Attention Outputs

- The computation can be represented as

- $Q \cdot K^{\top}$ is $n \times n$

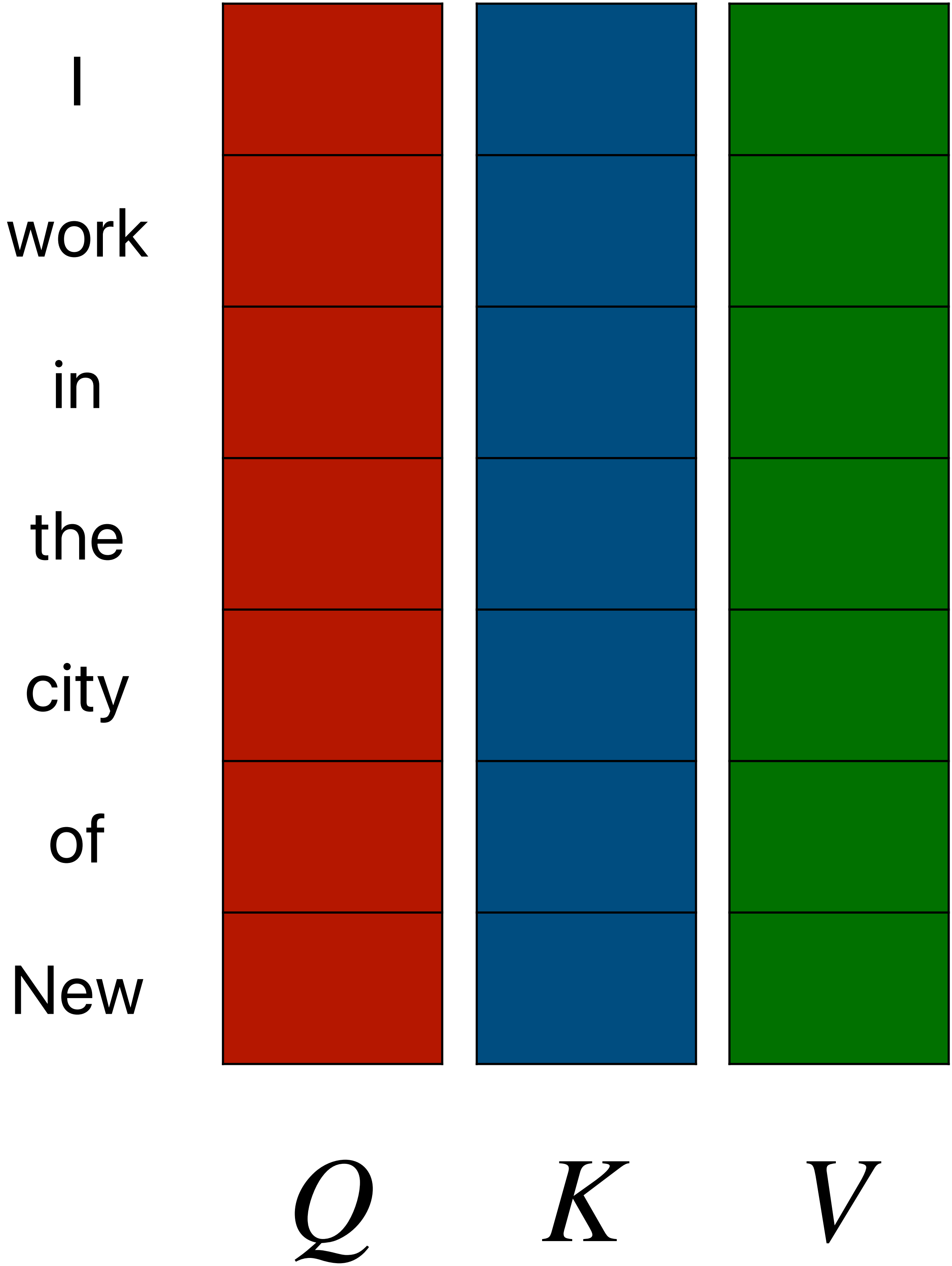
- diagonal of $D = \mathrm{LT}(\exp(Q \cdot K^\top)) \cdot 1_n$

