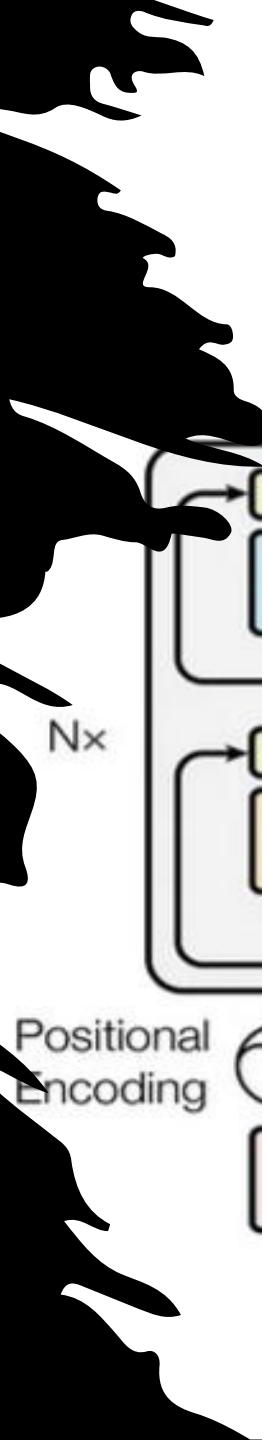
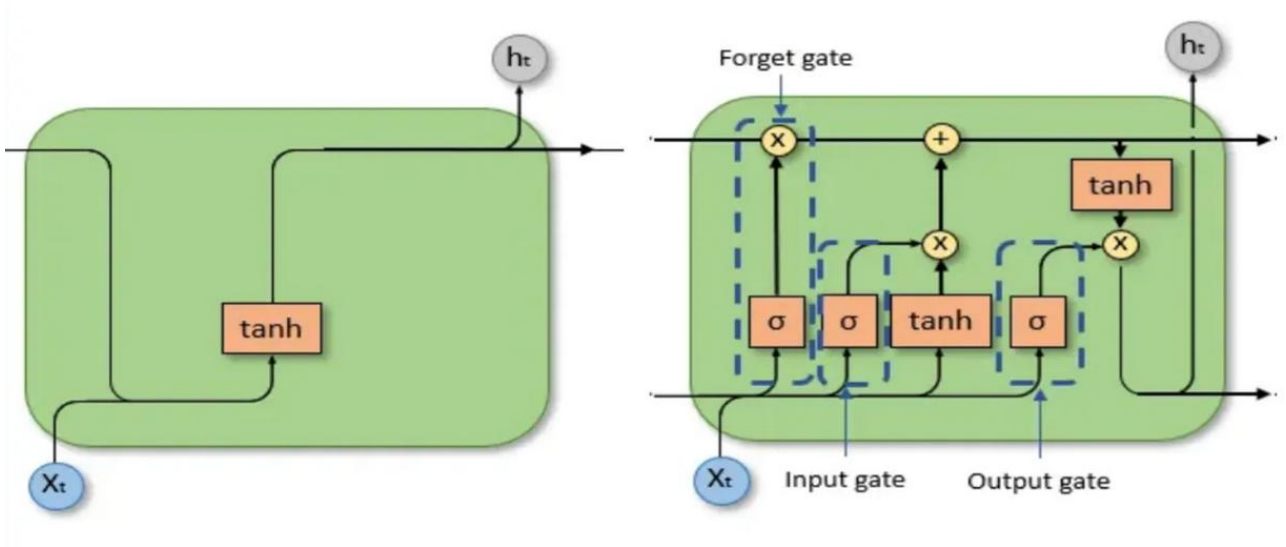
“ATTENTION IS ALL YOU NEED”

TRANSFORMERS AND SELF-ATTENTION

PRESENTER: SHAKKYA RANASINGHE



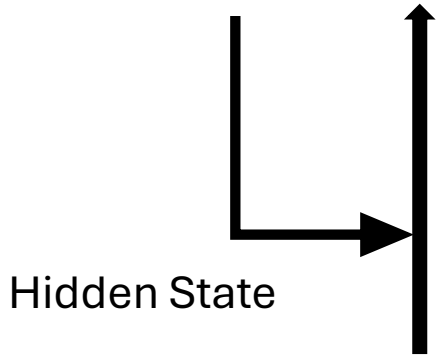
Once there was RNN's....



- Sequential Data
- Retained Information through Hidden States
- Improved with the use of cell states and gates into GRU and LSTM
  - Scalability
  - "Context"

x_{t-1} x_t

The quick brown fox jumps over the lazy dog.



“Attention is all you need”

