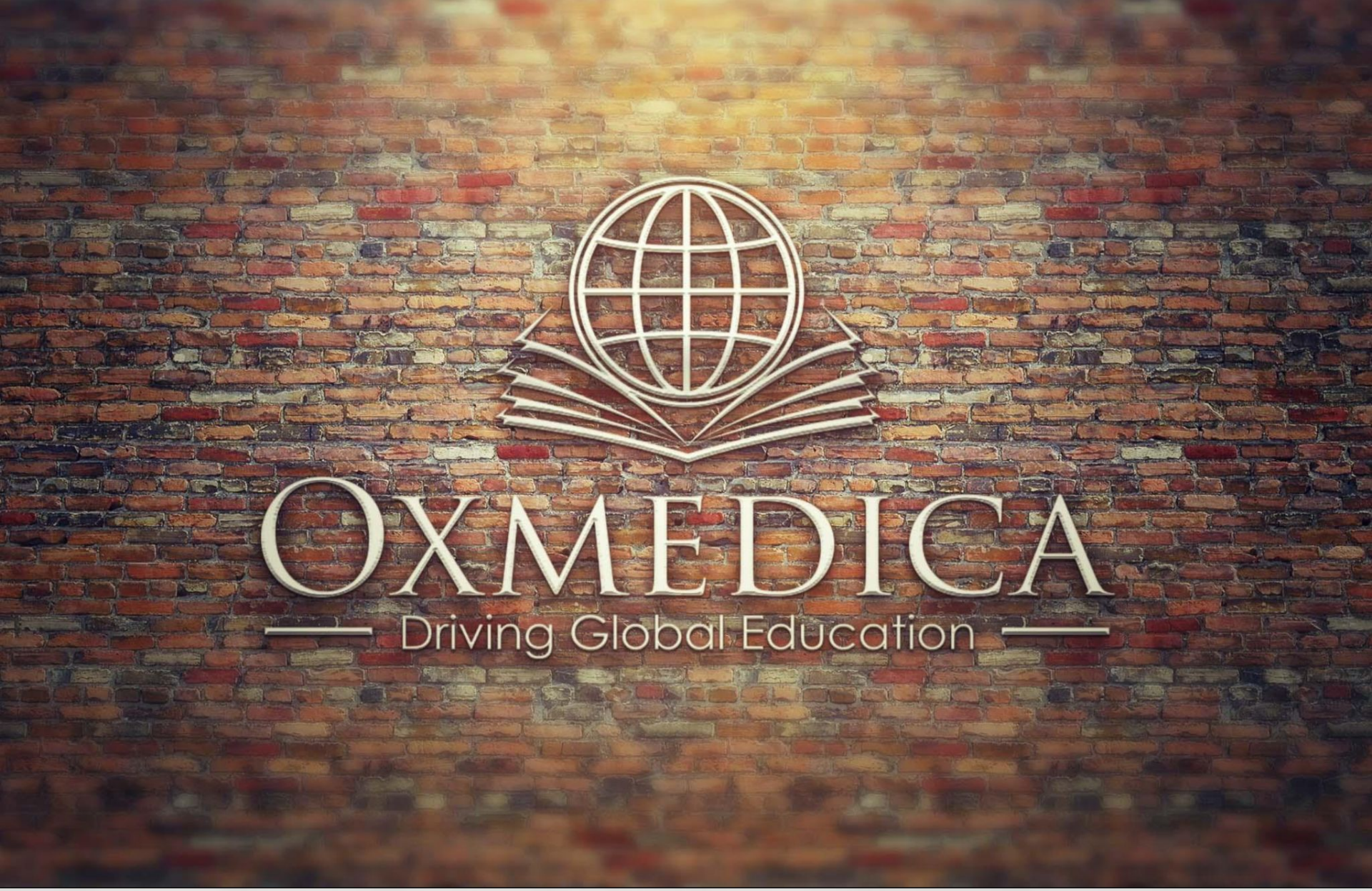


Large Language Models

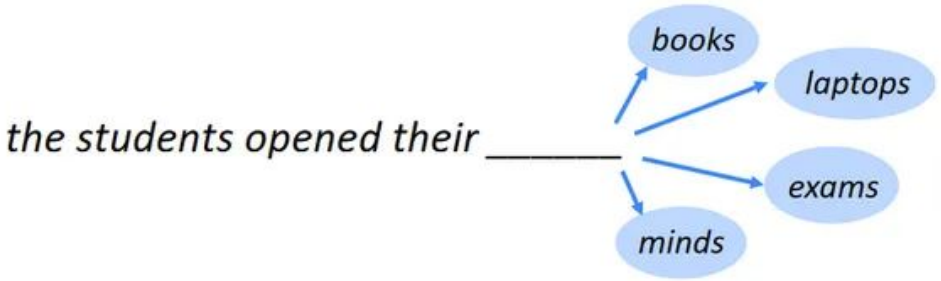
Overview

1 A technical overview of LLMs

2 Guest lecture

A technical overview of LLMs

1 Next token (subword) prediction



GOAL: Train models to predict the next token (word)

source

1 Next token (subword) prediction

TASK: AI and Big Data 1 vs GPT-4o

- 10 sentences with the final word missing.
- You have to write down what you think the final word is
- Compare your prediction with ChatGPT

source

1 Next token (subword) prediction

1. The sun set behind the mountains, casting a warm glow over the [redacted]
2. She carefully opened the old, dusty book, revealing pages filled with ancient [redacted]
3. The cat jumped onto the windowsill, watching the birds outside with keen [redacted]
4. After weeks of preparation, the team finally launched their new product to great [redacted]
5. The sound of laughter filled the air as children played in the [redacted]
6. He took a deep breath and stepped onto the stage, ready to deliver his [redacted]
7. The smell of freshly baked bread wafted through the kitchen, making everyone's mouth [redacted]
8. They decided to take a spontaneous road trip, exploring the countryside and discovering hidden [redacted]
9. The storm raged outside, but inside, the family gathered around the fireplace, enjoying the [redacted]
10. She received an unexpected letter in the mail, bringing news that would change her life [redacted]

