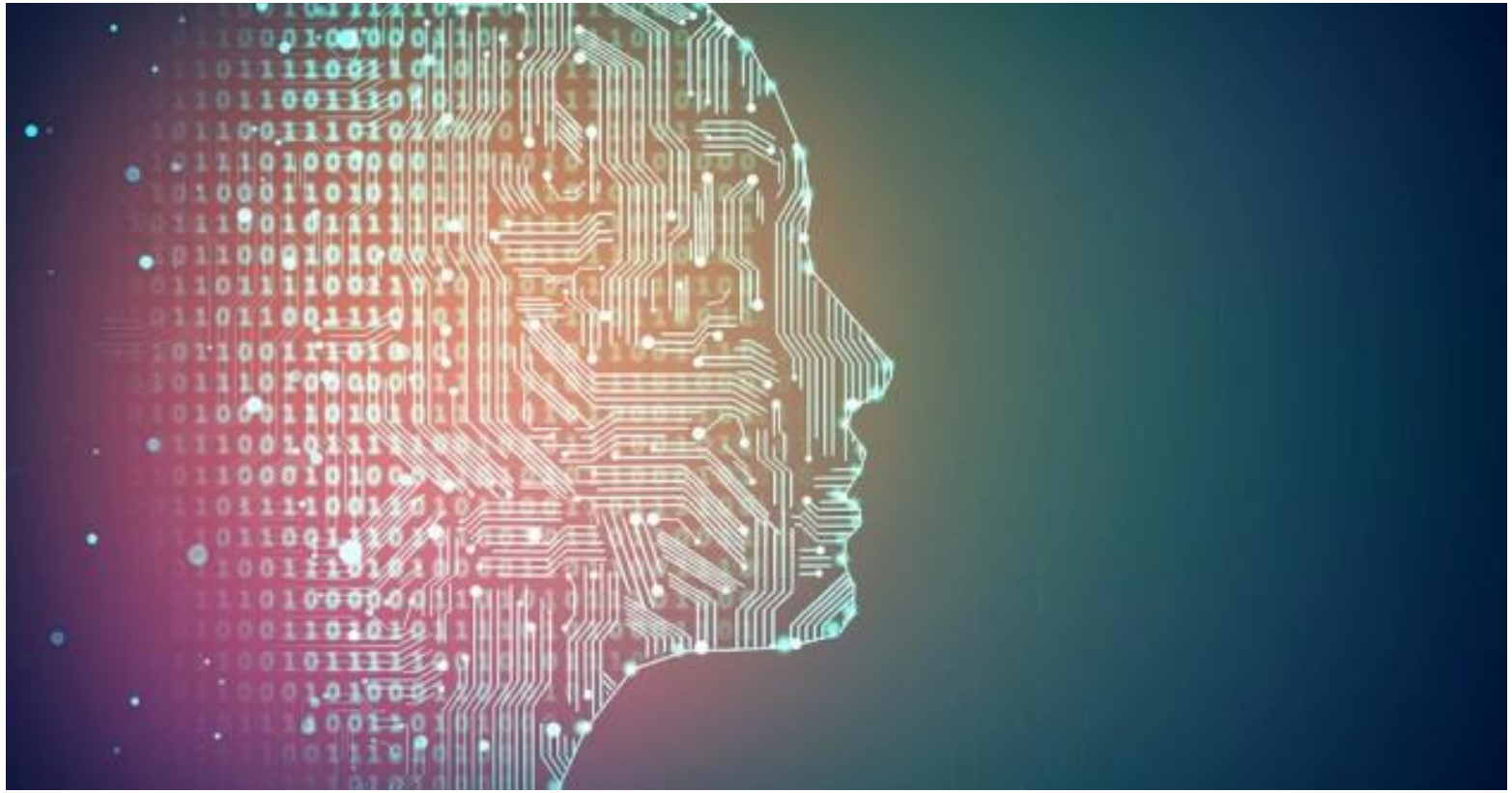
Generative AI: Transformers: Architecture
PART I: Overview

