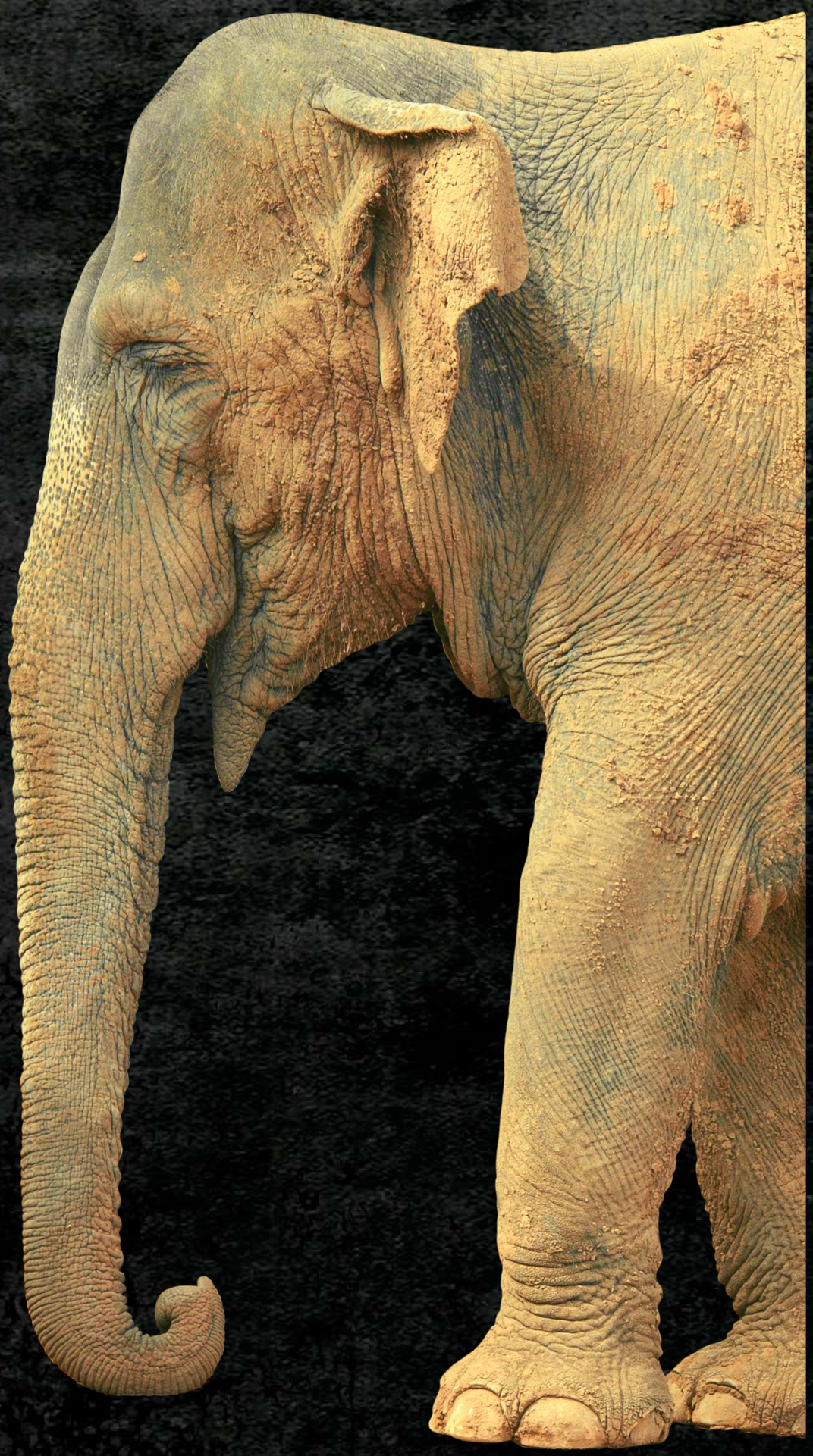
01

A model with 175 billion parameters



TECHNICAL GUIDE TO GPT-3

@learn.machinelearning

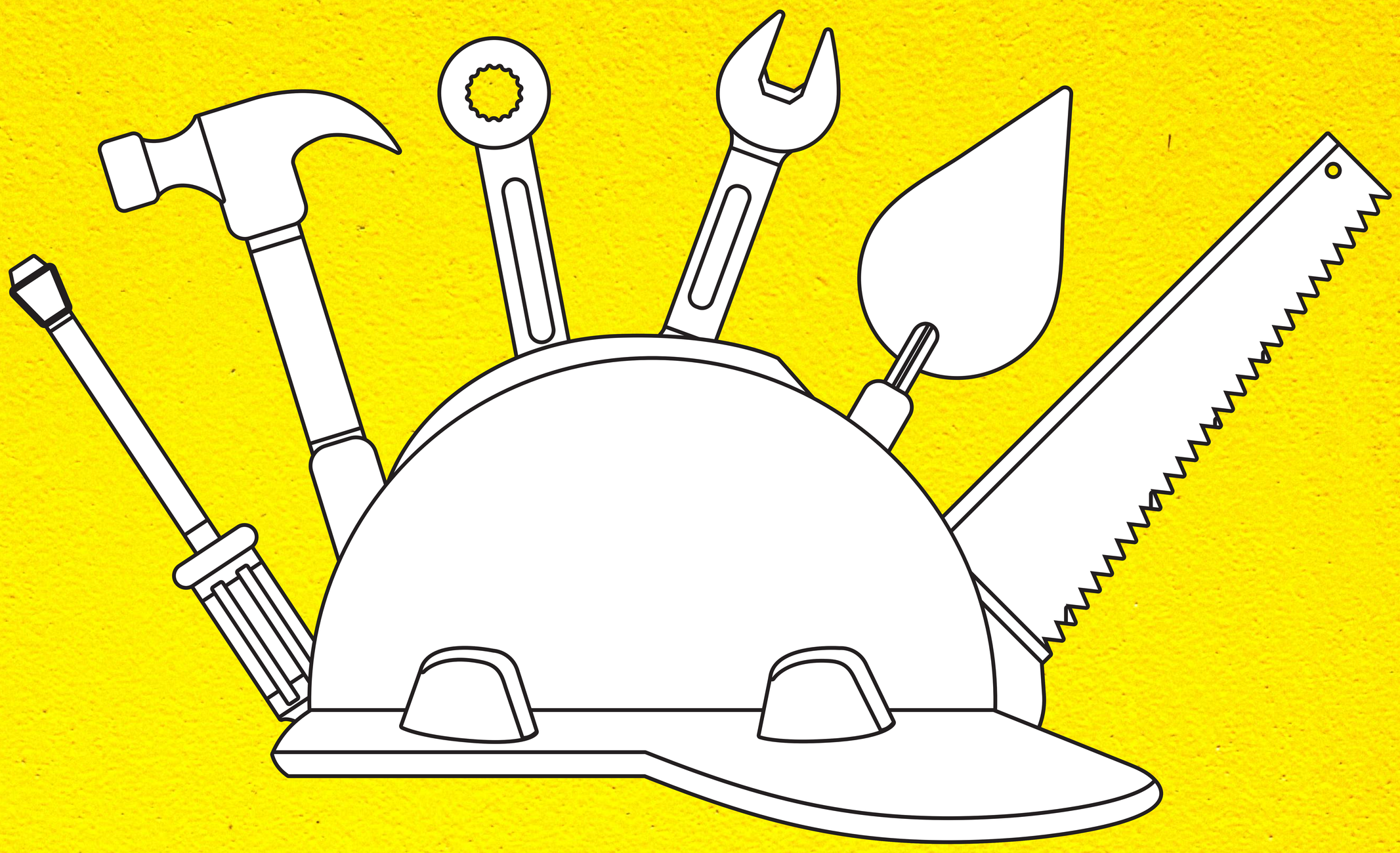
02

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

DATA USED TO TRAIN GPT-3

- GPT-3 175B is trained with 499 Billion tokens.
- While GPT-2 1.5B is trained with 40GB of Internet text, which is roughly 10 Billion tokens (conversely assuming the average token size is 4 characters).
- So GPT-3 175B has a lower data compression ratio of 2.85 in comparison to GPT-2 1.5G which has 6.66.

