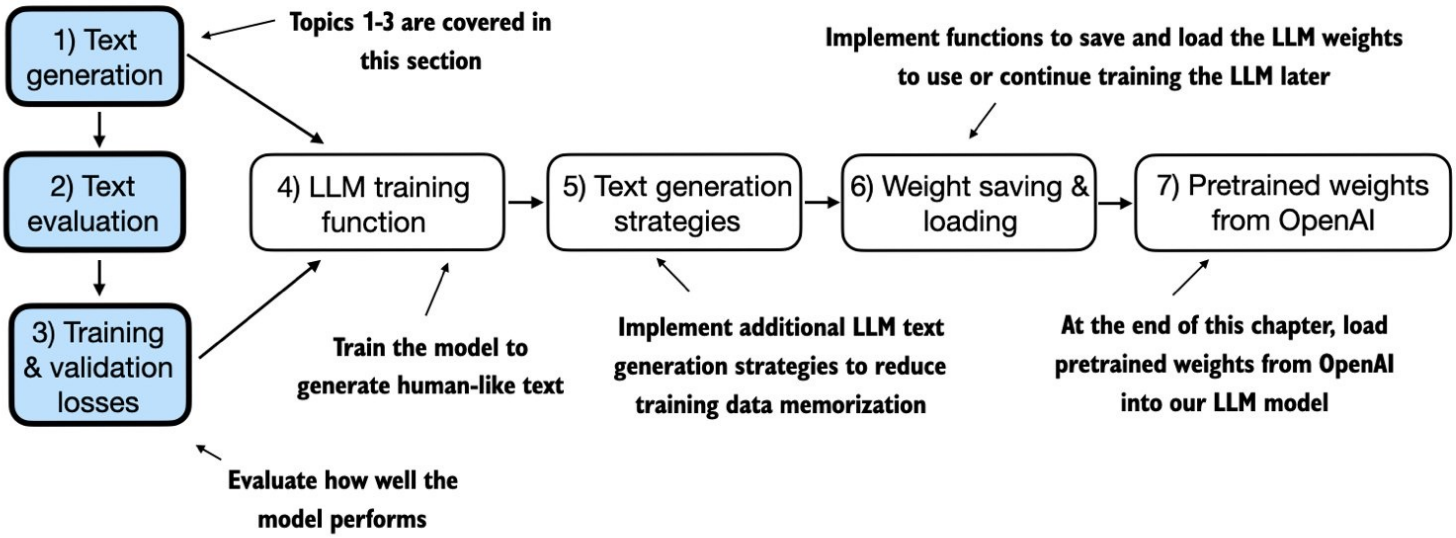
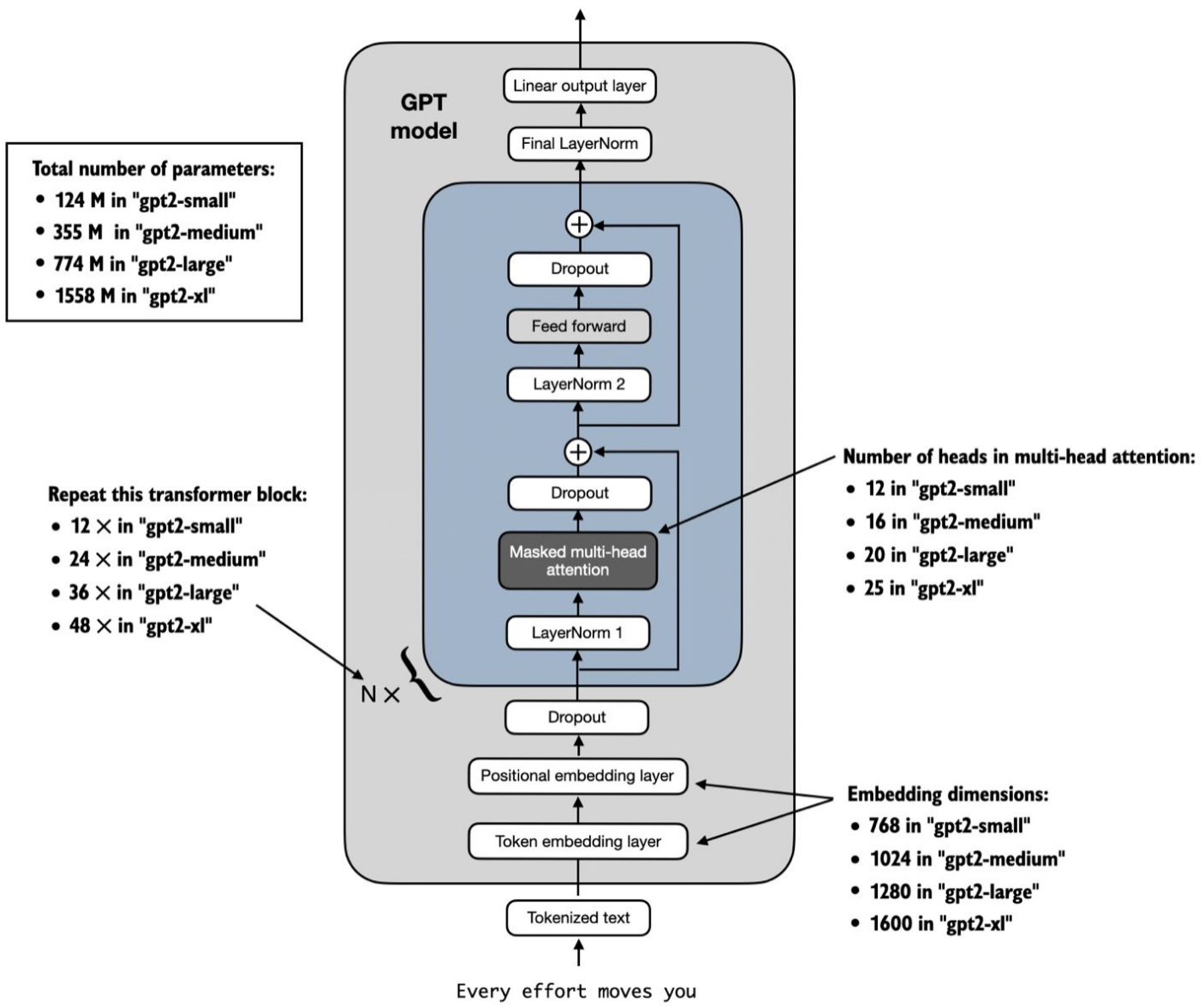
# Entire 124 m paramter GPT-2 model