Ashish Aswani et al. 2017

Machine Translation Paper

Natural Language Processing

Computer Vision

Generative Models

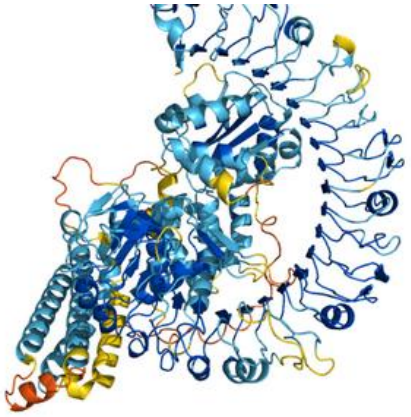
Translation

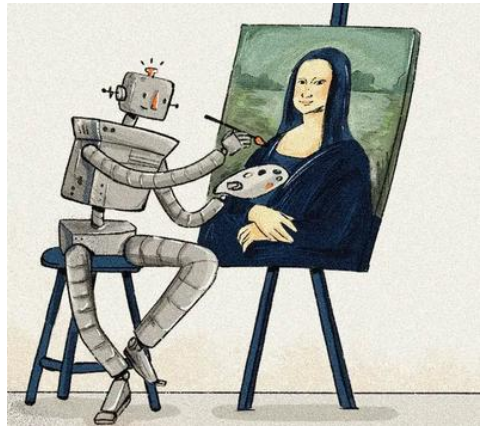
Reinforcement Learning

Speech Recognition

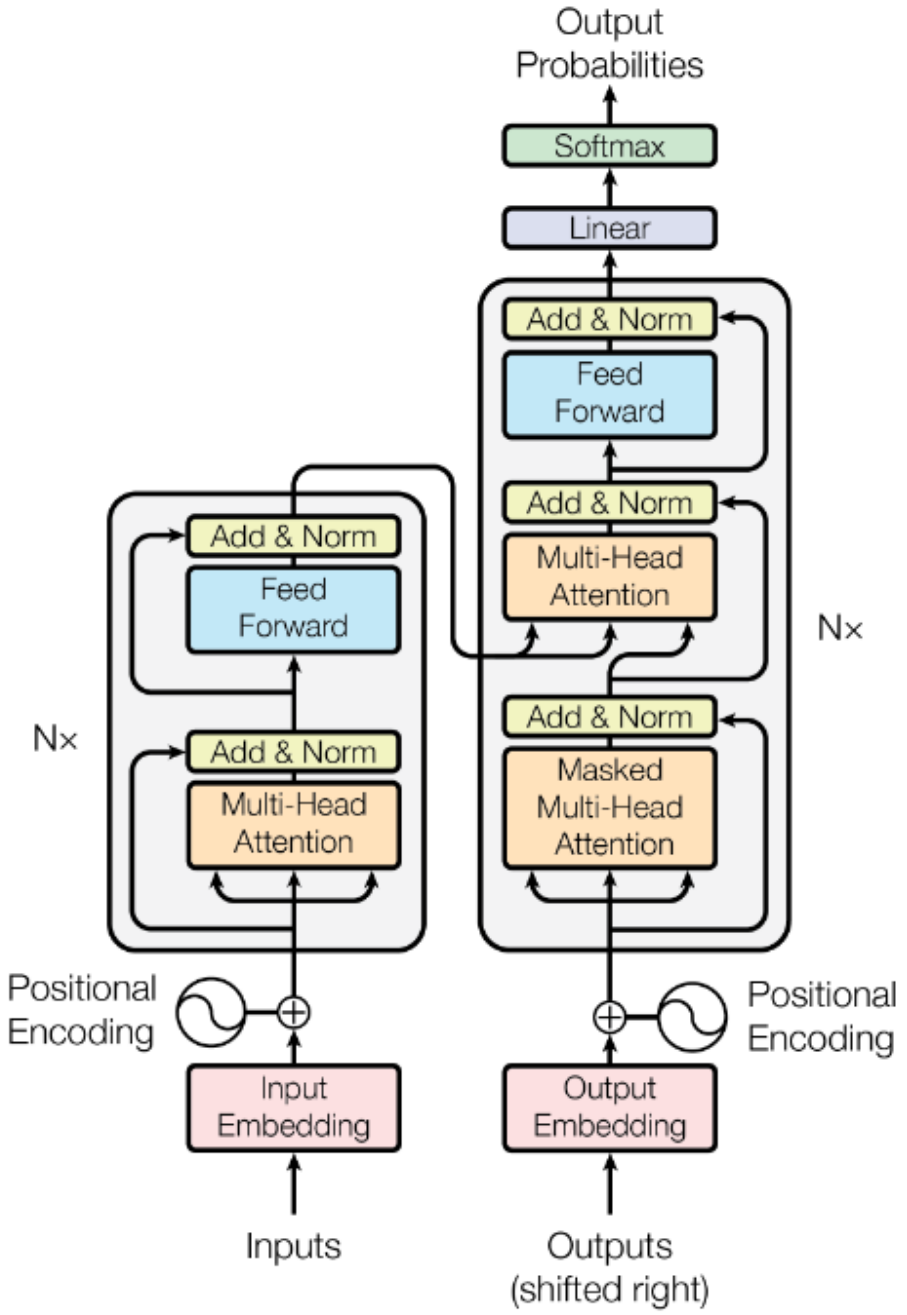
Evolution of Transformer Architecture

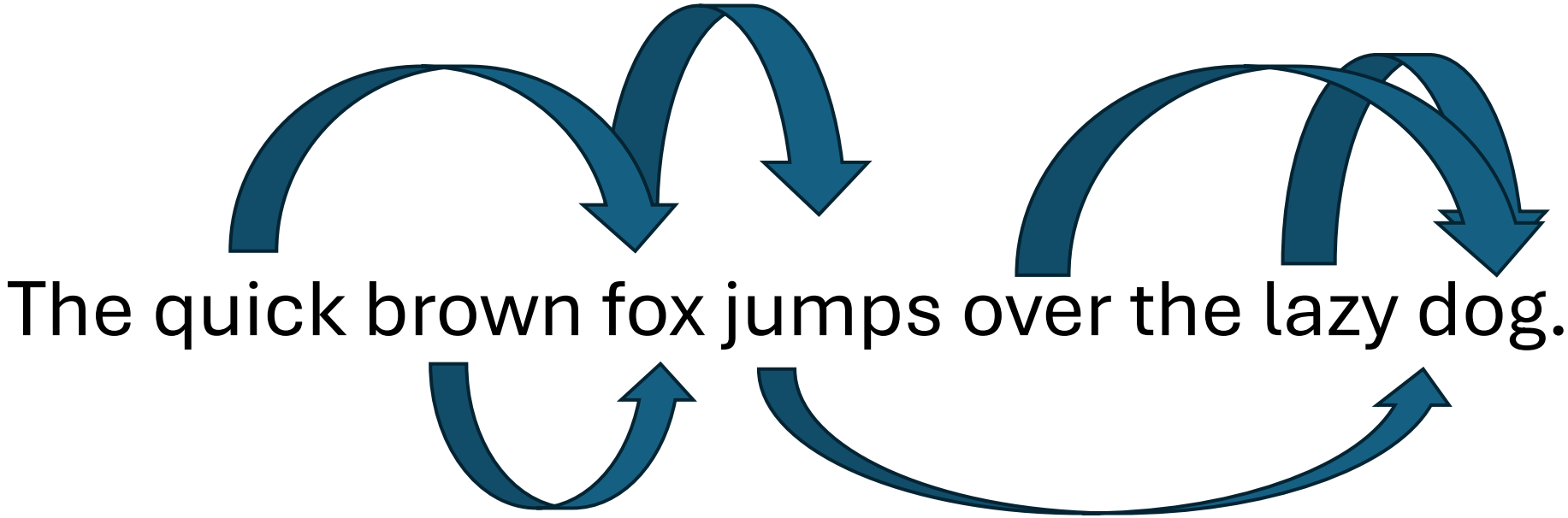
- Conversational AI and Agents
  - ChatGPT, Claude,
- Image Generation Models
  - DALL-E, SORA
- Healthcare
  - Alpha Fold
- Coding
  - Github Copilot
- Computer vision
  - Tesla Optimus





