

Lecture-4. Basic Intro To Transformer

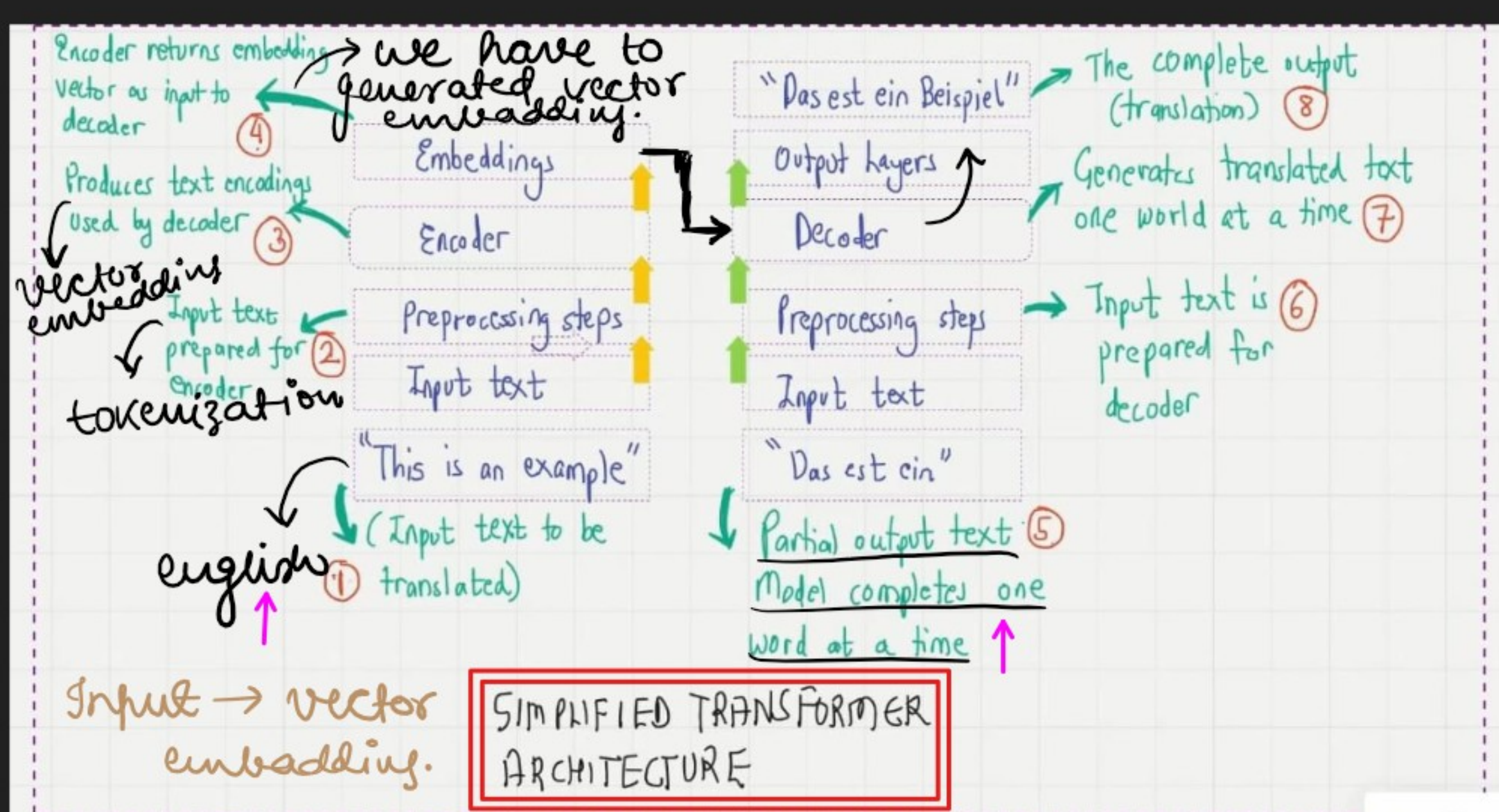
1. Most modern LLMs rely on the the transfer architecture → deep neural network architecture introduced in 2017 paper.

→ Attention is all you need.

Original Transformer = developed for machine translation.

Translating English text to German and French

2.



- Tokenize the data

Fine Tuning is fun for all!

