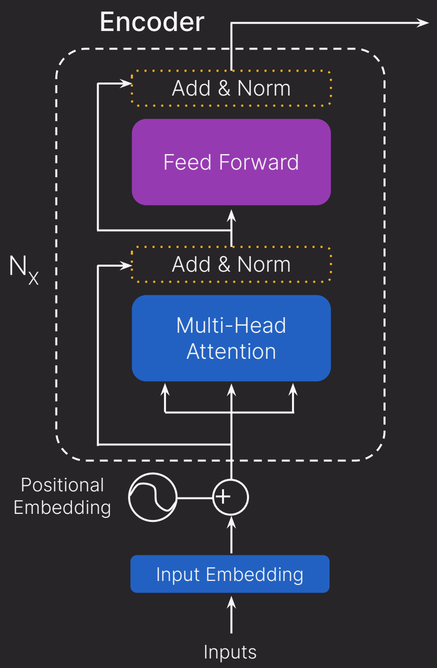
Recap



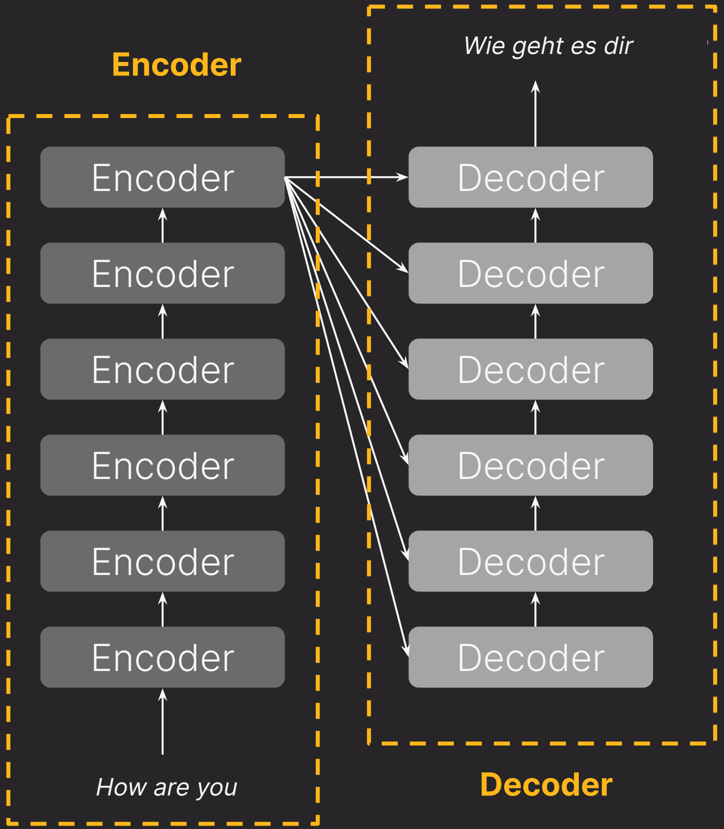
Encoder
Add & Norm
Feed Forward
N_x
Add & Norm
Multi-Head Attention
Positional Embedding
Input Embedding
Inputs

Attention Mechanism and Transformers

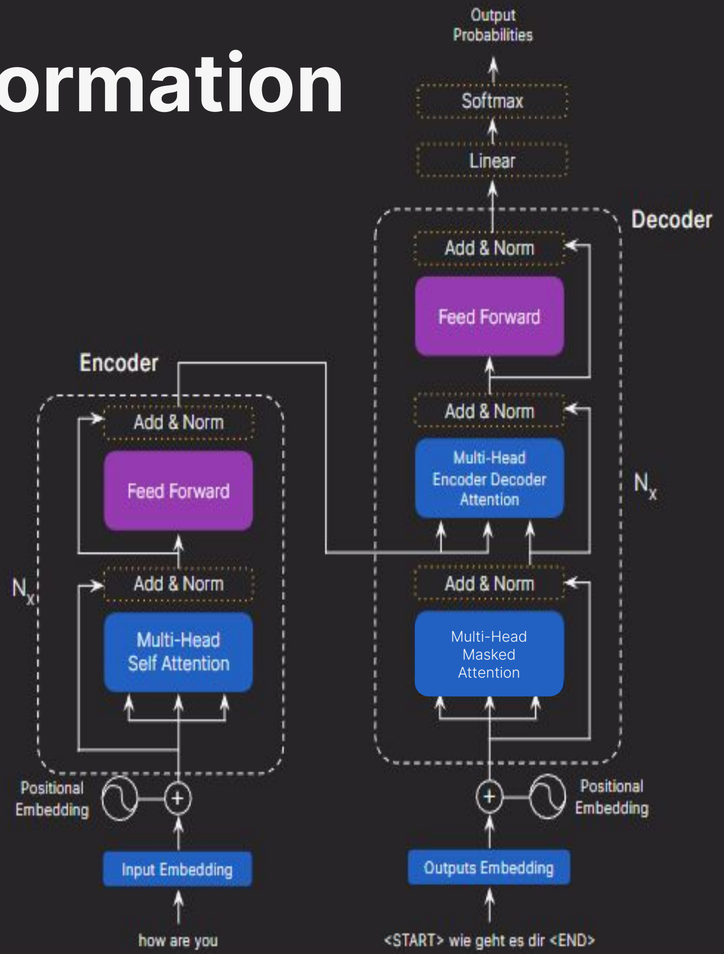
Video 8 : Flow of information in Transformers