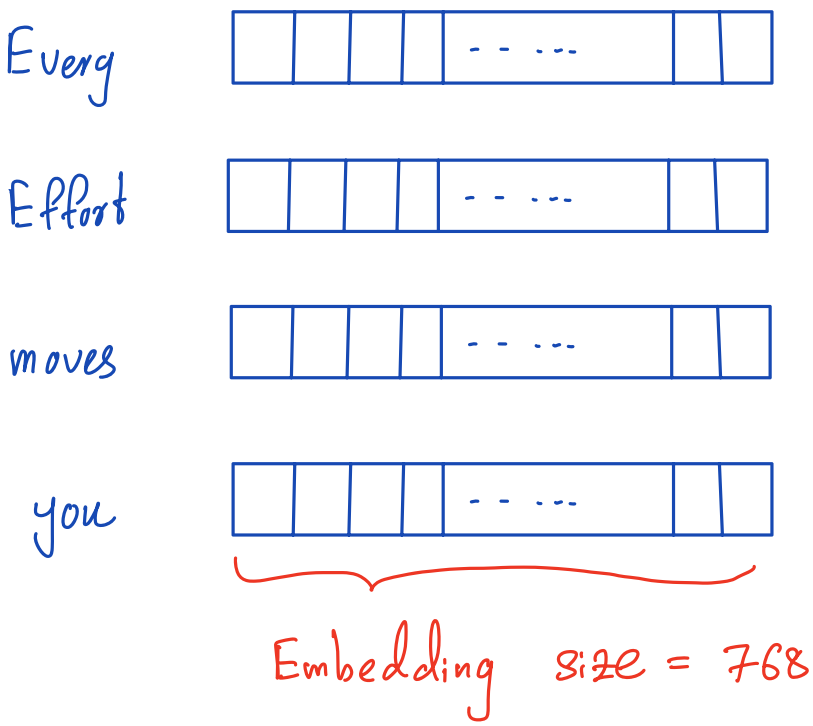



Every Effort moves you

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

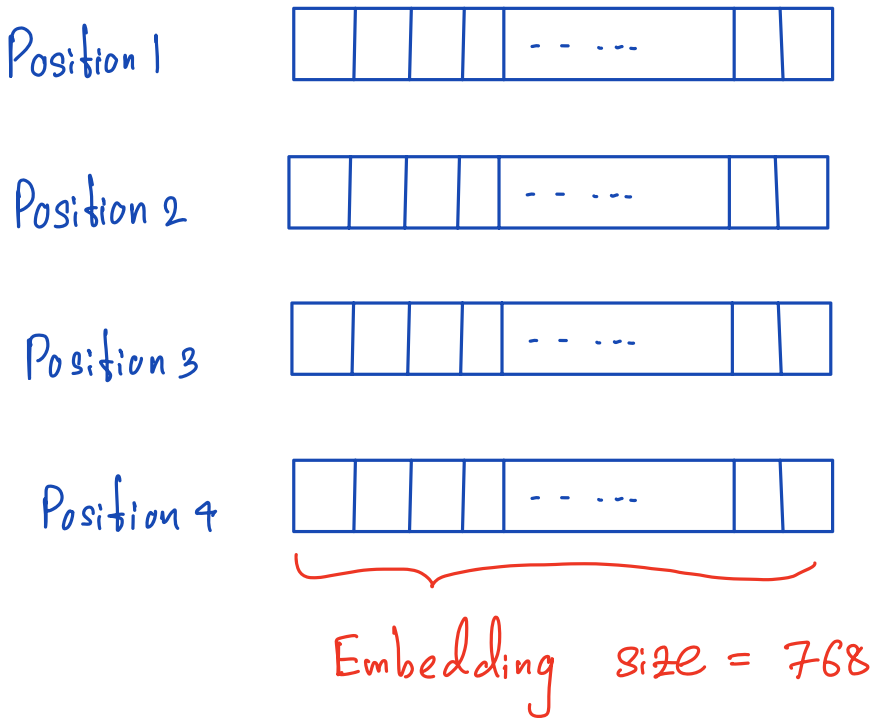
## ① Token Embedding:

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

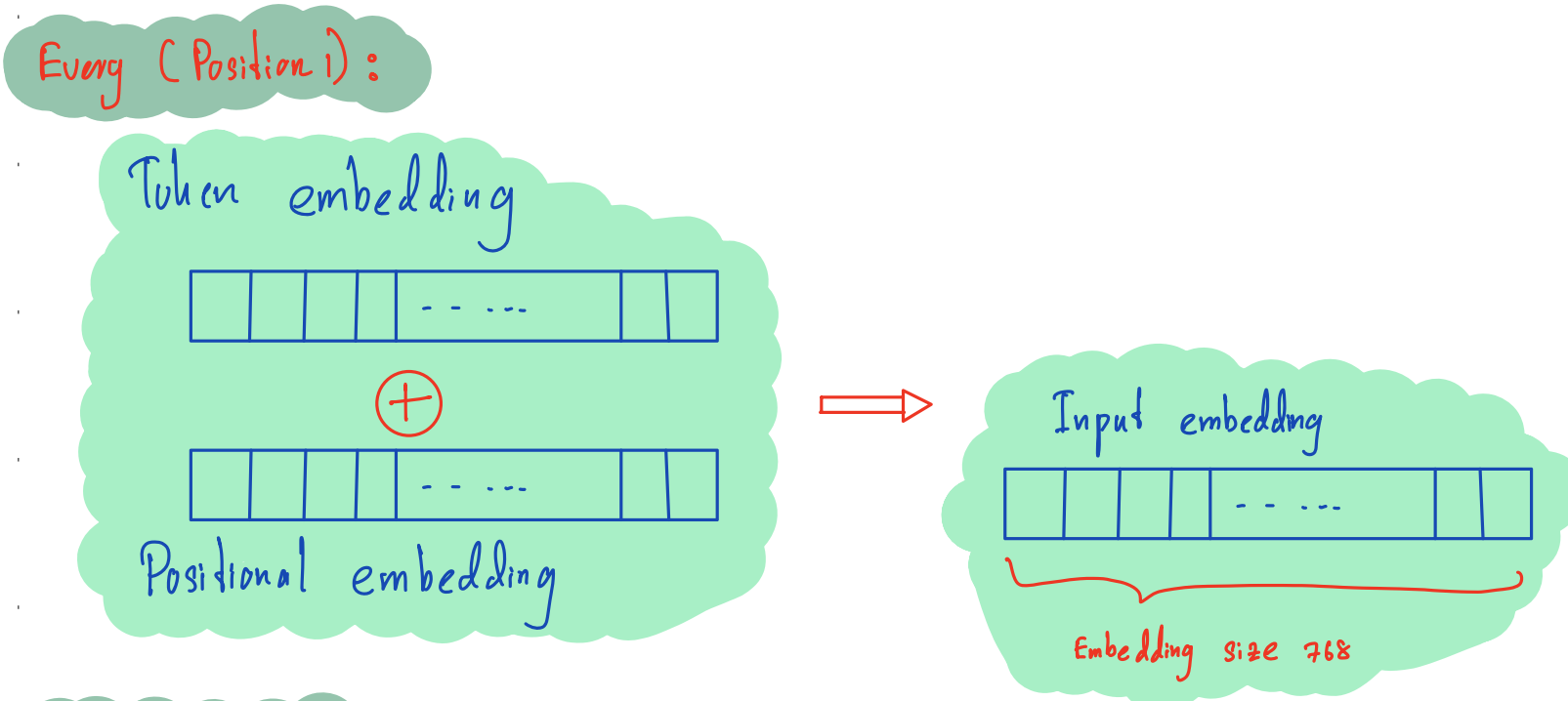


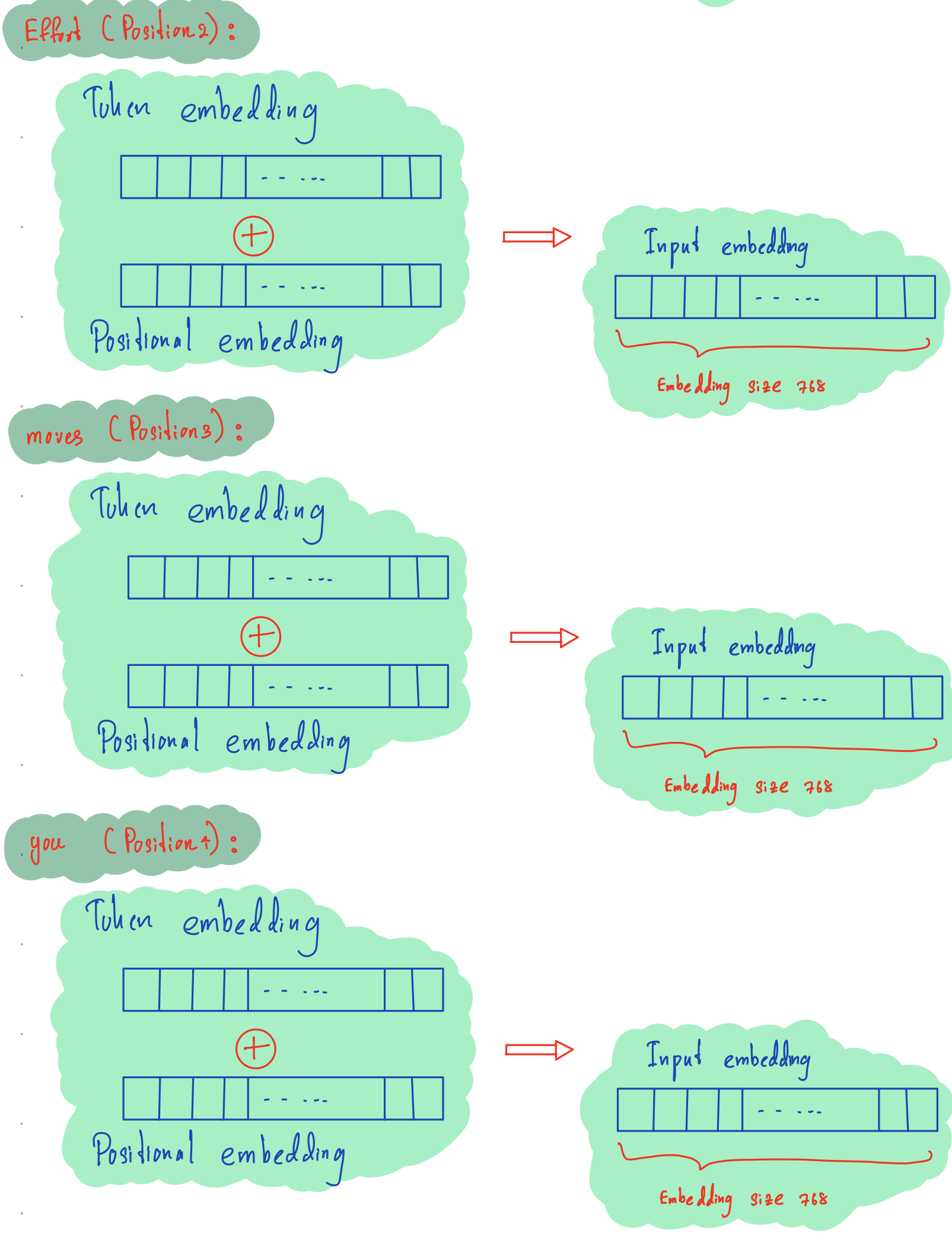
## ② Positional Embedding

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