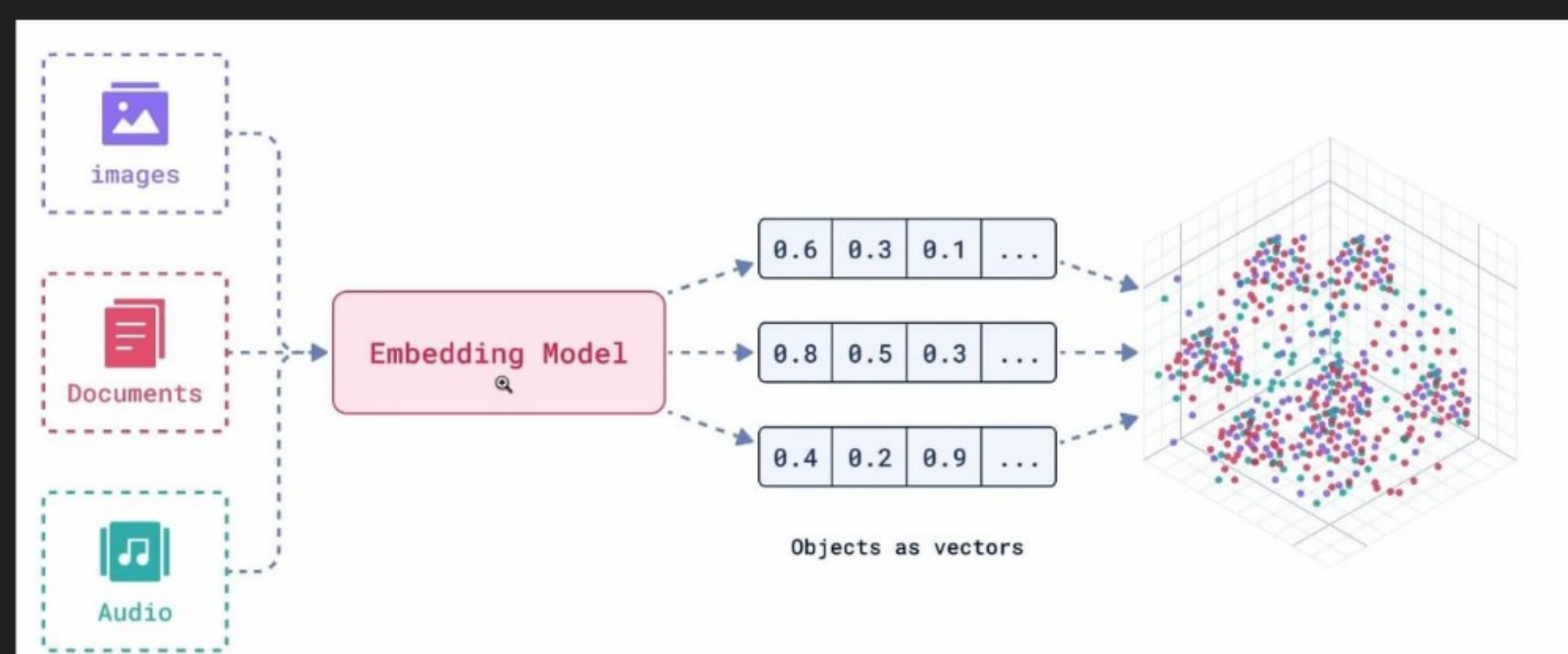
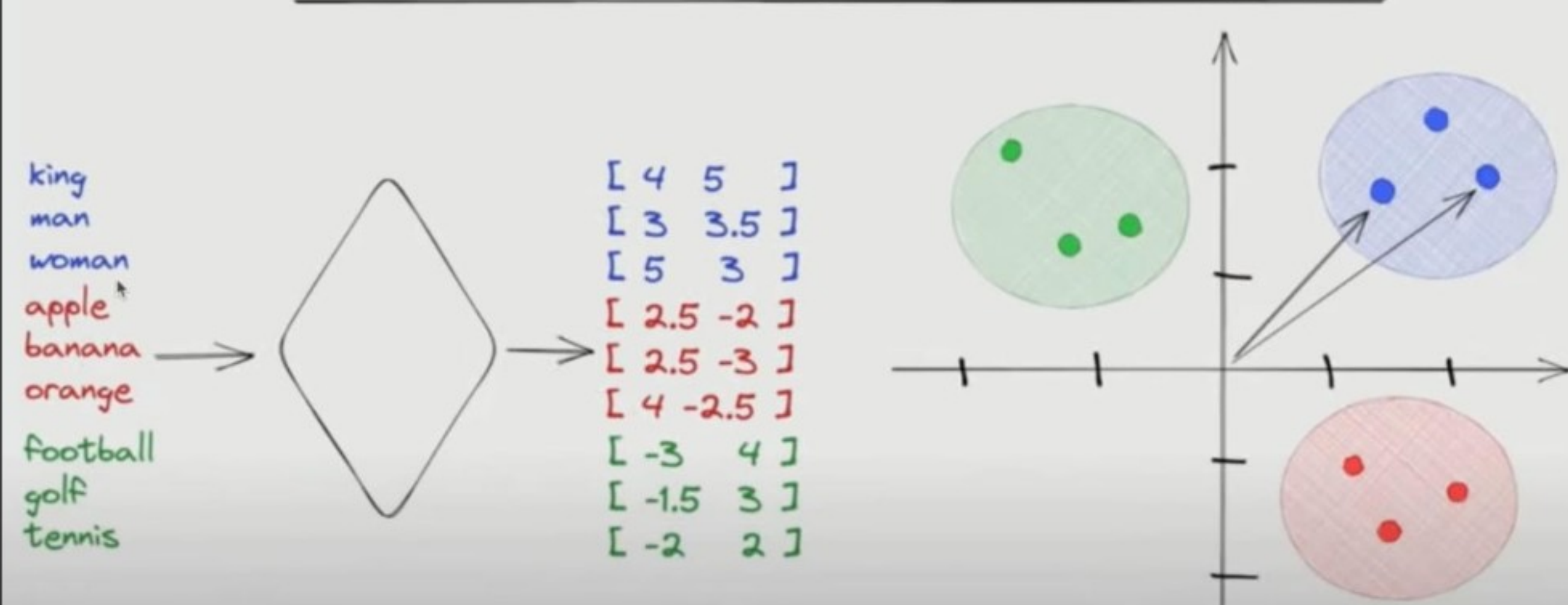
[34389, 13932, 278, 318, 1257, 329, 477, 0]

