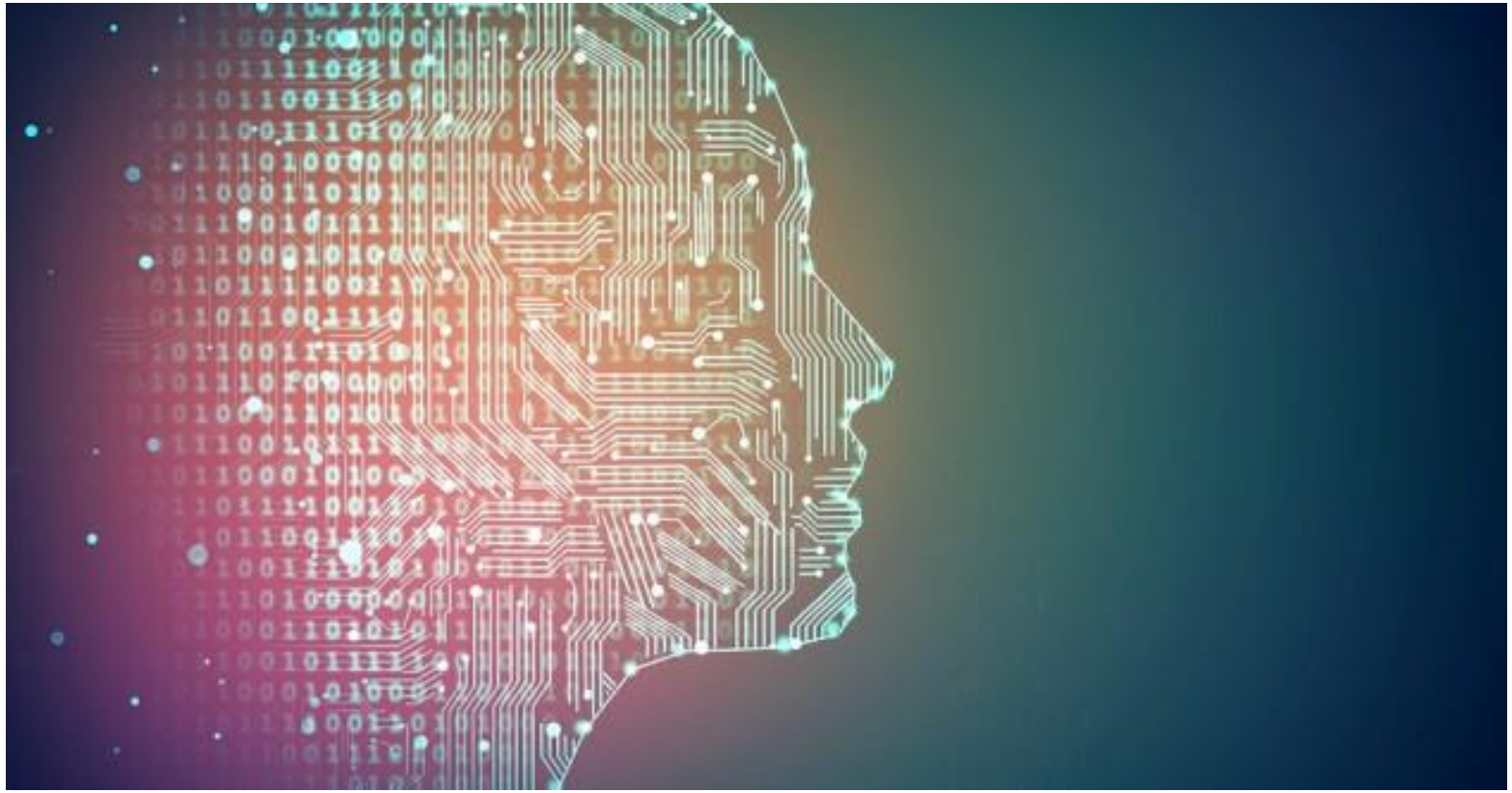
Generative AI: Transformers: Architecture
PART III: Decoder

Thuan L Nguyen, PhD

2: Generative AI: LLM: Transformers: Architecture



AI Deep learning (Source: mindovermachines.com)

3: Generative AI: LLM: Transformers: Architecture

1. Generative AI: Transformers: Architecture: Overview
2. Generative AI: Transformers: Architecture: Encoder-Decoder Structure
3. Generative AI: Transformers: Architecture: Revolutionary Core Features
4. Generative AI: Transformers: Architecture: Input Embedding & Positional Encoding
5. Generative AI: Transformers: Architecture: Attention & Multi-Header Attention
6. Generative AI: Transformers: Architecture: Encoder & Encoding Stack of Layers
7. Generative AI: Transformers: Architecture: Decoder & Decoding Stack of Layers

