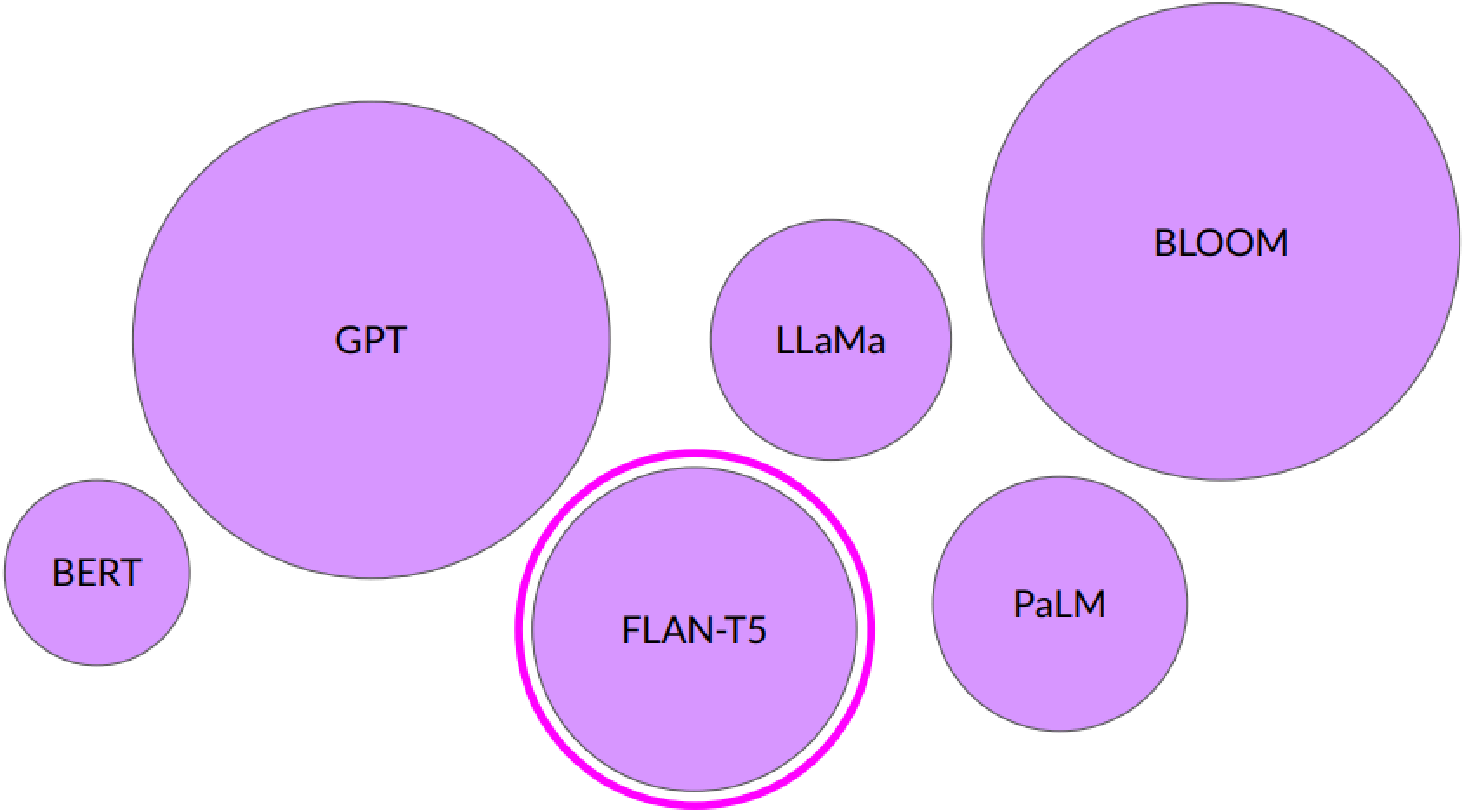
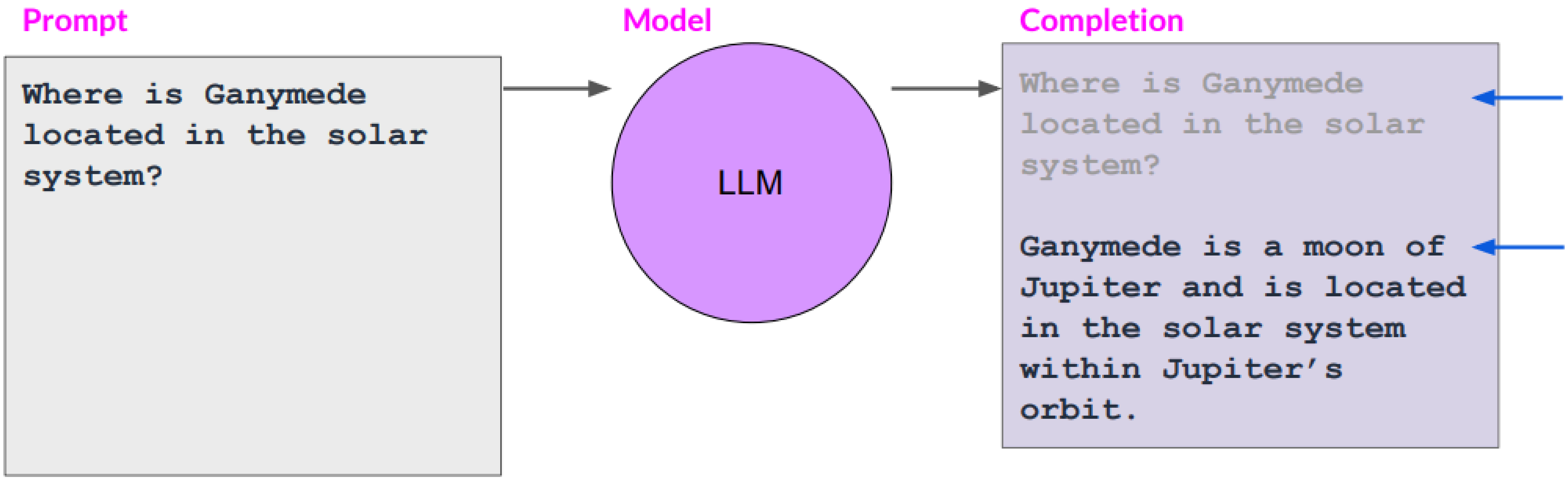
Large Language Models



Prompts and completions



Where is Ganymede located in the solar system?

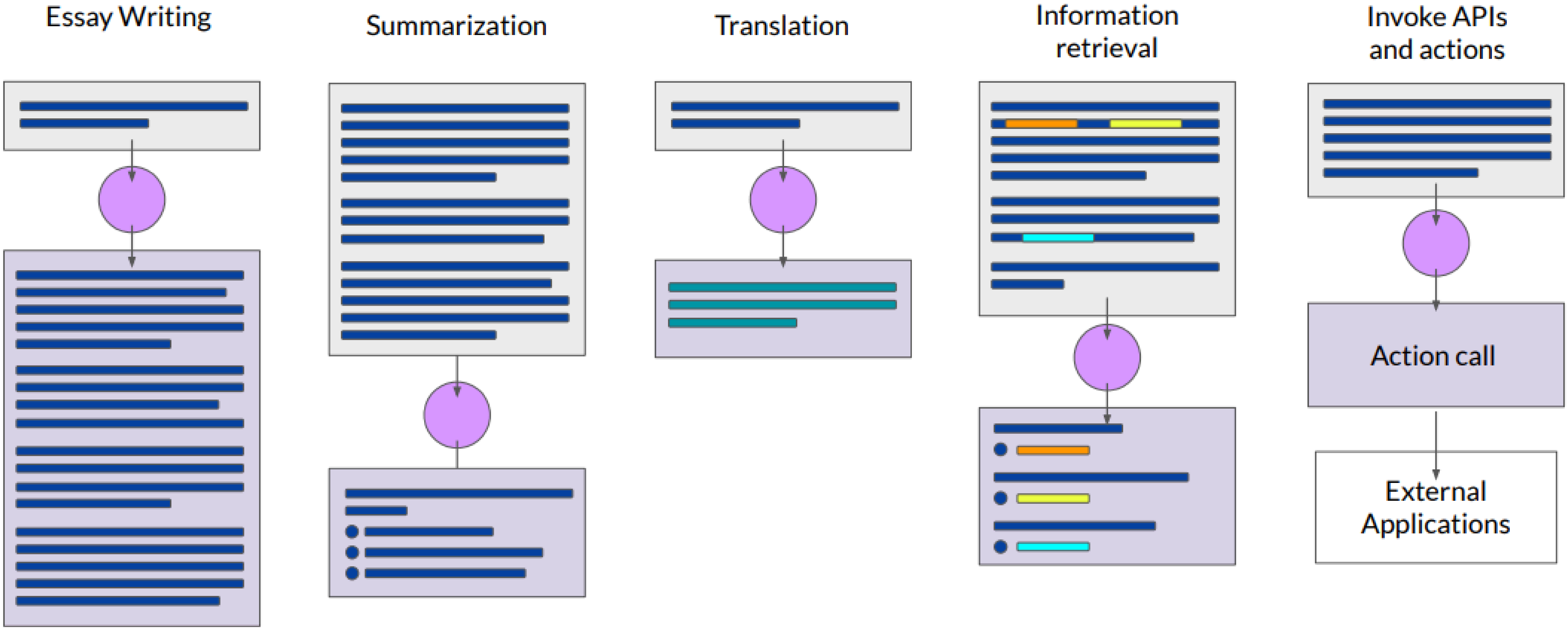
Where is Ganymede located in the solar system?

Ganymede is a moon of Jupiter and is located in the solar system within Jupiter's orbit.

