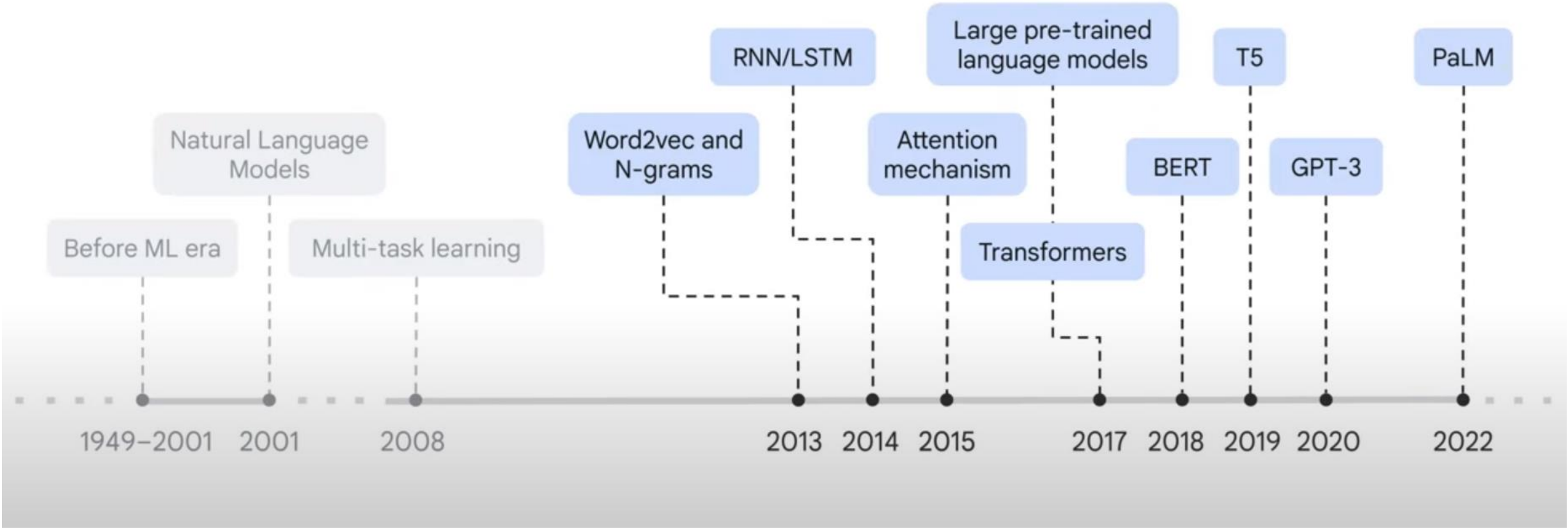
Natural Language Processing

AIGC 5501

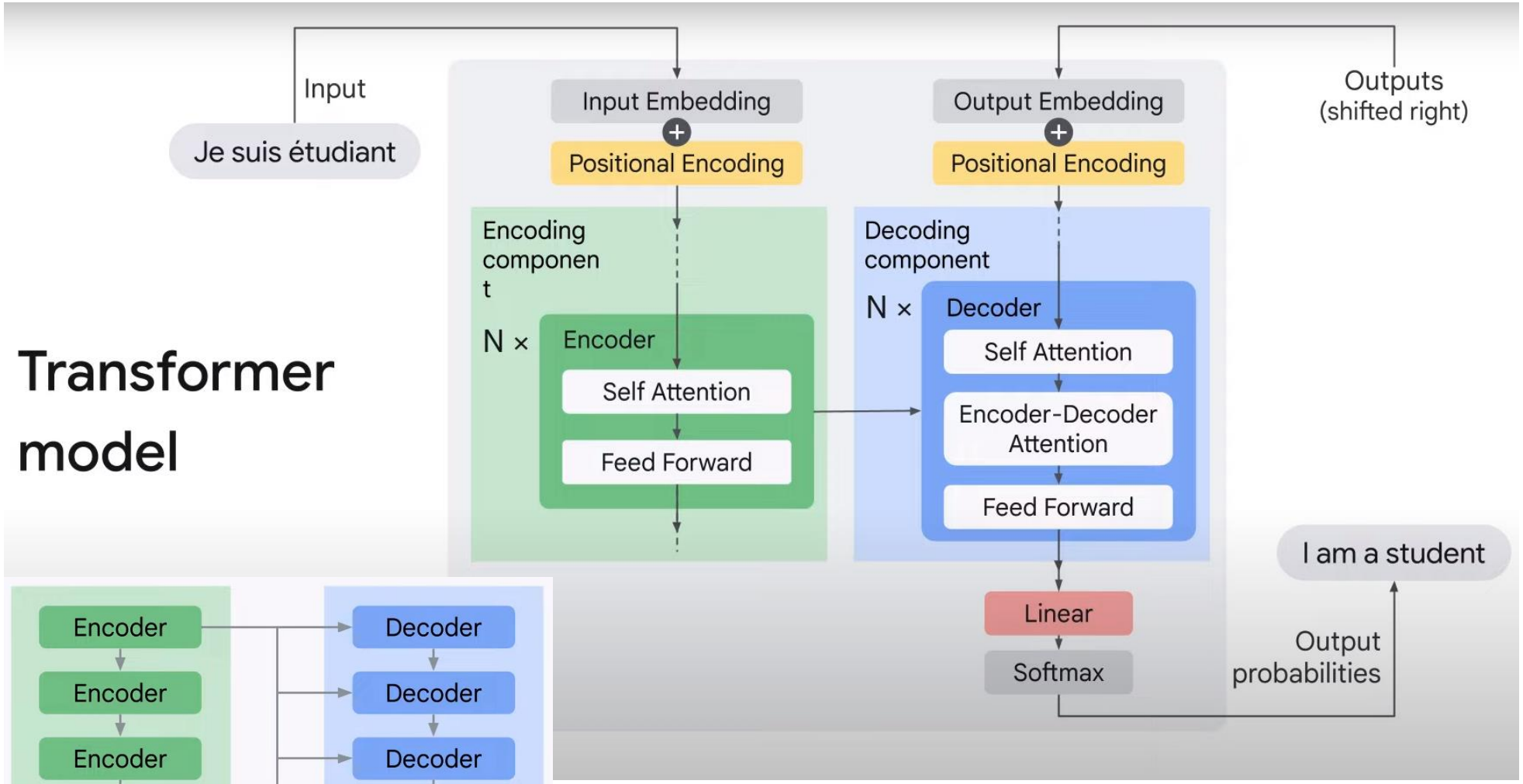
BERT & GPT

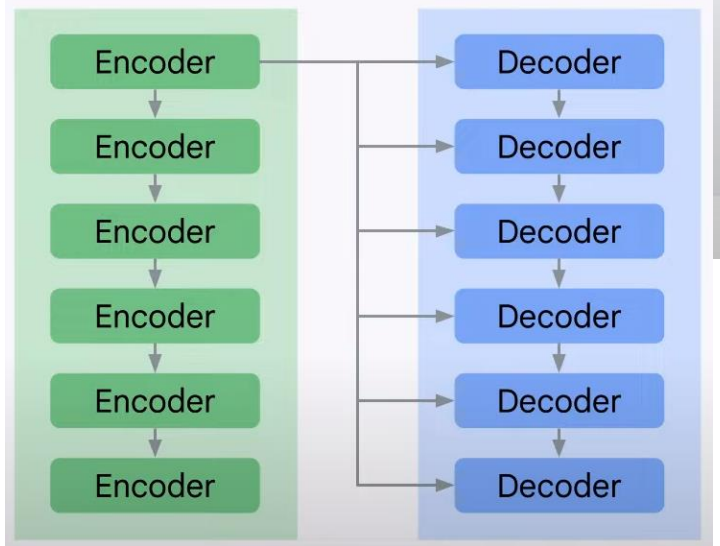
Instructor: Ritwick Dutta
Email: ritwick.dutta@humber.ca