Encoding

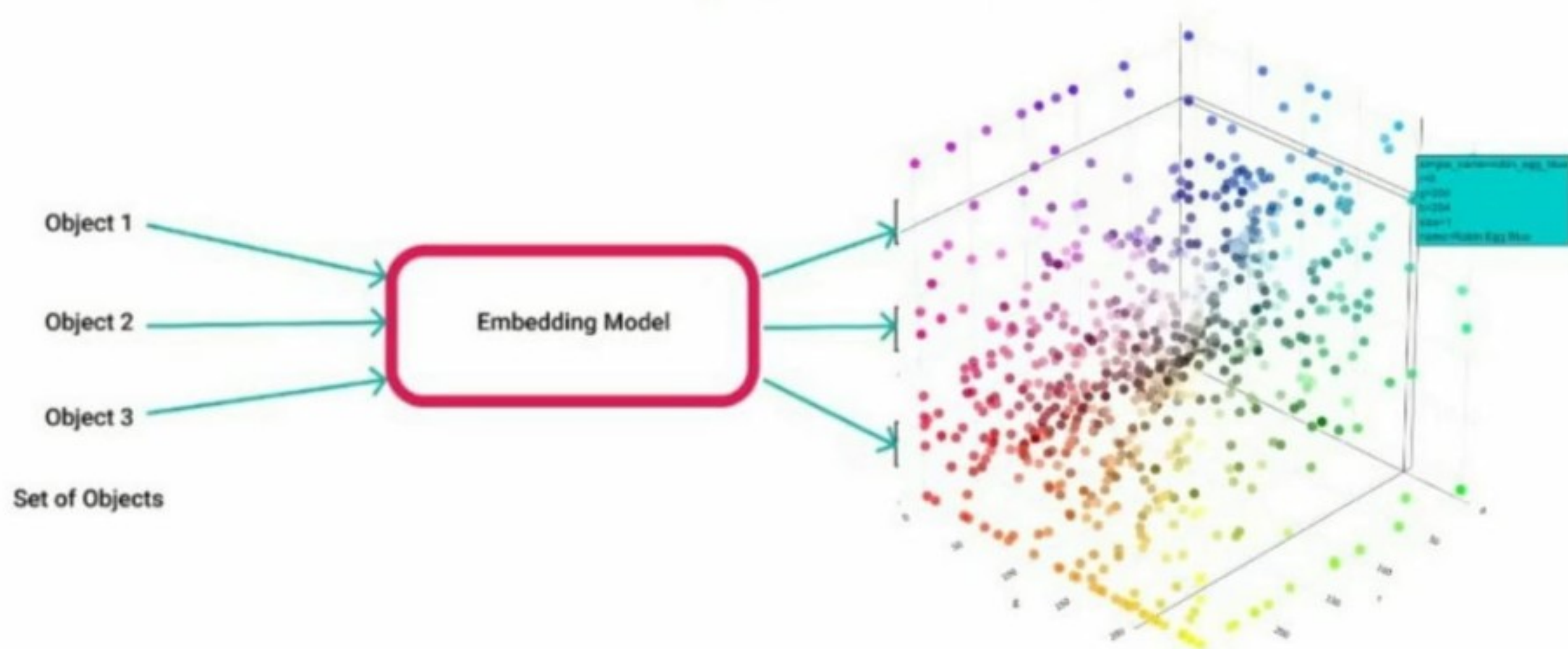
Decoding

Fine Tuning is fun for all!