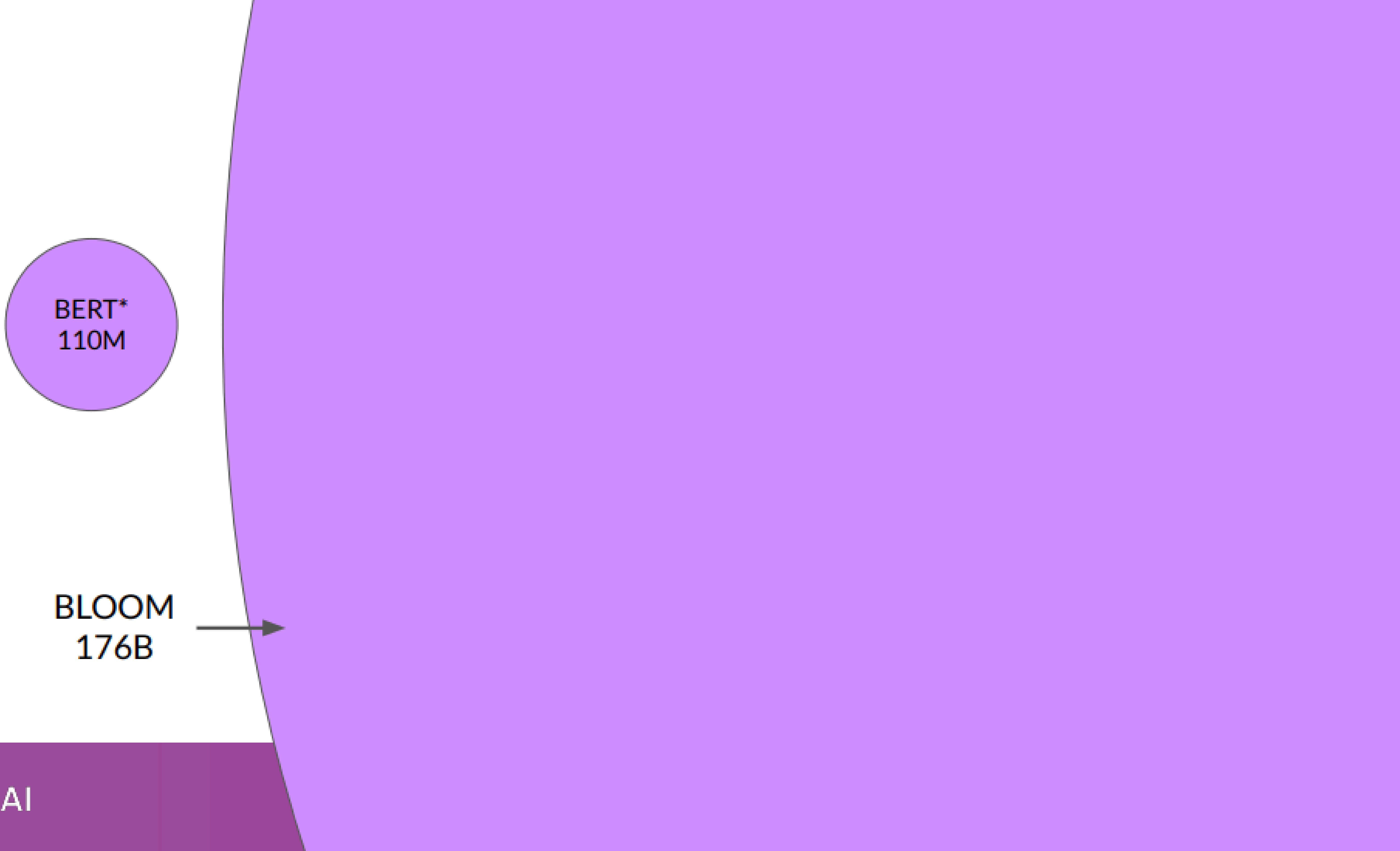
Context window

- typically a few 1000 words.

LLM use cases & tasks



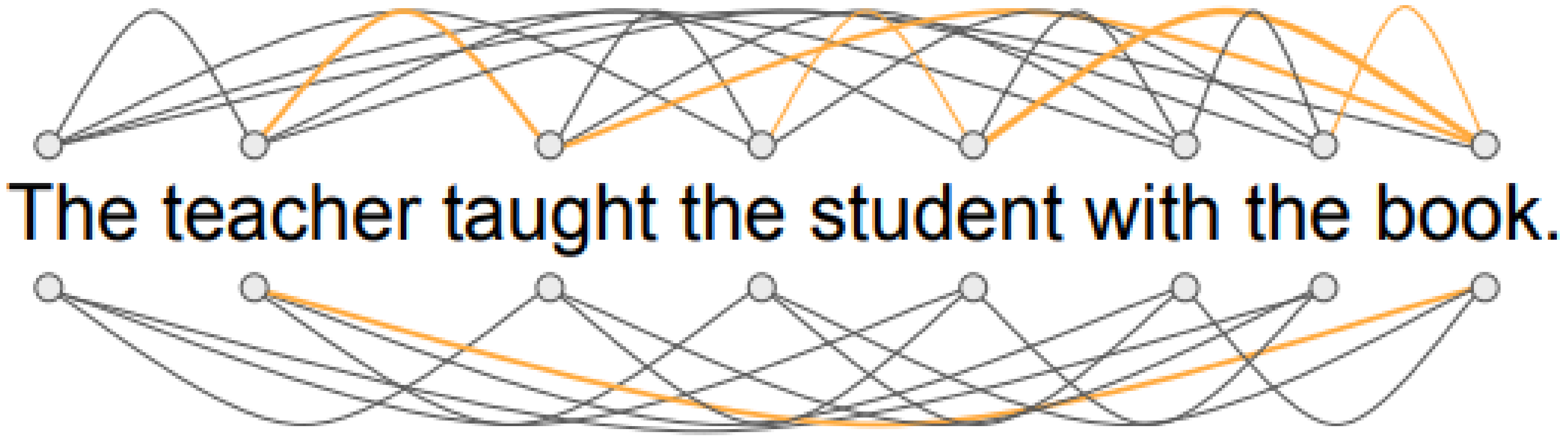
The significance of scale: language understanding



*Bert-base

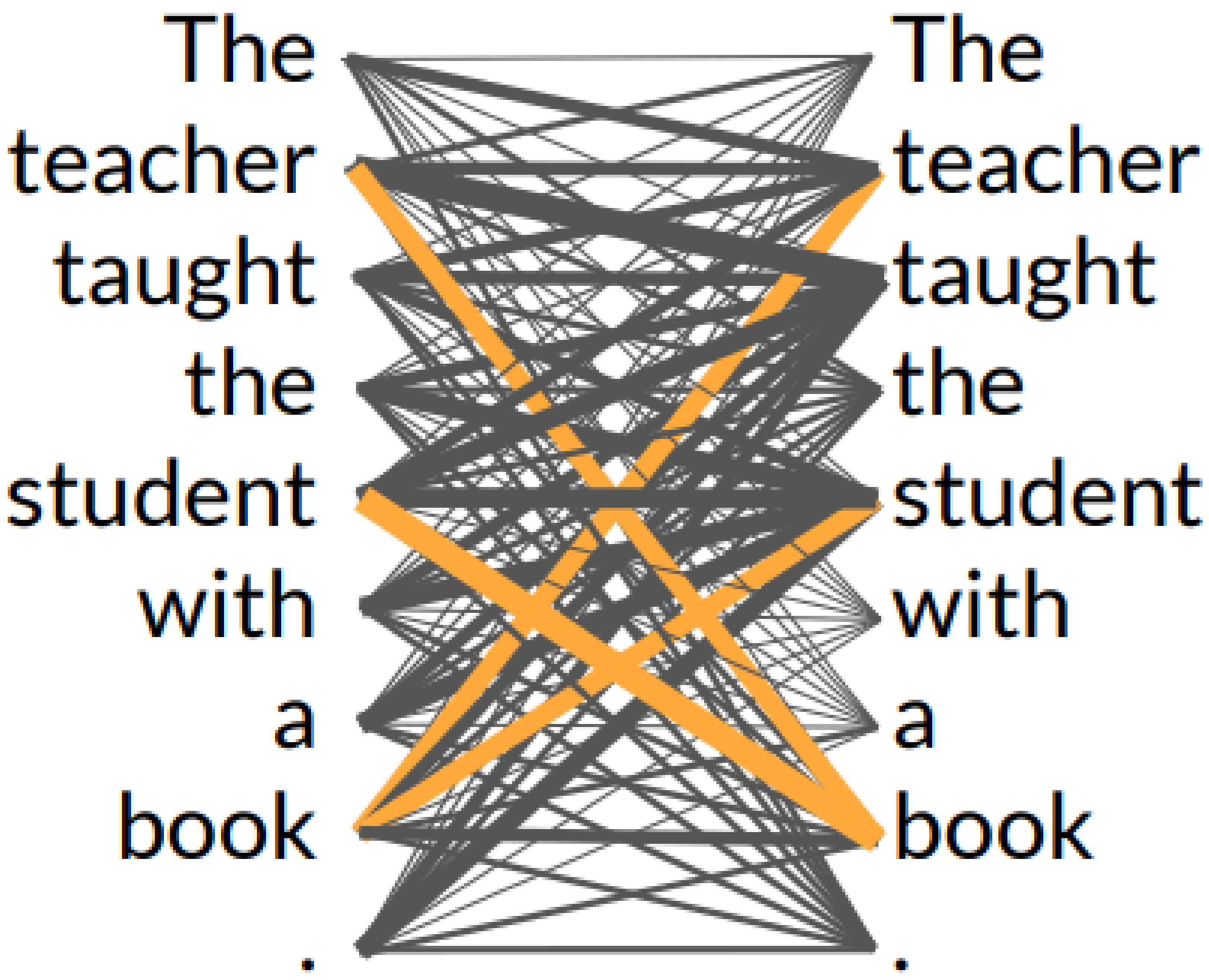
How LLMs work - Transformers architecture

Transformers

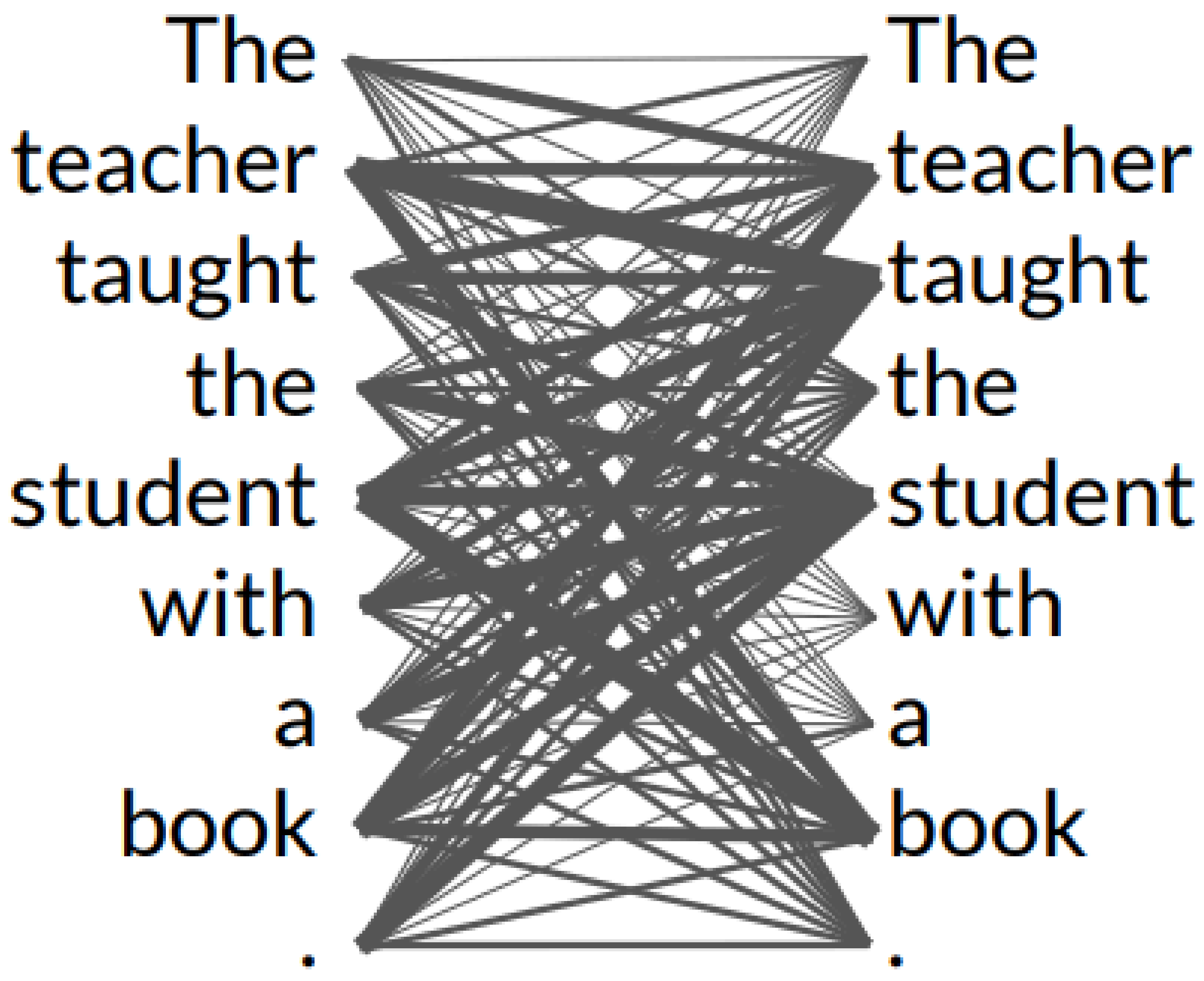


The teacher taught the student with the book.

