A closer look into the transformer architecture

April 28, 2025

Did you complete the reading assignments?

Estimating (GPU) RAM usage of a LLM

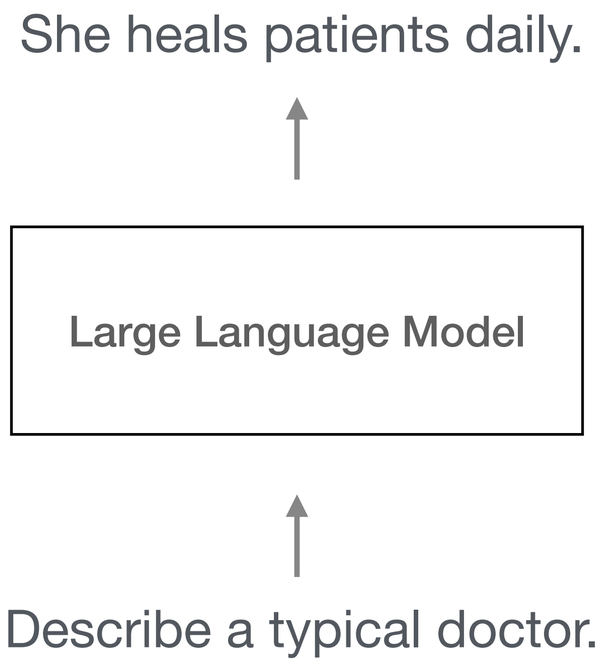
- I have a 7 billion parameter model. How much RAM do I need to host it?

- RAM Usage = 7 billion params x 2 bytes per parameter = 14 billion bytes*
  - **Kilo** bytes = 2^{10} = 1,024 ≈ **1 thousand bytes**
  - **Mega** bytes = 2^{20} = 1,048,576 ≈ **1 million bytes**
  - **Giga** bytes = 2^{30} = 1,073,741,824 ≈ **1 billion bytes**
  - **Tera** bytes = 2^{40} = 1,099,511,627,776 ≈ **1 trillion bytes**

- Unlike regular ML models where a parameter is 4 bytes (32 bit float), LLMs are usually 2 bytes

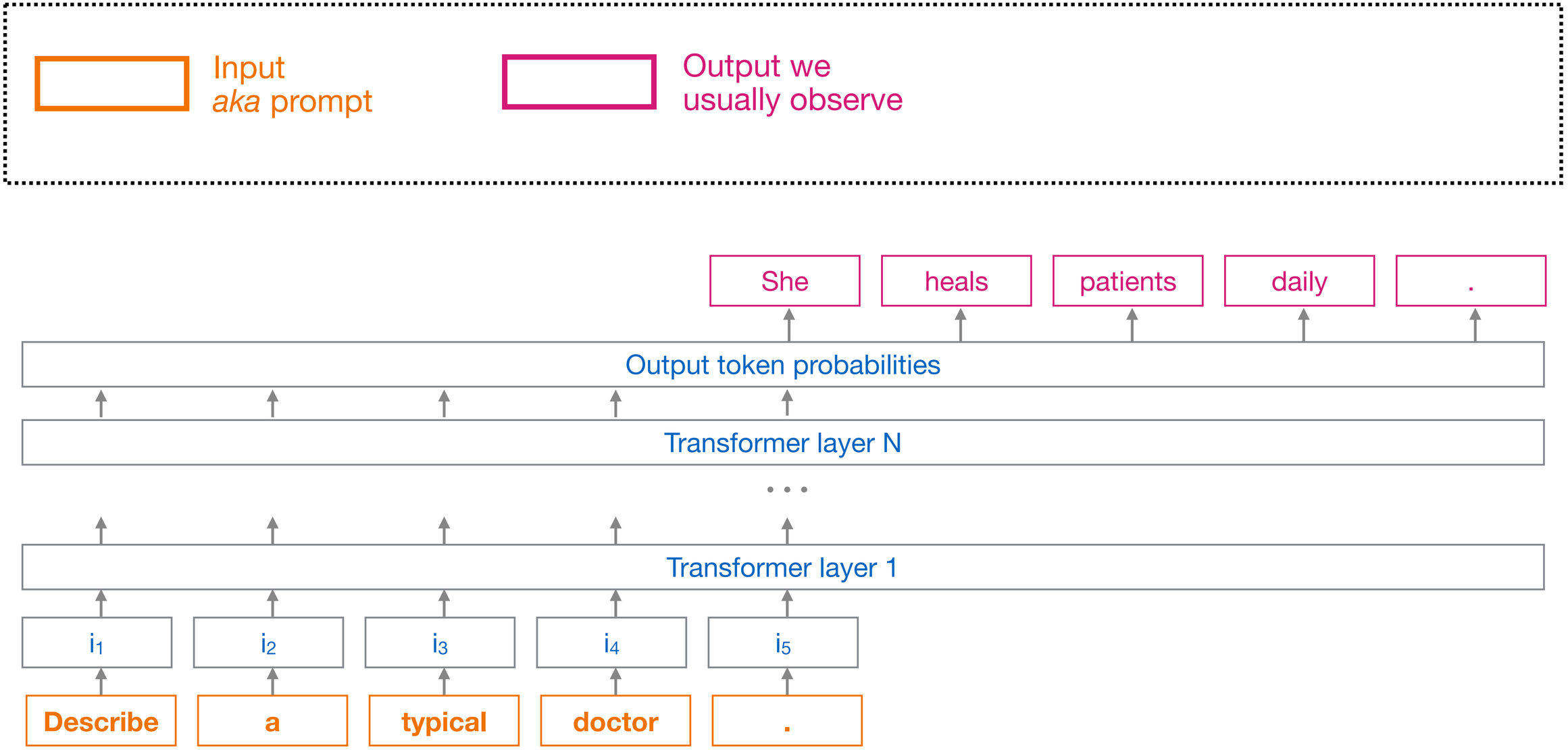
* Actually we need a bit more for the intermediate computations

