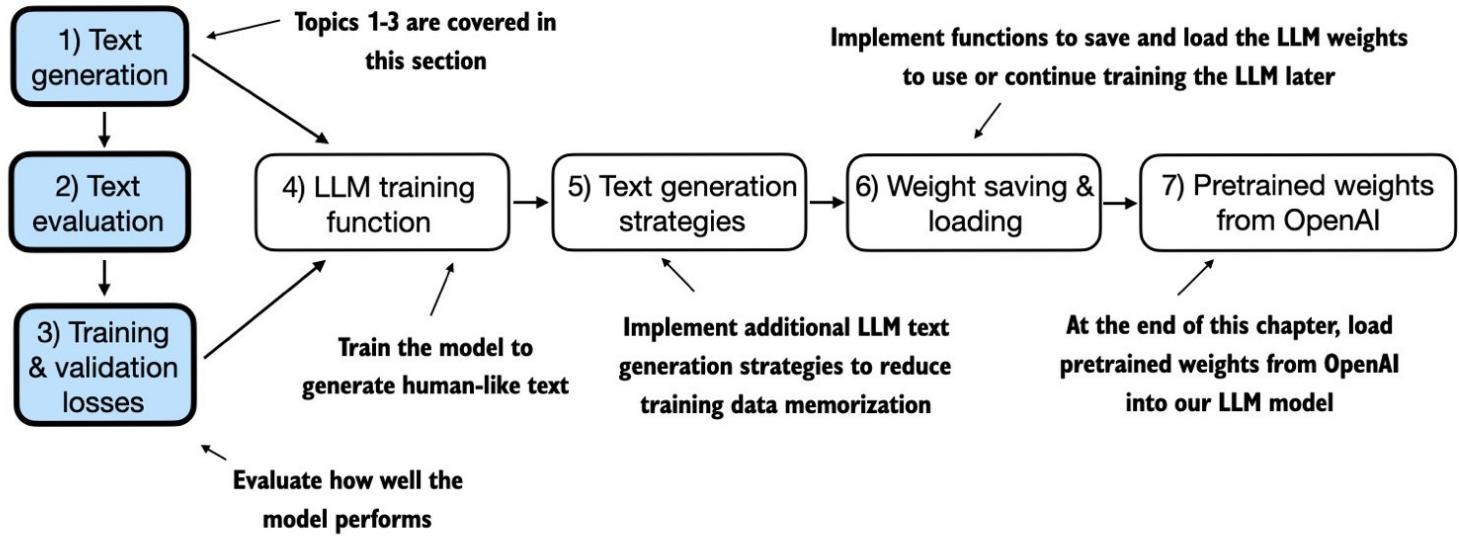


Entire 124 m parameter GPT-2 model

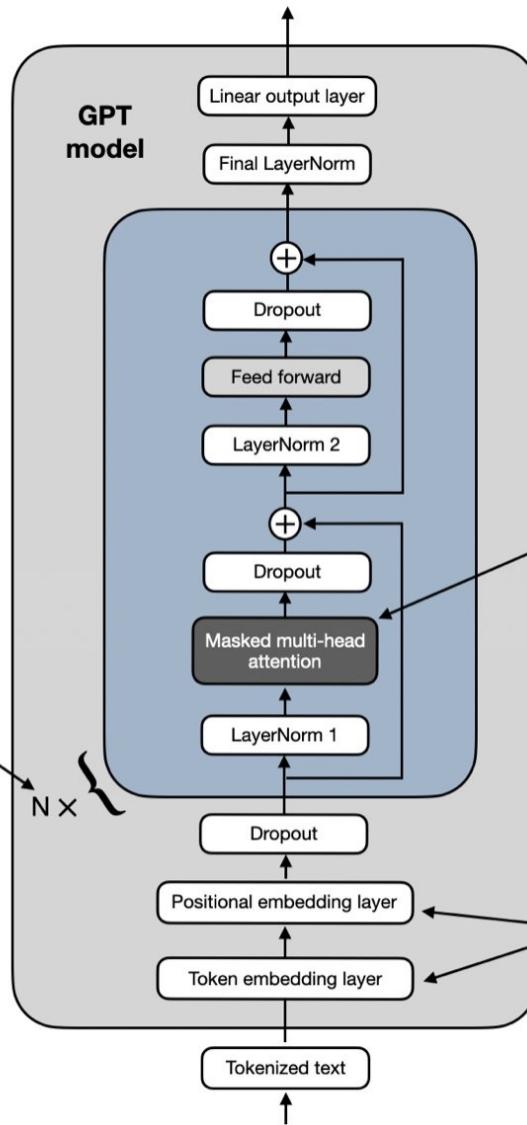


Total number of parameters:

- 124 M in "gpt2-small"
- 355 M in "gpt2-medium"
- 774 M in "gpt2-large"
- 1558 M in "gpt2-xl"

Repeat this transformer block:

- 12 X in "gpt2-small"
- 24 X in "gpt2-medium"
- 36 X in "gpt2-large"
- 48 X in "gpt2-xl"



Every Effort moves you

Input batch:

```
tensor([[6109, 3626, 6100, 345],  
       [6109, 1110, 6622, 257]])
```

① Token Embedding:

Token embedding is a way to convert words (or tokens) into numbers that a machine learning model can understand.

Every



Effort



moves



you



Embedding size = 768

② Positional Embedding

Positional embedding tells the model where each word appears in a sentence.

Position 1



Position 2



Position 3



Position 4



Embedding size = 768

③ Input embedding = Token Embedding + Positional Embedding

Every (Position i) :

Token embedding



Positional embedding



Input embedding



Embedding size 768

Effort (Position 2):

Token embedding



Positional embedding



Input embedding



Embedding size 768

moves (Position 3):

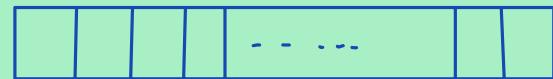
Token embedding



Positional embedding



Input embedding



Embedding size 768

you (Position 4):

Token embedding



Positional embedding



Input embedding



Embedding size 768

④ Dropout:

Prevent overfitting or Improve Generalization

Input Embedding

Every



Effort



moves

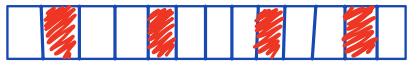
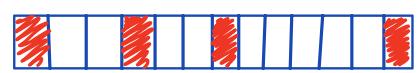
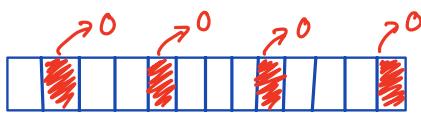


you



Embedding size = 768

Randomly turn off
some elements to 0



⑤ Transformer:

