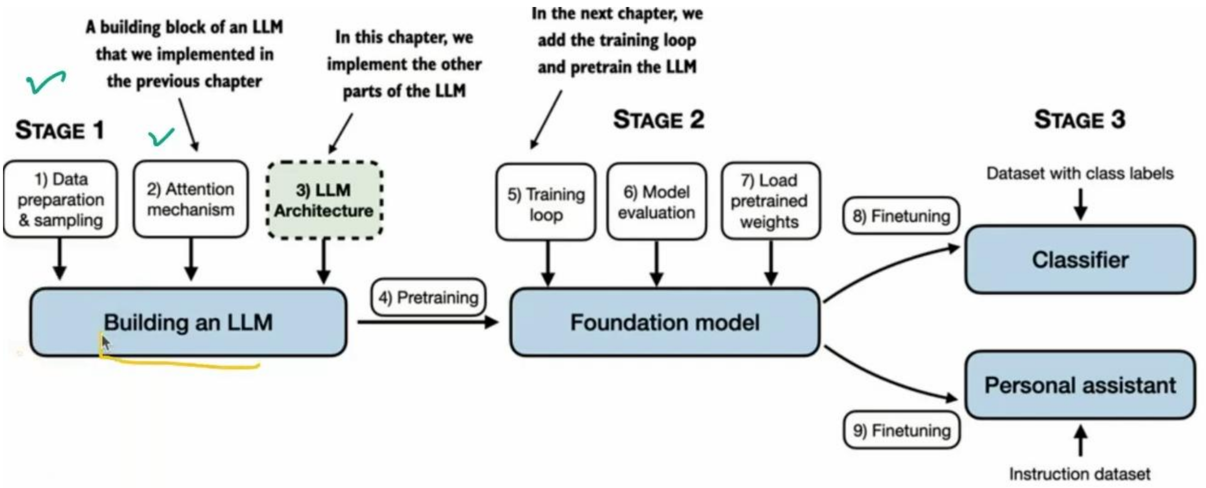
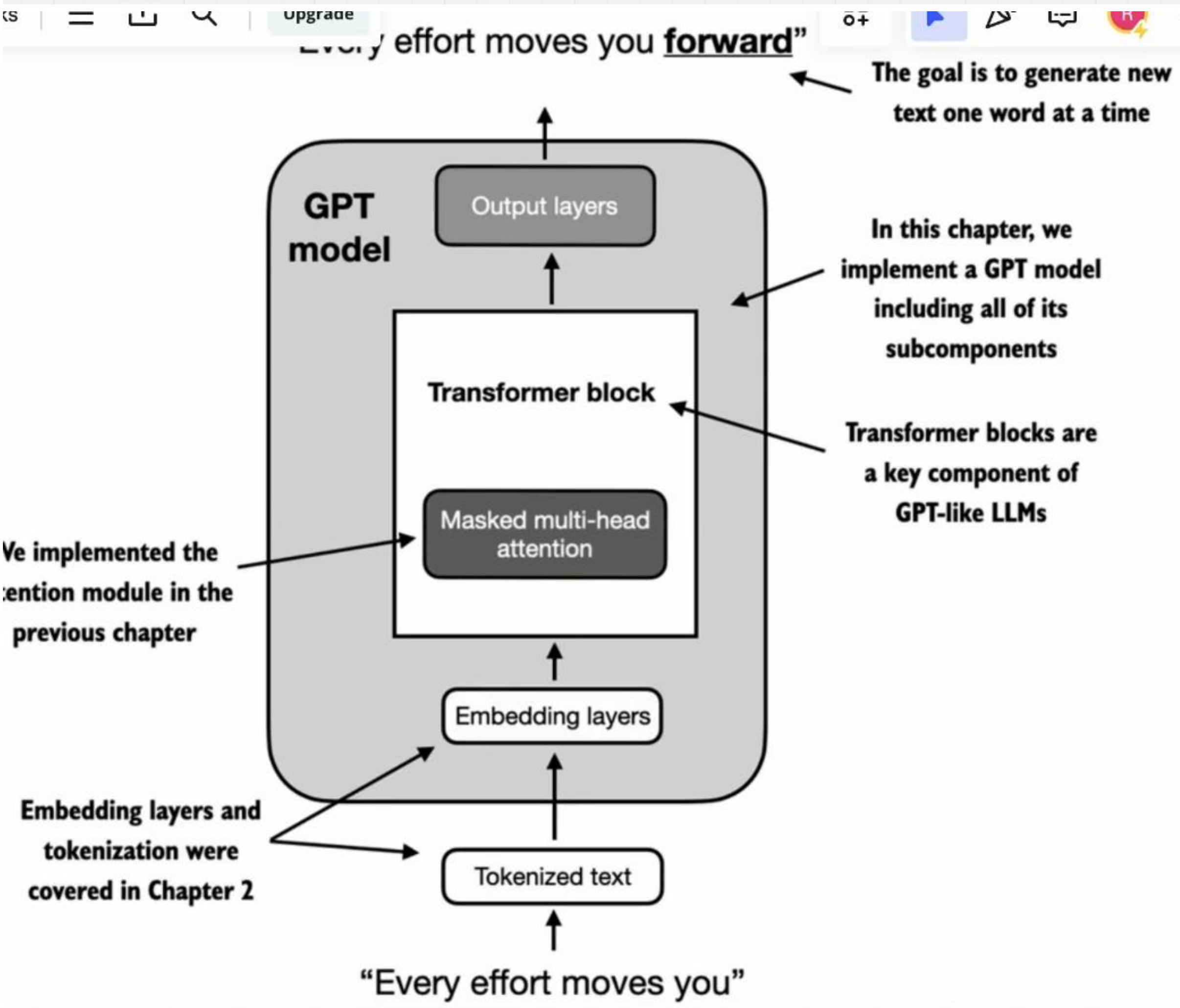