source

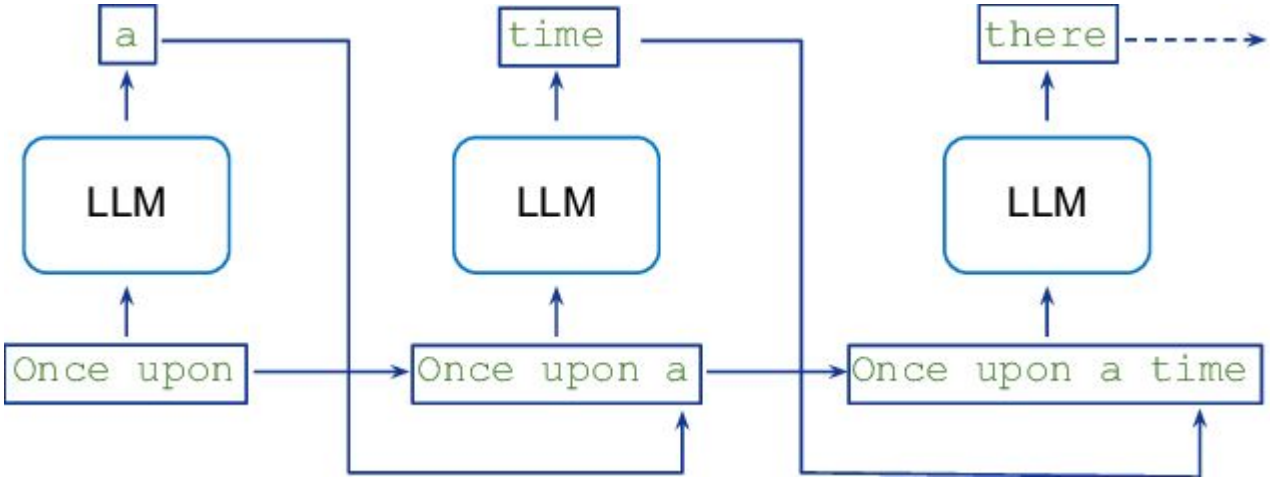
1 Next token (subword) prediction

8/10

1. The sun set behind the mountains, casting a warm glow over the **valley**
2. She carefully opened the old, dusty book, revealing pages filled with ancient **texts**
3. The cat jumped onto the windowsill, watching the birds outside with keen **interest**
4. After weeks of preparation, the team finally launched their new product to great **success**
5. The sound of laughter filled the air as children played in the **park**
6. He took a deep breath and stepped onto the stage, ready to deliver his **speech**
7. The smell of freshly baked bread wafted through the kitchen, making everyone's mouth **water**
8. They decided to take a spontaneous road trip, exploring the countryside and discovering hidden **gems**
9. The storm raged outside, but inside, the family gathered around the fireplace, enjoying the **warmth**
10. She received an unexpected letter in the mail, bringing news that would change her life **forever**

source

2 Autoregressive Language Modelling



Many forward passes!
(expensive!)