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

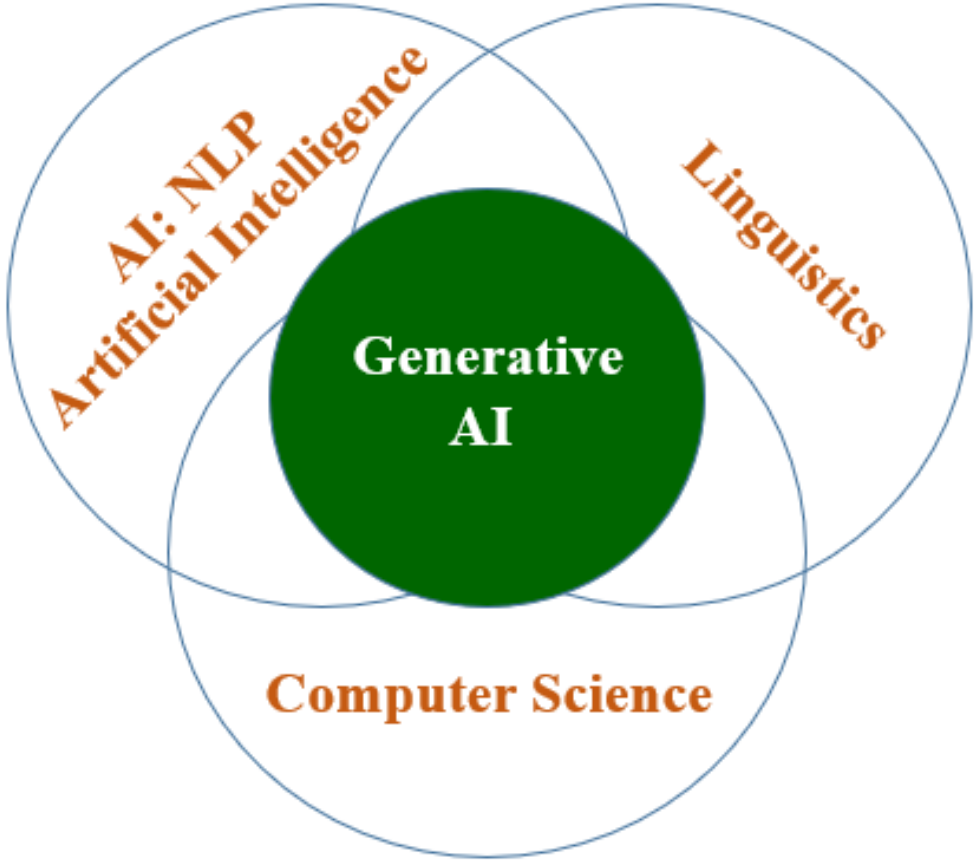
What is It?

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

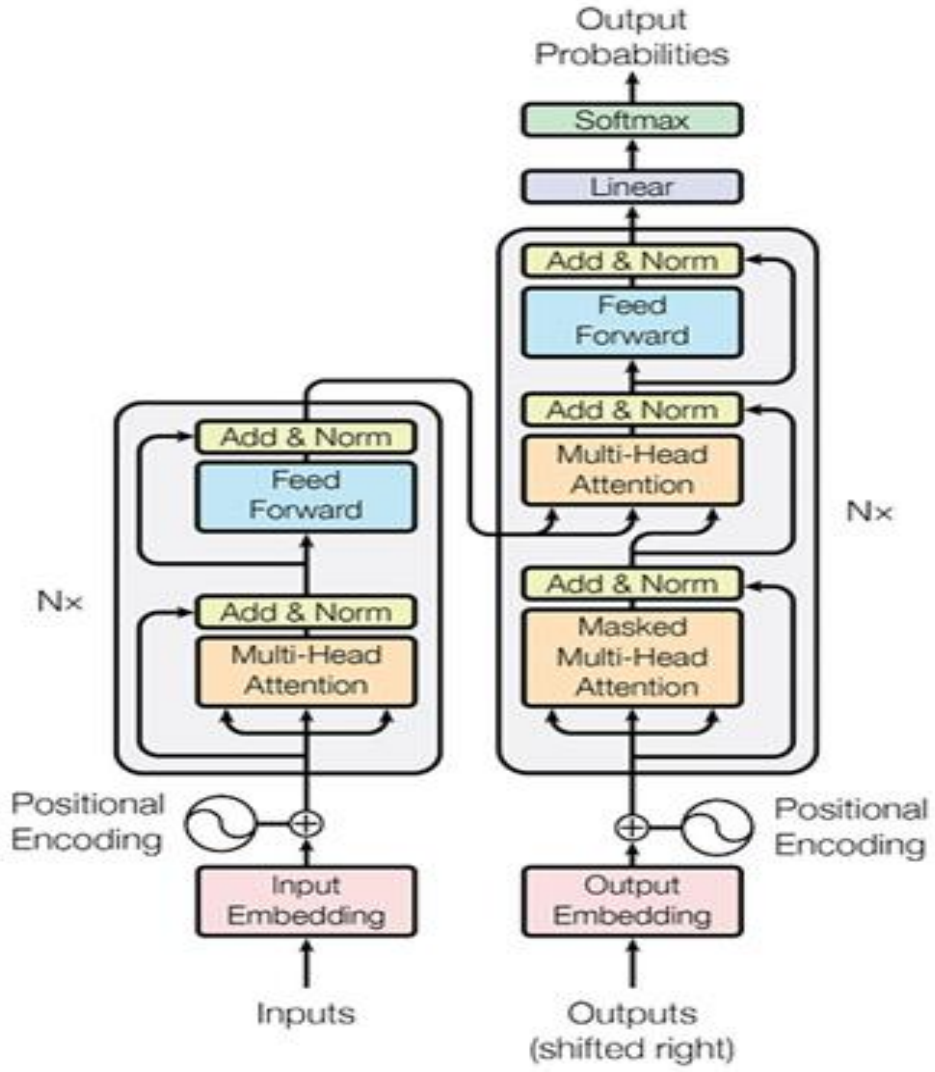
“**Attention Is All You Need!**”

