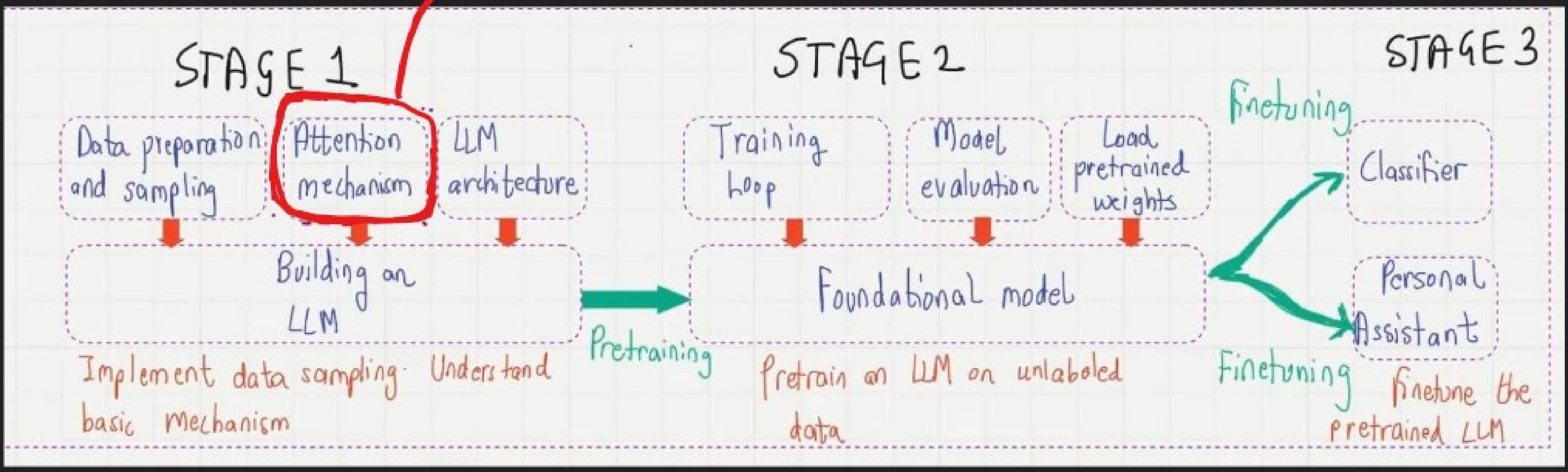
Lecture-13: Attention mechanism

Introduction

we are going to look at this sub-section



"The cat that was sitting on the mat, which was next to the dog, jumped"

(Attention)

Here, cat, sitting, jumped these three are taking lots of more attention.

That's why called attention mechanism

1. The 4 types of Attention Mechanism

Simplified self Attention

A simplified self attention technique to introduce the broad idea.

Self Attention

self attention with trainable weights, that form the basis of the mechanism used in LLM

Casual Attention

A type of self attention used in LLMs that allows the model to only consider previously and current input in a sequence

multi-head Attention ☆☆☆

An extension of self attention and casual attention that enables that model to simultaneousy attend to information from different representation subplaces.

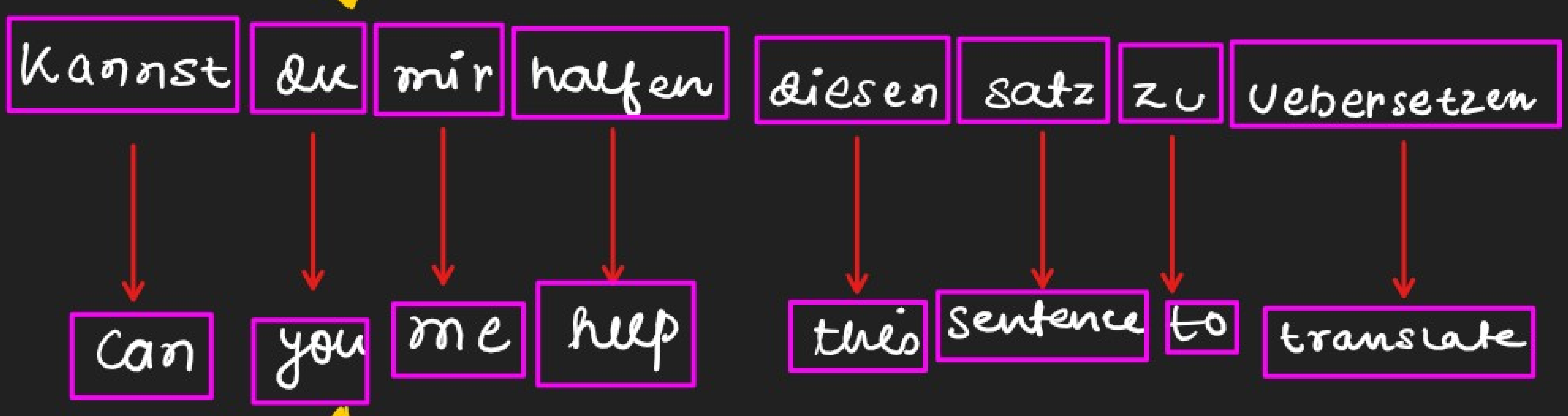
↪ LLM attend to various input data in parallel