4: Generative AI: LLM: Transformers: Architecture

Artificial Intelligence: Generative AI

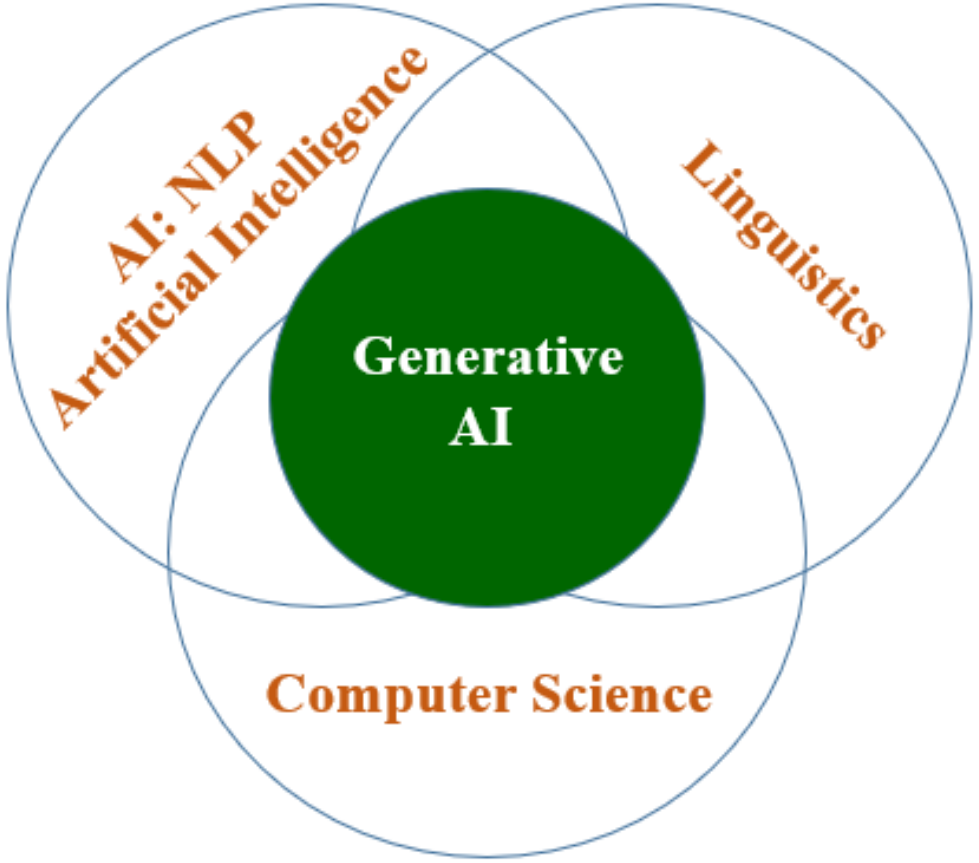
What is It?

Generative AI: A category of artificial intelligence focused on using AI deep learning models to generate new contents, including text, images, audio, video, and more. The contents are novel but look realistic and may be indistinguishable from human-created ones.

5: Generative AI: LLM: Transformers: Architecture

Artificial Intelligence: Generative AI: LLM

Foundational Sciences & Technologies



Generative AI is based on the NLP technologies such as Natural Language Understanding (NLU) and Conversational AI (AI Dialogues) - Those among the most challenging tasks AI needs to solve.

6: Generative AI: LLM: Transformers: Architecture

Artificial Intelligence: Generative AI: LLMs

Large Language Models

Large Language Models (LLMs) are revolutionary AI **Deep Learning neural networks** that excel in natural language understanding (NLU) and content generation.

- “LARGE" in LLMs refers to the vast scale of data and parameters used to train them, allowing LLMs to develop a comprehensive understanding of language.
- Being particularly transformer-based models trained on massive text datasets using deep learning techniques,
- Able to learn complex language patterns, capture nuances like grammar and tone, and generate coherent and contextually relevant text

7: Generative AI: LLM: Transformers: Architecture

Artificial Intelligence: Generative AI: LLMs

Large Language Models: Transformers

***Photo Source**: **Prompt** by Thuan L Nguyen – Generated by **Generative AI Text-To-Images**: Google-DeepMind ImageFX*

8: Generative AI: LLM: Transformers: Architecture

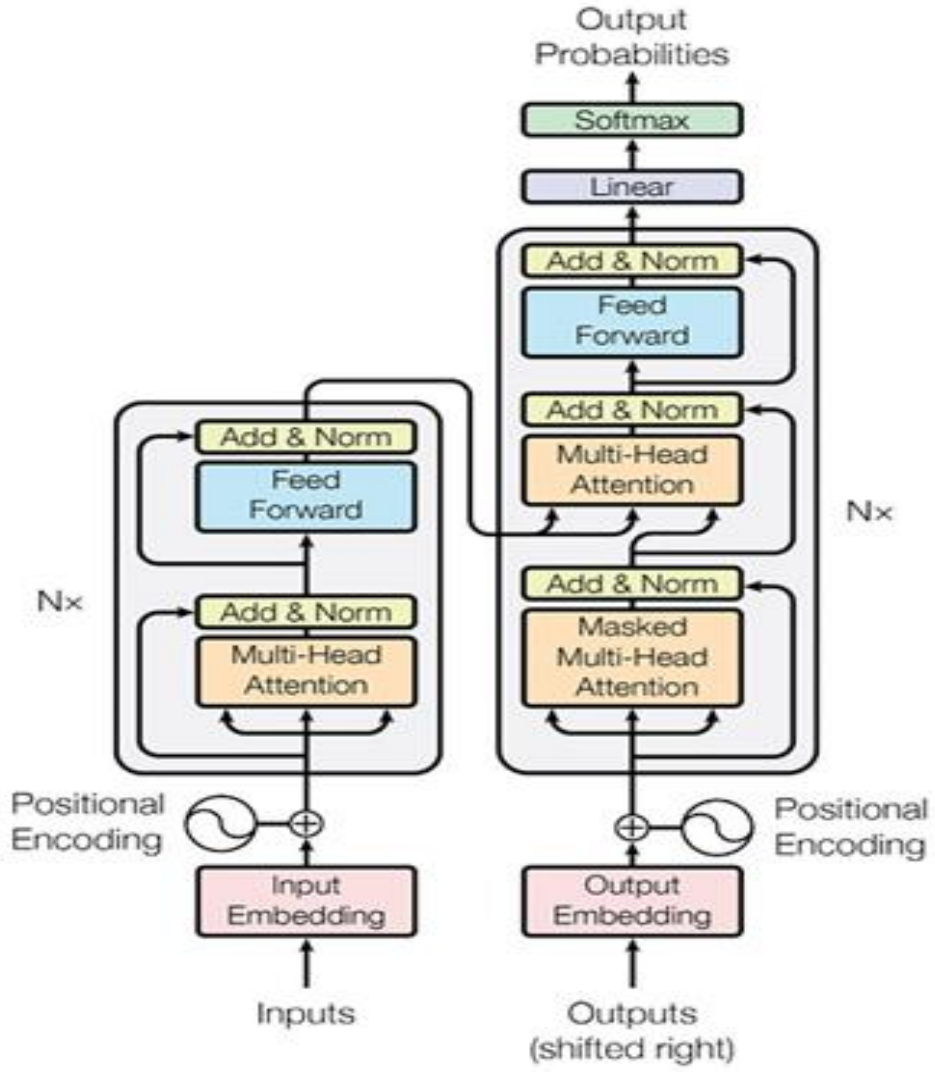
“**Attention Is All You Need!**”

