

## Lecture-4. Basic Intro To Transformer

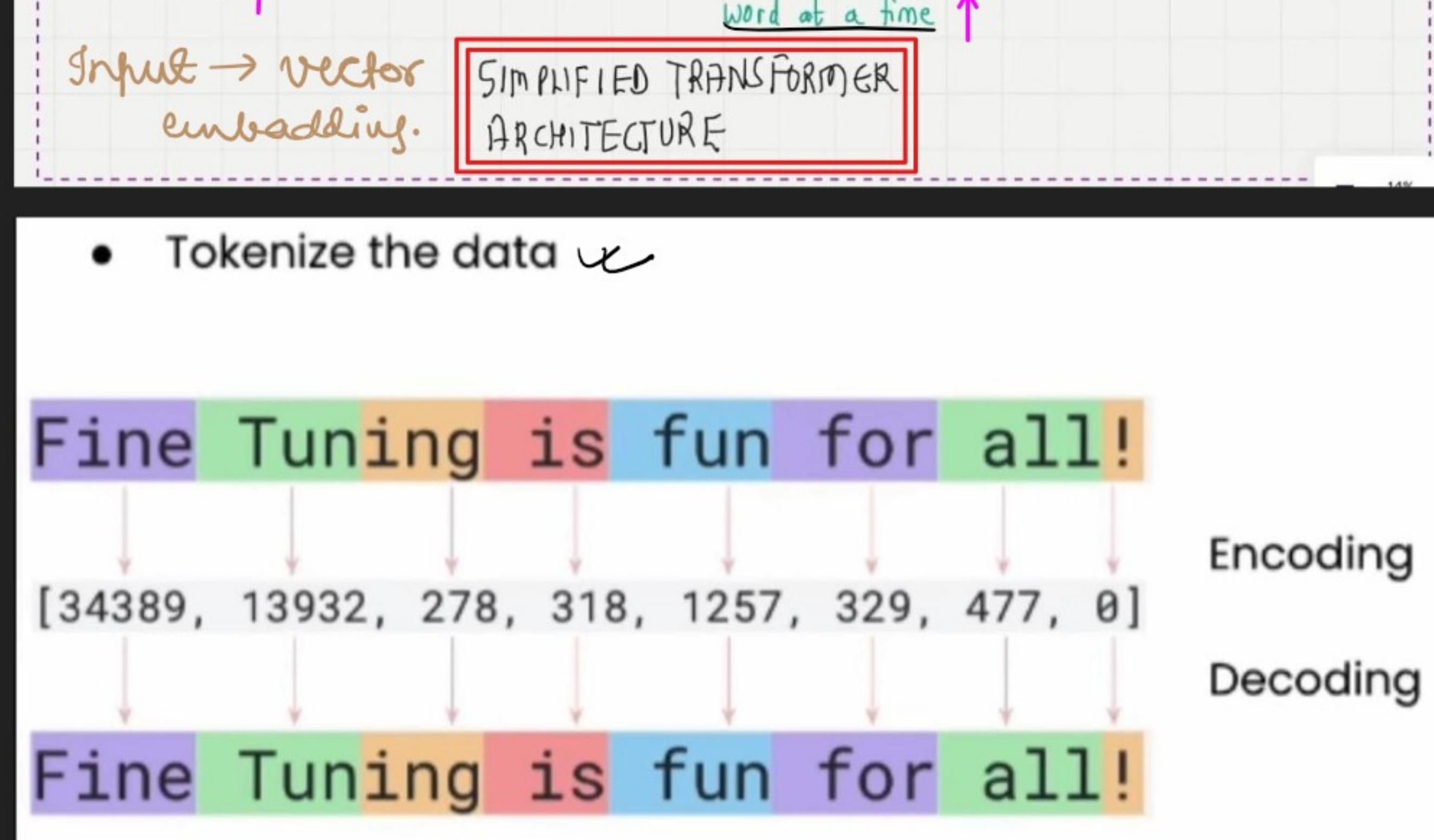
- Most modern LLMs rely on the transfer architecture → deep neural network architecture introduced in 2017 paper.

→ Attention is all you need.

