



Day 12



Overview

1

Revision time

2

Assessment II

3

Finish slides on training language models
(especially want to do the tokenizer task)

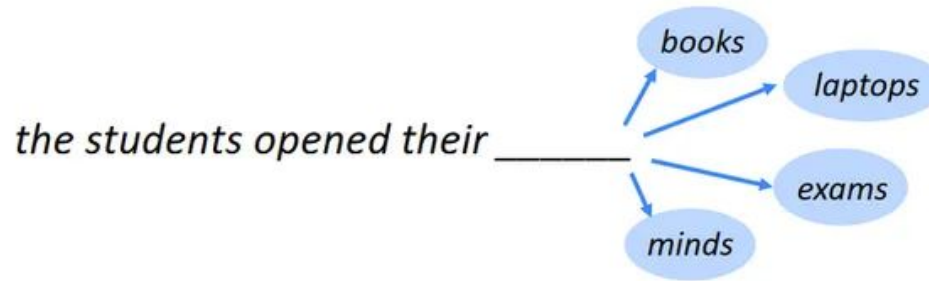
Revision Time

Assessment II

A technical overview of LLMs

1

Next token (subword) prediction



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

[source](#)



CONFIDENTIAL: Oxmedica Ltd. 10347756

OXMEDICA
The Oxford Medical Education Company



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

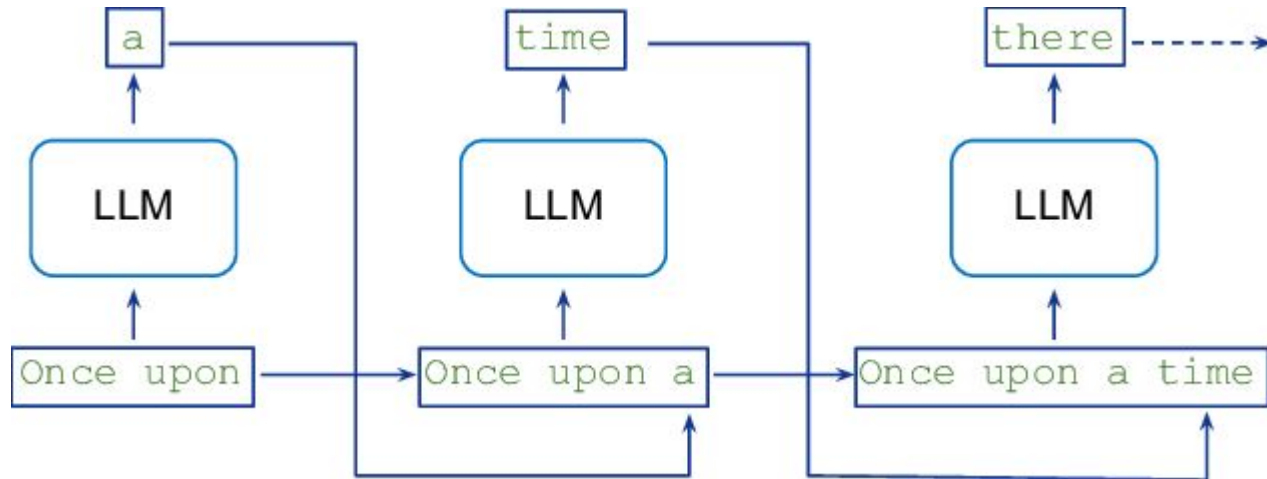
8/10

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](#)



CONFIDENTIAL: Oxmedica Ltd. 10347756



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

DoiT is a great company to work for.

36 character are represented as
11 token

Tokenized

Do i T is a great company to work for .

Tokenized ID

5211 72 51 318 257 1049 1664 284 670 329 13

3

Tokenization

DoiT is a great company to work for.

36 character are represented as
11 token

Tokenized

Do i T is a great company to work for .

Tokenized ID

5211 72 51 318 257 1049 1664 284 670 329 13

3

Tokenizer playground

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

Task in pairs: 10 mins

- Explore how words in English get tokenized. Try the GPT-3.5 tokenizer vs the GPT-4o tokenizer.
- 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?
- Try the same phrase but with the GPT-4o tokenizer. What do you notice? What do you think one of the main innovations of GPT-4o was?

