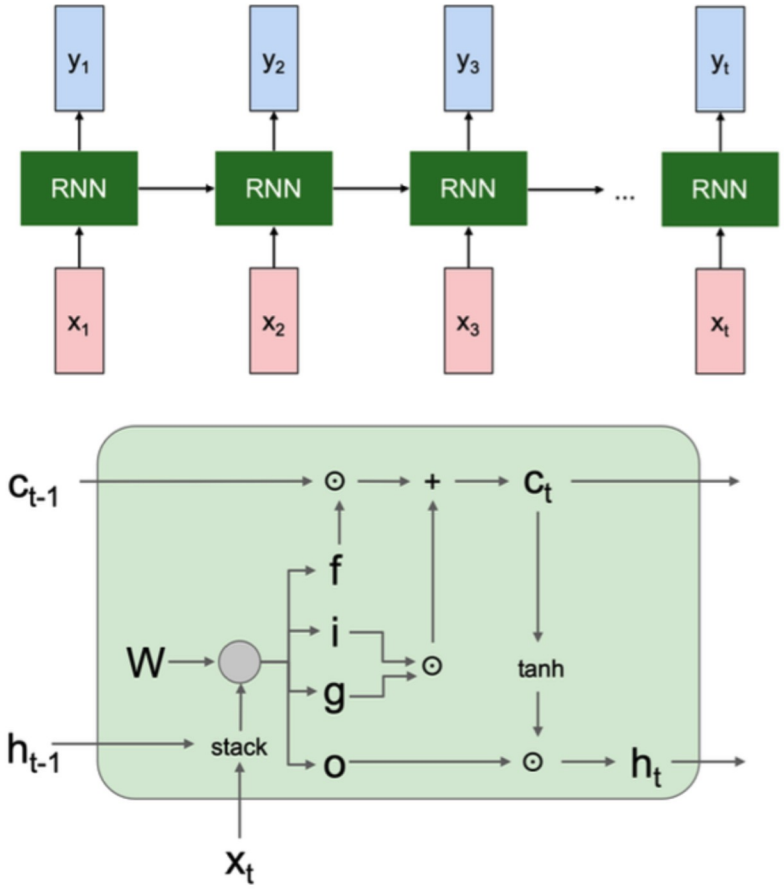
WALC 2024
Applied AI

Large Language Models (LLMs) at the Edge
Transformers Introduction

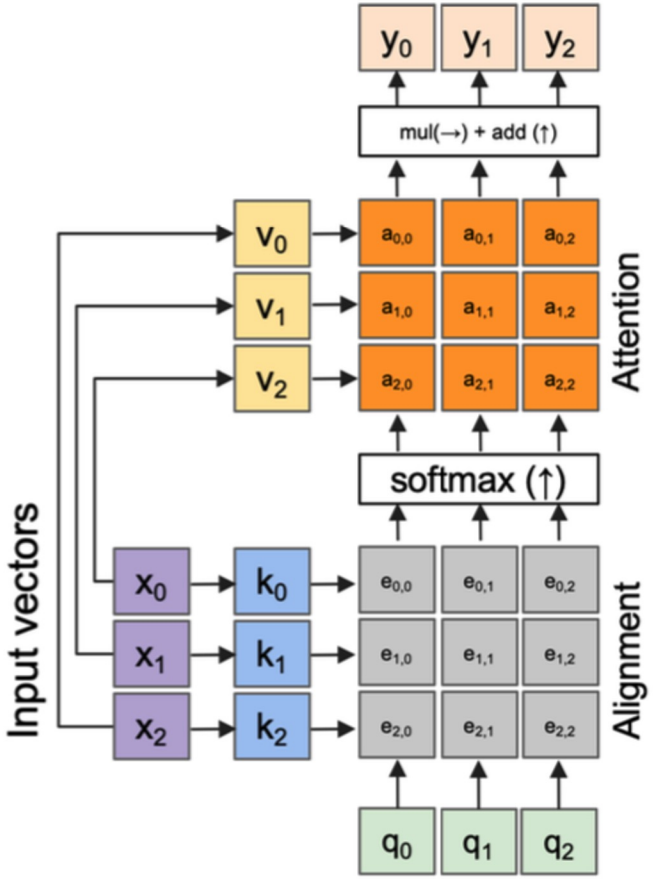
Prof. Marcelo J. Rovai
rovai@unifei.edu.br