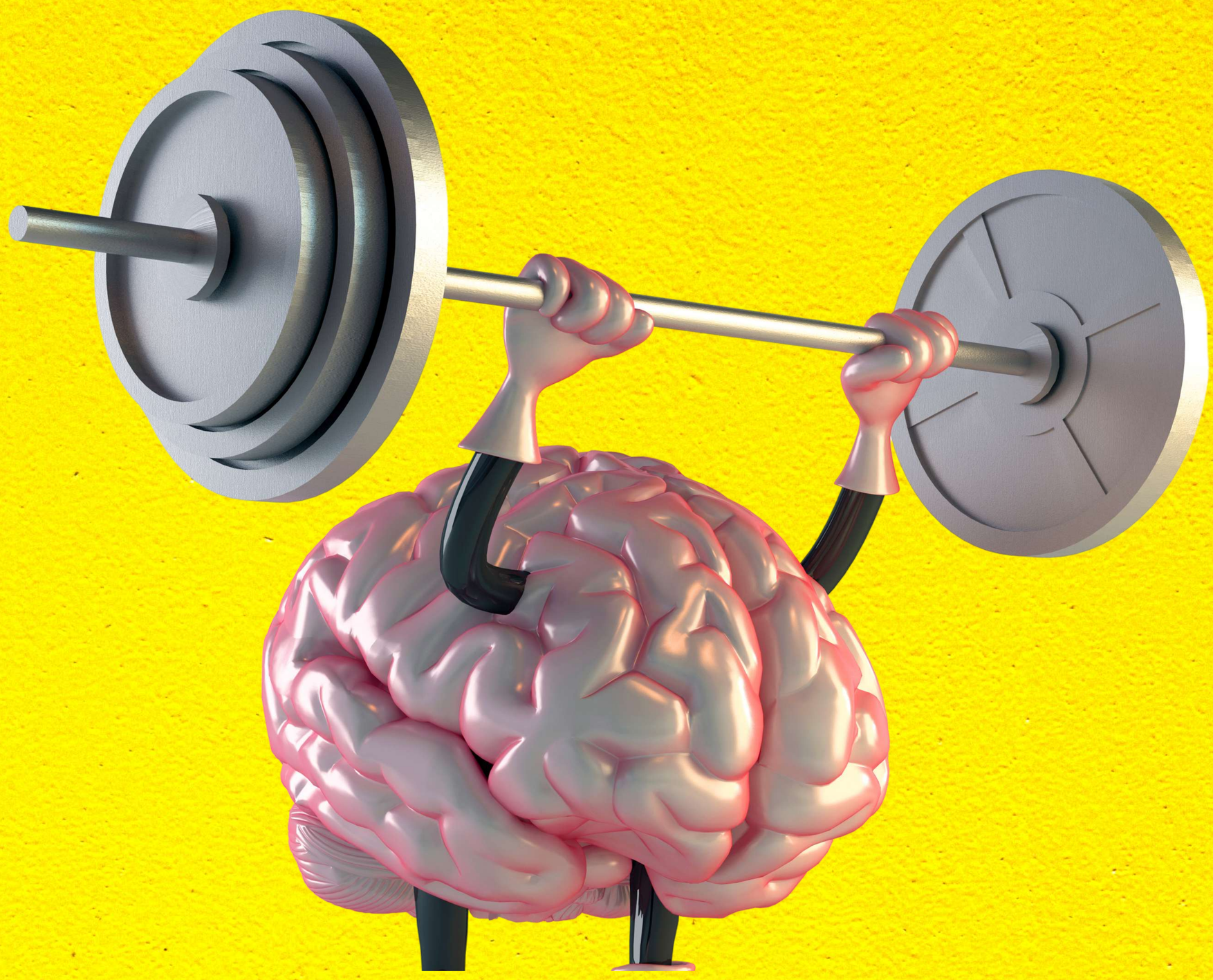
03



ARCHITECTURE OF GPT-3

- The OpenAI team used the same model and architecture as GPT-2 that includes modified initialisation, pre-normalisation, and reversible tokenisation along with alternating dense and locally banded sparse attention patterns in the layers of the transformer.

