arjumand.younus@ucd.ie

@ArjumandYounus

Introduction to Large Language Models - Some Exploration With ChatGPT API

Dr. Arjumand Younus

UCD DUBLIN

Google EMEA Anita Borg Scholar

Women Techmakers

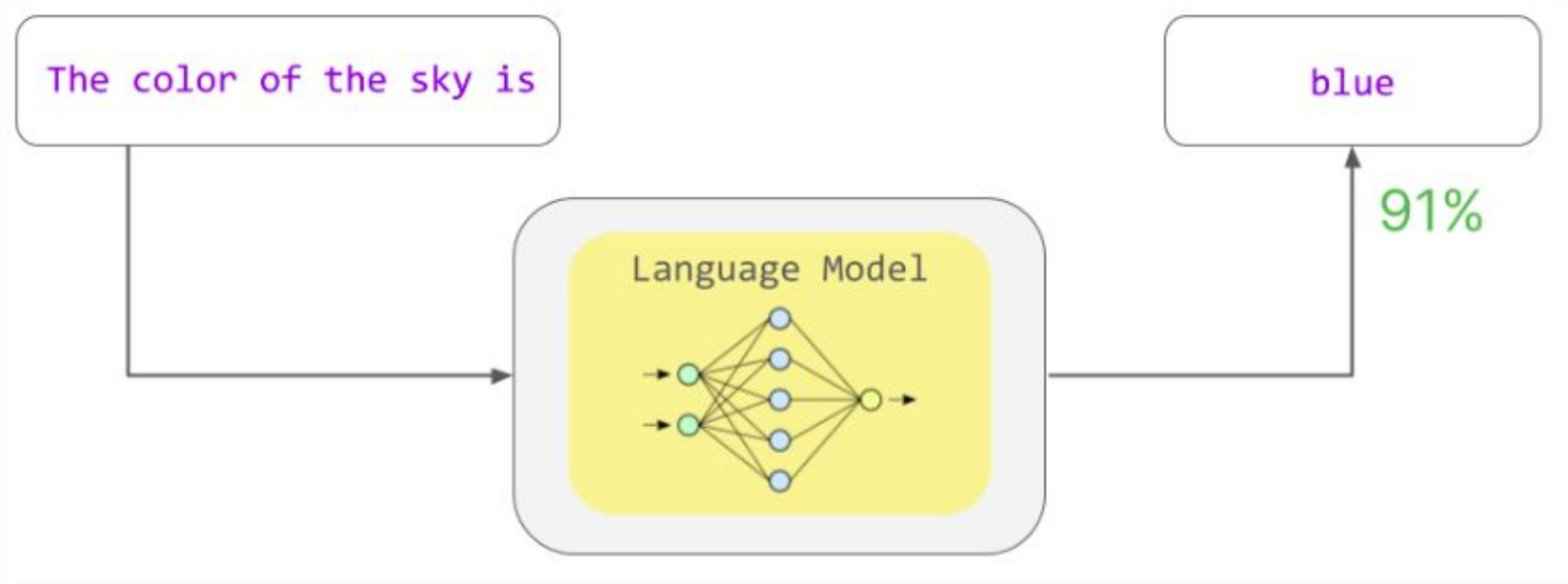
WOMEN IN RESEARCH IRELAND

Women in Research Ireland

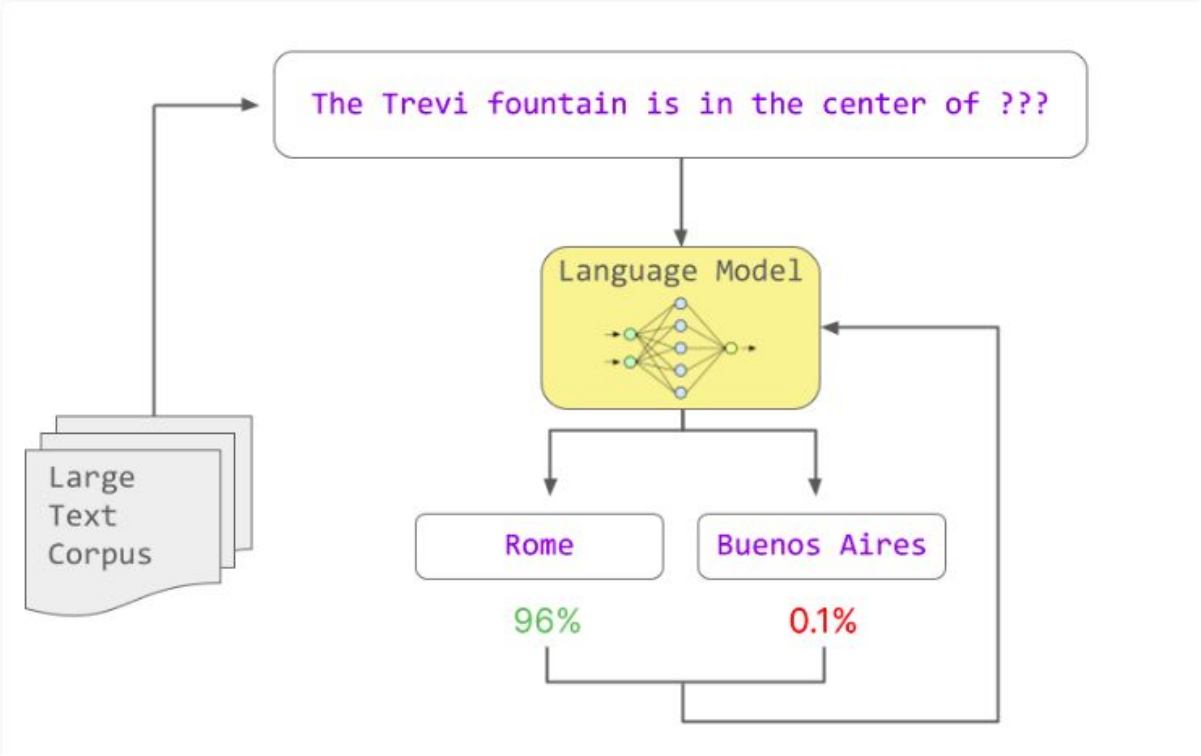
Generative AI and LLMs

- Ability to produce novel content
- Hard to distinguish from material produced by humans
- The generation of language modeled as a computational process
    - Power comes from massive computational architectures, training data, feedback mechanisms