Analytics Vidhya

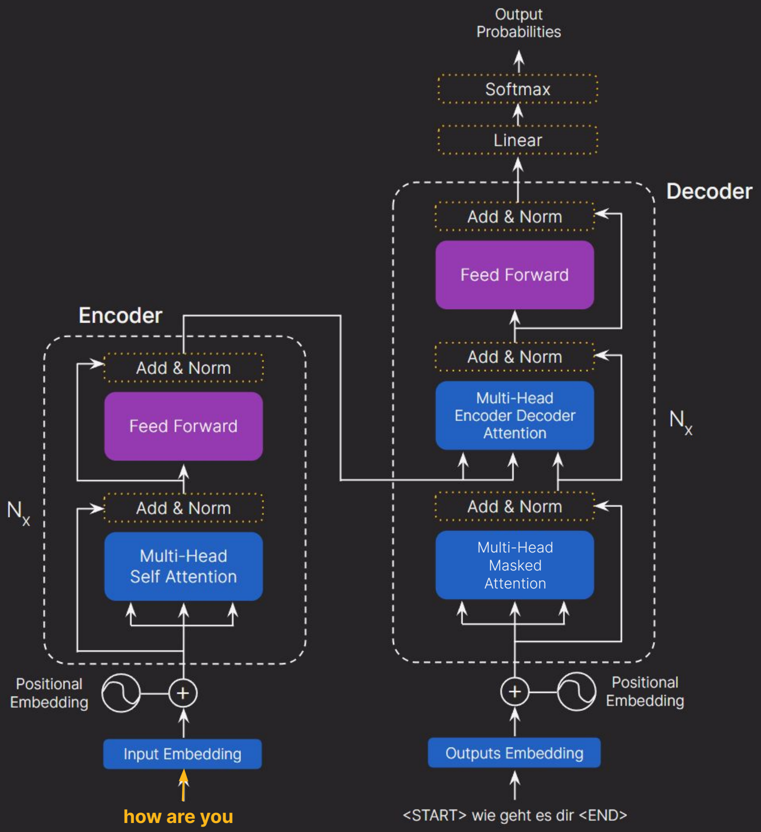
Transformers



The Flow of Information

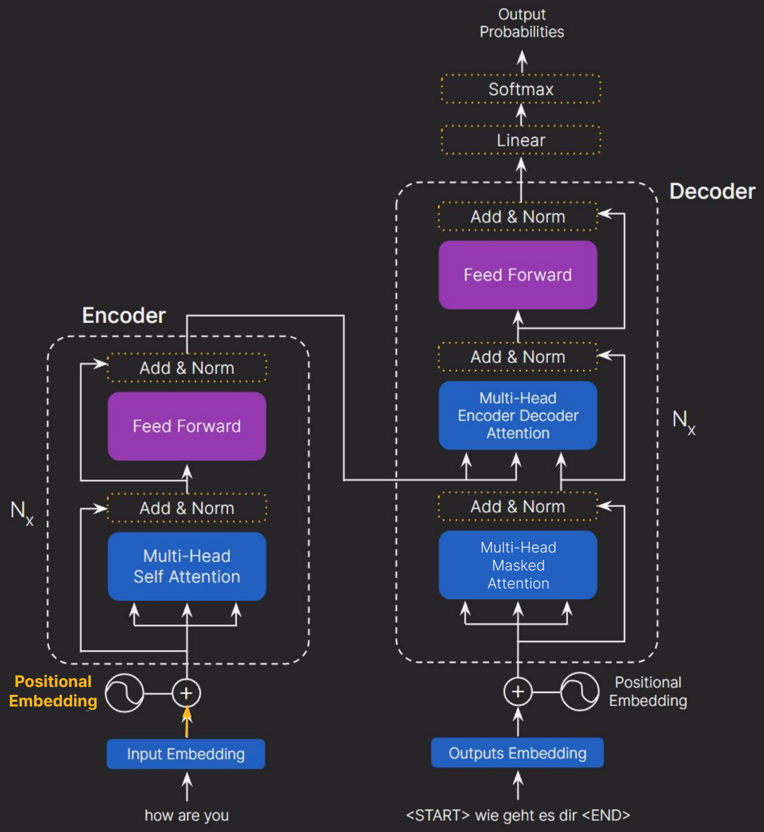


The Flow of Information



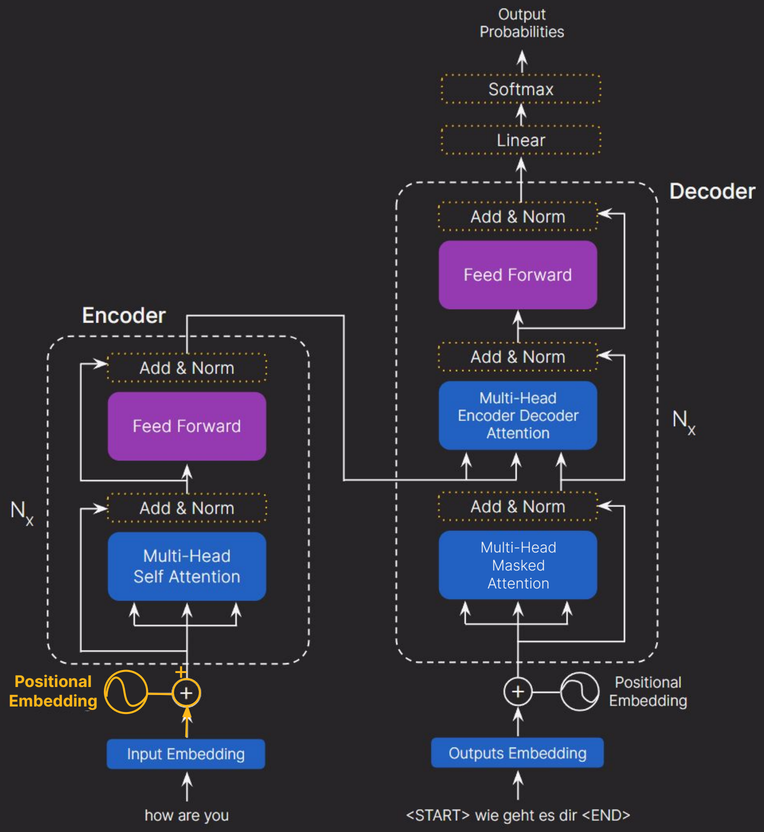
- Word embeddings of tokens goes as input in encoder.

The Flow of Information



- The transformers use **positional embedding** to capture sequential information of the tokens.