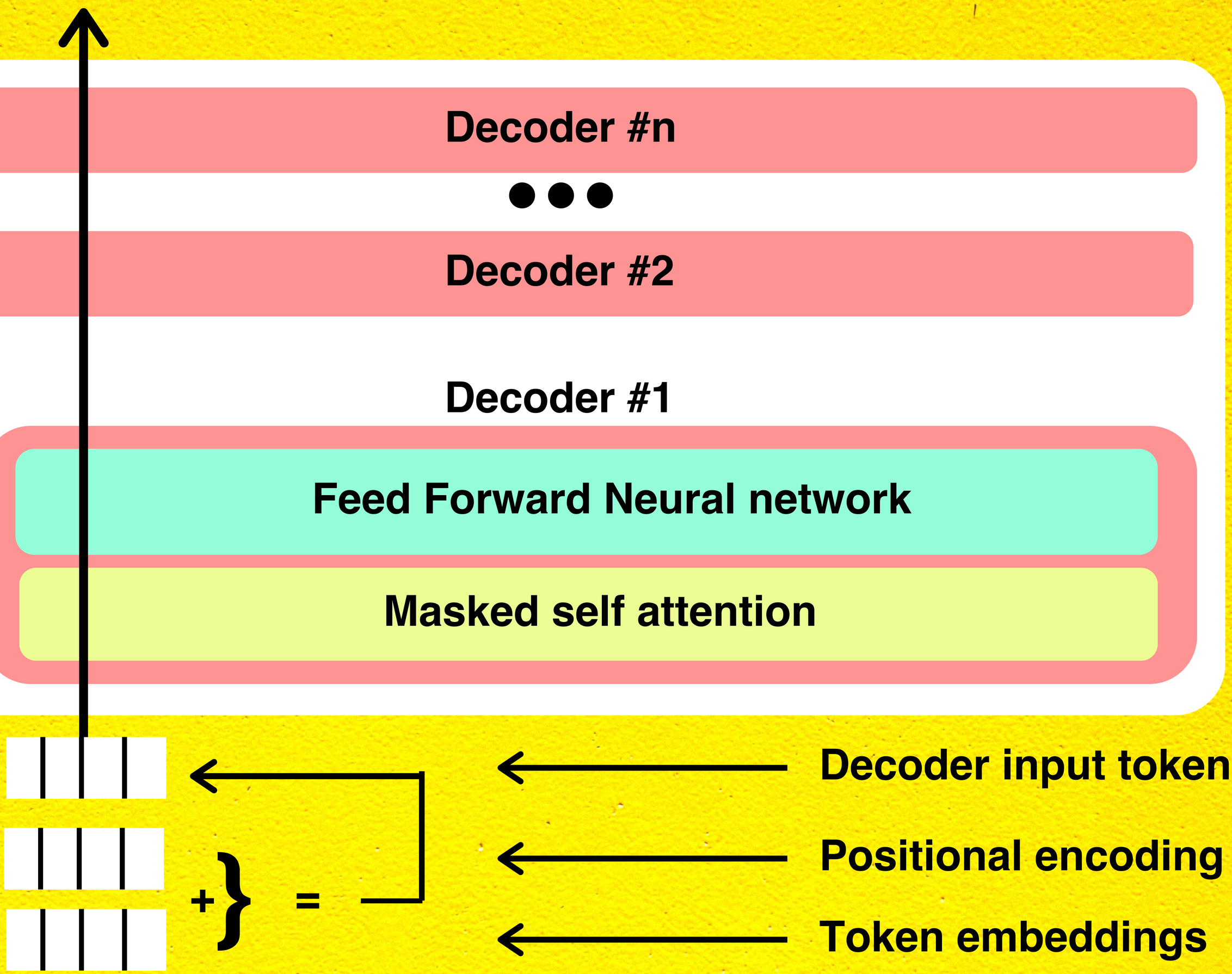
04



TRAINING OF GPT-3

- Because of the architecture it eventually helped GPT-3 (175 B) model to use 3.2 M as batch size for its training.
- It was given that all these models are trained on the same number of tokens and trained on V100 GPU'S having high bandwidth provided by Microsoft.