***Photo Source**: **Prompt** by Thuan L Nguyen – Generated by **Generative AI Text-To-Images**: Google-DeepMind ImageFX*

9: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: Architecture

Encoder-Decoder



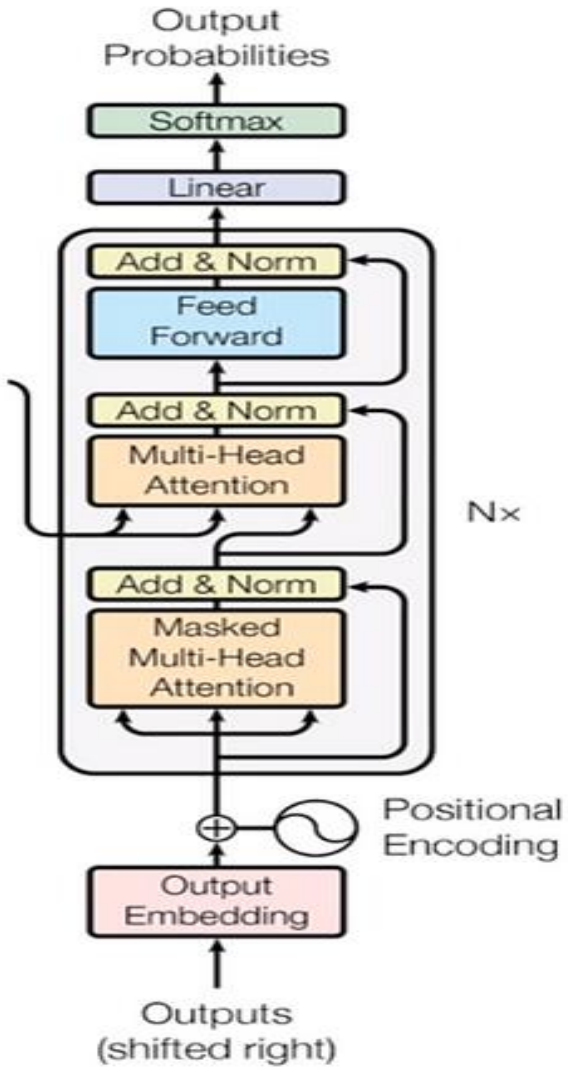
The Transformer - model architecture.

source: arXiv:1706.03762

10: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: Architecture

Decoder



source: arXiv:1706.03762

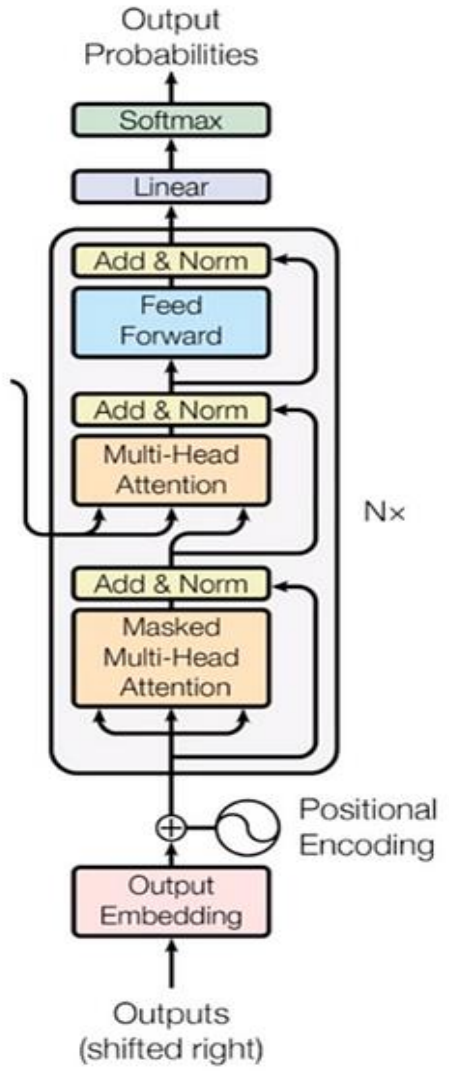
11: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: Architecture

Decoder

Main Layers of the Transformer: Decoder

- **Masked Multi-Head Attention**: Ensure, during output generation, that dependencies only consider previously generated words (auto-regressive property).
- **Encoder-Decoder Attention**: Connect decoder to encoder output, helping it focus on relevant areas of the input sequence.
- **Layer Normalization**
- **Feed-Forward Network**
- **Residual Connections**
- **Final Linear Layer and Softmax**: Transform the decoder's output into logits (raw predictions) and further into a probability distribution over the vocabulary.