3

Tokenization → Embeddings

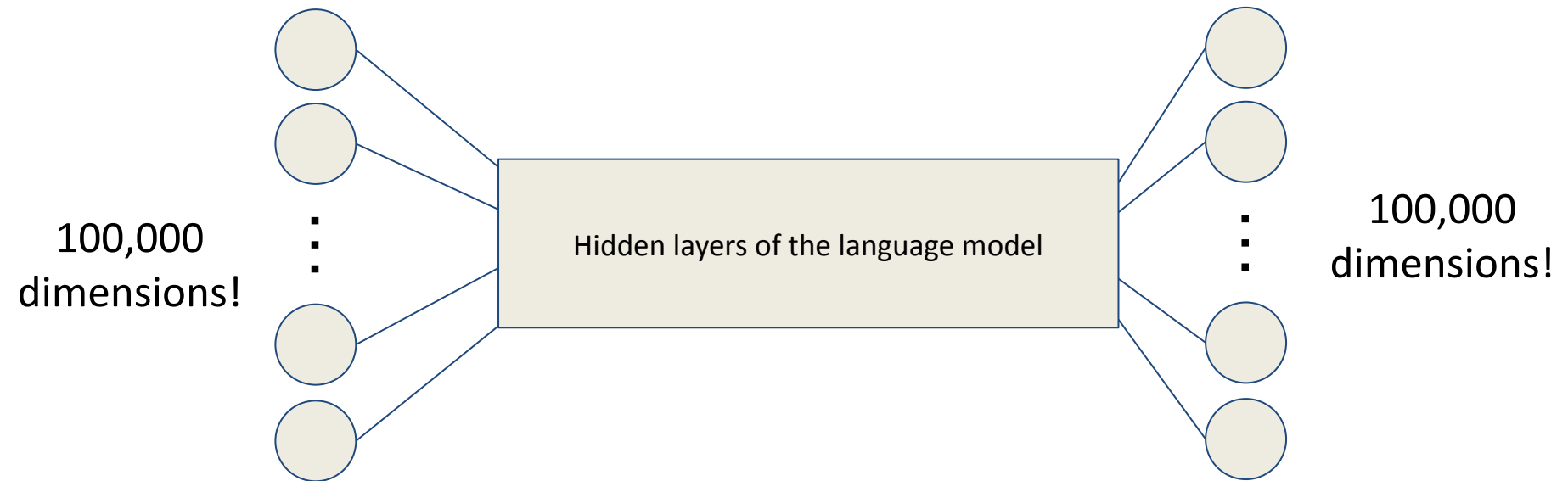
- Say there are 100,000 tokens in the vocab..