Large Language Models: Under the Hood



Large Language Models: Under the Hood



Large Language Models: Text Generation Process

Supervised Learning (x -> y)

Restaurant reviews sentiment classification

Input x	Output y
The pastrami sandwich was great!	Positive
Service was slow and the food was so-so.	Negative
The earl grey tea was fantastic.	Positive

Supervised Learning (x -> y)

Restaurant reviews sentiment classification

Input x	Output y
The pastrami sandwich was great!	Positive
Service was slow and the food was so-so.	Negative
The earl grey tea was fantastic.	Positive
Best pizza I've ever had!	Positive

Not a New Technology

- Based on an architecture called Transformer Model

Before Transformers...

- Recurrent Neural Networks
  - Limited by amount of computer and memory needed to perform well at generative tasks

Challenge of Natural Language

I took my money to the bank.

River bank?

The teacher's book?

The teacher taught the student with the book.

The student's book?

Transformers

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

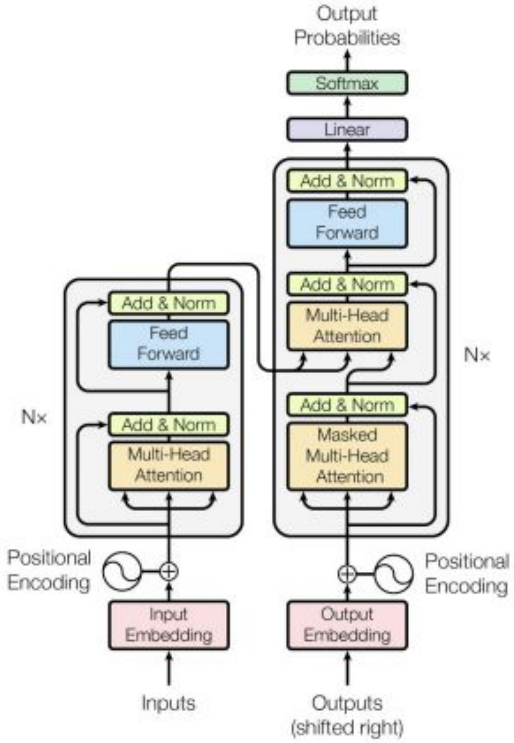
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

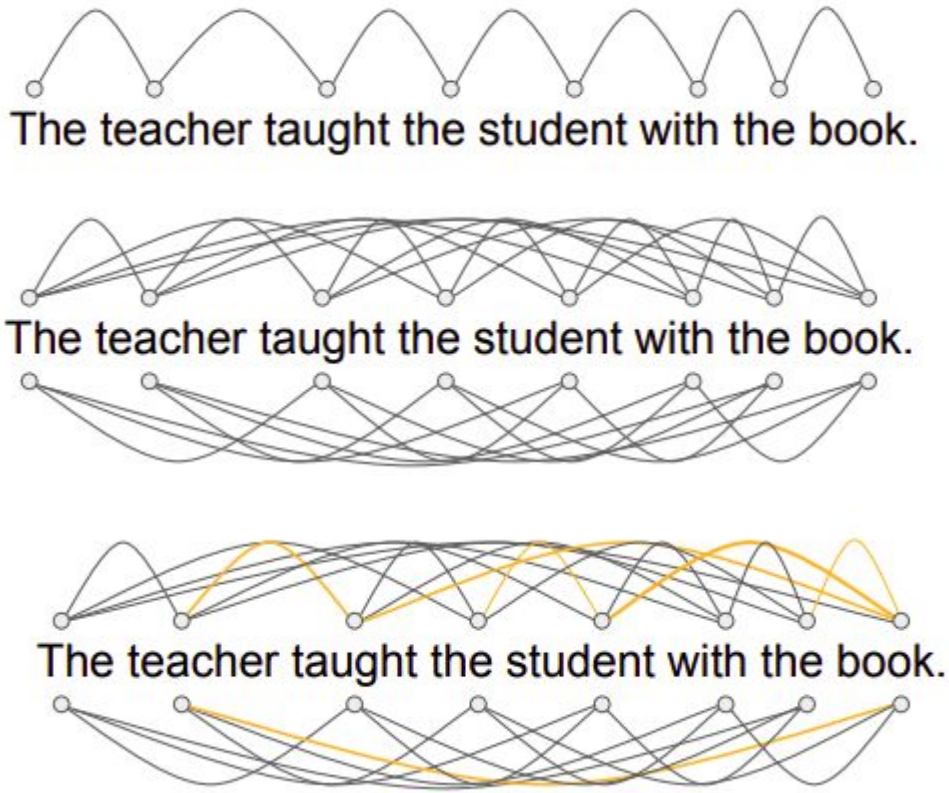
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

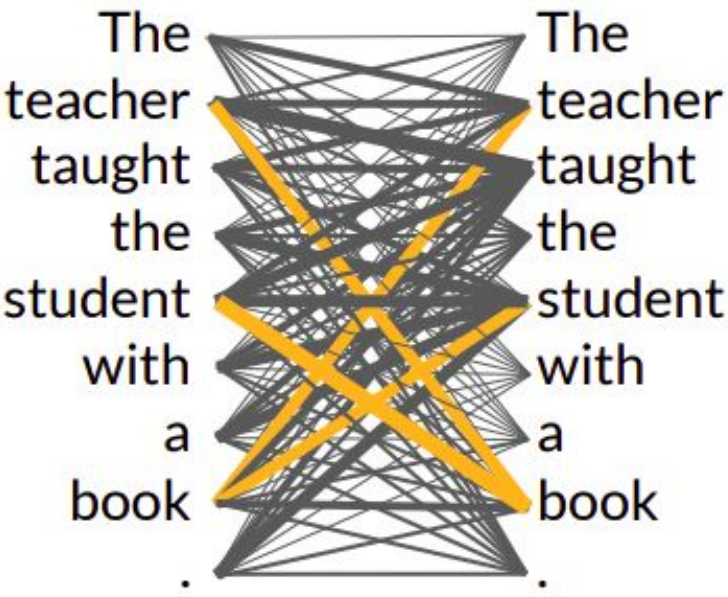
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to