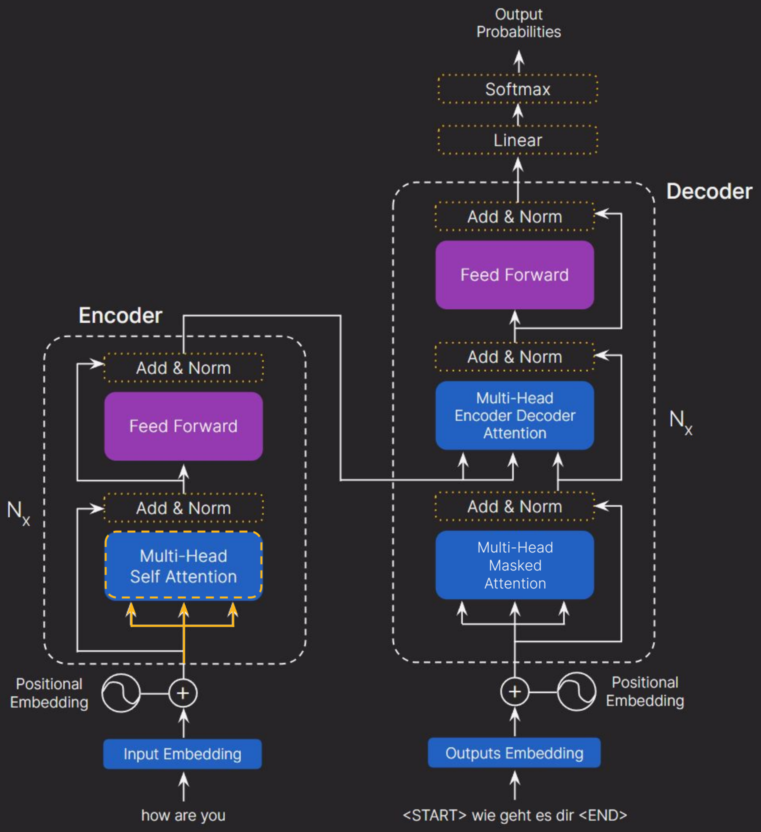
The Flow of Information



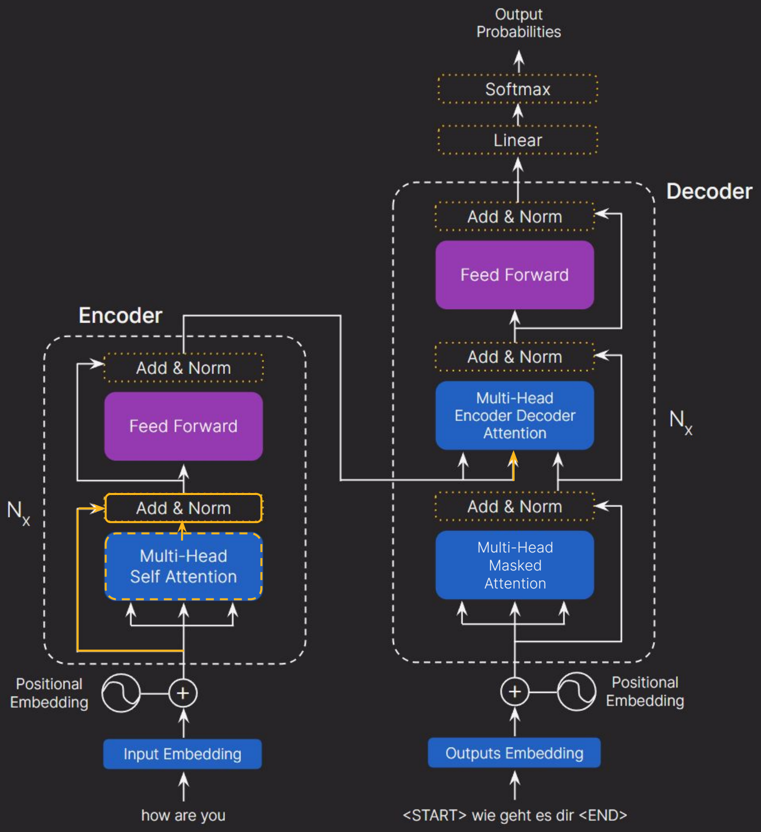
- The model uses positional encoding with word embeddings to generate a unique vector for each token.