Recap: How a transformer works

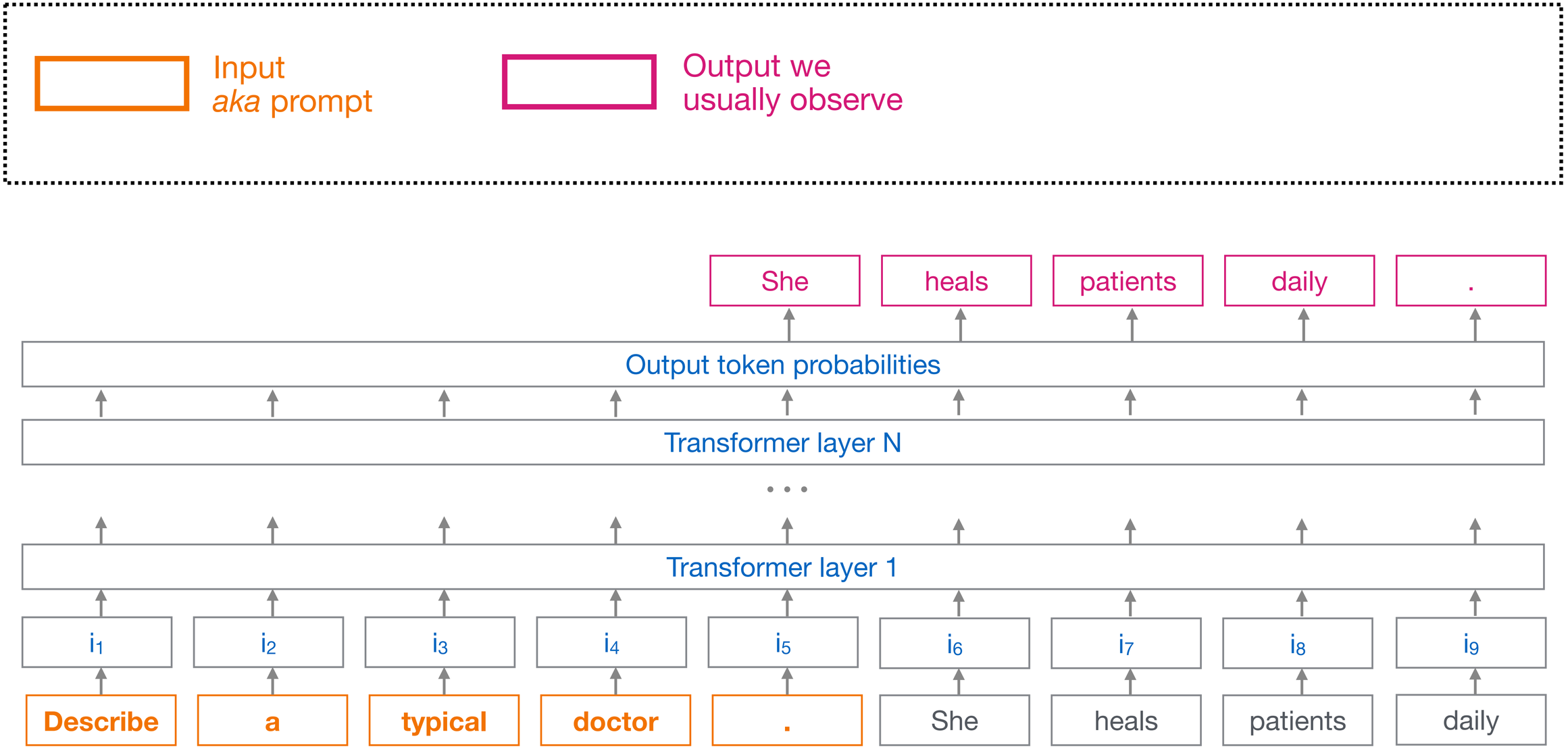


*Fictitious Example

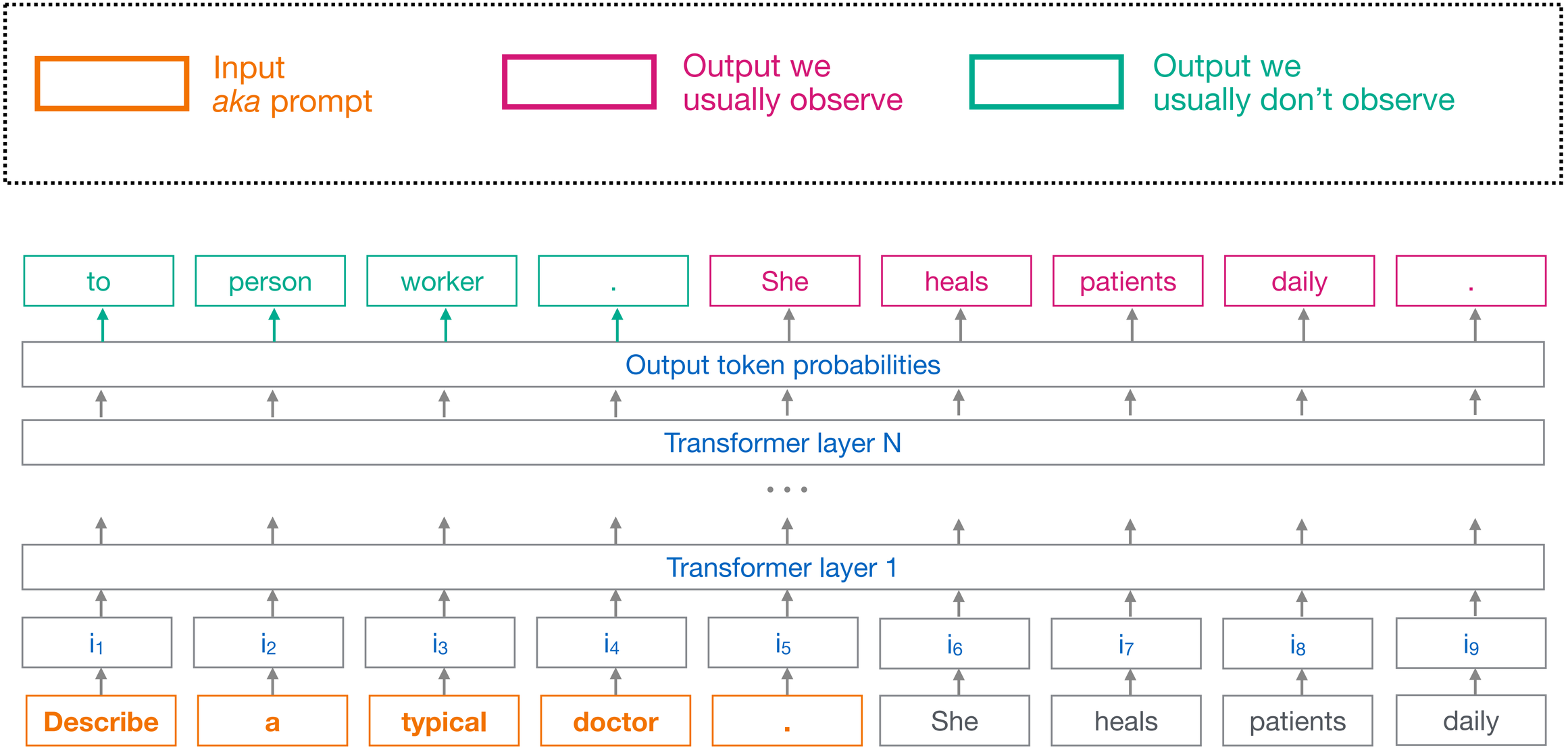
Recap: LLMs and causal language modeling



Recap: LLMs and causal language modeling

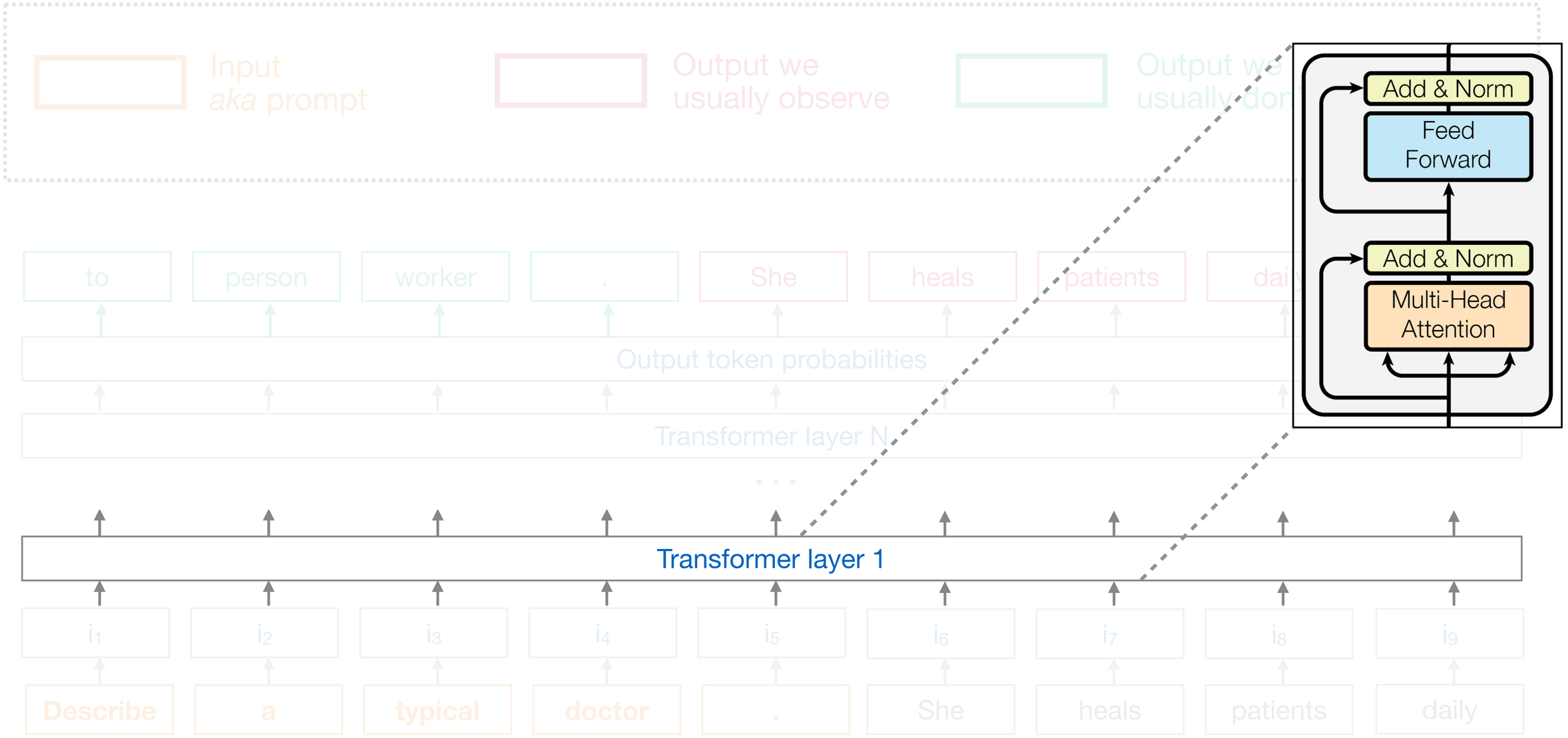


Recap: LLMs and causal language modeling



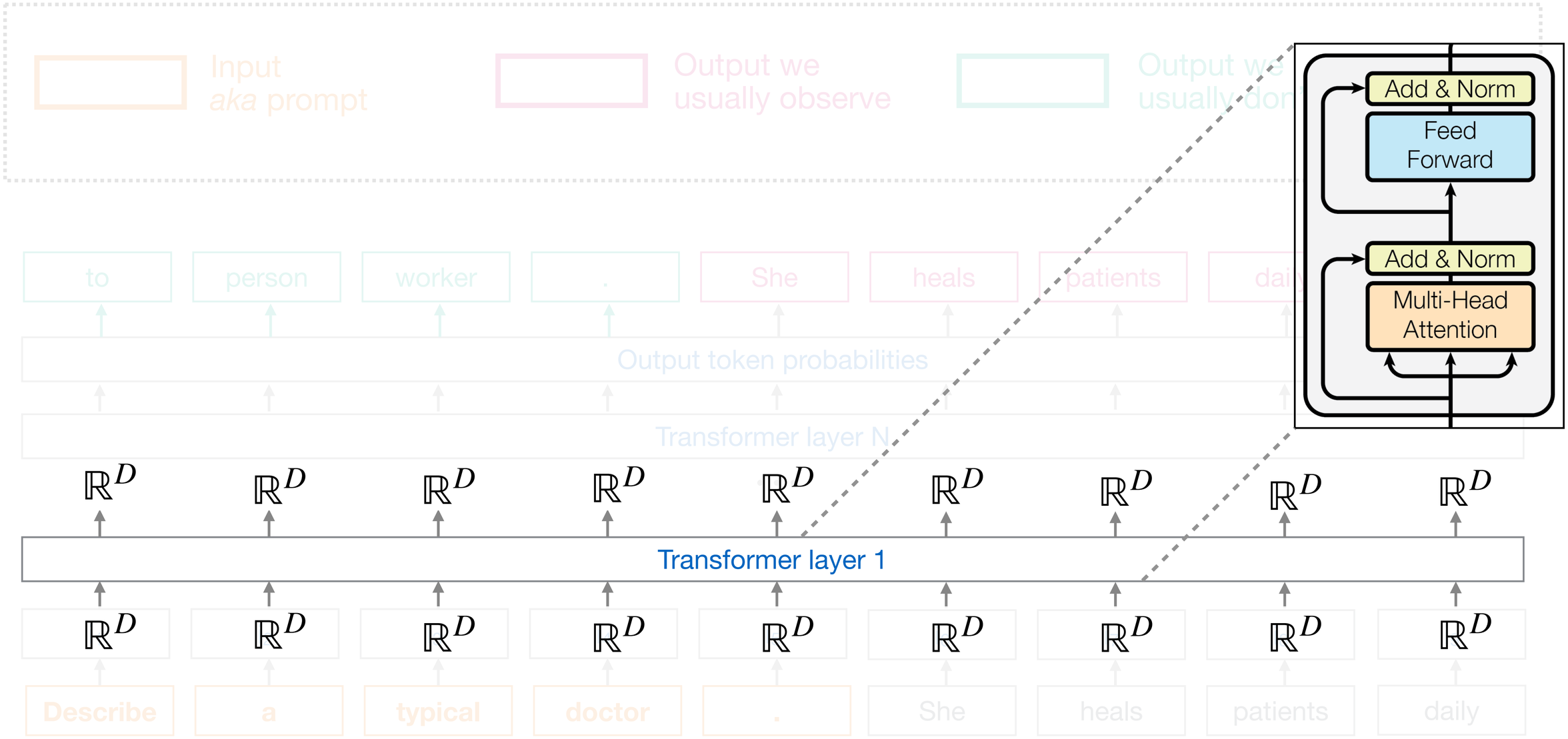
Exercise

Recap: LLMs and causal language modeling



[Vaswani et al.]