source

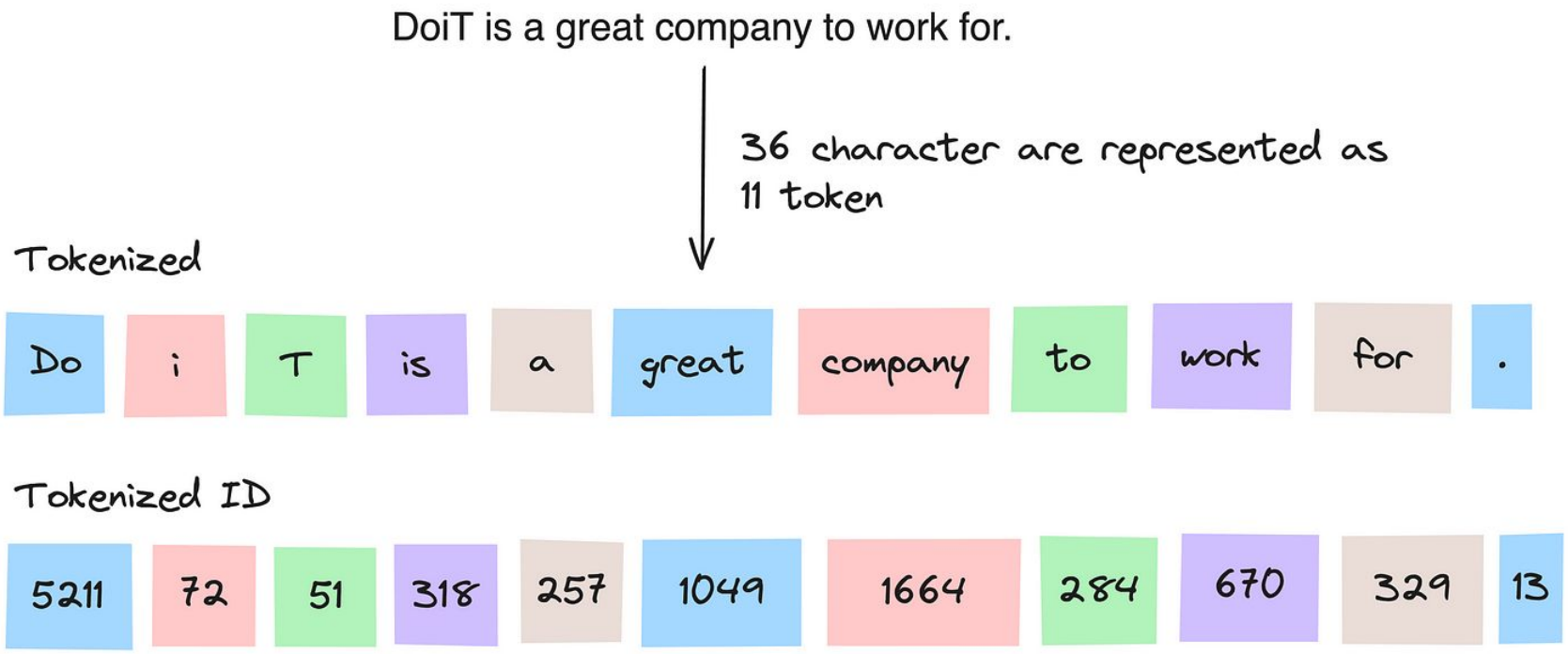
3 Tokenization

- A token is a subword
- There are a fixed list of subwords
- Tokenization is the process of breaking text into subwords

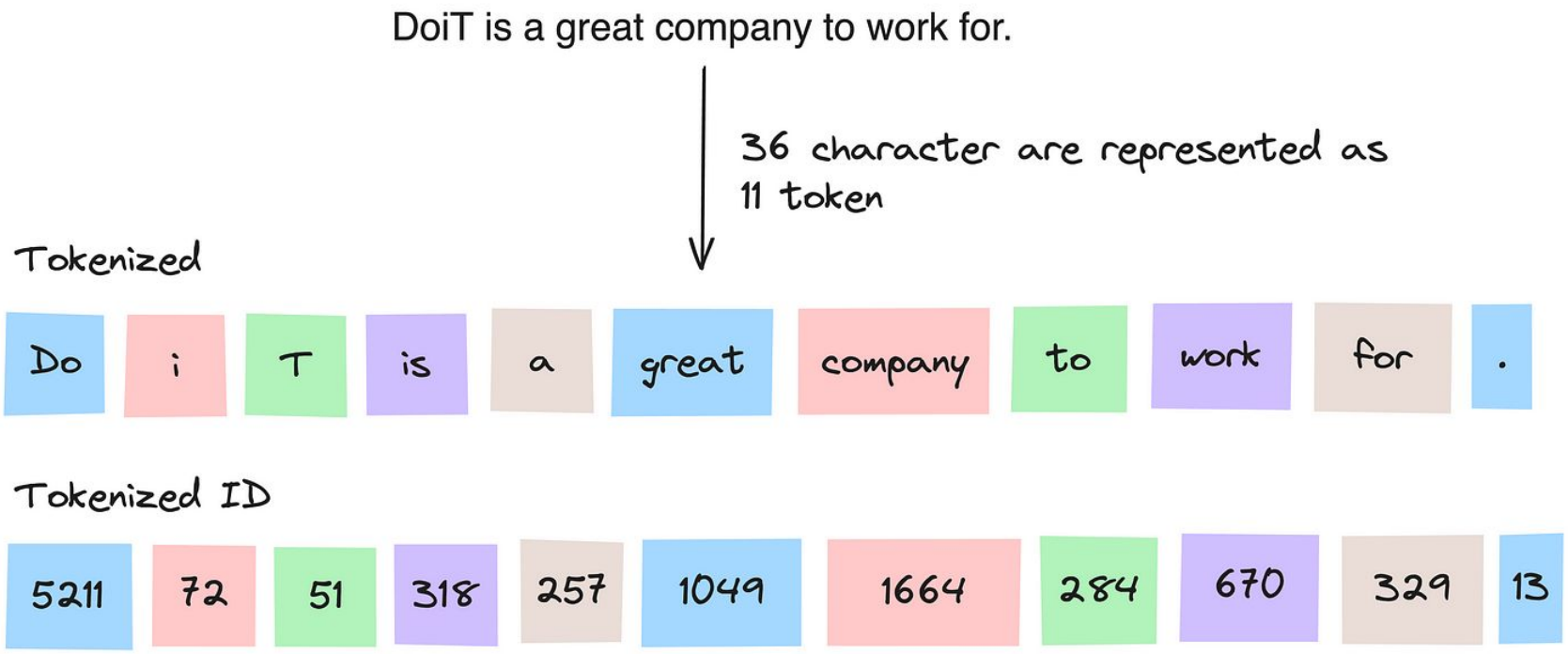
```
Two annoying things about OpenAI's tokenizer playground: (1) it's
 capped at 50k characters, and (2) it doesn't support GPT-4 or GPT-3.5
...

So, I built my own version w/ Transformers.js! It can tokenize the
 entire "Great Gatsby" (269k chars) in 200ms! ���
```