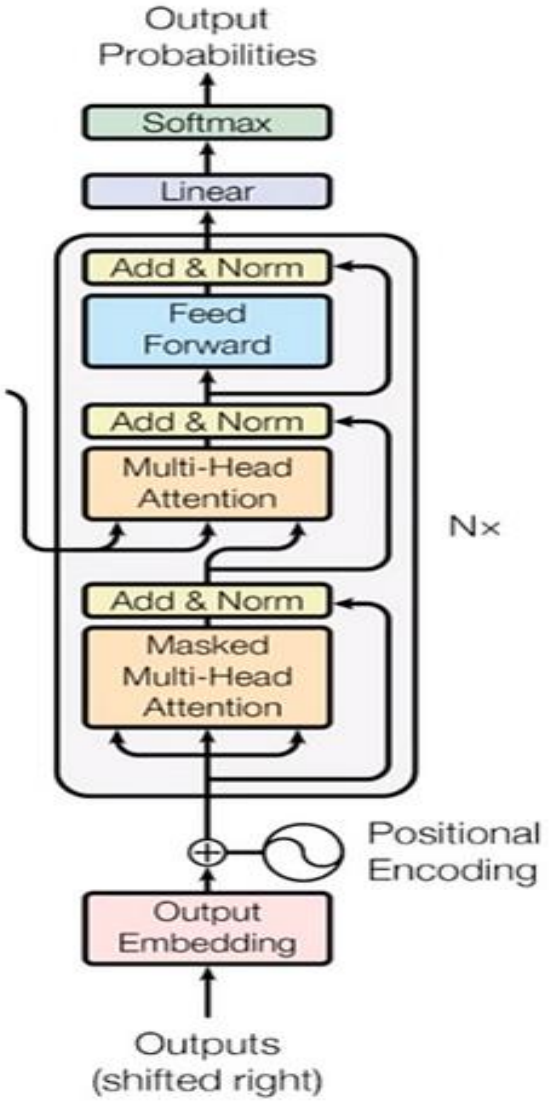


source: arXiv:1706.03762

12: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: Architecture

Decoder



source: arXiv:1706.03762

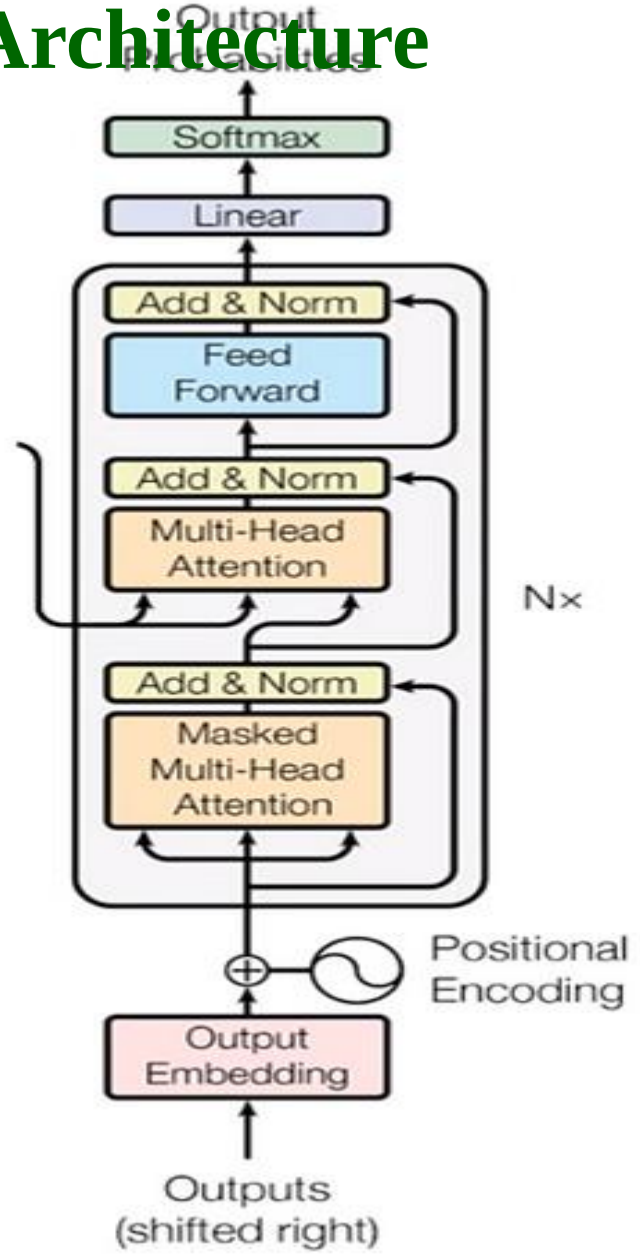
13: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

Decoder

It is **assumed** that it is in the process of training a **translator** for English to the French language

- Give an English sentence along with its translated French sentence for the model to learn.
- The **English** sentence passes through **Encoder** Block, and the **French** sentence pass through the **Decoder** Block.