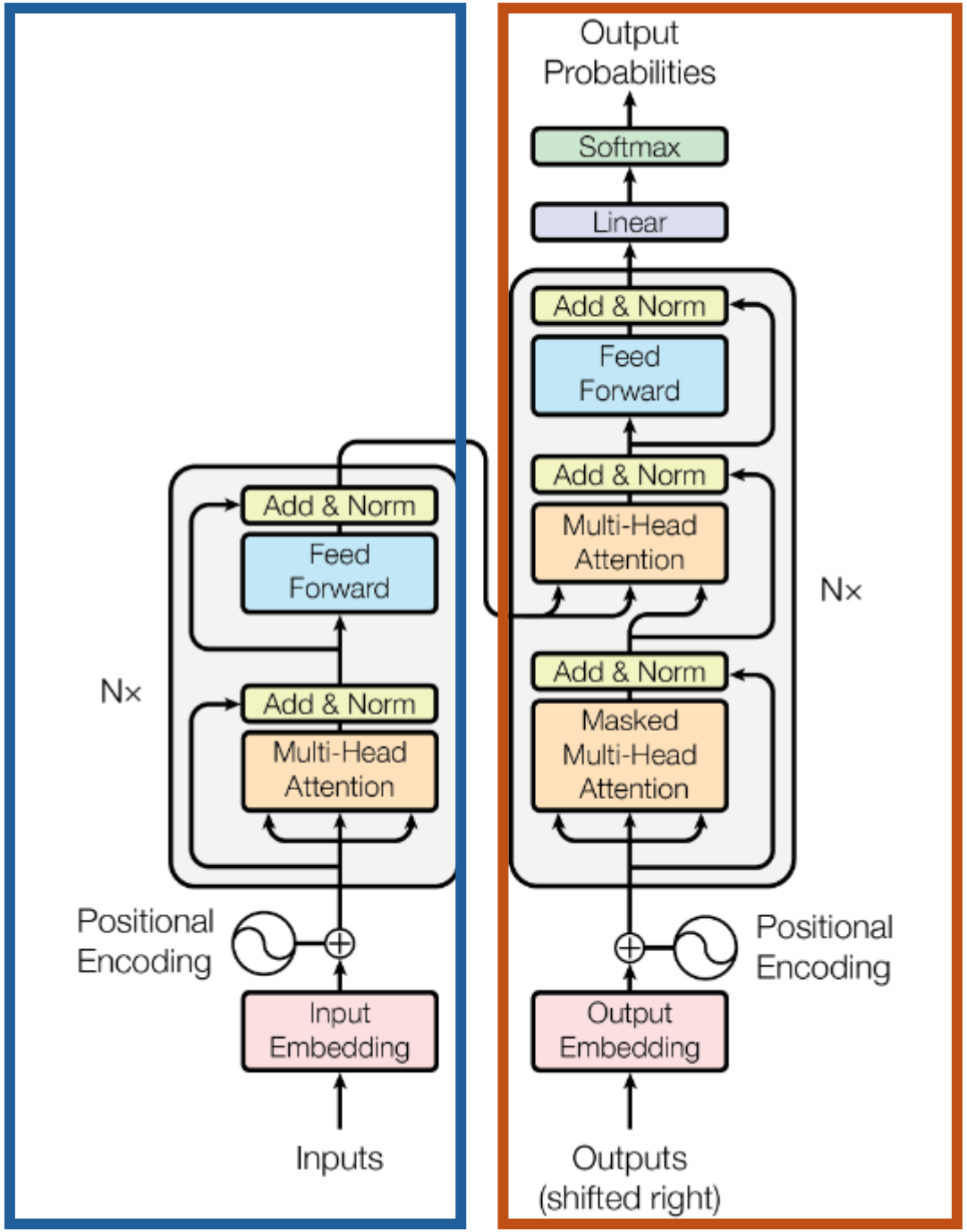
Transformer Architecture



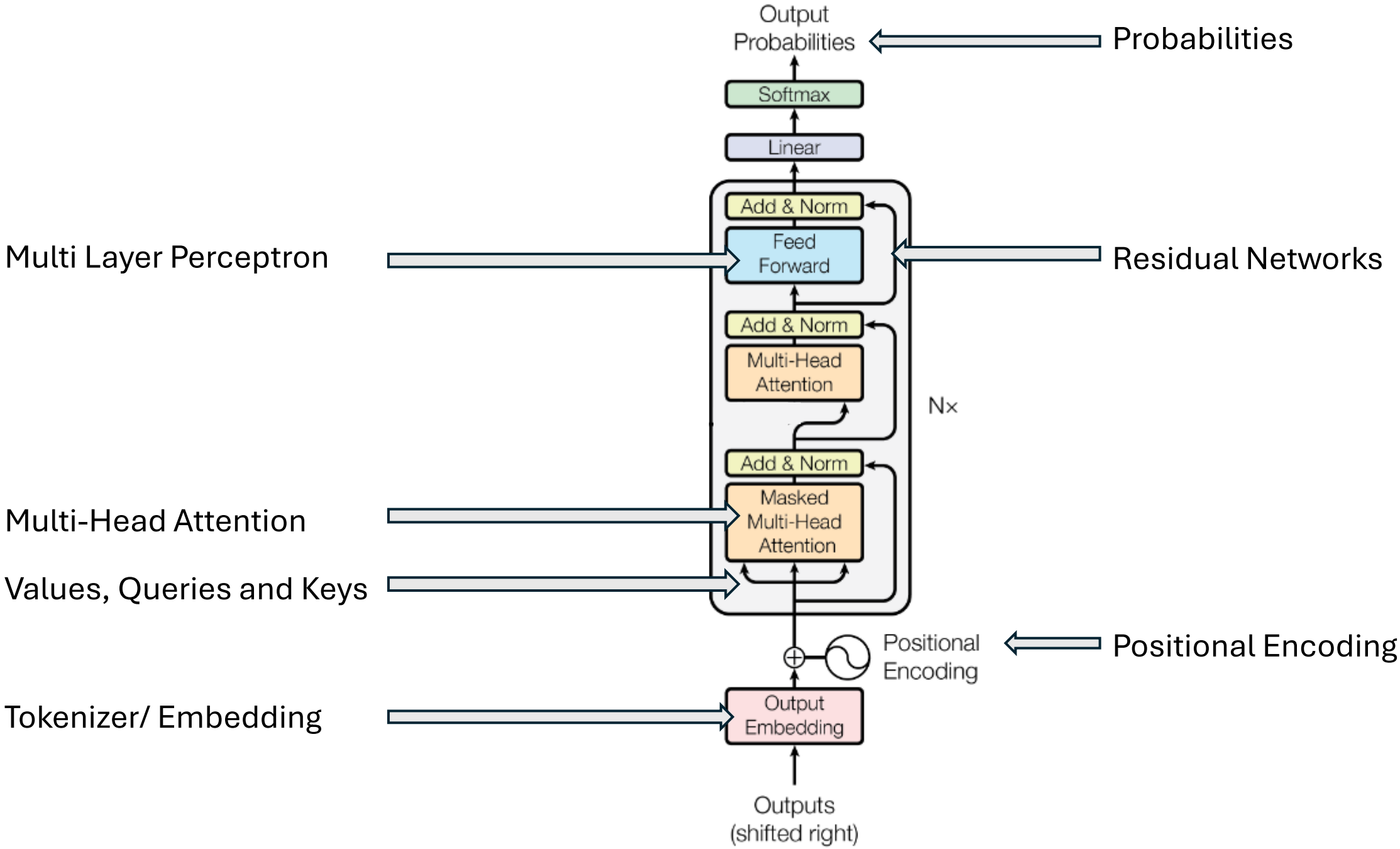


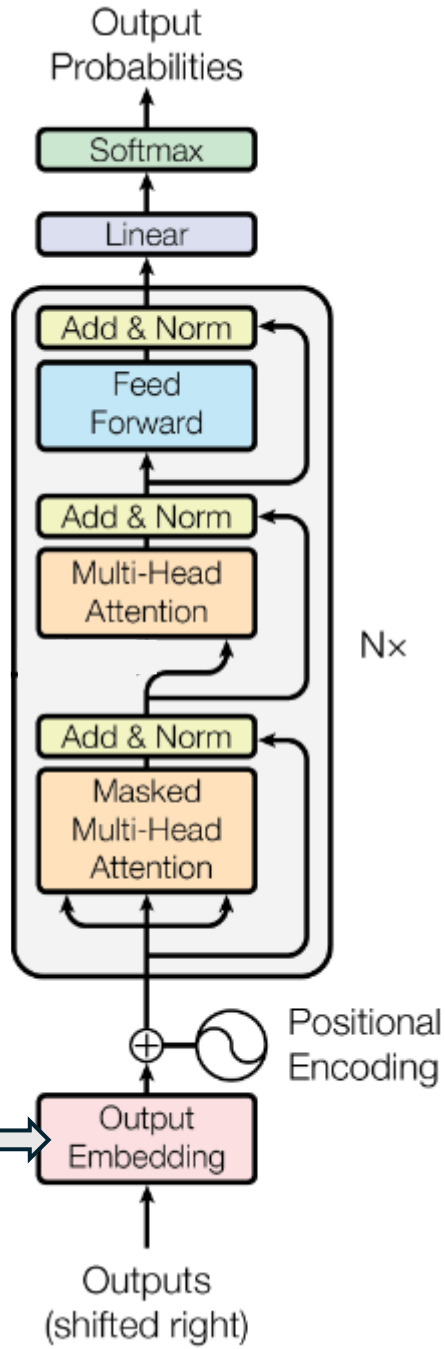
The quick brown fox
The quick brown fox jumps
The quick brown fox jumps over
The quick brown fox jumps over the
The quick brown fox jumps over the lazy
The quick brown fox jumps over the lazy dog

Encoder



Decoder


Output Probabilities
Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Residual Networks
Multi Layer Perceptron
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Masked Multi-Head Attention
Multi-Head Attention
Values, Queries and Keys
Positional Encoding
Positional Encoding
Output Embedding
Tokenizer/ Embedding
Outputs (shifted right)

