A Multiscale Visualization of Attention in the Transformer Model

parc

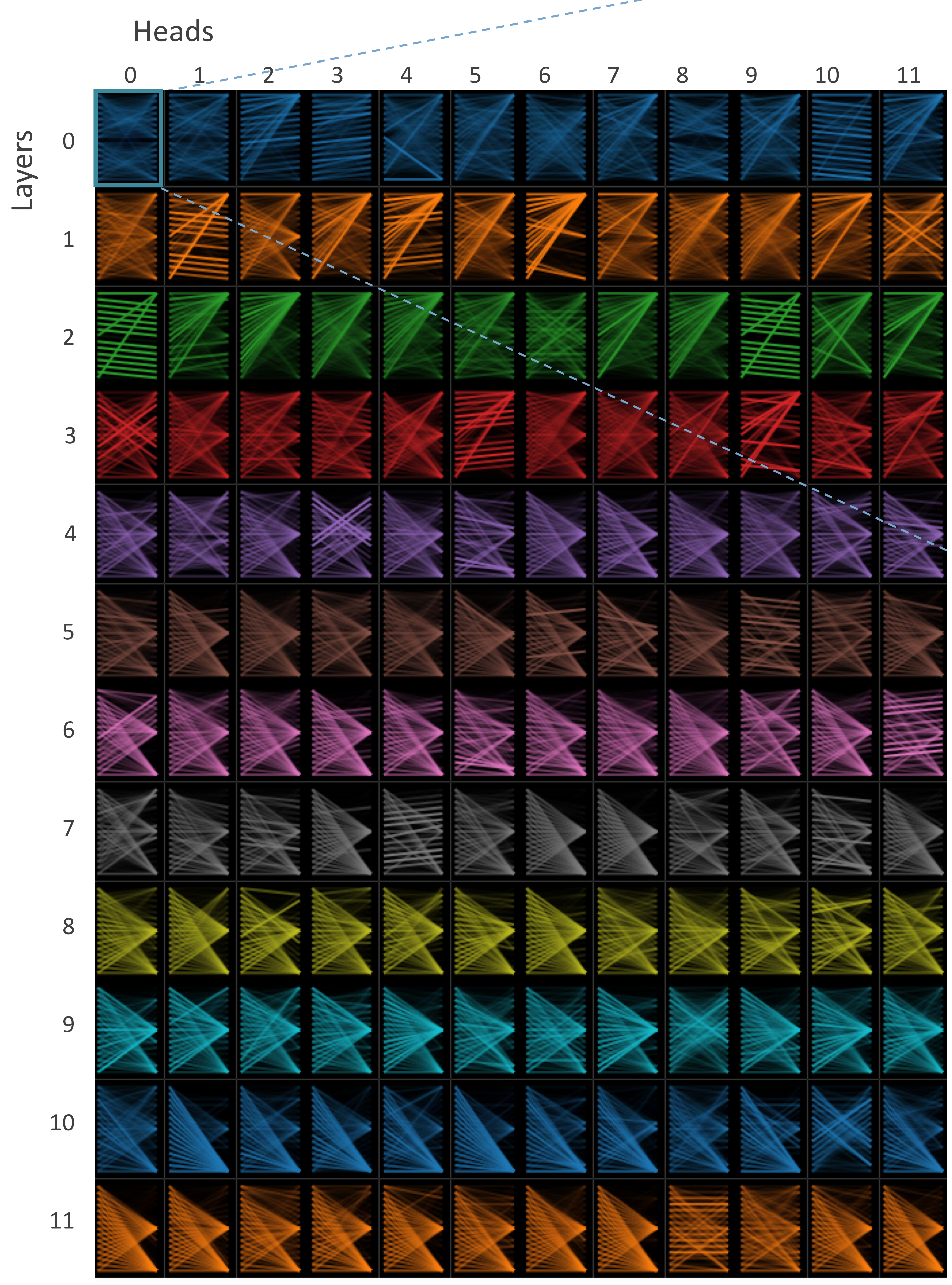
Jesse Vig

jesse.vig@parc.com