2. The problem with modeling long subsequence

(a) what is the problem with architecture without the attention mechanism which came before LLMs?

* Let's consider a language translation model:

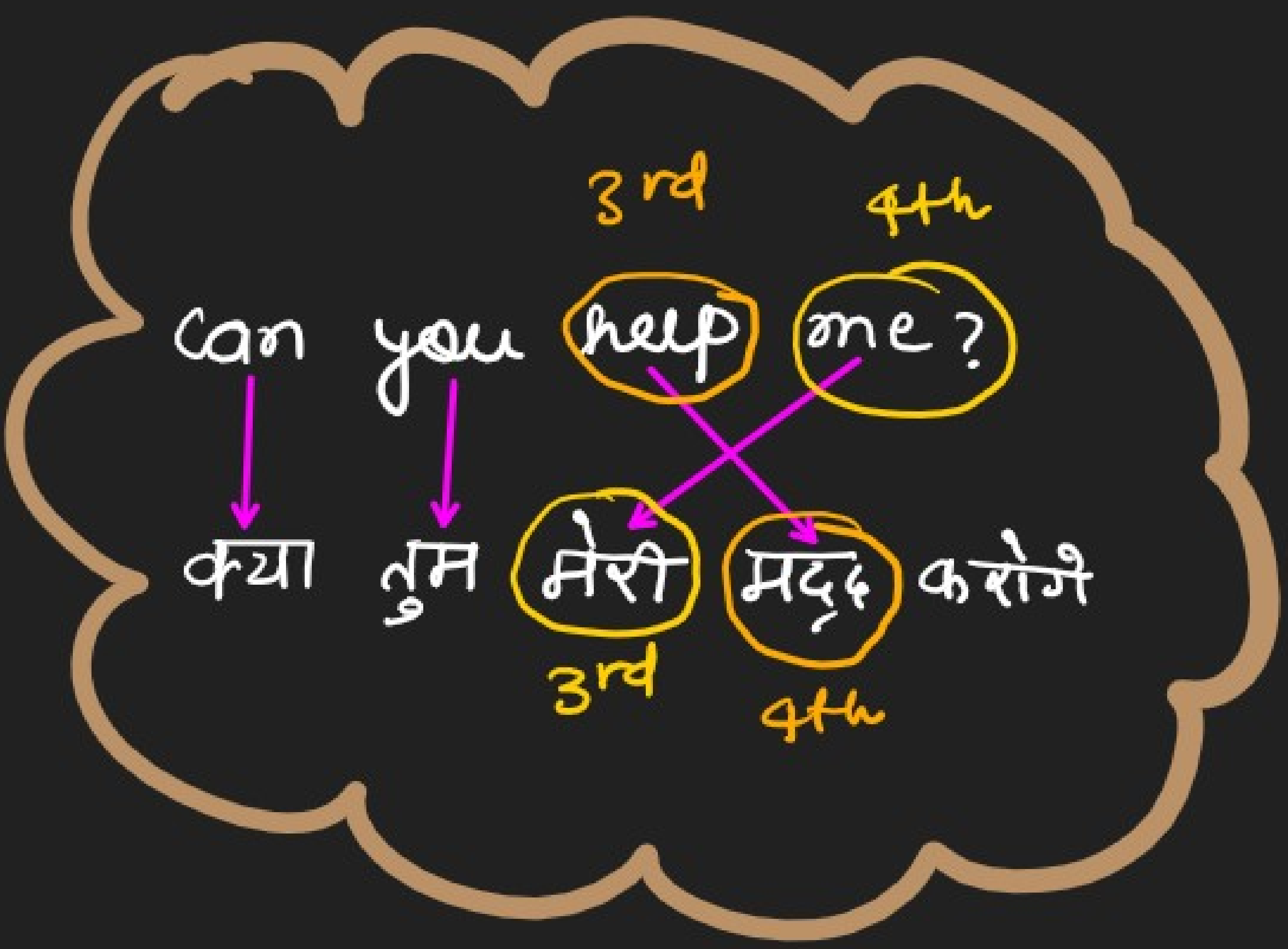
German input sentence to translate



The word by word translation result in a grammatically incorrect sentence

so,

- word by word translation does not work.



- The translation process requires contextual understanding and grammer alignment

(b) To address the issue that we can't translate text word by word, it is common to use a neural network with two sub-module

