الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

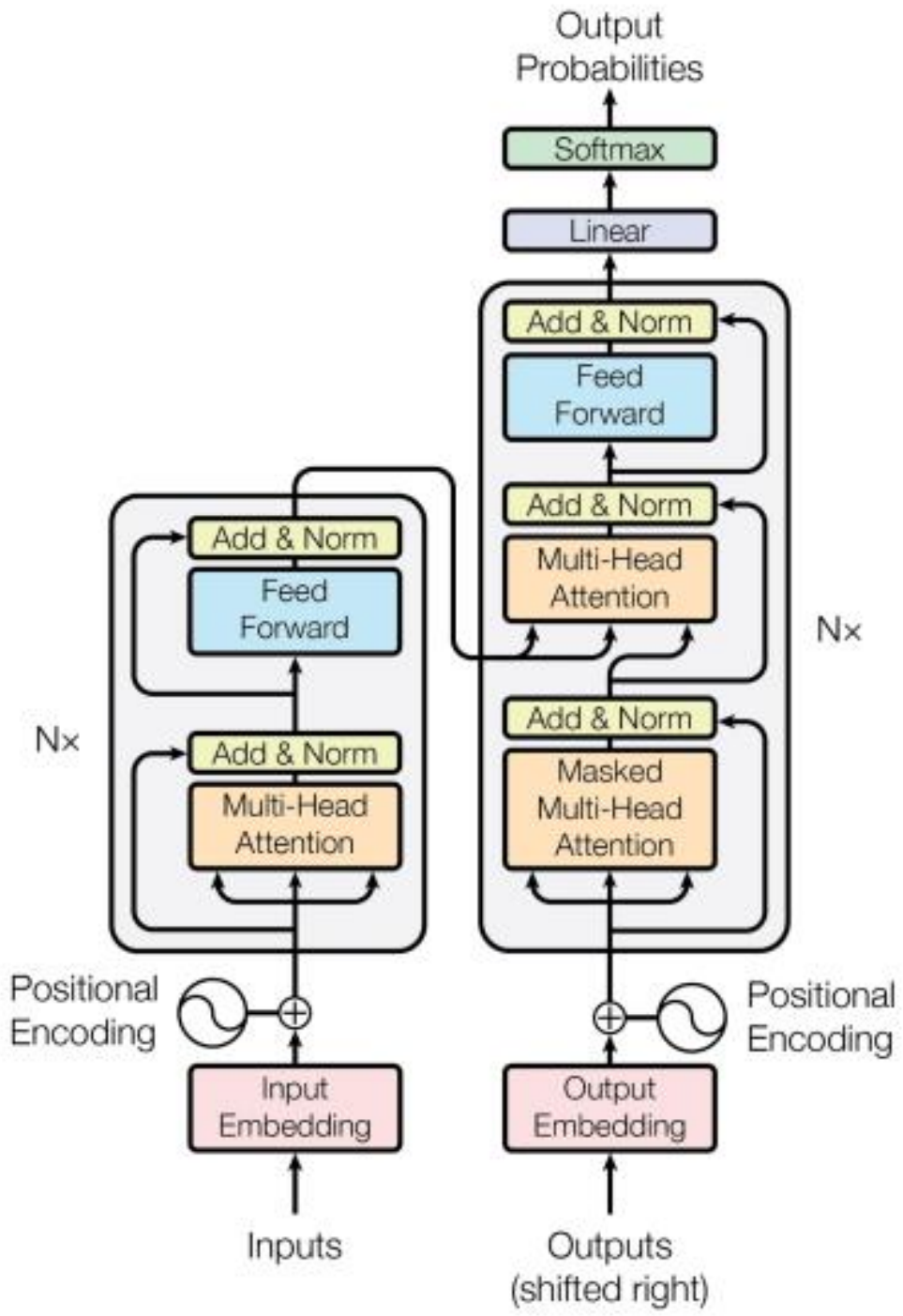
الذكاء الصنعي العملي	كلية الهندسة	المحاضرة 4