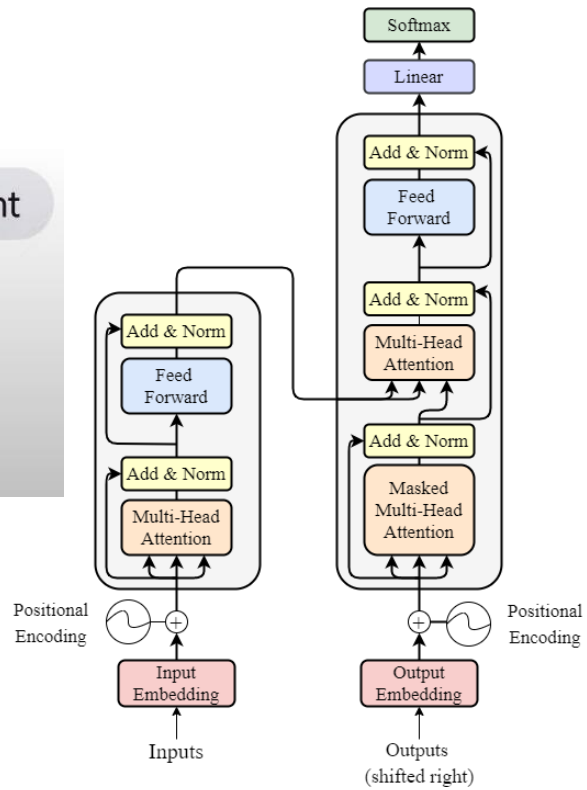
Language Modelling History



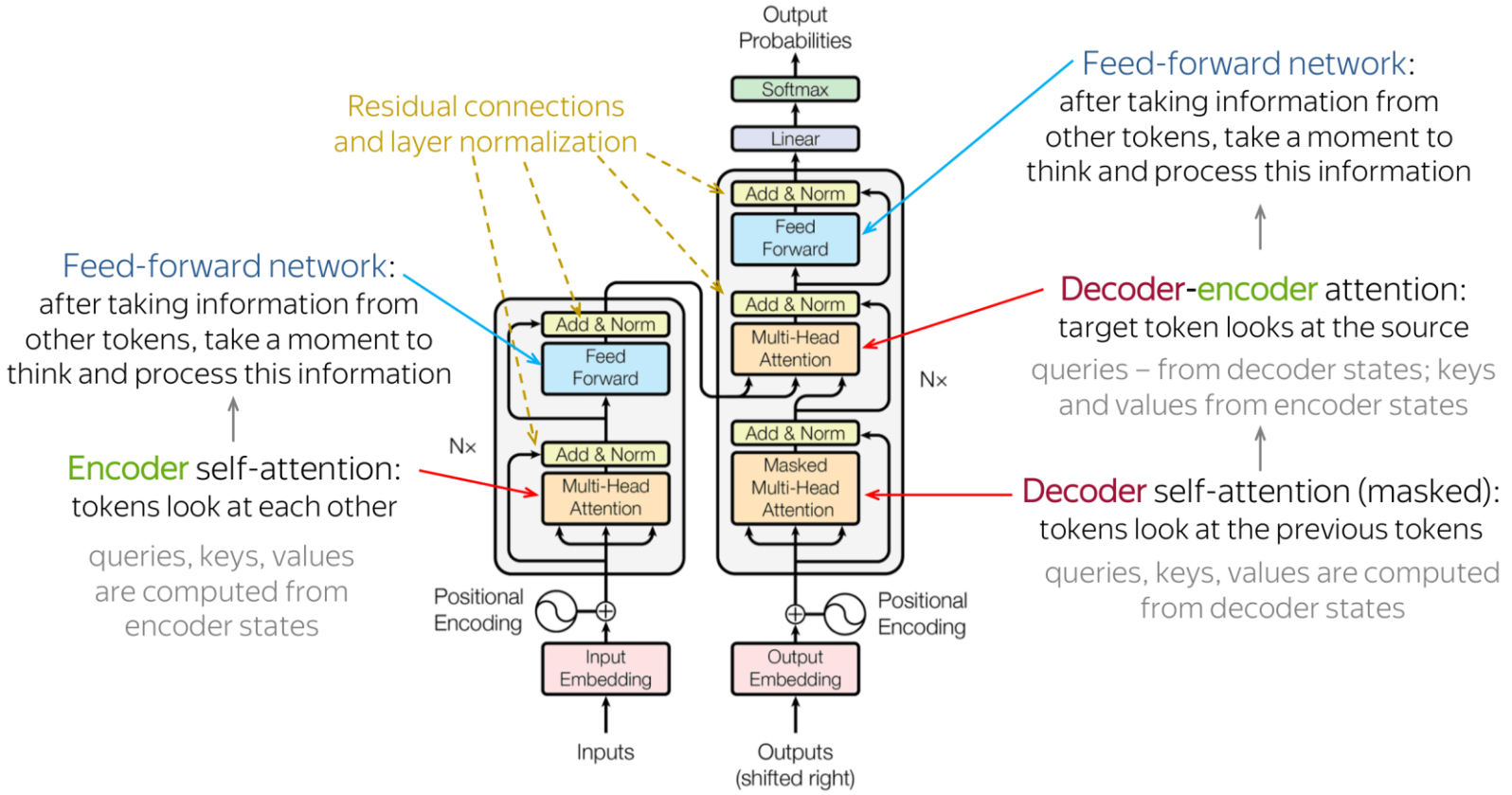
Transformer model





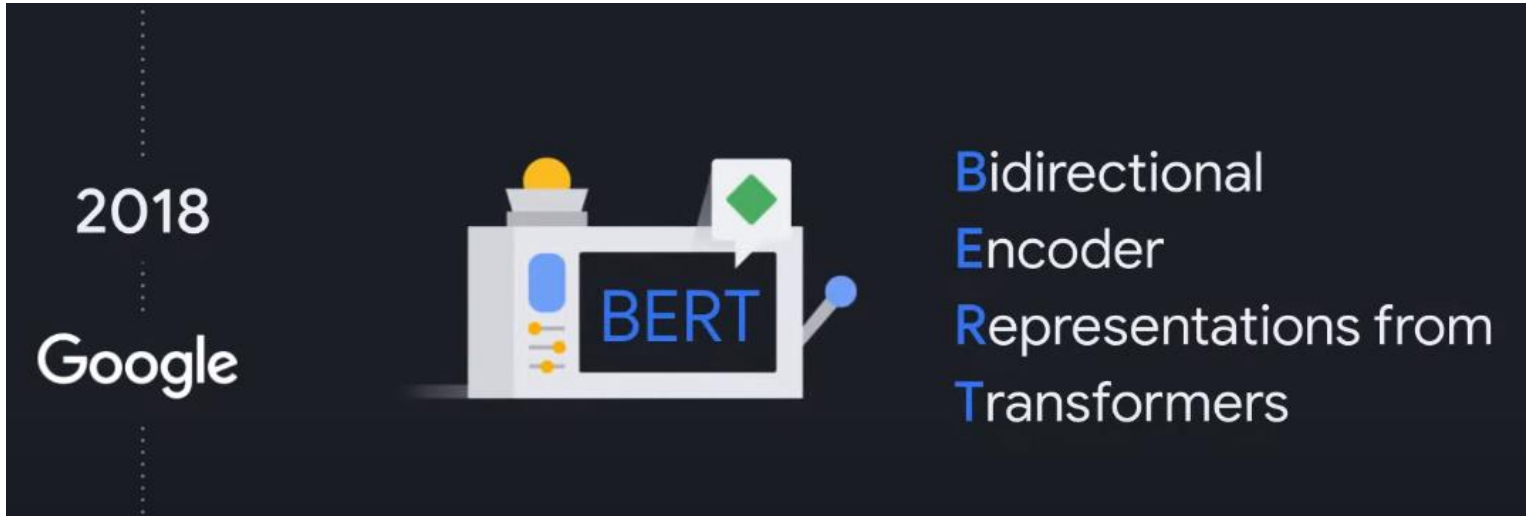


Pre-trained Transformer Models



There's multiple variations of transformer models now

BERT



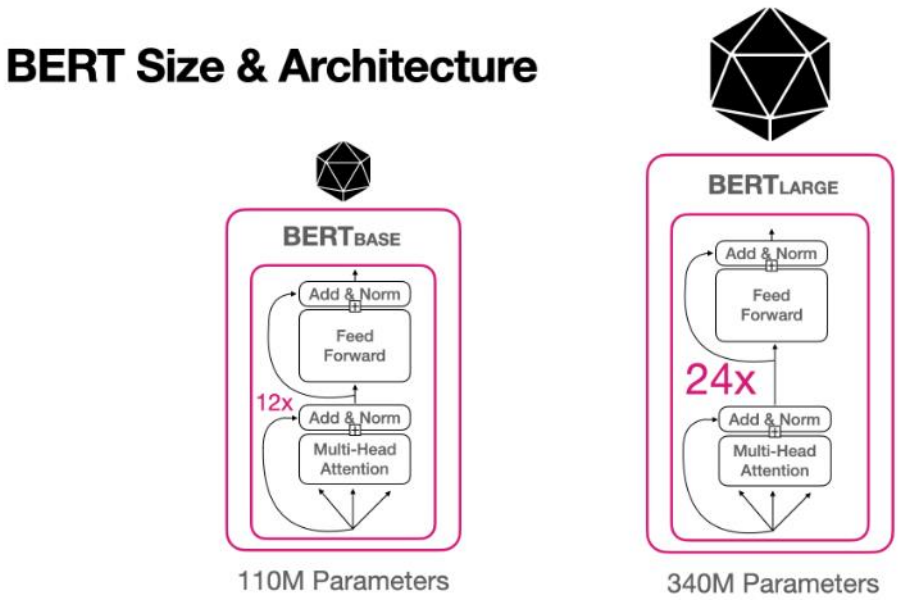
- A popular encoder only architecture – BERT
- It is one of the trained transformer models (one million steps – TPU)