See whiteboard

3

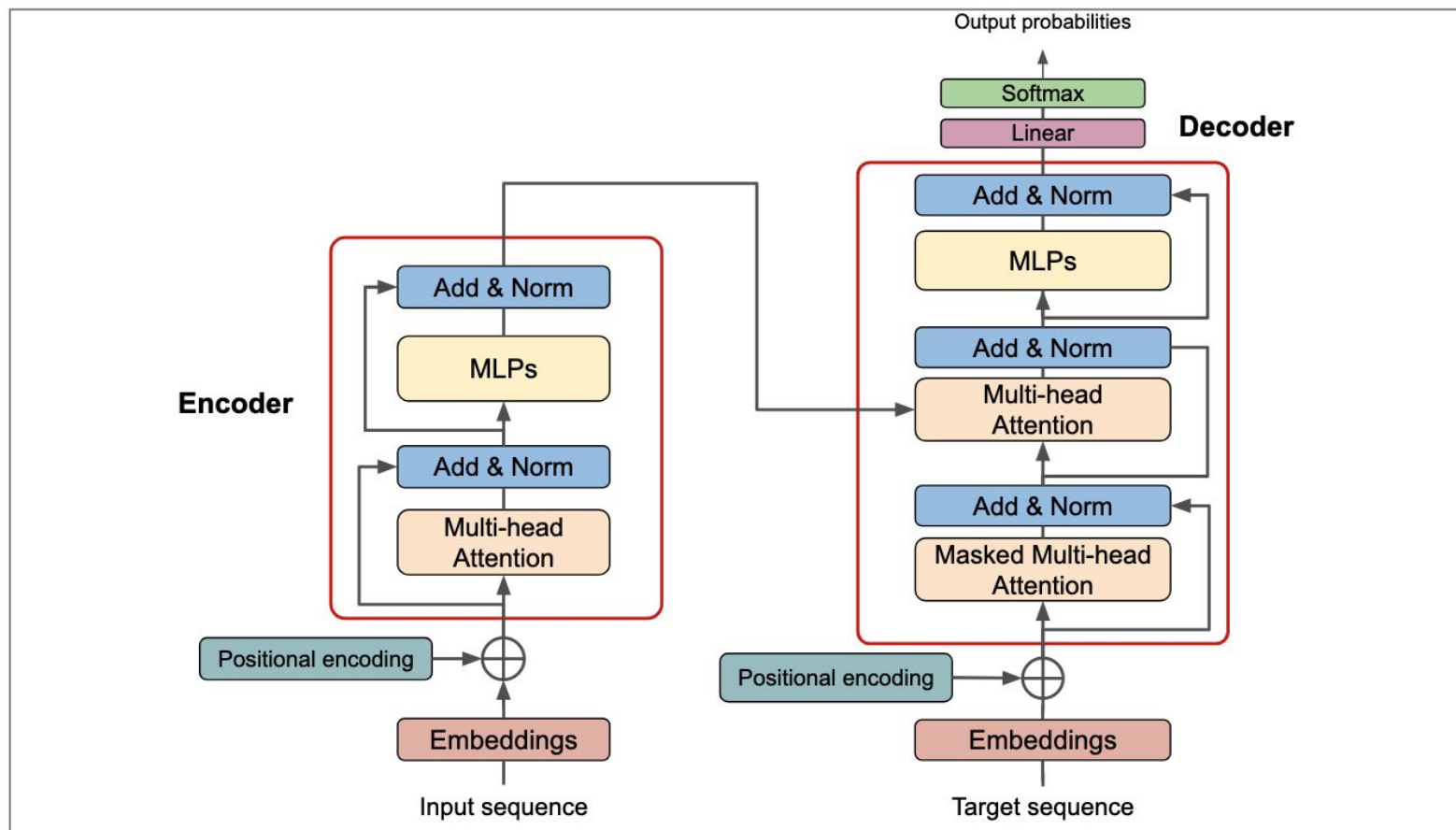
Embeddings → model input

- 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

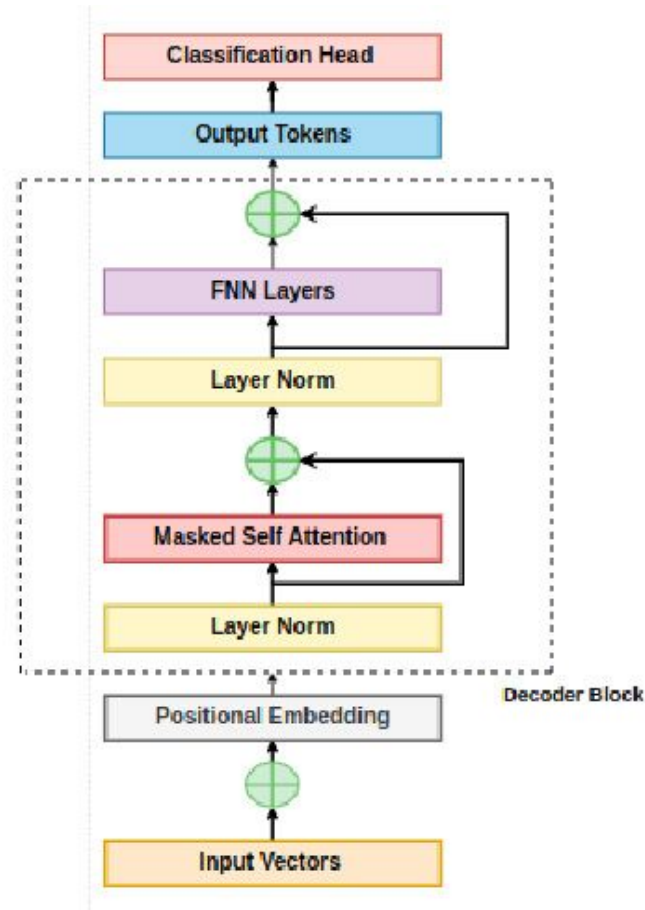
Model architecture



[source](#)

4

Actual model architecture for GPT models



[source](#)

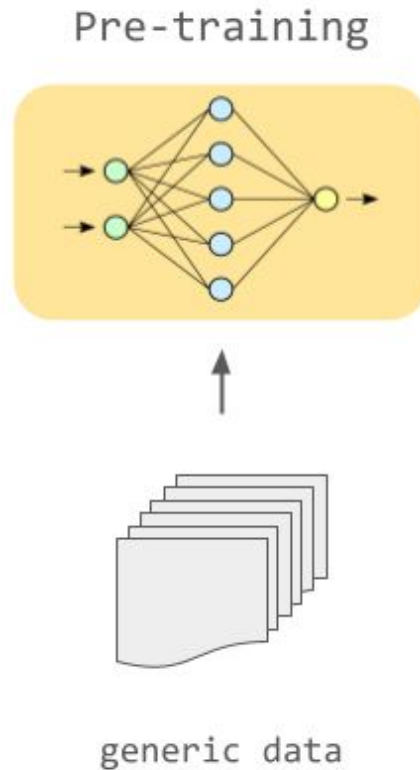


CONFIDENTIAL: Oxmedica Ltd. 10347756



5

Turning the whole internet into tokens

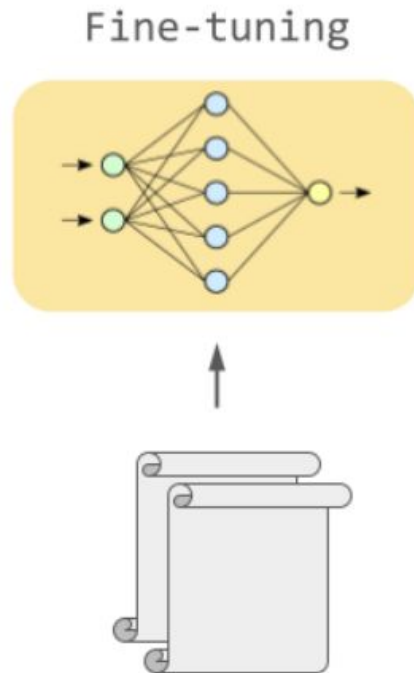


“Pretraining”