Transformers: Decoder Part

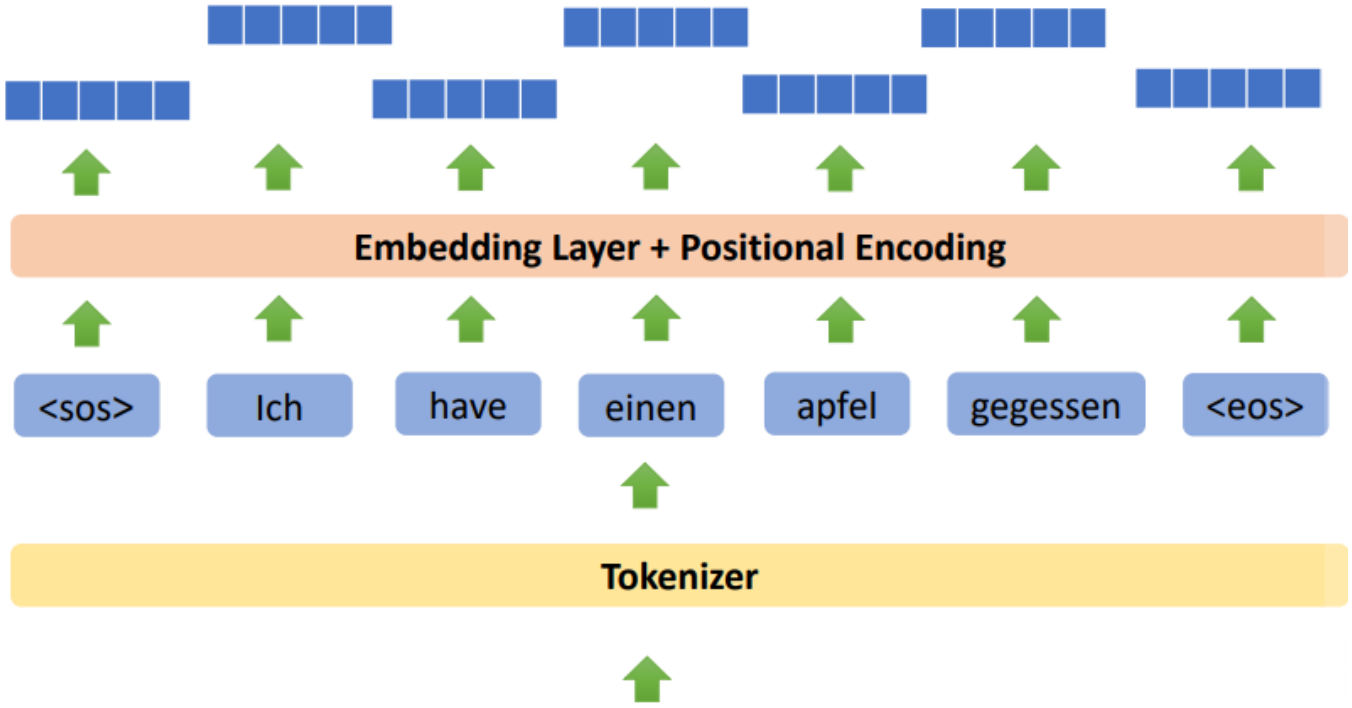
د. رياض سنبل

Transformers

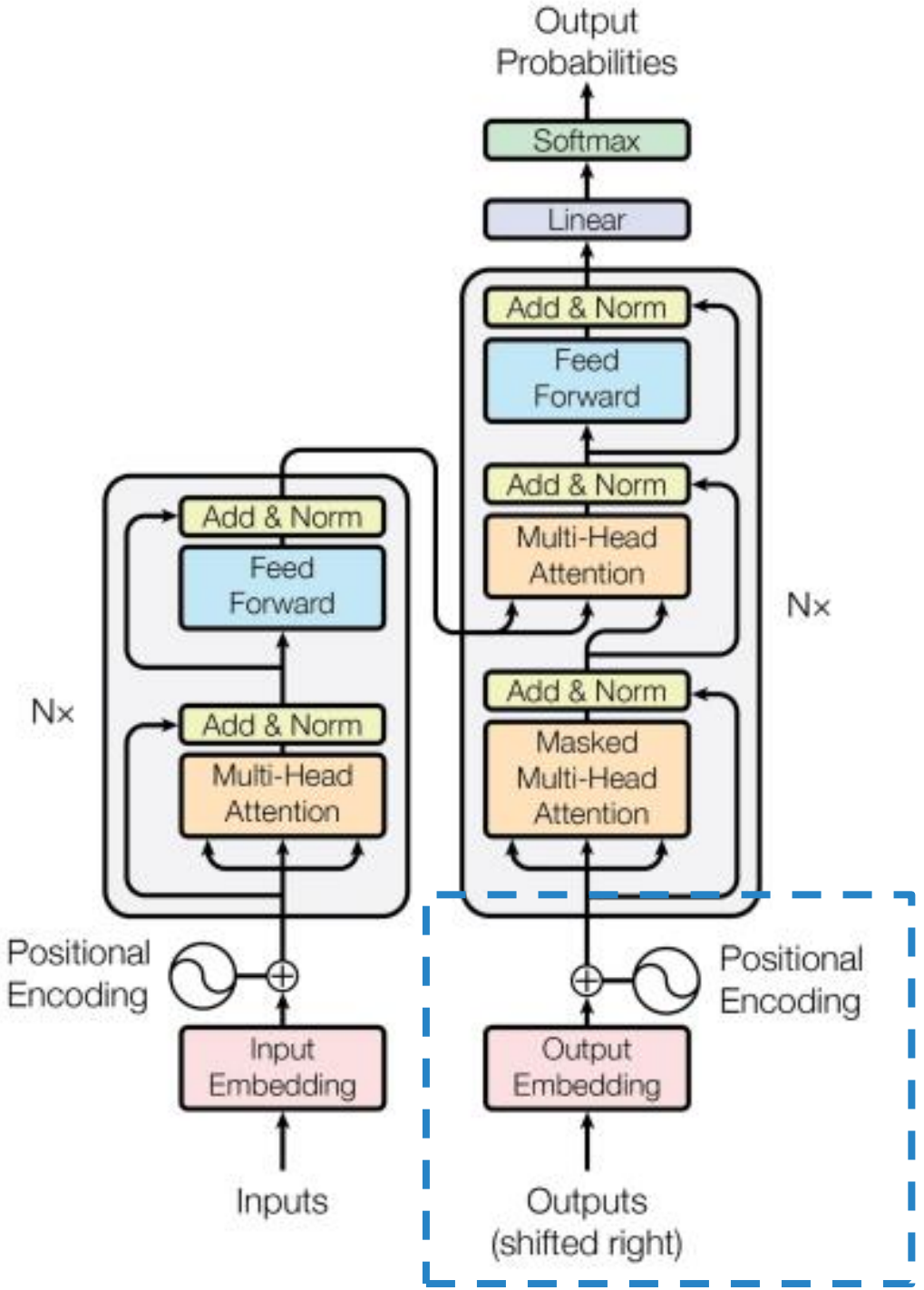
- Transformers are a type of neural network architecture that transforms or changes an input sequence into an output sequence.
- They do this by learning context and tracking relationships between sequence components.
- And break the problem into two parts:
  - An encoder (e.g., Bert)
  - A decoder (e.g., GPT)



