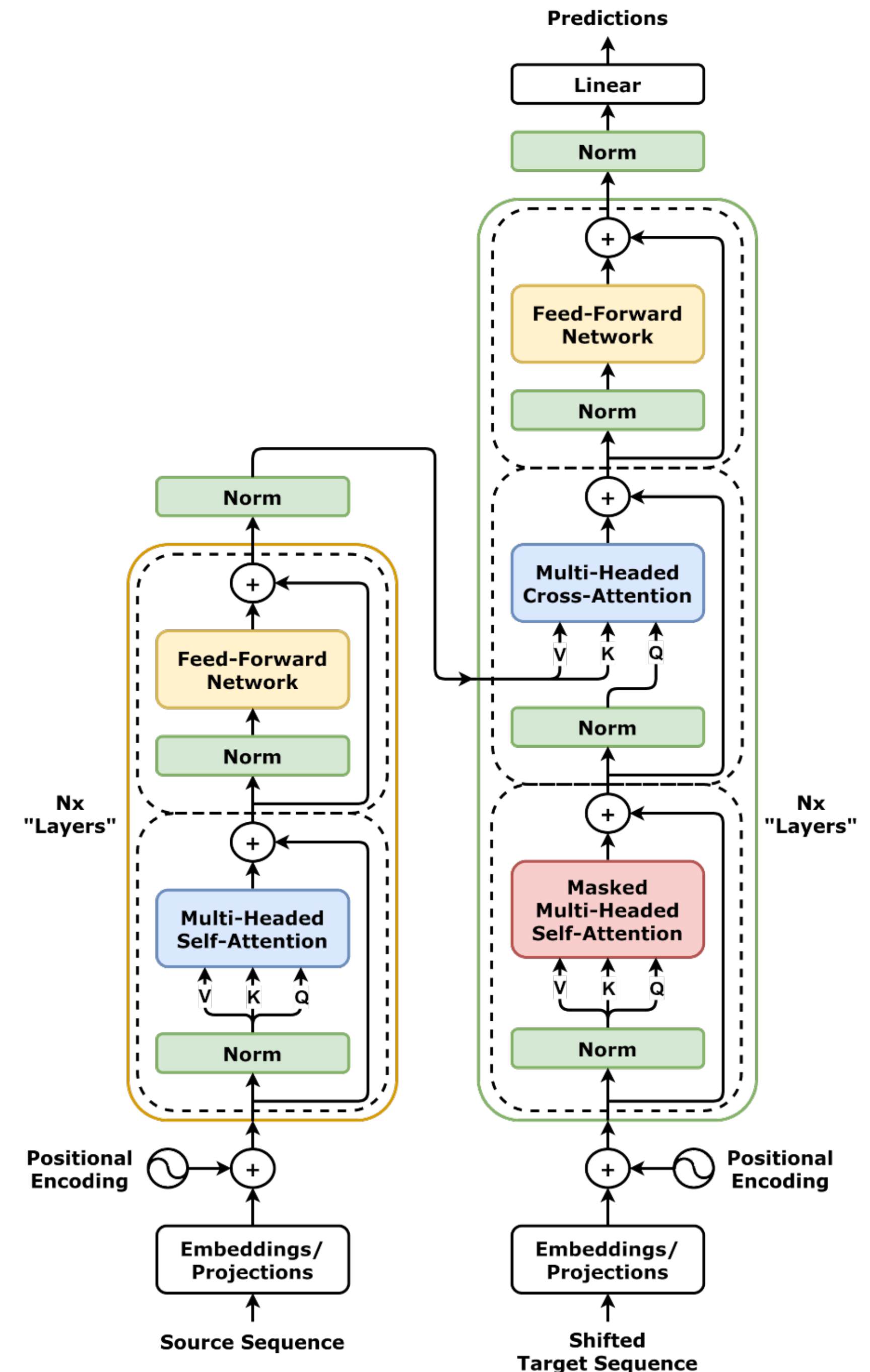


Transformer Architecture

- “Attention Is All You Need”
- Published by researchers at Google in 2017
- More parameters = more computing power needed = better results



Tokenization Example

This is some text that has been divided into tokens using OpenAI's tokenization algorithm for GPT-4

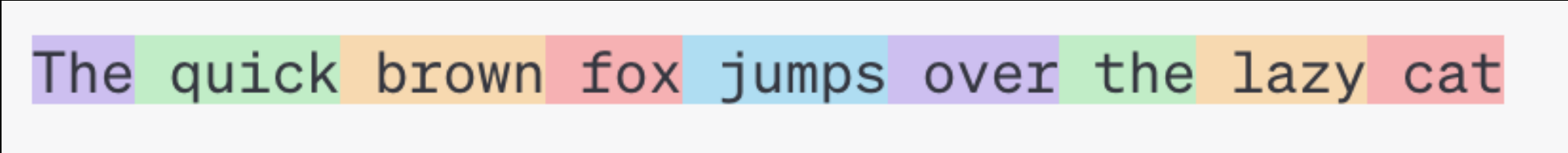
- [2028, 374, 1063, 1495, 430, 706, 1027, 18255, 1139, 11460, 1701, 5377, 15836, 596, 4037, 2065, 12384, 369, 480, 2898, 12, 19]
- LLM's are trained to output the next token number. A simple lookup table converts the number into a piece of text

Training Process

- Get sequence of text from training data, split into input and output tokens

• Input:  Output: 

- Have model make prediction based on input sequence



- Compare model output to actual output (“ cat” vs. “ dog”) to get error value
- Back propagation: run the model in reverse, adjusting weights based on error value
- Repeat many many times

Generating Output