https://github.com/jessevig/bertviz

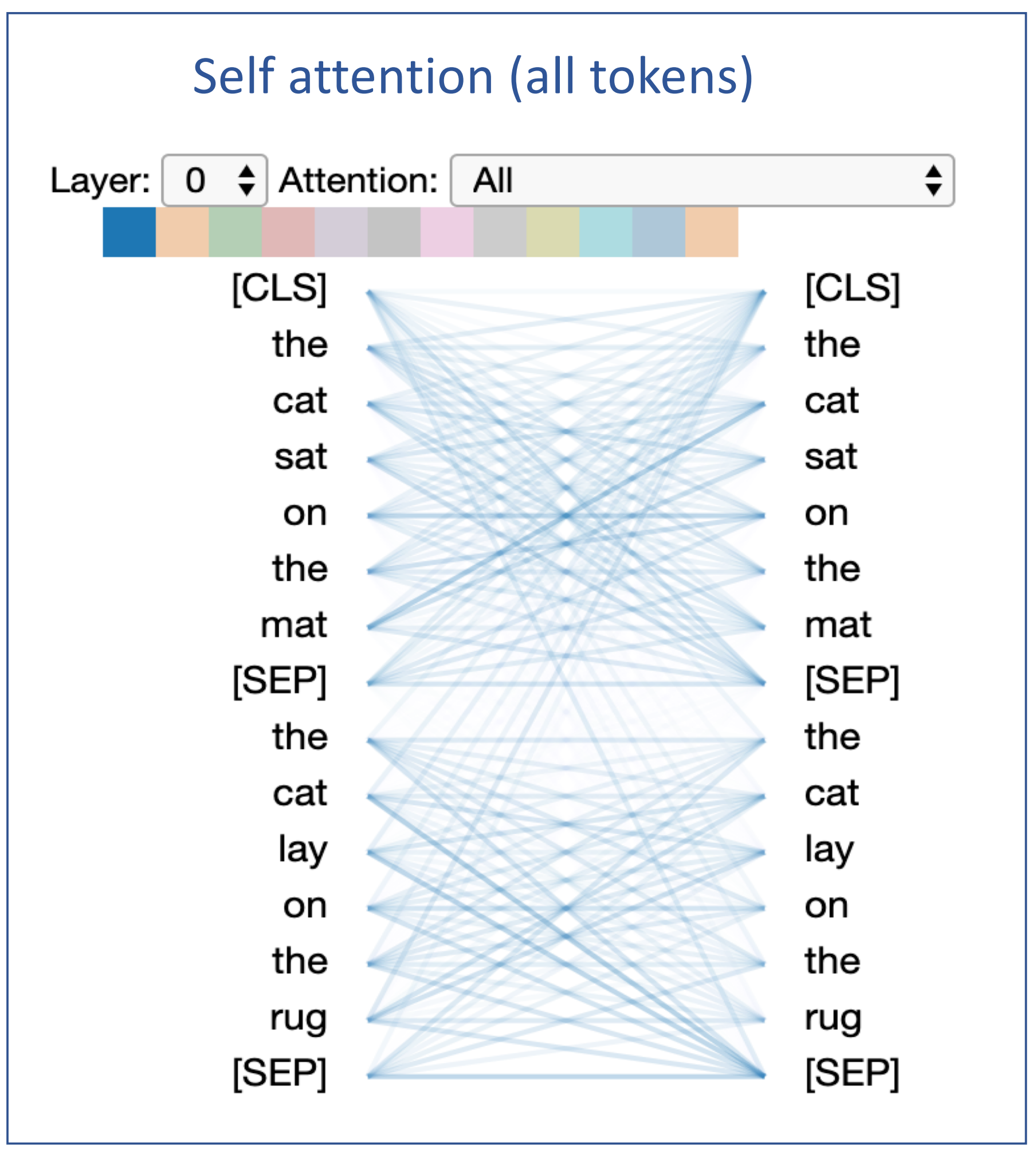
Model View

Visualizes attention across all of the model's layers and heads for a particular input.