- Pretrained on 3.3 billion words (Wikipedia + Toronto BooksCorpus)
- Uses a new language modeling objective - **masked language modeling**
- Can fine-tune the pretrained language model for a wide variety of NLP tasks

TPU – Tensor Processing Unit

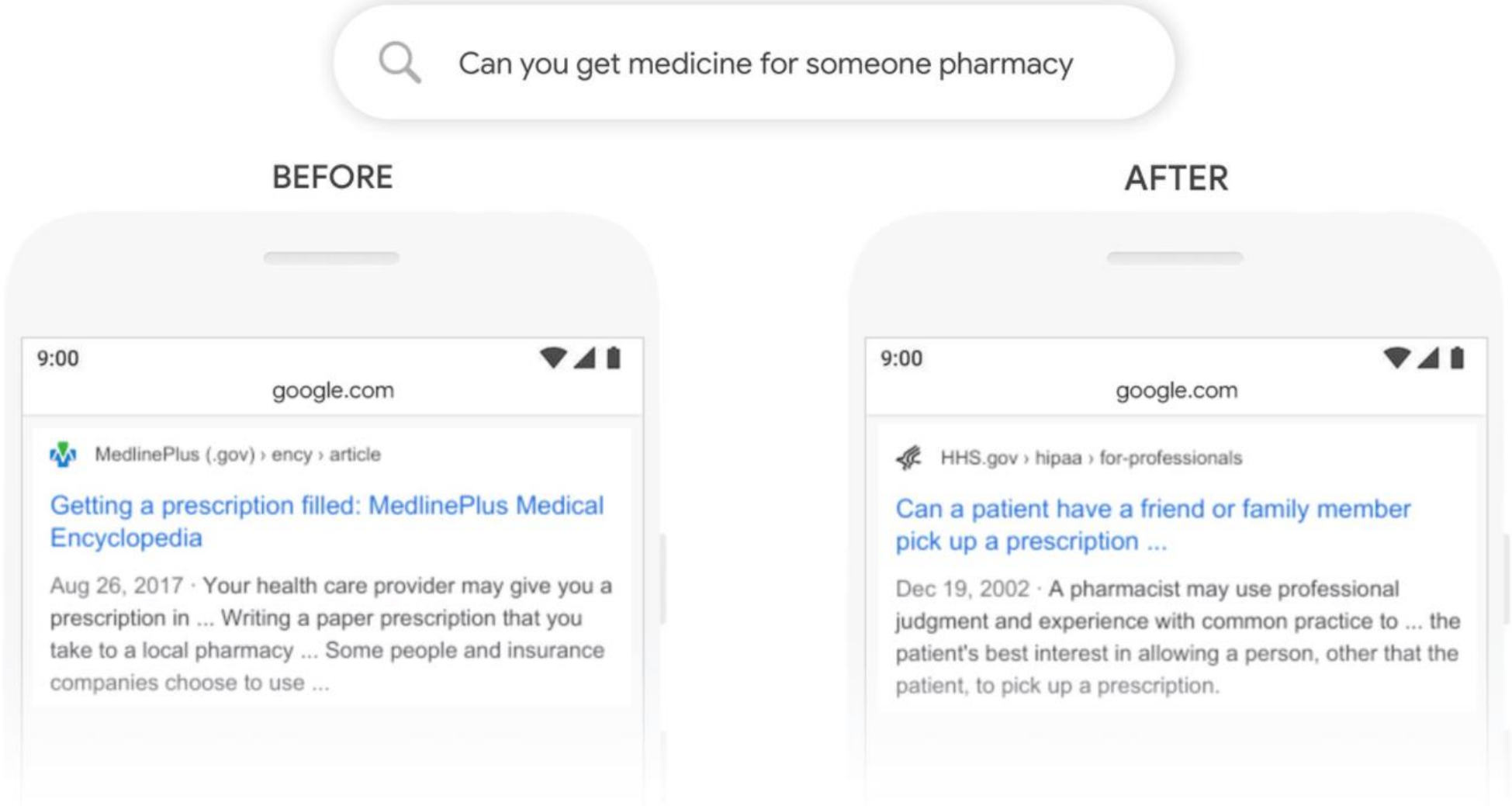
BERT Versions

	Transformer Layers	Hidden Size	Attention Heads	Parameters	Processing	Length of Training
BERTbase	12	768	12	110M	4 TPUs	4 days
BERTlarge	24	1024	16	340M	16 TPUs	4 days