P = Positional Embedding + Word embedding

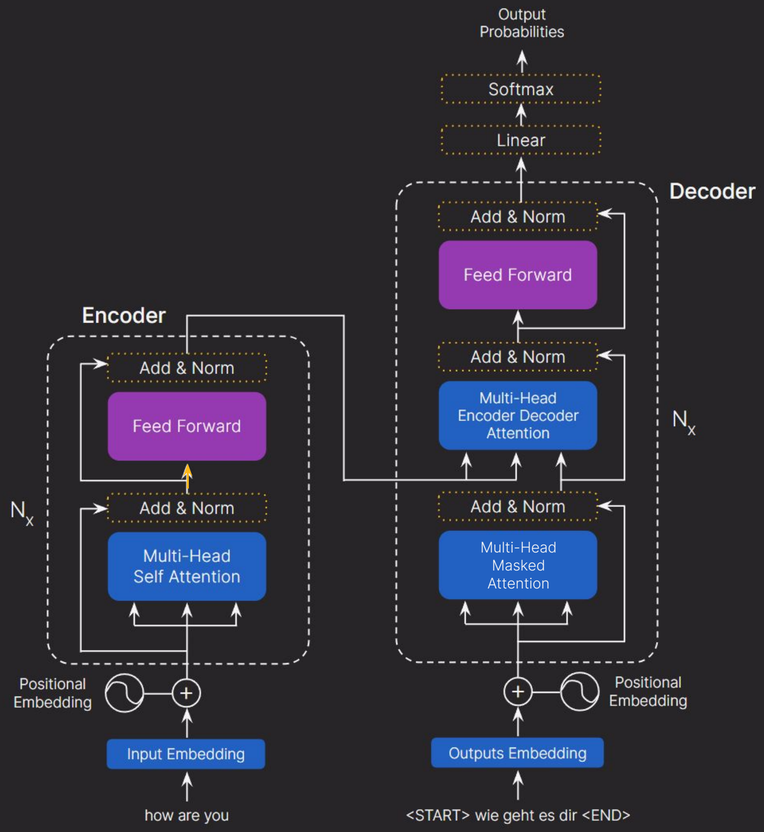
The Flow of Information



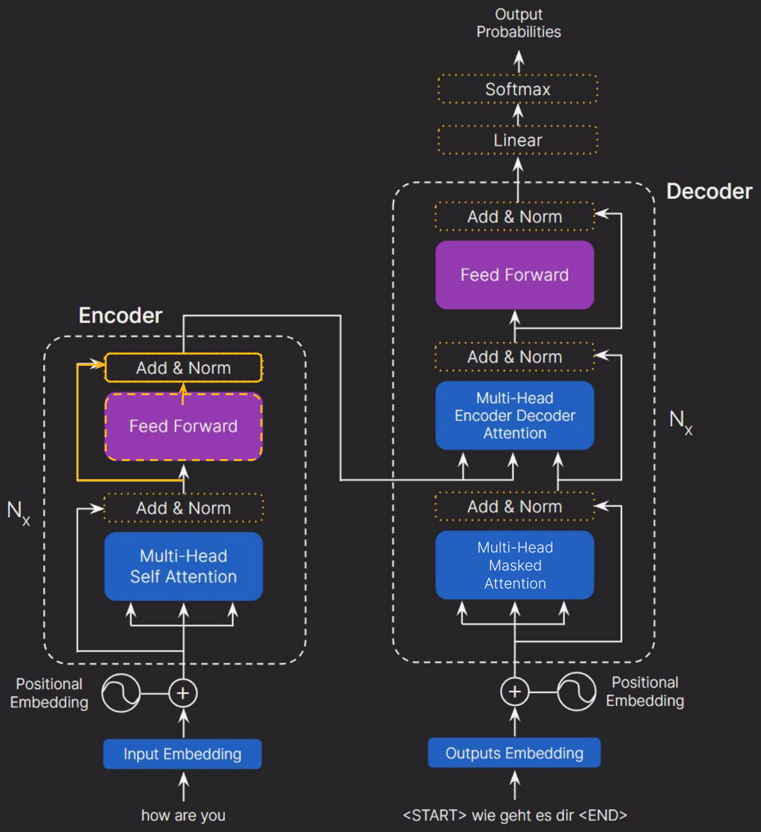
The Flow of Information



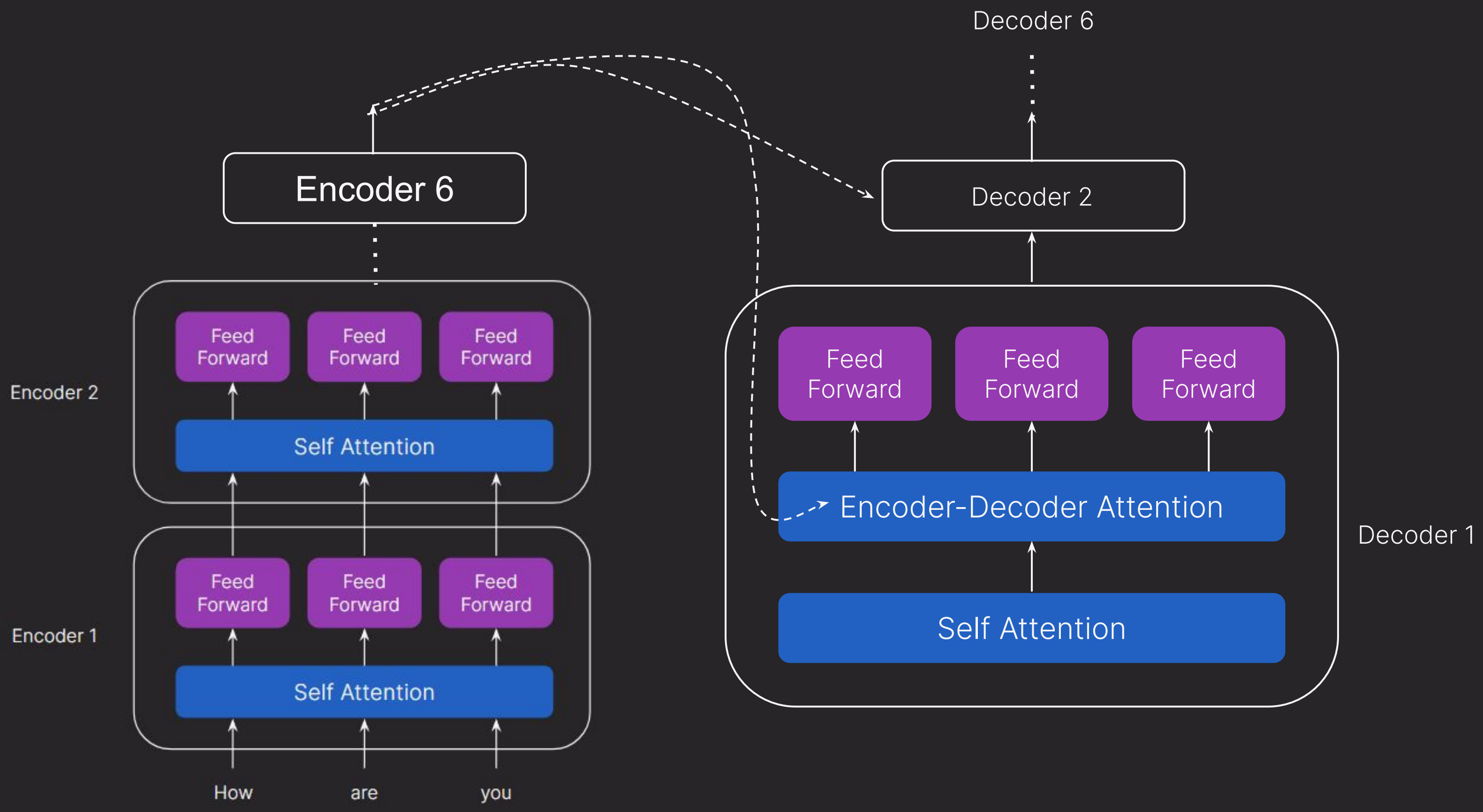
The Flow of Information



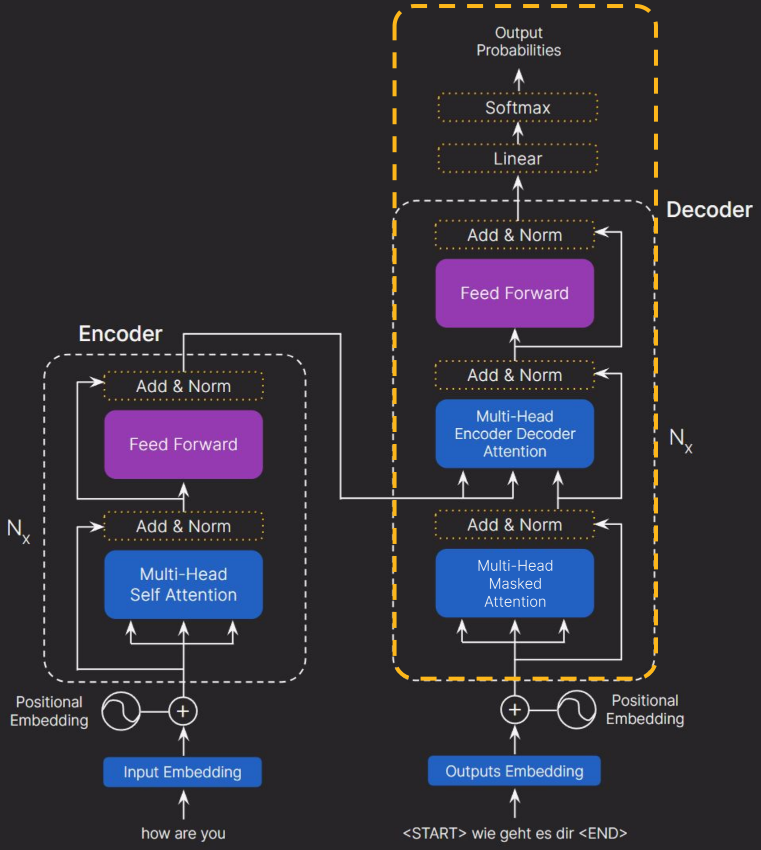