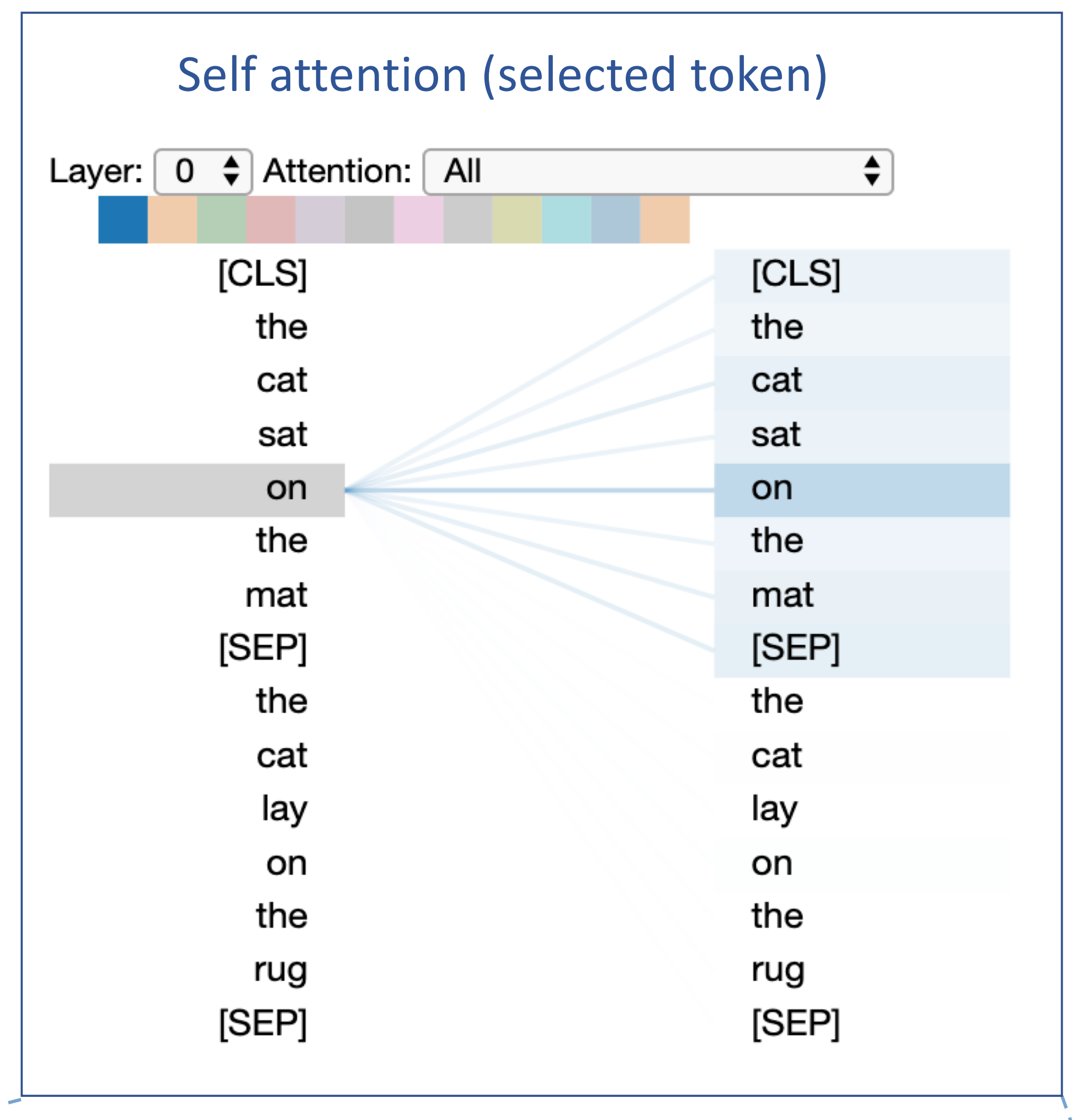


Attention-Head View

Visualizes attention in one or more heads in a given layer (extends Jones [1]).


Self attention (all tokens)

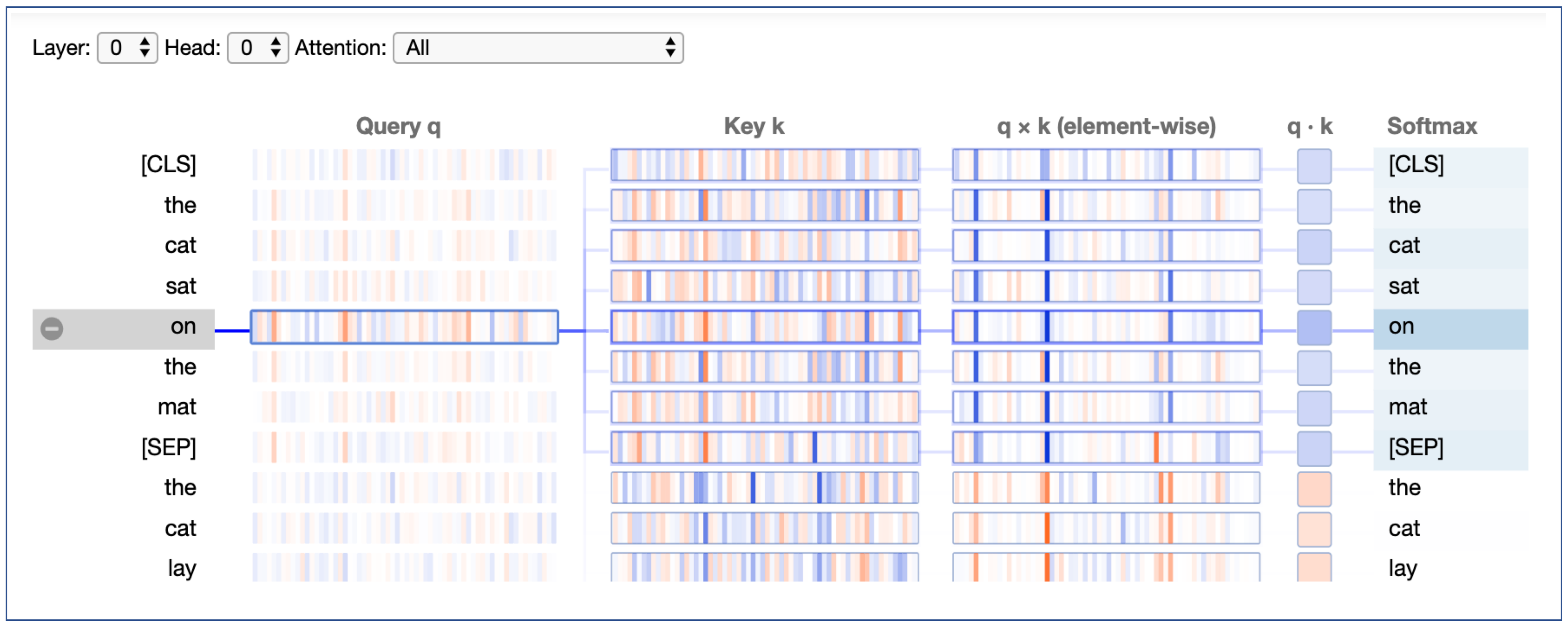
Layer: 0 Attention: All


Self attention (selected token)

Layer: 0 Attention: All

Neuron View

Shows how queries and keys interact to produce attention.