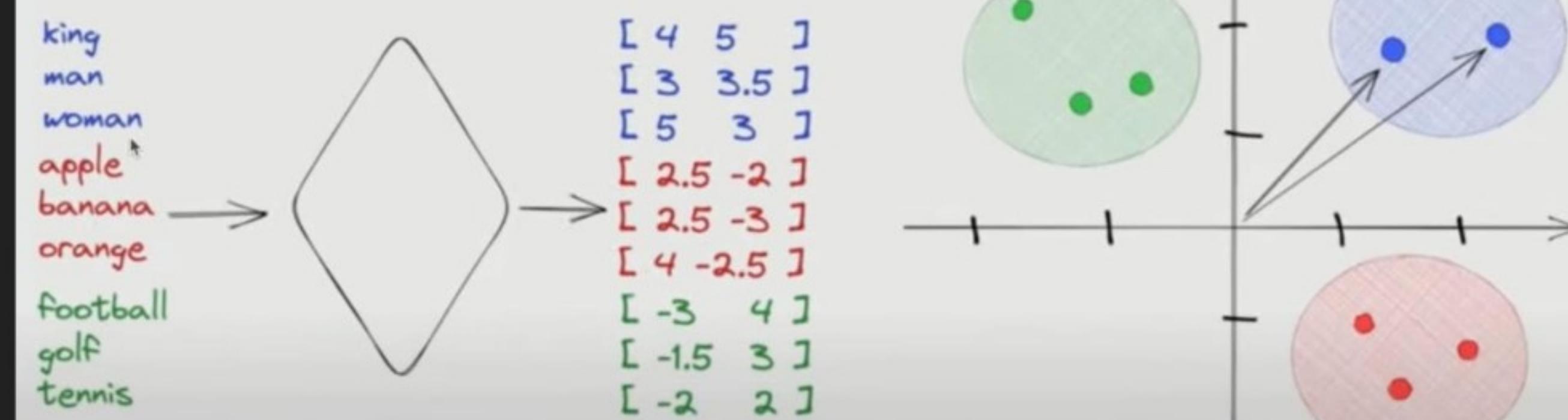
original Transformer = Developed for machine translation.



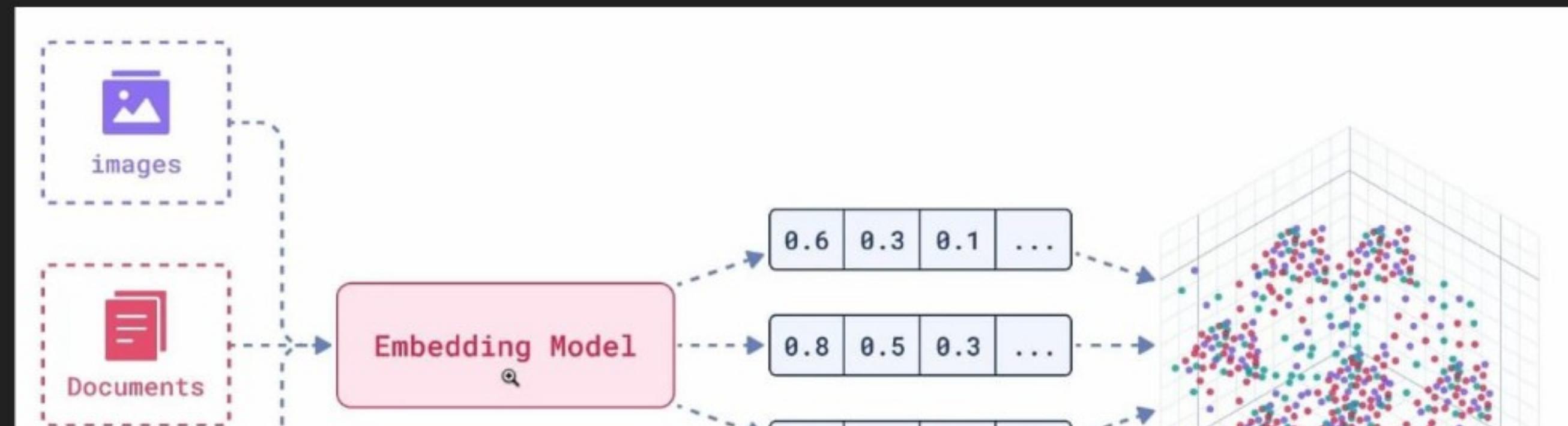
2.