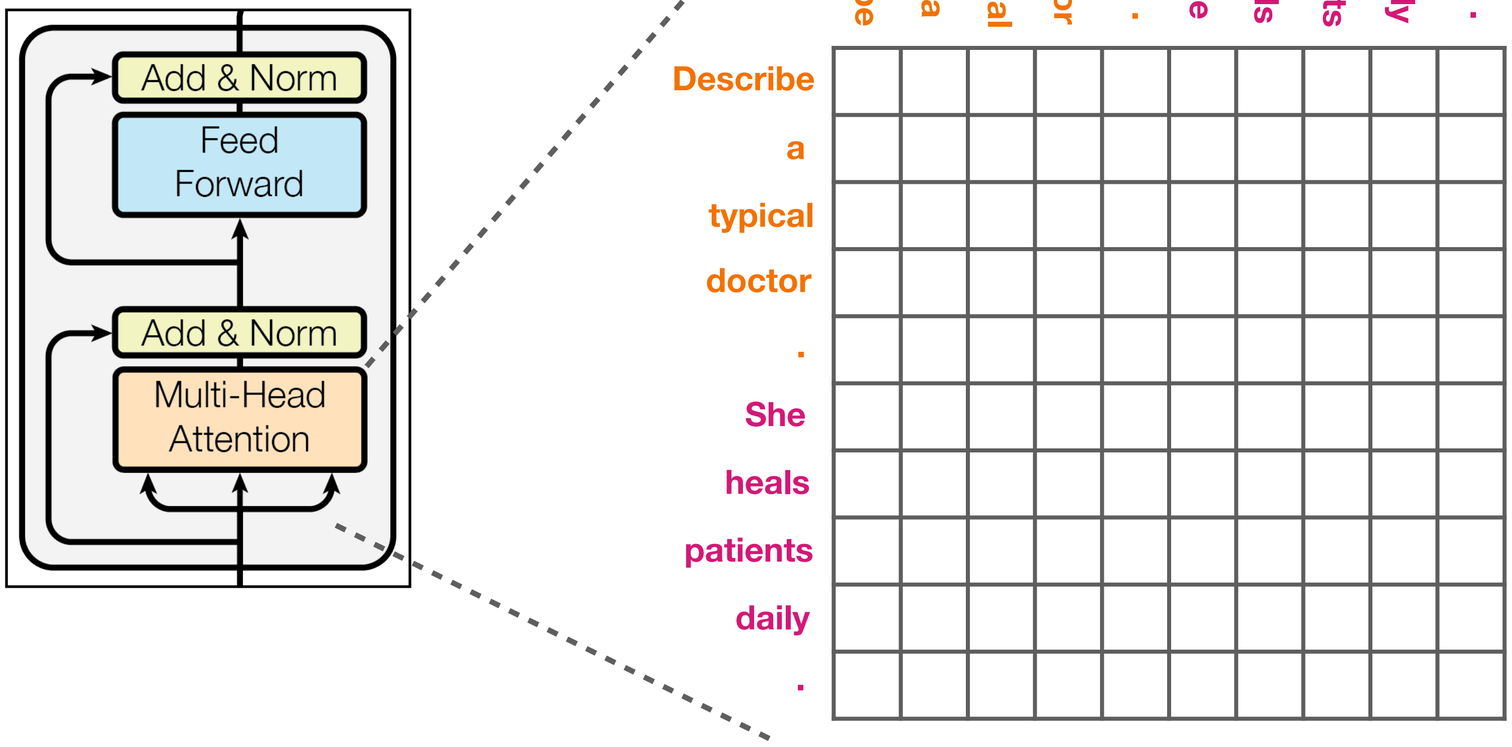
Recap: LLMs and causal language modeling



[Vaswani et al.]

Self attention between all token pairs

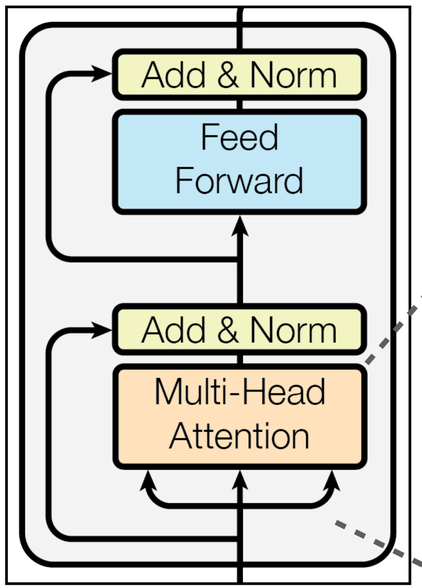
When processing *token i*, how much each *token j* should contribute?

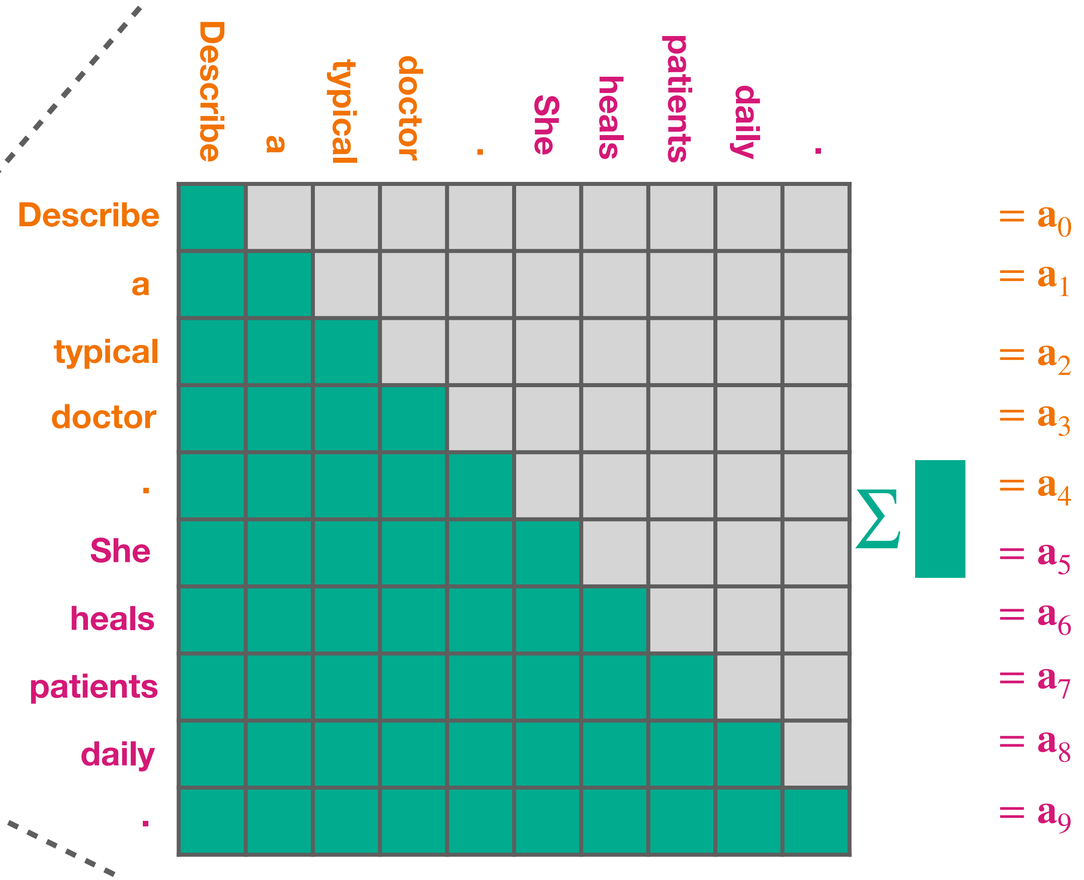


[Vaswani et al.]

Self attention between all token pairs

When processing *token i*, how much each *token j* should contribute?





Causal LLMs

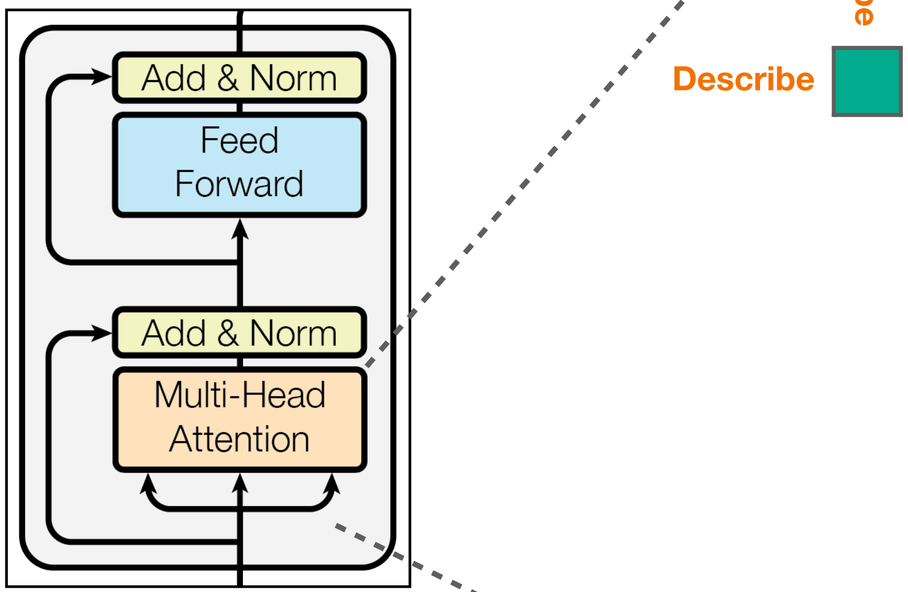
Attention of *token i* depends only on *current or previous* tokens

$$\mathbf{a}_i = \sum_{j \le i} \text{attn}(i, j)$$

[Vaswani et al.]

Self attention between all token pairs

When processing *token i*, how much each *token j* should contribute?



Describe

Describe

$= \mathbf{a}_0$

$\sum$

Causal LLMs

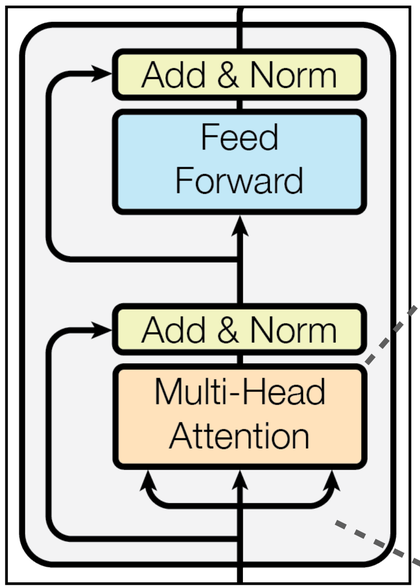
Attention of *token i* depends only on *current or previous* tokens

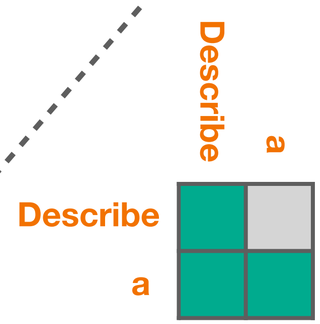
$$\mathbf{a}_i = \sum_{j \leq i} \text{attn}(i, j)$$

[Vaswani et al.]