- Expensive
- Takes a long time!

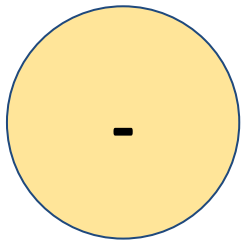
6

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

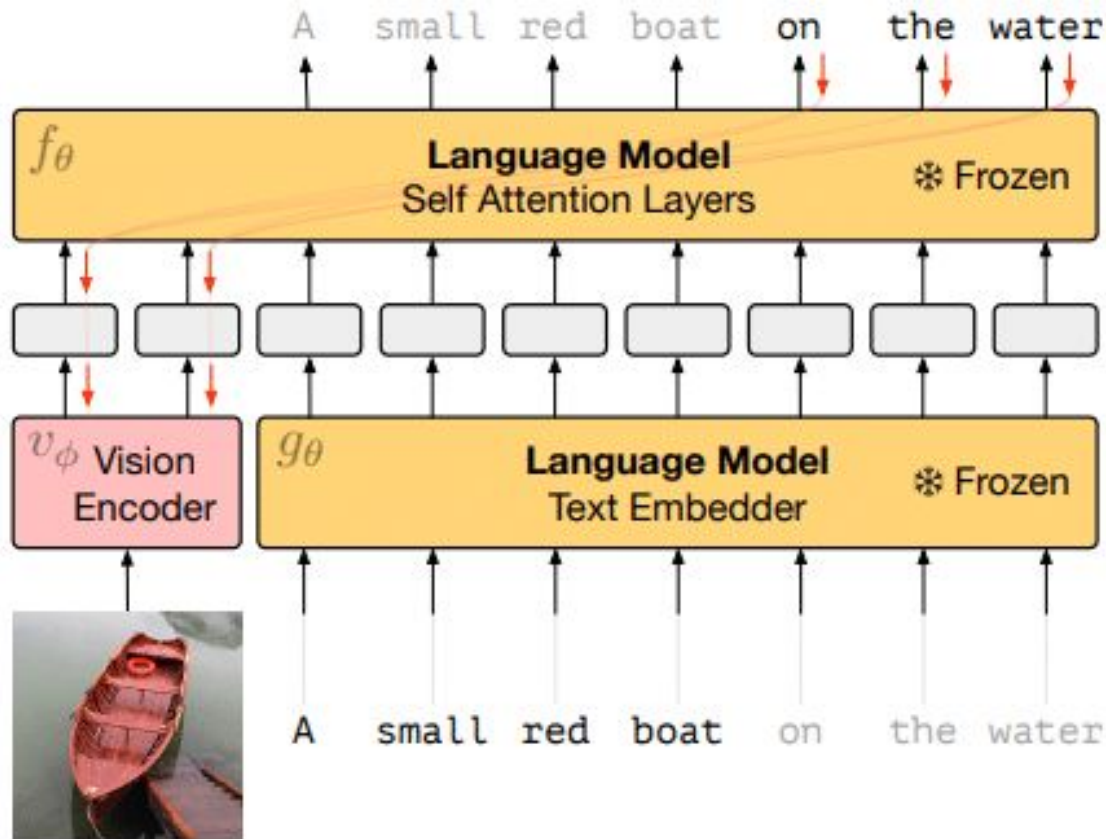
1. Size of the vocab (number of categories)
2. Number of blocks in the transformer (layers)
3. Dimensionality of the hidden layers (“the residual stream”)
4. Amount of data it was trained on

[source](#)



7

Multimodality - exactly the same!



[source](#)

Extension

1



SolidGoldMagikarp

1



What does the string " SolidGoldMagikarp" refer to?



The word "distribute" refers to the act of distributing or spreading something out among a group of people or things. It can also refer to the way in which something is distributed or arranged. For example, a company may distribute its products to different retailers, or a teacher may distribute assignments to students. Additionally, it can mean to give out something in an orderly way, such as distributing flyers or pamphlets.  

[source](#)



CONFIDENTIAL: Oxmedica Ltd. 10347756



2

What are special tokens?

The MoE Architecture