The Flow of Information



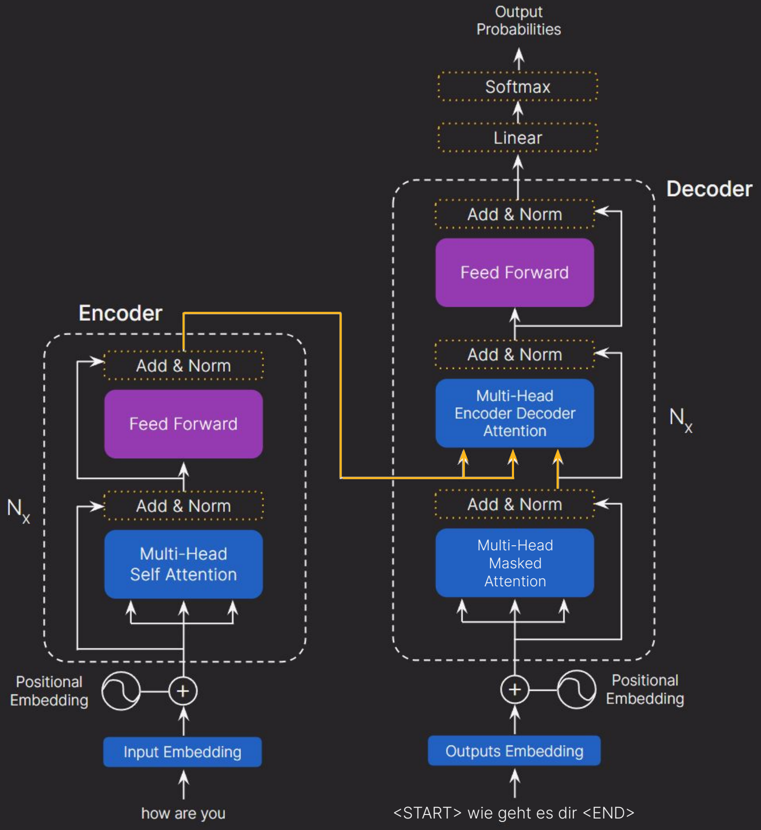
Multi-Head Self Attention



The Flow of Information



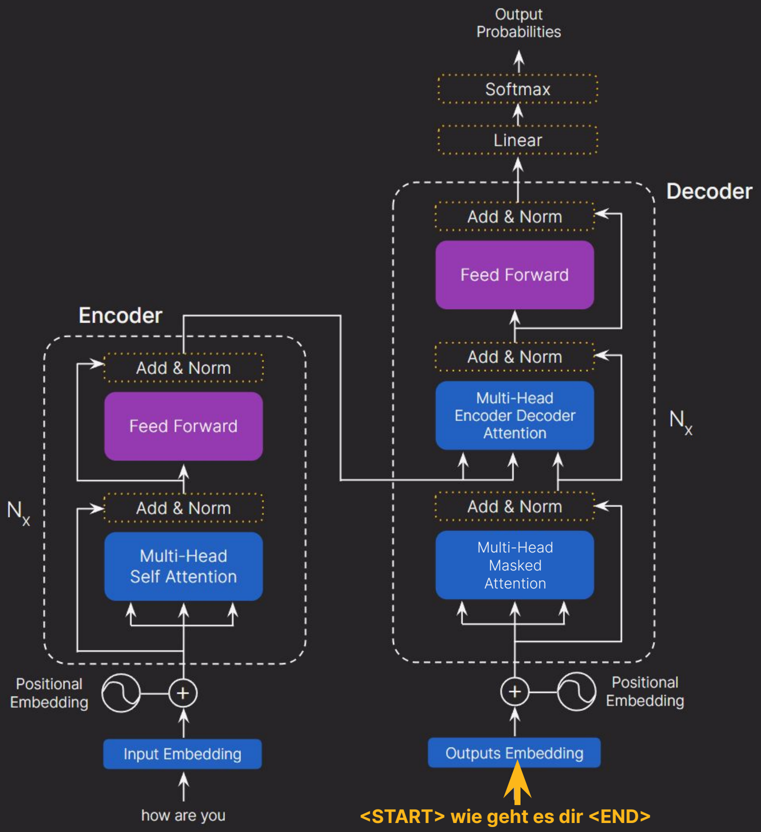
The Flow of Information



There are **2 inputs** to the decoder from the encoder:

- **First input:** Output matrix serves as the query and key matrix