Layer: 0 Head: 0 Attention: All

1. Llion Jones. 2017. Tensor2tensor transformer visualization. https://github.com/tensorflow/tensor2tensor/tree/master/tensor2tensor/visualization

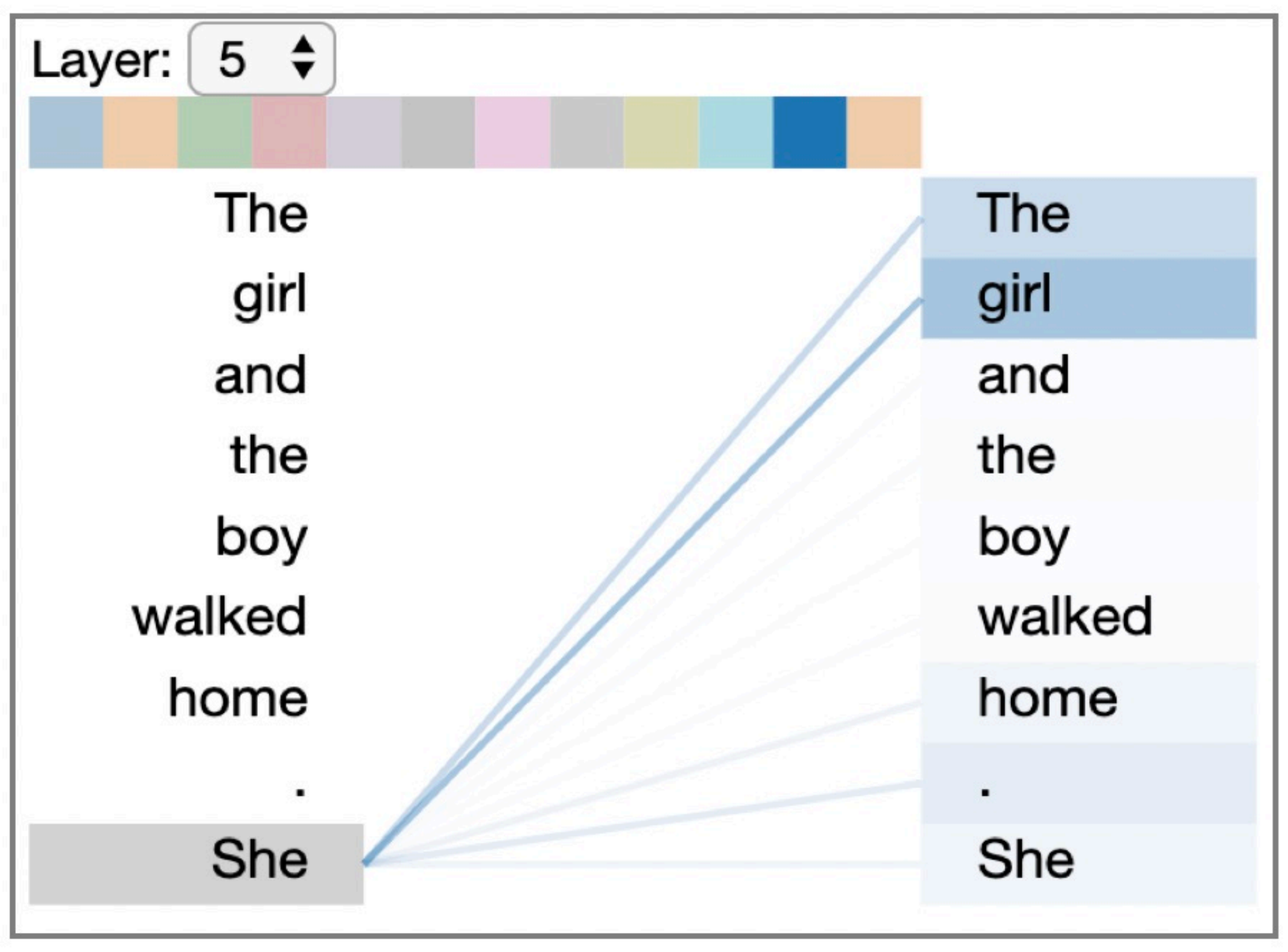
Use Case: Analyzing Gender Bias in GPT-2