UNIFEI - Federal University of Itajuba, Brazil
TinyML4D Academic Network Co-Chair

Recap: Models Beyond DNN and CNN



Recurrent neural network



Attention mechanism / Transformers

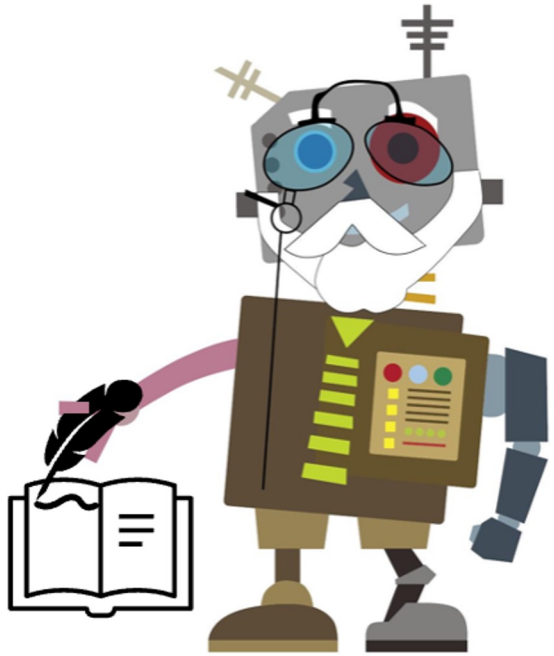
Machado de Assis Bot with RNN - GRU

The *robot writer* model is a **Recurrent Neural network (RNN/GRU)**. The model, with 4M parameters, was trained with a **150-characters sequence** from seven of his books: *Memorias Posthumas de Braz Cubas, Dom Casmurro, Quincas Borba, Papeis Avulsos, A Mão e a Luva, Esaú e Jacob*, and *Memorial de Ayres*.

Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(1, 150, 256)	29,952
gru (GRU)	(1, 150, 1024)	3,938,304
dense (Dense)	(1, 150, 117)	119,925

Total params: 4,088,181 (15.60 MB)
Trainable params: 4,088,181 (15.60 MB)
Non-trainable params: 0 (0.00 B)



A LUVA DE CASMURRO II

A missa do coupé e um presente e o governo devia cazar logo no papel, a morte do autor, e todos os seus considerados de alegria. Era um espirito de vinte e cinco annos, e eu não estou alguns passos no cerebro, como de outra cousa. Deus me disse:

--Não digo que não. Se eu tivesse a intenção de um probosito. Palha acudiu a mulher, não havia nada. A noite vinha tambem para o seminario, tinha o aspecto do partido recto e de restaurar a minha mãe e do pae, pela primeira vez, a menor destinada a dispensar o chapéo, esperou que não vinhas com as suas mãos de creanças. A manhã della chegasse a baroneza e a maneira desta divida. Parece que é casada.

--Está bom, perdoa-lhe de todos os lados, a vida de que o comprar para o meu quarto de hora, e contavam com o fim de a anterior, e, a parede pouco tempo a alma de pessoas que definitivamente lhe interessam a menos para mim. De quando em quando, esses dous annos de conversação para o fim de deixar nenhuma pessoa que se dispersasse; mas não falo de uma cousa nem lhe pedia com a mão tremula, como se ella quizesse. Eu, apertando-lhe a mão, aliás o principio do governo, a proposito disso, com a desattenção de Estevão, e eu começou a aborrecel-o, e a solidão podia ser melhor, e a sympathia colloca da mãe, e não se sabe calar o enterro no meio do lagem, o que iam-se apanhados no chão, e para a mulher, não tendo visto, nem a mesma cousa.