3 Tokenization



3 Tokenization



3 Tokenizer playground

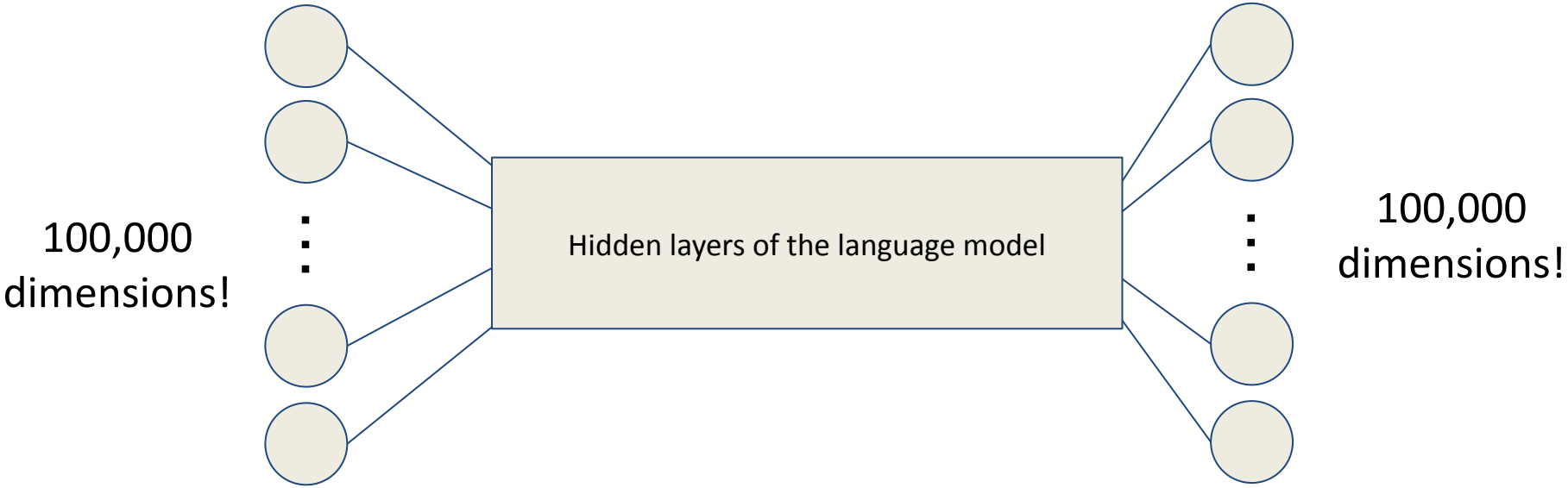
https://tiktokenizer.vercel.app/?model=gpt-3.5-turbo