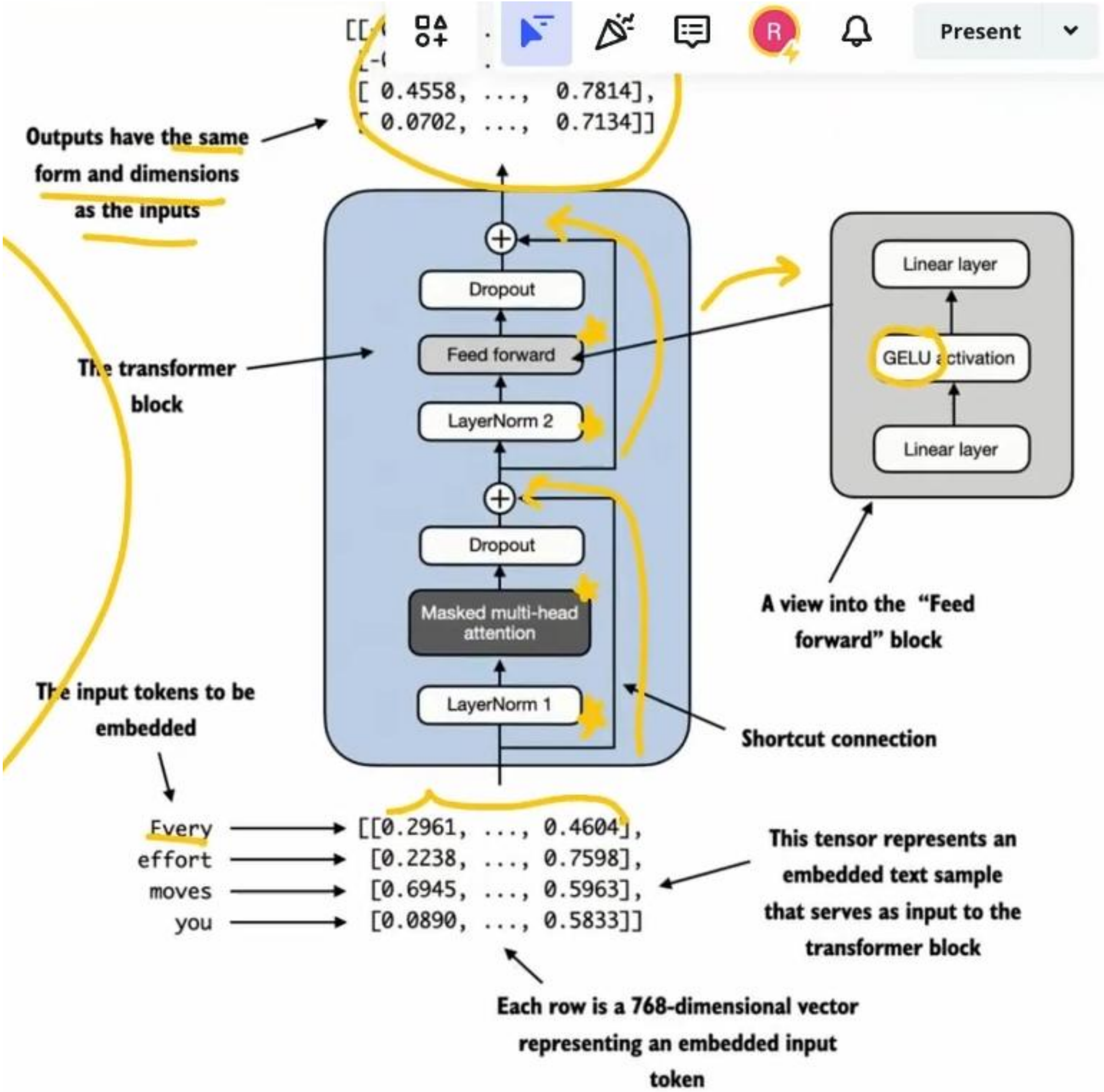