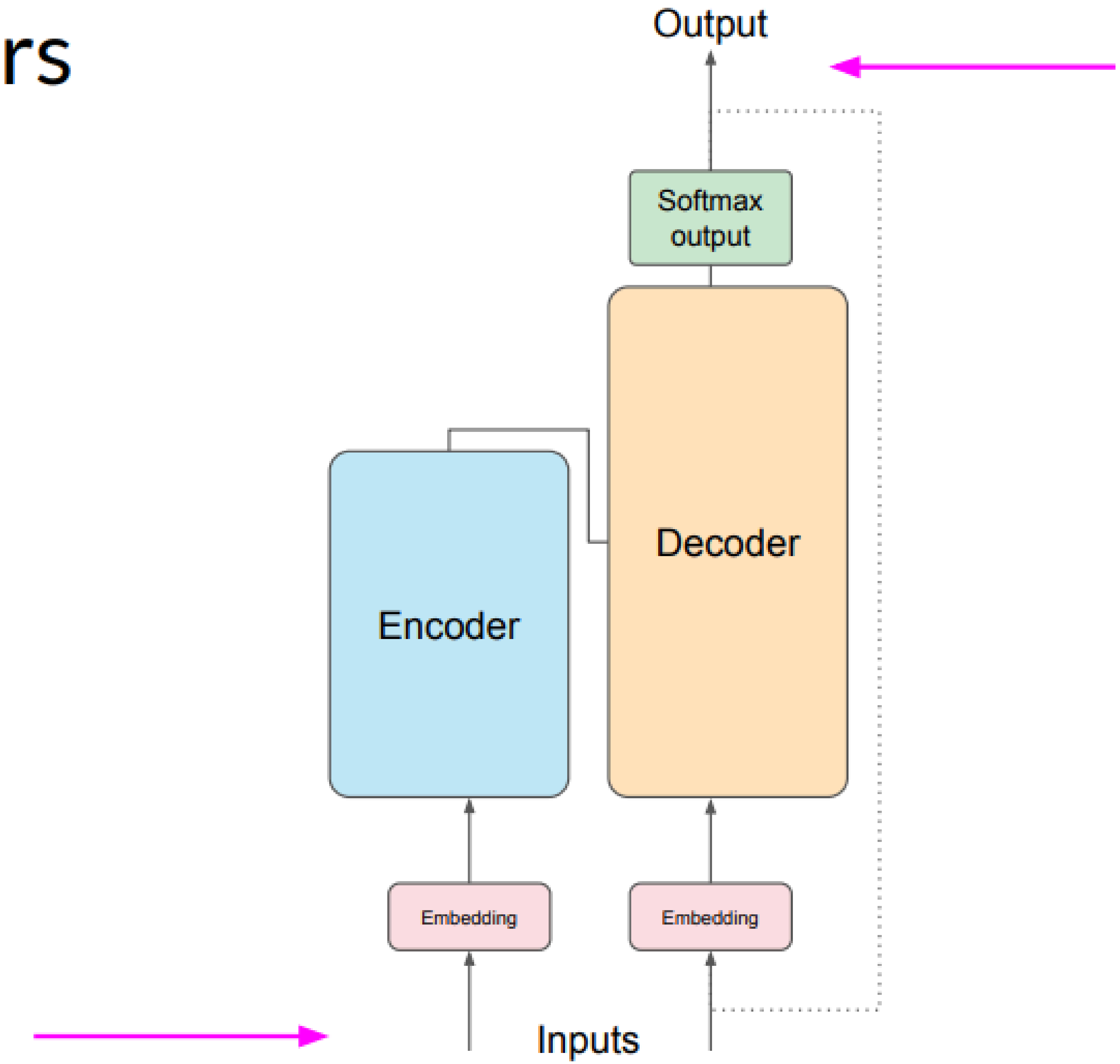
Self-attention



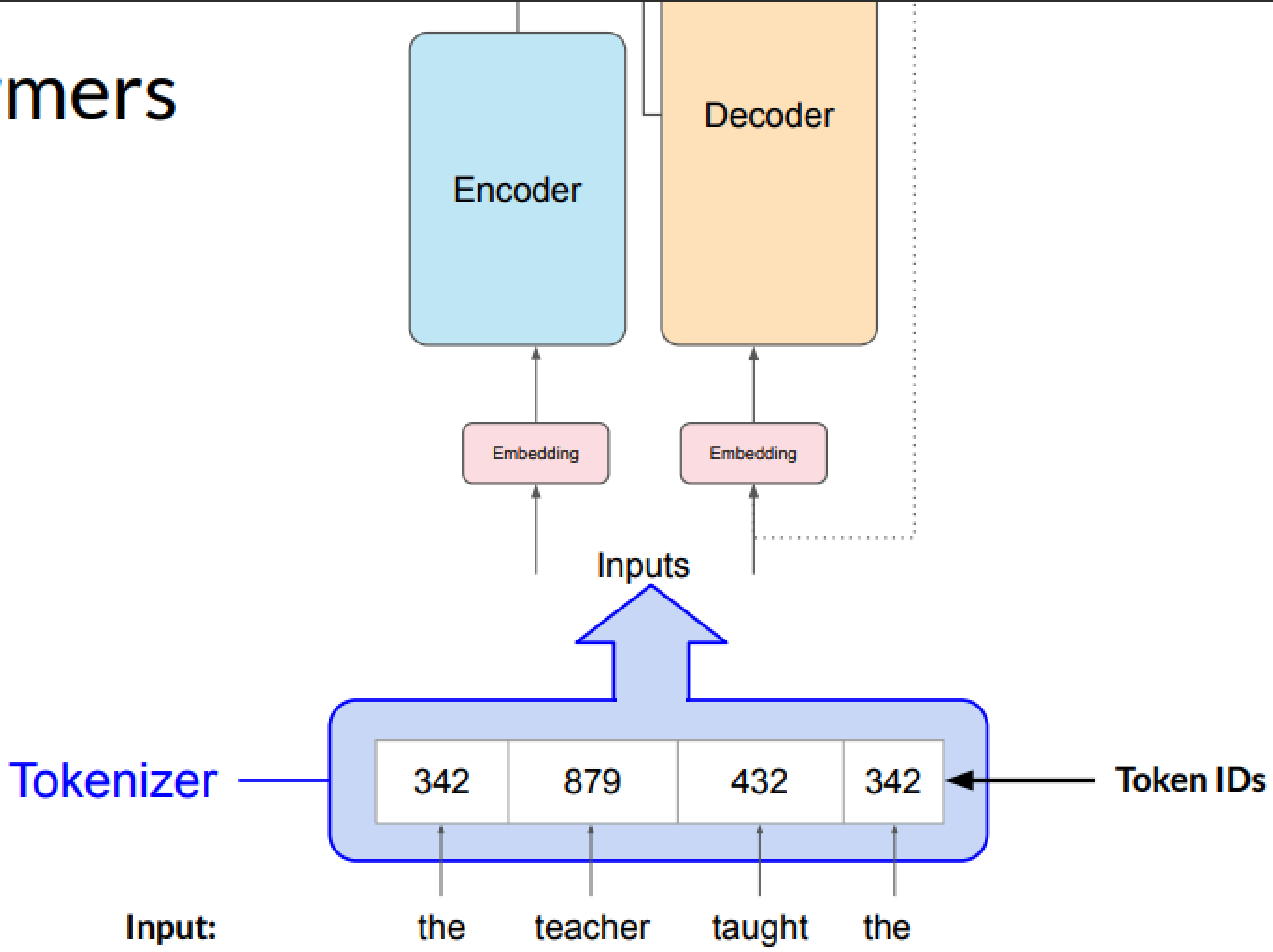
Self-attention



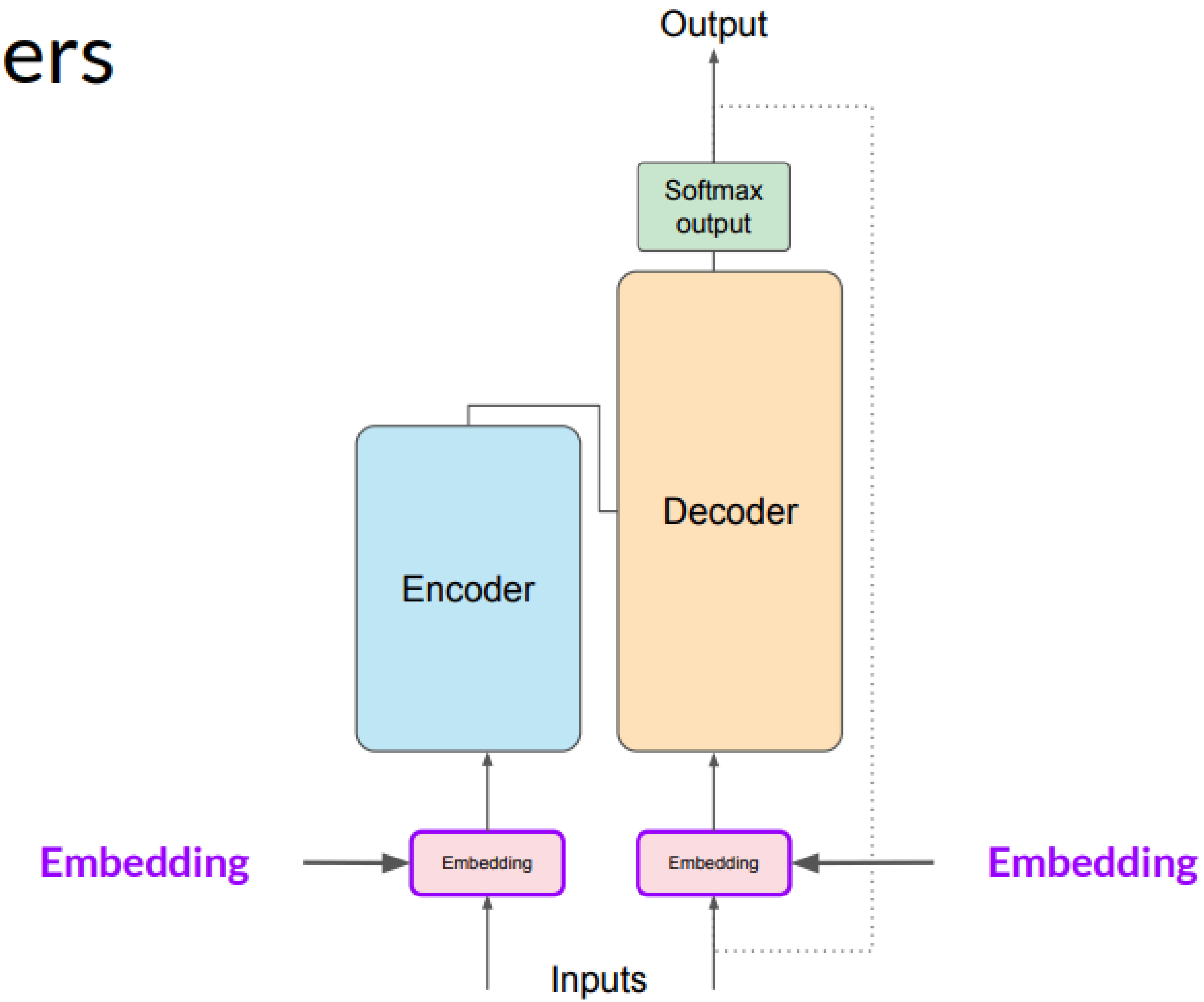
Transformers



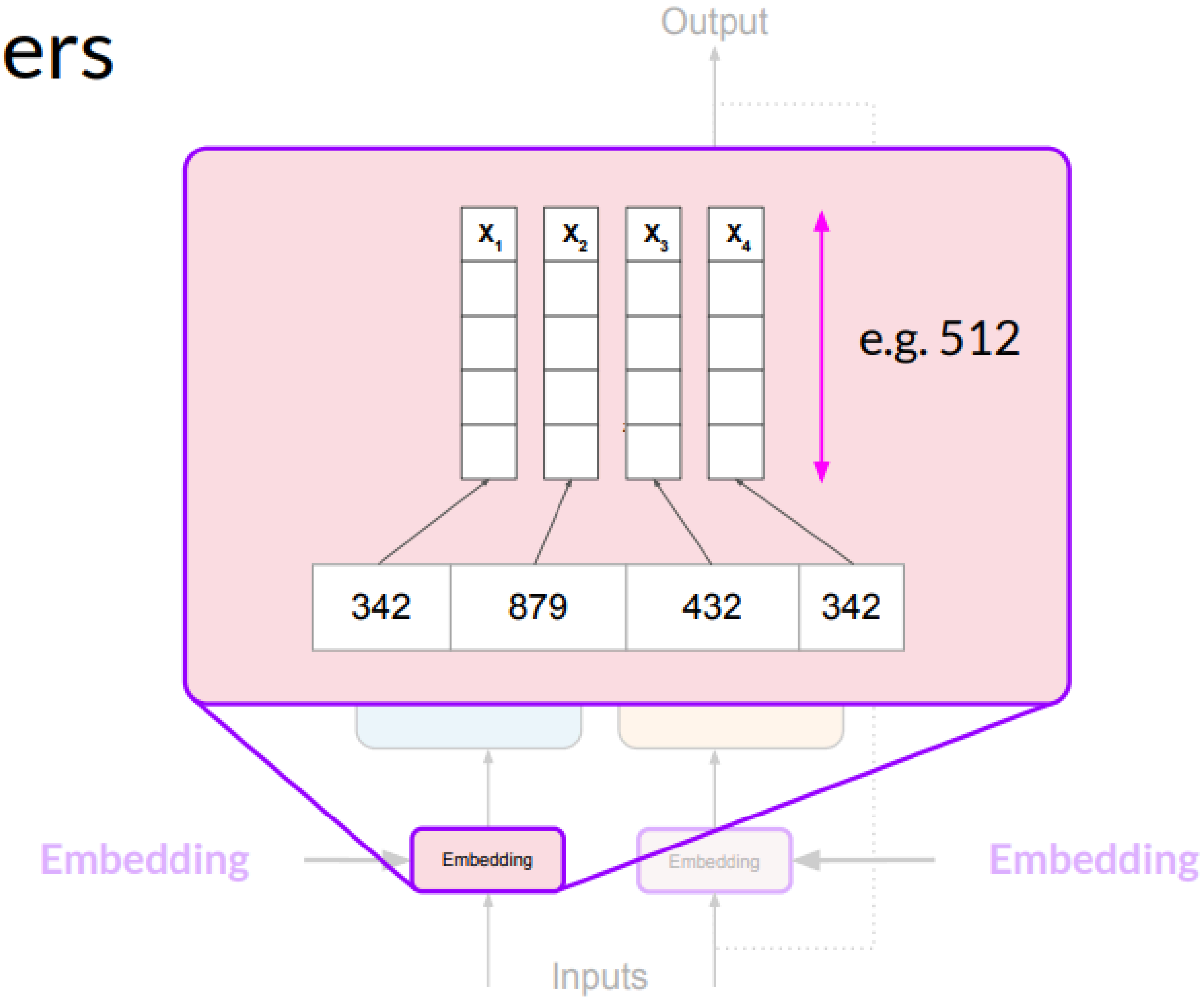
Transformers



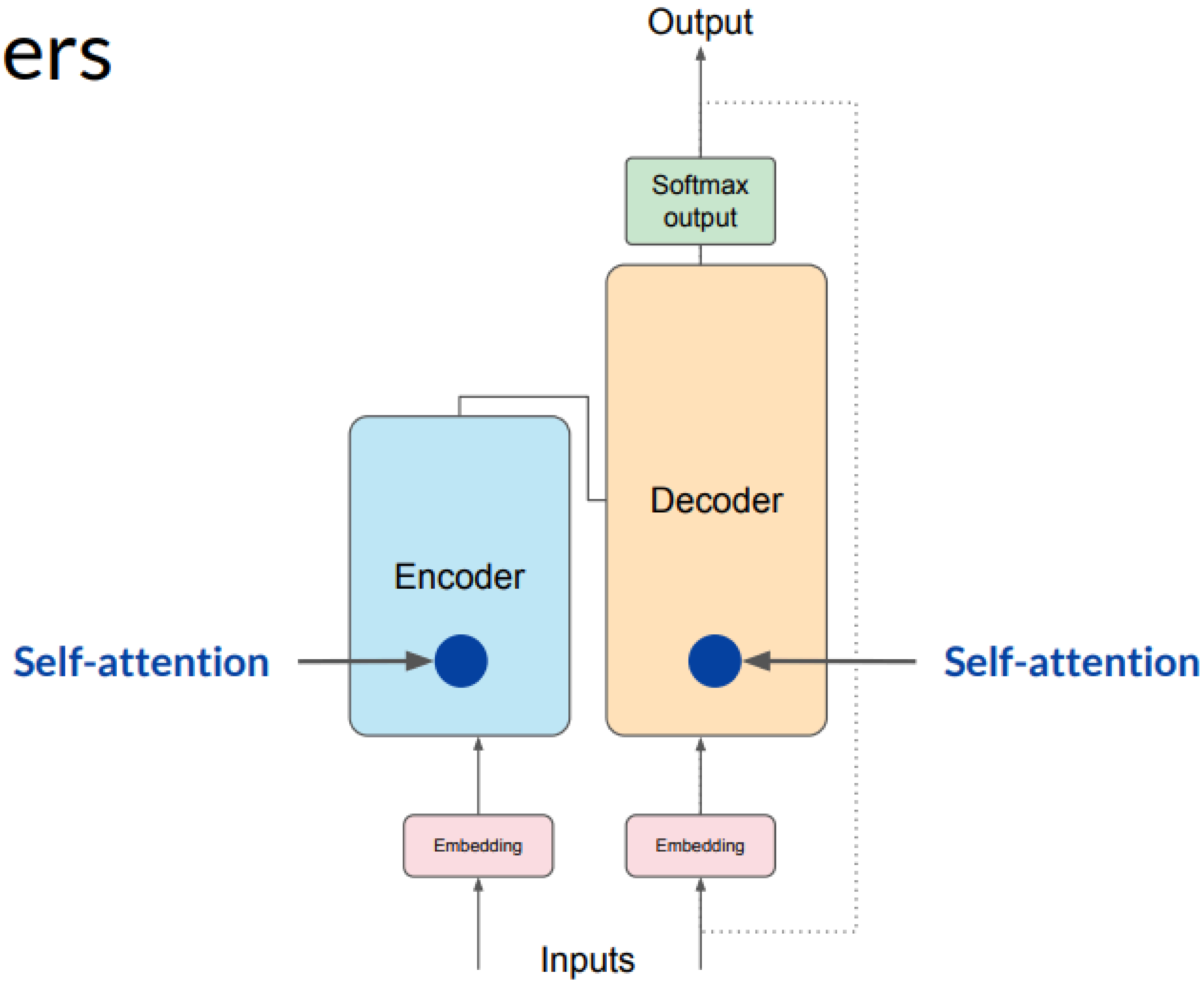
Transformers



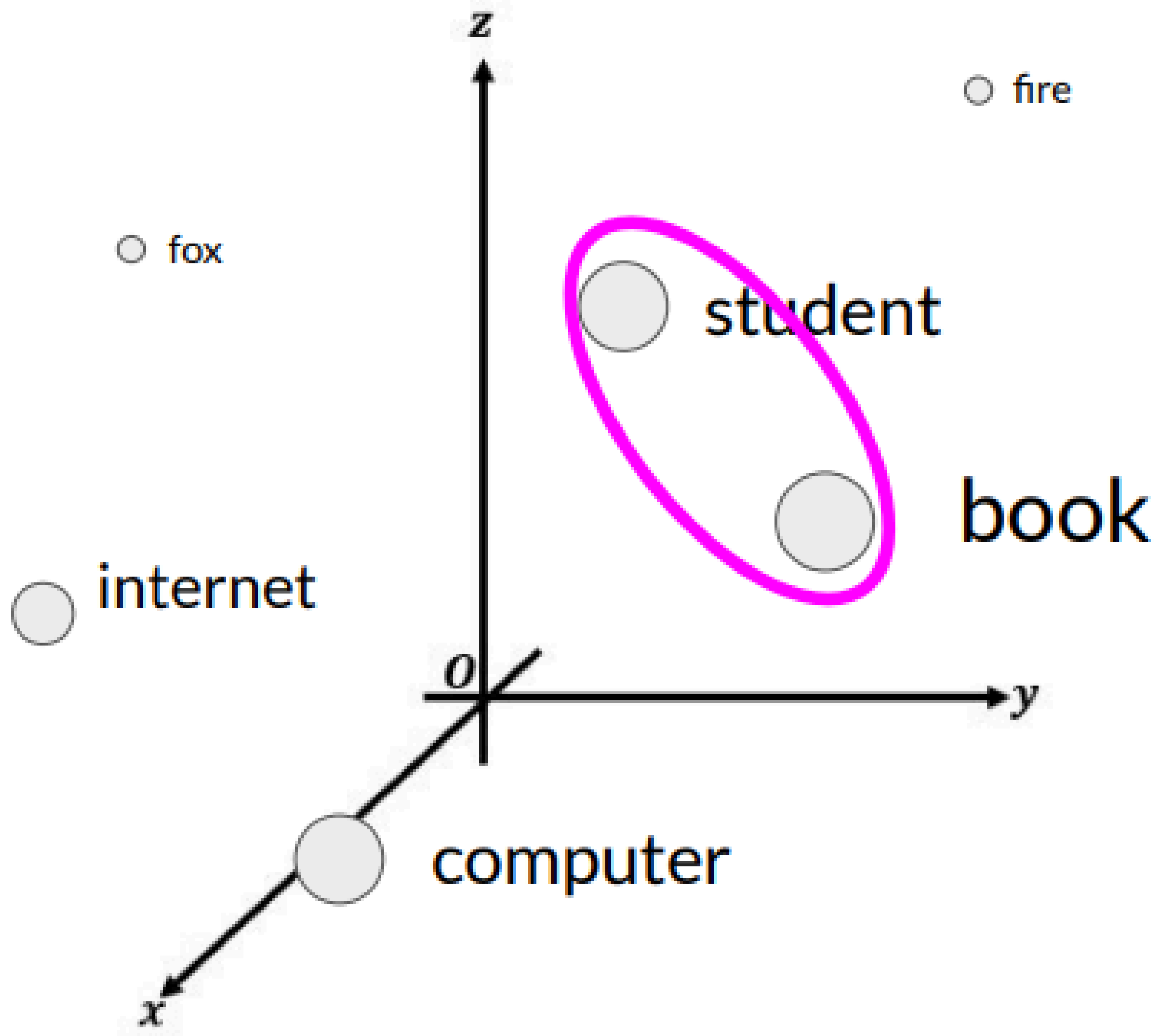