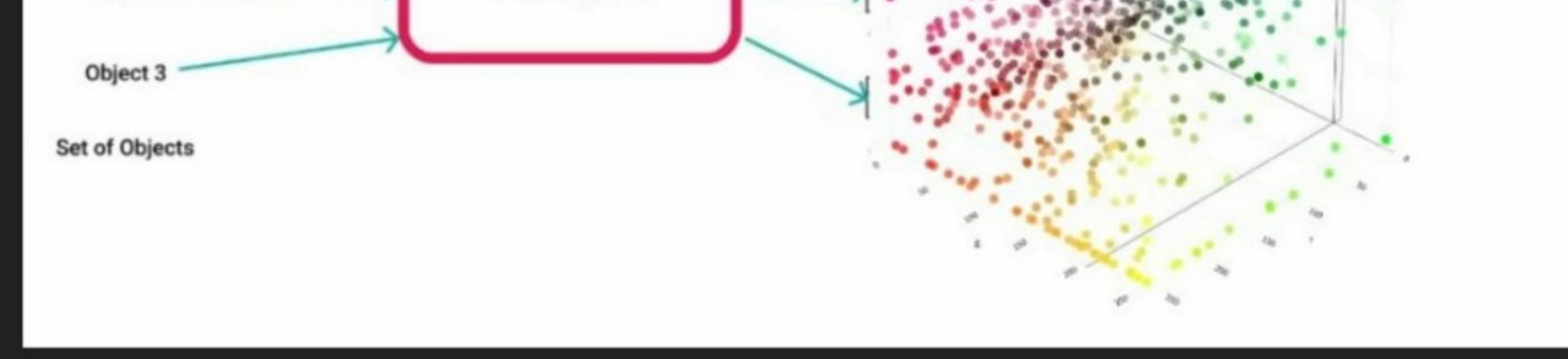
- Tokenize the data ↵



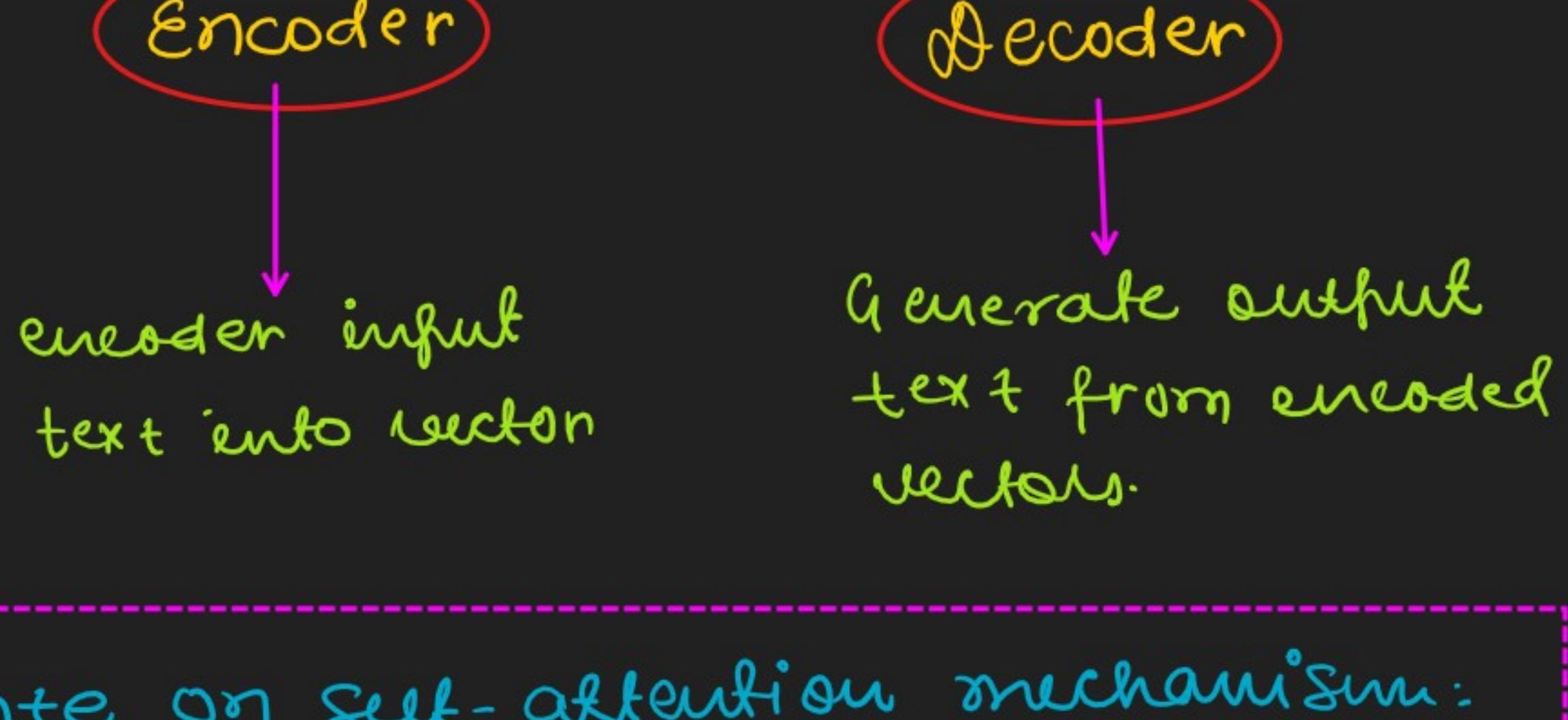
↙ vector embeddings (2D example)