- Naively computing takes $O(n^2)$ time

- Prohibitive when n is large

13

$$D^{-1} \cdot \mathsf{LT}(\exp(Q \cdot K^{\mathsf{T}})) \cdot V$$

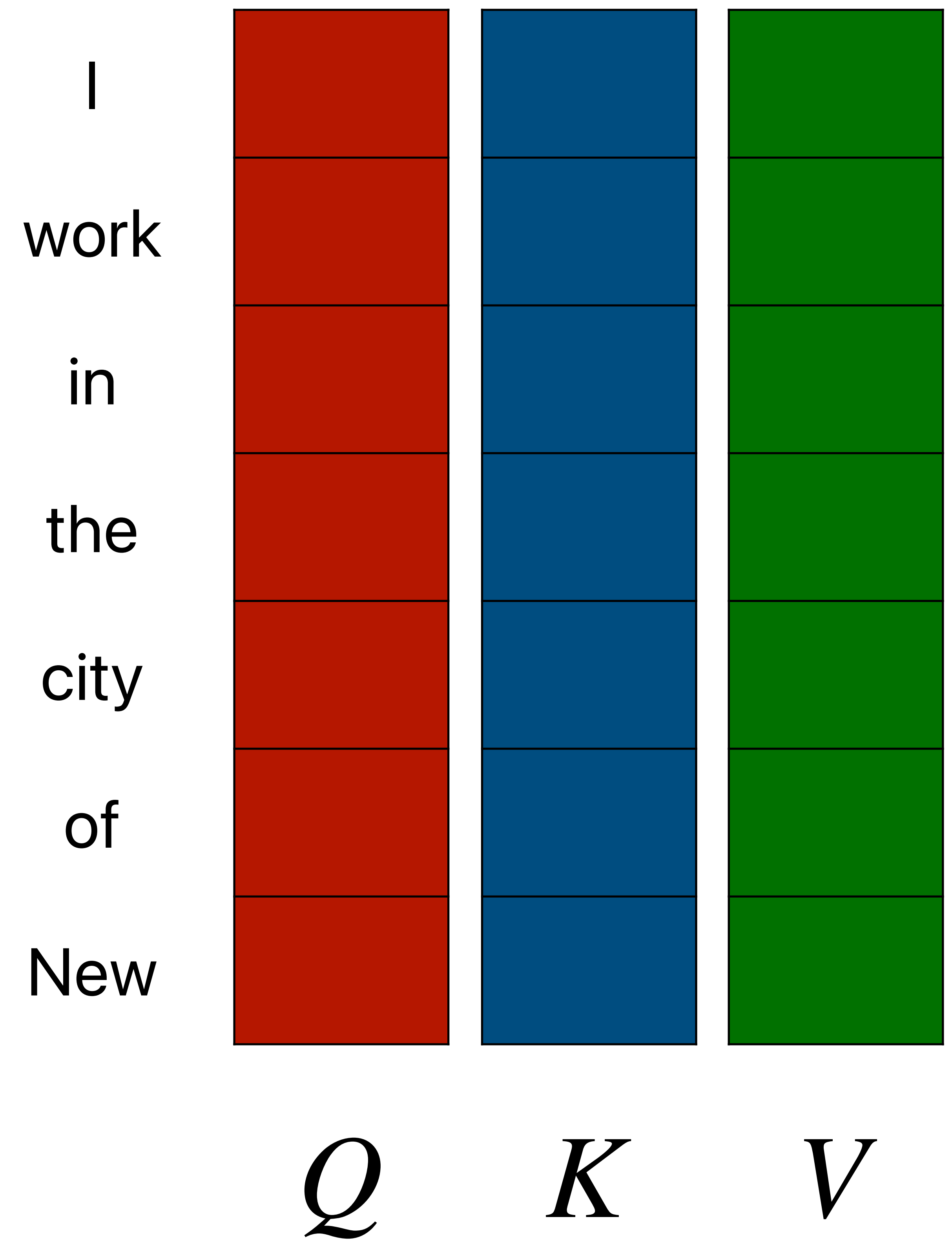

I
work
in
the
city
of
New
Q
K
V

Attention Outputs

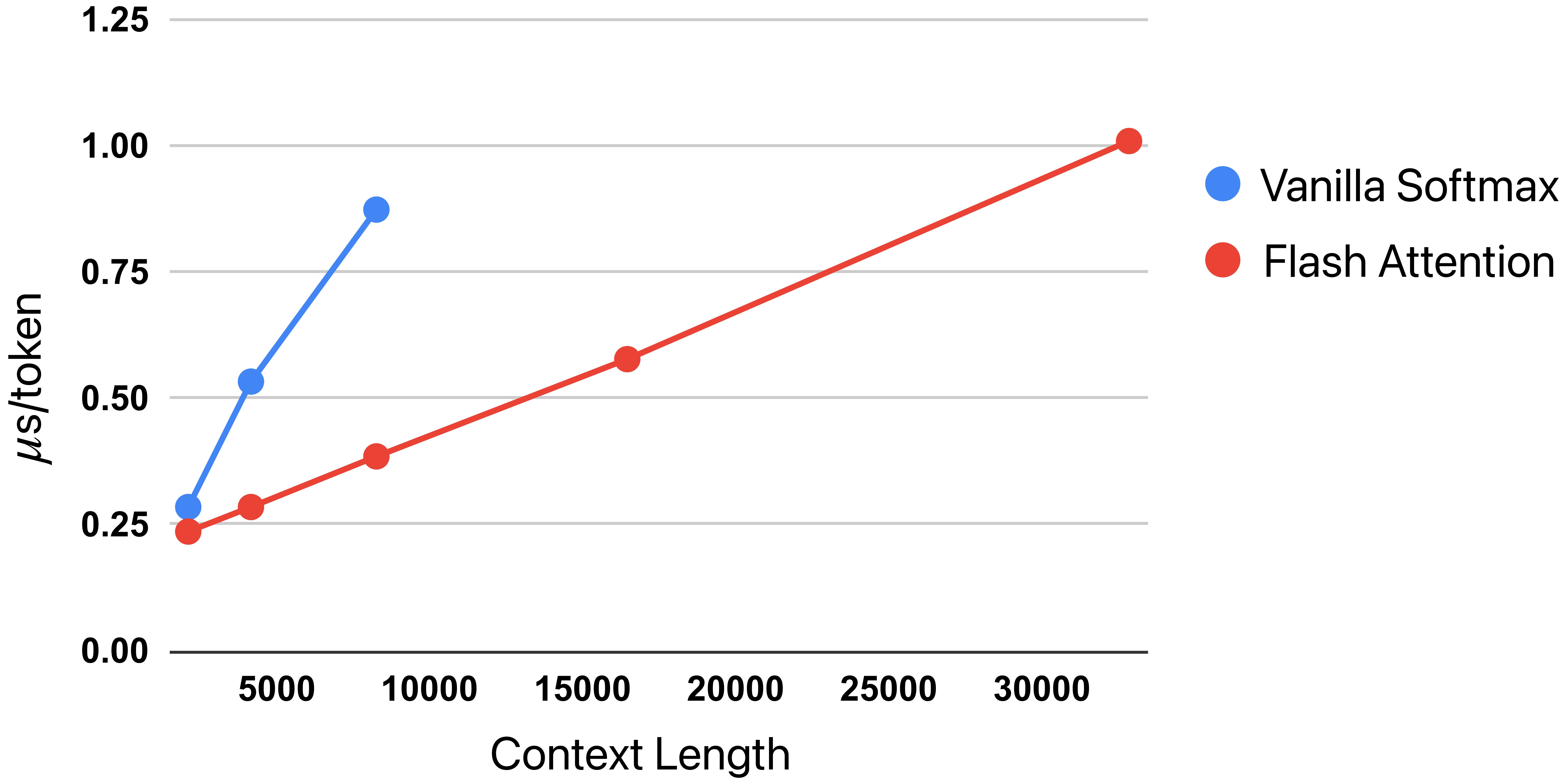
- The computation can be represented as

$$D^{-1} \cdot \mathsf{LT}(\exp(Q \cdot K^{\mathsf{T}})) \cdot V$$

  - $Q \cdot K^{\mathsf{T}}$ is $n \times n$
  - diagonal of $D = \mathsf{LT}(\exp(Q \cdot K^{\mathsf{T}})) \cdot 1_n$
  - Naively computing takes $O(n^2)$ time
    - Prohibitive when n is large



Train Step Latency Per Token



Each token in the training examples looks back at the whole context