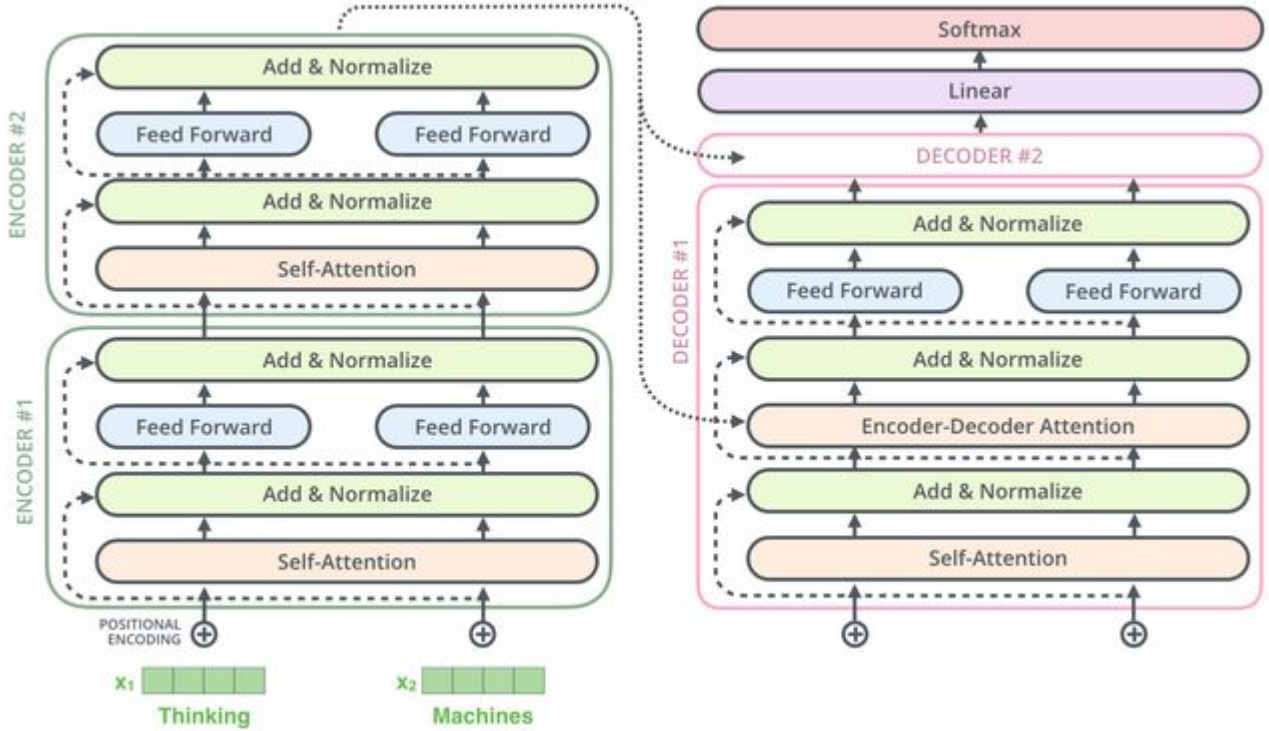
GPT and BERT

- **Zishan Ahmad & Deeksha Varshney**
IIT Patna

Outline

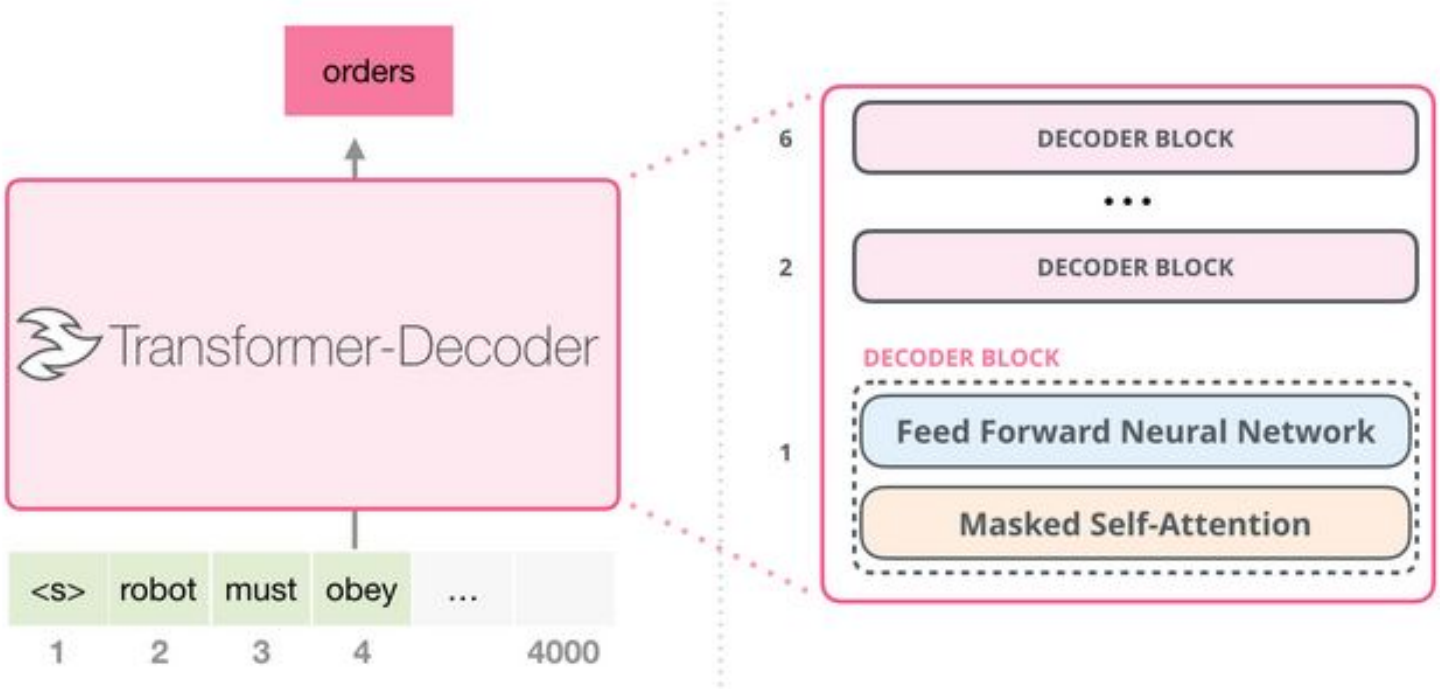
- Model Architecture
- GPT-2
  - Causal Language Modelling
  - Sequence Generation
- BERT
  - Masked Language Modelling
  - Next Sentence Prediction
  - Sequence Generation

Model Architecture (Transformer)



Generative Pre-trained Transformer (GPT)