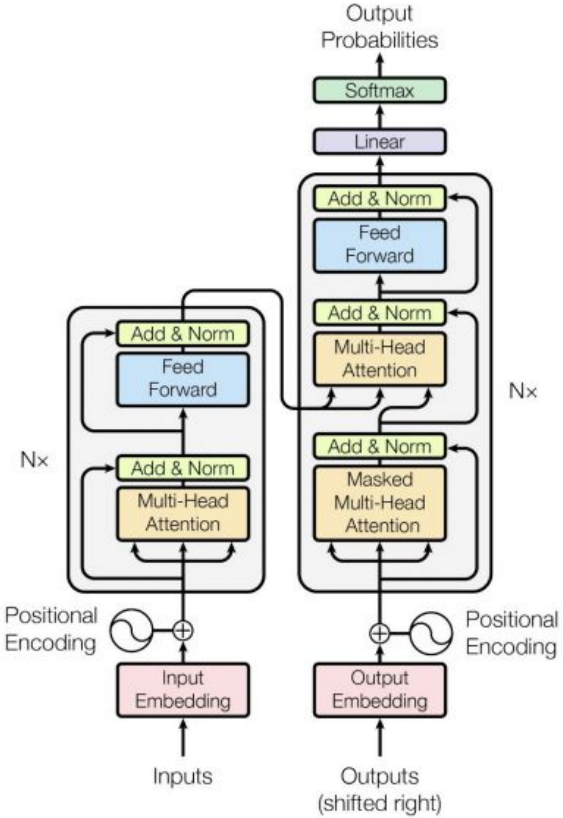
Transformers



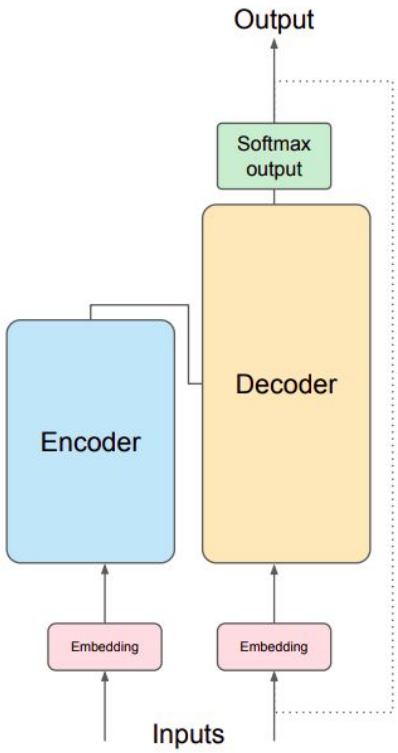
Transformers



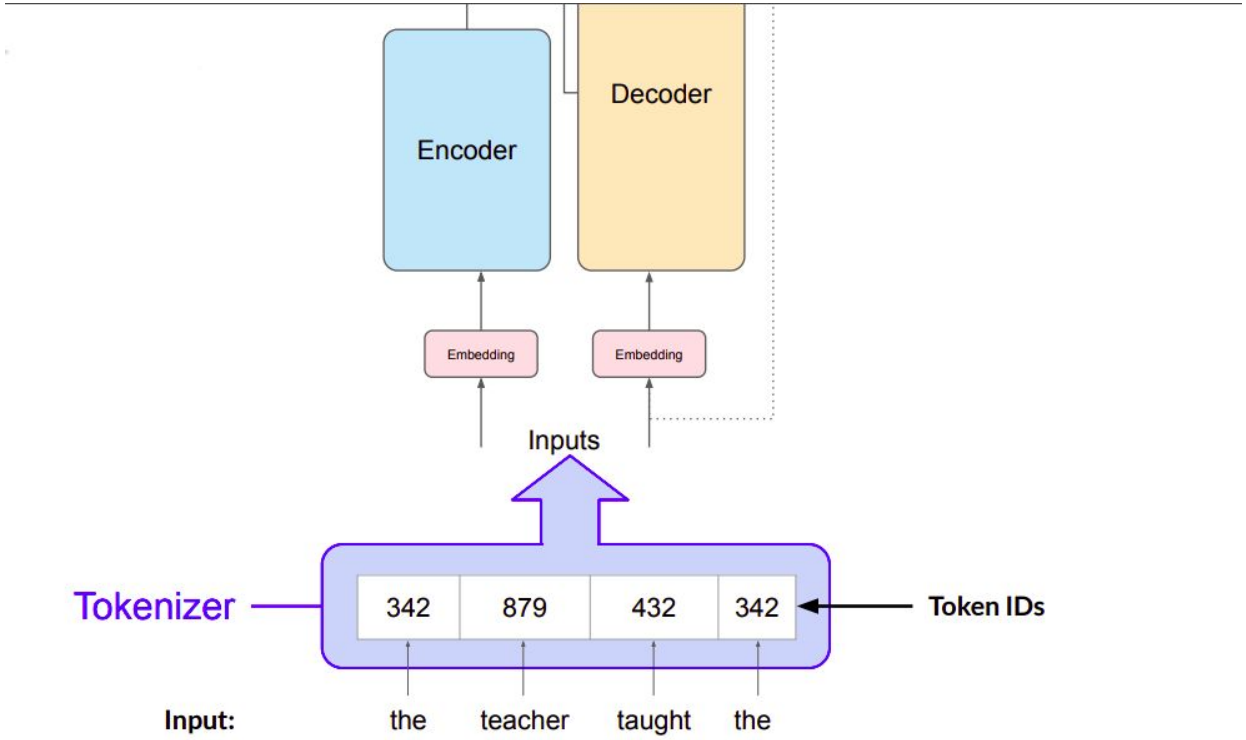
Transformers: How They Work



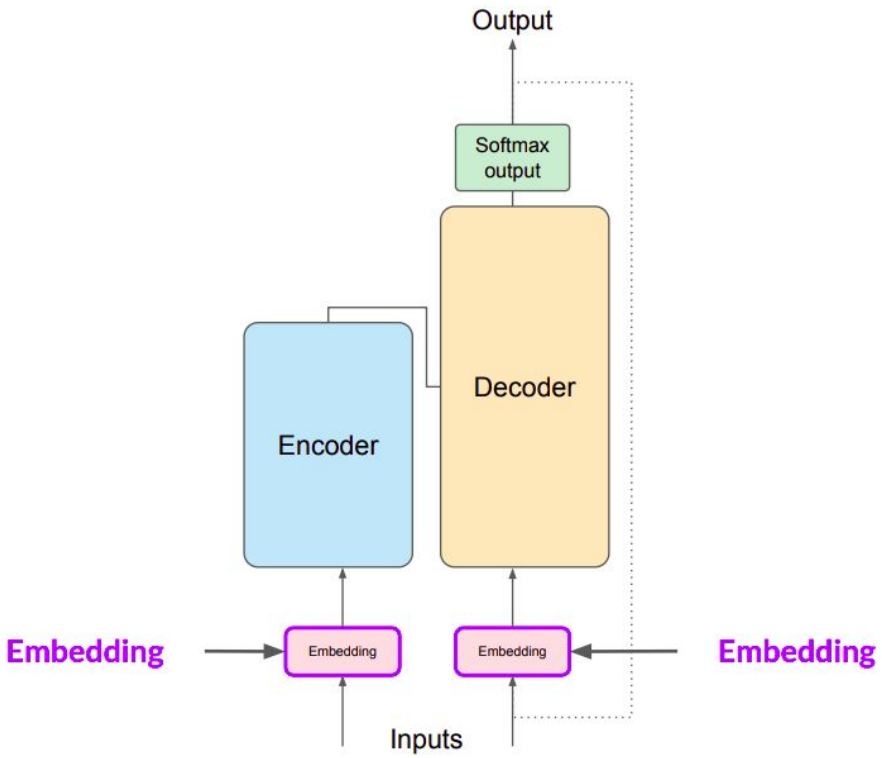
Transformers: How They Work



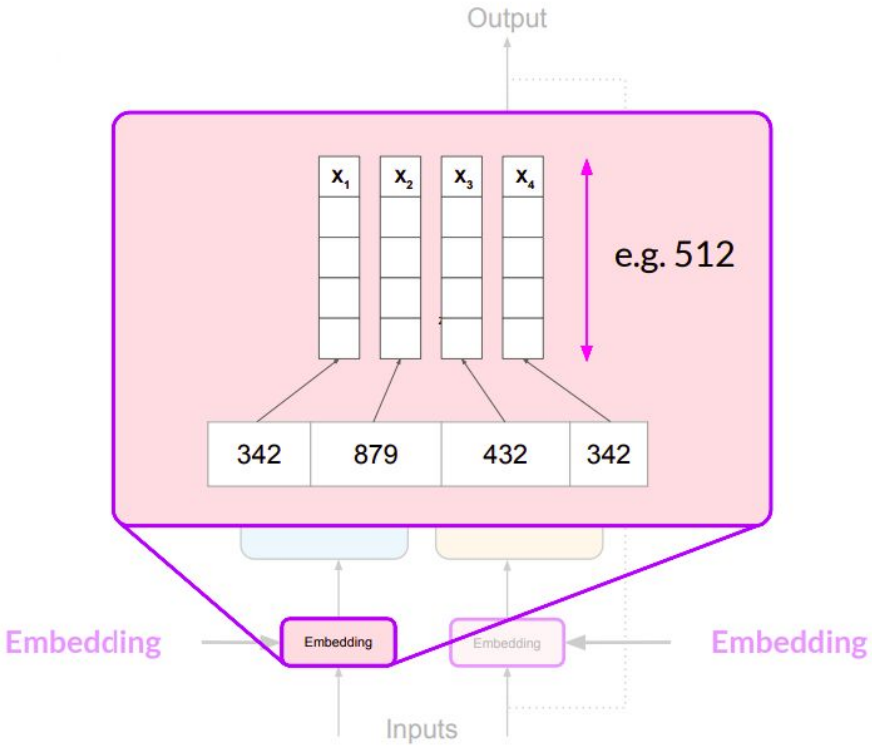
Transformers: How They Work



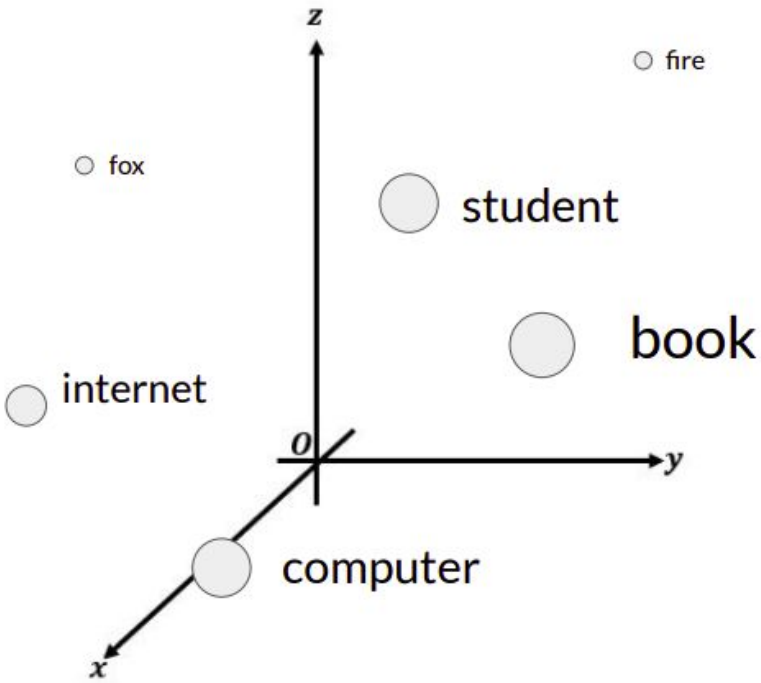
Transformers: How They Work