***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

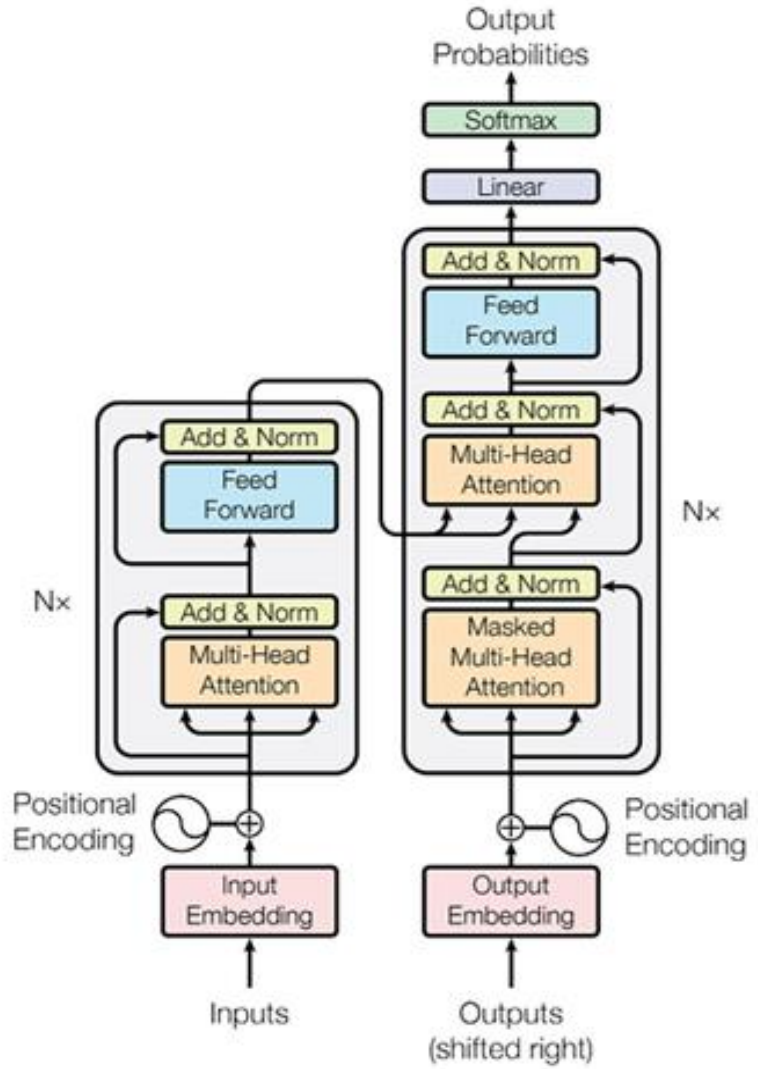
LLM: Transformers: Architecture

Encoder-Decoder

Foundational Revolutionary Core Features:

- **Parallelize Computations**: Transformers discard the sequential processing paradigm of earlier recurrent neural networks (RNNs) in favor of parallel computation facilitated by attention mechanisms. This dramatically speeds up training and inference.

