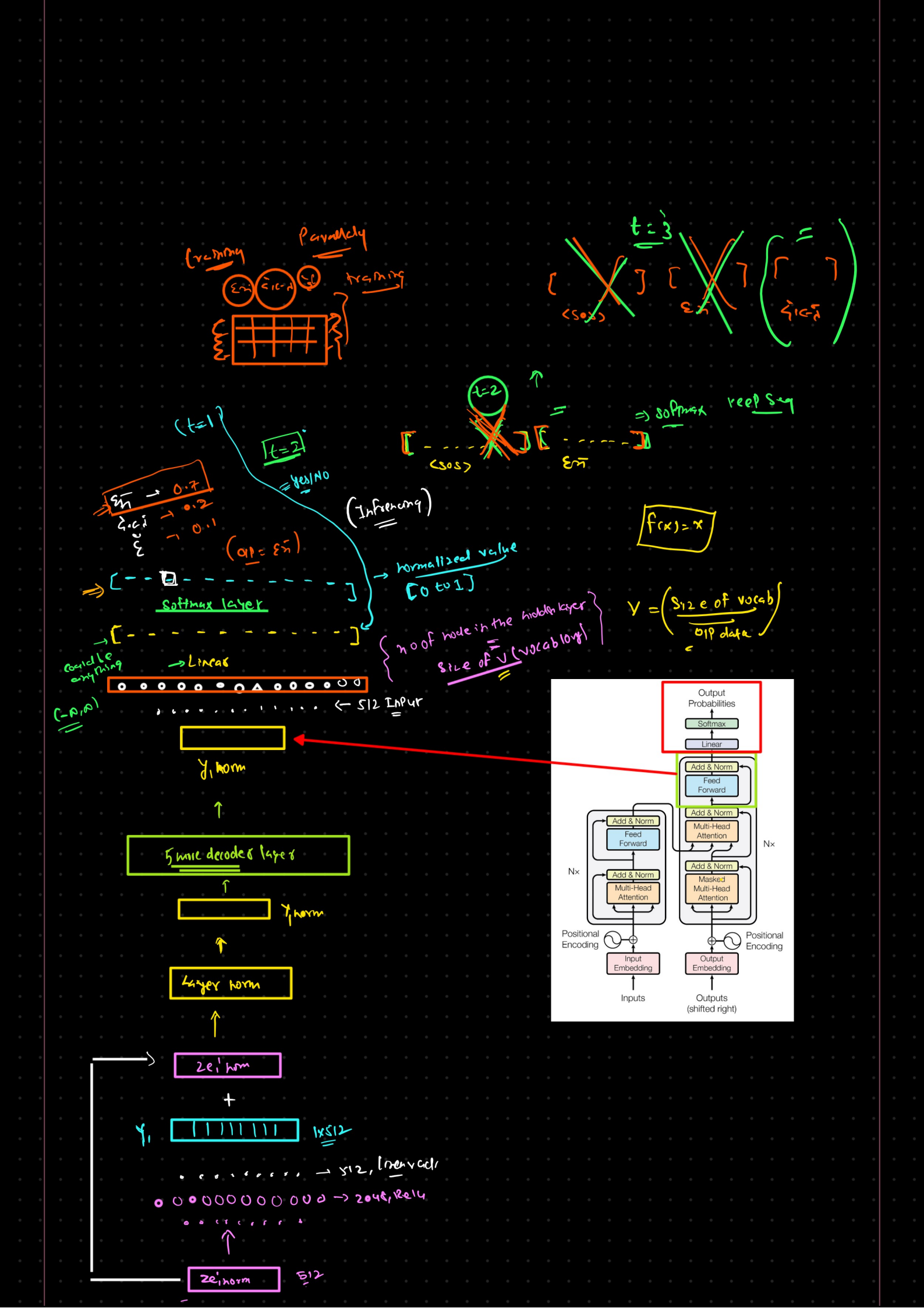
→ Prediction

⇒ ML, DL

OP ?