Output Embedding



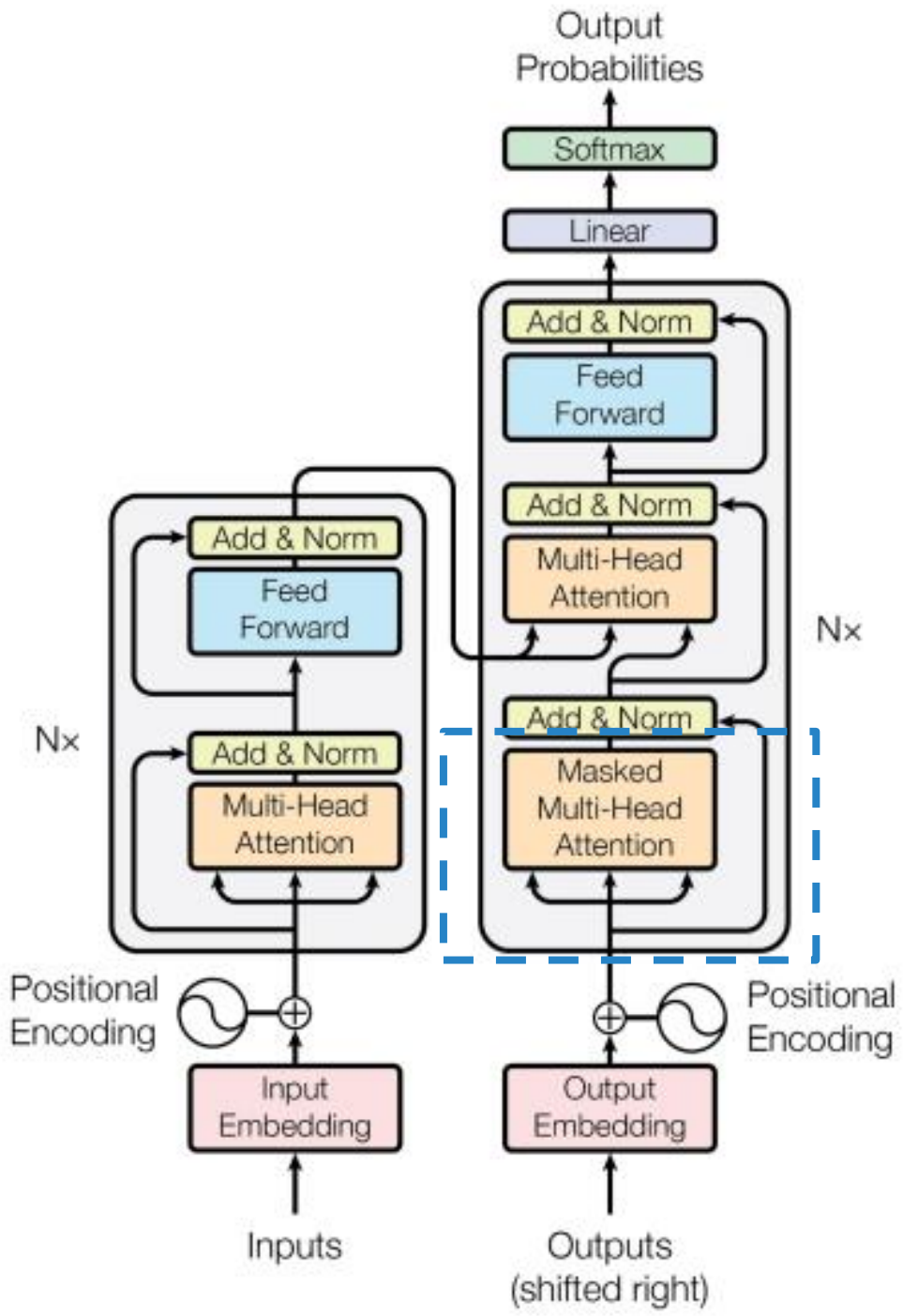
Generate Target Emebeddings



Masked Attention

1	<sos>	Ich	have	einen	apfel	gegessen	<eos>
2	<sos>	Ich	have	einen	apfel	gegessen	<eos>
3	<sos>	Ich	have	einen	apfel	gegessen	<eos>
4	<sos>	Ich	have	einen	apfel	gegessen	<eos>
5	<sos>	Ich	have	einen	apfel	gegessen	<eos>
6	<sos>	Ich	have	einen	apfel	gegessen	<eos>
7	<sos>	Ich	have	einen	apfel	gegessen	<eos>

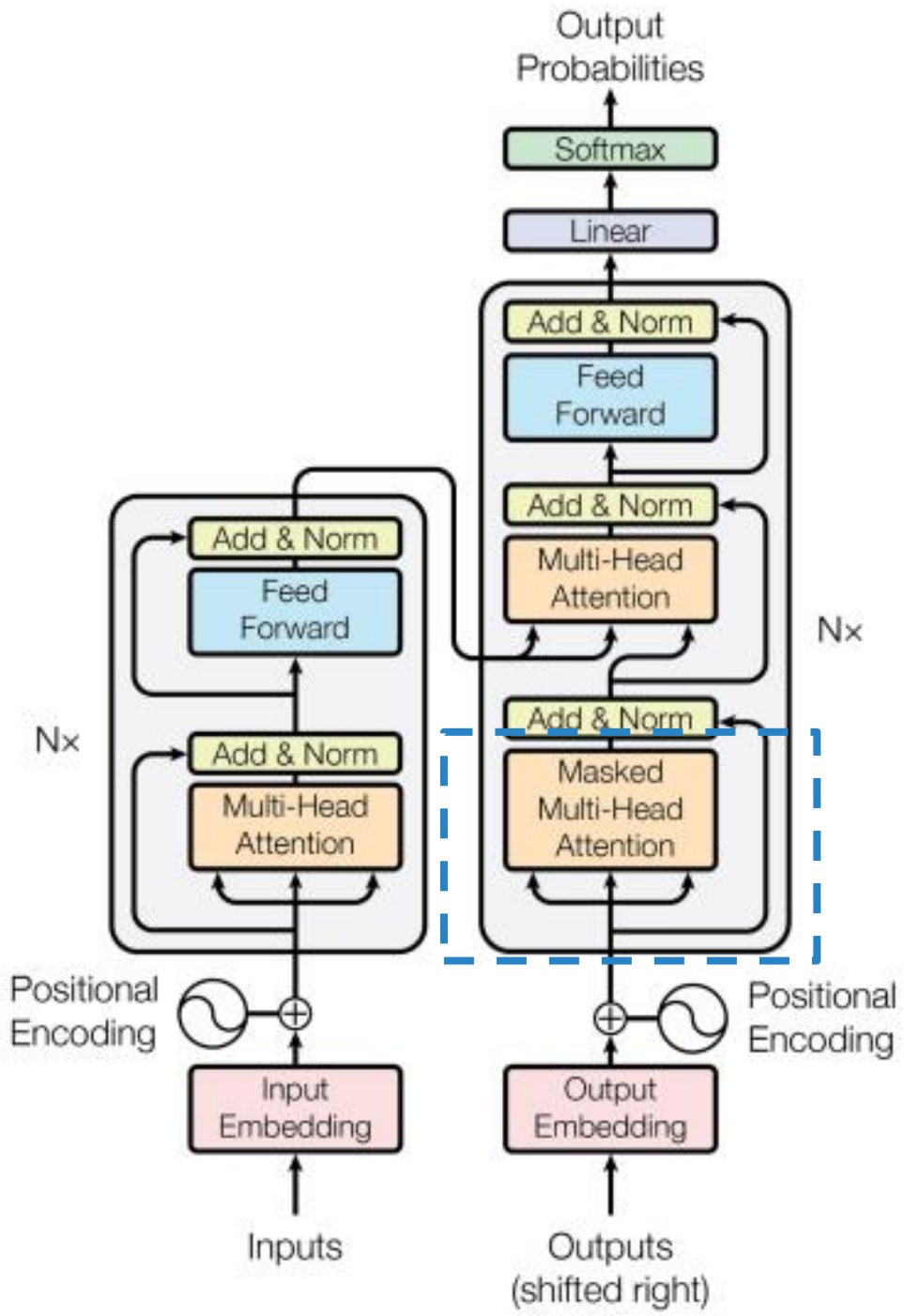
Mask the available attention values ?