- **Handle Long-Range Dependencies**: RNNs tend to struggle with long-range dependencies within sequences. The Transformer, through its **attention mechanism**, directly calculates relationships between any two elements in the sequence, regardless of distance, fostering a better grasp of context.



The Transformer - model architecture.

source: arXiv:1706.03762

11: Generative AI: LLM: Transformers: Architecture

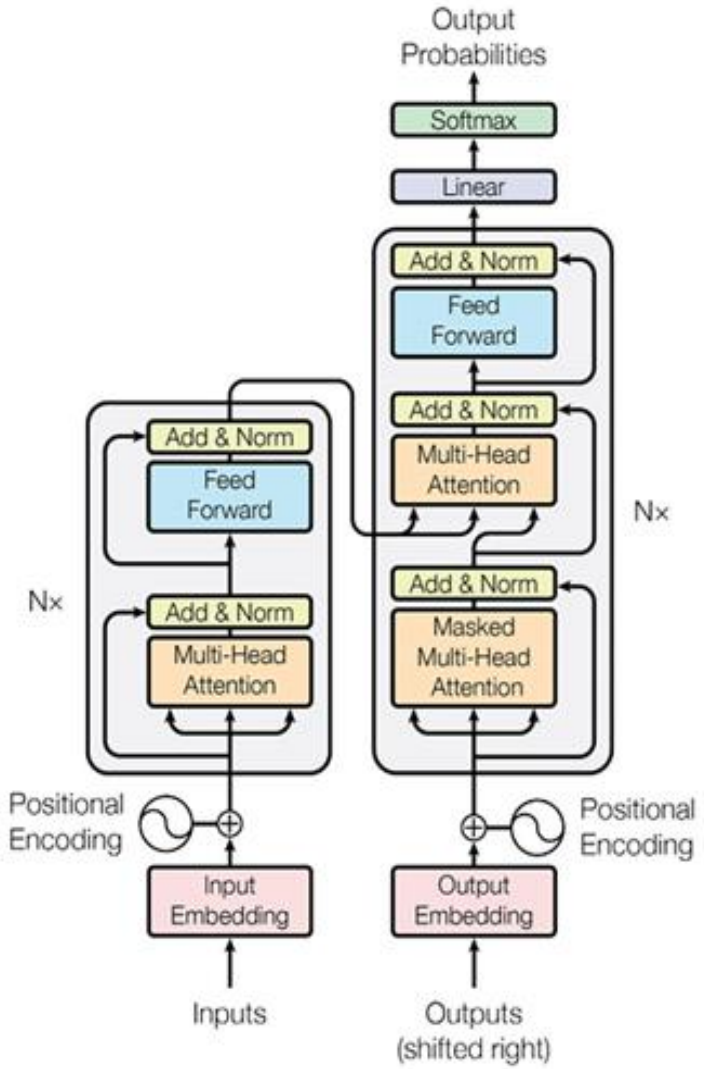
LLM: Transformers: Architecture

Encoder-Decoder

Revolutionary Architecture of the Transformer:

Encoder-Decoder Structure: The Transformer generally employs an encoder-decoder structure.

- **Encoder**: The encoder's job is to process the input sequence (for example, a sentence of text) and generate a rich, contextualized representation of that input.
- **Decoder**: The decoder takes the encoder's output and, in an auto-regressive manner (one step at a time), generates the target sequence (for example, a translation of the sentence).