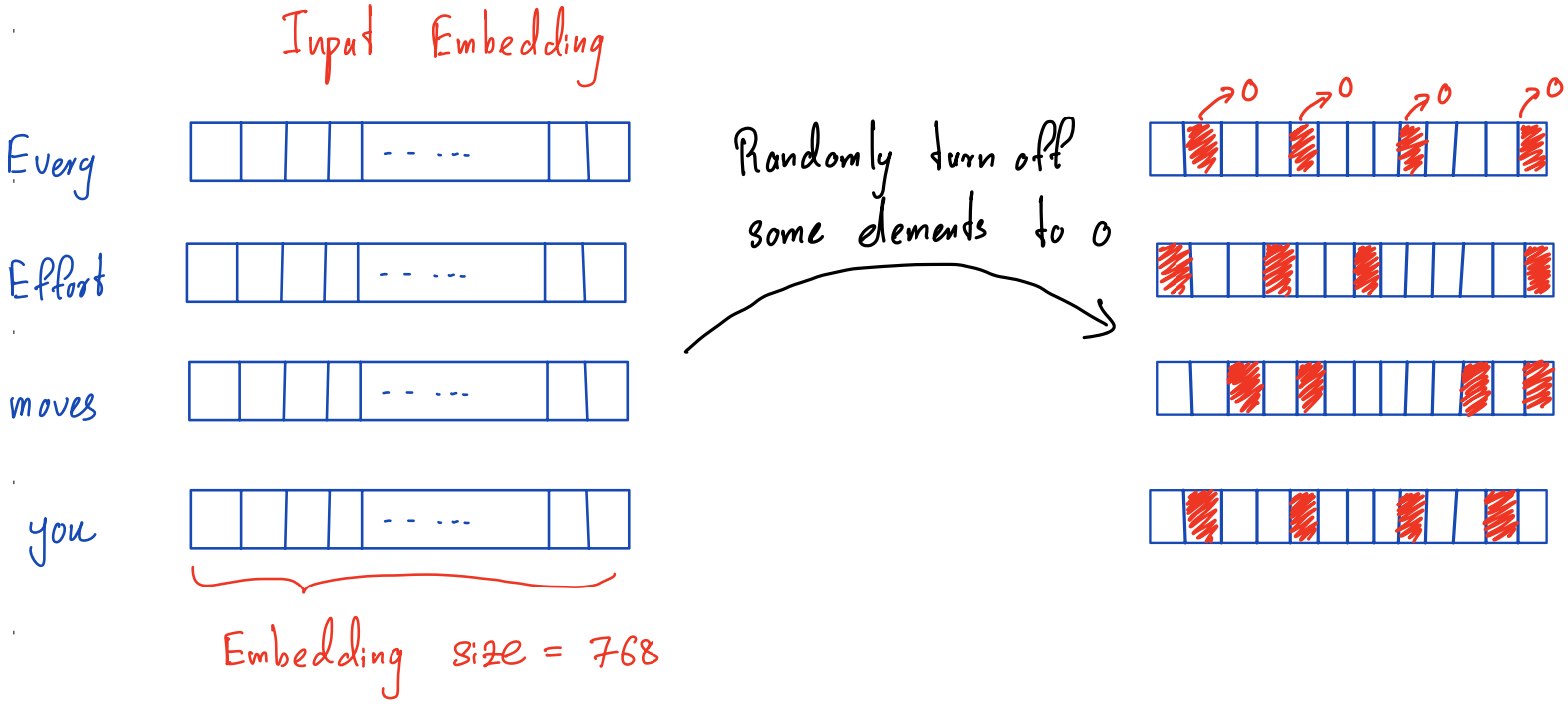



## ③ Input embeding = Token Embeding + Positional Embeding

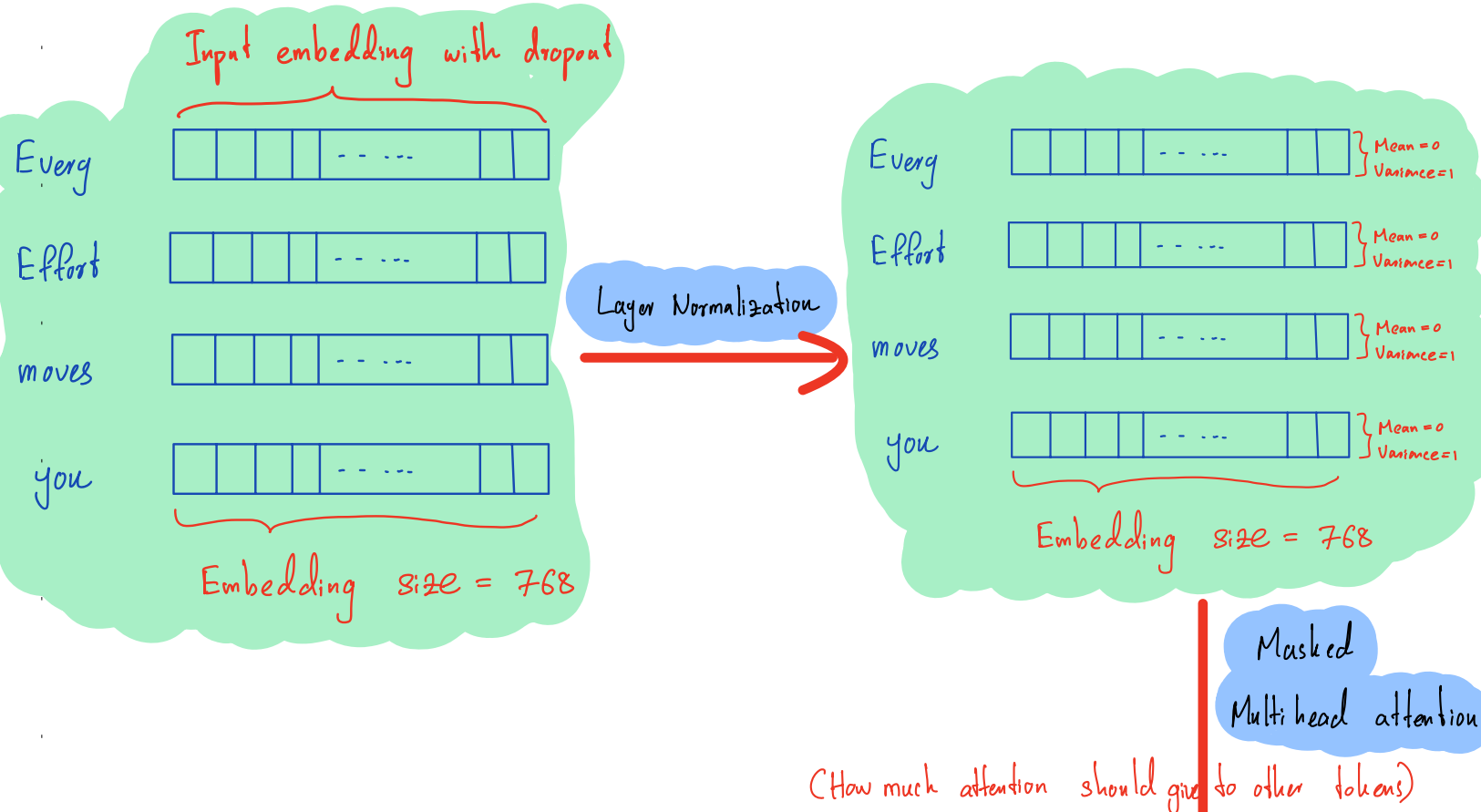



Effort (Position 2):
Token embedding
Positional embedding
Input embedding
Embedding size 768
moves (Position 3):
Token embedding
Positional embedding
Input embedding
Embedding size 768
you (Position 4):
Token embedding
Positional embedding
Input embedding
Embedding size 768