Input embedding with dropout

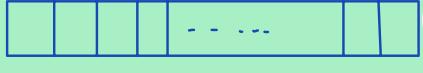
Every



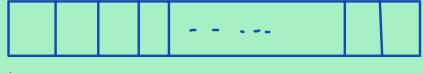
Effort



moves



you



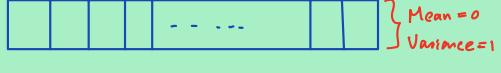
Embedding size = 768

Layer Normalization

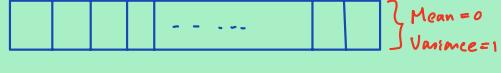
Every



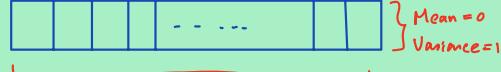
Effort



moves



you

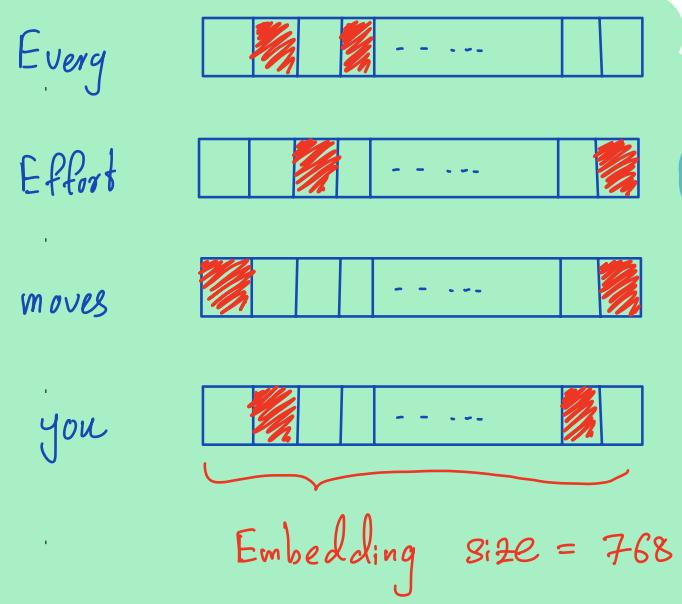


Embedding size = 768

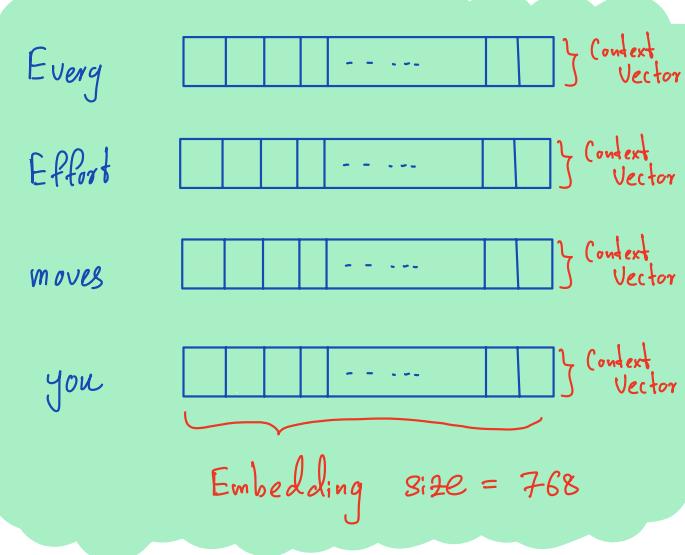
Masked

Multi head attention

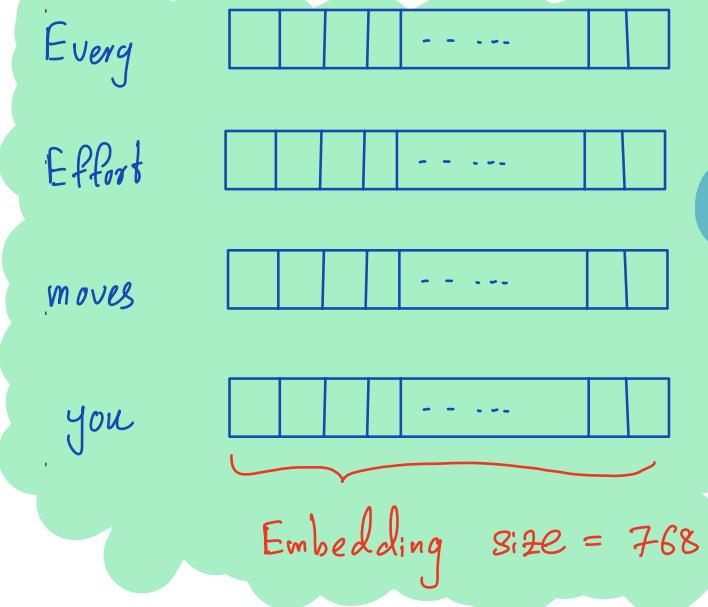
(How much attention should give to other tokens)



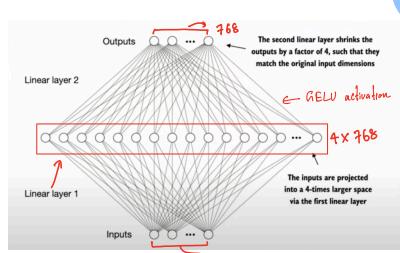
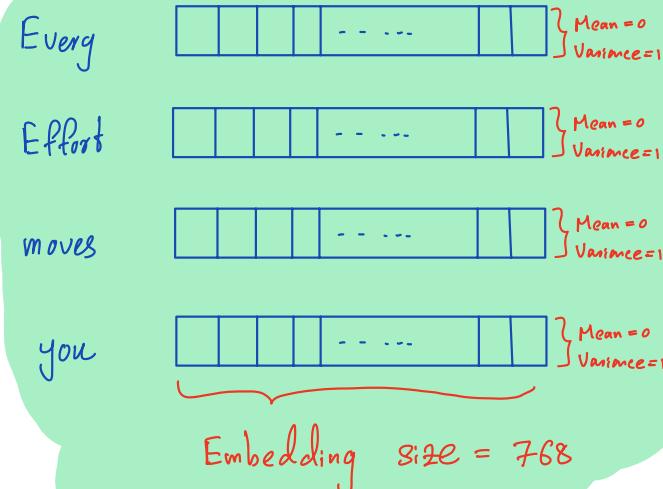
Dropout