Transformers



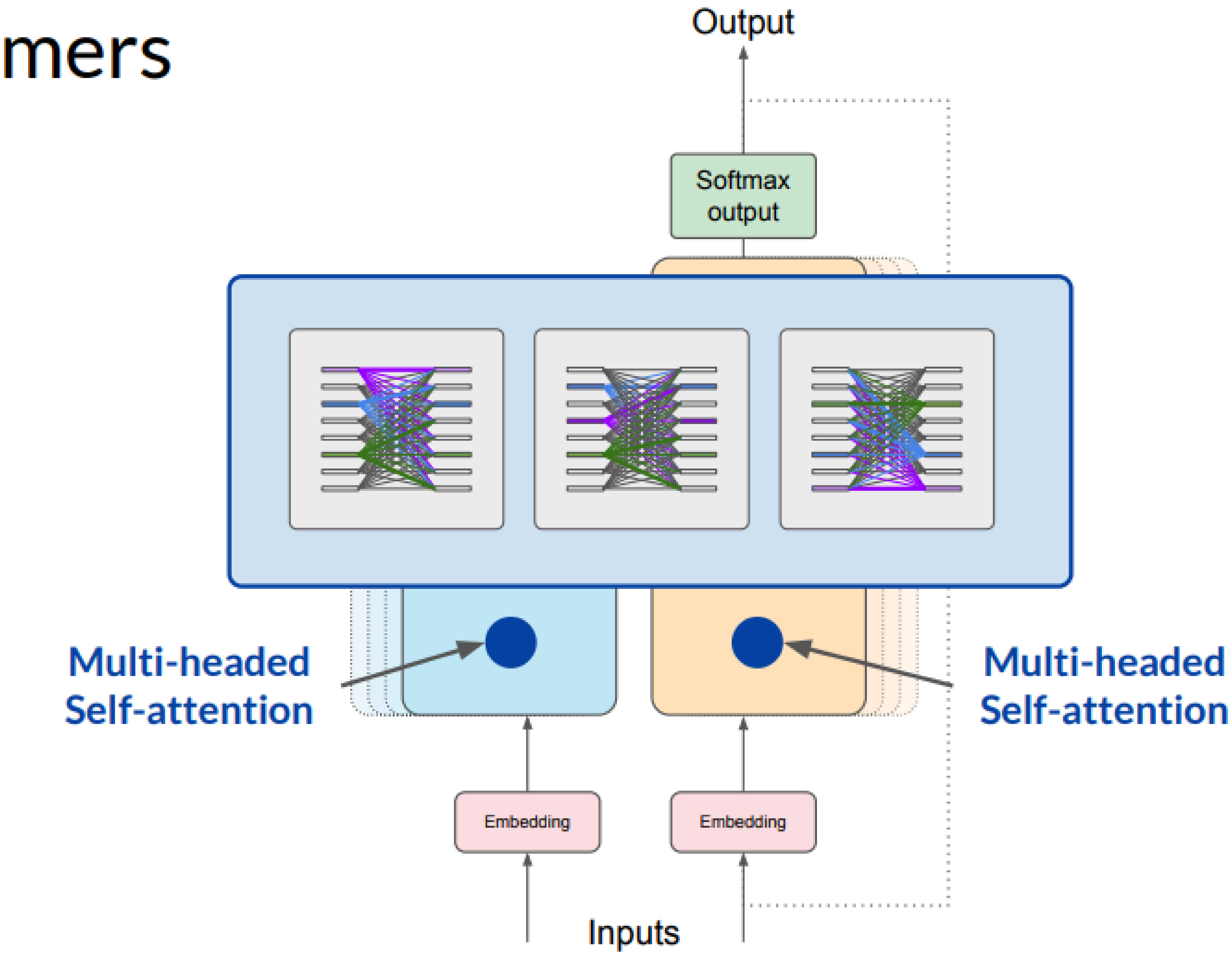
Transformers



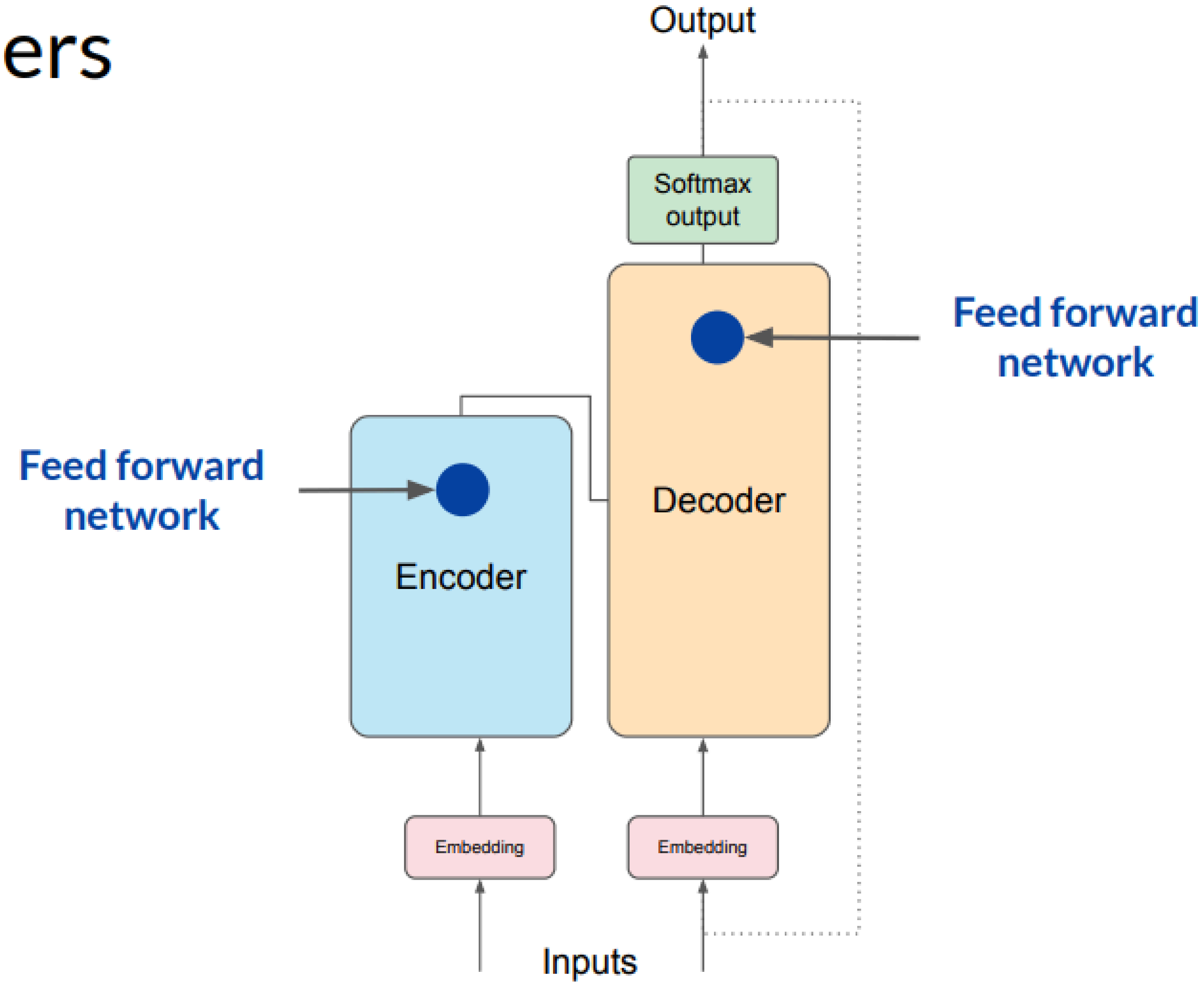
Transformers



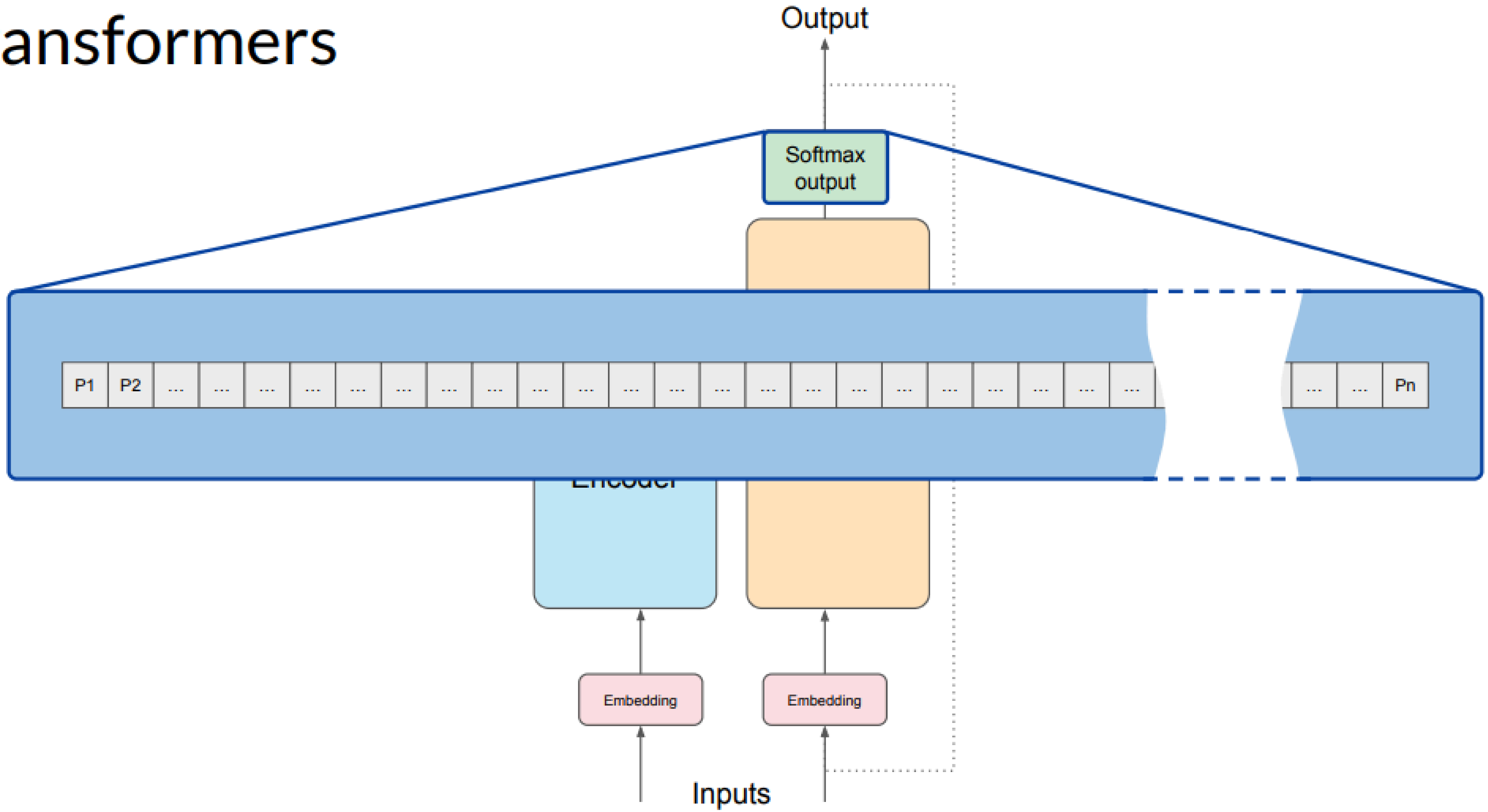
Transformers



Transformers

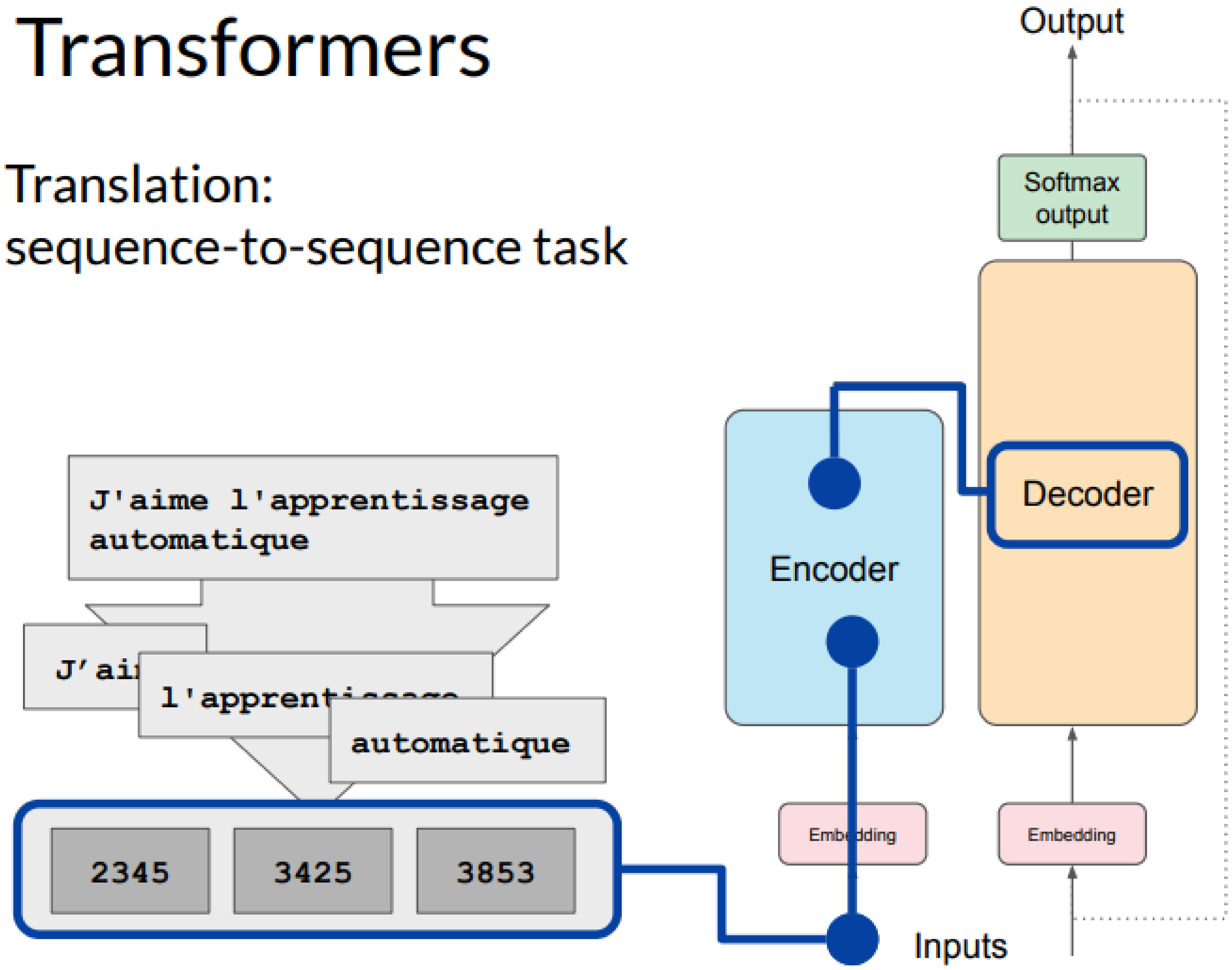


Transformers



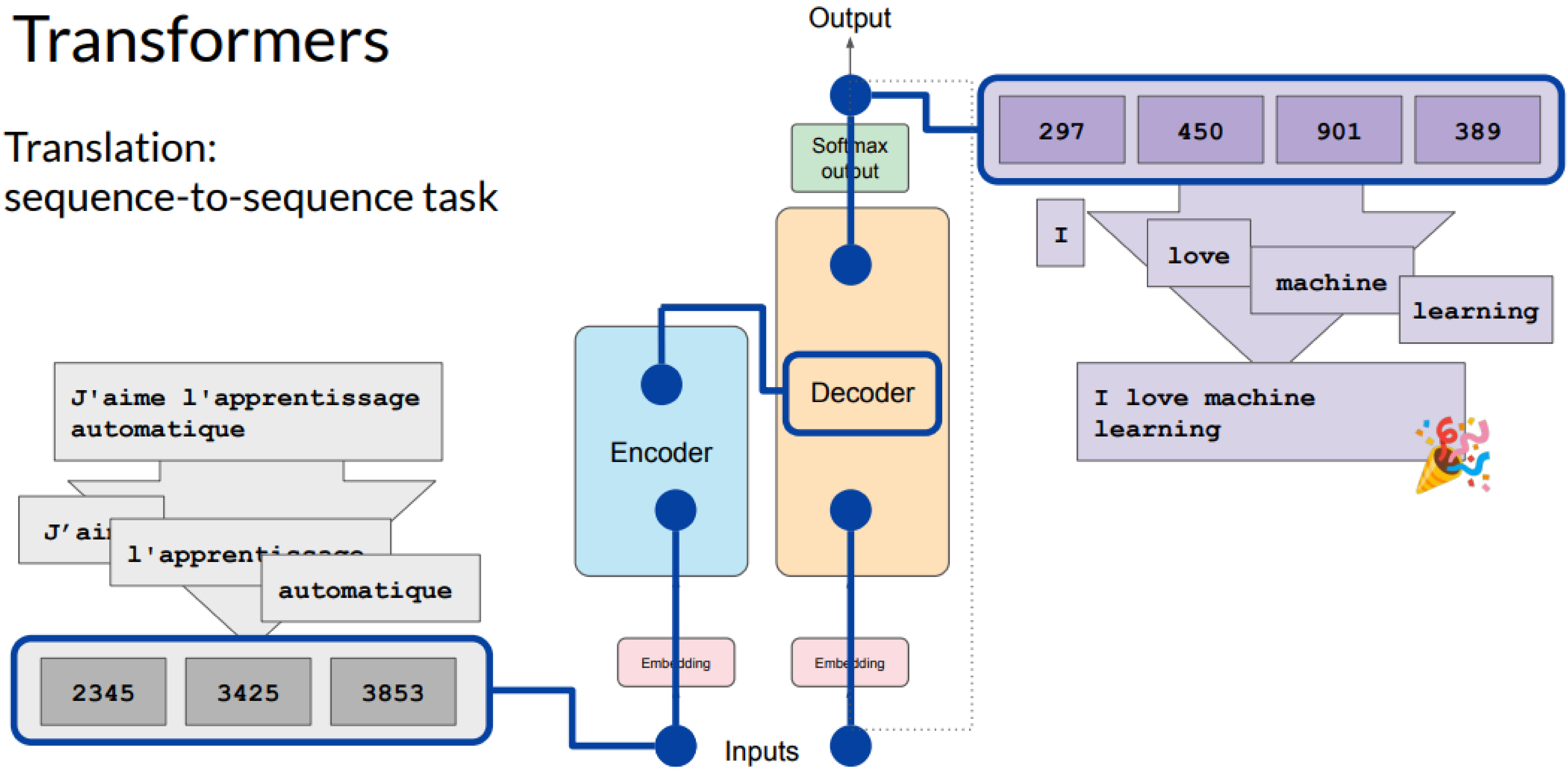
Transformers