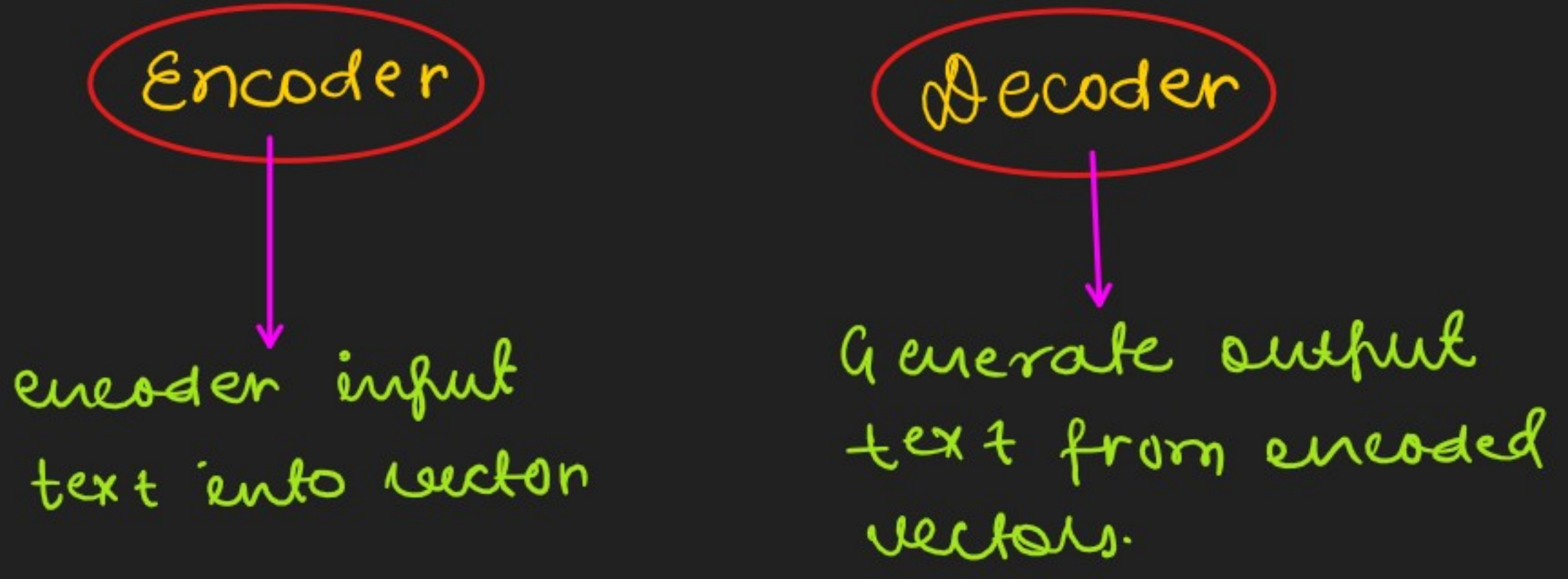
vector embeddings (2D example)



INTRODUCTION TO VECTOR EMBEDDINGS



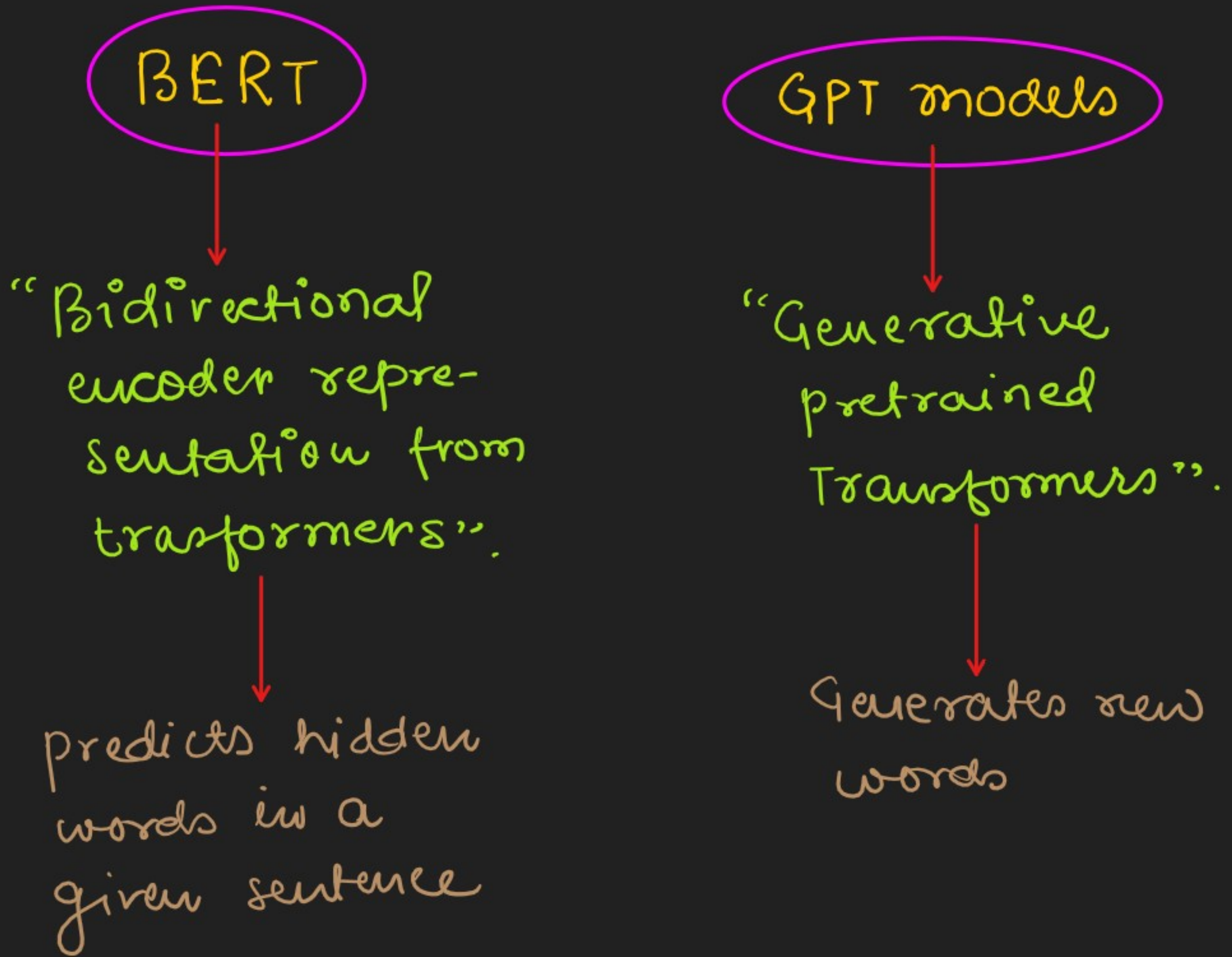
3. Transformer architecture consist of:



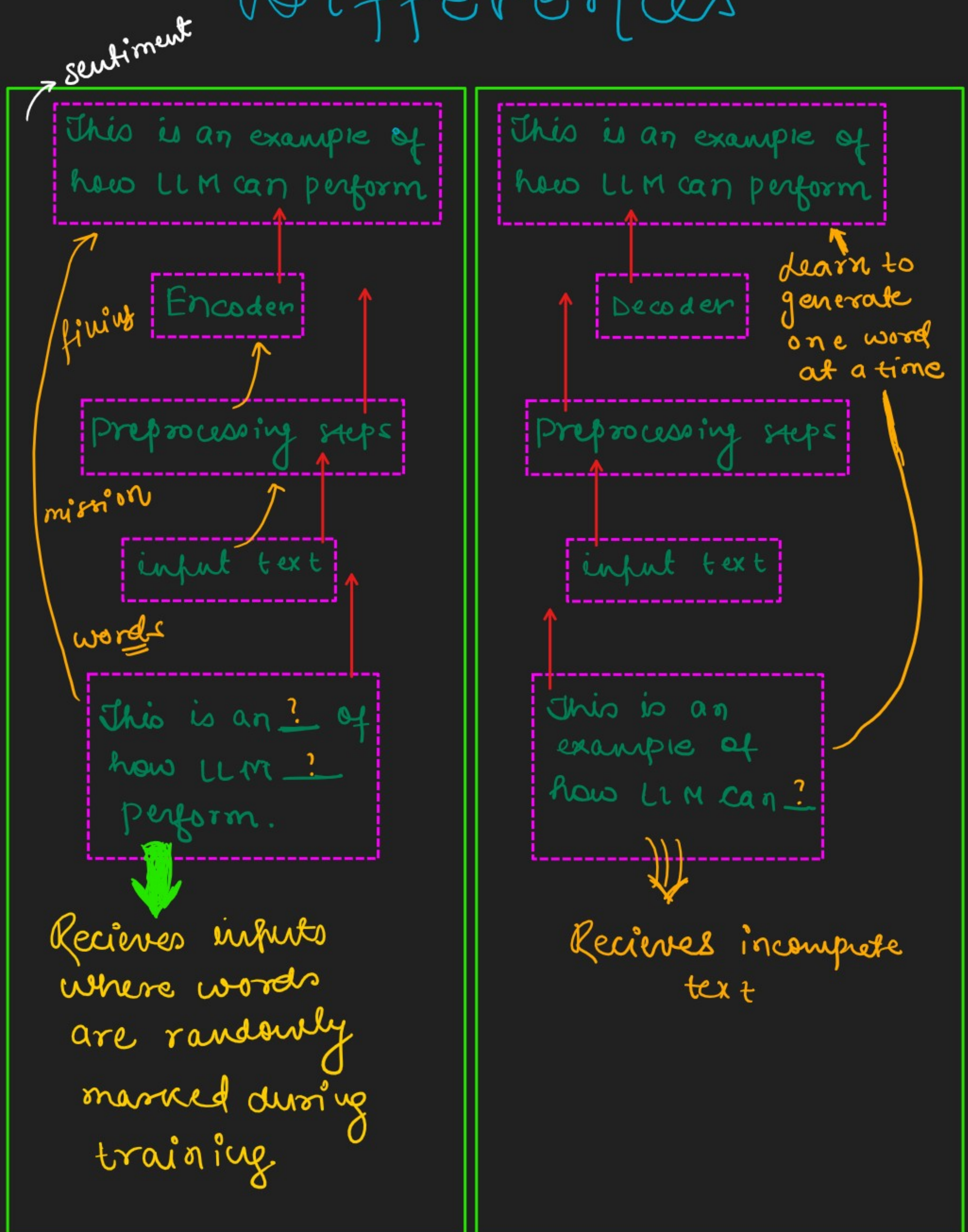
A note on self-attention mechanism:

- Key part of transformer: Allow model to weigh importance of different words/tokens relative to each other.
- Enables models to capture long range dependencies.
- we will look at this in detail later.

4. Later variations of transformer architecture



Differences

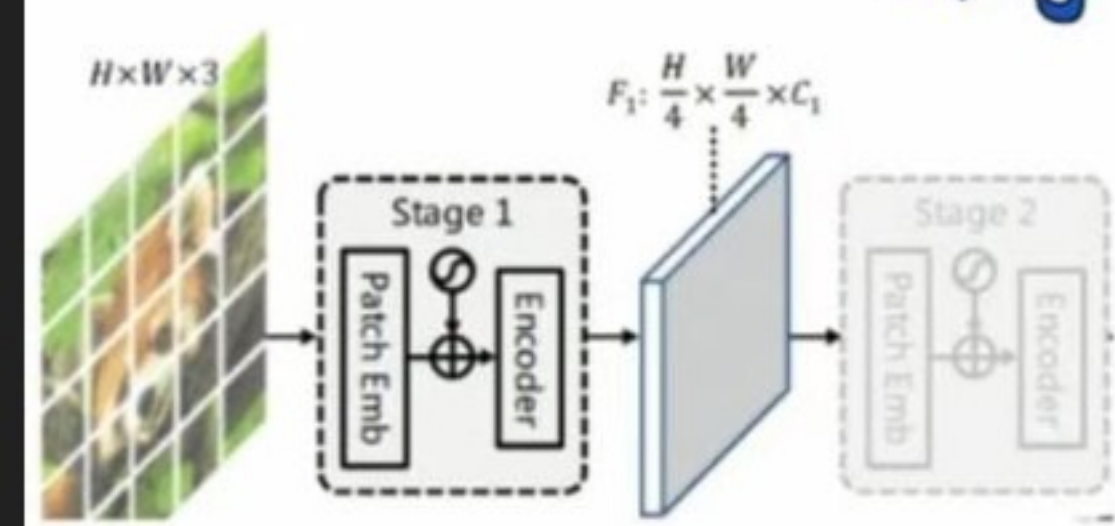


5. Transformers vs LLMs

→ Not all transformers are LLMs.

→ Transformers can also be used for computer vision.

Transformers in computer vision



Self-attention (Vaswani et al.)

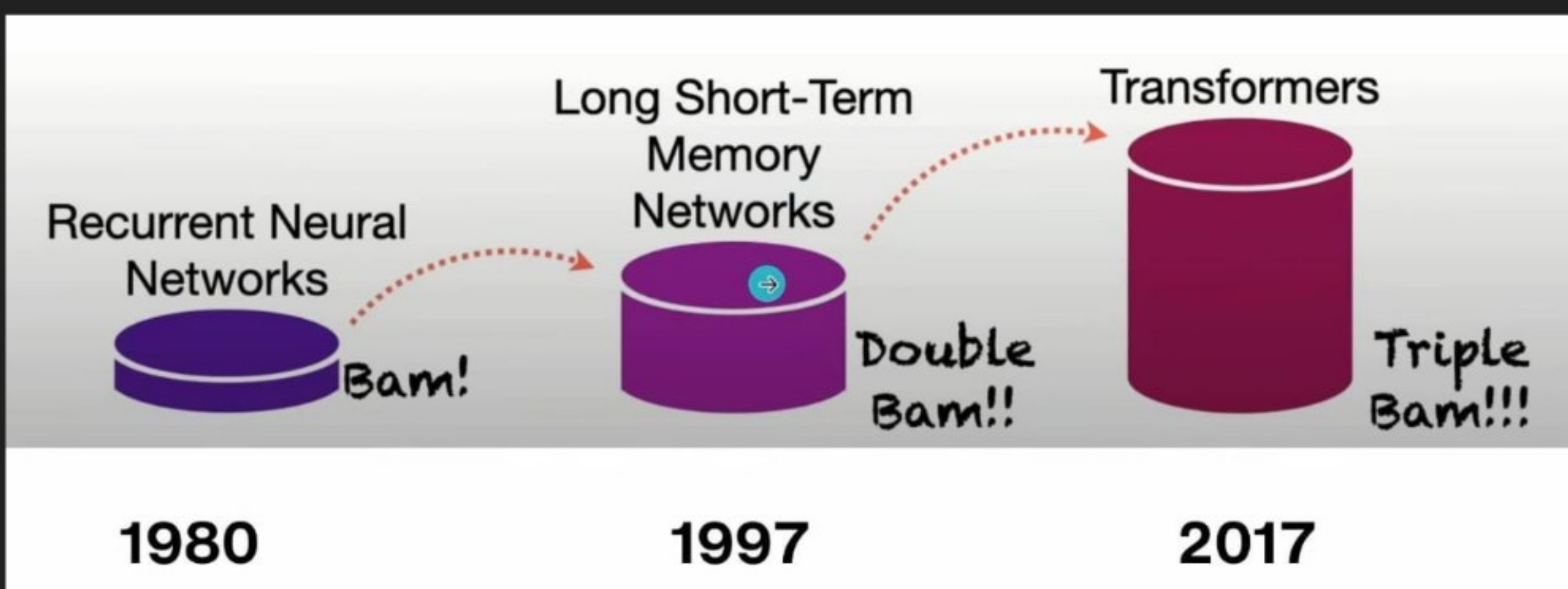
$$\mathcal{A}(K, Q) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right)$$