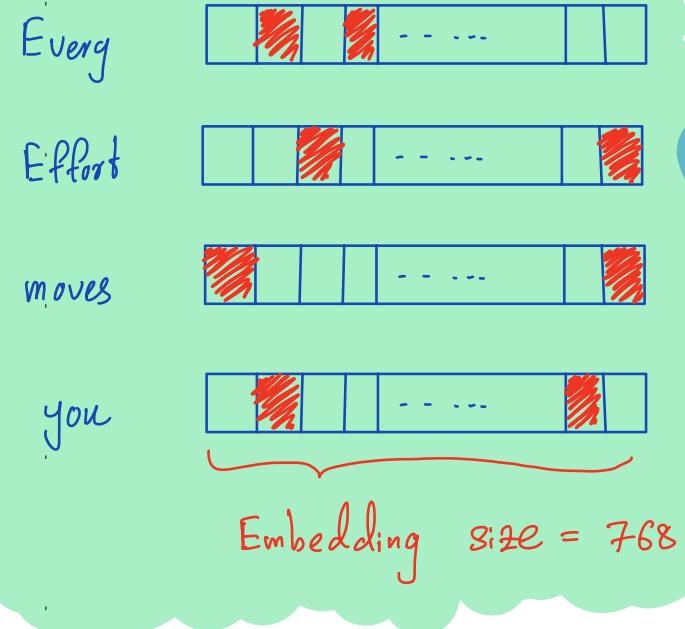
Shortcut Connection



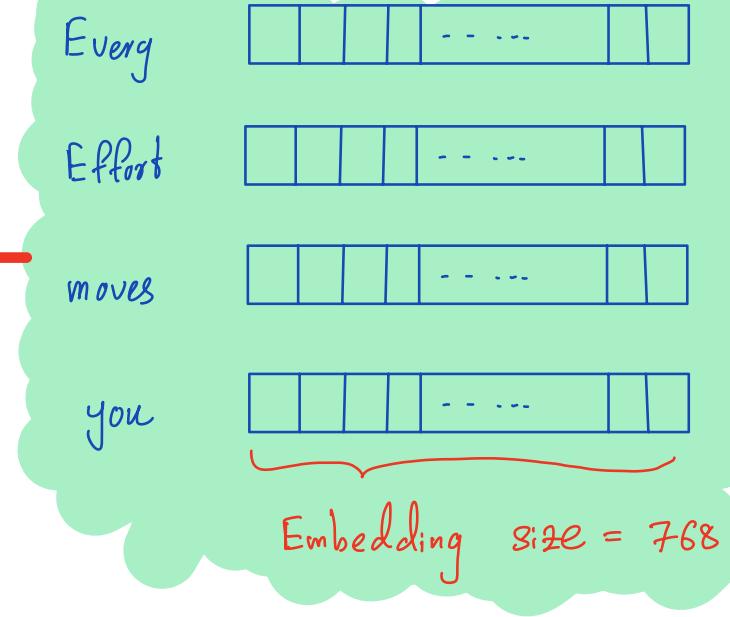
Layer Norm



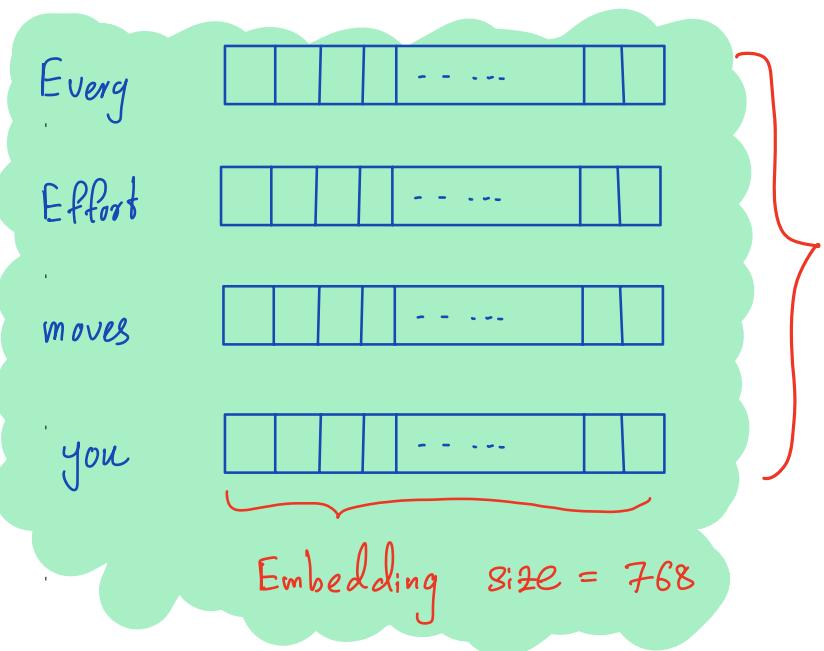
Feed Forward Neural Network



Dropout

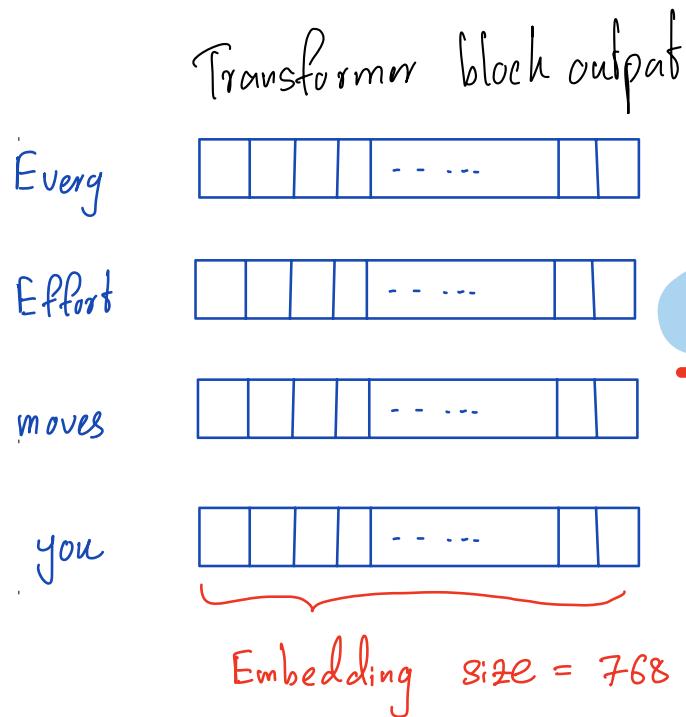


Shortcut Connection

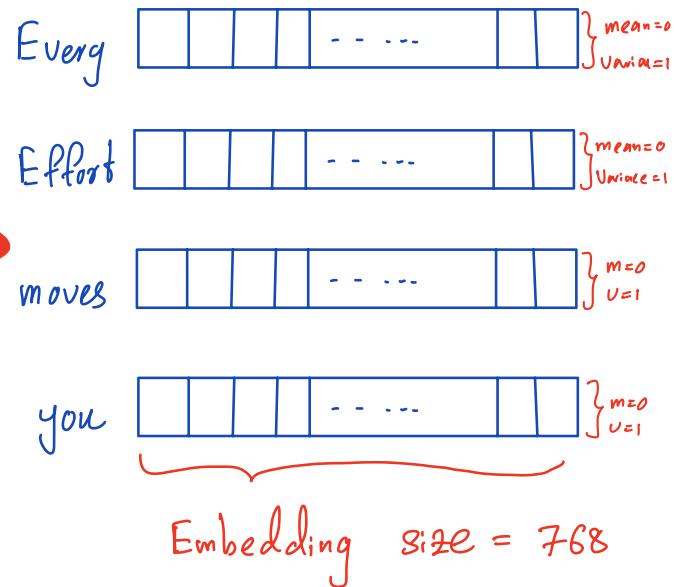


Transformer Block output

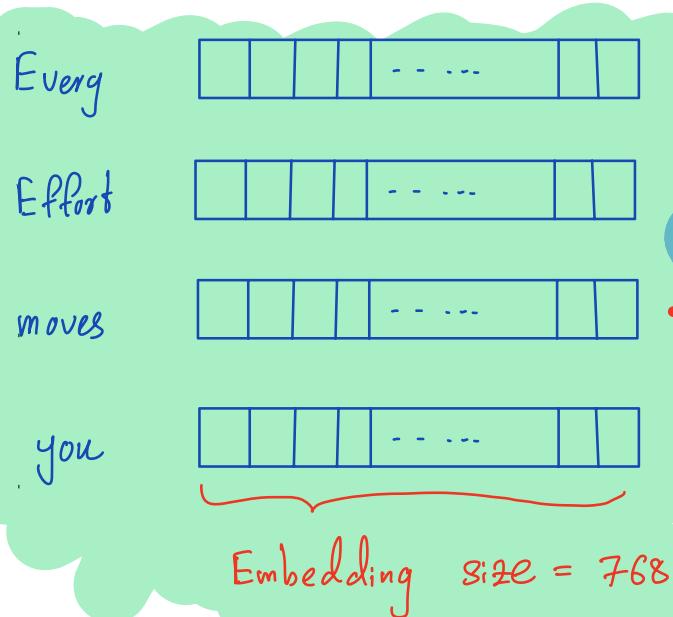