Translation:
sequence-to-sequence task



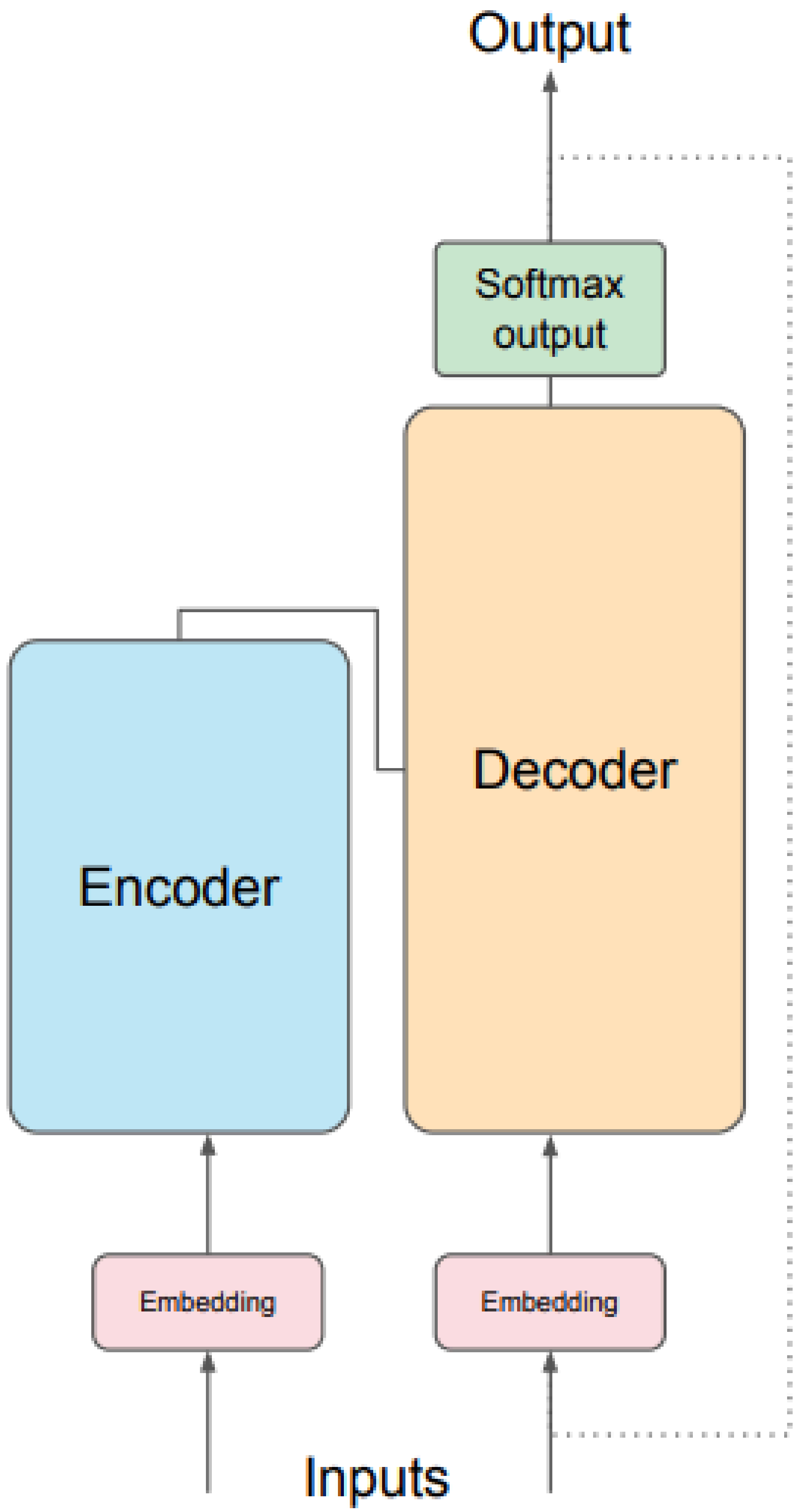
Transformers

Translation:
sequence-to-sequence task



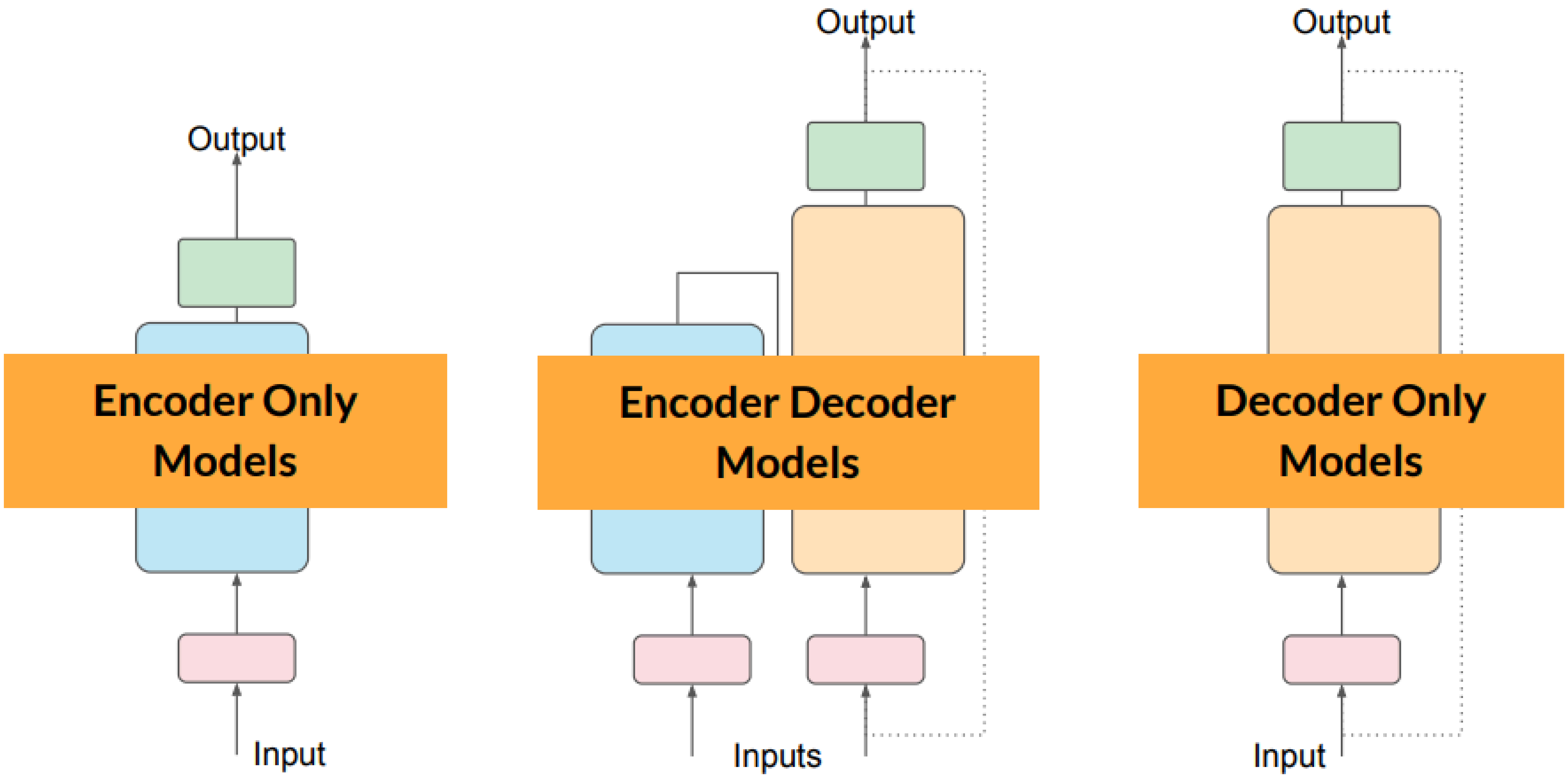
Transformers

Encoder
Encodes inputs ("prompts") with contextual understanding and produces one vector per input token.

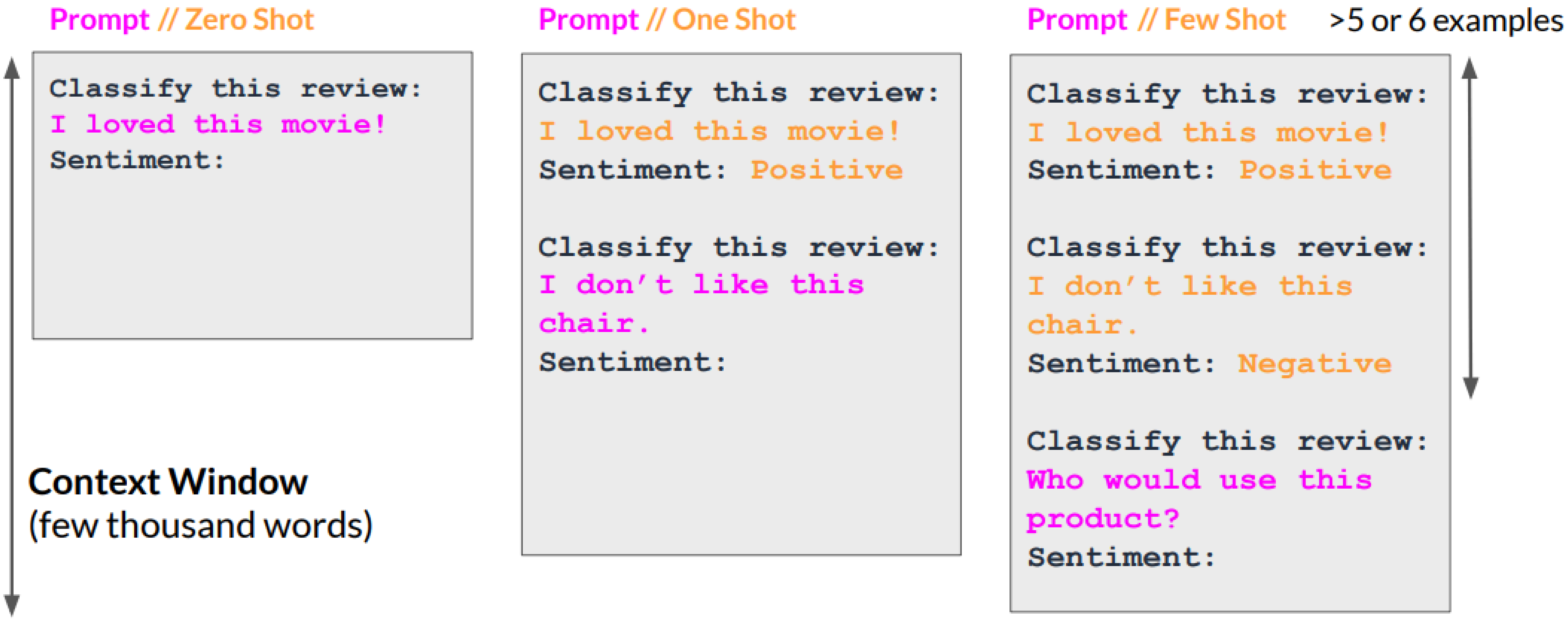