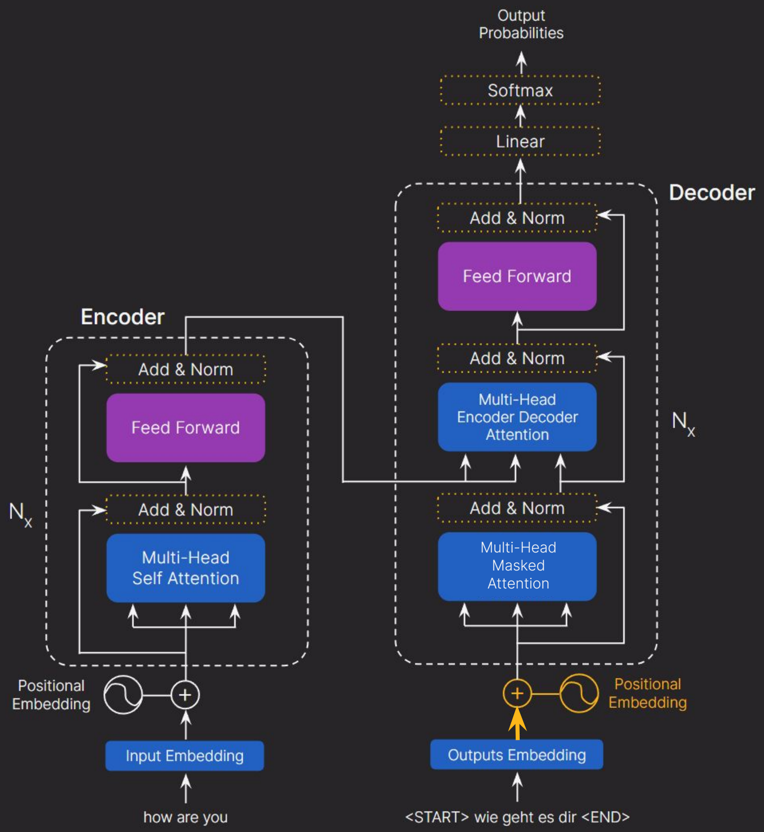
The Flow of Information



There are **2 inputs** to the decoder from the encoder:

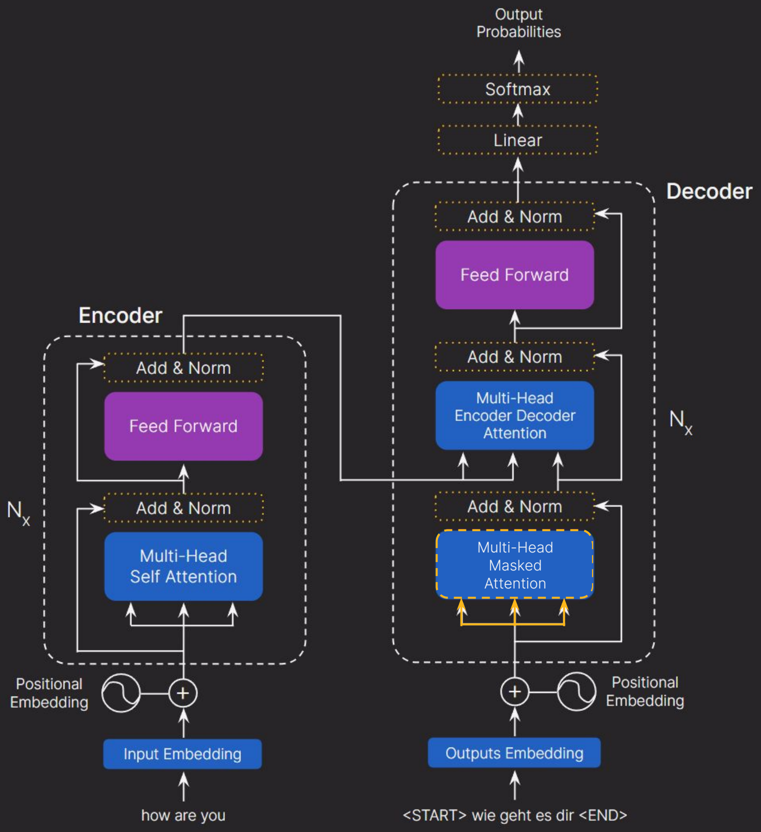
- **First input:** Output matrix “R” which serves as the query and key matrix
- **Second input:** Predicted text