The Transformer - model architecture.

source: arXiv:1706.03762

12: Generative AI: LLM: Transformers: Architecture

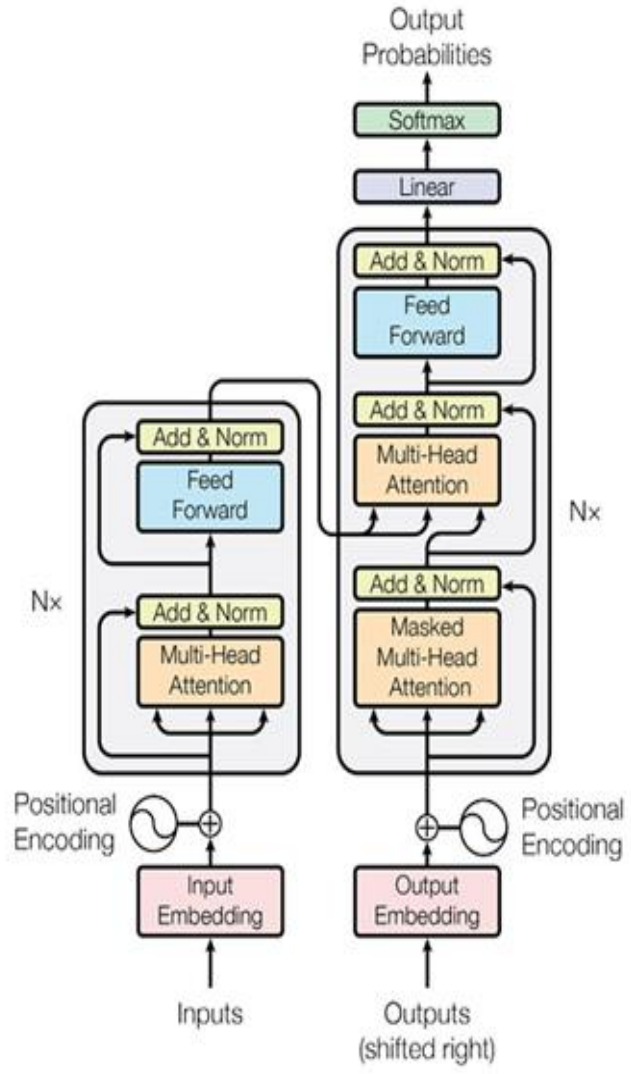
LLM: Transformers: Architecture

Encoder-Decoder

Revolutionary Architecture of the Transformer :

Multi-Head Attention Mechanism: Heart and Mind of the Transformer. It works as follows:

- **Queries, Keys, and Values**: For each word in the input sequence, the Transformer creates three vectors: a **query vector (Q)**, **a key vector (K)**, and a **value vector (V)**. These vectors are learned representations of the word itself.
- **Scaled Dot-Product Attention**: The **query vector** of a word is compared to the key vectors of all words in the sequence, including itself. Similarity scores (via dot product) are calculated, and softmax is applied to normalize these scores into probabilities. These represent how much **"attention"** a word should **pay** to each **other** word.
- **Weighted Sum**: Finally, the value vectors are multiplied by their corresponding attention scores and summed up. This becomes the **attention-enhanced output representation** of the **current** word.
- **Multi-Head**: Multiple sets of these Q, K, V transformations happen in parallel (forming "heads"), creating multiple diverse representations of each word. These representations are concatenated and linearly projected to yield a final unified representation for each word.



The Transformer - model architecture.

source: arXiv:1706.03762

13: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: Architecture

Encoder-Decoder

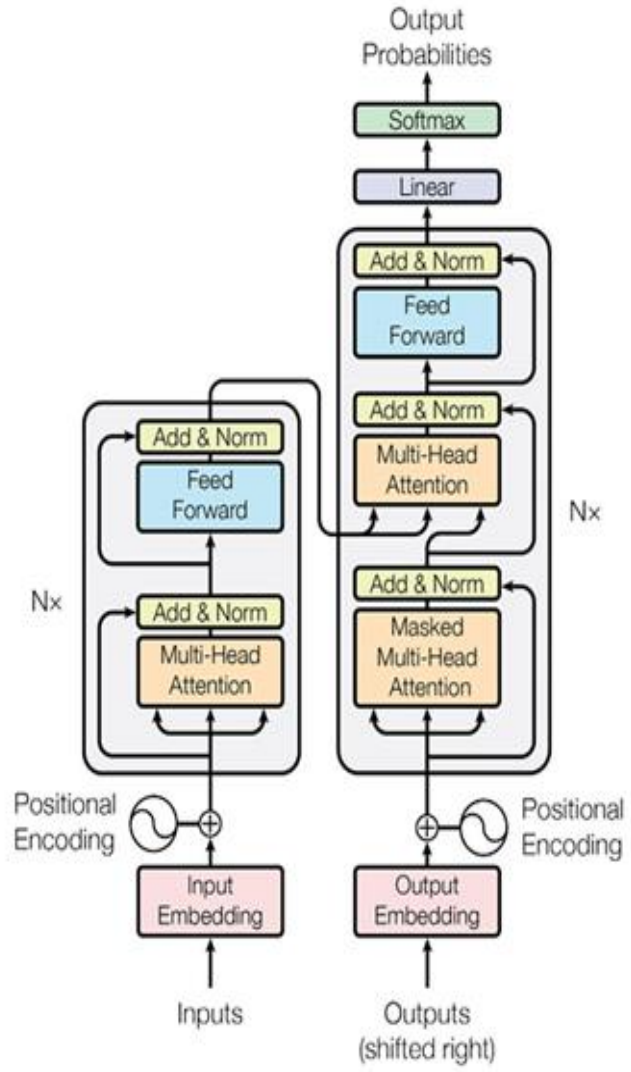
Revolutionary Architecture of the Transformer :