	$BERT_{BASE}$	$BERT_{LARGE}$	Transformer
Layers	12	24	6
Feedforward networks (hidden units)	768	1024	512
Attention heads	12	16	8



BERT powers Google Search



BERT Use Cases

BERT can be used on a wide variety of language tasks:

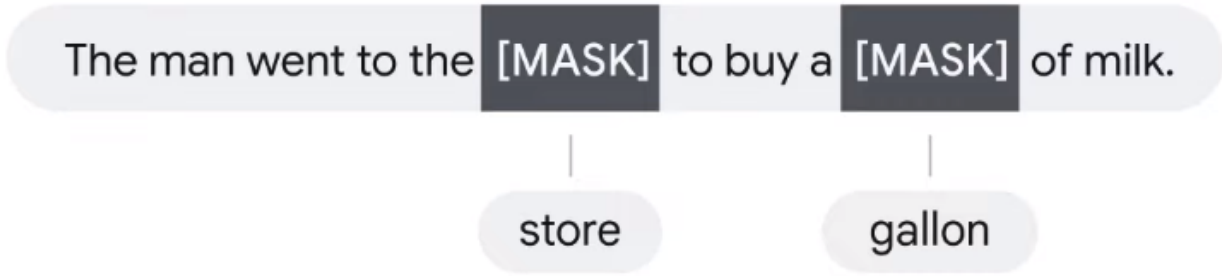
- Can determine how positive or negative a movie's reviews are. (Sentiment Analysis)
- Helps chatbots answer your questions. (Question answering)
- Predicts your text when writing an email (Gmail). (Text prediction)
- Can write an article about any topic with just a few sentence inputs. (Text generation)
- Can quickly summarize long legal contracts. (Summarization)
- Can differentiate words that have multiple meanings (like 'bank') based on the surrounding text. (Polysemy resolution)

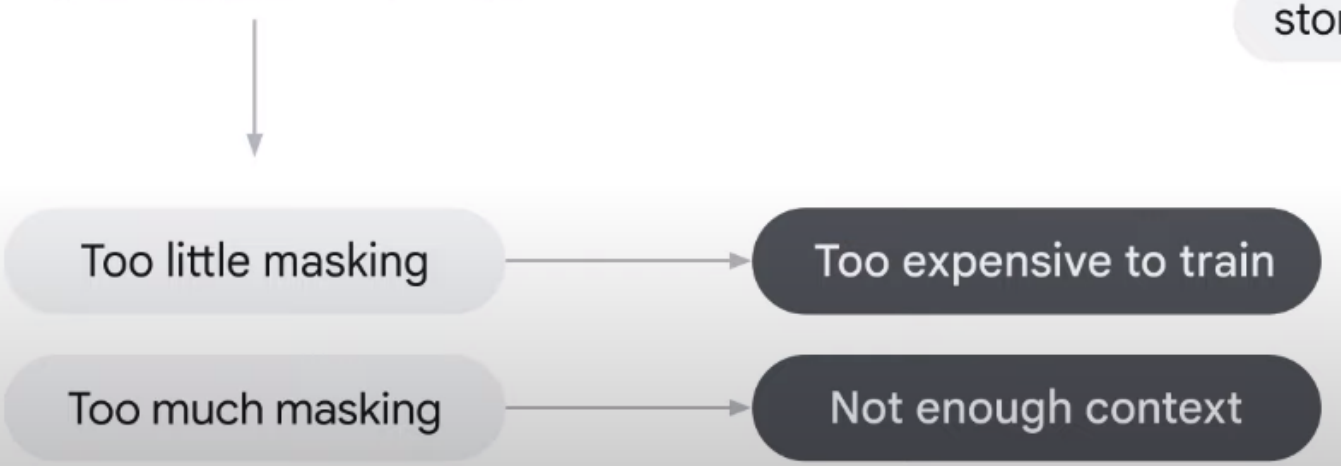
How BERT Works!

1 Masked language modeling (MLM)

Mask out k% of the input words, and then predict the masked words

- Recommendation use k = 15%





How BERT Works!

2 Next sentence prediction (NPS)

Binary classification task

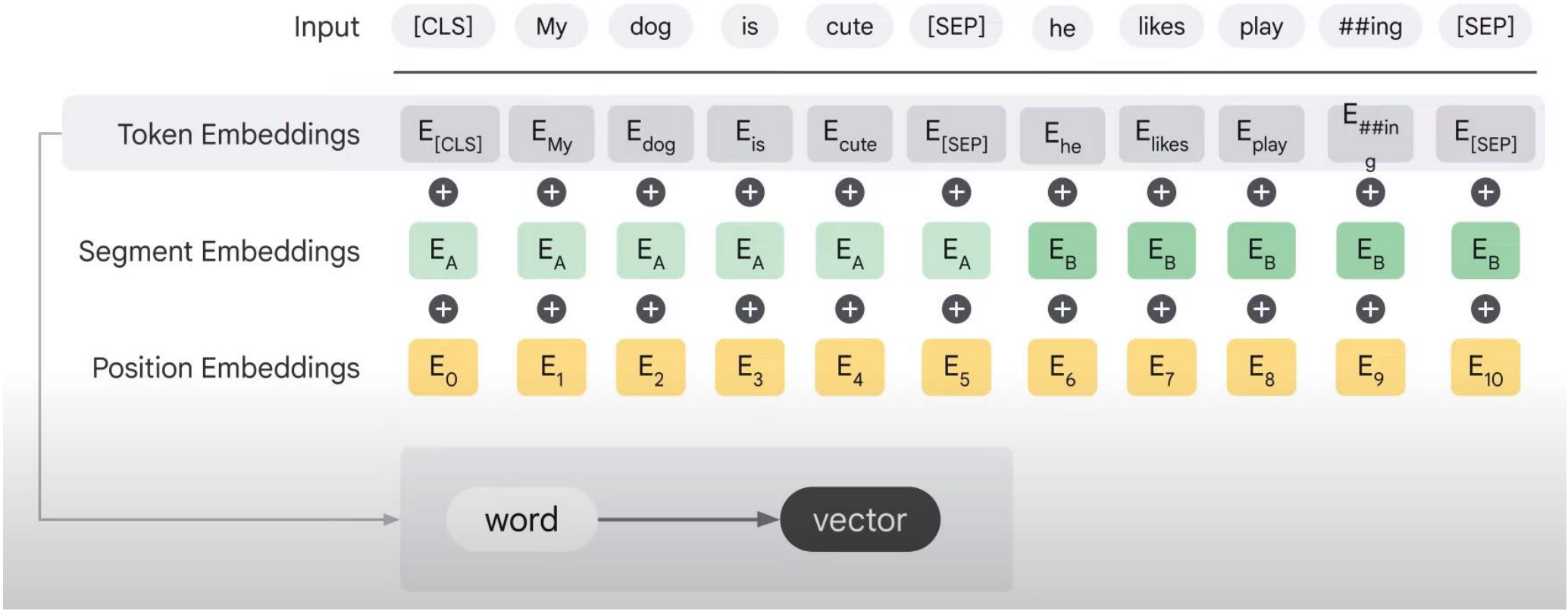
Learn the relationships between sentences and predict the next sentence given the first one.