⑥ Layer Normalization (after the transformer output)



Layer Norm

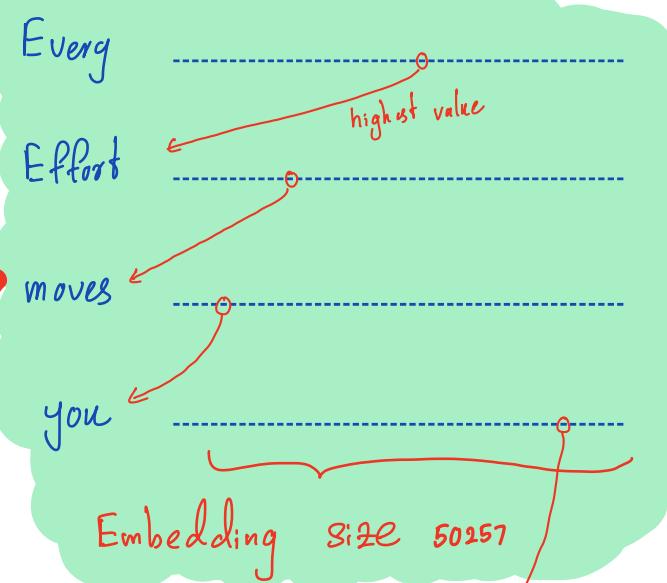


⑦ Output Head



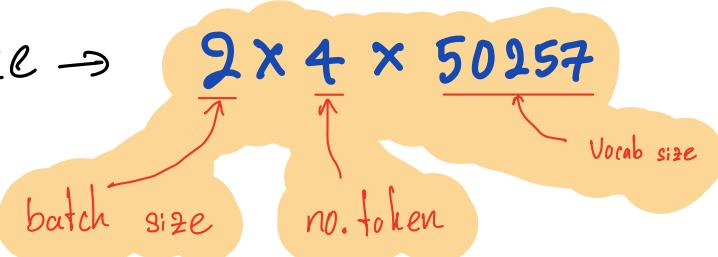
Neural Network Layer
(768 x 50257)

Output logist for one text sample



Each element corresponds to the probability of it being the next token.

Output tensor value → $2 \times 4 \times 50257$



The diagram illustrates the dimensions of an output tensor. It consists of three orange cloud-like shapes. The first shape contains the text "batch size" with a red arrow pointing to the first dimension "2". The second shape contains the text "no. token" with a red arrow pointing to the second dimension "4". The third shape contains the text "Vocab size" with a red arrow pointing to the third dimension "50257".