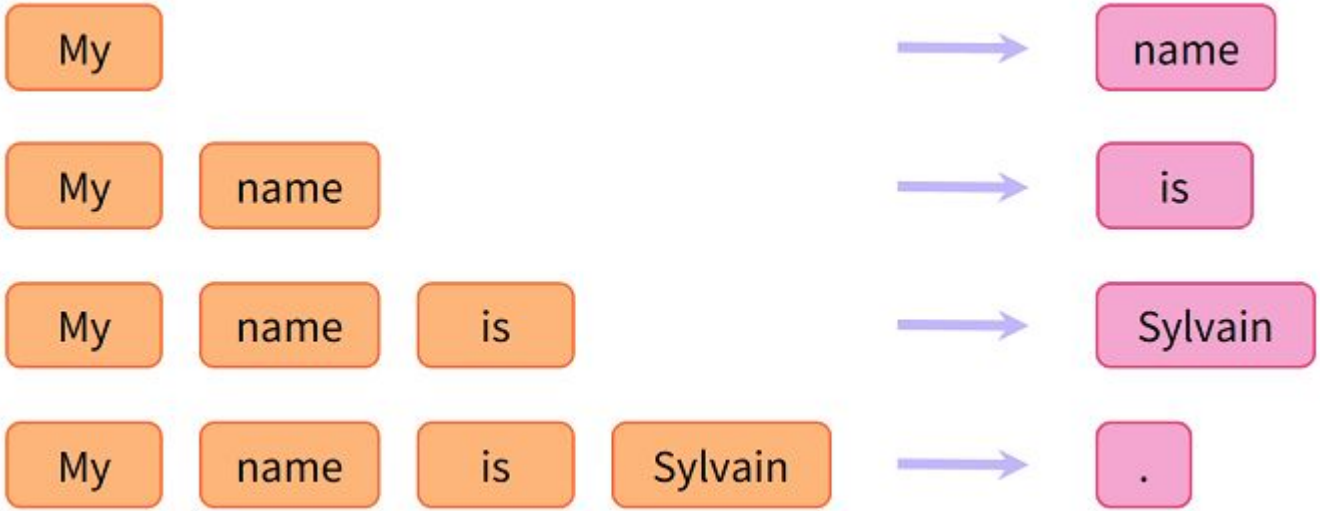
GPT: Model Architecture



GPT: Causal Language Modelling

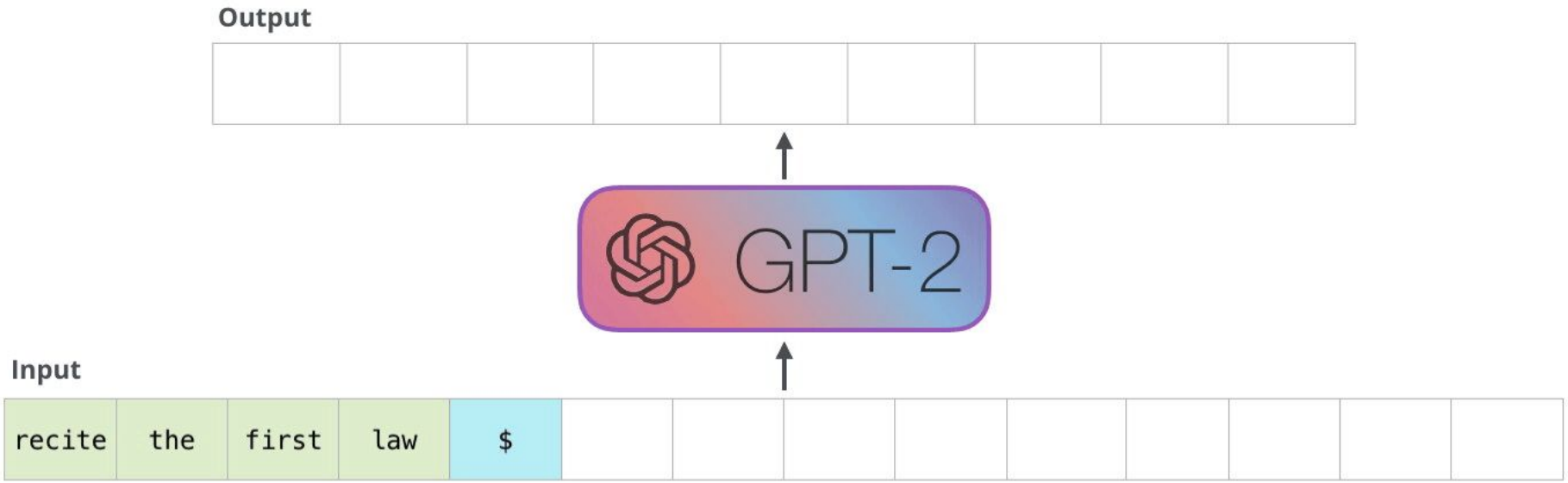
GPT is a transformer model pre-trained on causal language modelling task

- *Causal Language Modelling (CLM)* is the task of predicting the next token/word given the previous sequence of token/word



GPT: Sequence Generation

Given an input how does GPT generate the output?



Codes

GPT:
https://colab.research.google.com/drive/1faCRg3KeJRmI-hkrRDHSDv5bOh8ksb9N?usp=sharing

WE:
https://colab.research.google.com/drive/1fBp7mQZzgQvjinmu4QmpIASX9B5WtgRA?usp=sharing