The Flow of Information



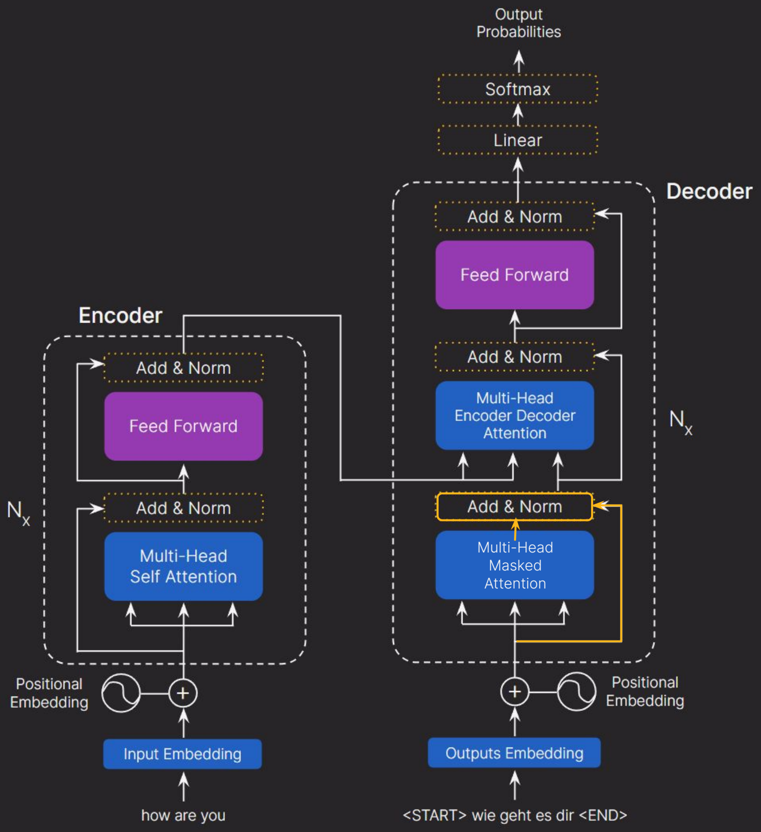
- Inputs are embedded just like the encoder part.

The Flow of Information



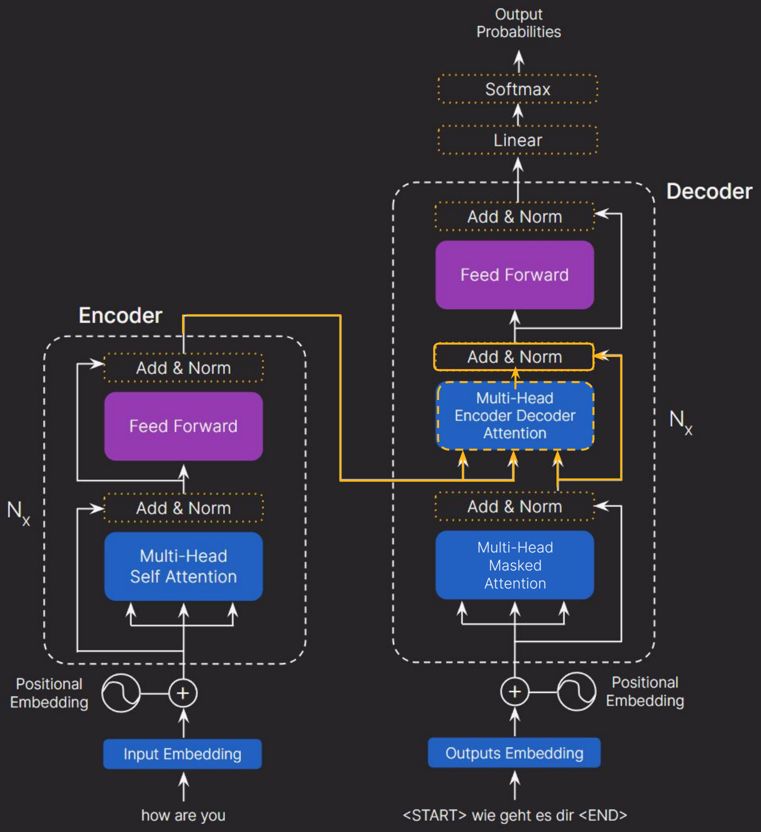
- **Masked multi-headed attention** is performed.
- This method focuses on relevant sentence parts without previewing future words.

The Flow of Information



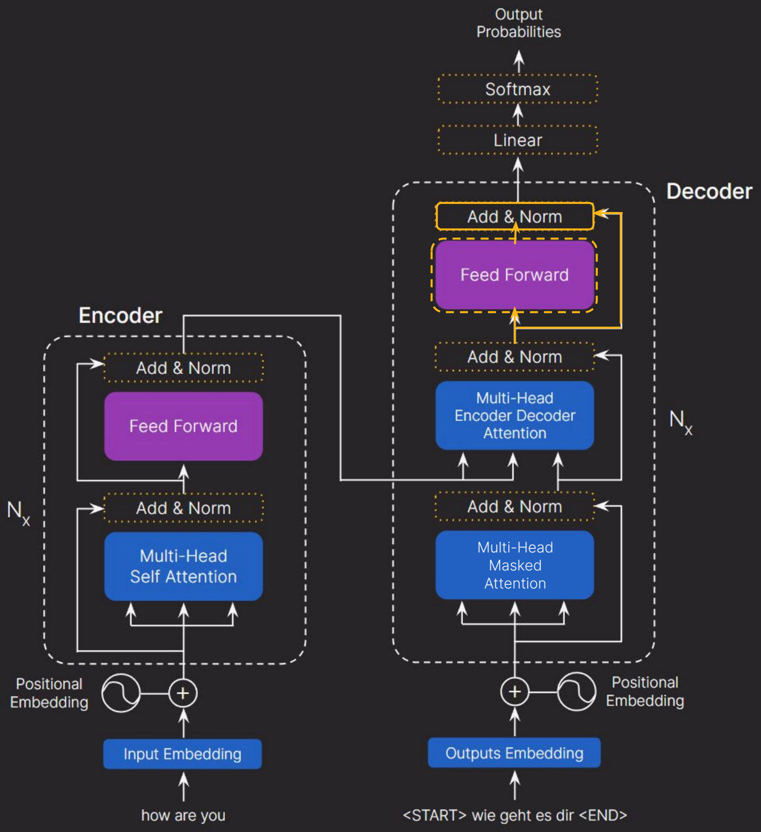
- The resultant matrix is added to the original matrix of the decoder part and then normalized.