Generative AI (GenAI)

Generative AI is an artificial intelligence system capable of creating new, original content across various mediums such as text, images, audio, and video. These systems learn patterns from existing data and use that knowledge to generate novel outputs that didn't previously exist.

Large Language Models (LLMs), Small Language Models (SLMs), and multimodal models can all be considered types of GenAI when used for generative tasks.

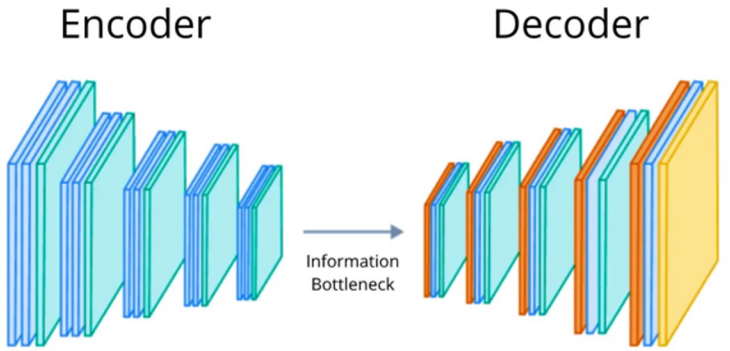
LLM / SLM

Large Language Model / Small Language Models

LLMs are **specialized deep learning models designed to understand and generate human language**, used for tasks like translation, summarization, and generating human-like text responses. SLMs are the same, but use a simpler, less resource-intensive approach (smaller in size).

Deep Learning models (or artificial neural networks)

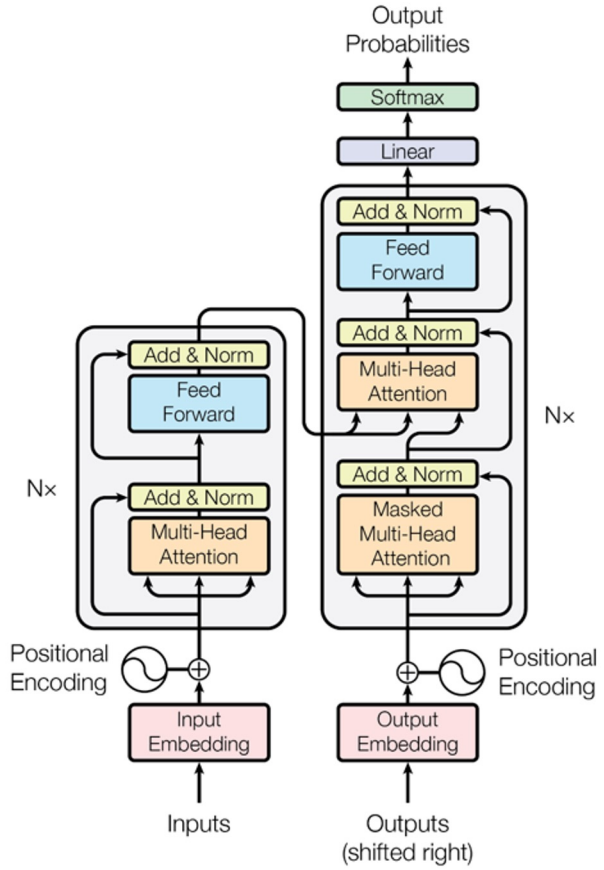
- **Autoencoders**: Used primarily for unsupervised learning tasks such as dimensionality reduction and feature extraction, autoencoders learn to compress data from the input layer into a shorter code and then reconstruct the output from this representation.



- **Transformer Models**: Highly effective in handling sequences, transformers use mechanisms like self-attention to weigh the importance of different words in a sentence, regardless of their position. The Transformer architecture, while innovative, can be seen as a derivative of earlier deep learning models, particularly those based on the concept of sequence modeling. However, the most direct lineage can be traced to the sequence-to-sequence (seq2seq) models that utilize **encoder-decoder** architectures. These earlier seq2seq models were often built using **recurrent neural networks (RNNs)** or their more advanced variants like **LSTMs (Long Short-Term Memory Networks) or GRUs (Gated Recurrent Units).**