After learning about the attention mechanism, let us learn about the LLM architecture now.



Zoom Into the Transformer block



(3) What we are yet to learn:

(a) Transformer blocks

(4) We will scale upto the size of a small GPT-2 model → 124 million parameters

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Sutskever [**1]

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting

competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al., 2017), reading comprehension systems (Jia & Liang, 2017), and image classifiers (Alcorn et al., 2018) on the diversity and variety of possible inputs highlights some of the shortcomings of this approach.

Our suspicion is that the prevalence of single task training on single domain datasets is a major contributor to the lack of generalization observed in current systems. Progress towards robust systems with current architectures is likely to require training and measuring performance on a wide range of domains and tasks. Recently, several benchmarks

This was corrected to 124 later

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

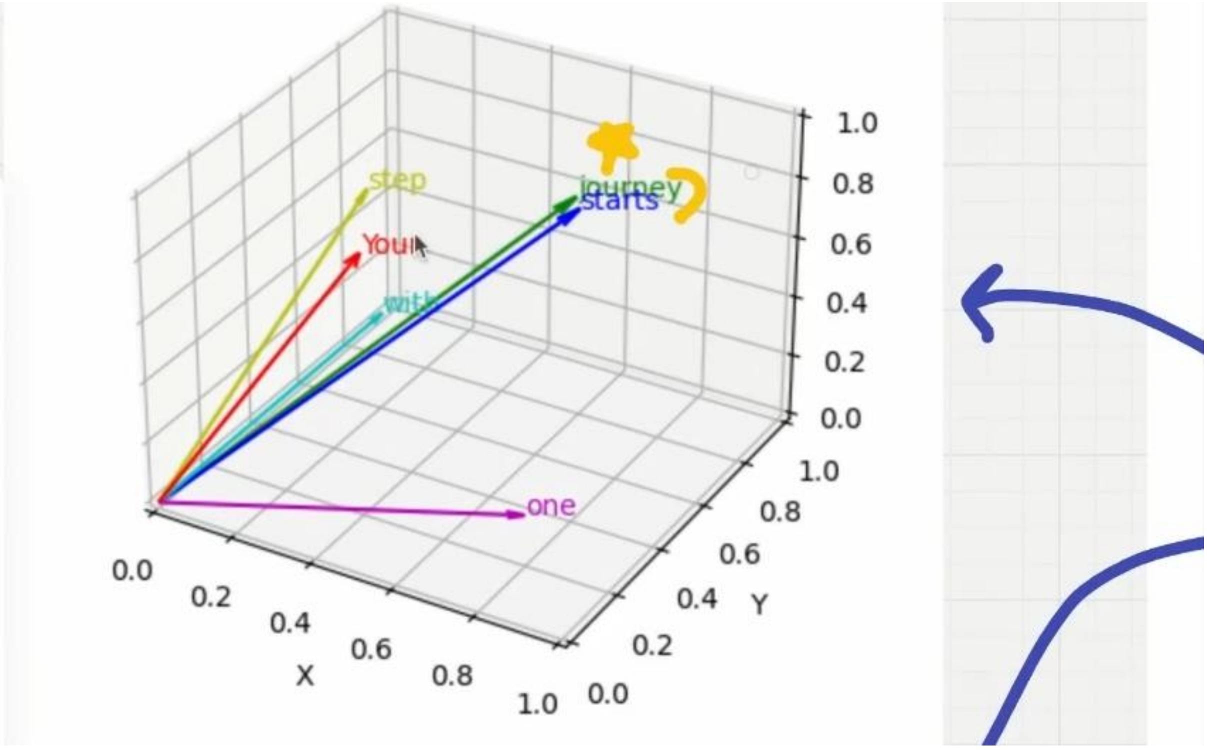
⑤ OpenAI has made GPT-2 weights public.
GPT-3,4 weights have not yet been made public.