$\mathcal{A} \in \mathbb{R}^{N \times N}$

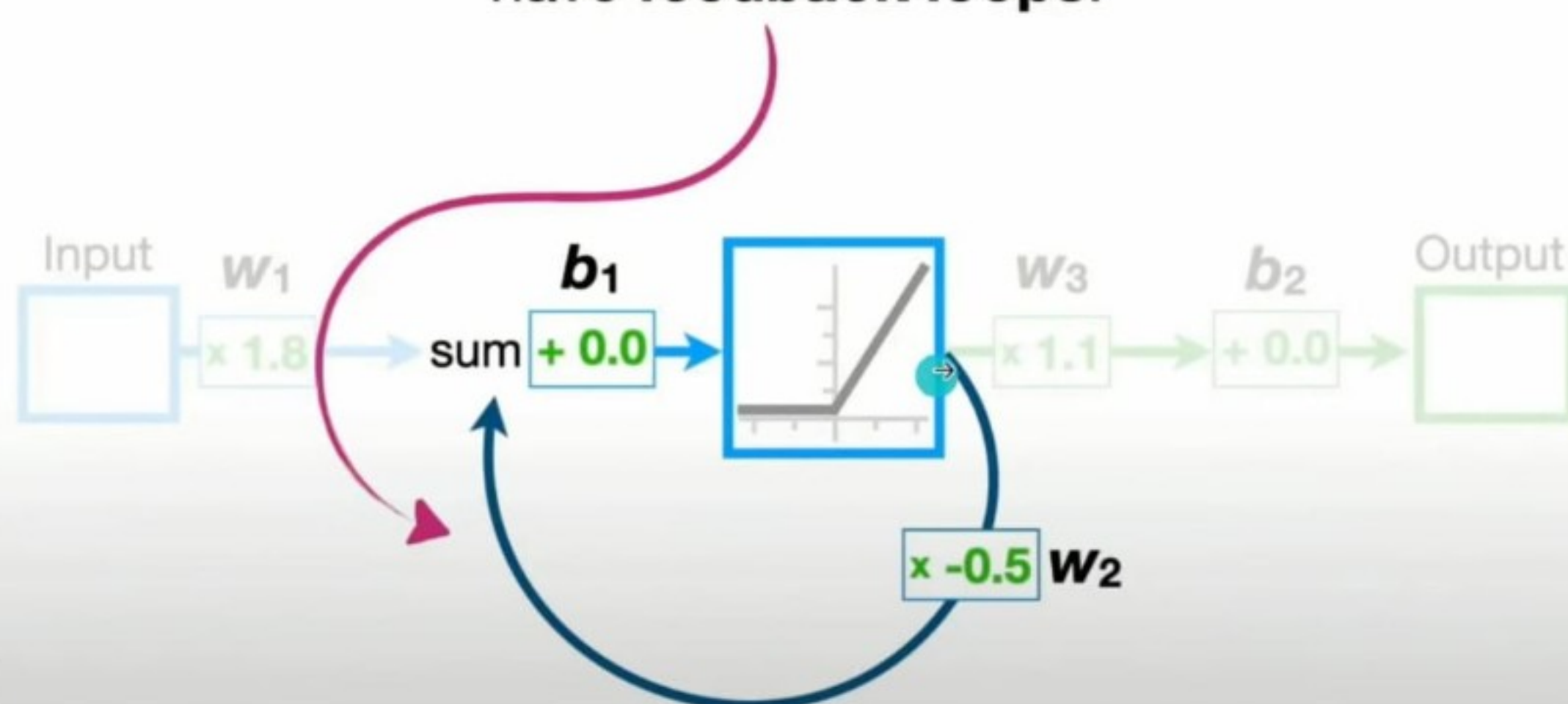


→ **Music** Not all LLMs are transformers.