Attention – The Game Changer:

- The self-attention mechanism allows the model to weigh the importance of different words in an input sequence.
- Thereby, the transformer can effectively understand the interconnected nuances of language that can span further distances within a sentence or paragraph.

Multi-Head Self-Attention: **Where the magic happens:**

- Self-attention allows each word within the input sequence to "attend" to all other words and determine how relevant each is in providing context.
- Multi-head attention performs this calculation multiple times in parallel, with each 'head' learning to focus on different aspects of the relationships between words.
- This multi-faceted approach enriches the representation of each word.



The Transformer - model architecture.

source: arXiv:1706.03762

14: Generative AI: LLM: Transformers: Architecture

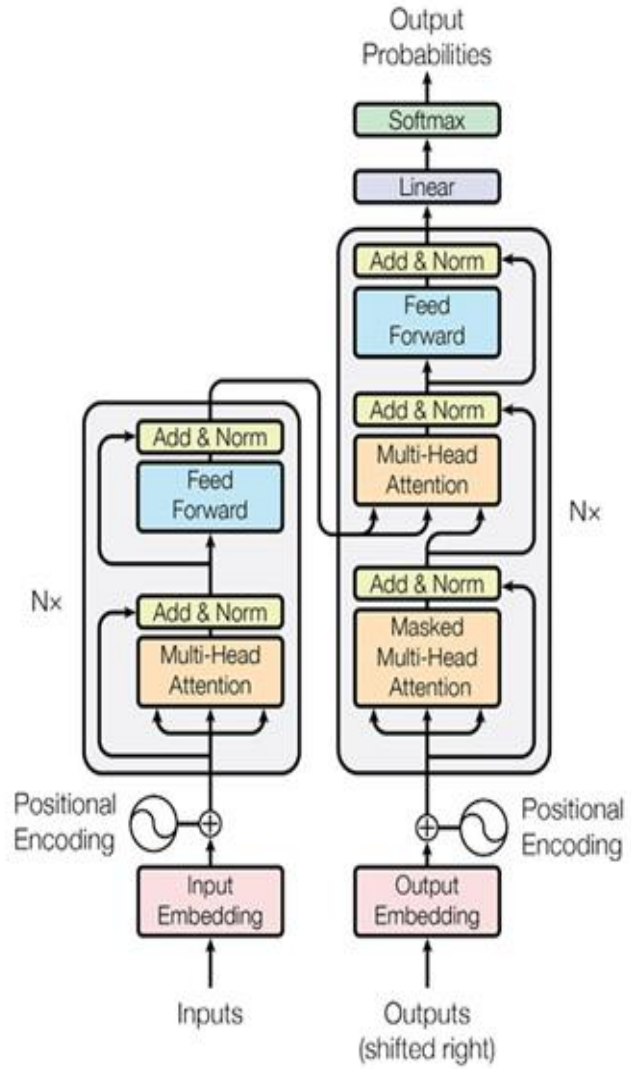
LLM: Transformers: Architecture

Encoder-Decoder

Revolutionary Architecture of the Transformer :

Positional Encoding: Because Transformers don't process sequences recurrently, the position of words within a sequence needs to be explicitly encoded. The Transformer injects positional information through vectors called positional encodings, with unique representations for each position in the sequence.

Feed-Forward Networks: Both the encoder and decoder layers include point-wise, fully-connected feed-forward neural networks. These networks add further non-linear transformations to the attention-based output, enhancing the model's expressiveness.