Sentence A	The man went to the store.
Sentence B	He bought a gallon of milk.
Label	IsNextSentence

Sentence A	The man went to the store.
Sentence B	Penguins are flightless.
Label	NotNextSentence

A binary classification task
This helps Bert perform at a sentence level

BERT input embeddings



BERT is designed to process input sequences upto a length of **512**.

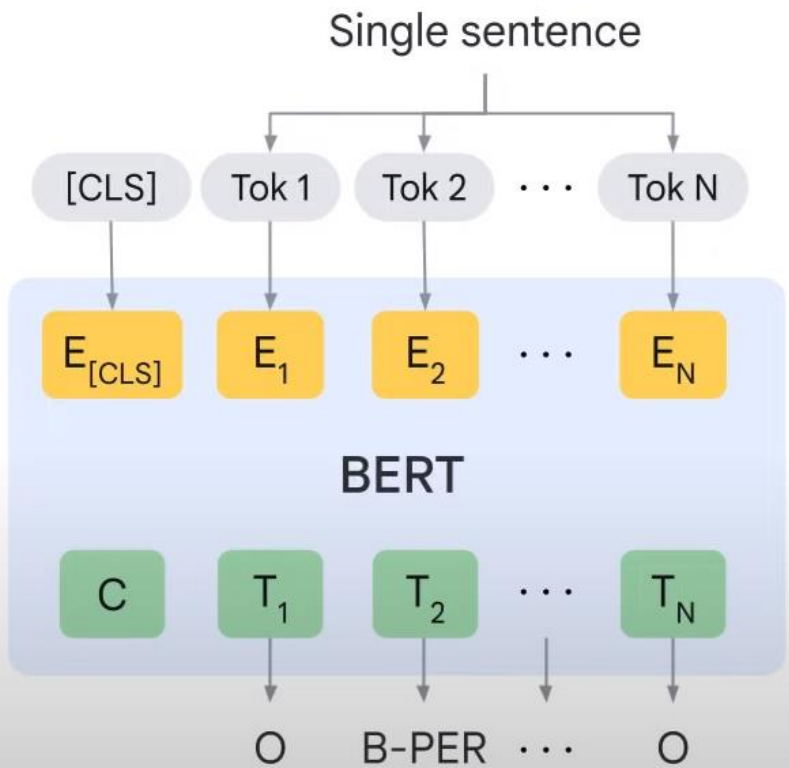
Three different kinds of embedding need to be fed into the model for the input sentence

1. **Token Embedding** – A representation of each token as an embedding in the input sentence. The words are transformed into vector representations of certain dimensions.
2. **Segment Embeddings** – Uses the special token - [SEP] to separate two different splits of the sentence.
3. **Position Embedding** – learn the order input sequence of the words in the sentence

You can use BERT for various downstream tasks, or example:

- Single sentence classification
- Sentence pair classification
- Question answering
- Single sentence tagging tasks

Single sentence tagging tasks: CoNLL-2003 NER



BERT is Open Source

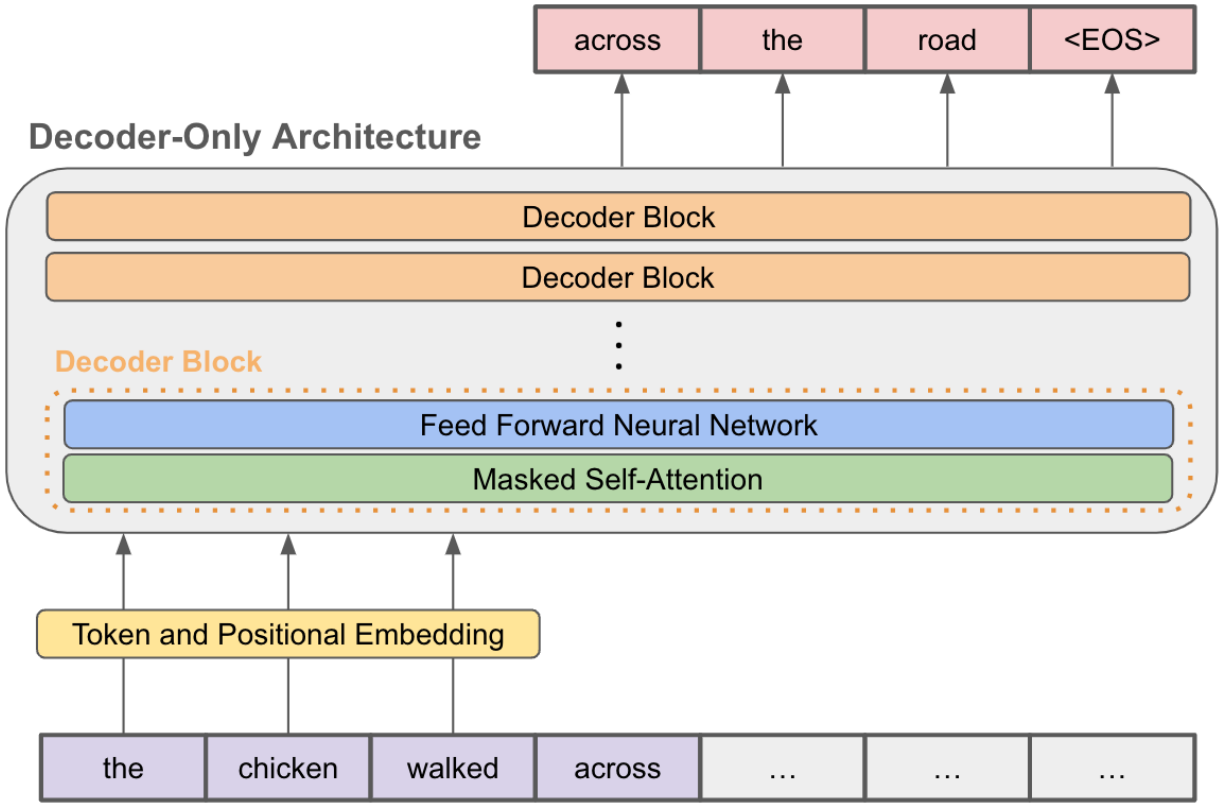
Unlike other large learning models like GPT-3, BERT's source code is publicly accessible (view BERT's code on Github) allowing BERT to be more widely used all around the world. This is a game-changer!