Self attention between all token pairs

When processing *token i*, how much each *token j* should contribute?





$= \mathbf{a}_0$

$= \mathbf{a}_1$

$\sum$ 

Causal LLMs

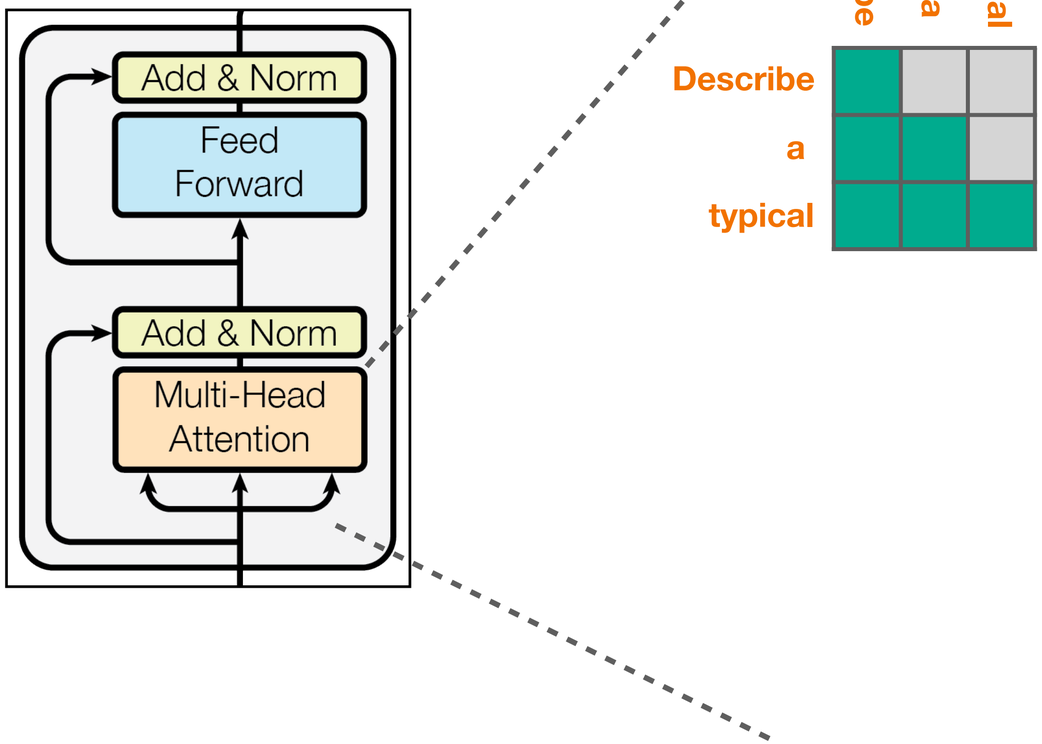
Attention of *token i* depends only on *current or previous* tokens

$$\mathbf{a}_i = \sum_{j \leq i} \text{attn}(i, j)$$

[Vaswani et al.]

Self attention between all token pairs

When processing *token i*, how much each *token j* should contribute?



$= \mathbf{a}_0$

$= \mathbf{a}_1$

$= \mathbf{a}_2$

$\sum$

Causal LLMs

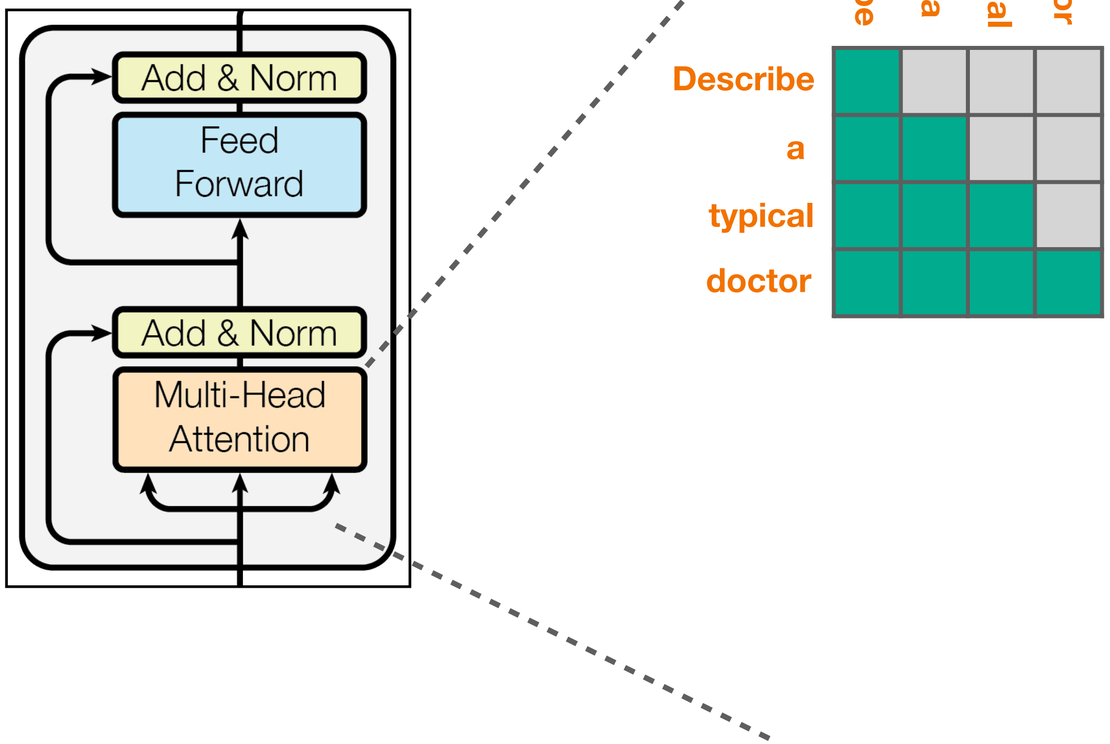
Attention of *token i* depends only on *current or previous* tokens

$$\mathbf{a}_i = \sum_{j \leq i} \text{attn}(i, j)$$

[Vaswani et al.]

Self attention between all token pairs

When processing *token i*, how much each *token j* should contribute?



$$= \mathbf{a}_0$$
$$= \mathbf{a}_1$$
$$= \mathbf{a}_2$$
$$= \mathbf{a}_3$$

$$\sum$$

Causal LLMs

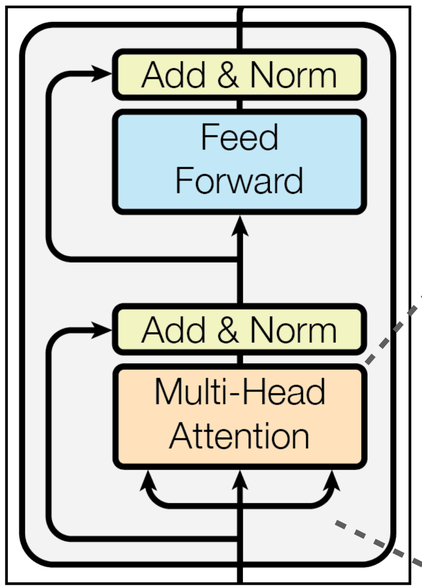
Attention of *token i* depends only on *current or previous* tokens

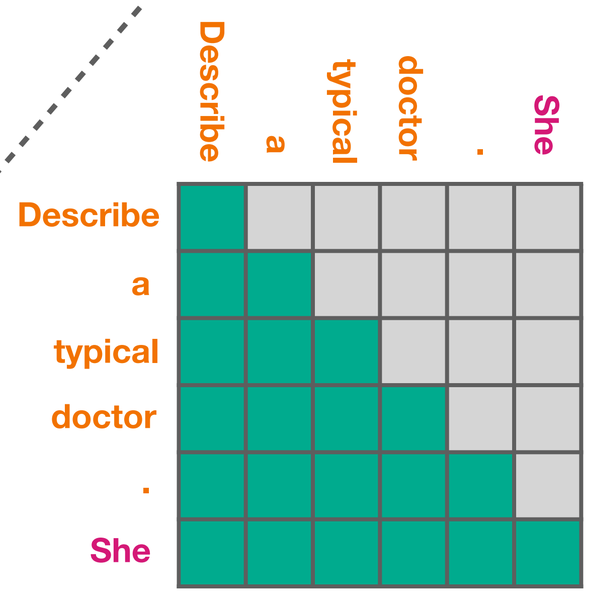
$$\mathbf{a}_i = \sum_{j \le i} \text{attn}(i, j)$$

[Vaswani et al.]

Self attention between all token pairs

When processing *token i*, how much each *token j* should contribute?





$$= \mathbf{a}_0$$
$$= \mathbf{a}_1$$
$$= \mathbf{a}_2$$
$$= \mathbf{a}_3$$
$$= \mathbf{a}_4$$
$$= \mathbf{a}_5$$

Causal LLMs

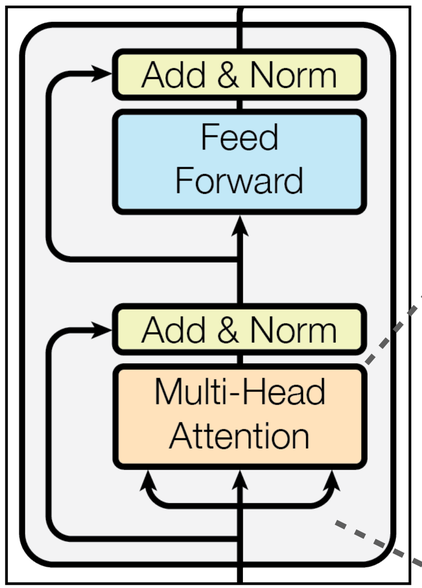
Attention of *token i* depends only on *current or previous* tokens