Masked Attention

1	<sos>	- ∞	- ∞	- ∞	- ∞	- ∞	- ∞
2	<sos>	Ich	- ∞	- ∞	- ∞	- ∞	- ∞
3	<sos>	Ich	have	- ∞	- ∞	- ∞	- ∞
4	<sos>	Ich	have	einen	- ∞	- ∞	- ∞
5	<sos>	Ich	have	einen	apfel	- ∞	- ∞
6	<sos>	Ich	have	einen	apfel	gegessen	- ∞
7	<sos>	Ich	have	einen	apfel	gegessen	<eos>

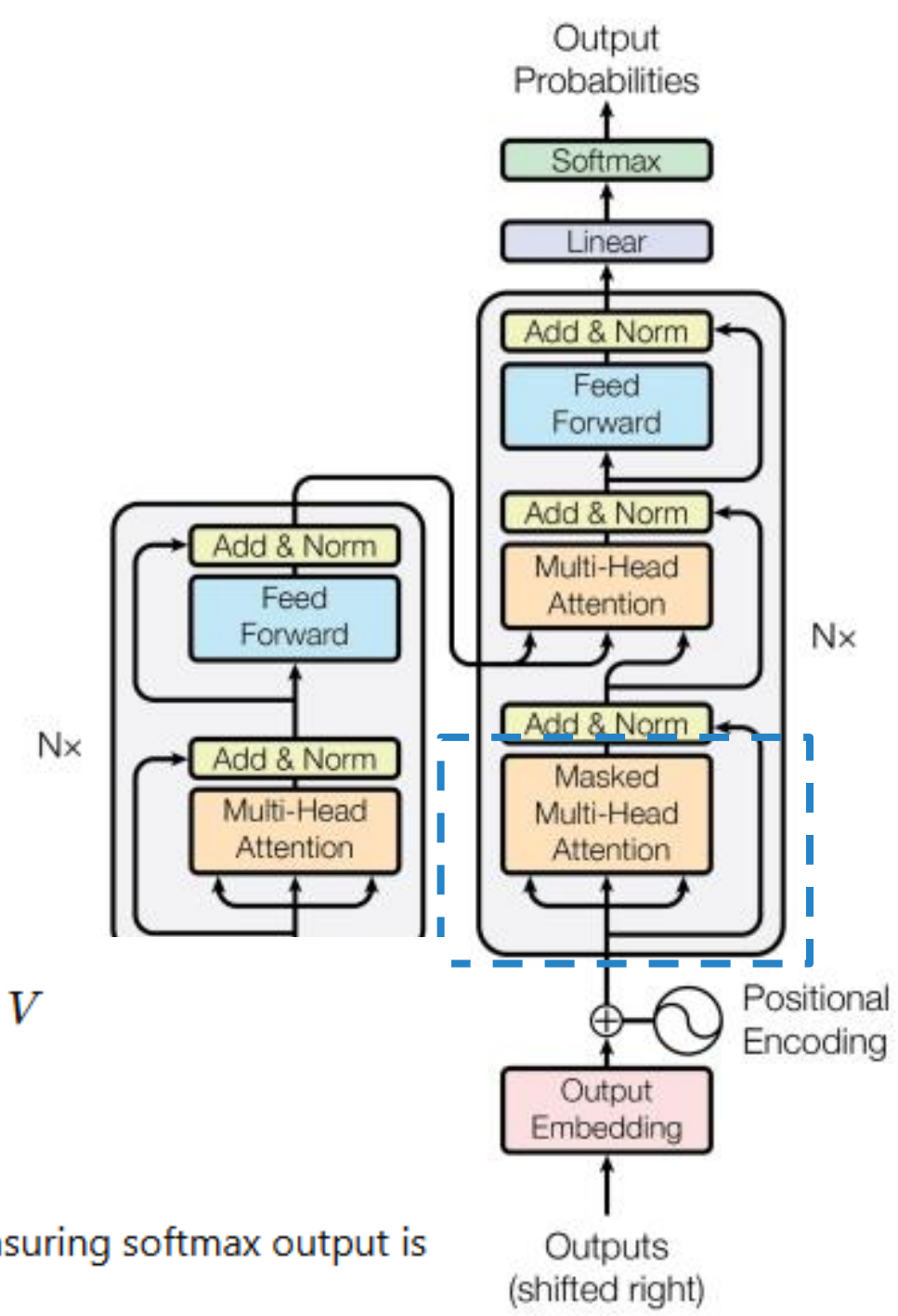
Softmax -> - ∞ -> 0



Masked Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V$$

- $Q, K, V \in \mathbb{R}^{T\times d_k}$: represent queries, keys, and values.
- M: a mask matrix with $-\infty$ in positions corresponding to future tokens, ensuring softmax output is zero for those.