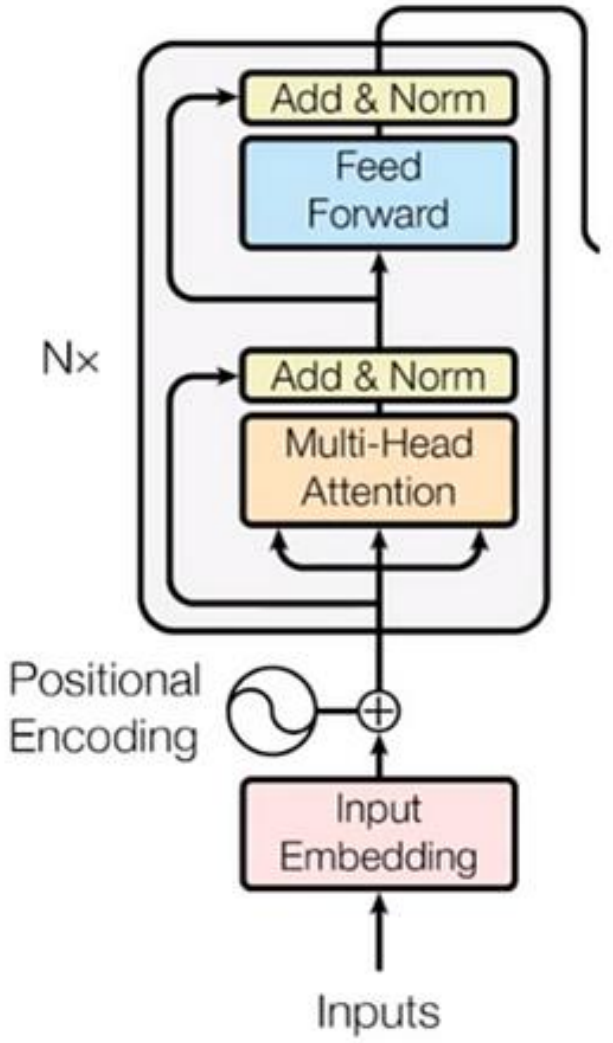
The Transformer - model architecture.

source: arXiv:1706.03762

15: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: Architecture

Encoder



source: arXiv:1706.03762

16: Generative AI: LLM: Transformers: Architecture

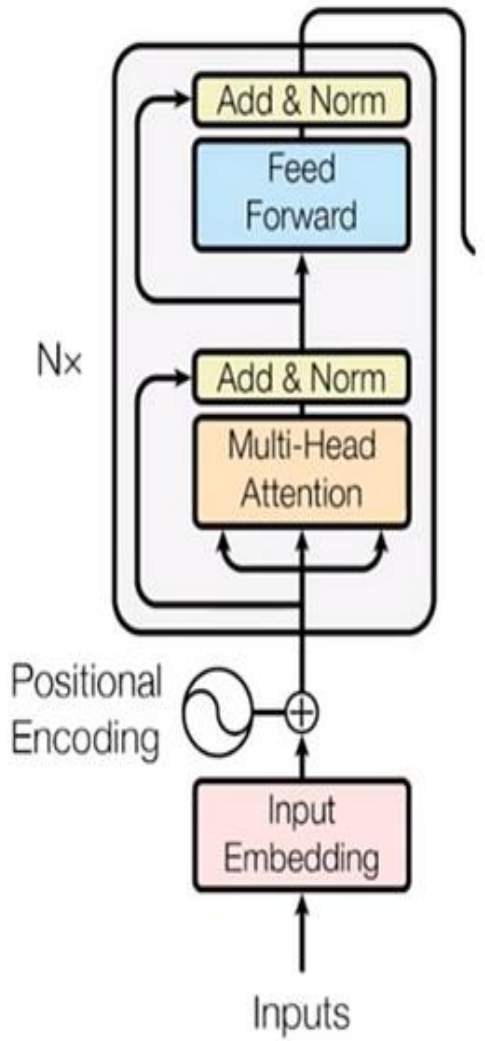
LLM: Transformers: Architecture

Encoder

Main Layers of the Transformer: Encoder

- **Input Embeddings**: The input words are converted into word embeddings (dense vector representations).
- **Multi-Head Attention (Self-Attention)**: Words "attend" to each other within the input sequence.
- **Layer Normalization**: Help stabilization during training.
- **Feed-Forward Network**: Further refine the representations of words
- **Residual Connections**: Shortcut connections add the input to the output of each sub-layer, making training more manageable.