source: arXiv:1706.03762

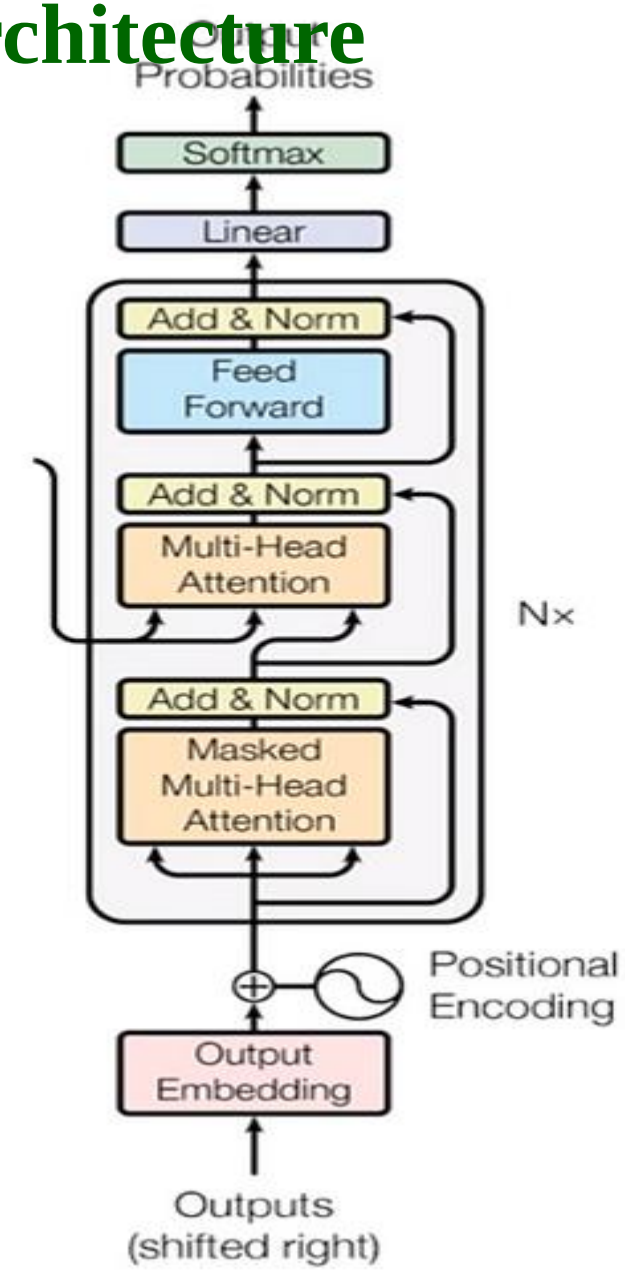
14: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

Decoder

First, **Embedding** Layer and **Positional** Encoder transform the words into respective **vectors**.

- Similar to what has been done in the Encoder.



source: arXiv:1706.03762

15: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

Decoder

Masked Multi-Head Attention (Self-Attention)

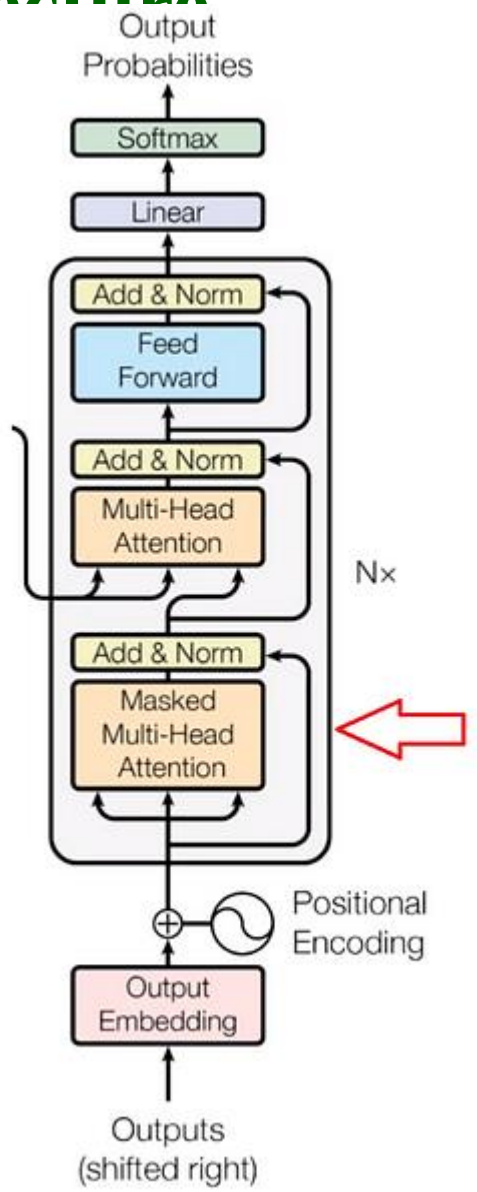
TO DO: Find out how relevant a particular word (French) is w.r.t to other words in that sentence.

OUTPUT: **Attention vectors**

- Each vector represents the **relevance** of a word (French) w.r.t other words.

Every word → An **attention vector**

- Capture the **contextual relationship** between the word (French) and other words in that sentence.



source: arXiv:1706.03762

16: Generative AI: LLM: Transformers: Architecture

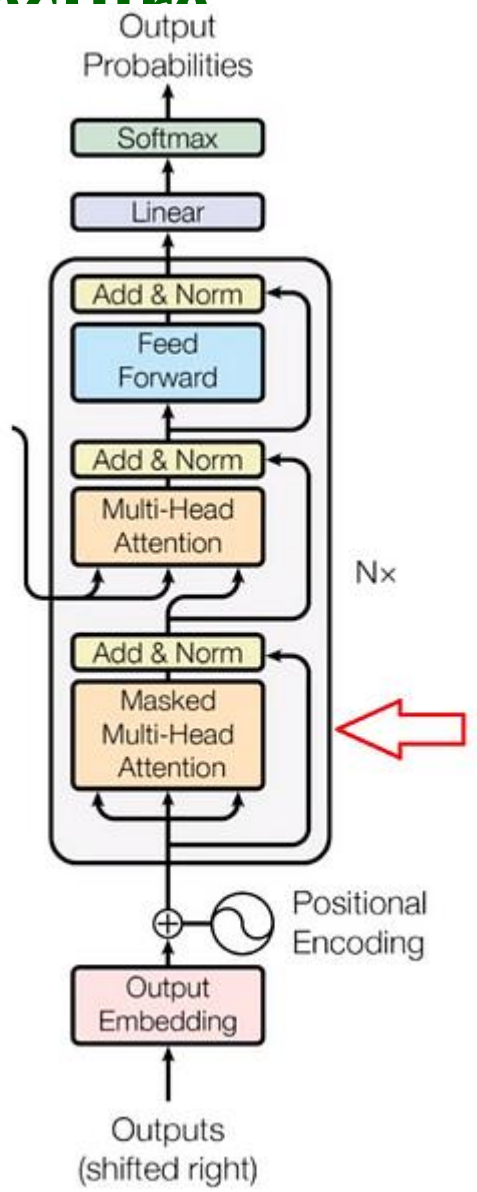
LLM: Transformers: How Does it Works?