RNN
rus

German text → ENCODER → Read and process text

DECODER → Translate text
Eng text

(c) Before transformers, Recurrent Neural Network (RNNs) were the most popular encoder-decoder architecture for language translation.

Implemented in 1980

(d) RNN: output from previous step is fed as input to current text.

(e) Here's how the encoder-decoder RNN works:

Input text

↓

Encoder (Process input text sequentially)

↓

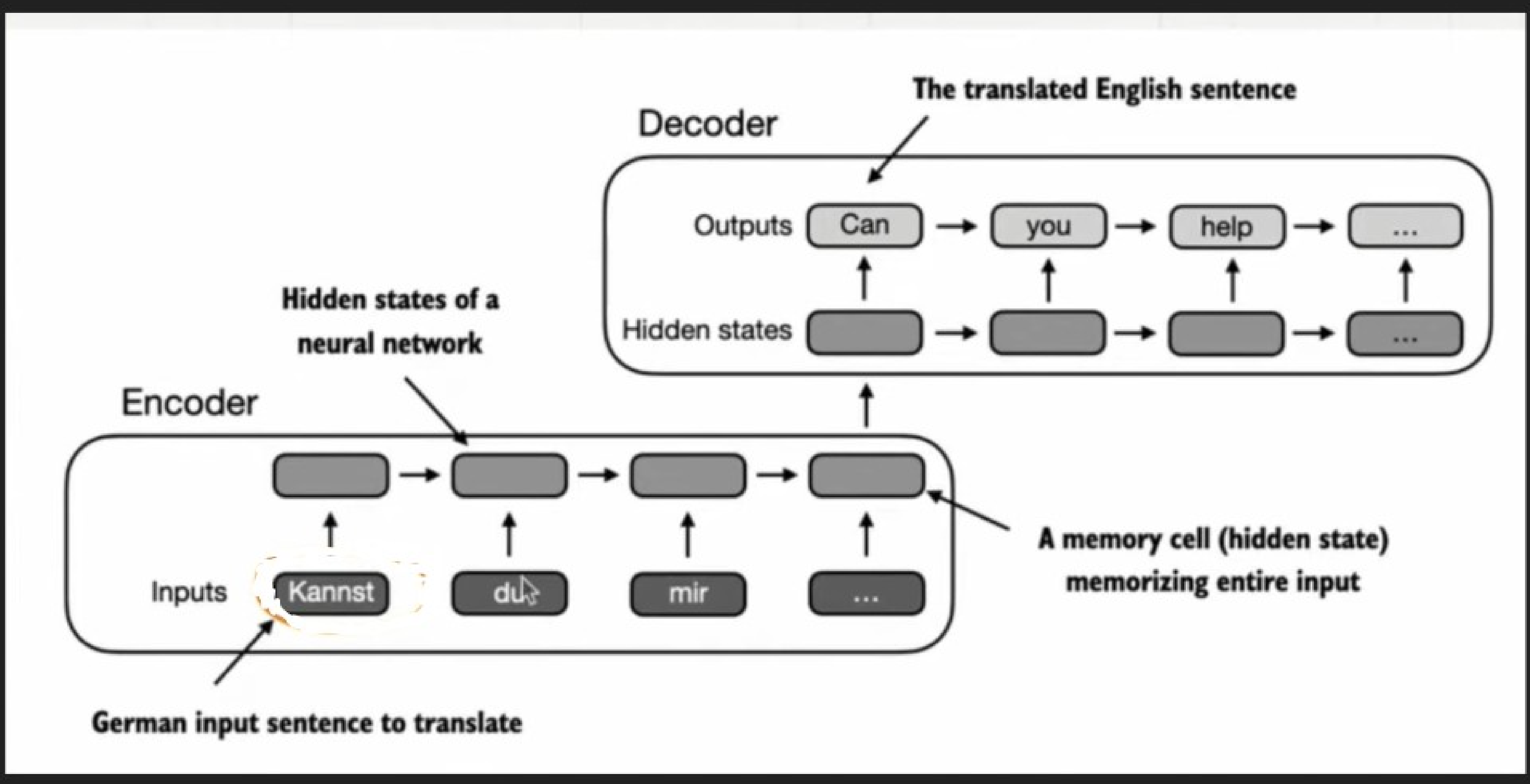
updates hidden State at each step (Internally values at hidden layer)