(ze1')
LayerNorm
ze1'norm
ANN
Znorm
+
ze1
Cross Attention
w1* V(we) + w2* V(are) + w3 * V(friend) → 1 vector
V(friend)
K(friend)
w3
V(are)
K(are)
q<SOS>
w2
w1
V(we)
K(we)
WQ
WK
WV
z1'norm <SOS>
LayerNorm
z1'
we are friend
x1
+
z1
q
k
v
(Masked Self Attention)
x1
PE
embedding
<SOS>
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Positional Encoding
Input Embedding
Inputs
Nx
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Output Embedding
Outputs (shifted right)


Parallely
training
(running
t=3
<sos>
t=2
softmax
keep
(t=1)
t=2
Yes/No
(Inferencing)
0.7
0.2
0.1
(all = ξn)
normalized value
[0 to 1]
f(x) = x
Softmax layer
no of node in the hidden layer
size of V (vocablory)
V = (size of vocab / o/p data)
Linear
← 512 Input
Yi norm
5 nos decoder layers
Layer norm
Zei norm
+
1x512
→ 512, Linear
→ 2048, Relu
Zei norm
512
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Multi-Head Attention
Masked Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

transformer decoder = (Parallely)

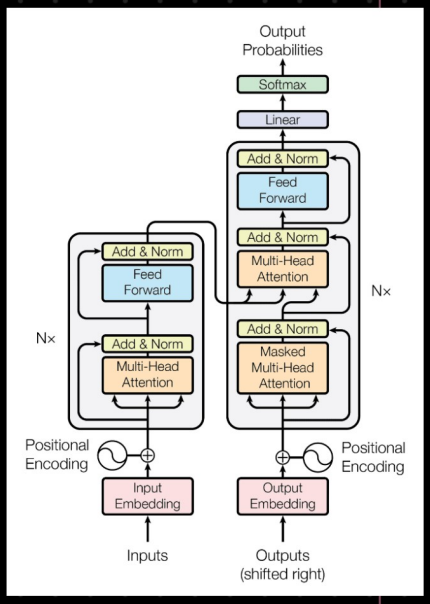
↳ it is a non autoregressive at the time of training

↳ It is autoregressive at the time of prediction

Input: We are friends

O/P: (<sos> ...)