## ④ Dropout : Prevent overfiting or Improve Generalization




## ⑤ Transformer :



Every

Effort

moves

you

Embedding size = 768

Dropout

Every

Effort

moves

you

Context Vector

Context Vector

Context Vector

Context Vector

Embedding size = 768

Shortcut Connection

Every

Effort

moves

you

Embedding size = 768

Layer Norm

Every

Effort

moves

you

Mean = 0
Variance = 1

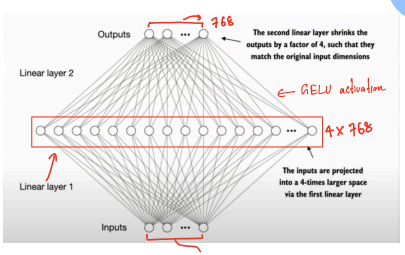
Mean = 0
Variance = 1

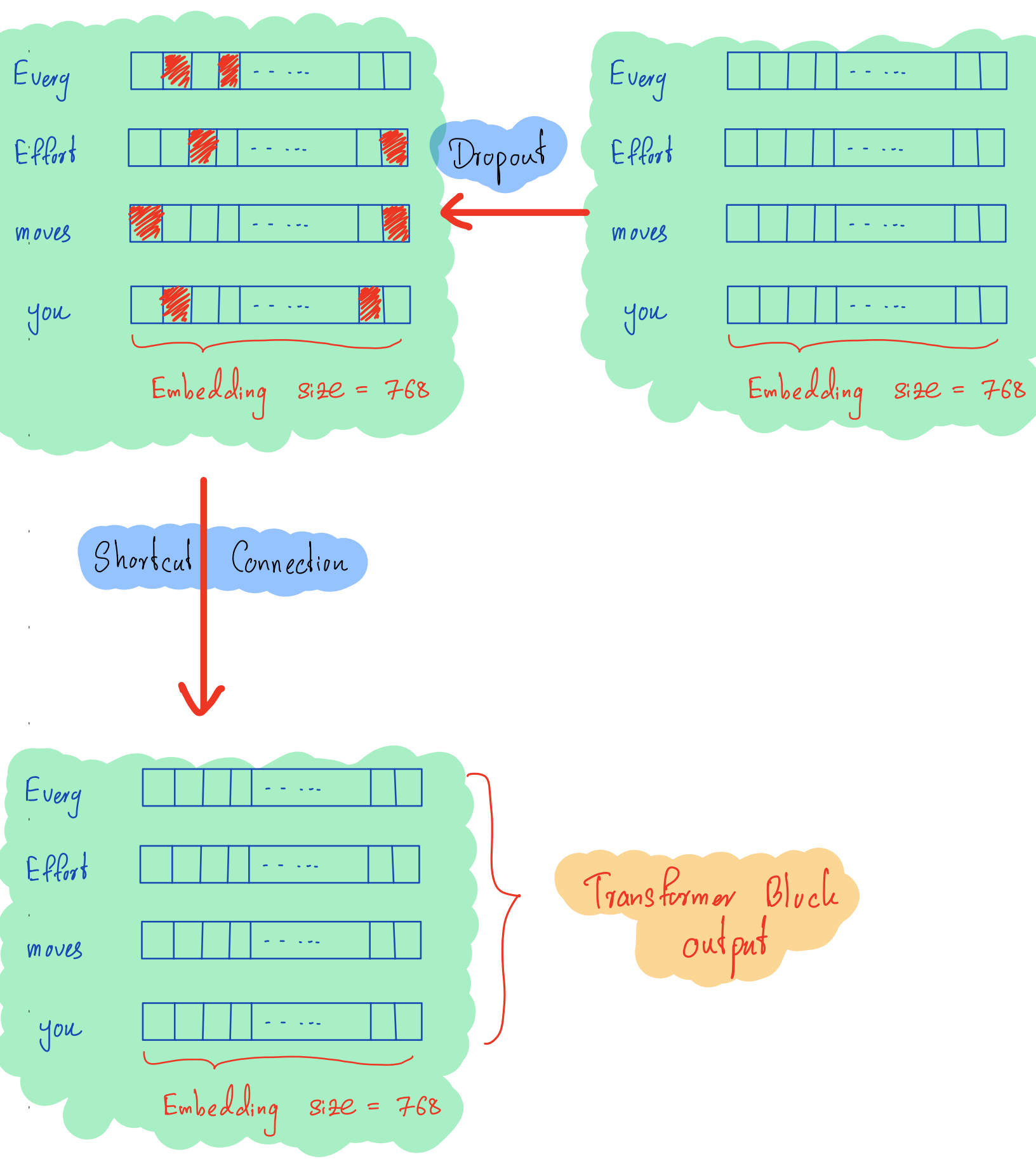
Mean = 0
Variance = 1

Mean = 0
Variance = 1

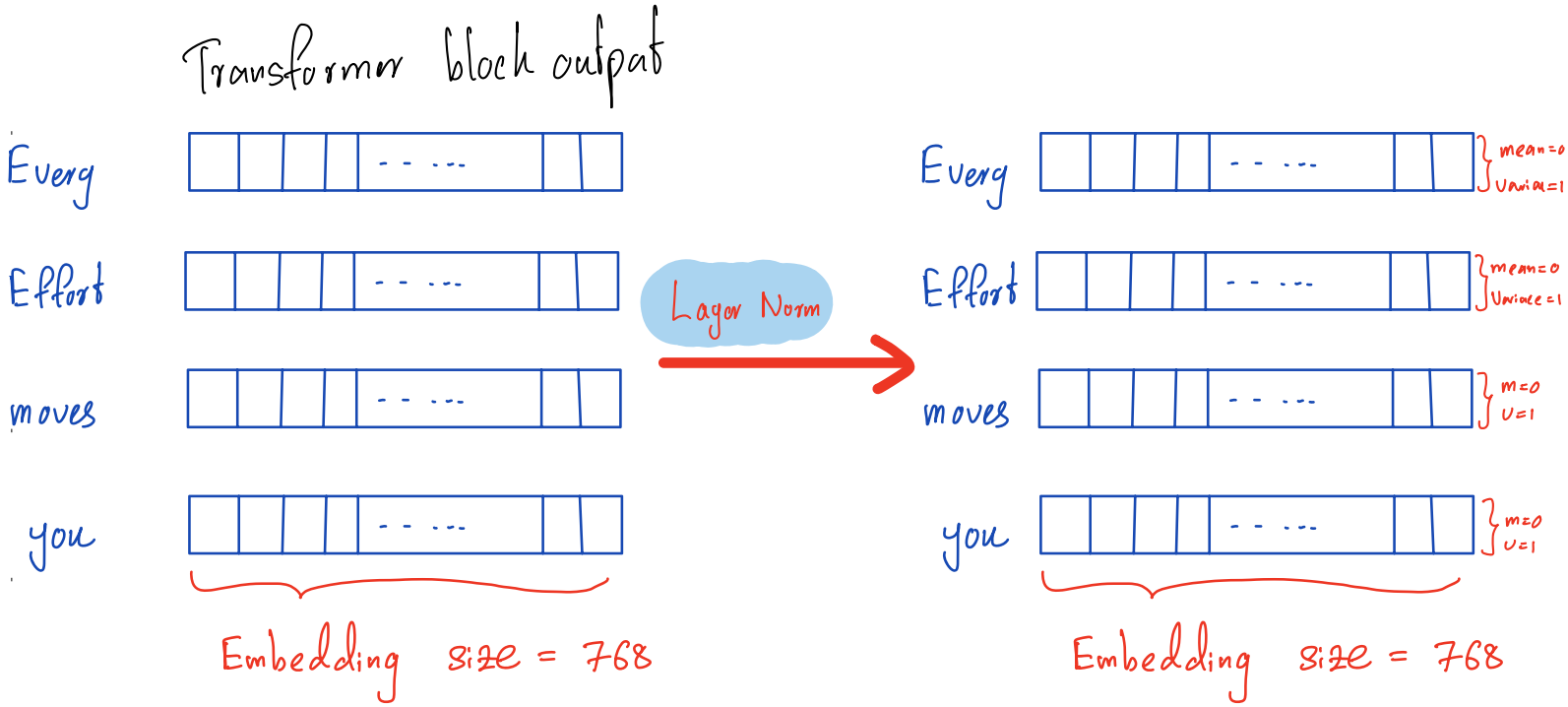
Embedding size = 768

Feed forward Neural Network