05



TRANSFORMER - DECODER ACRHITECTURE

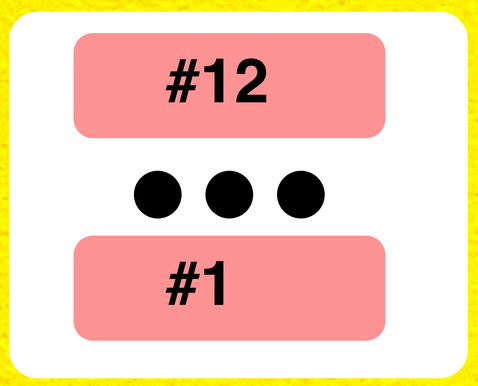
learn.machinelearning

A BIT MORE ABOUT ARCHITECTURE

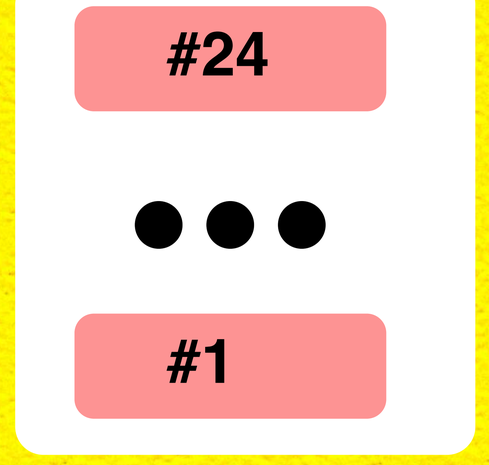
- GPT-2 or GPT-3 uses Transformer - Decoder acrhitecture.
- The flow looks like first tokens are converted into woed embeddings and then added with positional embeddings(specific pattern about the position of each word).
- This is applied to all tokens in a sequence which is limited to 2048.
- Then the input token is passed through first token and masked self attention layer.
- Masked Self Attention is similar to Self-attention expect that in the former one model can’t see the words after the particular current word.
- For each attention head, scores of tokens are calculated and this final matrix is sent to the feed-forward network.
- In feed forward network again there are two layers.
- The output of the feed forward network is the output of the particular decoder layer.

DIFFERENT GPT - 3 MODELS

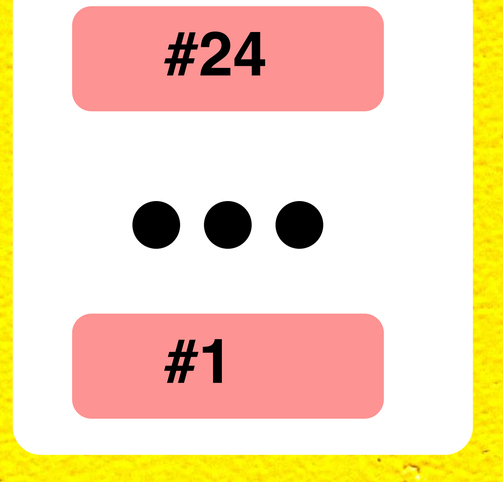
- The team developed 8 different models varying in parameters, layers and model size ranging from 125 Million Parameters to 175 Billion Parameters.



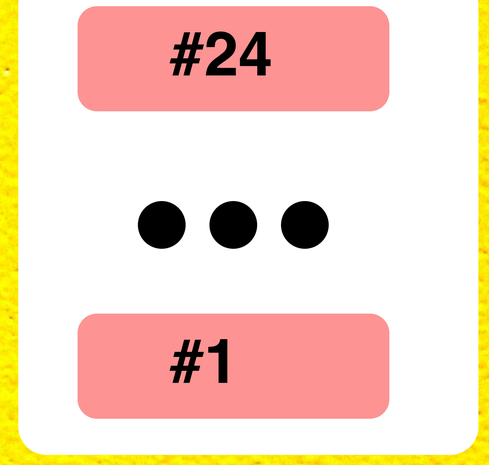
GPT-3 small
12 Decoder layers
125M parameters



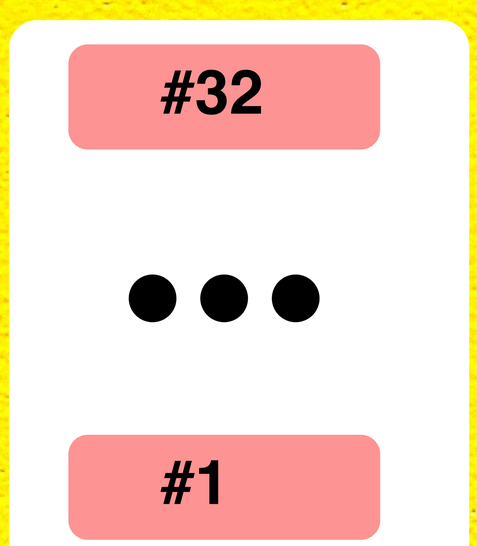
GPT-3 medium
24 Decoder layers
350M parameters