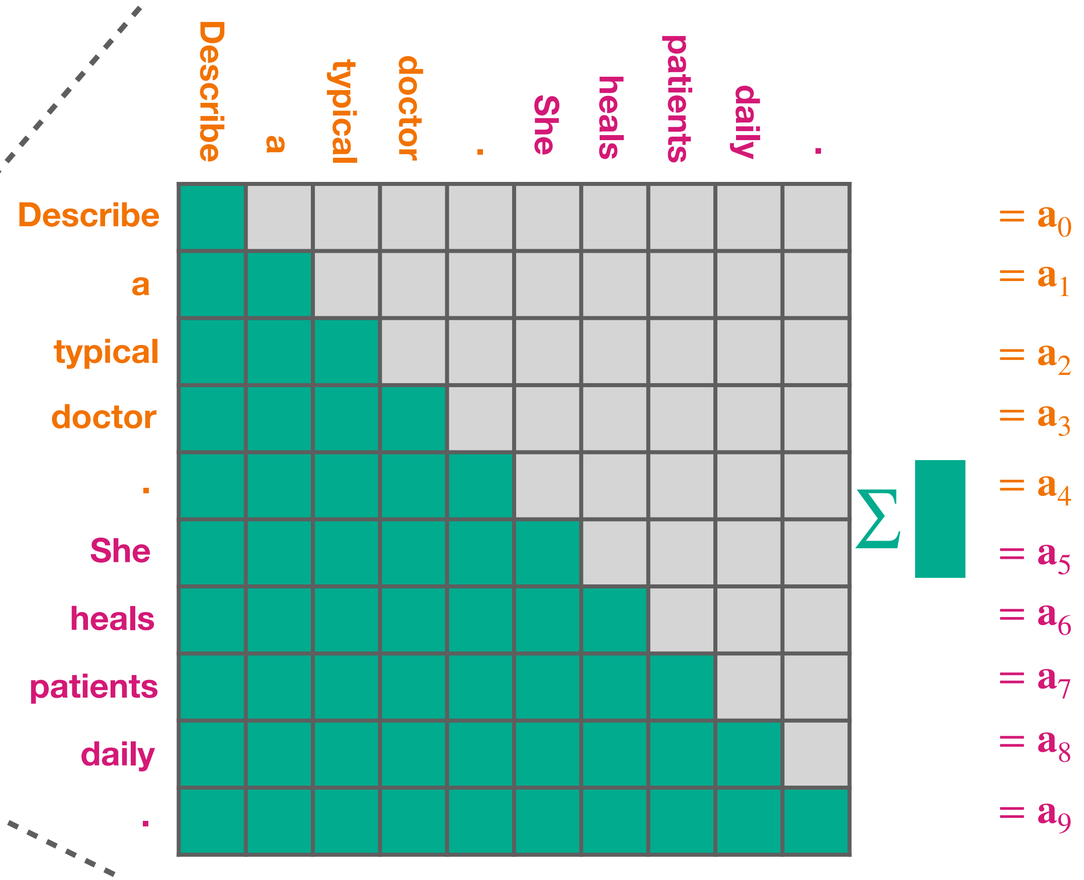
$$\mathbf{a}_i = \sum_{j \le i} \text{attn}(i, j)$$

[Vaswani et al.]

Self attention between all token pairs

When processing *token i*, how much each *token j* should contribute?





Causal LLMs

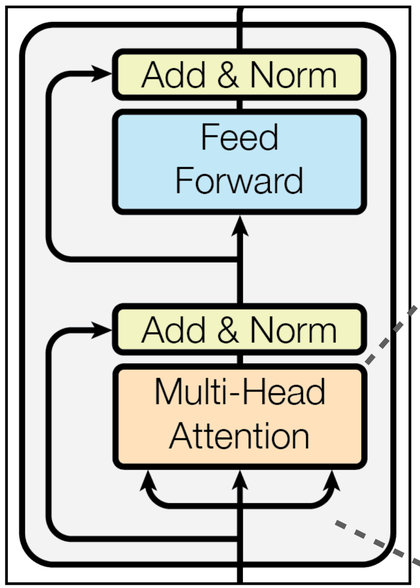
Attention of *token i* depends only on *current or previous* tokens

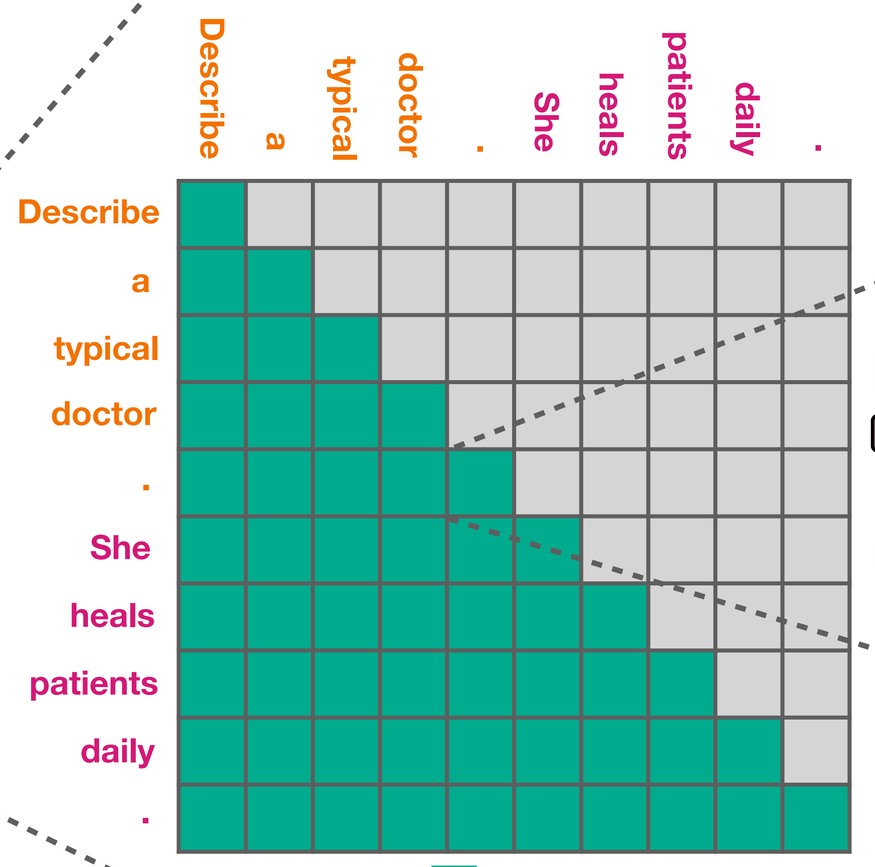
$$\mathbf{a}_i = \sum_{j \le i} \text{attn}(i, j)$$

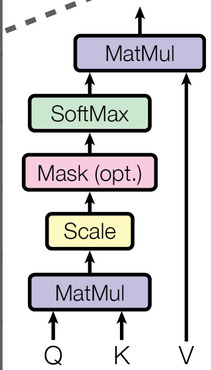
[Vaswani et al.]

Self attention between all token pairs

When processing *token i*, how much each *token j* should contribute?







$$\text{attn}(i,j) = \text{softmax}(\frac{Q_i K_j^T}{\sqrt{d_k}})V_j$$

$$\text{attn}(i,j) \in \mathbb{R}^D$$

$$\mathbf{a}_i = \sum_{j \le i} \text{attn}(i,j)$$

[Vaswani et al.]

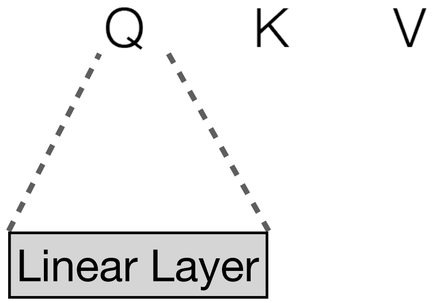
The nitty gritty

$Q_i \in \mathbb{R}^D$

$K_i \in \mathbb{R}^D$

$V_i \in \mathbb{R}^D$

$T_i \in \mathbb{R}^D$

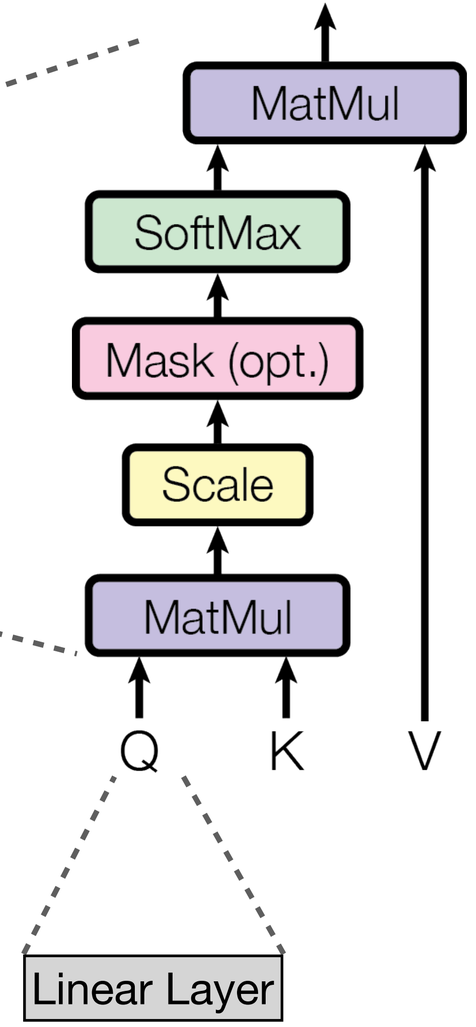


Vaswani et al.

The nitty gritty

$\text{attn}(i, j) \in \mathbb{R}^D$

$$\text{attn}(i, j) = \text{softmax}(\frac{Q_i K_j^T}{\sqrt{d_k}}) V_j$$



$Q_i \in \mathbb{R}^D$

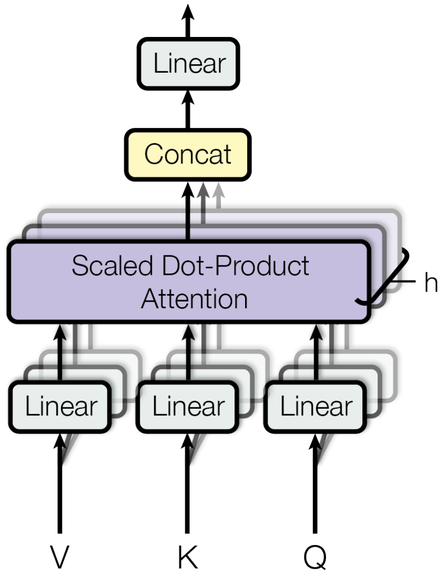
$K_i \in \mathbb{R}^D$

$V_i \in \mathbb{R}^D$

$T_i \in \mathbb{R}^D$

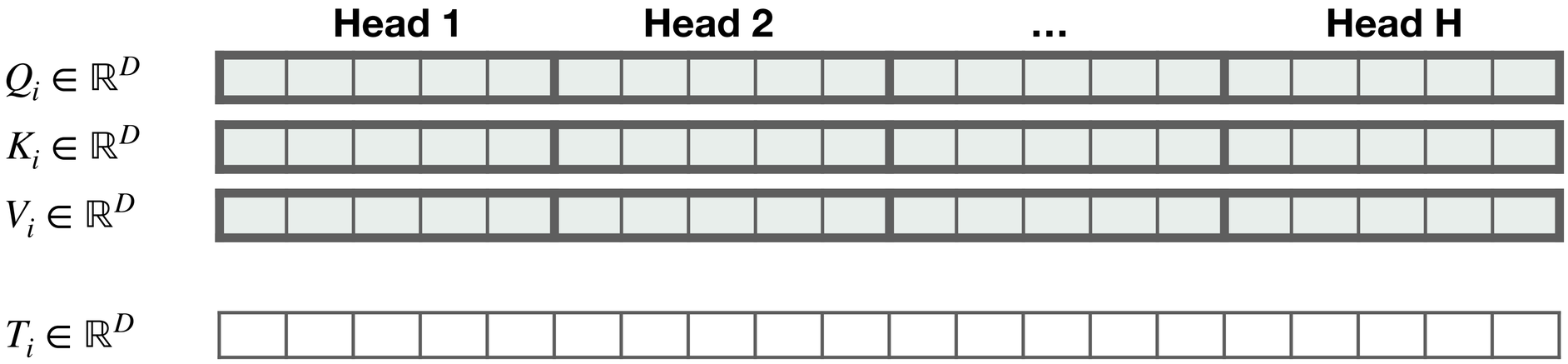
Vaswani et al.