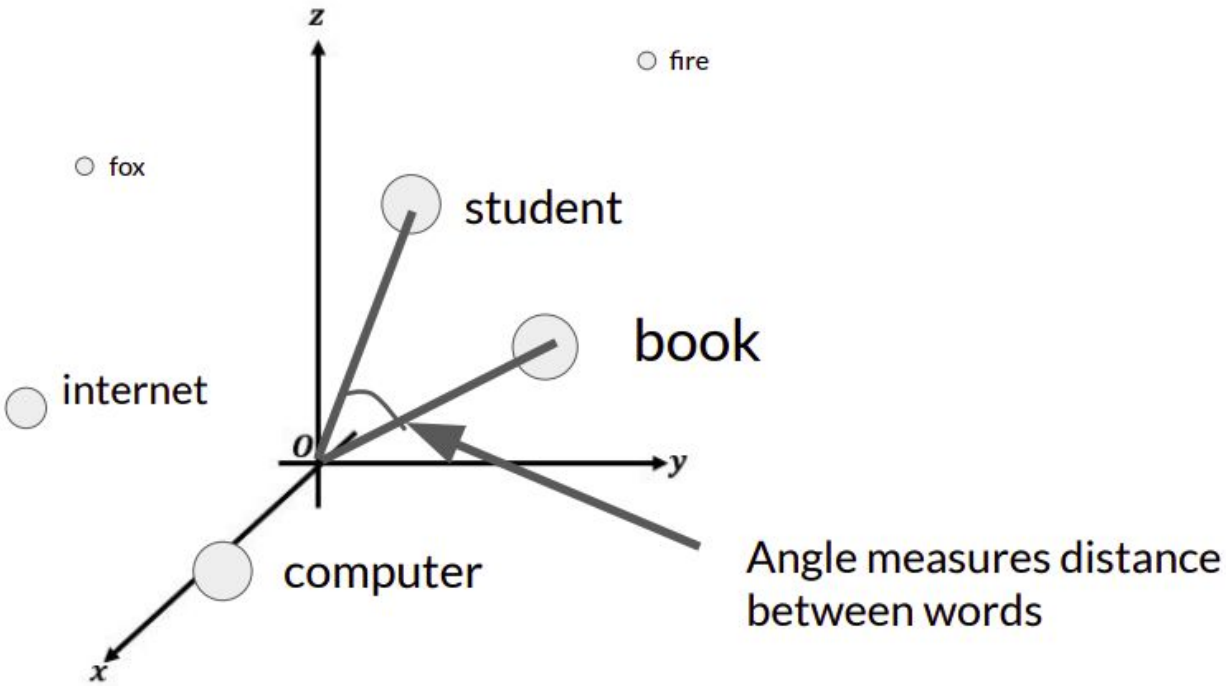
Transformers: How They Work



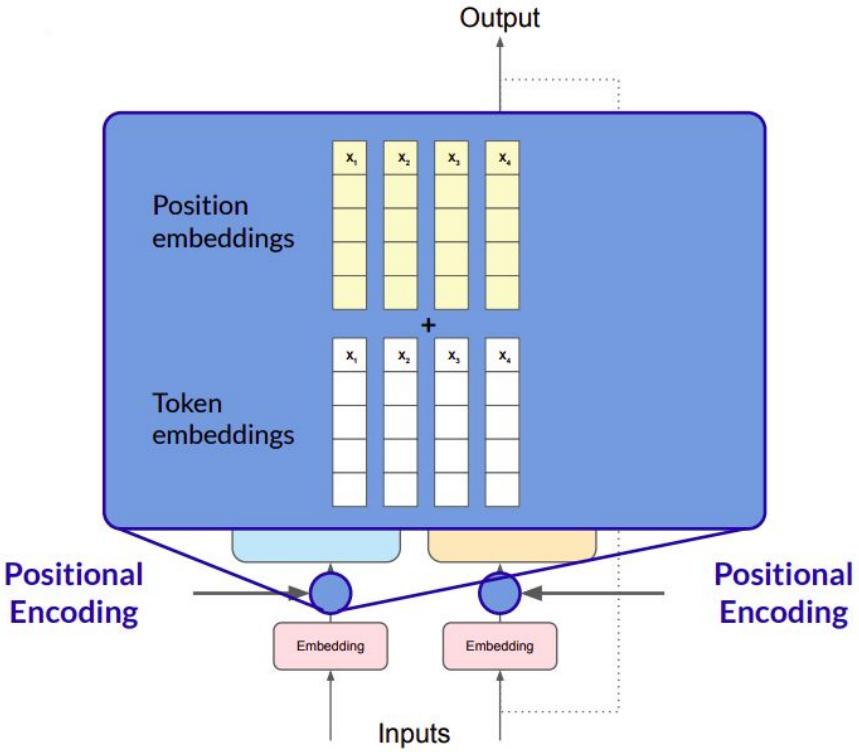
Transformers: How They Work



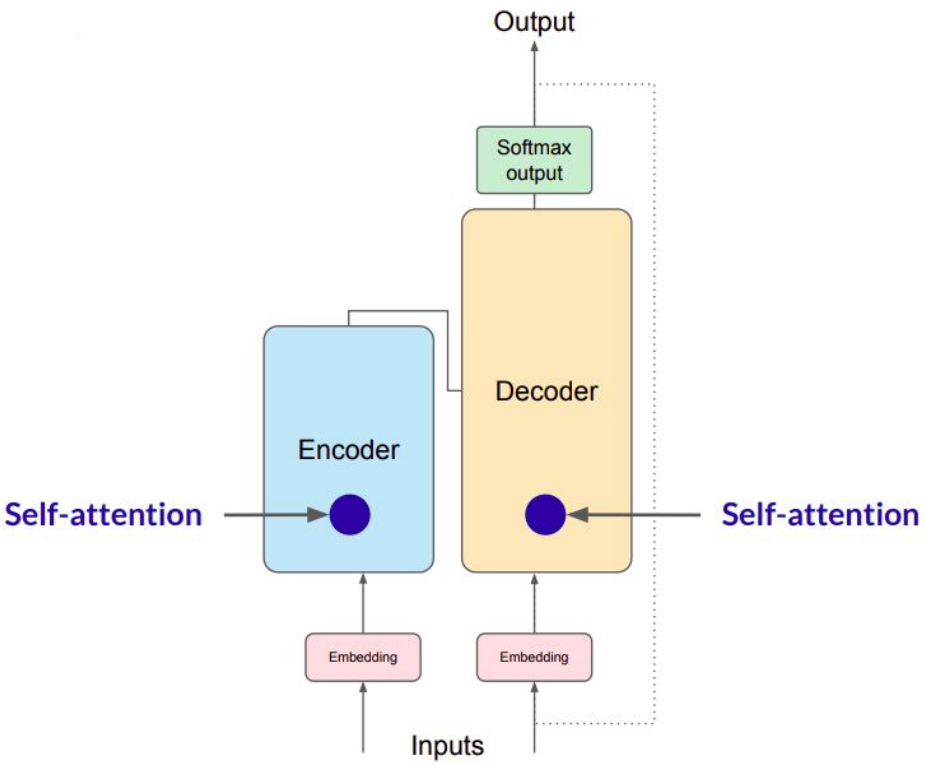
Transformers: How They Work



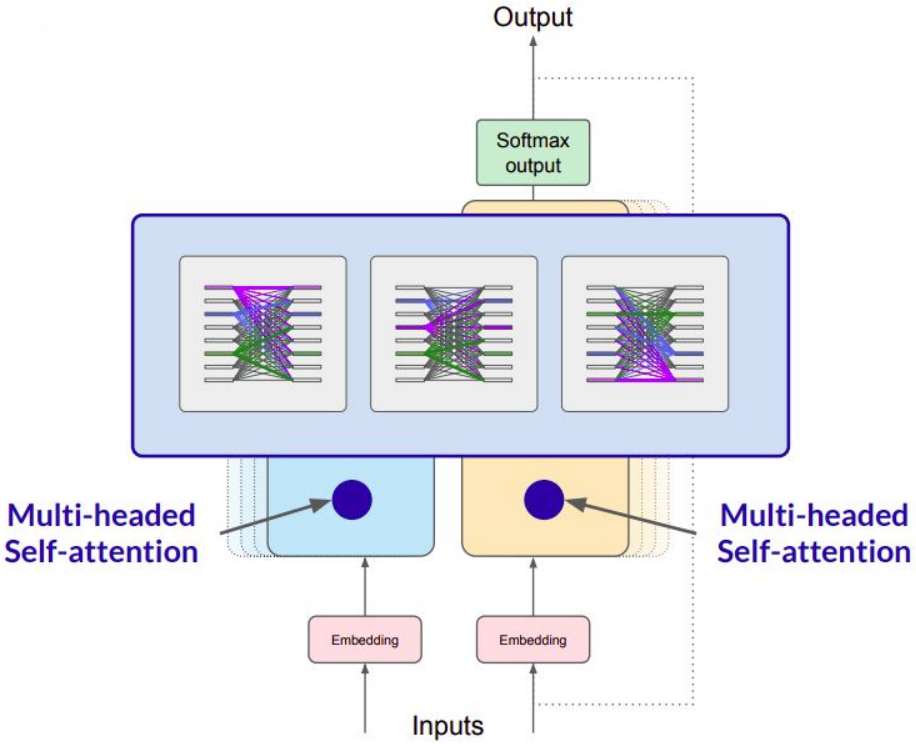
Transformers: How They Work



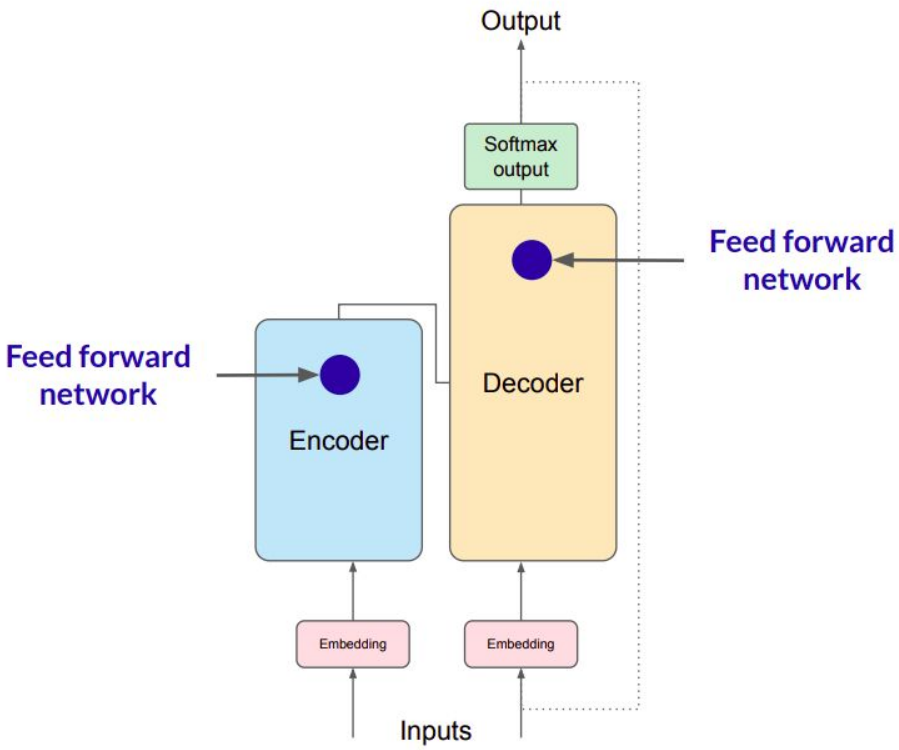
Transformers: How They Work



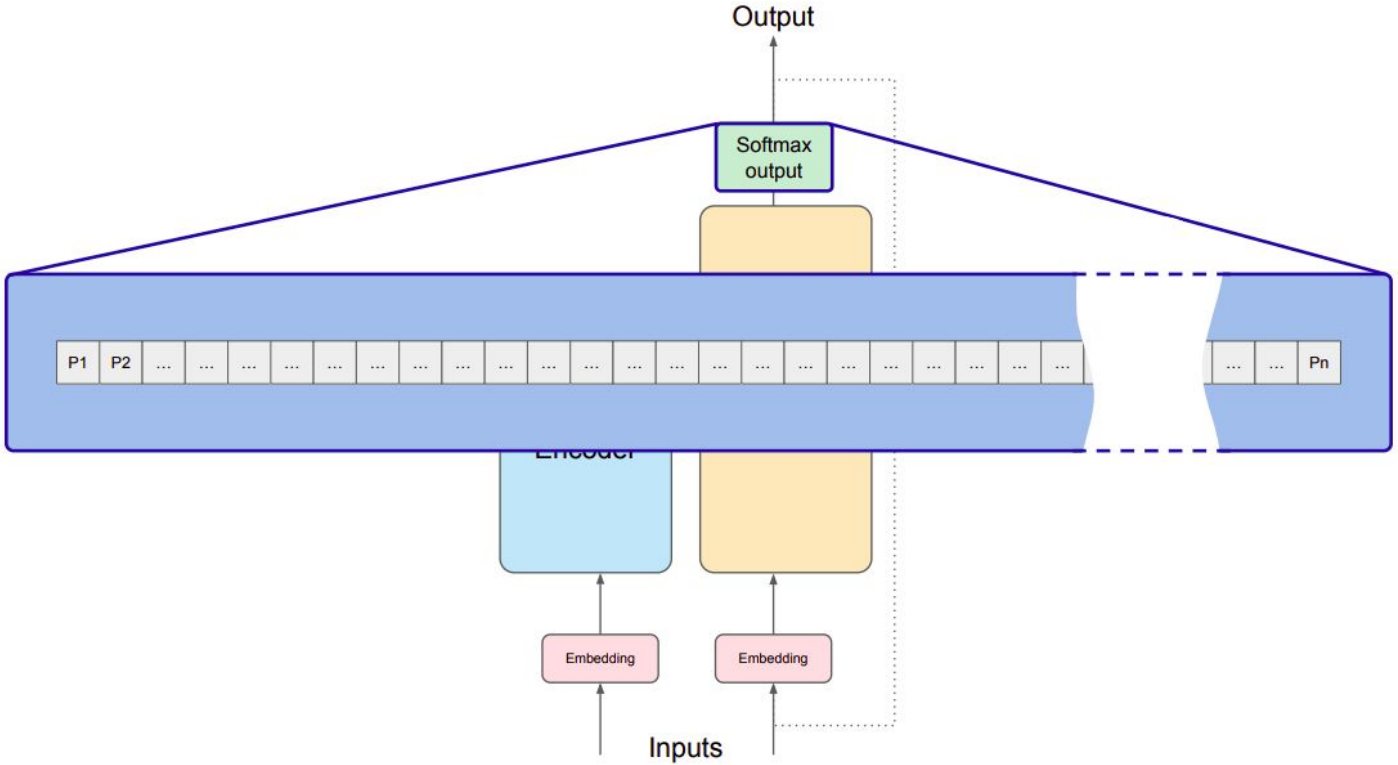
Transformers: How They Work