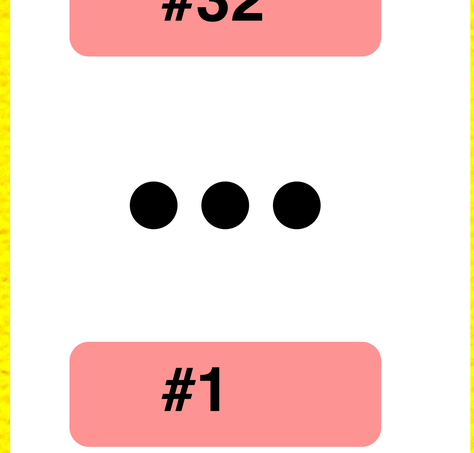
GPT-3 Large
24 Decoder layers
760M parameters



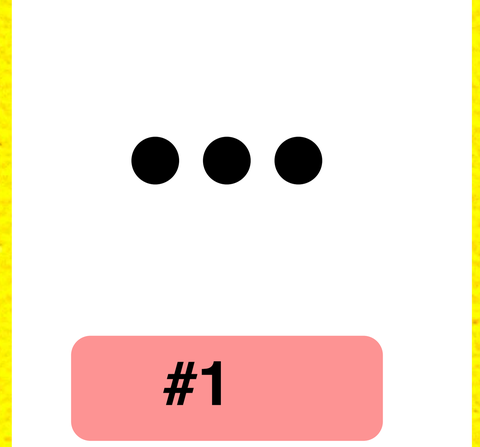
GPT-3 XL
24 Decoder layers
1.3B parameters



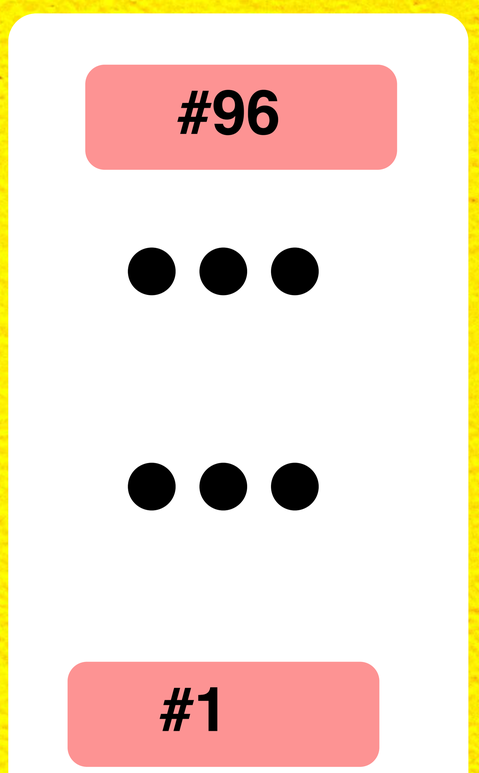
GPT-3 2.7B
32 Decoder layers
2.7B parameters



GPT-3 6.7B
32 Decoder layers
6.7B parameters



GPT-3 13B
40 Decoder layers
13B parameters



GPT-3 170B or GPT-3
96 Decoder layers
170B parameters

0

Translate English to French:

Cheese =>

← **Task description**

← **Prompt**

1

Translate English to French:
Sea otter => loutre de mer

Cheese =>

← **Task description**

← **Example**

← **Prompt**

1+

Translate English to French:
Sea otter => loutre de mer
peppermint => menthe poivrée

← **Task description**

← **Example - 1**

← **Example - 2**

EVALUATION OF GPT - 3

- Evaluation of GPT-3 is done under 3 conditions
- Few-shot learning: In addition to the model description, the model sees a few examples of the task.
- One-shot learning: The model sees one example of the task.
- Zero-shot learning: The model predicts the answer given only a natural language description of the task.
- GPT-3 achieved promising results on all three

09

APPLICATIONS OF GPT - 3

- GPT-3 as an Author
- GPT-3 as a blogger
- Search Engine
- Spreadsheet auto-completion
- Simple English to SQL
- English to Layout generator
- Full blow UI design
- Text to CSS
- Text to LaTeX
- English to Keras
- Text to manim

learn.machinelearning

RESOURCES

CLICK THE LINKS TO GET RESOURCES

@learn.machinelearning

- Find all GPT-3 resources here