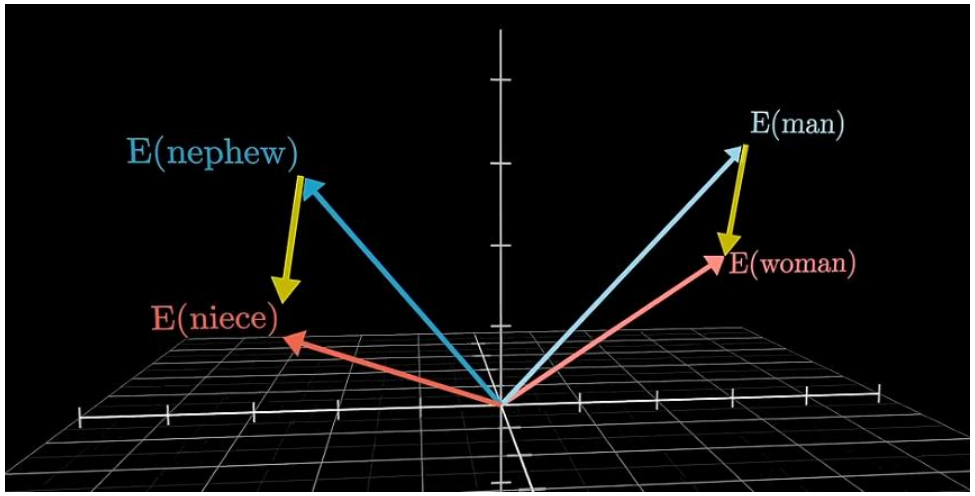


Positional Encoding

Tokenizer/ Embedding

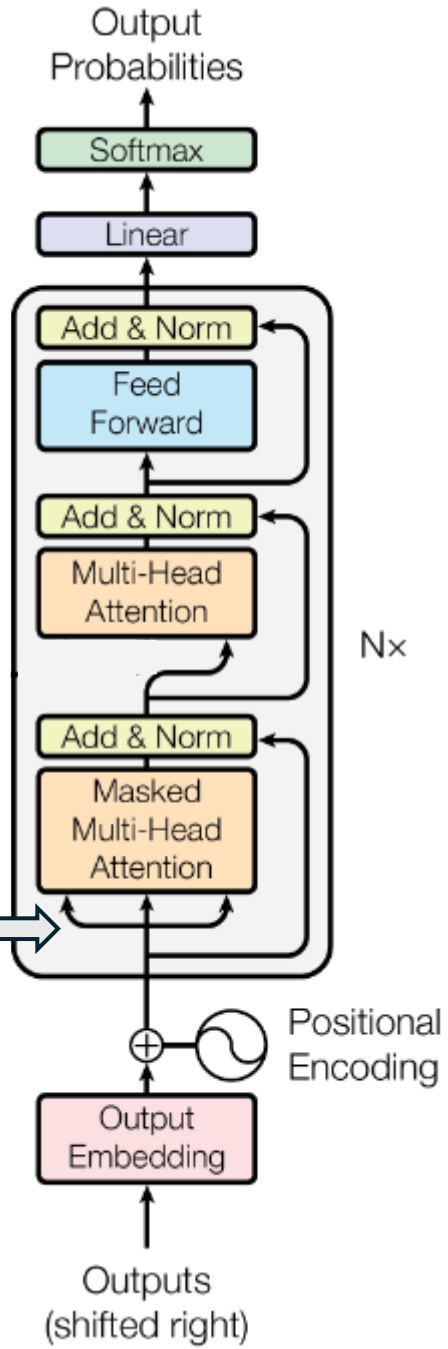
The quick brown fox jumps over the lazy dog

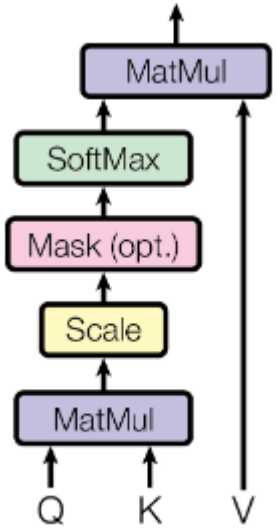
$$\begin{vmatrix} 0.4571 \\ 0.9855 \\ 1.3480 \\ -80.230 \\ -0.2392 \end{vmatrix}$$