↓

final hidden state (encoder tries to capture sentence meaning)

↓

Decoder uses the final hidden state to generate translate sentence (one word at a time) (decoder also updates it's hidden state at each time)


The translated English sentence
Decoder
Outputs
Can
you
help
...
Hidden states
...
Hidden states of a neural network
Encoder
A memory cell (hidden state) memorizing entire input
Inputs
Kannst
du
mir
...
German input sentence to translate

(f) The encoder process the entire input text into a hidden state (memory cell).

Decoder takes this hidden state to produce an output.

(g) BIG ISSUE:

RNN cannot directly access earlier hidden states from the encoder, during the decoding phase.

↓

It relies solely on current hidden state

↓

This leads to a loss of context, especially in, complex sentence where dependencies might span long distance.

Encoder compress entire input sequence into a single hidden state vector

↓

If the sentence is very long, it becomes very different for the RNN to capture all information in a single vector

↓

main draw back of RNN

The cat that was sitting on the mat, which was next to the dog, jumped.

↓

Le chat qui était assis sur le tapis qui était à côté du chien, a sauté

↓

Here, the key action "Jumped" depends on the subject "cat" but also on understanding the longer dependencies.

↓

that was sitting on the mat, next to the dog

RNN decode might struggle with this.

3. Capturing data dependency with attention mechanism

(a) RNN work fine for translating short sentences, but don't work for long text as they don't have direct access to previous words in the input.

(b) One major shortcoming in the approach is that:

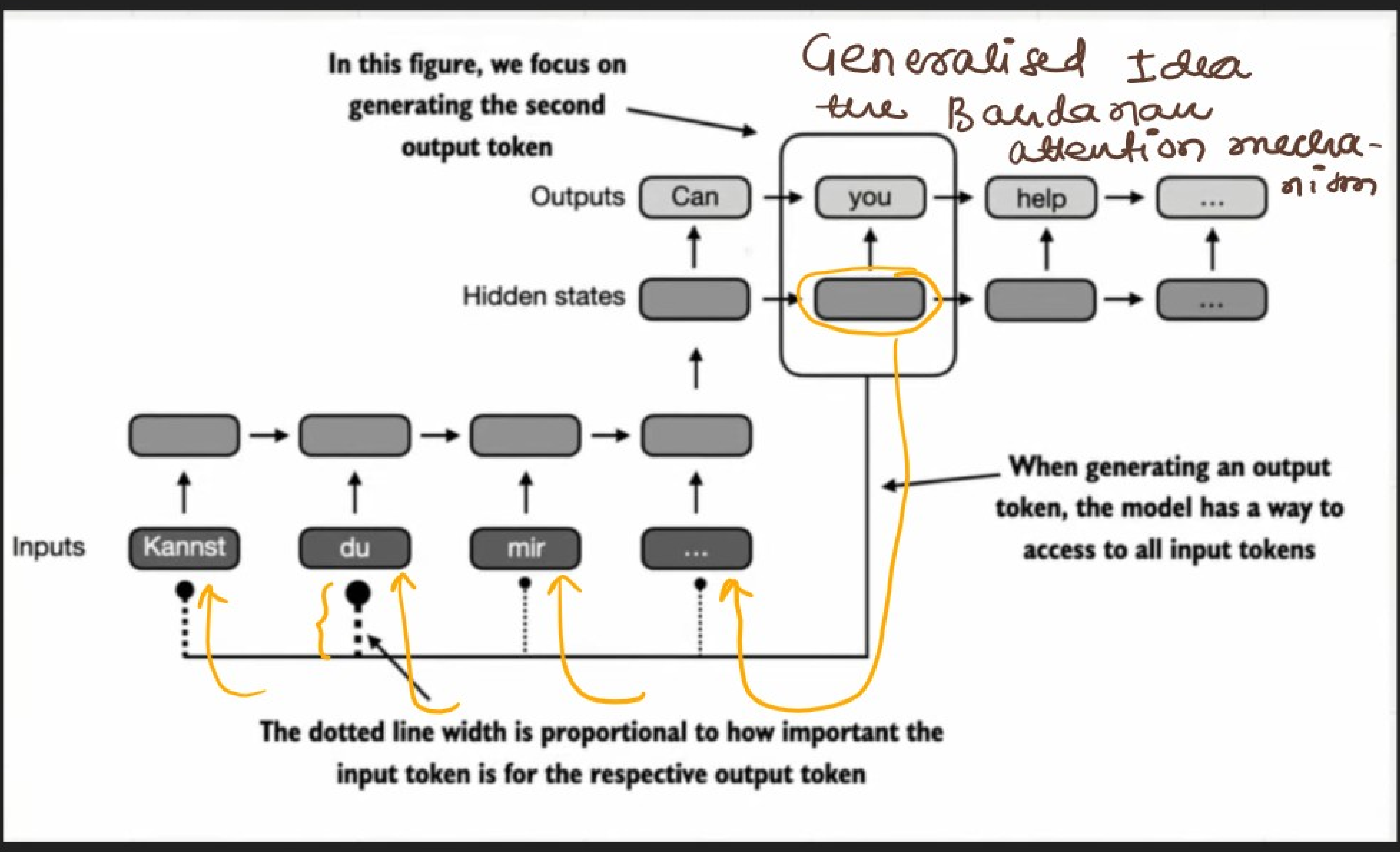
RNN must remember the entire encoded input in a single hidden state before passing it to the decoder.

(c) In 2014, researcher develop the so called:

"Bahdanau attention mechanism of RNN"

↓

Modifies the encoder-decoder RNN such that decoder can selectively access different parts of the input sequence at each decoding step.


In this figure, we focus on generating the second output token

Generalised Idea the Bandanau attention mecha-nism

When generating an output token, the model has a way to access to all input tokens

The dotted line width is proportional to how important the input token is for the respective output token

- Using an attention mechanism, the text generating decoder part of the network can access all input token selectively.

- This means that some input token are more important than others for generating a given output token.

- This importance is determined by the so called attention weights.

(d) only 3 years later, researchers found that RNN architecture are not required for building deep neural networn for natural language and proposed the original transformer architecture; with a self attention mechanism inspired by the Bahdanau attention mechanism.

"The cat that was sitting on the mat, which was next to the dog, jumped"

↓

"Le chat qui était assis sur le tapis, qui était à côte du chien, a saute"