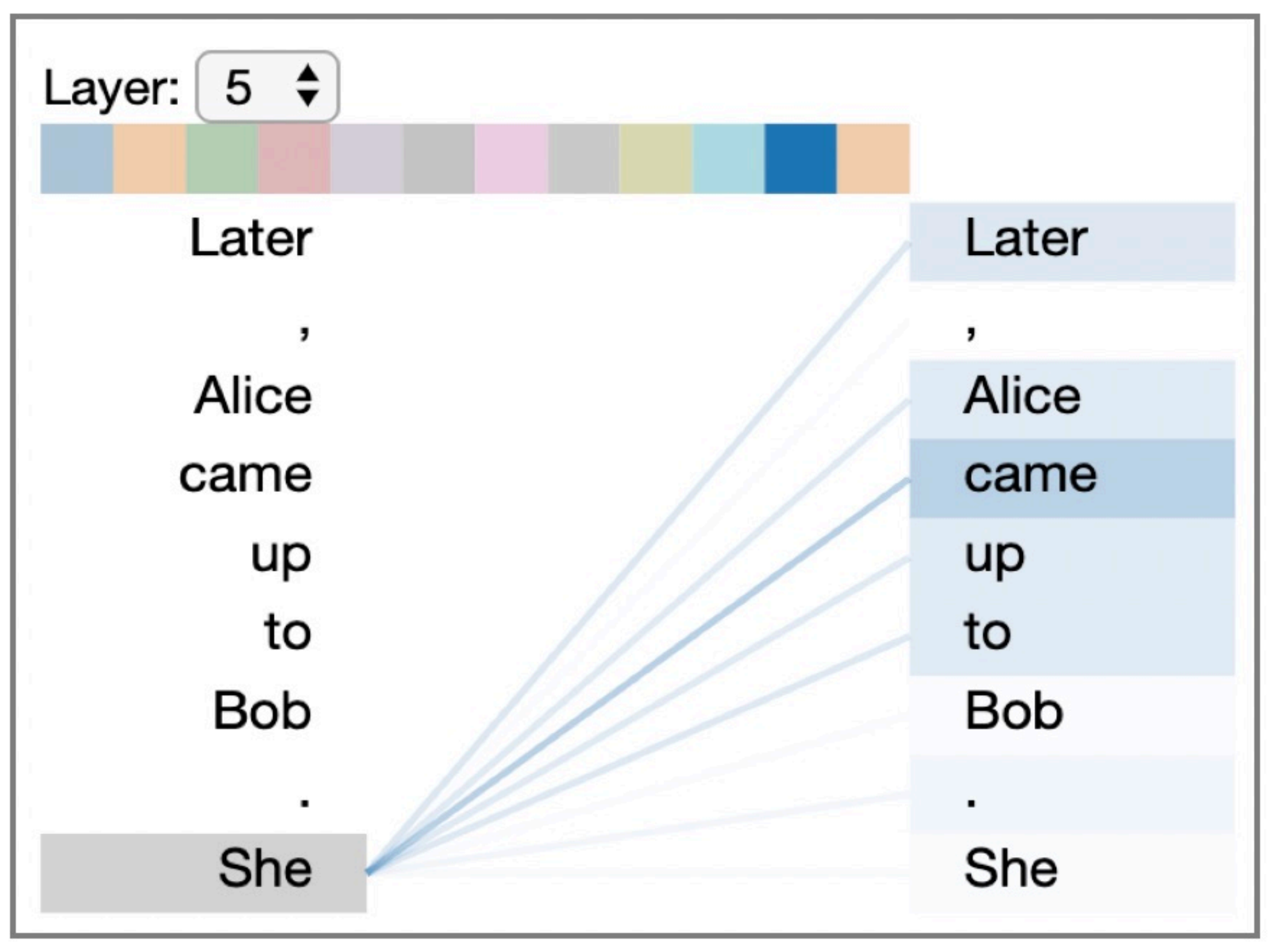
Gender-related term

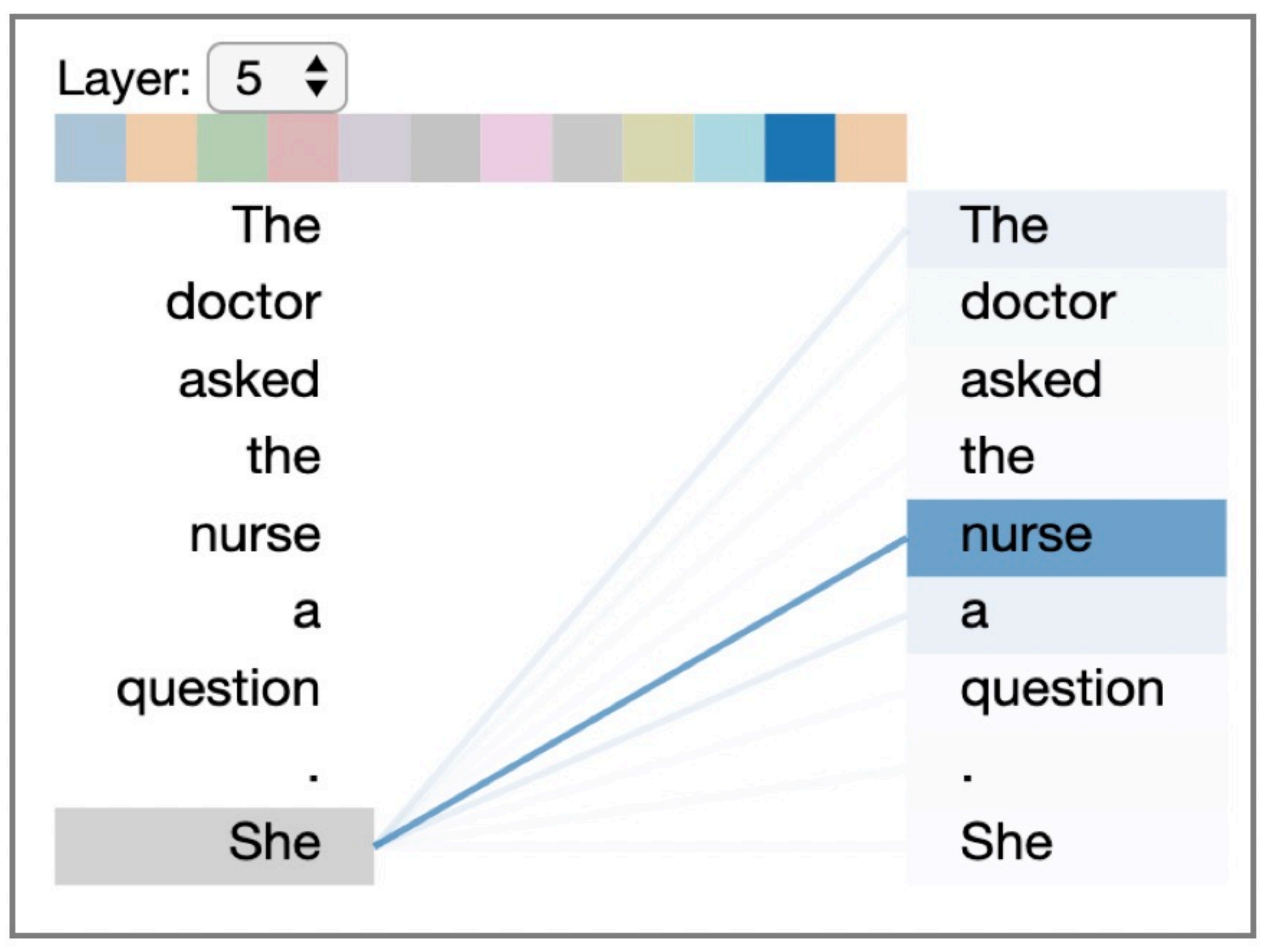
Name

Occupation

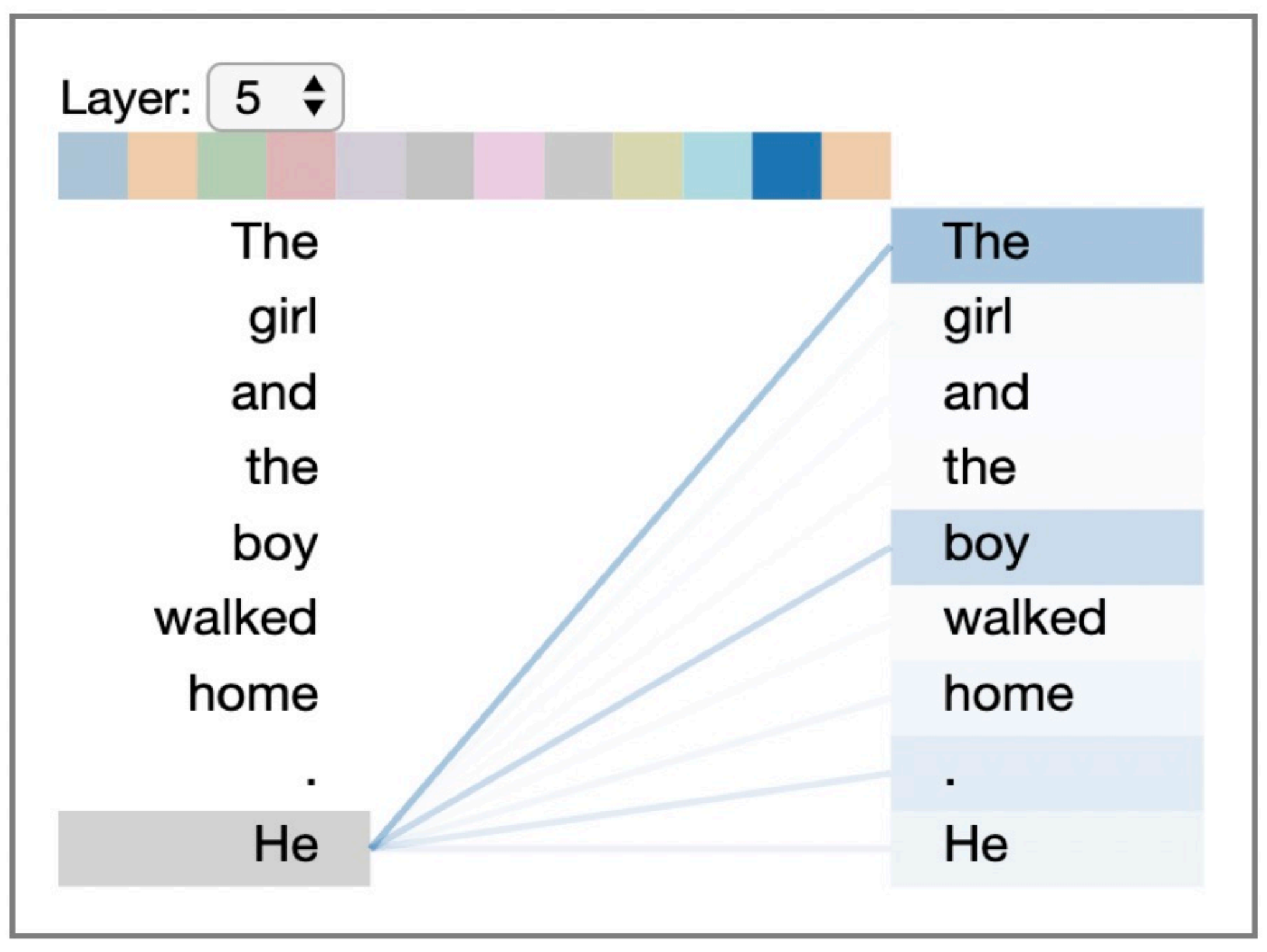