Decoder

Masked Multi-Head Attention (Self-Attention)

"**Masked**": Learning mechanism in artificial neural networks (ANN):

- Give an English word.
- First, the transformer translates the English word into the French version by itself using previous results.
- Then the transformer matches and compares the predicted French word with the actual correct French word that is fed into the decoder block.
- After comparing both, the transformer updates its matrix value, through which the transformer can learn the correct translation after several iterations.



source: arXiv:1706.03762

17: Generative AI: LLM: Transformers: Architecture

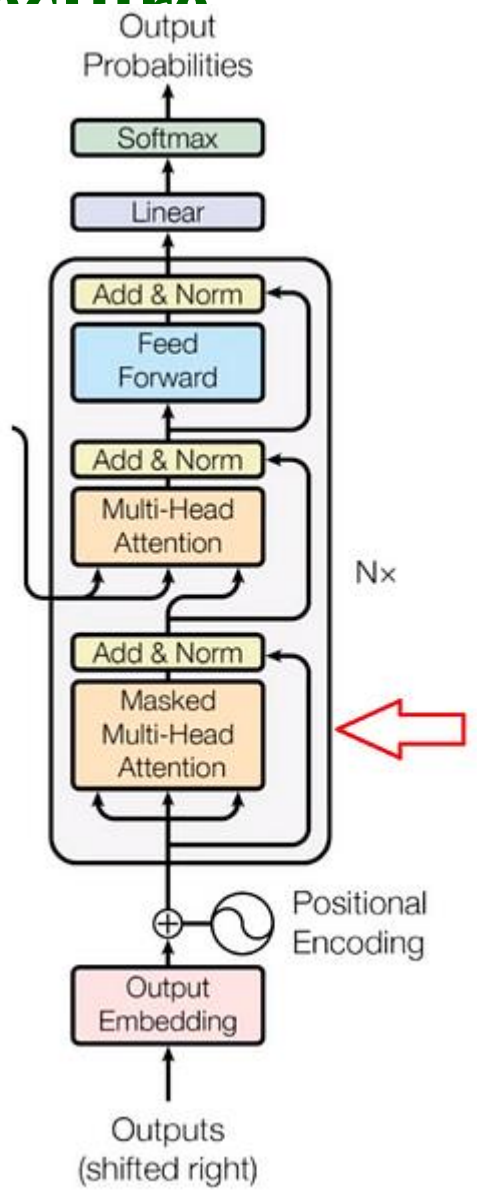
LLM: Transformers: How Does it Works?

Decoder

Masked Multi-Head Attention (Self-Attention)

"Masked": Learning mechanism in artificial neural networks (ANN):

- To teach the transformer and make it learn better: Hide or mask the next French word.
  - At first the transformer predicts the next word by itself using previous results, without knowing the real translated word.
    - Of course, it makes no sense if it is shown the next French word → Hide or mask the actual French word.



source: arXiv:1706.03762

18: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

Decoder

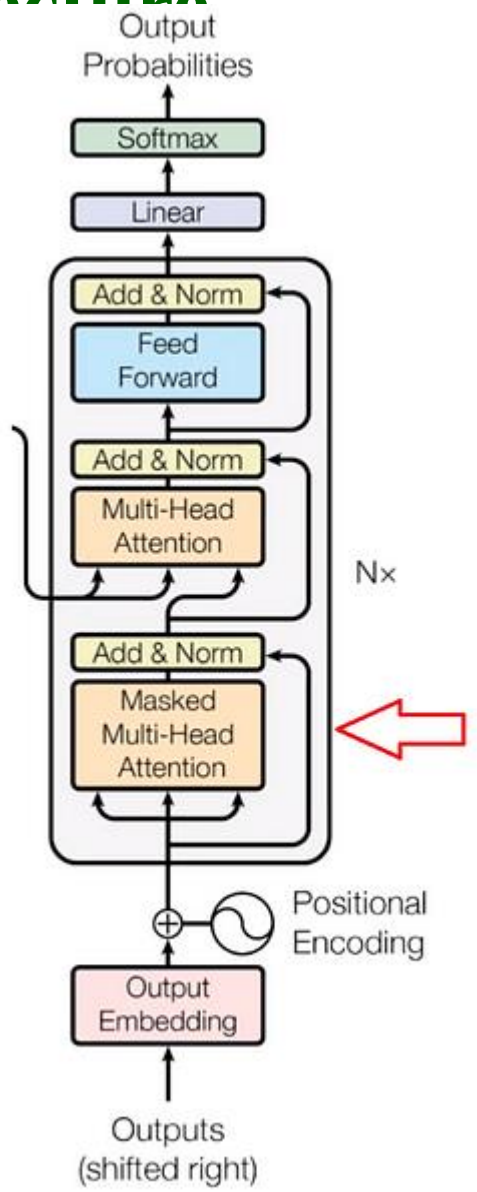
Masked Multi-Head Attention (Self-Attention)

"Masked": Learning mechanism in artificial neural networks (ANN):

ENCODER:
- It is possible to take any word from the English sentence.