Decoder
Accepts input tokens and generates new tokens.

Transformers



Summary of in-context learning (ICL)



Prompt // Zero Shot

```
Classify this review:
I loved this movie!
Sentiment:
```

Prompt // One Shot

```
Classify this review:
I loved this movie!
Sentiment: Positive

Classify this review:
I don’t like this
chair.
Sentiment:
```

Prompt // Few Shot

>5 or 6 examples

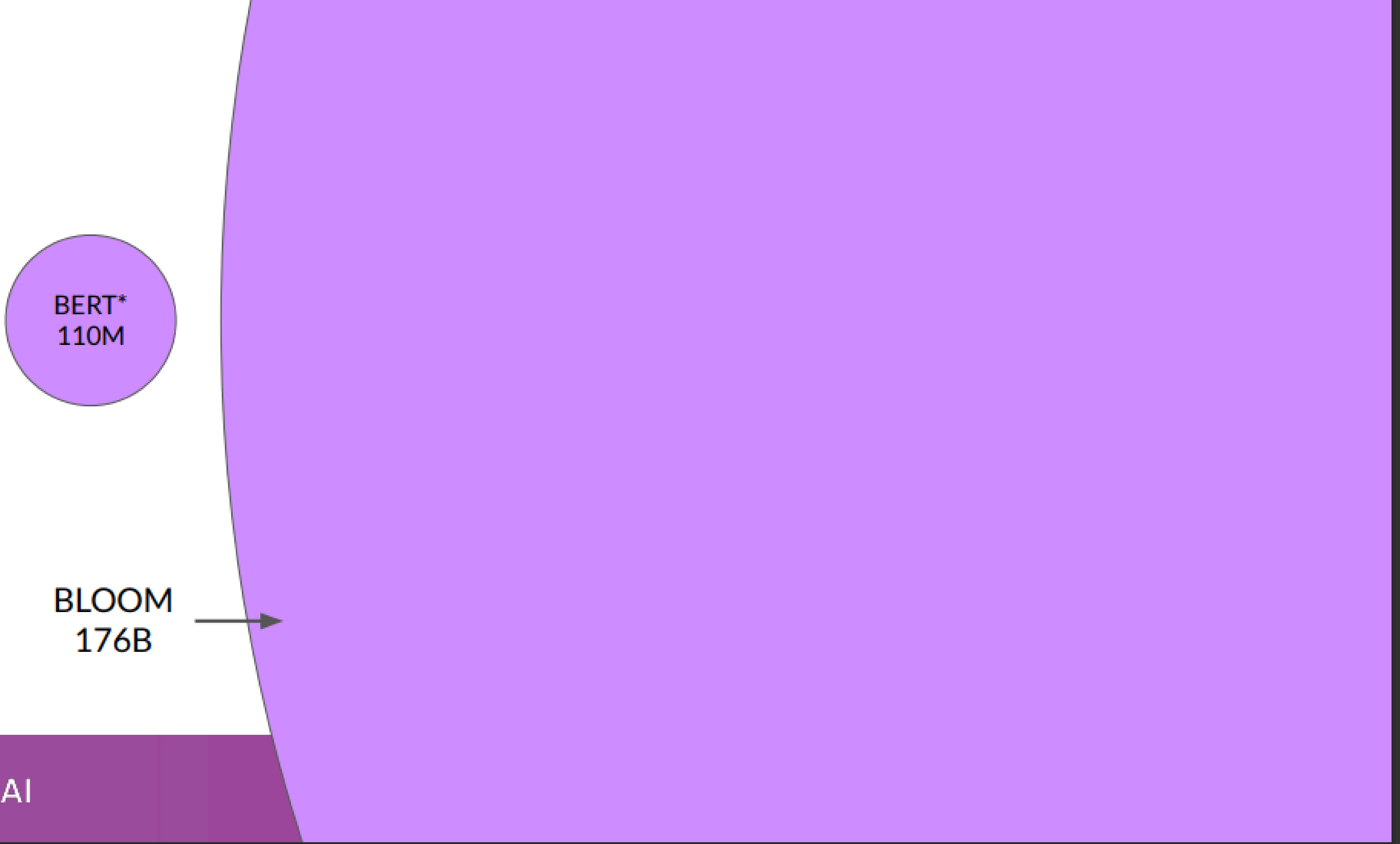
```
Classify this review:
I loved this movie!
Sentiment: Positive

Classify this review:
I don’t like this
chair.
Sentiment: Negative

Classify this review:
Who would use this
product?
Sentiment:
```