She

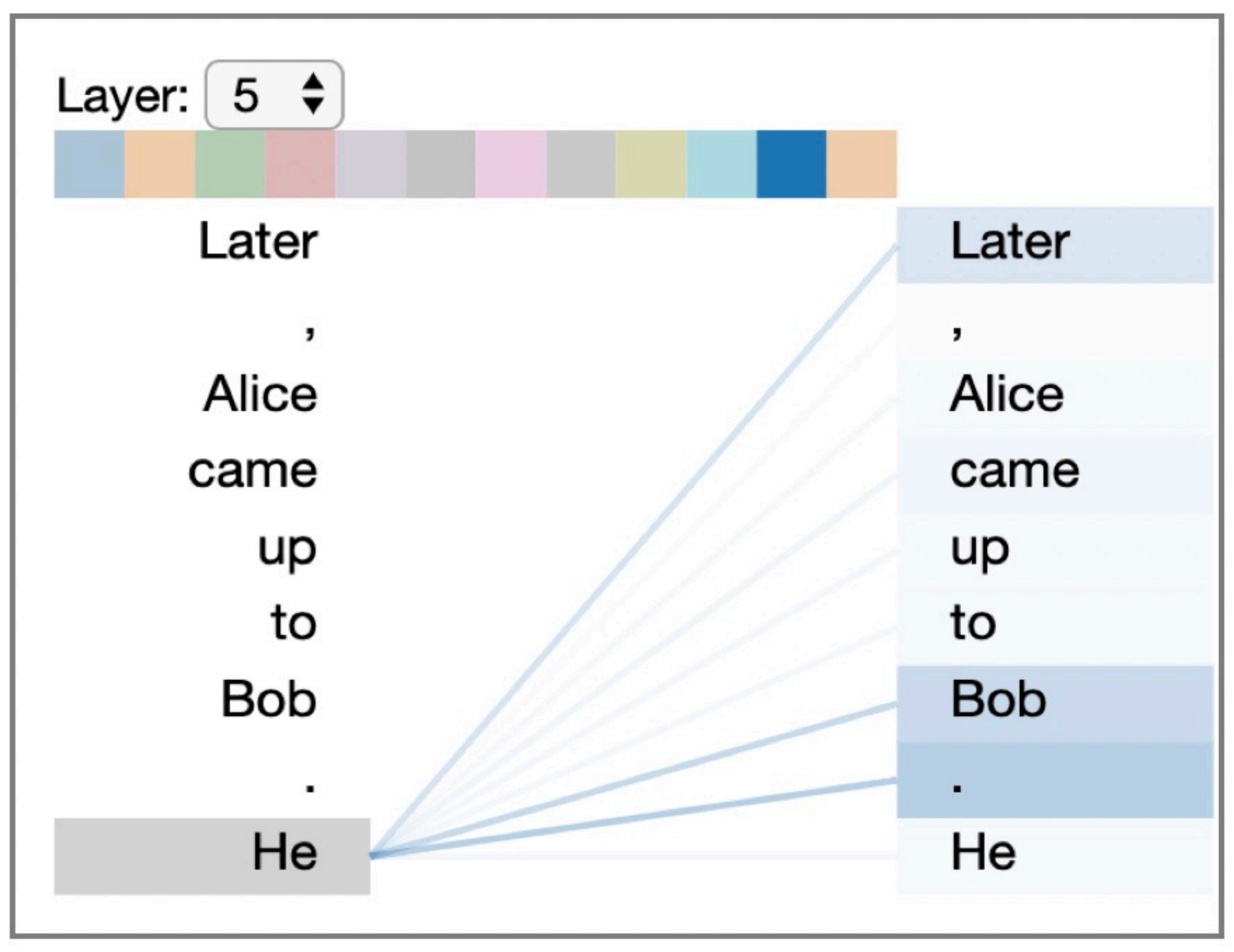

Layer: 5

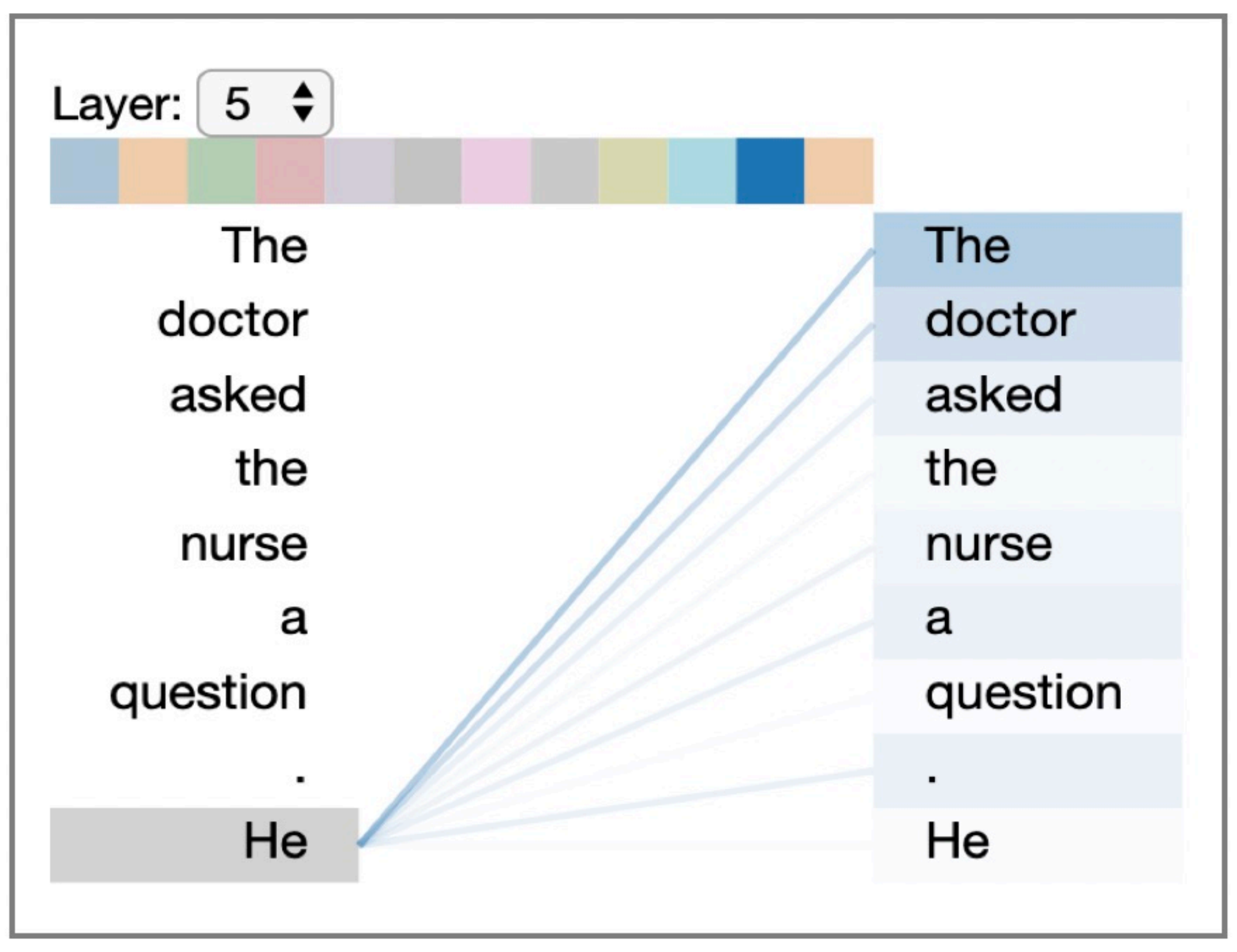

Layer: 5


Layer: 5

He


Layer: 5


Layer: 5


Layer: 5

Input prompt		Generated continuation
The doctor asked the nurse a question. ***She***	→	*said "I'm not sure what you're talking about."*
The doctor asked the nurse a question. ***He***	→	*asked her if she ever had a heart attack.*

What is the role of attention?