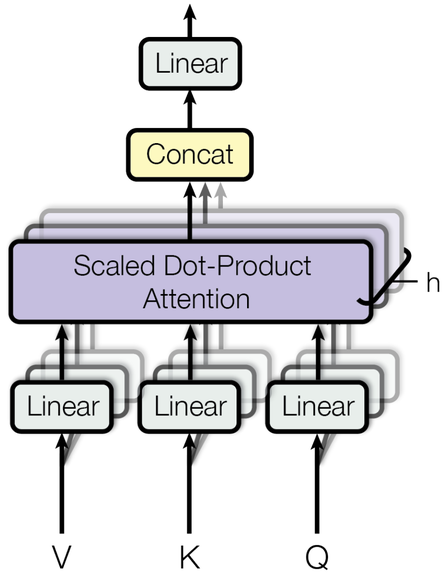
Attention is multi-headed



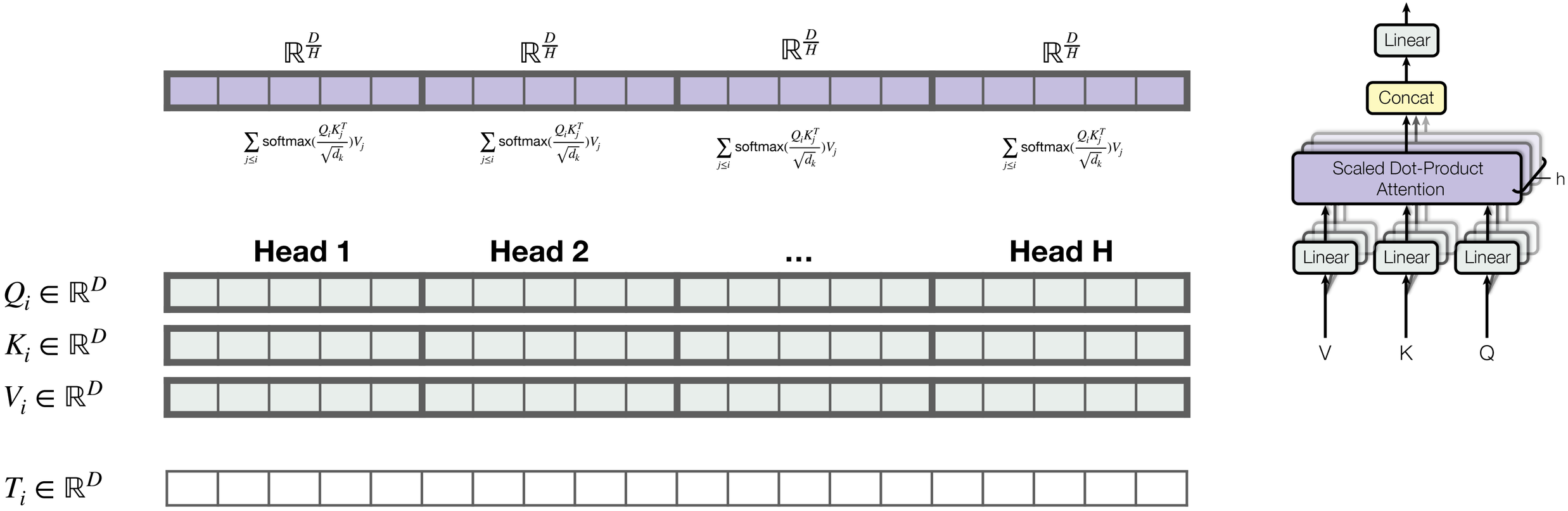
$T_i \in \mathbb{R}^D$

Attention is multi-headed

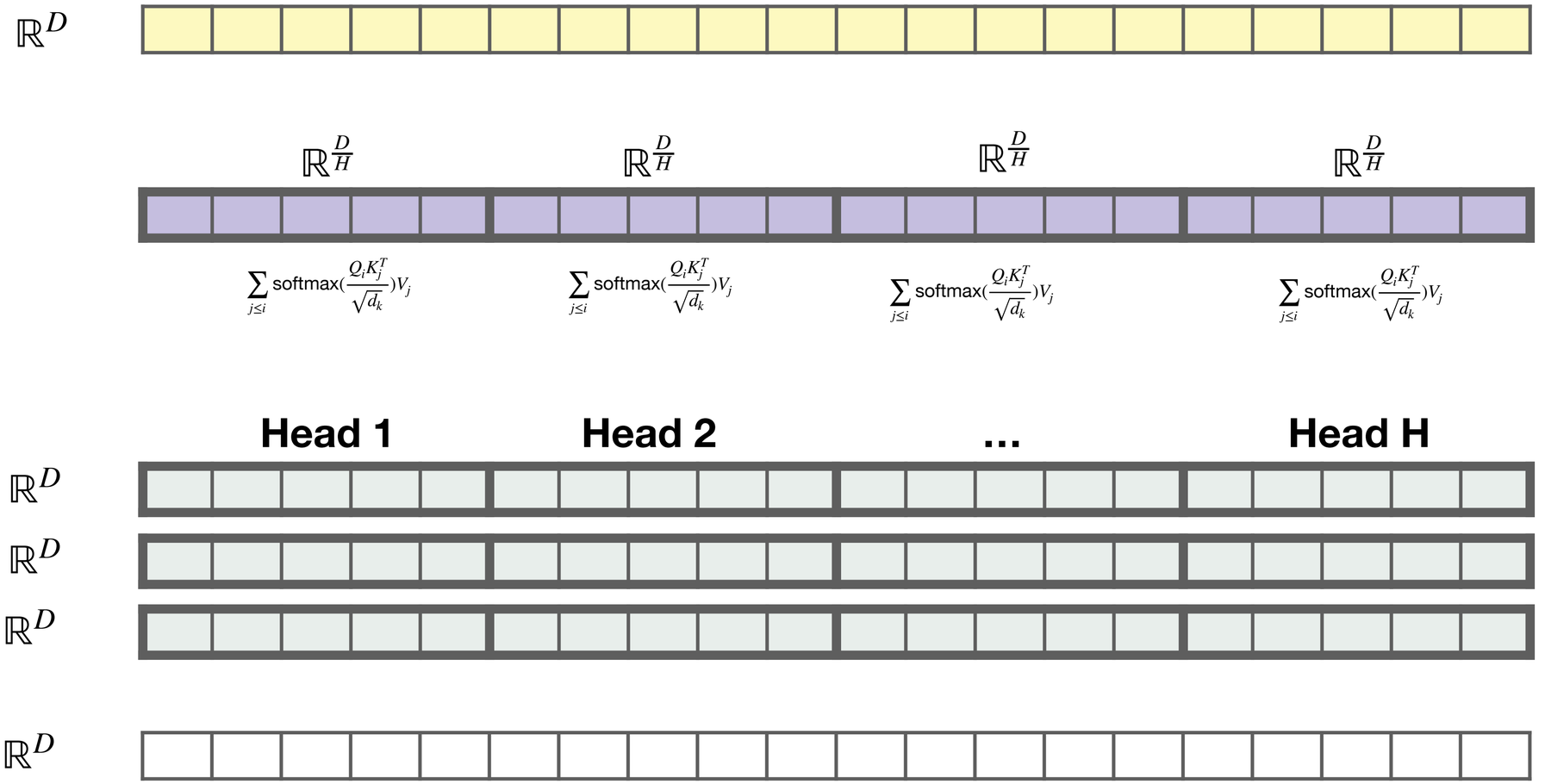


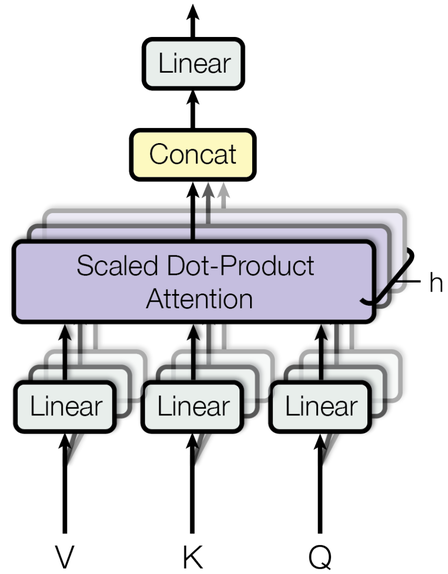


Attention is multi-headed



Attention is multi-headed





Attention is multi-headed

$A_i \in \mathbb{R}^D$

$\mathbb{R}^D$

$\mathbb{R}^{\frac{D}{H}}$ $\mathbb{R}^{\frac{D}{H}}$ $\mathbb{R}^{\frac{D}{H}}$ $\mathbb{R}^{\frac{D}{H}}$

$\sum_{j \le i} \text{softmax}(\frac{Q_i K_j^T}{\sqrt{d_k}}) V_j$ $\sum_{j \le i} \text{softmax}(\frac{Q_i K_j^T}{\sqrt{d_k}}) V_j$ $\sum_{j \le i} \text{softmax}(\frac{Q_i K_j^T}{\sqrt{d_k}}) V_j$ $\sum_{j \le i} \text{softmax}(\frac{Q_i K_j^T}{\sqrt{d_k}}) V_j$