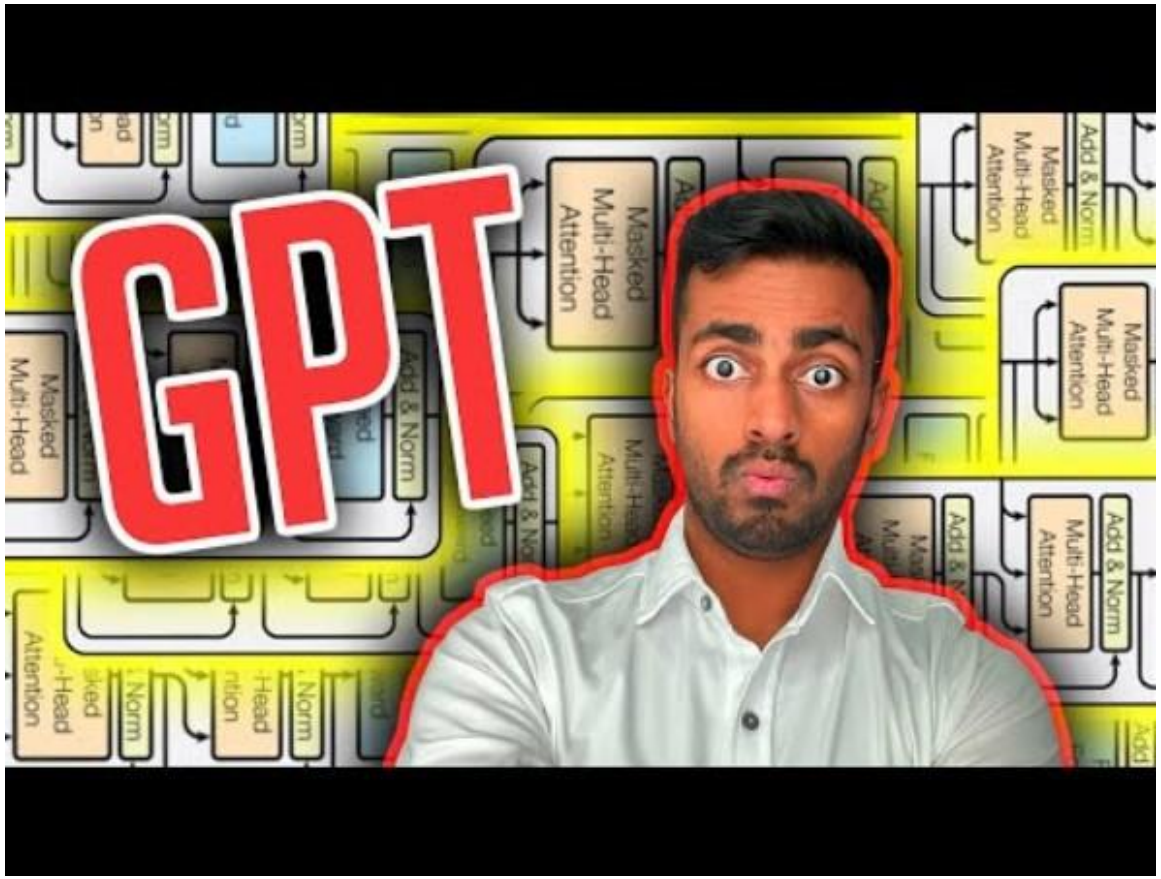
BERT models pre-trained for specific tasks:

- Twitter sentiment analysis
- Analysis of Japanese text
- Emotion categorizer (English - anger, fear, joy, etc.)
- Clinical Notes analysis
- Speech to text translation
- Toxic comment detection

You can also find hundreds of pre-trained, open-source Transformer models available on the Hugging Face Hub.

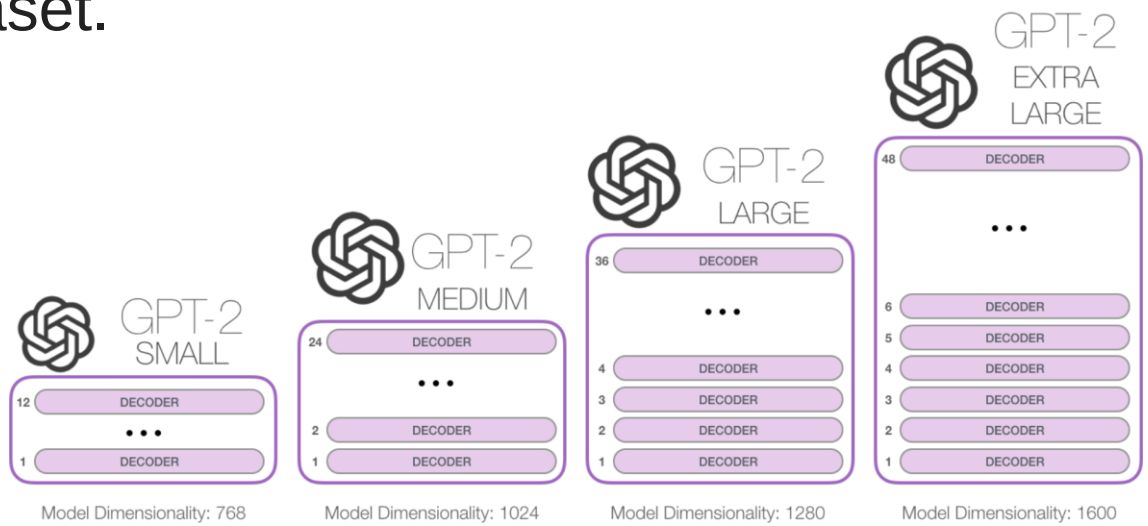
GPT - Explained





GPT - 2

- The GPT-2 is very similar to the decoder-only transformer
- Its however, a very large, transformer-based language model trained on a massive dataset.

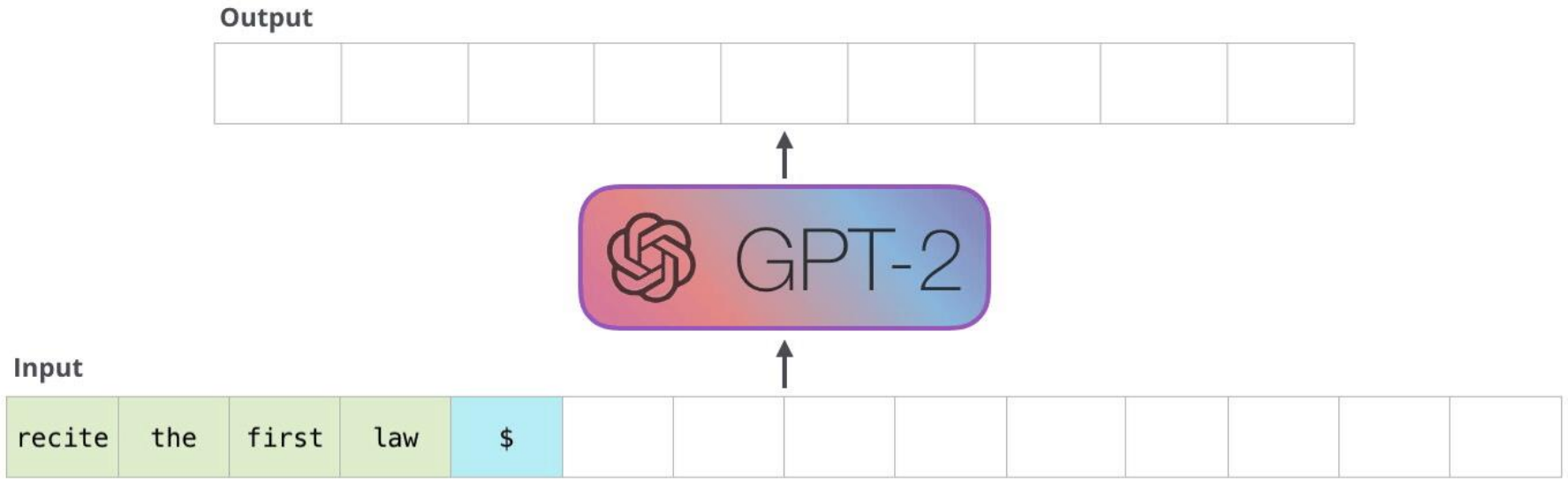


GPT Model

The way GPT models actually work is that after each token is produced, that token is added to the sequence of inputs.