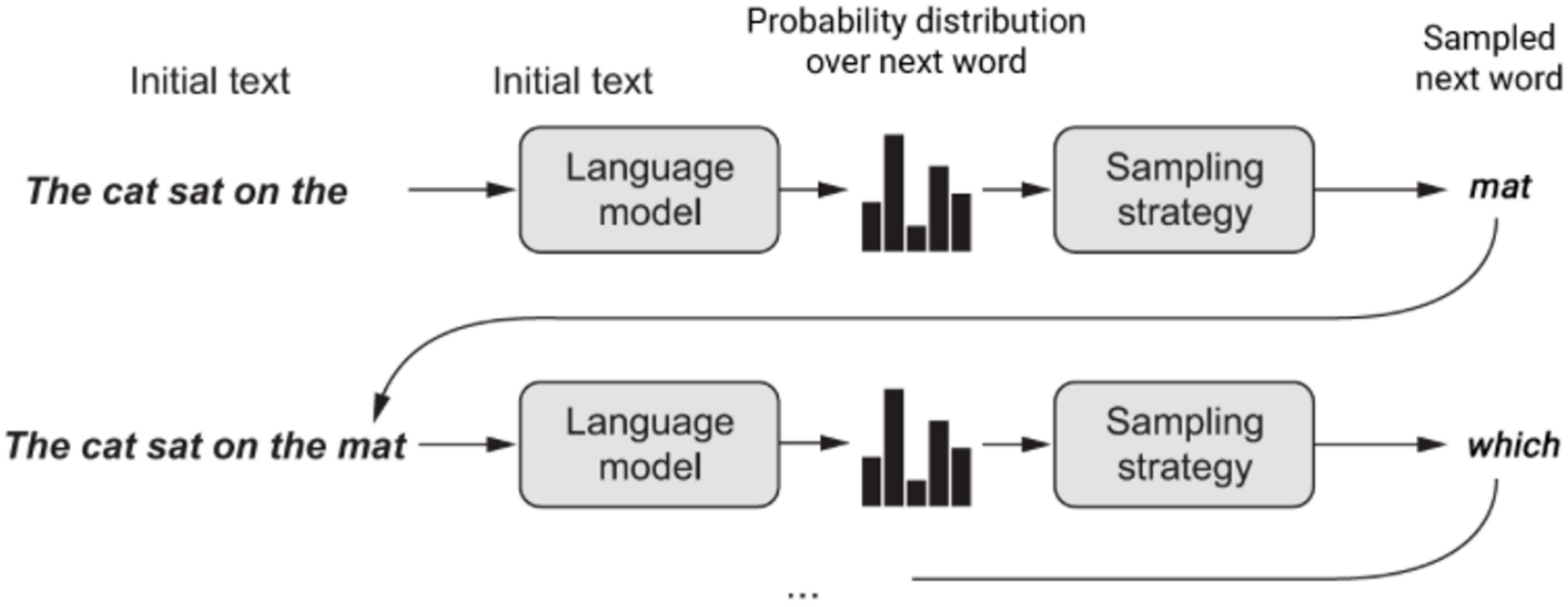
LLM/SLM – Large /Small Language Model

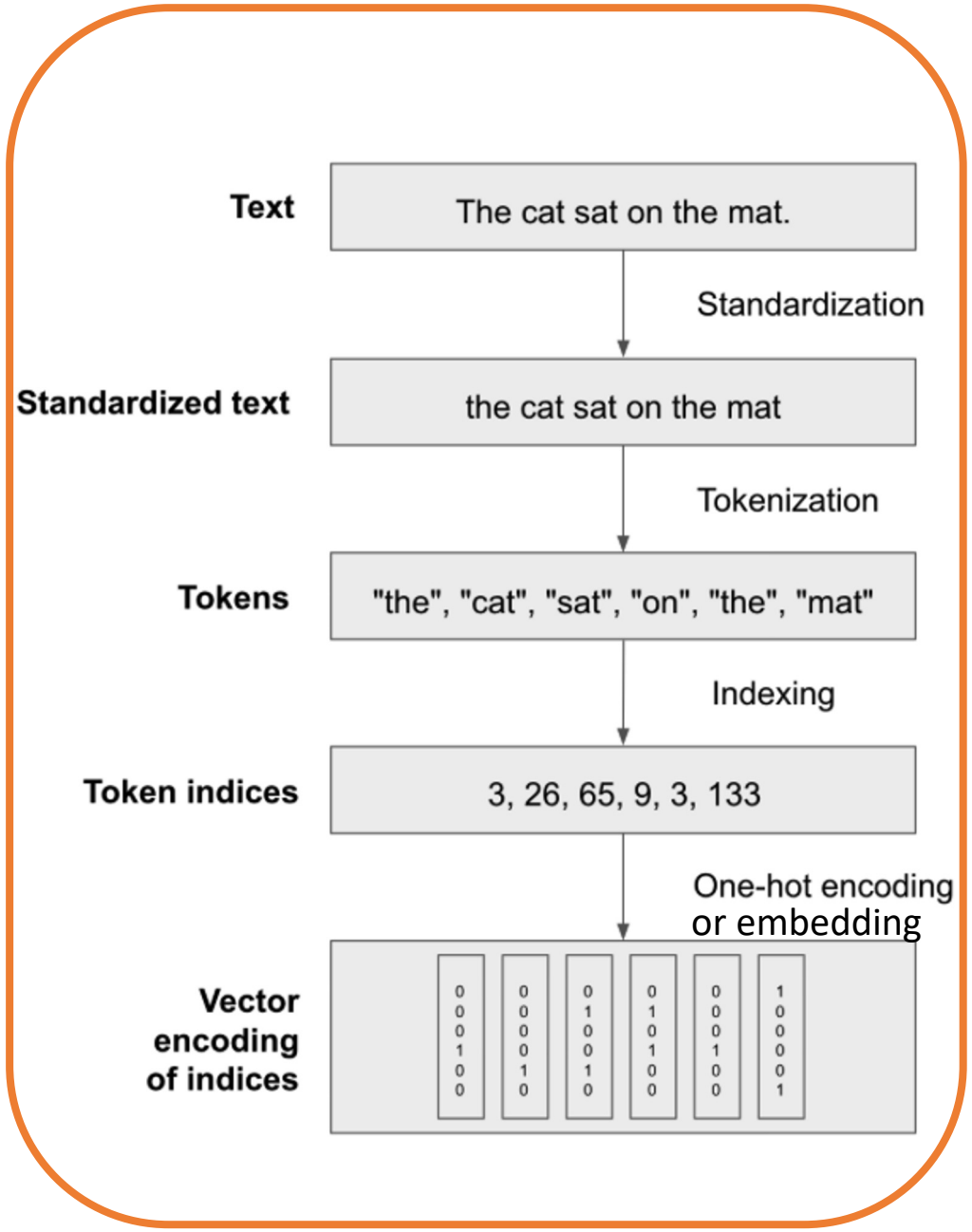
Large Language Models (LLMs) and SLMs are advanced neural networks based on the Transformer architecture that excel in understanding and generating human language. They represent a significant evolution from earlier sequence-based models like RNNs, which surpass them in handling long-range dependencies and parallel processing efficiency.