We will use these parameters

GPT_CONFIG_124M = {

```
GPT_CONFIG_124M = {
    "vocab_size": 50257,  # Vocabulary size
    "context_length": 1024,      # Context length
    "emb_dim": 768,          # Embedding dimension
    "n_heads": 12,           # Number of attention hea
    "n_layers": 12,          # Number of layers
    "drop_rate": 0.1,        # Dropout rate
    "qkv_bias": False        # Query-Key-Value bias
}
```

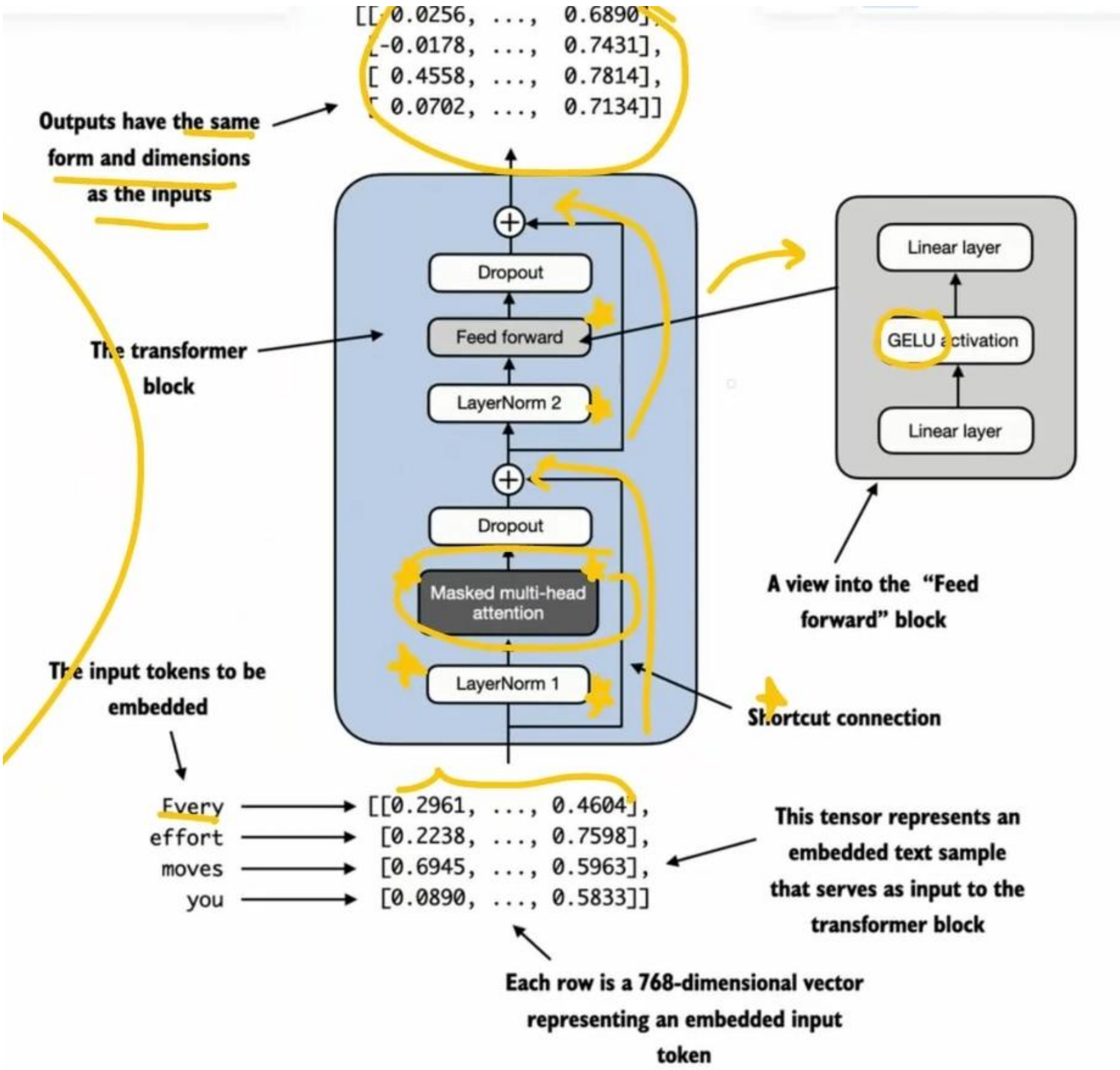
Context length: how many maximum words are used to predict the next word(here word means tokens)