DECODER:
- It is possible to only take the **previous** word of the **French** sentence.

Therefore, while performing the parallelization with matrix operation, it should ensure the matrix should **mask** the words appearing later by transforming them into o's so that the attention network (decoder) cannot use them.

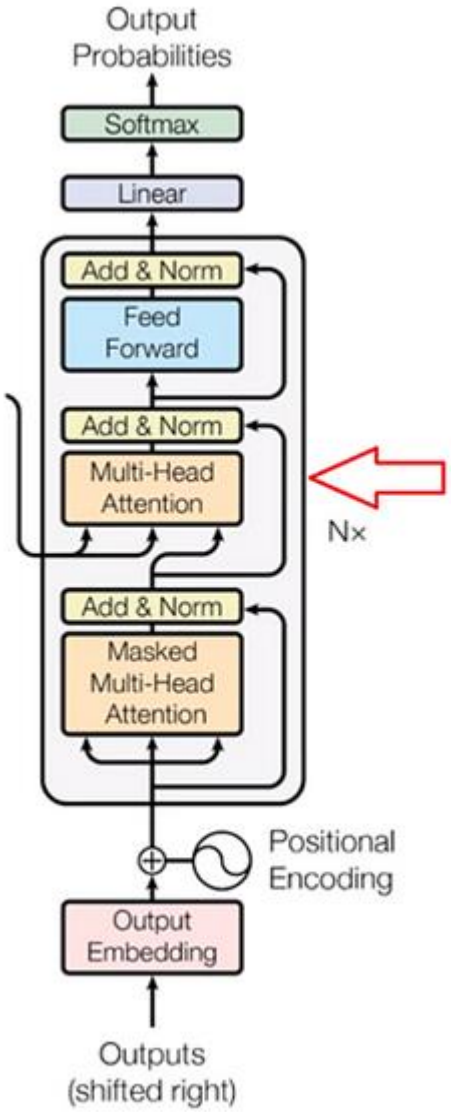


source: arXiv:1706.03762

19: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

Decoder



Decoder: Multi-Head Attention (Self-Attention)

a.k.a. **Encoder-Decoder Attention Block**:

- The resulting **attention** vectors from the previous **decoder** layer are passed into the **Decoder** Multi-Head Attention Block.
- The **attention** vectors from the **Encoder** Block are also passed into the **Decoder** Multi-Head Attention Block.
  - The results from the encoder block comes into the picture.
  - That's why it is called **Encoder-Decoder** Attention Block.

source: arXiv:1706.03762

20: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

Decoder

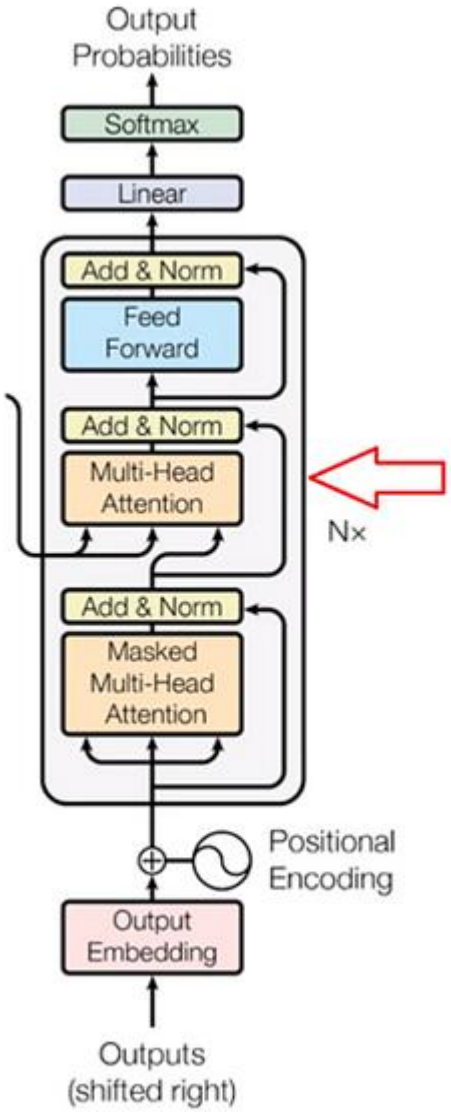
Decoder: Multi-Head Attention (Self-Attention)

a.k.a. **Encoder-Decoder Attention Block**:

- There is one vector of every word for each English and French sentence.
- This block (Multi-Head Attention in Decoder)
  - Do the mapping of English and French words
  - Find out the relation between English and French words.

Therefore, Multi-Head Attention in Decoder:

- Where the main English to French word mapping happens.



source: arXiv:1706.03762

21: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

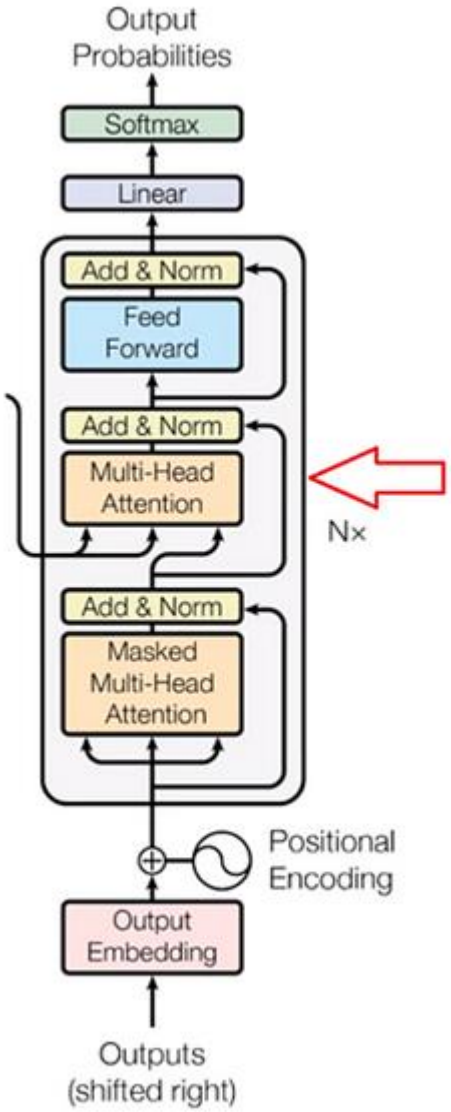
Decoder

Decoder: Multi-Head Attention (Self-Attention)