A Transformer has **stacked encoder layers** (usually 6 or more), with each layer **refining** the representations of words or contents from the previous.

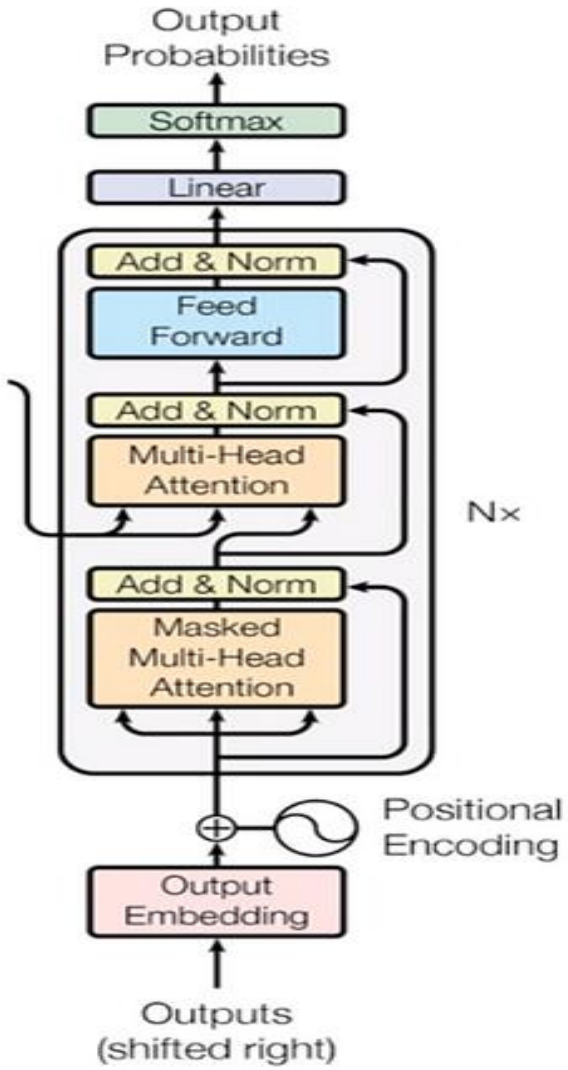


source: arXiv:1706.03762

17: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: Architecture

Decoder



source: arXiv:1706.03762

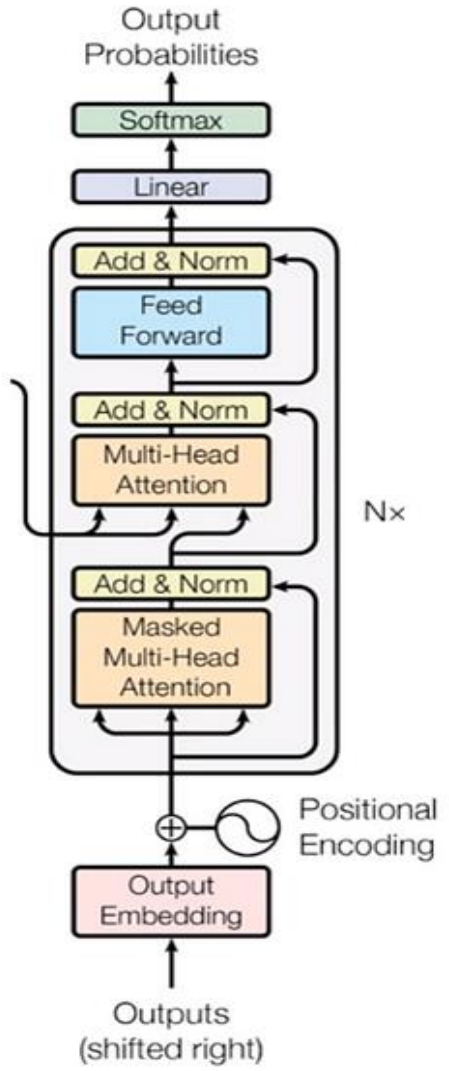
18: Generative AI: LLM: Transformers: Architecture

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