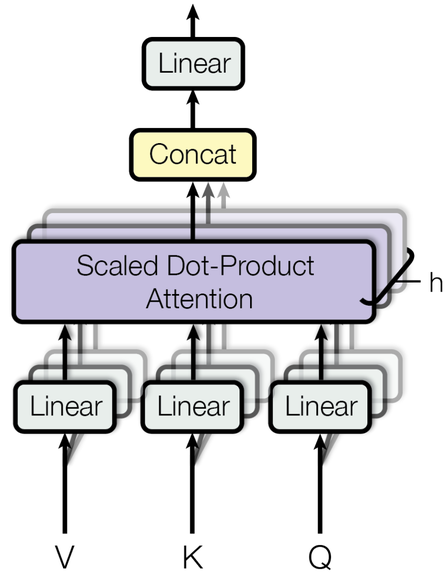
Head 1 **Head 2** **...** **Head H**

$Q_i \in \mathbb{R}^D$

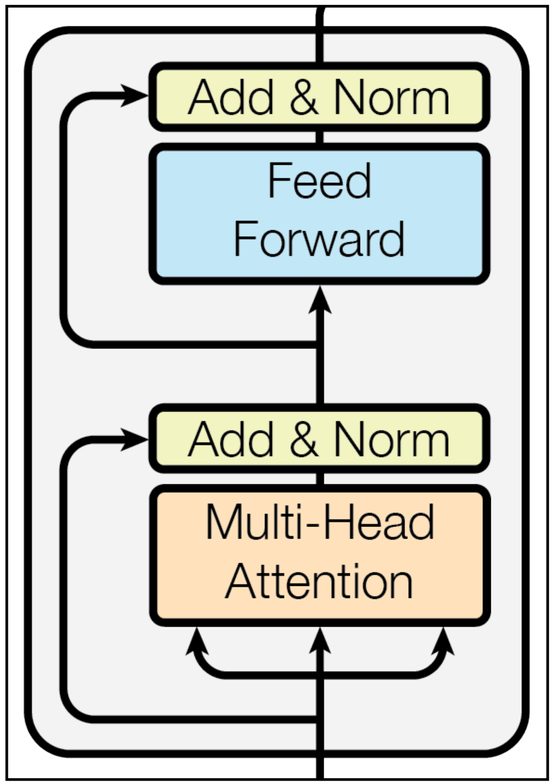
$K_i \in \mathbb{R}^D$

$V_i \in \mathbb{R}^D$

$T_i \in \mathbb{R}^D$

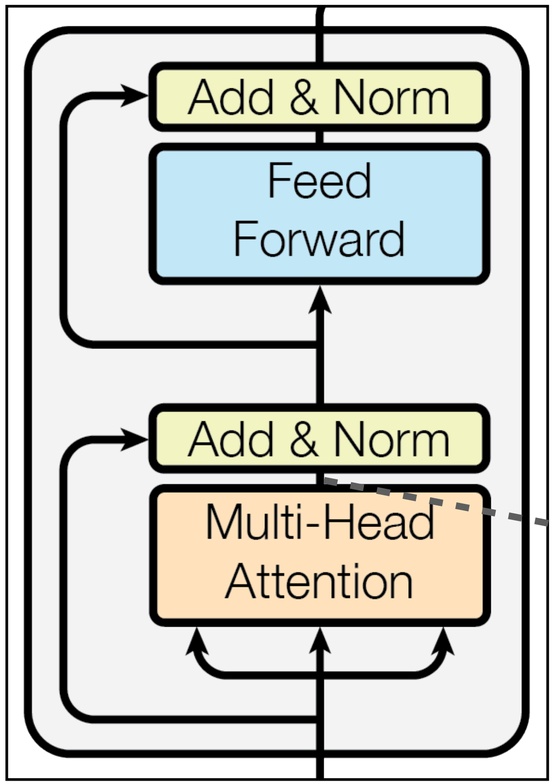


The full transformer layer



[Vaswani et al.]

The full transformer layer

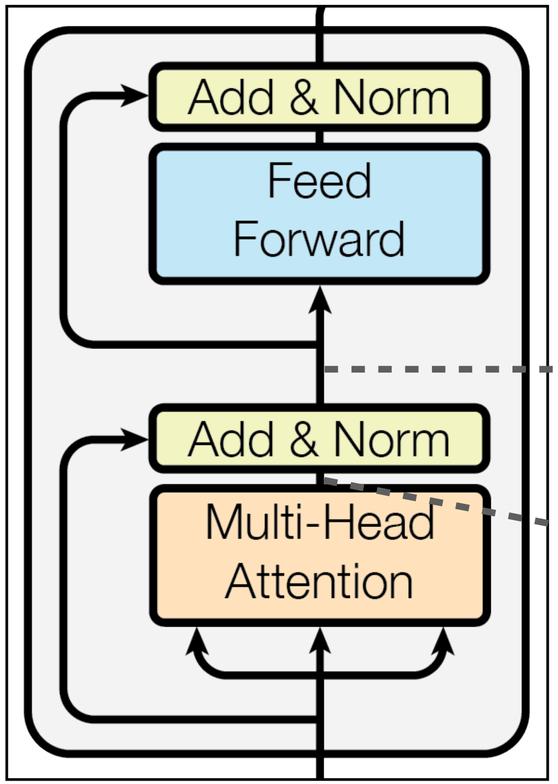


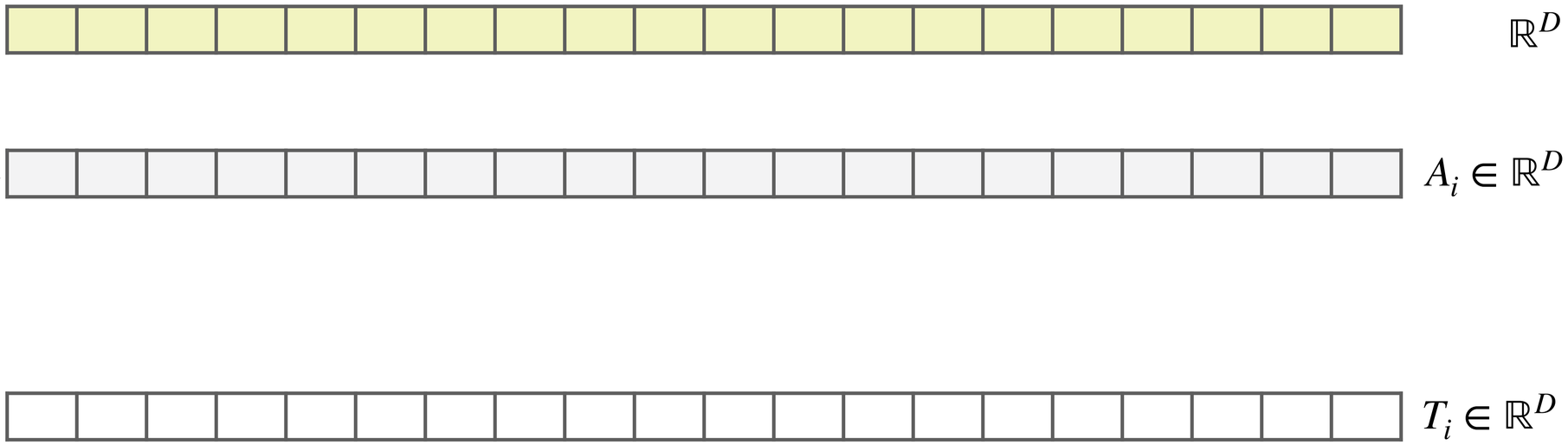
$A_i \in \mathbb{R}^D$

$T_i \in \mathbb{R}^D$

[Vaswani et al.]

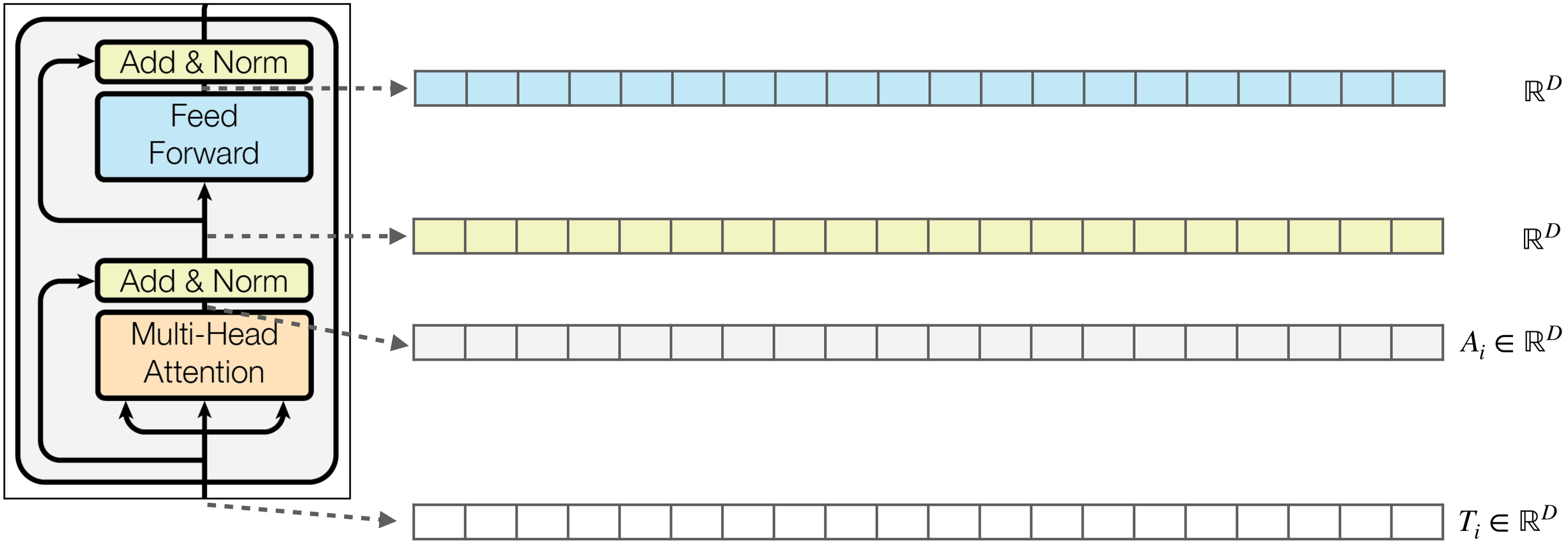
The full transformer layer





[Vaswani et al.]

The full transformer layer



[Vaswani et al.]

The full transformer layer

[Vaswani et al.]

The full transformer layer

Both input and the output are $\mathbb{R}^D$

[Vaswani et al.]