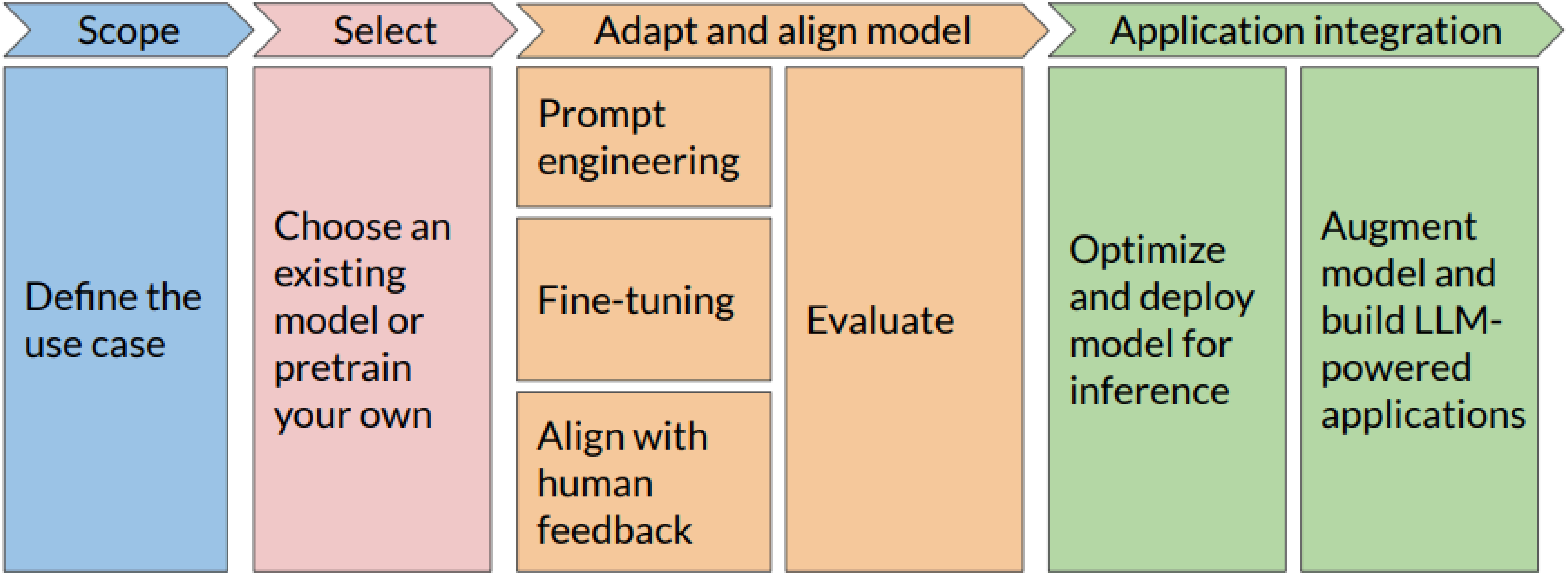
Context Window
(few thousand words)

The significance of scale: task ability



*Bert-base

Generative AI project lifecycle



Scope

Define the use case

Select

Choose an existing model or pretrain your own

Adapt and align model

Prompt engineering

Fine-tuning

Align with human feedback

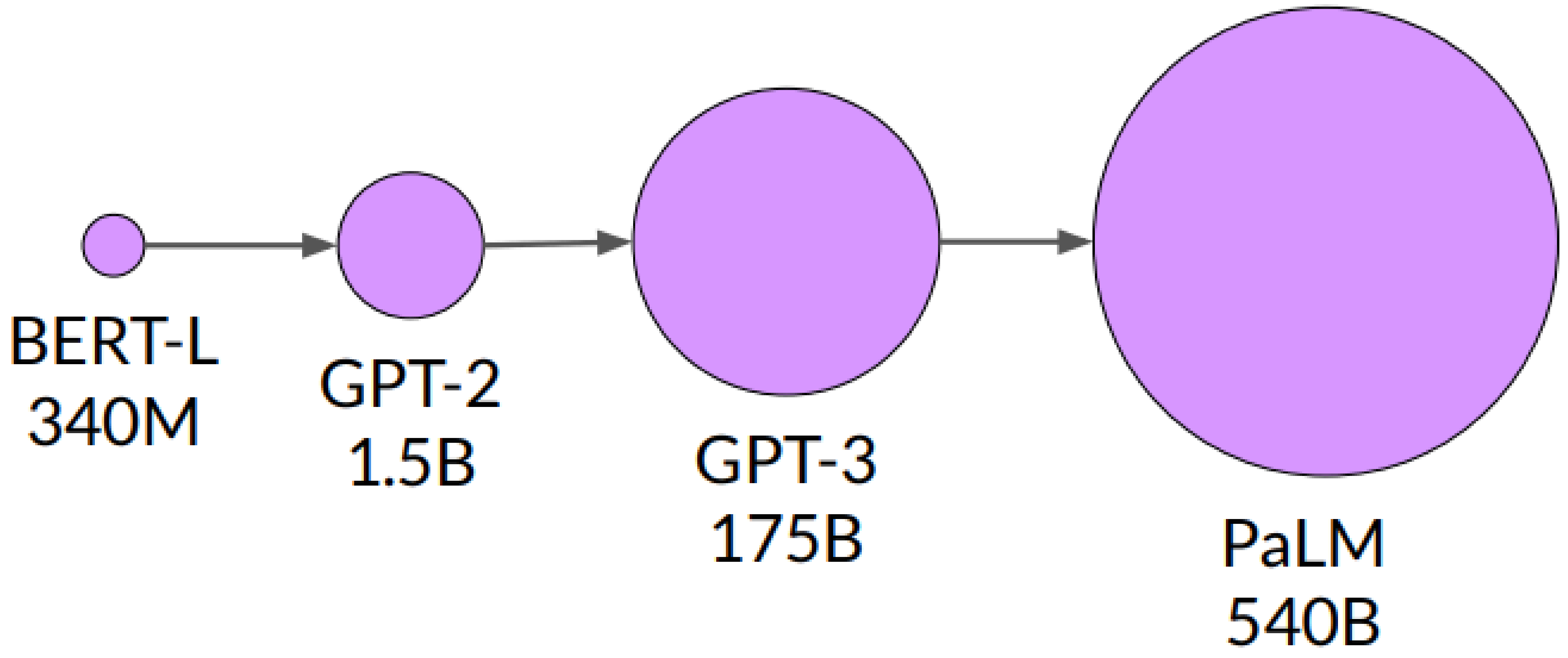
Evaluate

Application integration

Optimize and deploy model for inference

Augment model and build LLM-powered applications

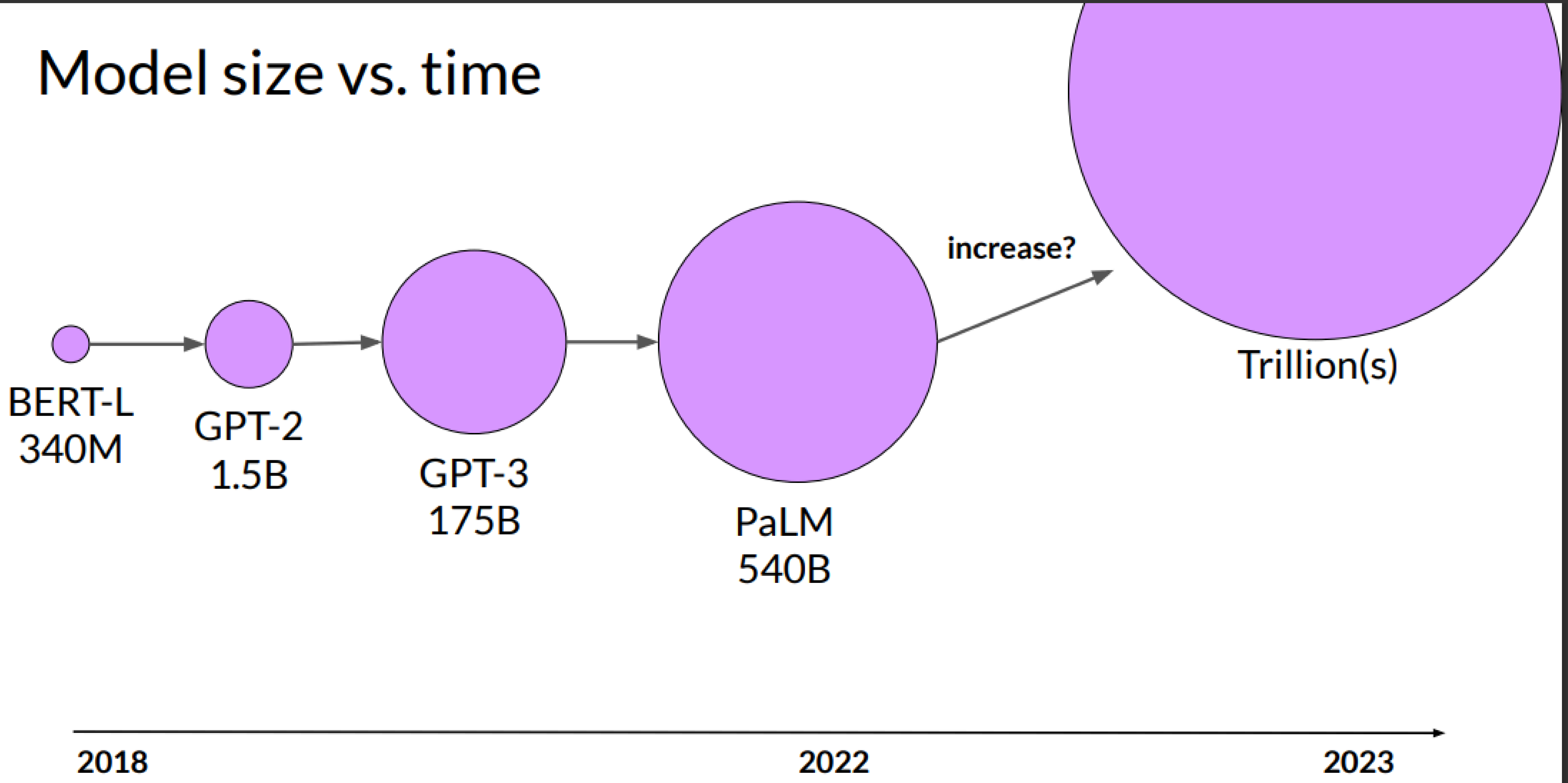
Model size vs. time



Growth powered by:

- Introduction of transformer
- Access to massive datasets
- More powerful compute resources

Model size vs. time



BERT-L
340M

GPT-2
1.5B

GPT-3
175B

PaLM
540B

increase?

Trillion(s)

2018

2022

2023