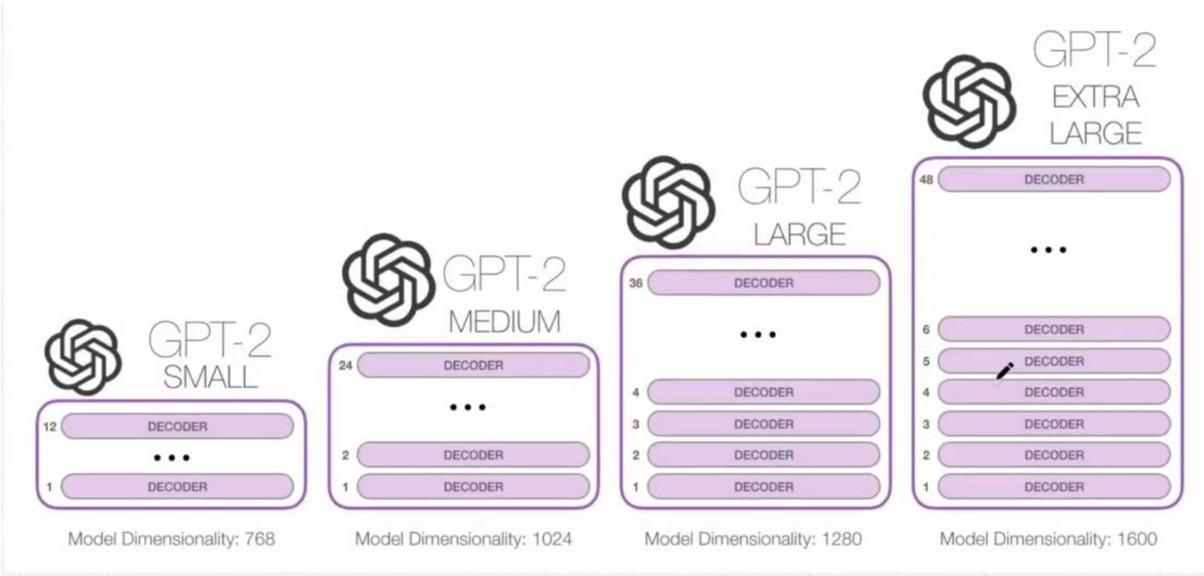


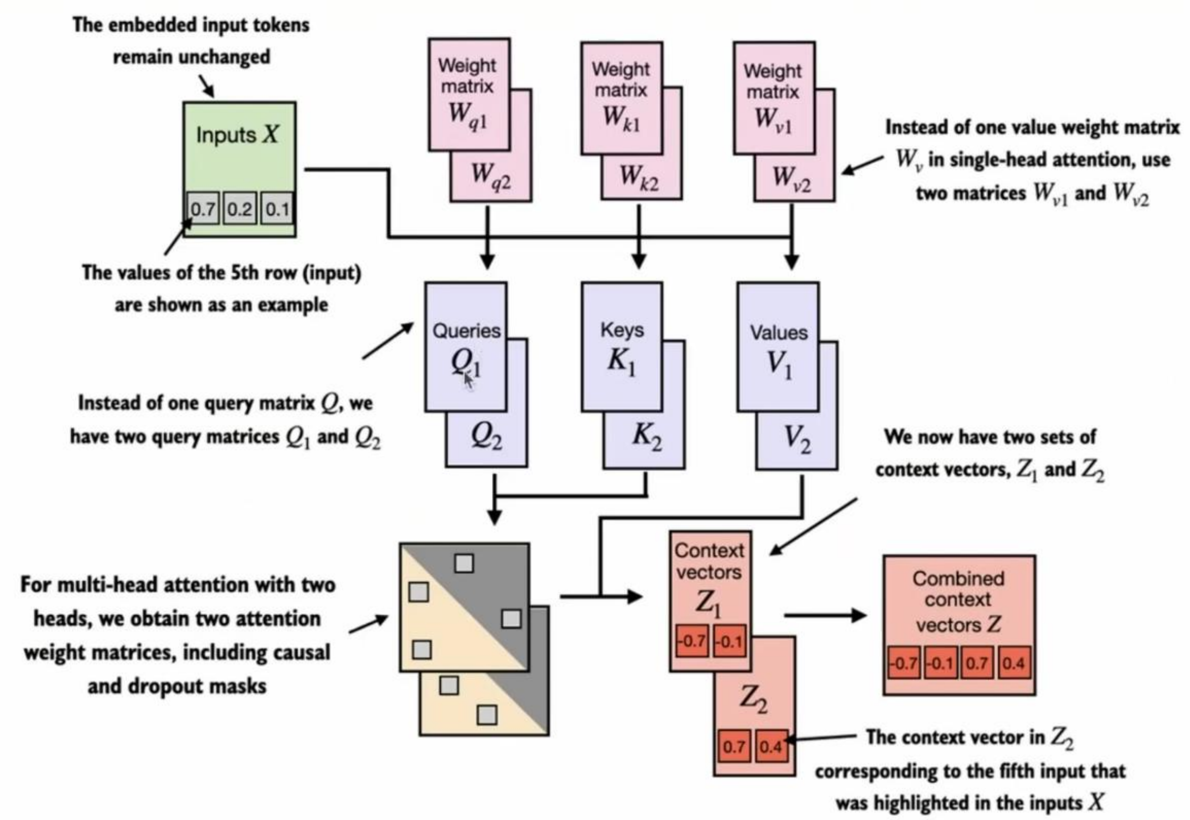
Ervery token in the vocabulary will we have projected in vector space(here we use 768) and the embedding shoud be such that the so meaning should be captured