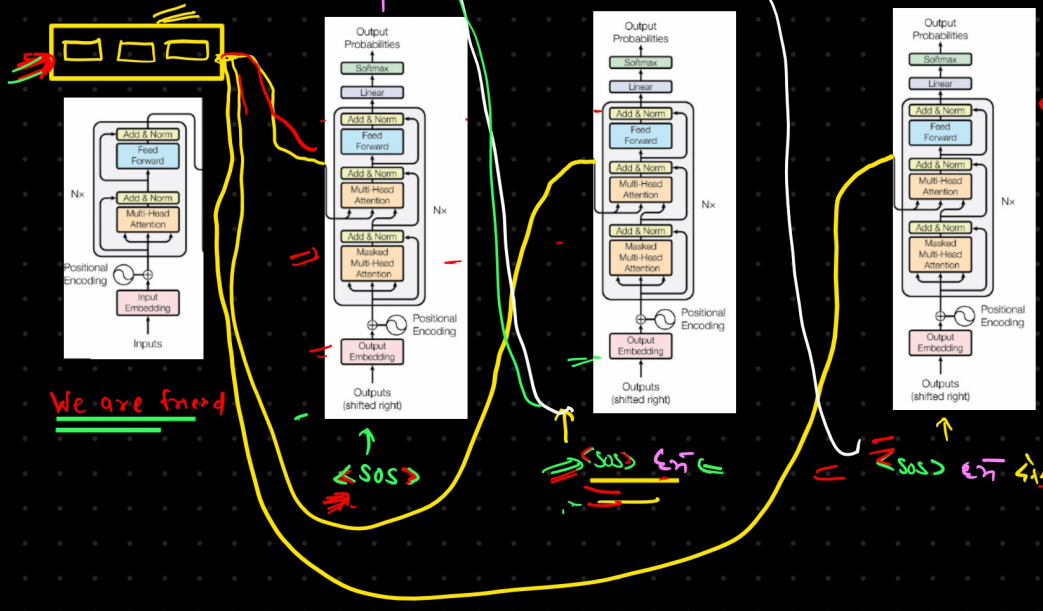
y_{pred}



<eos>

Output seq

t=1 t=2 t=3



We are friend

<sos>

6

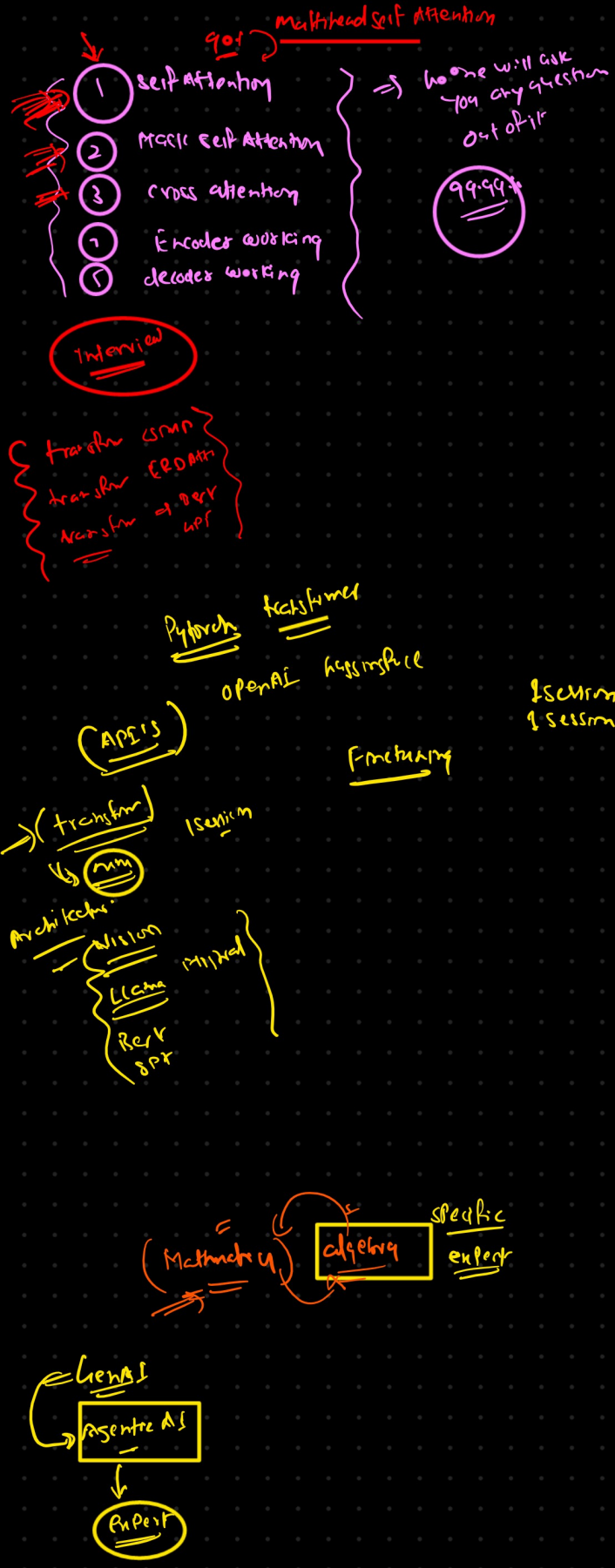
8

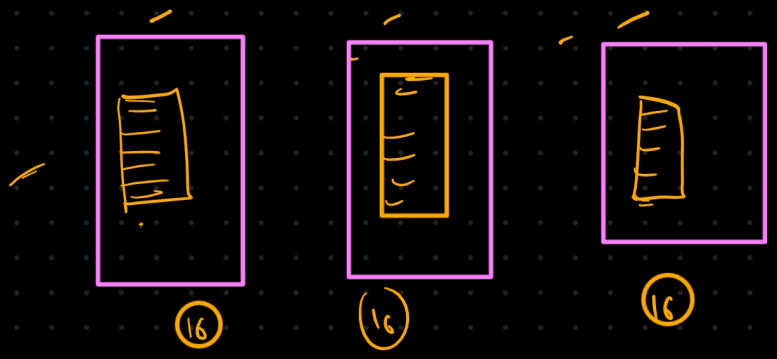
9

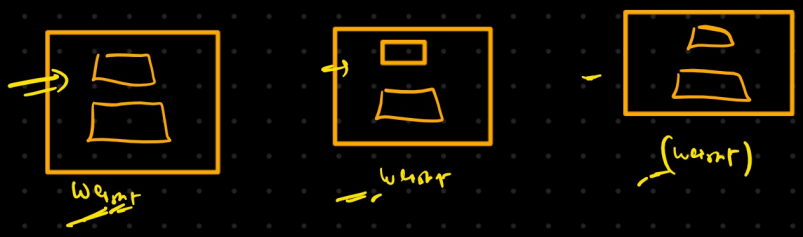
Masking => Data shift

at the training => masking

inferencing


90%
Multihead self Attention
1 Self Attention
2 Masked self Attention
3 cross attention
4 Encoder working
5 decoder working
=> how one will ask you any question out of it
99.99%
Interview
transformer scratch
transformer BERT Arch
transformer => GPT LLM
Pytorch transformer
OpenAI huggingface
1 Session Pytorch
1 Session APIs
1 session → transformer based model
APIs
Fine tuning
4 session → Fine tuning
{4-5}
8 → 4
transformer
1 session
many
Architecture
Mixed
Vision
Llama
Bert GPT
{6-8
Project/Coding
{Resume/ISO interview question
Mathematics
coding
specific expert
GenAI
Agentic AI
Expert





Generative Pretrain transformer