And that new sequence becomes the input to the model in its next step.

This is an idea called “auto-regression”. This is one of the ideas that made RNNs unreasonably effective.

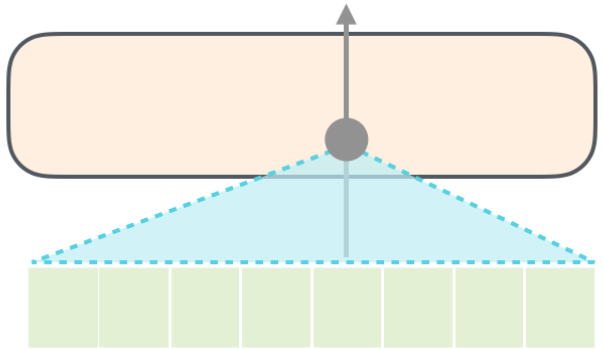


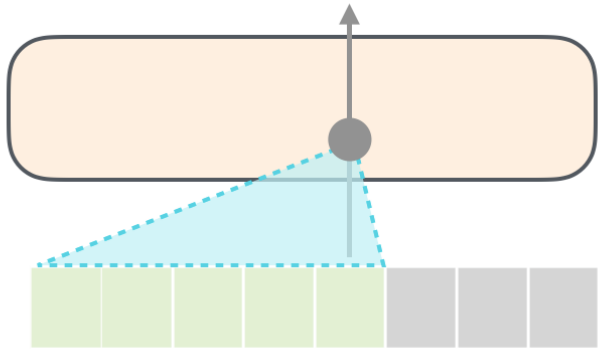
GPT Model

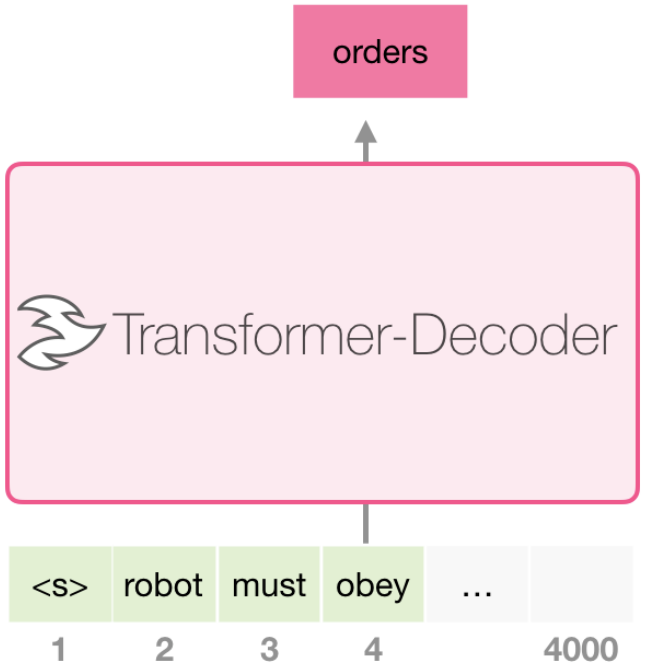
It's important that the distinction between self-attention (what BERT uses) and masked self-attention (what GPT-2 uses) is clear.

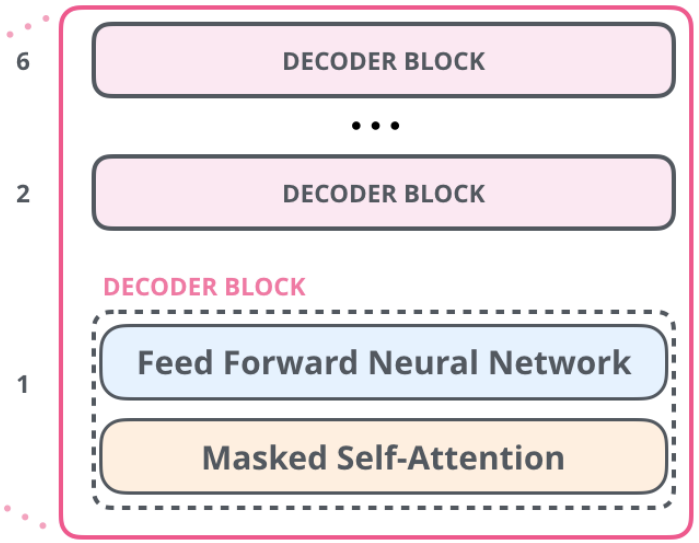
A normal self-attention block allows a position to peak at tokens to its right.

Masked self-attention prevents that from happening.

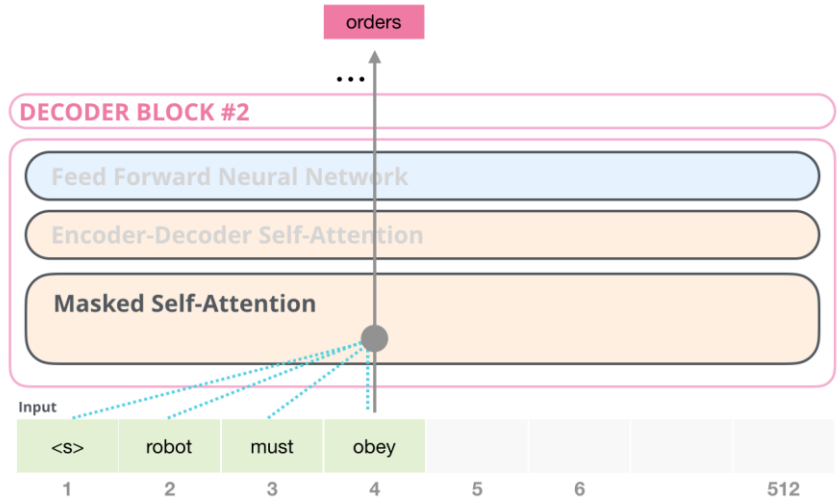








The blocks were very similar to the original decoder blocks, except they did away with that second self-attention layer.



GPT 3

A trained language model generates text.

We can optionally pass it some text as input, which influences its output.