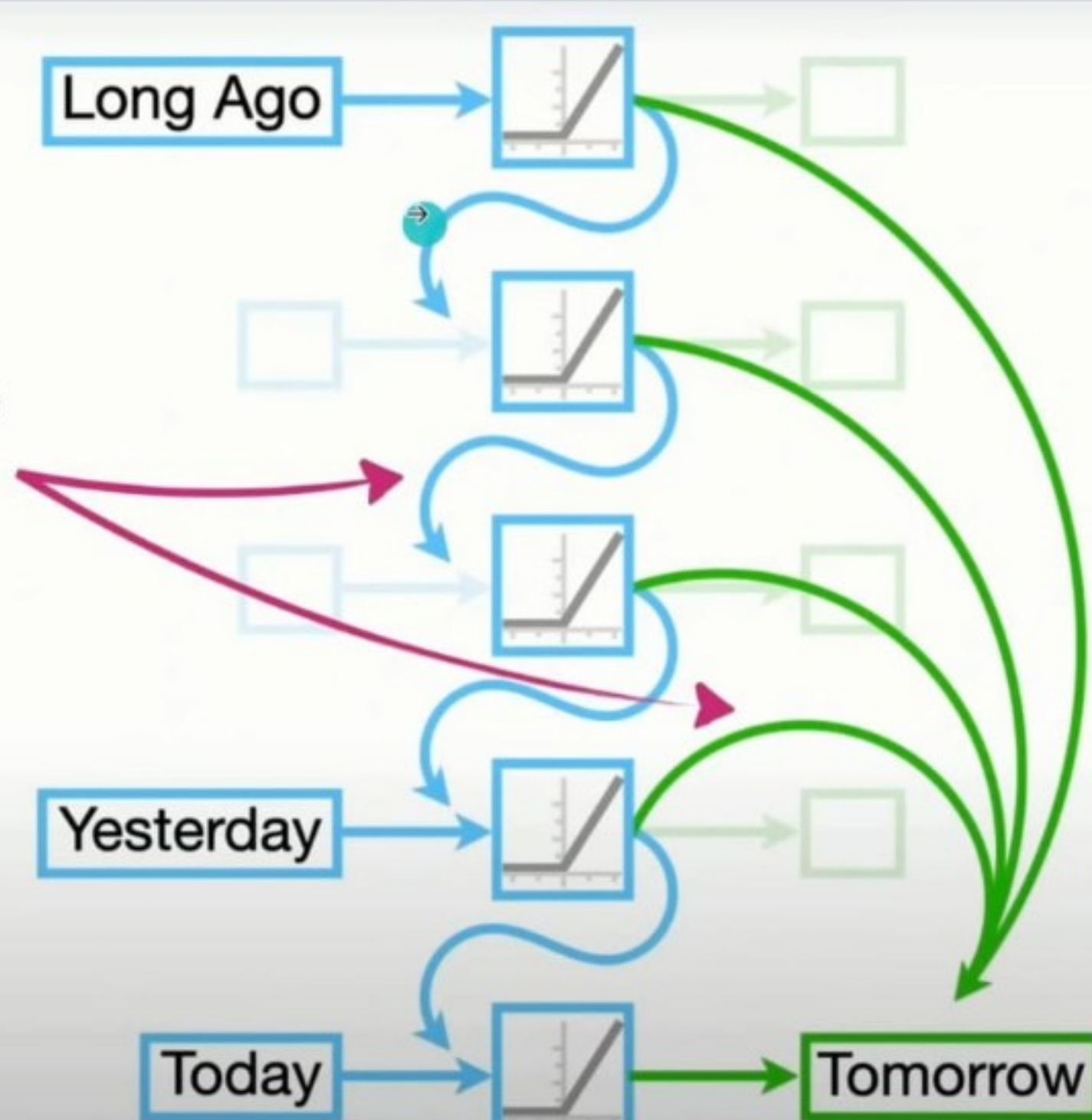
→ LLMs can be based on recurrent or convolutional architecture as well.



The big difference is that **Recurrent Neural Networks** also have **feedback loops**.



...**Long Short-Term Memory** uses two separate paths to make predictions about tomorrow.



First, this **green line** that runs all the way across the top of the unit is called the **Cell State** and represents the **Long-Term Memory**.

Now, this **pink line**, called the **Hidden State**, represents the **Short-Term Memories**.