The Flow of Information



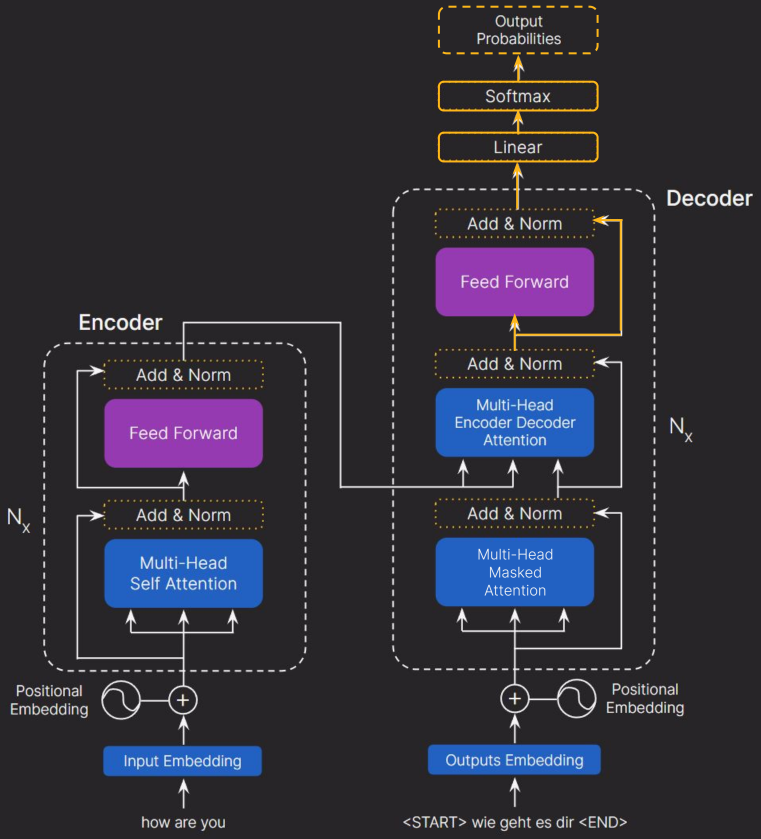
- 3 inputs to the multi-head encoder decoder attention:
  - Value vector from the encoder
  - Key vector from the encoder
  - Query vector after the first add and normalize step

The Flow of Information



- The output matrix is passed through a feed forward network
- It is added to the resultant matrix from earlier add and norm step to get the decoder stack output.

The Flow of Information

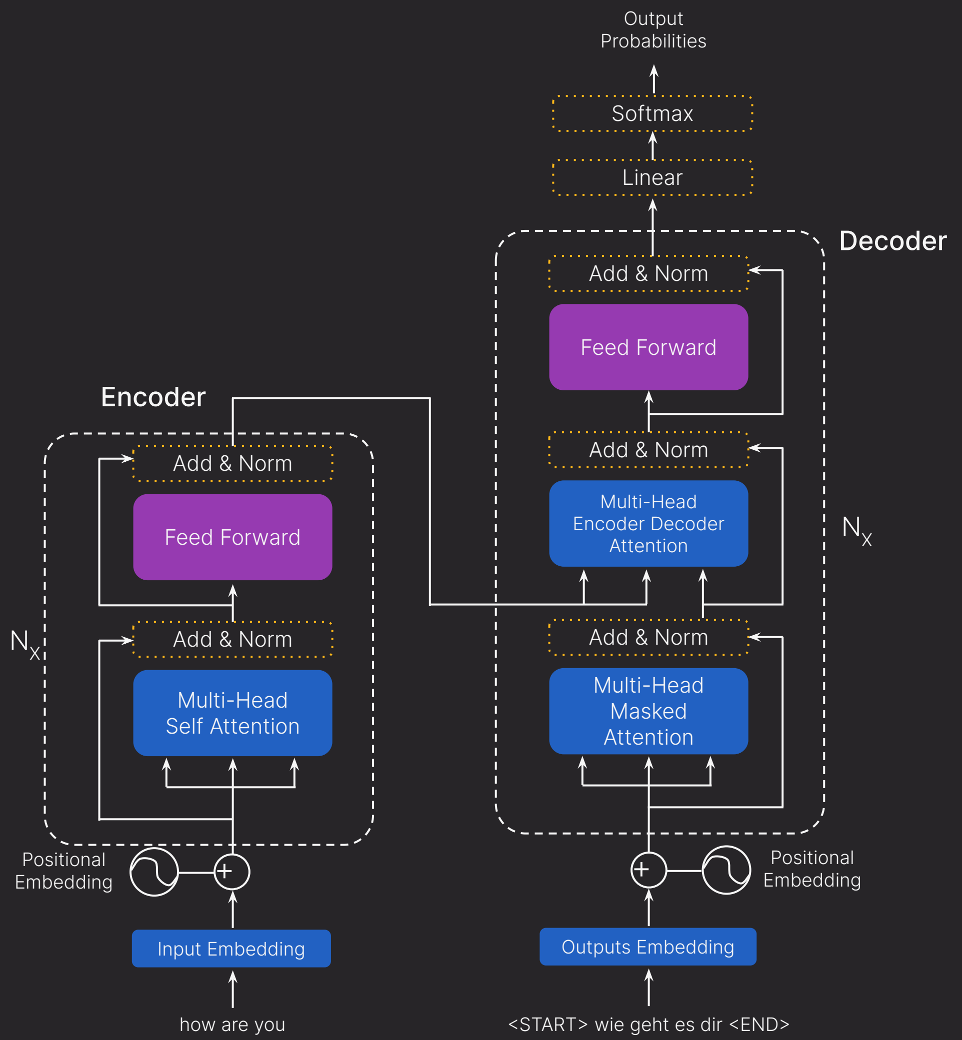


- Output of the decoder is passed through a linear layer followed by a softmax layer to get the prediction.

Analytics Vidhya

Up Next: Implementation of Transformers in Jupyter

IN AIR



Analytics Vidhya