*GPT3 Model embeddings has 12,288 dimensions

Values, Queries and Keys





Query: Question to be asked on the Data

Key: Answer Position of the Data

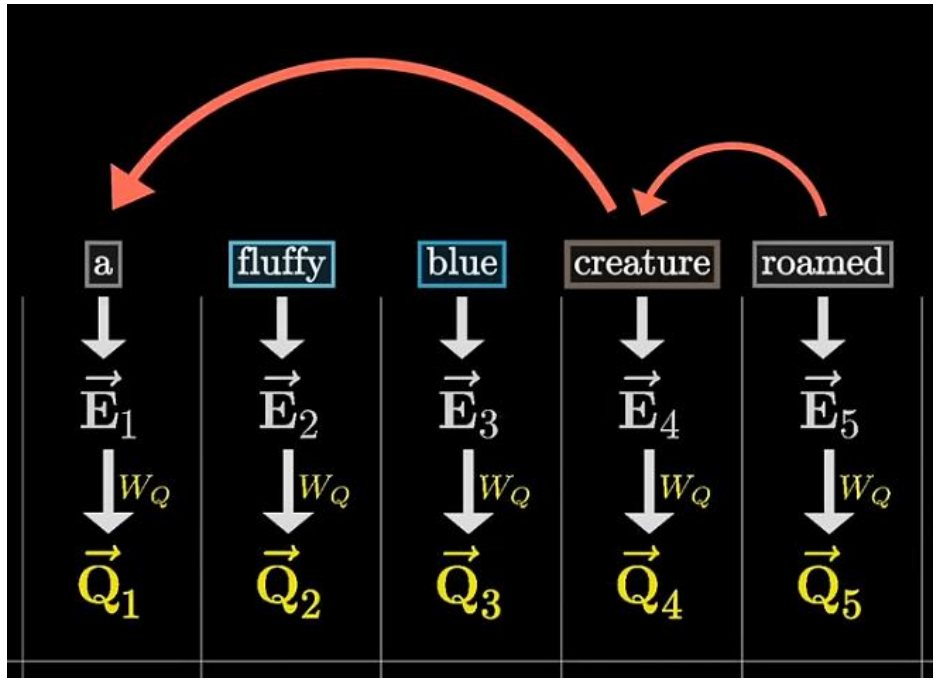
Dot Product

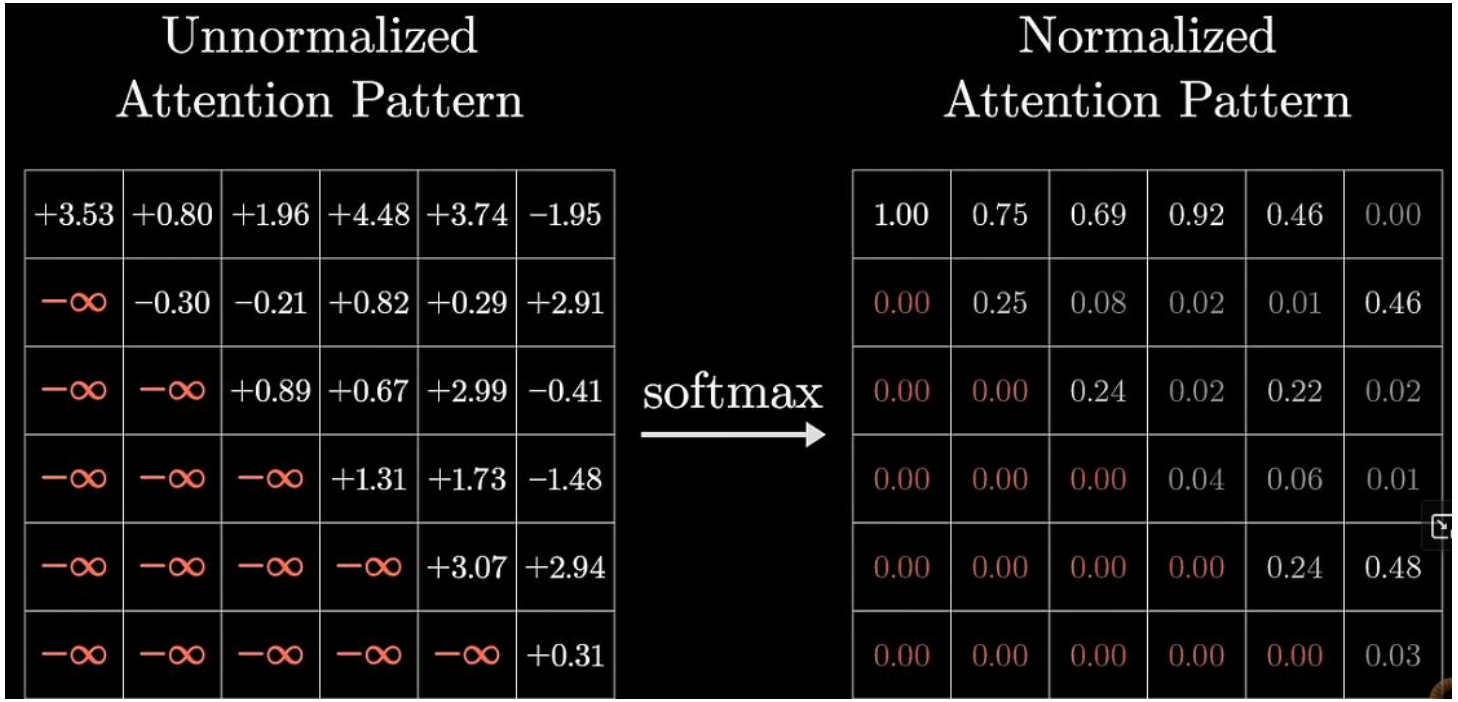
Value: Importance/relevance between Key & Query for each word: “Attention”

Softmax(Q, K, V)

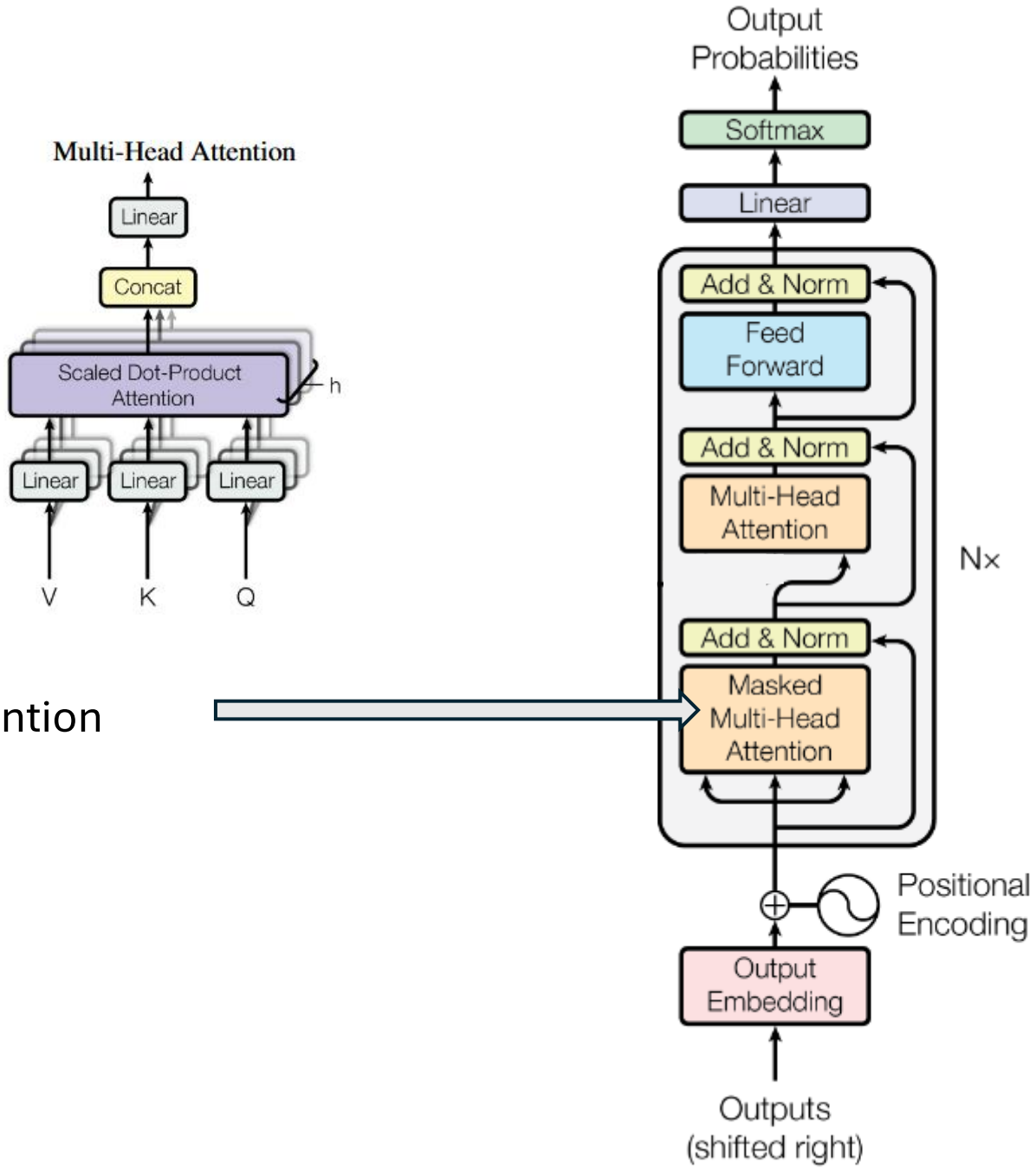
Probability Distribution of possible words