## INTRODUCTION TO VECTOR EMBEDDINGS



### 3. Transformer architecture consists of:



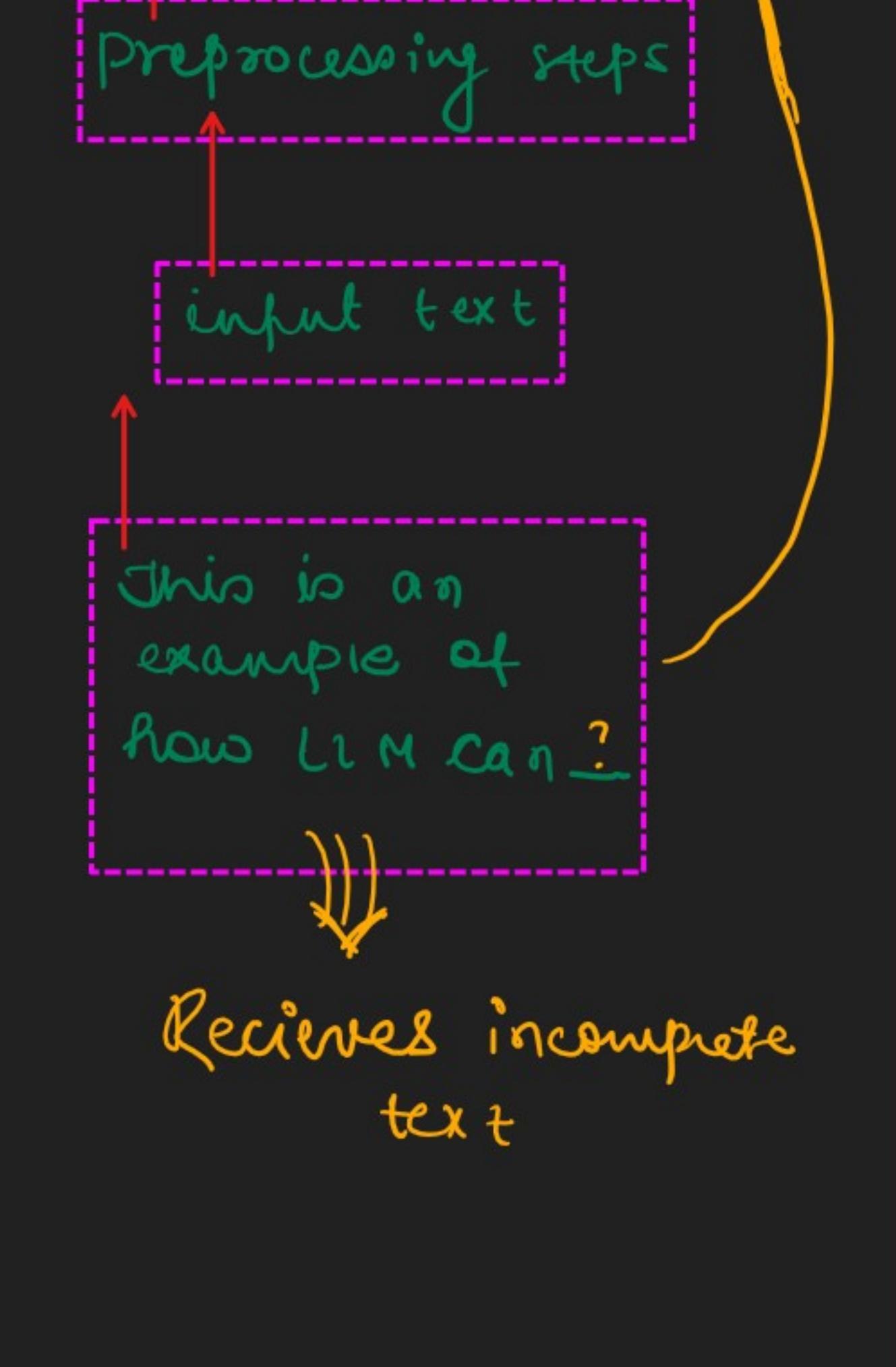
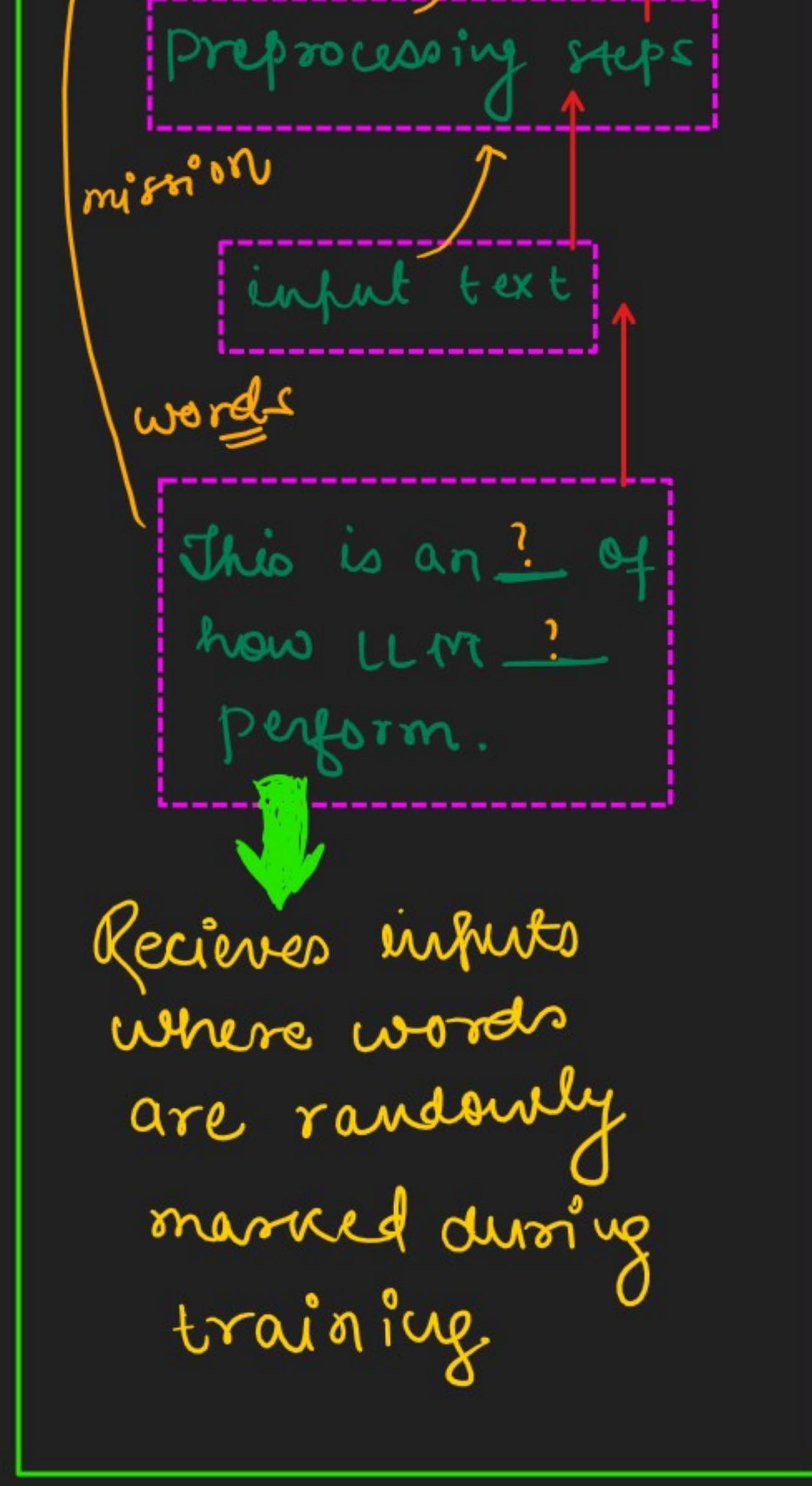
A note on self-attention mechanism:

- 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. Other variations of transformer architecture

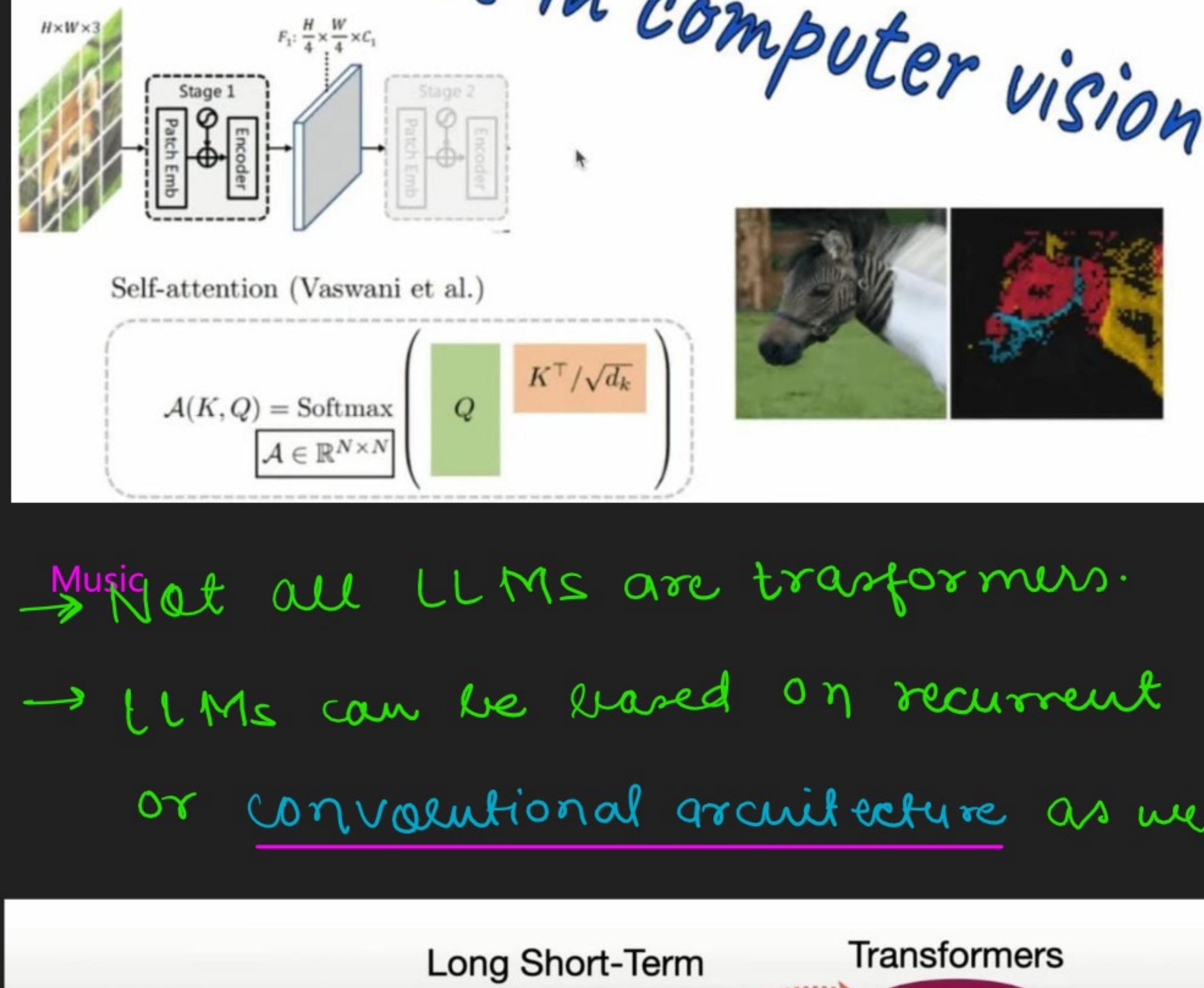