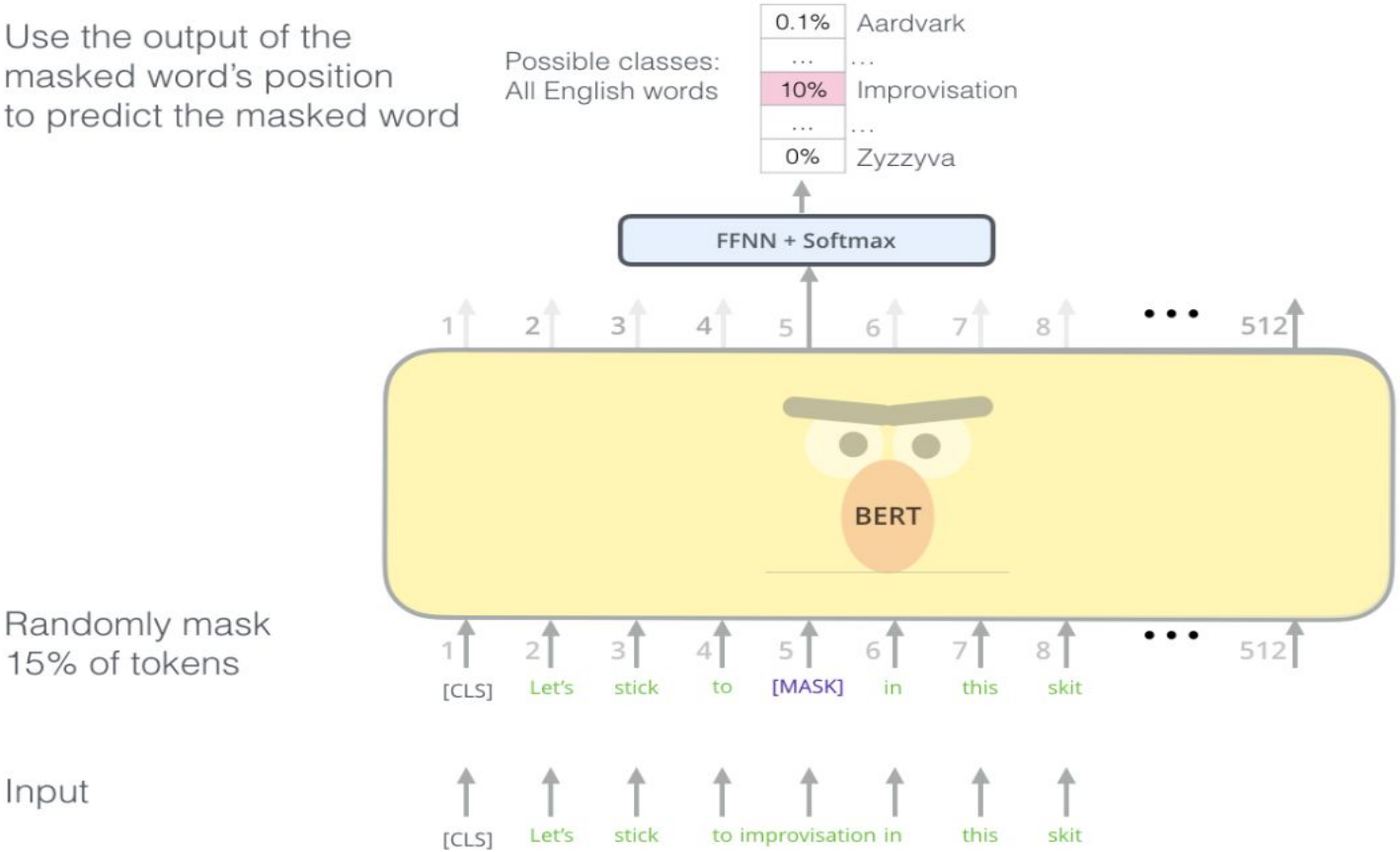
https://colab.research.google.com/drive/1EfnpvIGZp5SOHZCw4D3uDj-Z7hJ97KfJ?usp=sharing

https://colab.research.google.com/drive/10XCsBjW88b9pYiLgWADxVaDLhZHhSeVV?usp=sharing

https://colab.research.google.com/drive/1mD5kkHLP5BhYH5HYBxxXuBiijBtA3CcM?usp=sharing

BERT (Bidirectional Encoder Representations from Transformers)

Model Architecture



Masked Language Modeling

- The first input token is supplied with a special [CLS] token. CLS here stands for Classification.
  - `[CLS] The man went to the store. [SEP] He bought a gallon of milk.`
- They hid 15% of the words and used their position information to infer them for MLM.
  - `[CLS] The man MASK to the MASK. [SEP] He bought a MASK of milk.`

Masked Language Modeling

- Just like the vanilla encoder of the transformer, BERT takes a sequence of words as input which keep flowing up the stack.
- Each layer applies self-attention, and passes its results through a feed-forward network, and then hands it off to the next encoder.