Number of heads: Number of attention head

Number of layers: number of transformer blocks

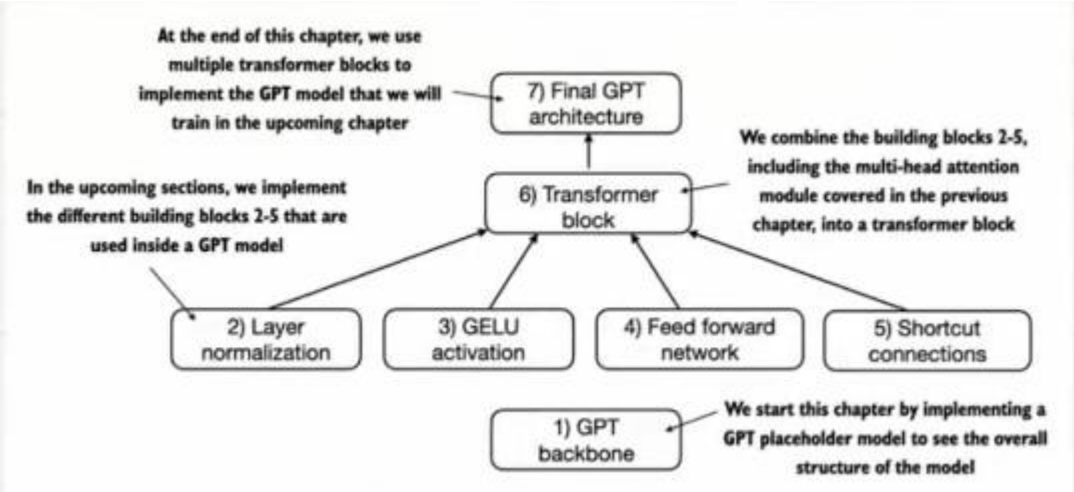




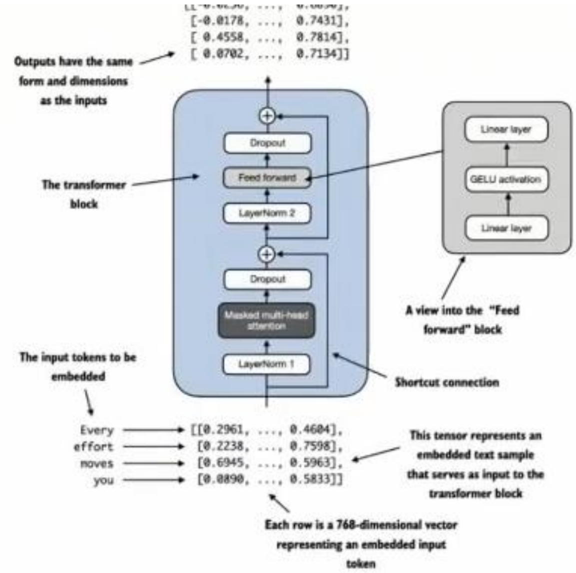


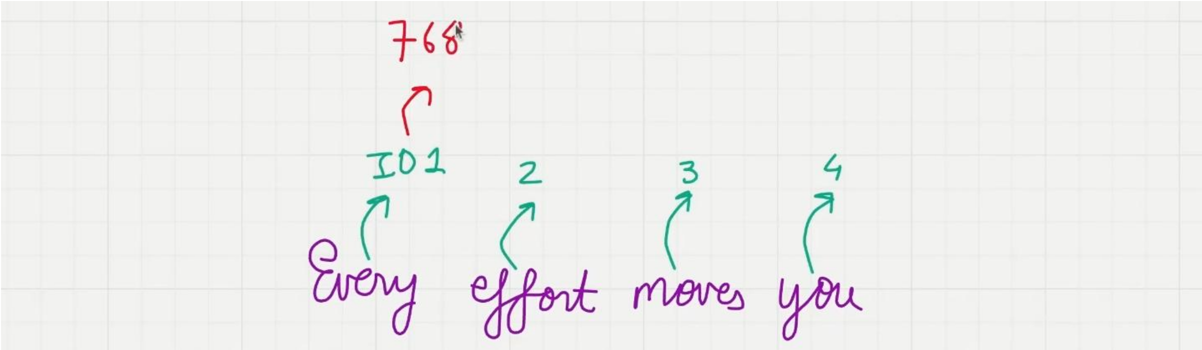