Task in pairs: 5 mins

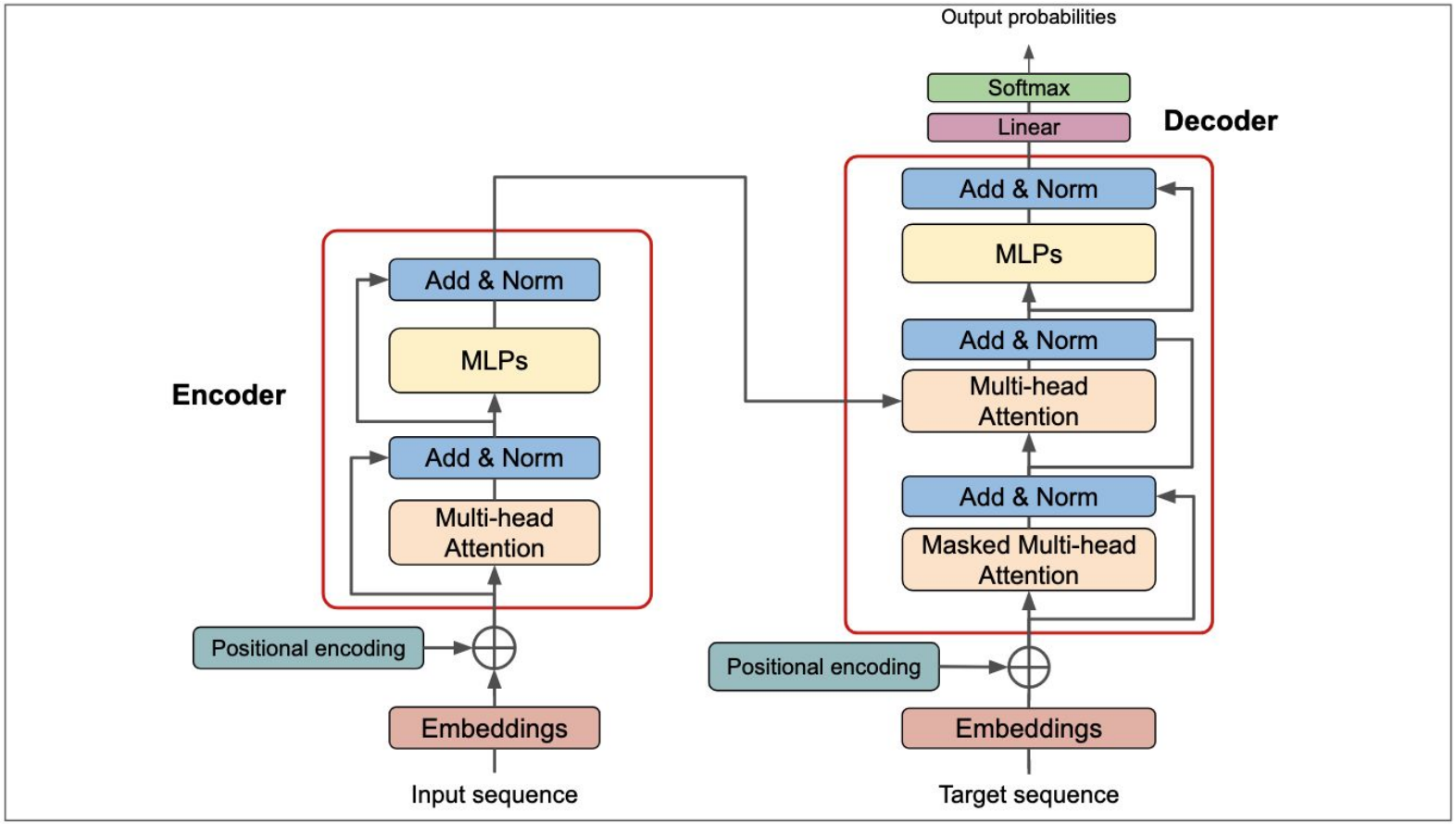
- Explore how words in English are tokenized. Try GPT 3.5 vs GPT 4o
- Can you find any weird results?
- Compare tokenization of the same phrase in both English and Arabic on GPT 3.5. What do you notice?