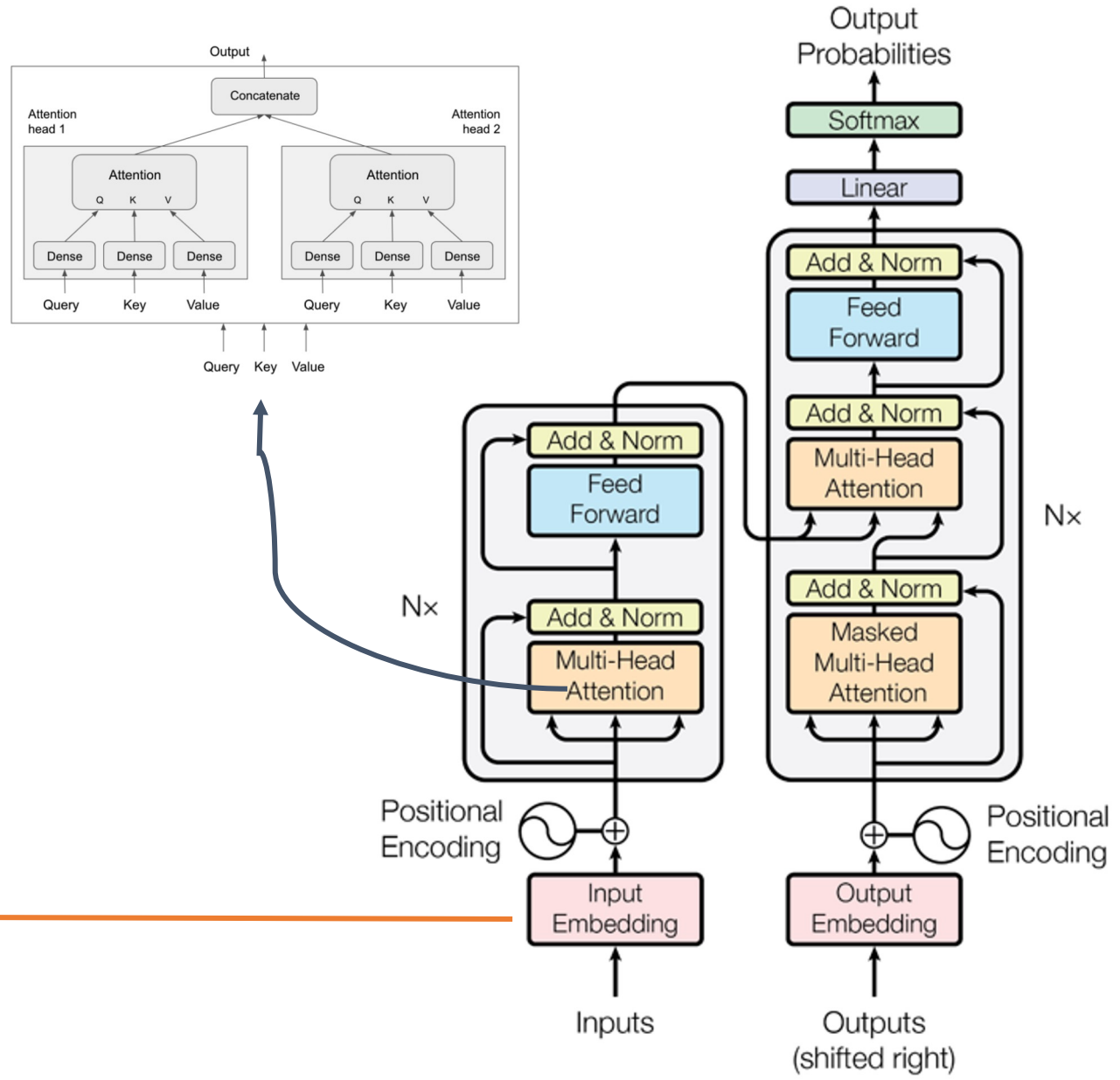


The Illustrated Transformer

LLM/SLM – Large /Small Language Model



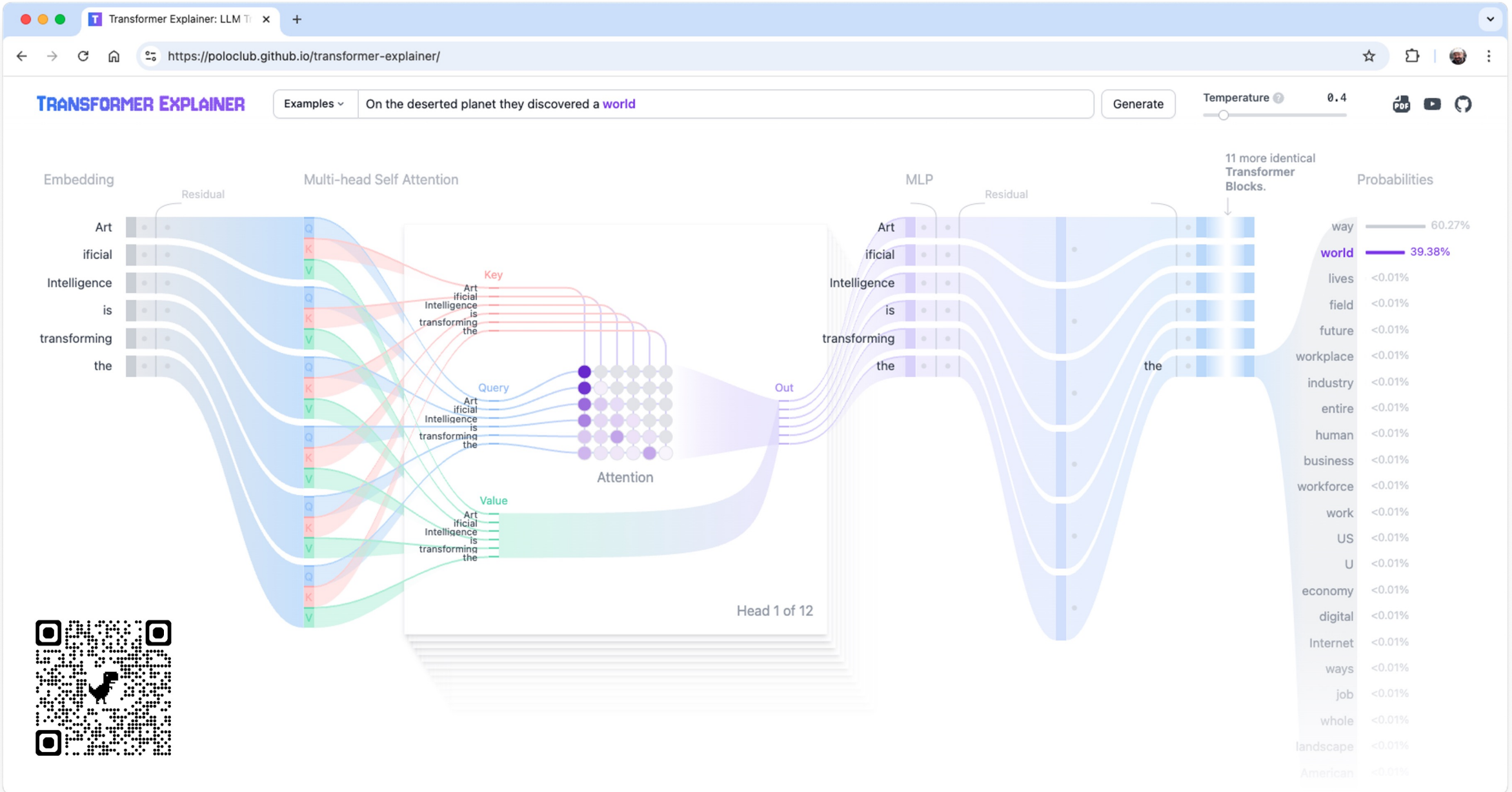




Encoder Decoder

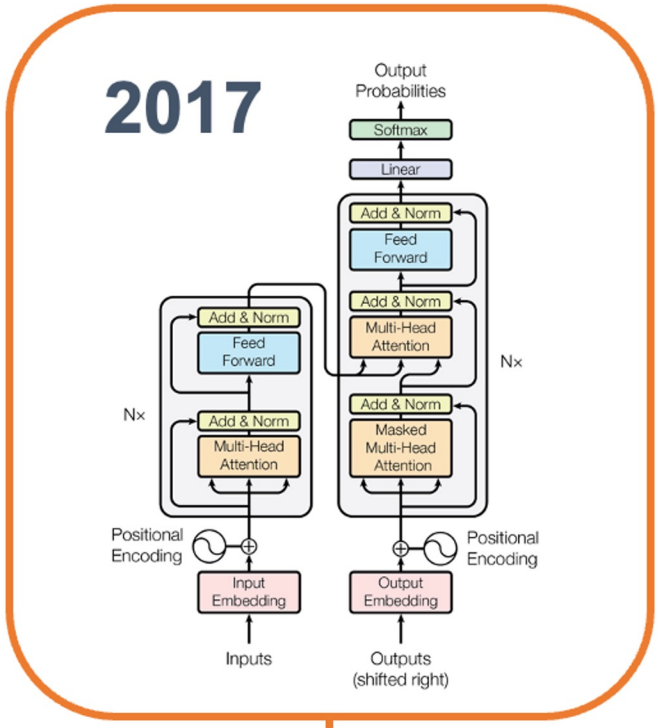