Masking: Following Words Cannot influence previous words

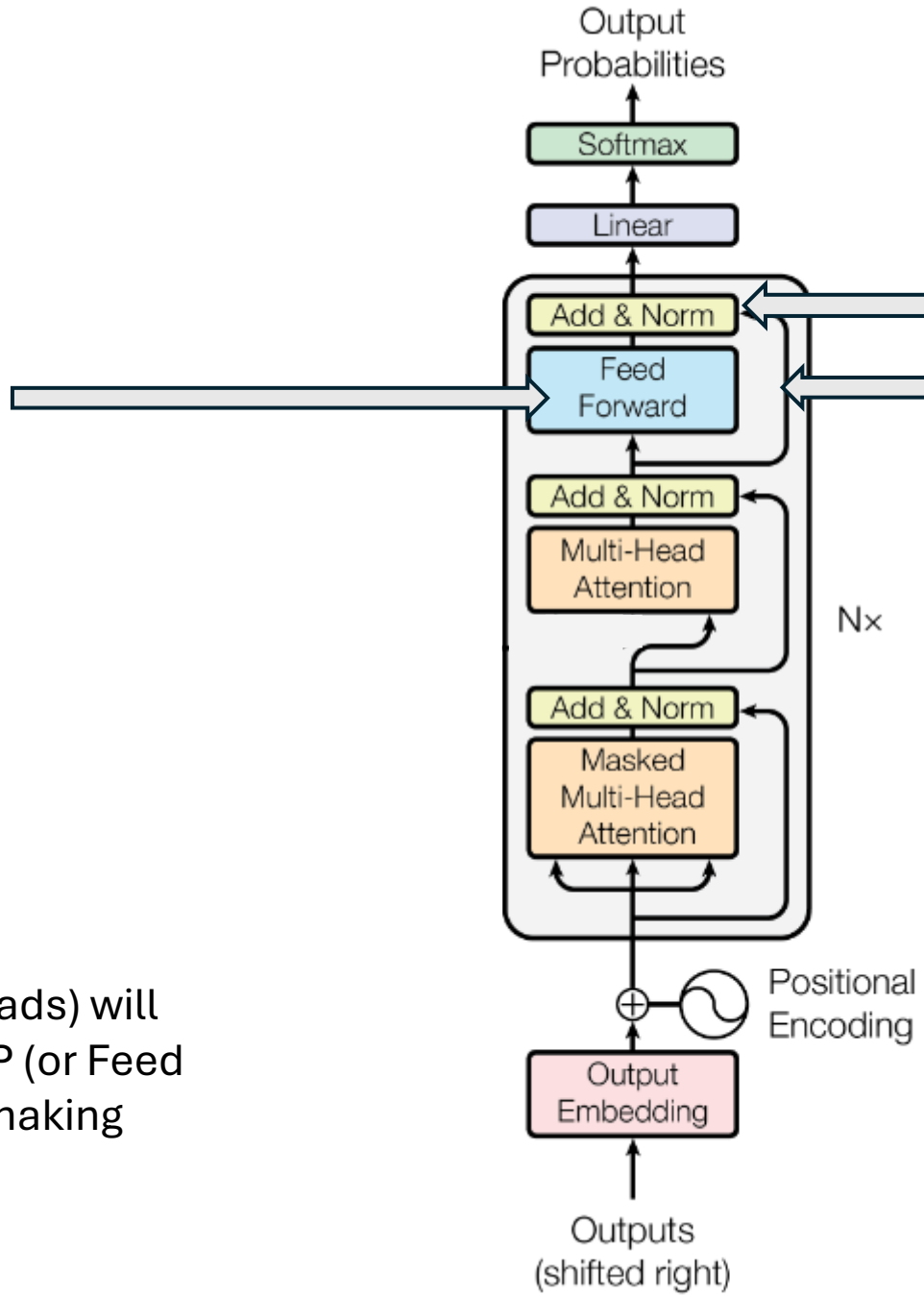




Multi-Head Attention



Multi Layer Perceptron

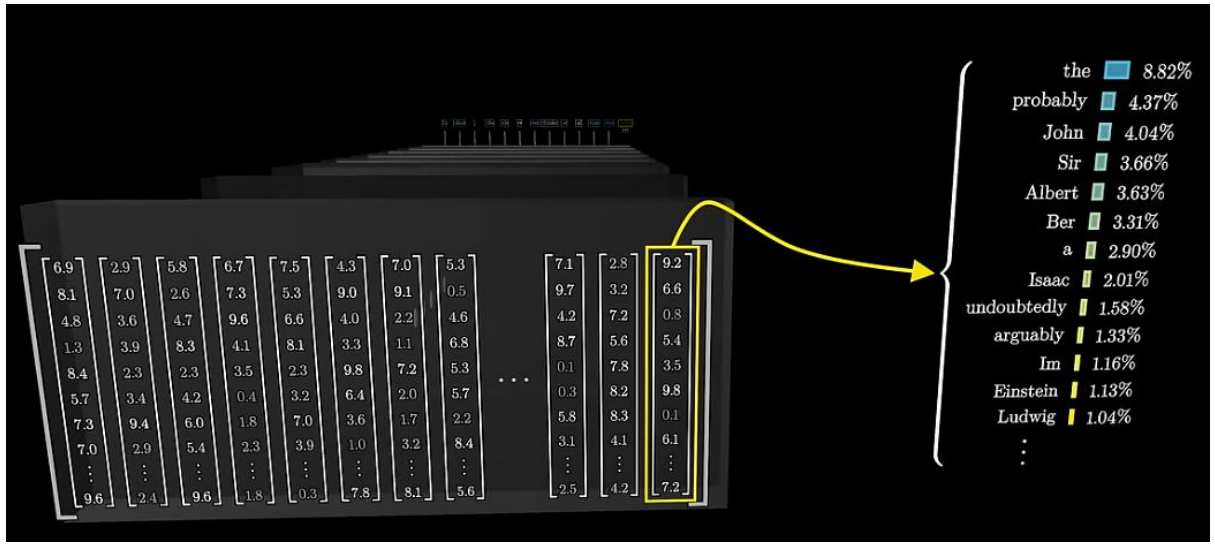


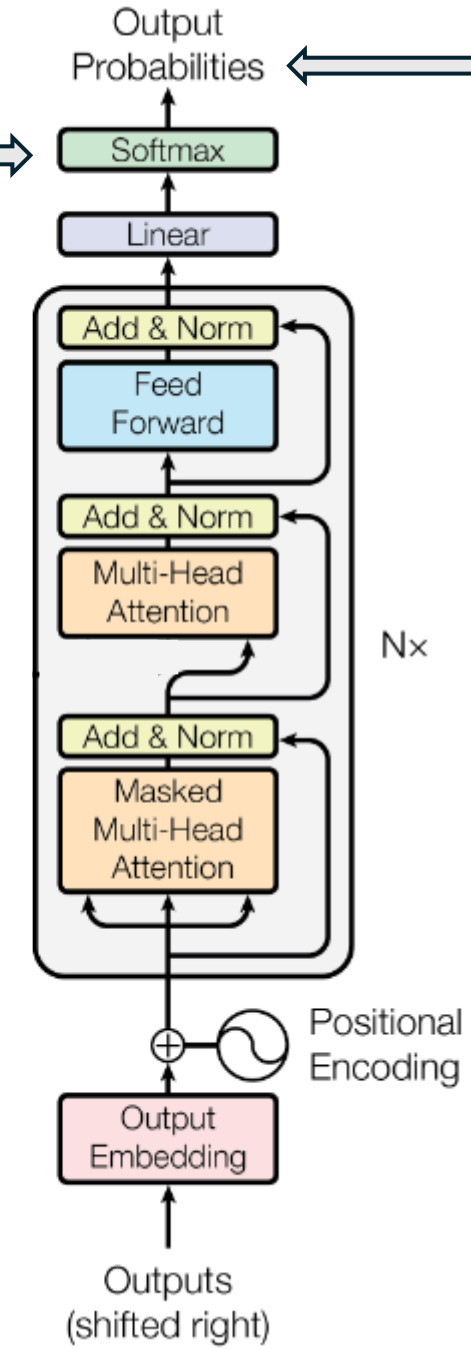
Layer Norms

Residual Networks

Self Attention blocks (or multi-heads) will gather the data, where