3 Tokenization → model input

- Say there are 100,000 tokens in the vocab
- Well… this is 100,000 dimensional input and 100,000 dimensional output of the neural network

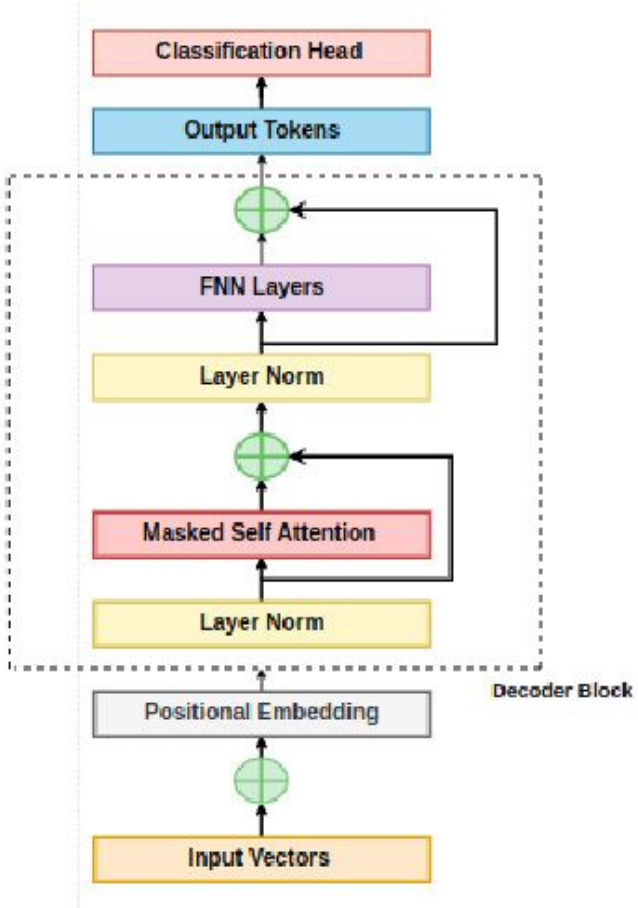


4 Model architecture



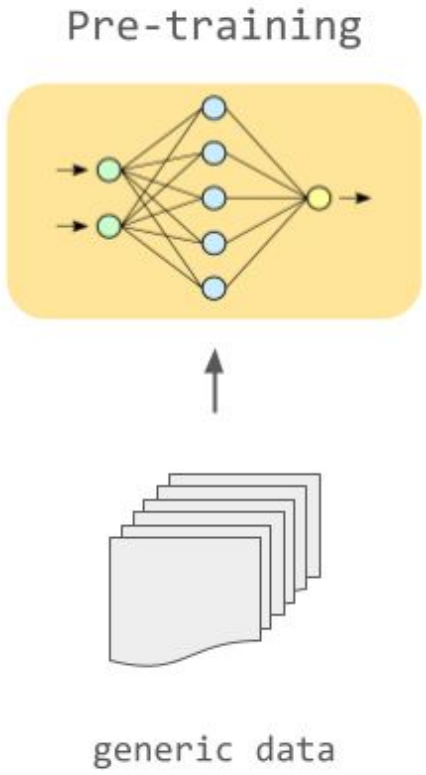
source

4 Model architecture



source

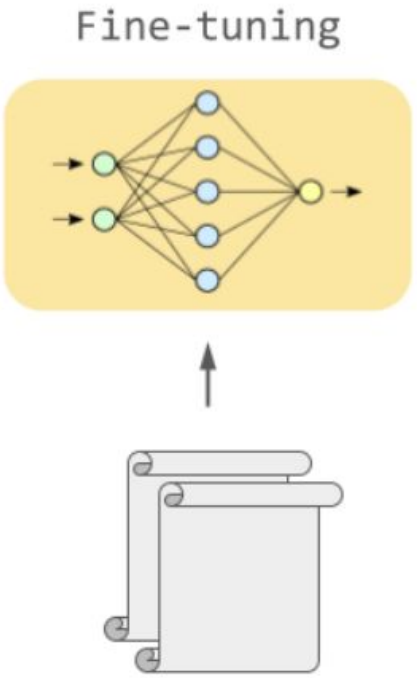
5 Turning the whole internet into tokens



"Pretraining"

- Expensive
- Takes a long time!