LLM/SLM – Large /Small Language Model

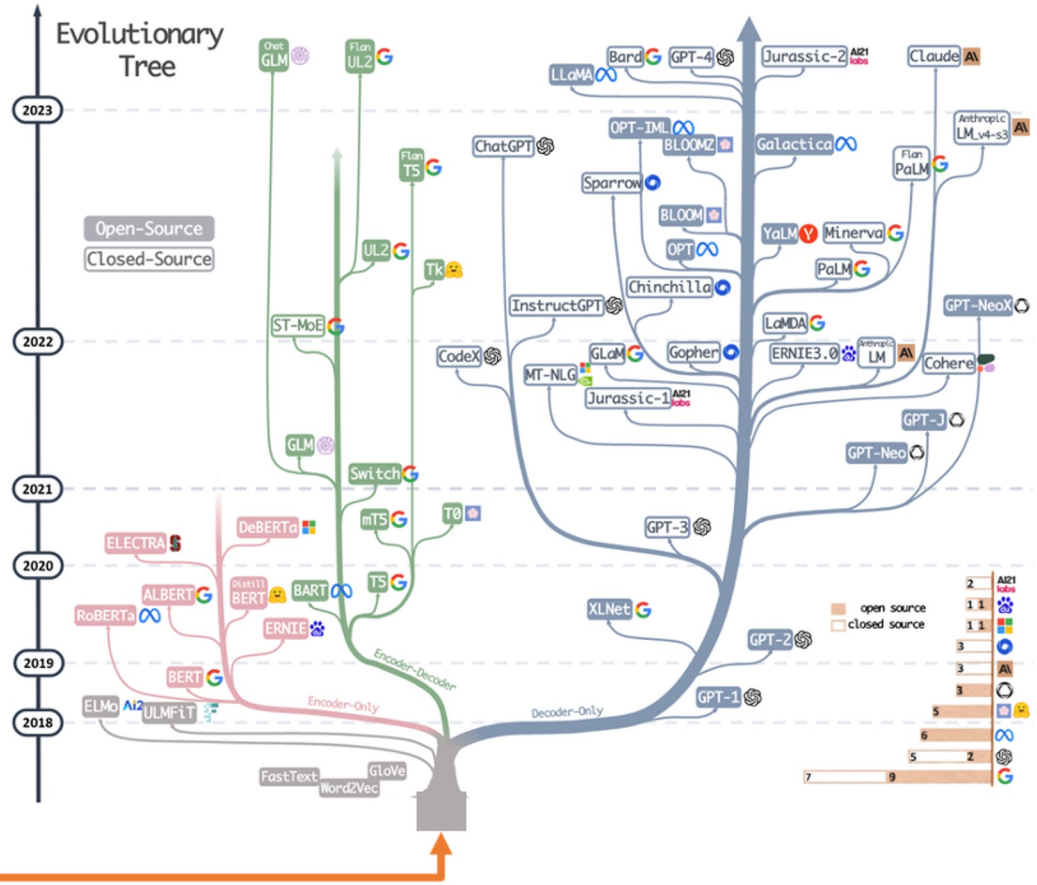
How large language models work



https://poloclub.github.io/transformer-explainer/

Transformers to LLMs and SLMs





Questions?


TINYML4D


WALC 2024
Online
11 al 15 de Noviembre
27°
Taller sobre
Tecnologías de Redes Internet
para América Latina y el Caribe
Fundación
EsLaRed
Organización General