### Differences

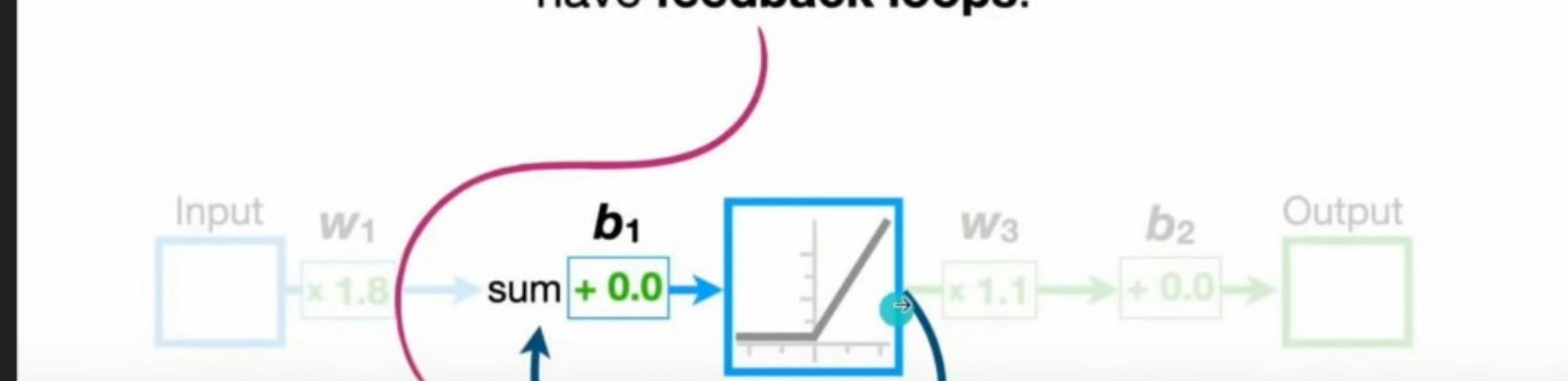


## 5. Transformers VS LLMs

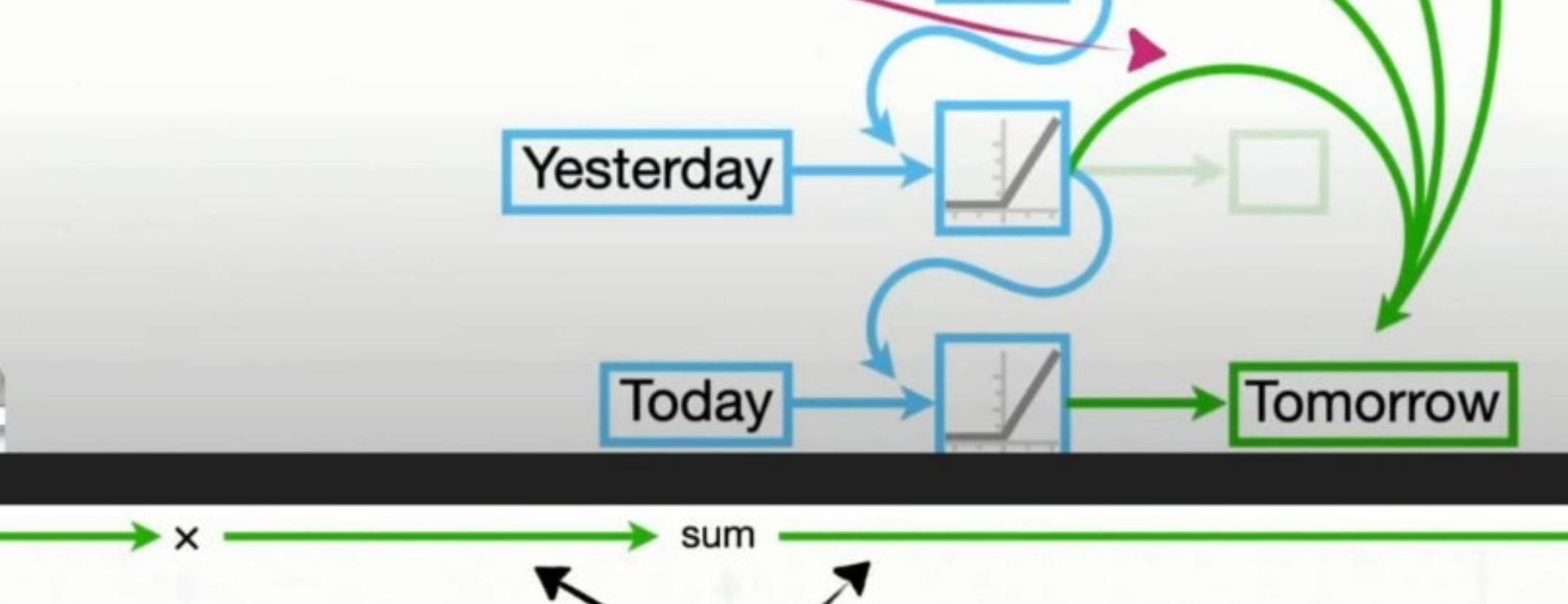
- Not all transformers are LLMs.
- Transformers can also be used for computer vision.