a.k.a. **Encoder-Decoder Attention Block**:

OUTPUTS:

- **Attention** vectors for **every** word in English and French sentences.
- **Each** vector represents the **relationship** with other words in both languages.



source: arXiv:1706.03762

22: Generative AI: LLM: Transformers: Architecture

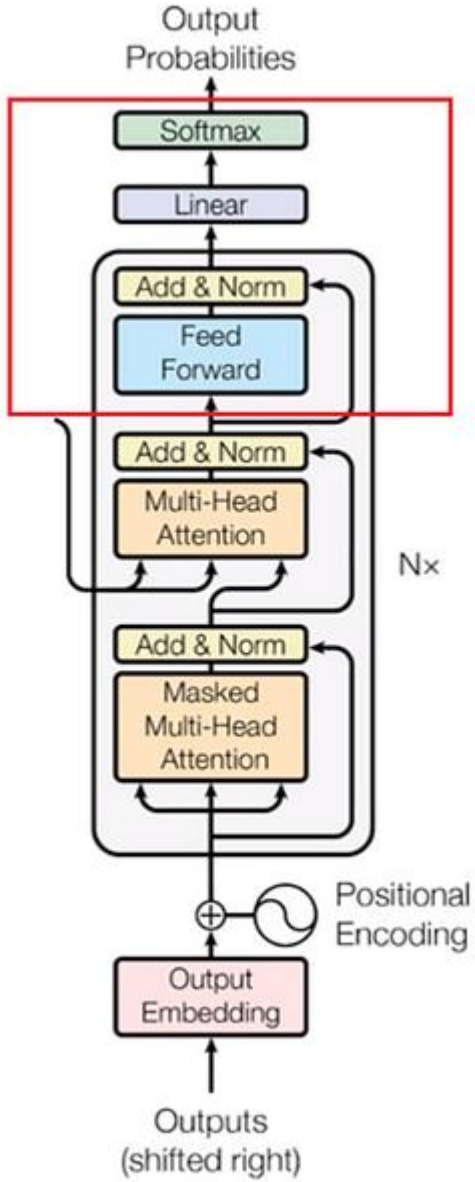
LLM: Transformers: How Does it Works?

Decoder

Decoder: Feed Forward Neural Networks (FFNN): Linear Layers

- **Each** attention vector is passed into a **feed-forward** neural network.
  - FFNN transforms output vectors into some form which is easily acceptable by another decoder block or a linear layer.

- A **Linear Layer**, a.k.a. Flattened Layer, is a feed-forward layer belonging to the feed forward neural network.
  - It is used to expand the dimensions into numbers of words in the French language after translation.



source: arXiv:1706.03762

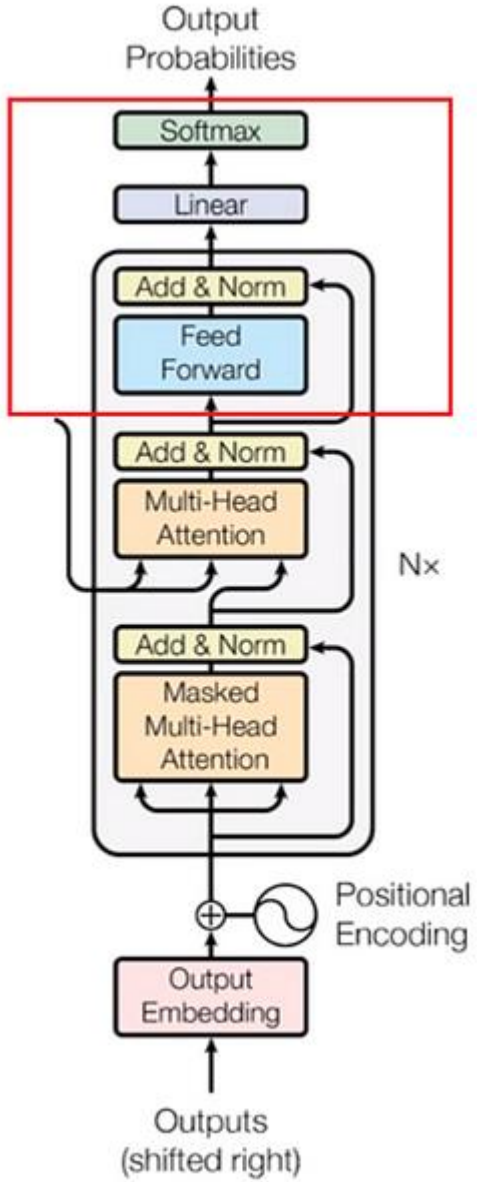
23: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

Decoder

Decoder: Softmax Layer

- The **outputs** from the last Feed Forward Linear Layer are passed through a **Softmax** Layer,
  - Transform the input into a **probability** distribution.
- The **final outputs** are resulting words produced with the **highest** probability after translation.



source: arXiv:1706.03762