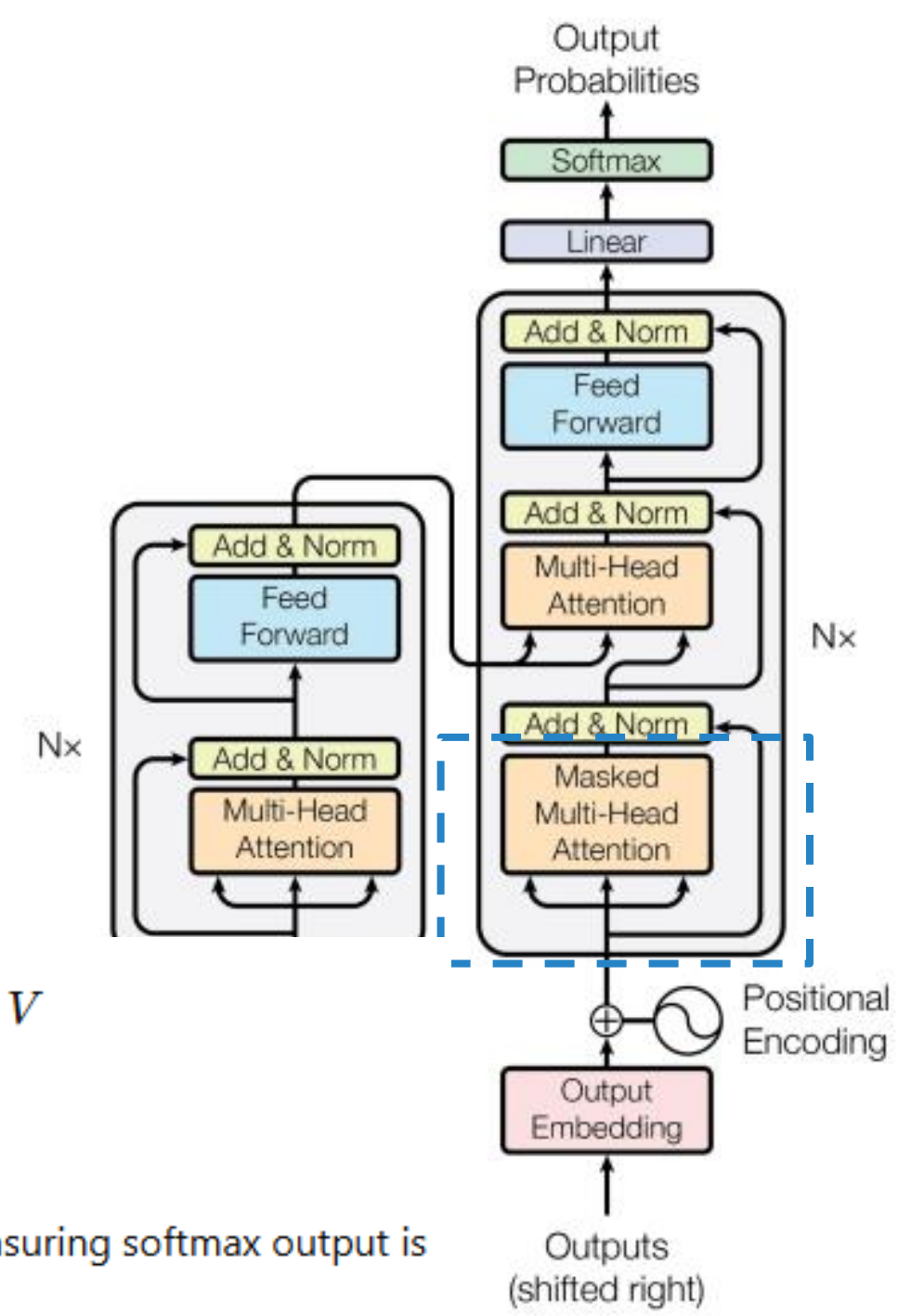
Masked Attention

$R^{T \times T}$ $R^{T \times d_h}$

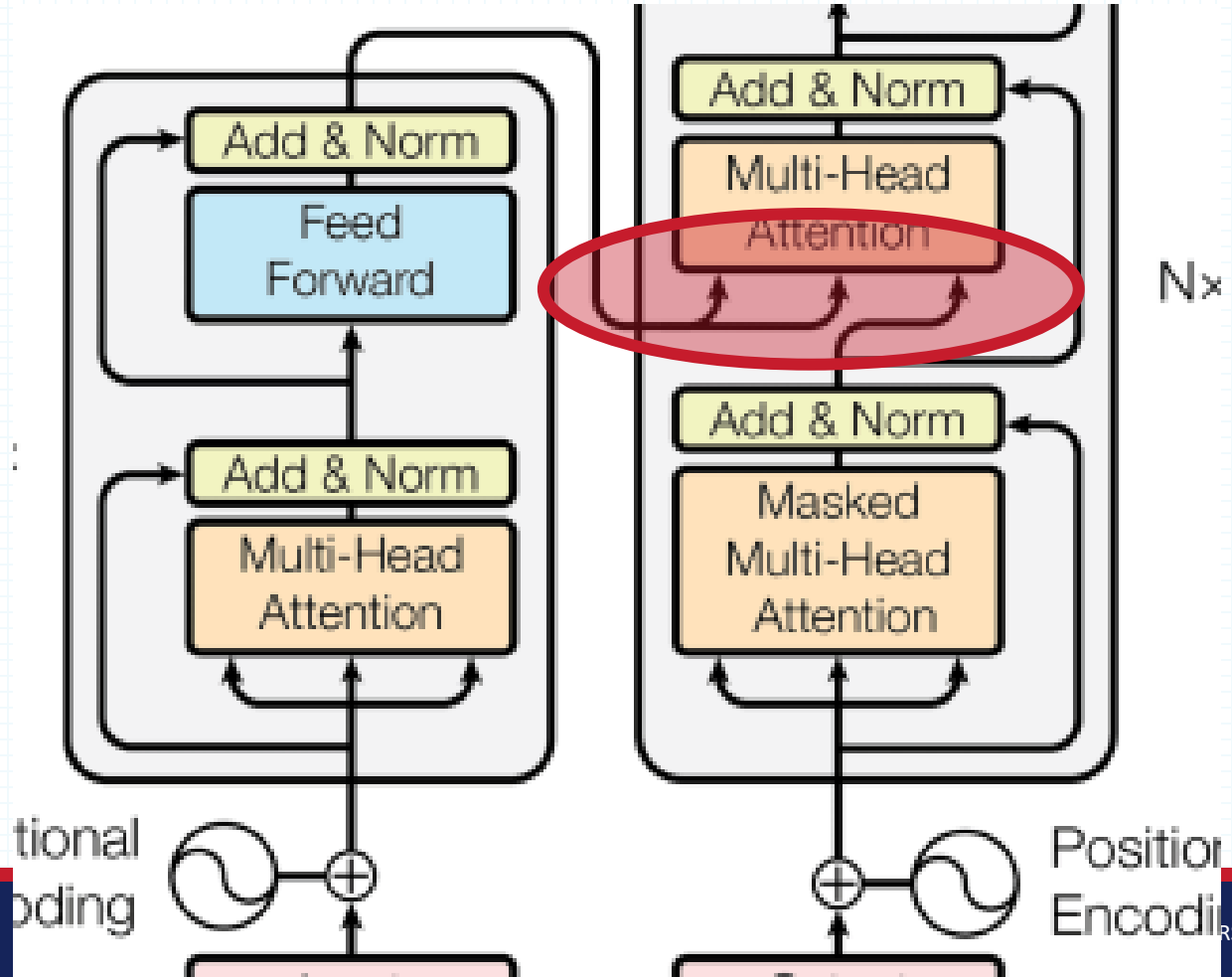
Masked Attention Values

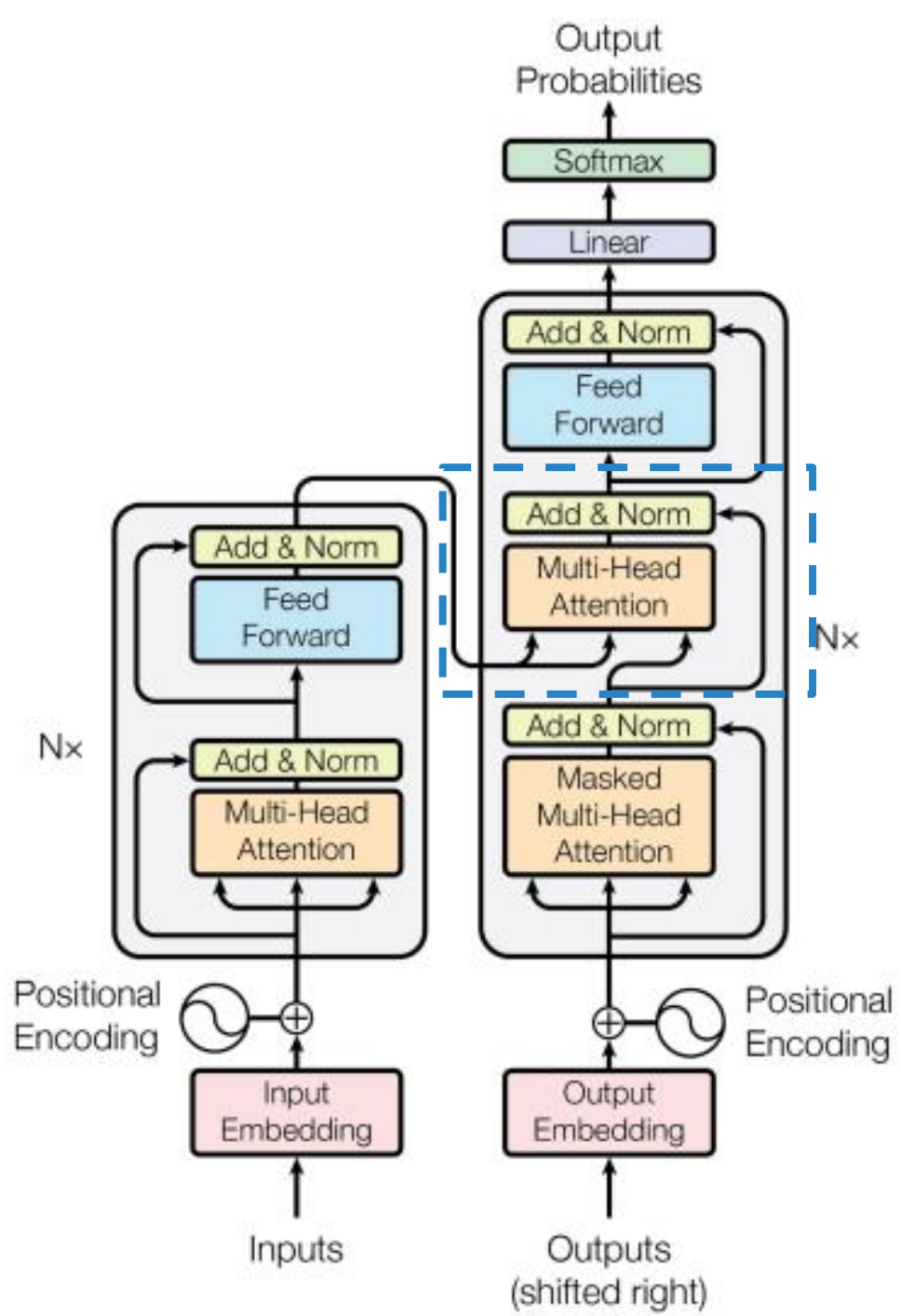
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V$$

- $Q, K, V \in \mathbb{R}^{T \times d_k}$: represent queries, keys, and values.
- M: a mask matrix with $-\infty$ in positions corresponding to future tokens, ensuring softmax output is zero for those.