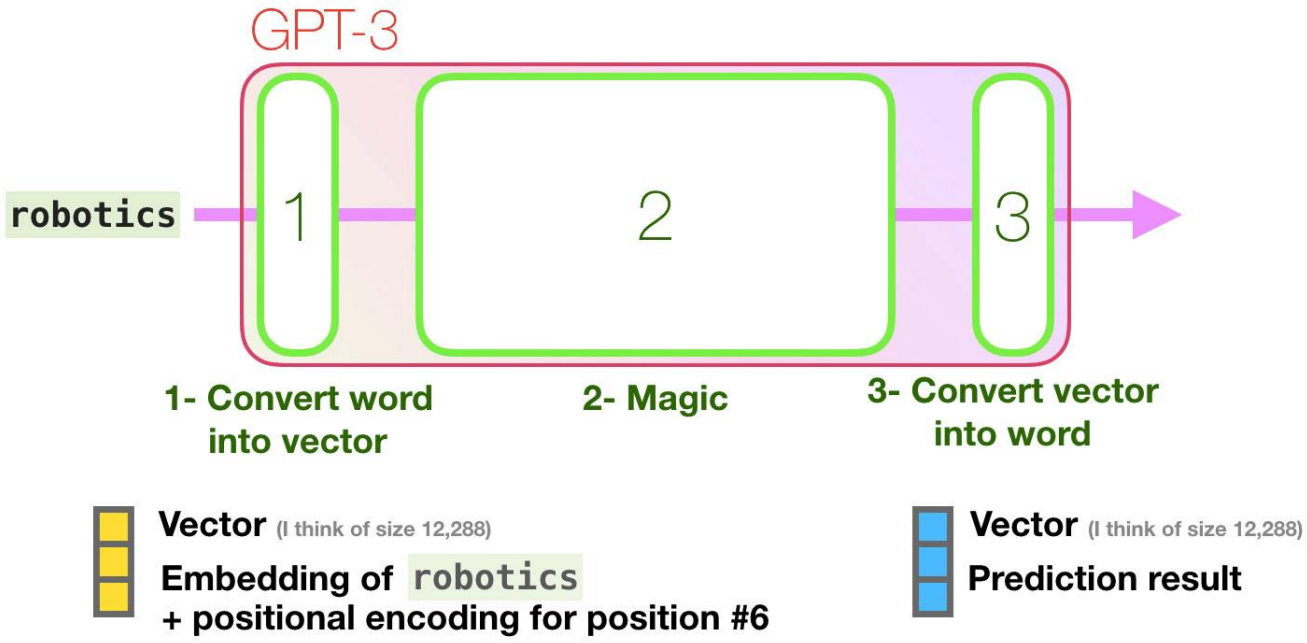
The output is generated from what the model “learned” during its training period where it scanned vast amounts of text.

It was estimated to cost 355 GPU years and cost $4.6m.

GPT3 is MASSIVE.
It encodes what it learns from training in 175 billion numbers (called parameters). These numbers are used to calculate which token to generate at each run.

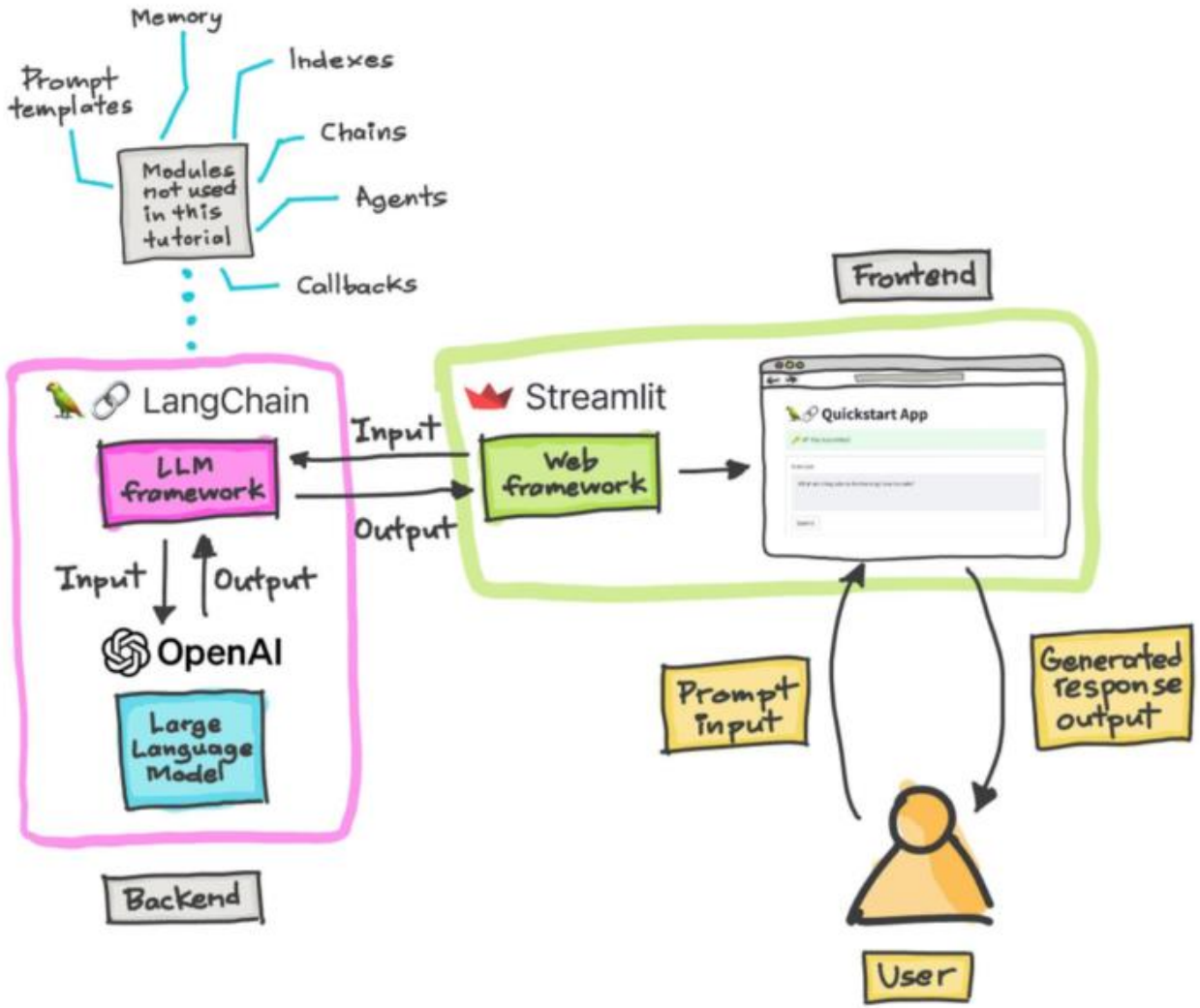
GPT3 is 2048 tokens wide. That is its “context window”. That means it has 2048 tracks along which tokens are processed.

https://jalammar.github.io/how-gpt3-works-visualizations-animations/



A step-by-step guide using any LLM, LangChain, and Streamlit

This is how it works under the hood:



https://blog.streamlit.io/langchain-tutorial-1-build-an-llm-powered-app-in-18-lines-of-code/

With Falcon:
https://www.youtube.com/watch?v=2OvPoSQanLs

LLM Use Cases

Streamlit Chatbot
https://www.youtube.com/watch?v=sBhK-2K9bUc

Recipe Generator App:
https://www.youtube.com/watch?v=RvRGtuGCAMY

Final Projects

a) Chatbot for Humber - students association assistant (Eng + French support)

b) Creative assistant for Humber marketing team (Eng + French support)

c) Document summarizer - PDF / word / reports (Eng + French support)

Demo Presentation Next Week – 10 mins per team