6 Making the model useful



"Post-training"

- Instruction tuning
- Safety finetuning
- Knowledge finetuning

-

TASK in pairs:

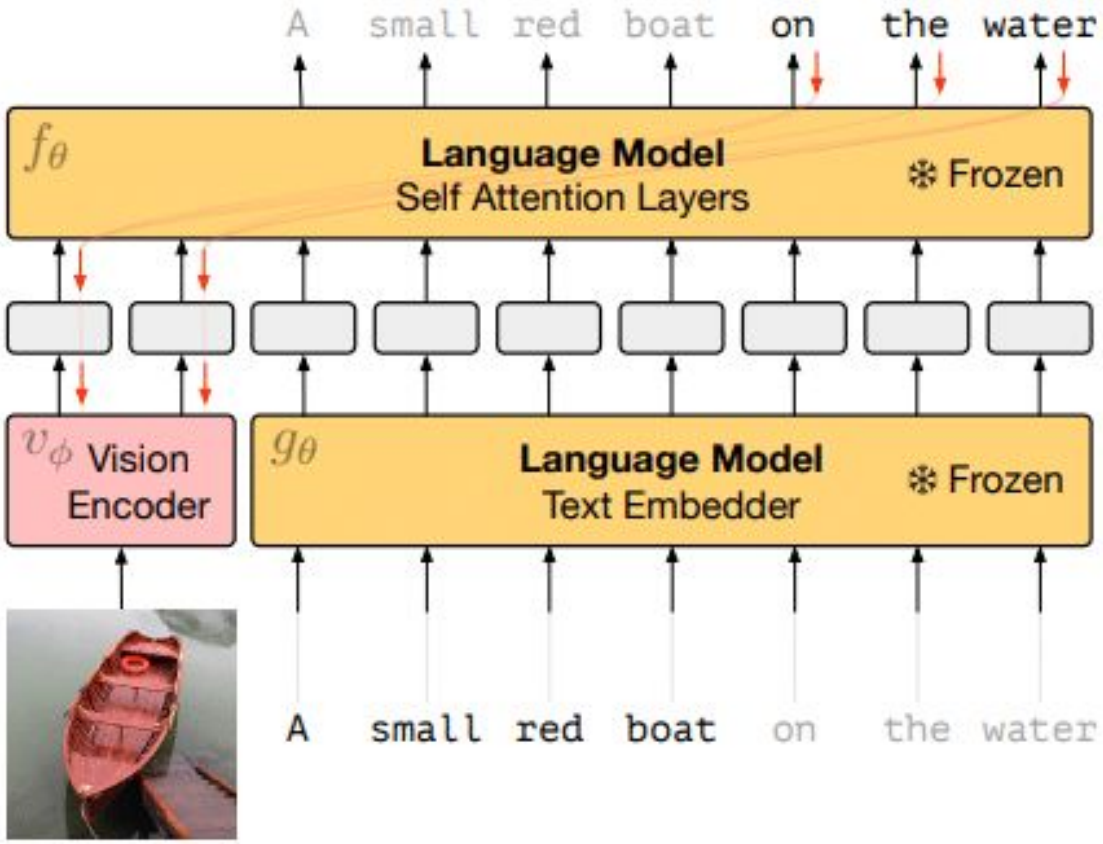
For the Llama-3-8bn-it model (released in April). Find the following information

1. Size of the vocab (number of categories)
2. Number of blocks in the transformer (layers)
3. Dimensionality of the hidden layers (“the residual stream”)
4. Amount of data it was trained on

source

7

Multimodality - exactly the same!



source

Guest Lecture