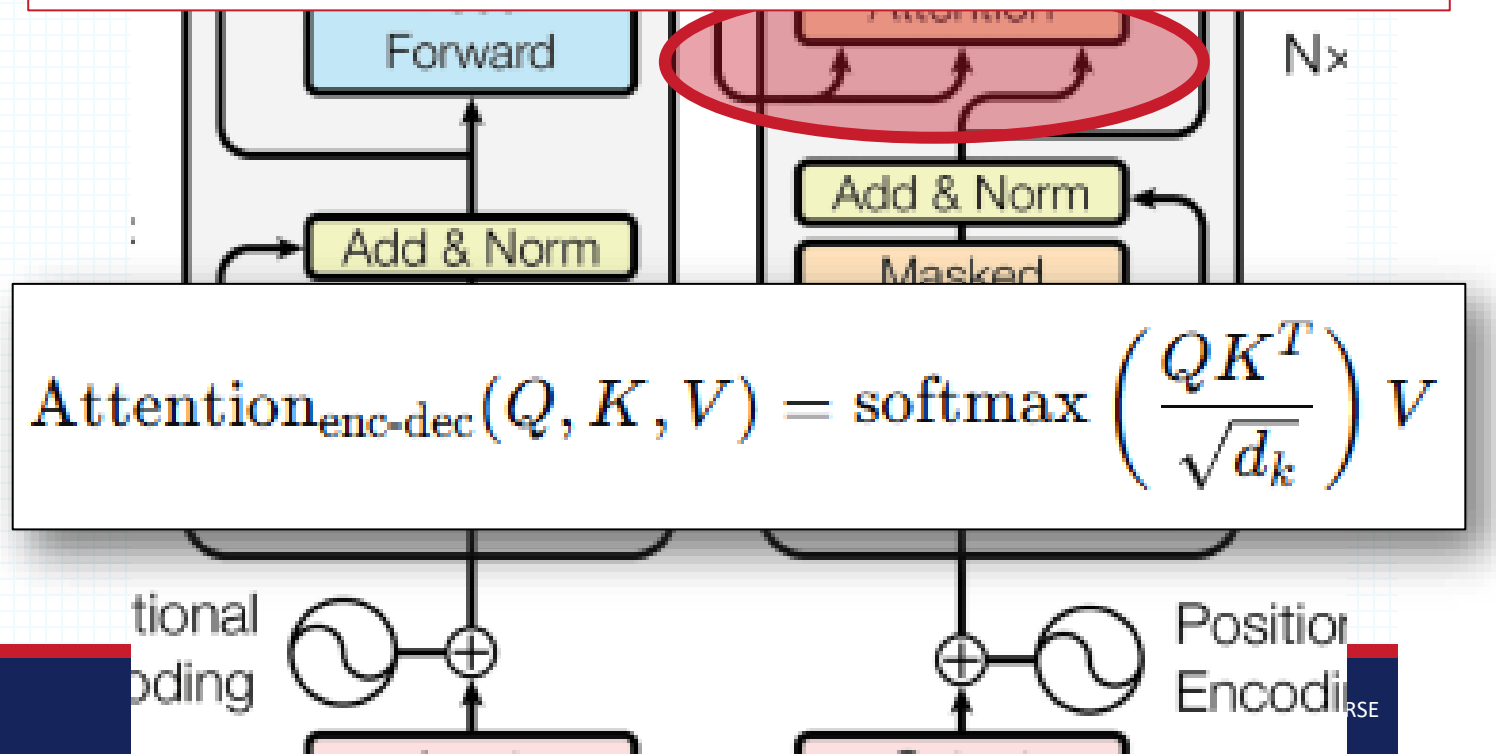
Encoder-Decoder Attention (Cross-Attention)



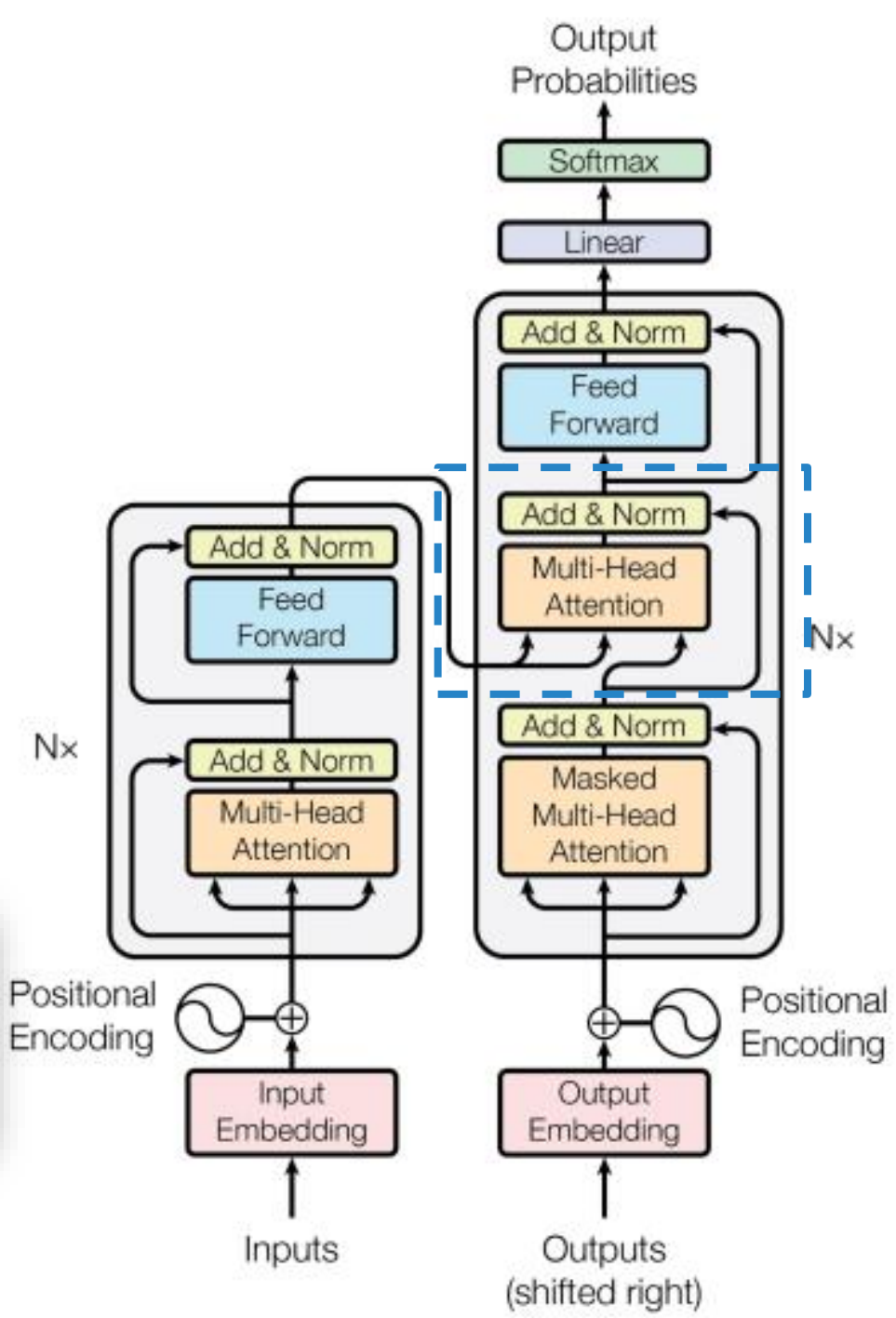


Encoder-Decoder Attention (Cross-Attention)