Using this configuration, we will build a GPT placeholder architecture (Dummy GPT model)

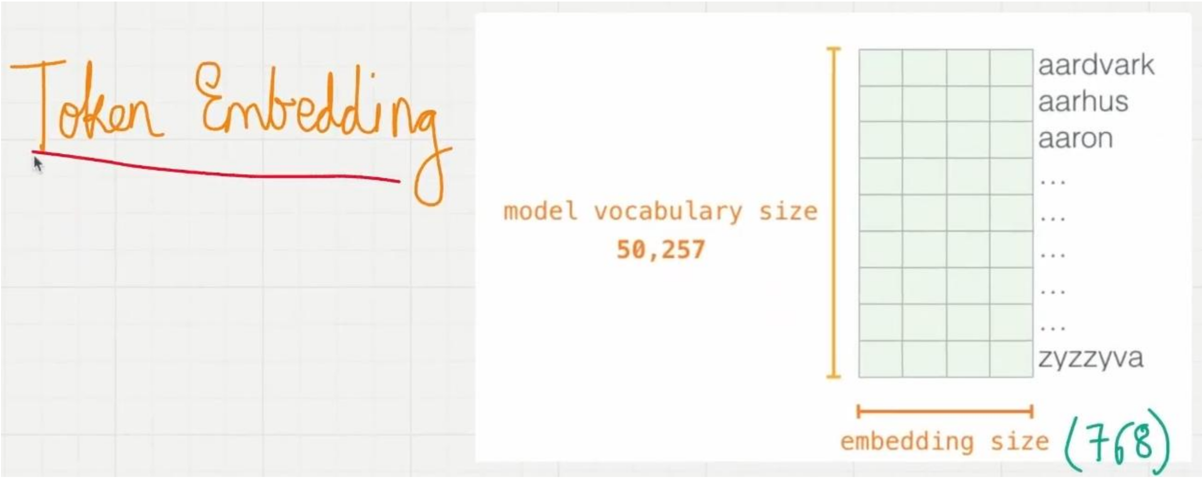
→ This will give us a birds eye view of how everything fits together.