Transformers: How They Work

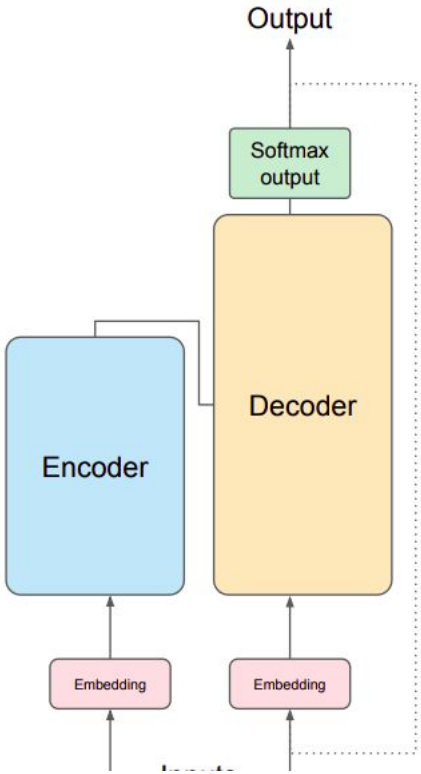


Transformers: How They Work



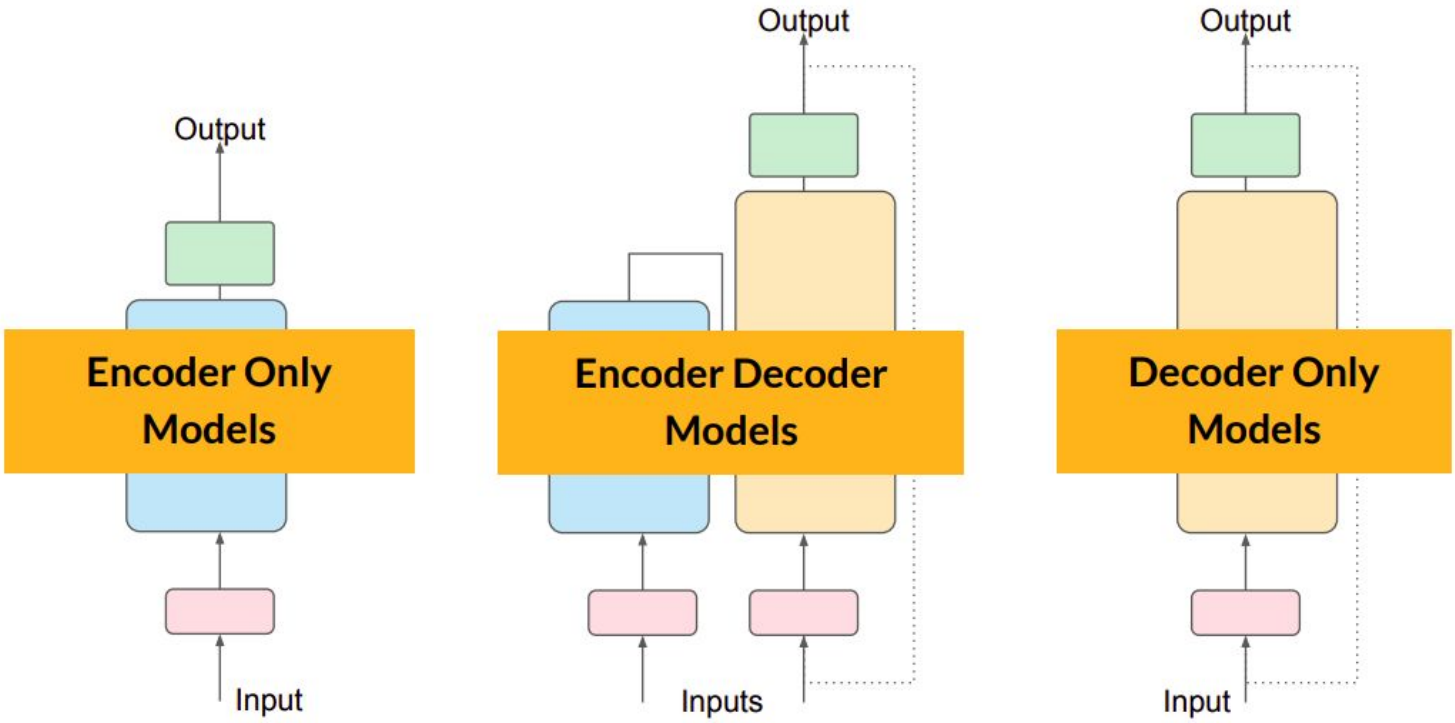
Encoder and Decoder Functionalities

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



Decoder
Accepts input tokens and generates new tokens.

Various Transformer Models



Large Language Models: How It Works