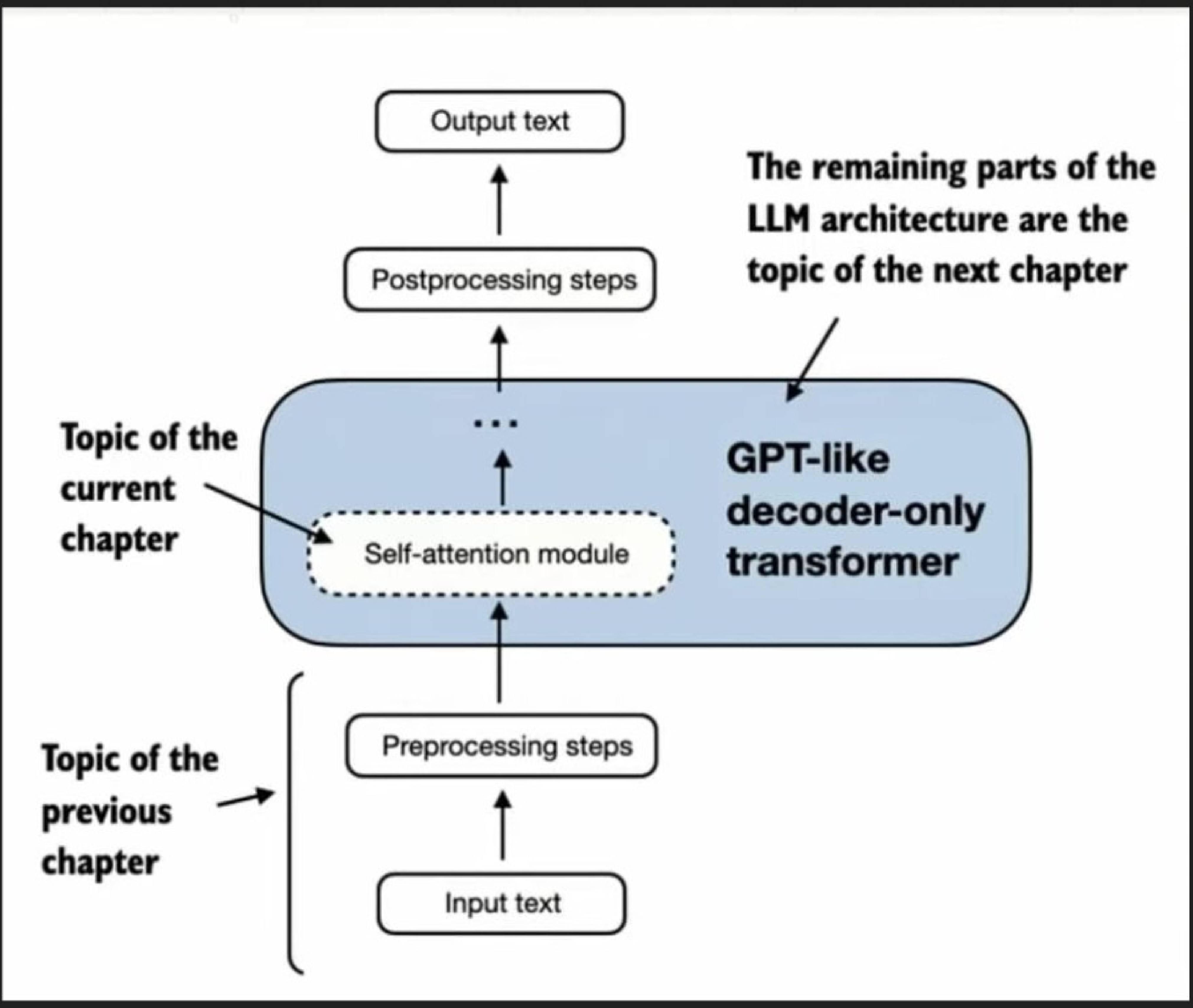
At each decoding step, the model can look back at the entire input sequence and decide which parts are most relevant to generate the current word.

when decoder is predicting "saute" the attention mechanism allow it to focus one part of input that corresponds to "jumped"

Dynamic focus on different parts of input sequence allows models to learn long range dependency more effectively.

(e) self attention is a mechanism that allow each position of the input sequence to attend to all position in the same sequence when computing the representation of a sequence.

(f) self attention is a key component of contemparary LLMs based on the transformer architecture, such as the GPT series.


Output text
Postprocessing steps
The remaining parts of the LLM architecture are the topic of the next chapter
...
GPT-like decoder-only transformer
Topic of the current chapter
Self-attention module
Preprocessing steps
Topic of the previous chapter
Input text

4. Attending to different parts of the input with self-attention

(a) In "self attention", the "self" refers to the mechanism ability to compute attention weights by relating different position in a single input sequence

(b) It means the relationship between various parts of the input itself, such as words in a sentence.

(c) This is in contrast to traditional attention mechanism, where the focus is an relationships b/w elements of 2 different sequence