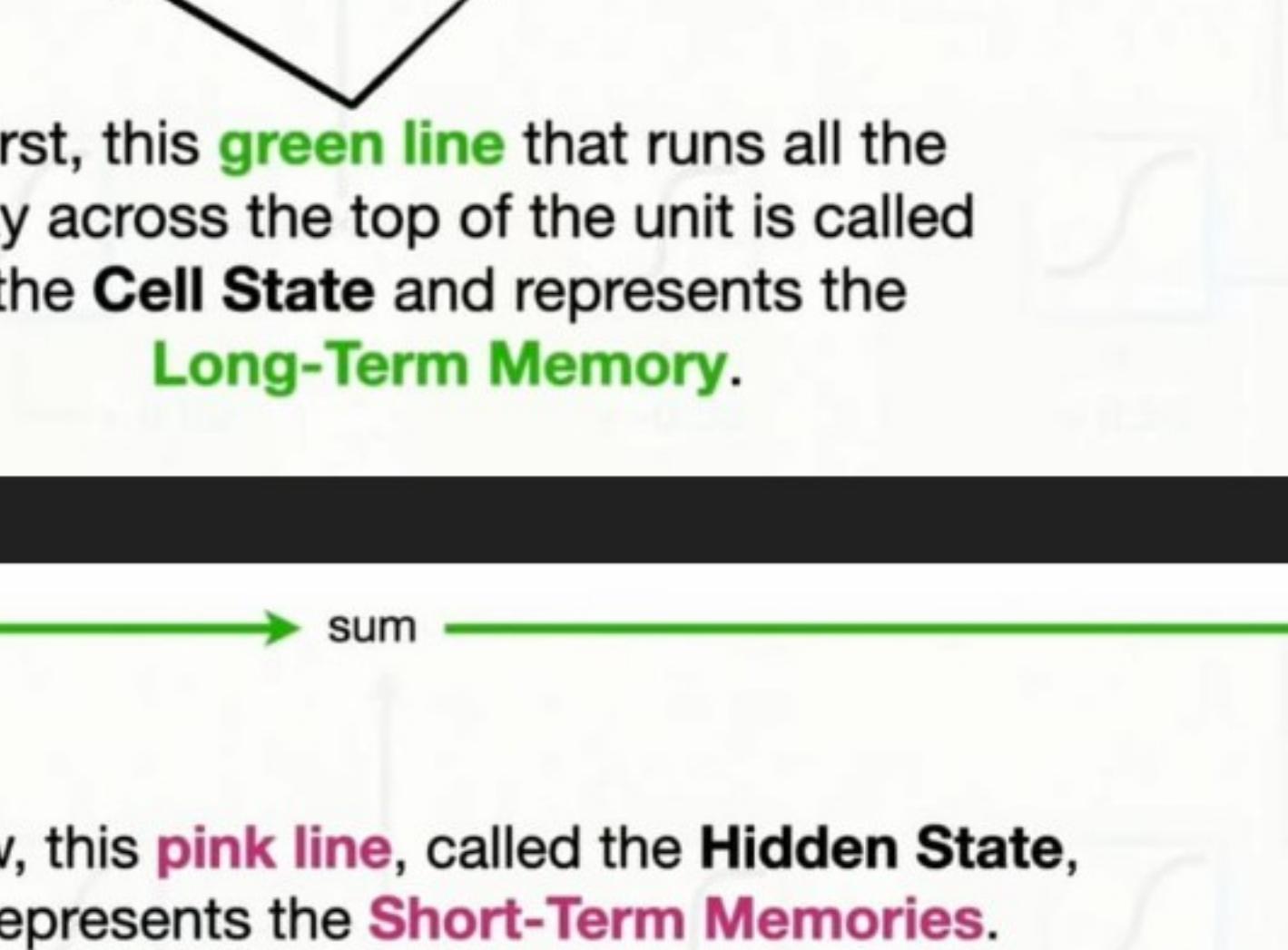
- Not all LLMs are transformers.
- LLMs can be based on recurrent or convolutional architecture as well.