A language model is built by using supervised learning (x→y) to repeatedly predict the next word.

My favorite food is a bagel with cream cheese and lox.

Input x	Output y
My favorite food is a	bagel
My favorite food is a bagel	with
My favorite food is a bagel with	cream

Types of LLMs

Base LLM

Predicts next word, based on text training data

Once upon a time, there was a unicorn
that lived in a magical forest with all her unicorn friends

What is the capital of France?
What is France's largest city?
What is France's population?
What is the currency of France?

Instruction Tuned LLM

Tries to follow instructions

What is the capital of France?
The capital of France is Paris.

Tokens

Learning new things is fun!

Prompting is a powerful developer tool.

Tokens

Learning new things is fun!

Prompting is a powerful developer tool.

lollipop

Tokens

Learning new things is fun!

Prompting is a powerful developer tool.

lollipop

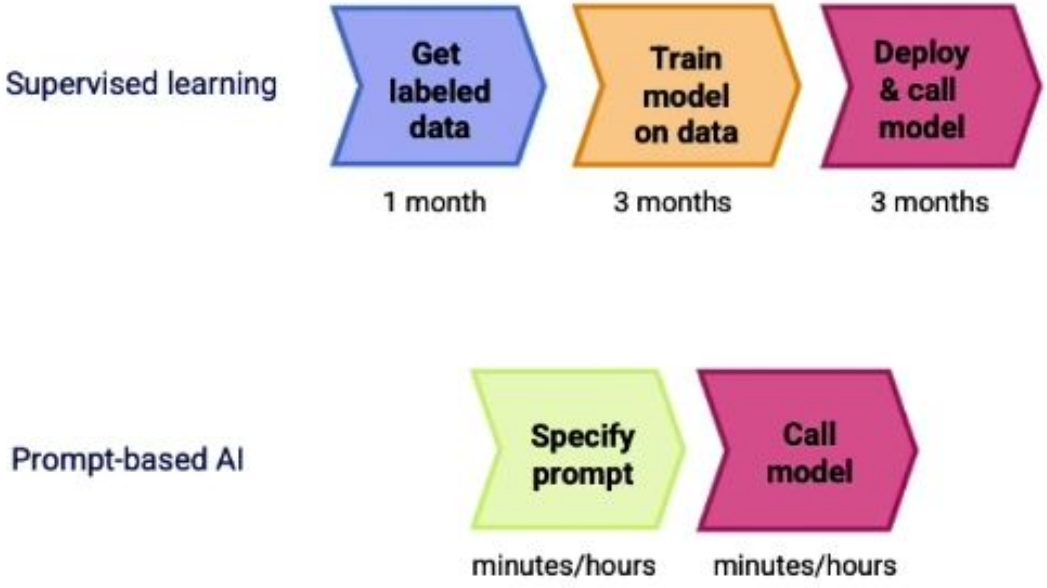
l-o-l-l-i-p-o-p

For English language input, 1 token is around 4 characters, or ¾ of a word.

Token Limits

- Different models have different limits on the number tokens in the input `context` + output completion
- gtp3.5-turbo ~4000 tokens

Application Development With ChatGPT



Thank You!!!