The outputs of the **Encoder** act as Keys and Values, and the output from the **Decoder's** self-attention acts as Queries.



$$\text{Attention}_{\text{enc-dec}}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

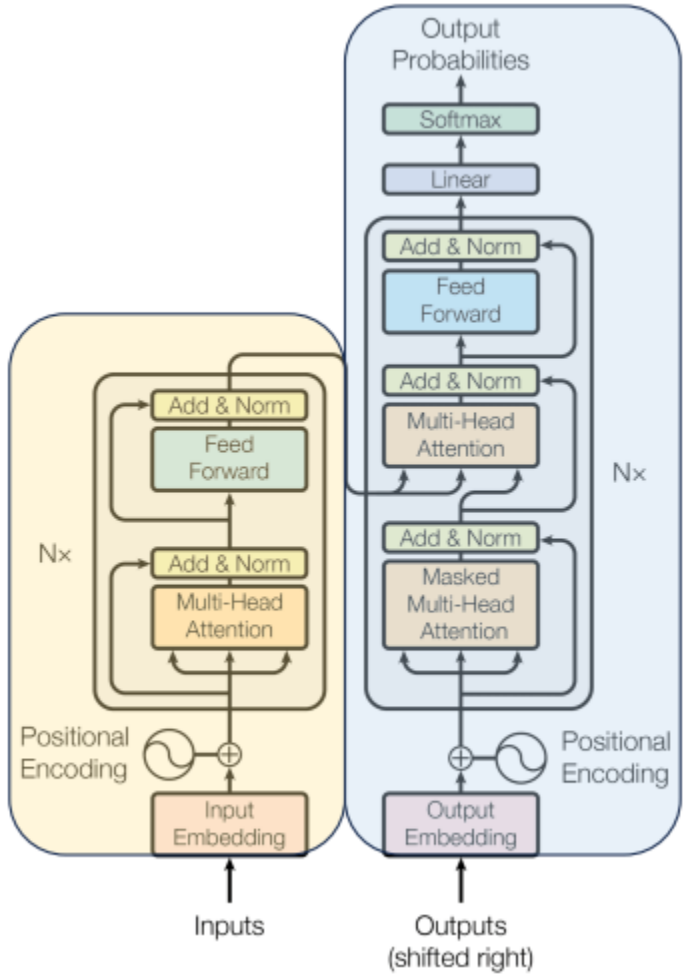


Transformers

- a Transformer architecture can be designed in **three different ways** depending on the task and objective:
- 1. **Encoder-Only Transformers**
- **2. Decoder-Only Transformers**
- 3. **Encoder-Decoder Transformers (Seq2Seq)**

BERT
Oct 2018

Representation



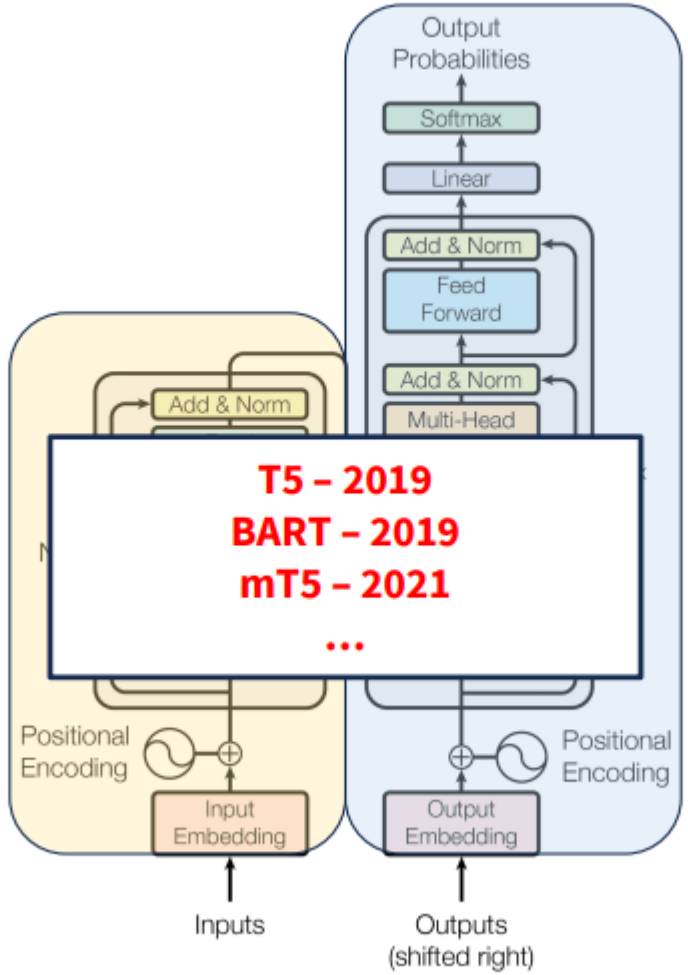
GPT
Jun 2018

Generation

Transformers… The LLM Era

BERT – 2018
DistilBERT – 2019
RoBERTa – 2019
ALBERT – 2019
ELECTRA – 2020
DeBERTa – 2020
…

Representation



GPT – 2018
GPT-2 – 2019
GPT-3 – 2020
GPT-Neo – 2021
GPT-3.5 (ChatGPT) – 2022
LLaMA – 2023
GPT-4 – 2023
…

Generation