as the MLP (or Feed Forward networks), will work on making sense of the data

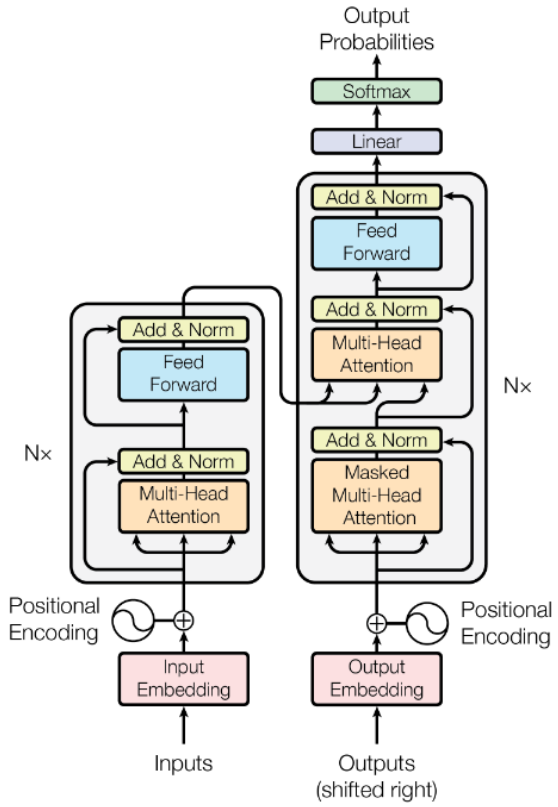
- Softmax
  - Rigid
- Temperature Softmax
  - Creativity !!





Probabilities

Thank you !!



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Reference: Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.