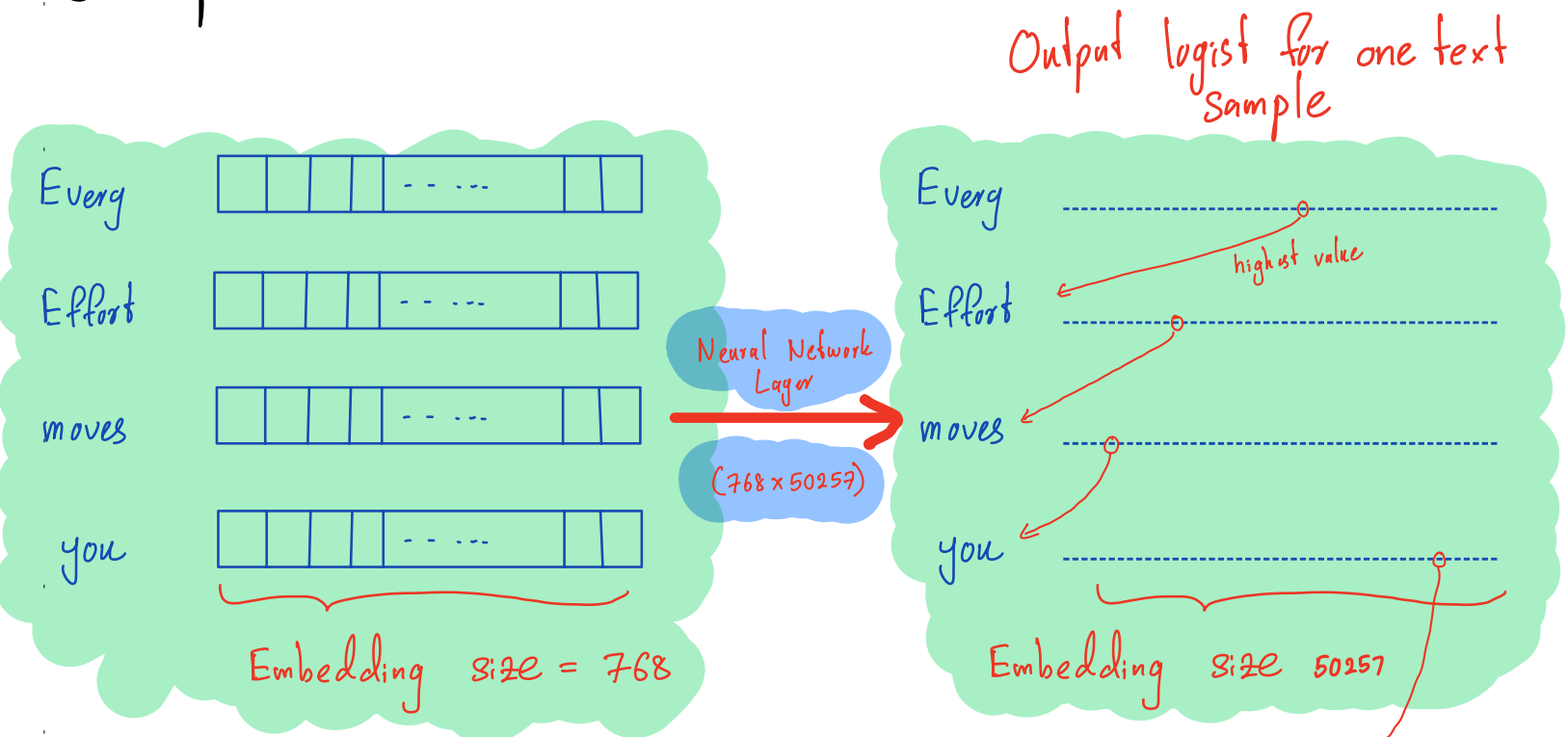



Every
Effort
moves
you
Embedding size = 768
Dropout
Every
Effort
moves
you
Embedding size = 768
Shortcut Connection
Every
Effort
moves
you
Embedding size = 768
Transformer Block output

## ⑥ Layer Normalization (after the transformer output)




## ⑦ Output Head



Output tensor value → 2 × 4 × 50257