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



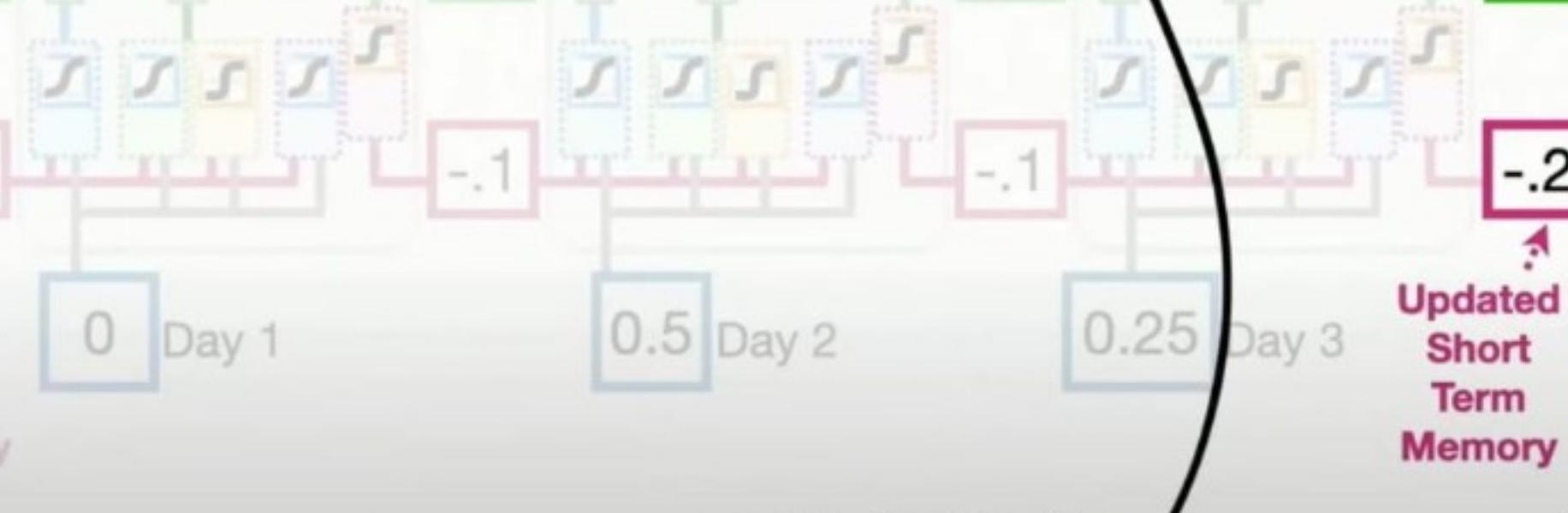
...Long Short-Term Memory uses two separate paths to make predictions about tomorrow.



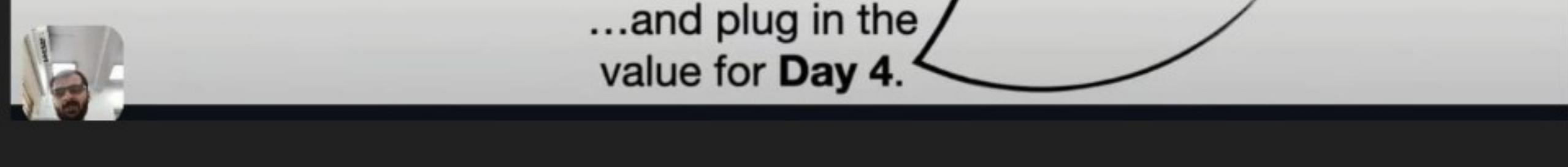
First, this **green line** that runs all the way across the top of the unit is called the **Cell State** and represents the **Long-Term Memory**.



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



...and plug in the value for **Day 4**.



Initial Long Term Memory  
Initial Short Term Memory  
Day 1  
Day 2  
Day 3  
Day 4  
Day 5  
High=1  
Low=0  
Value  
Day  
Updated Long Term Memory  
Updated Short Term Memory

...and plug in the value for **Day 4**.