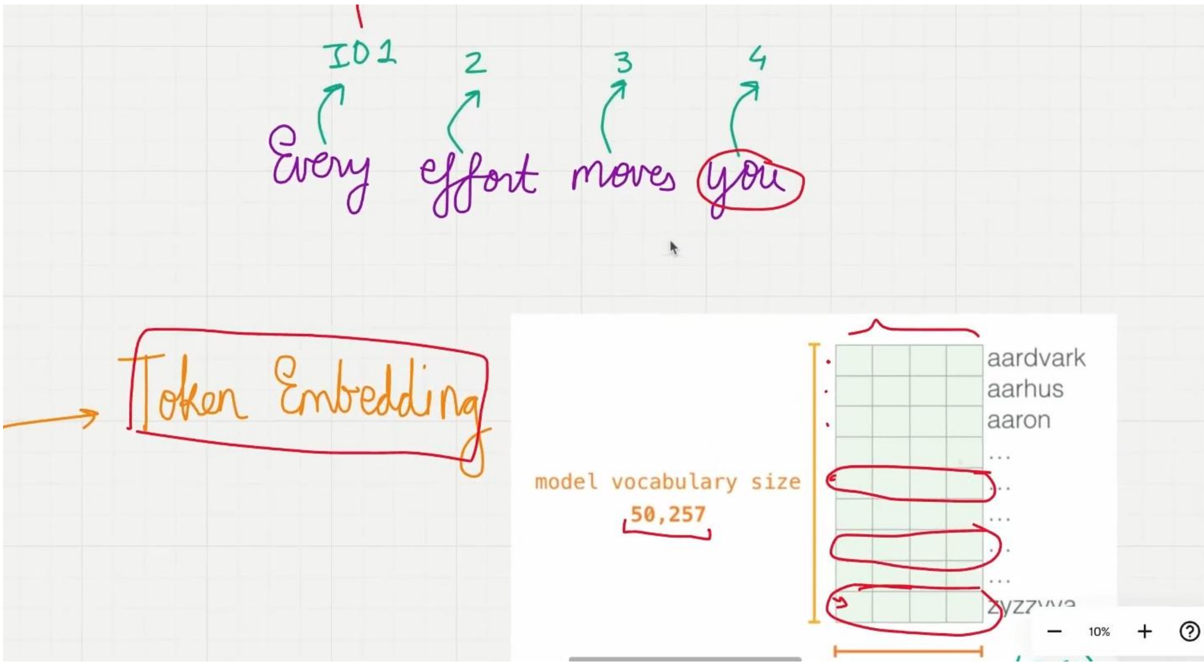


→ Mental model in which we will code the GPT architecture

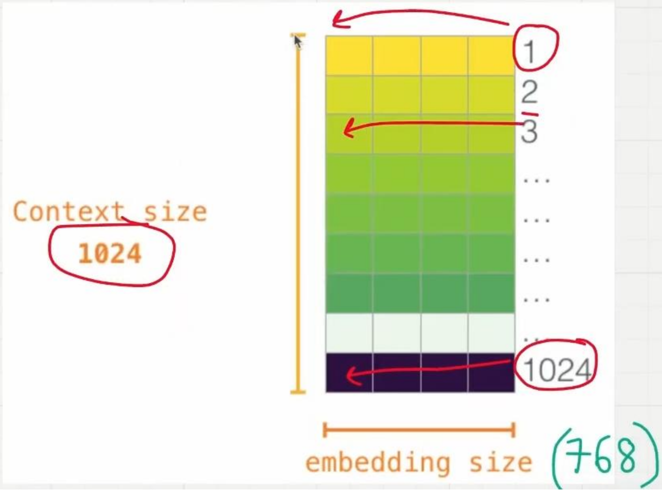



768
101
2
3
4
Every effort moves you


Token Embedding
model vocabulary size
50,257
aardvark
aarhus
aaron
...
zyzzyva
embedding size (768)


101
2
3
4
Every effort moves you
Token Embedding
model vocabulary size
50,257
aardvark
aarhus
aaron
...

Positional Embedding



Output logits for 1 text sample:

Token 1

Token 2

Token 3

Token 4

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

Vocabulary size (50257)