The full transformer layer

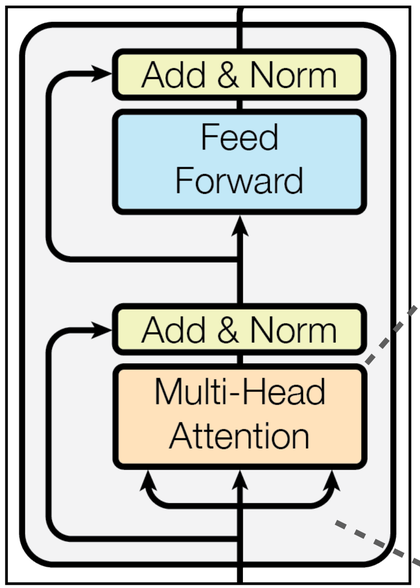
Both input and the output are $\mathbb{R}^D$

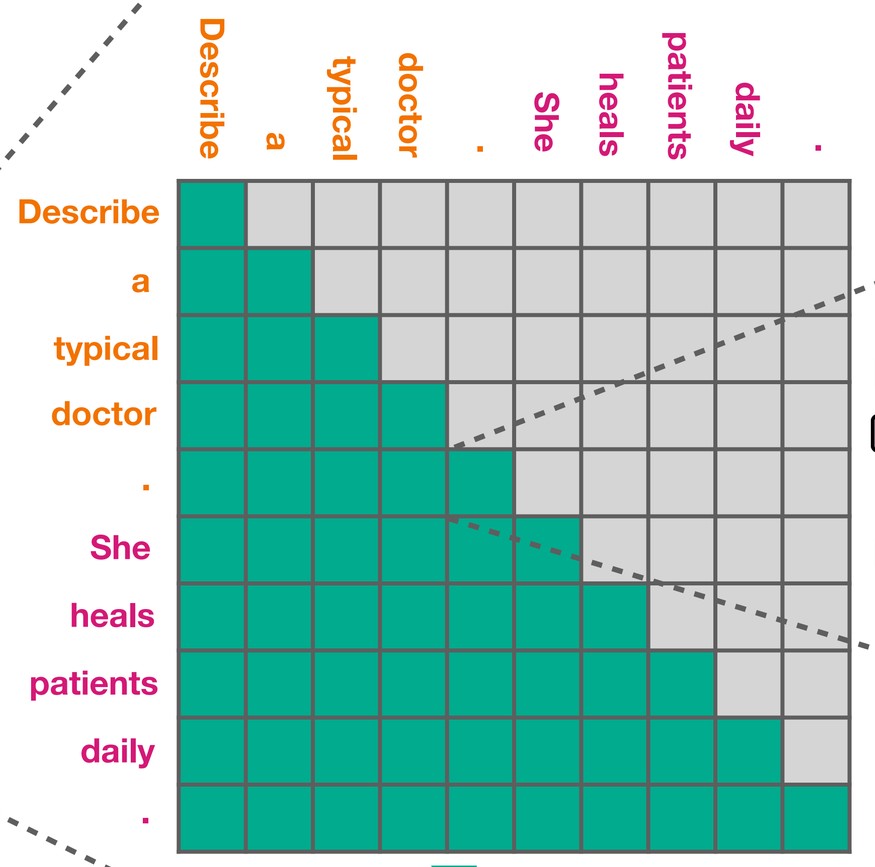
Which part is computationally most expensive?

[Vaswani et al.]

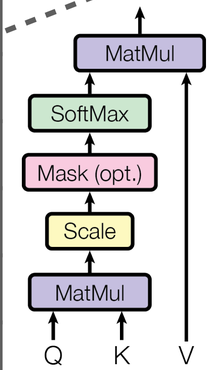
Self attention between all token pairs

When processing *token i*, how much each *token j* should contribute?





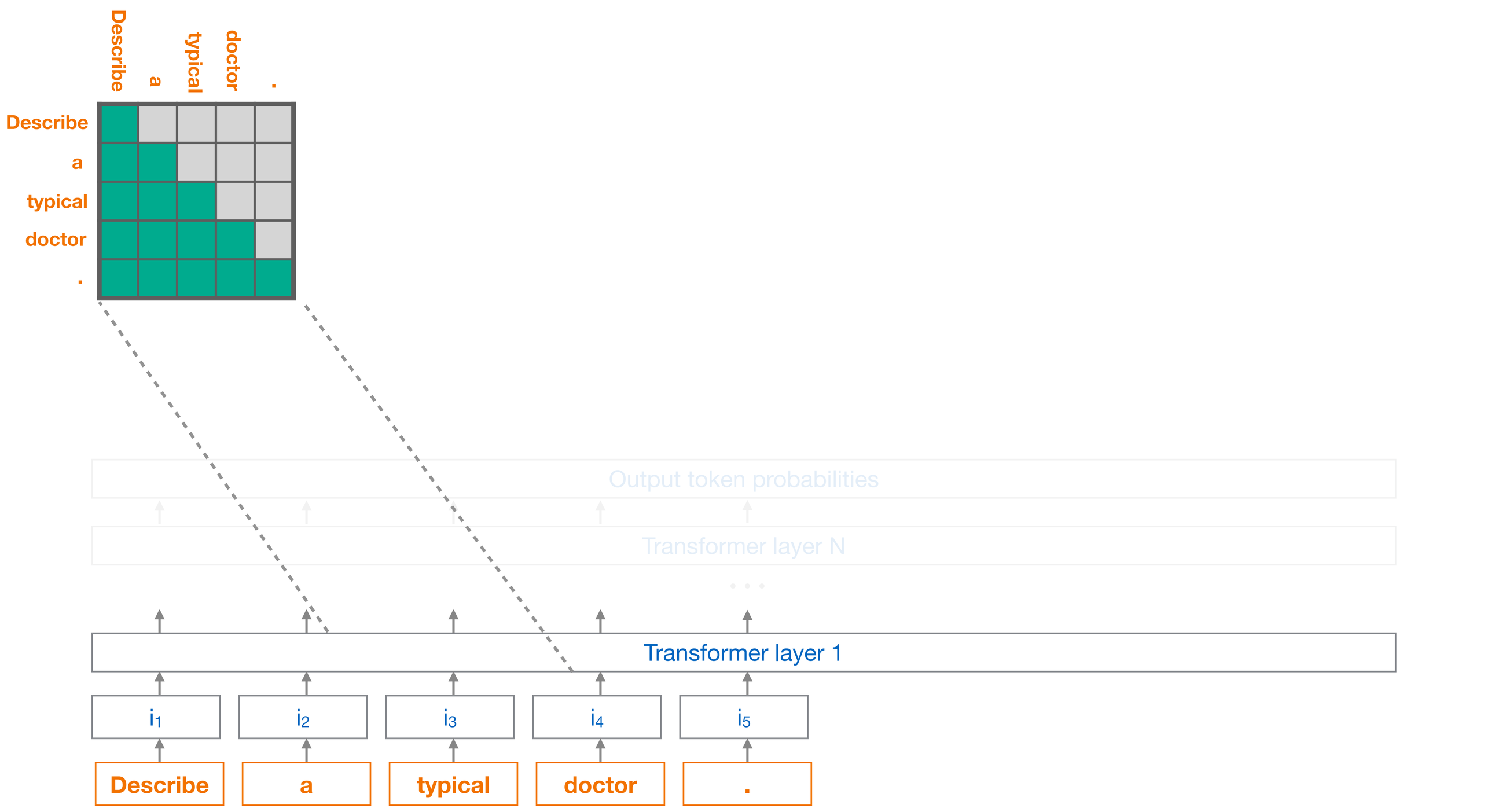
$$\mathbf{a}_i = \sum_{j \le i} \text{attn}(i, j)$$

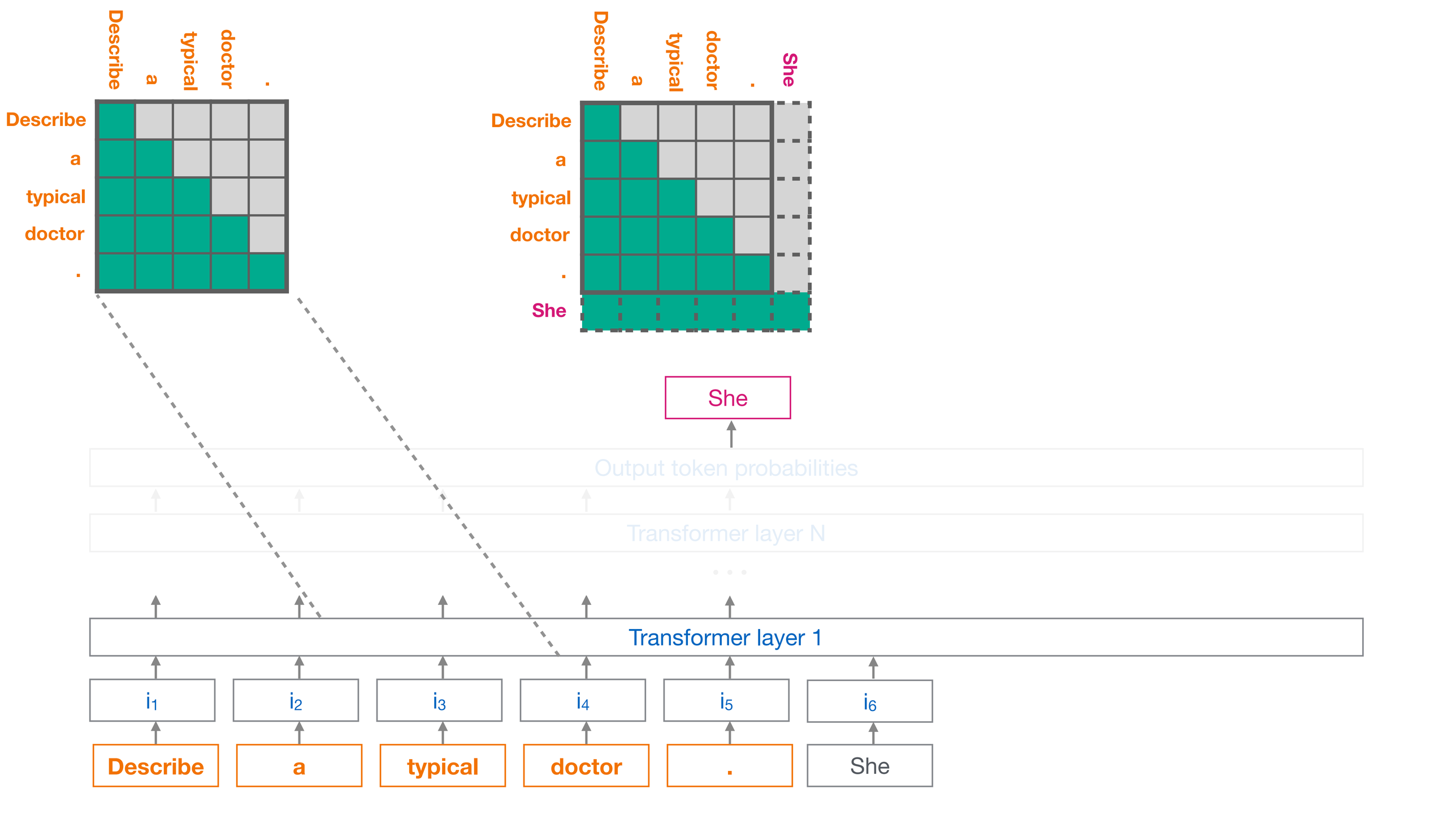


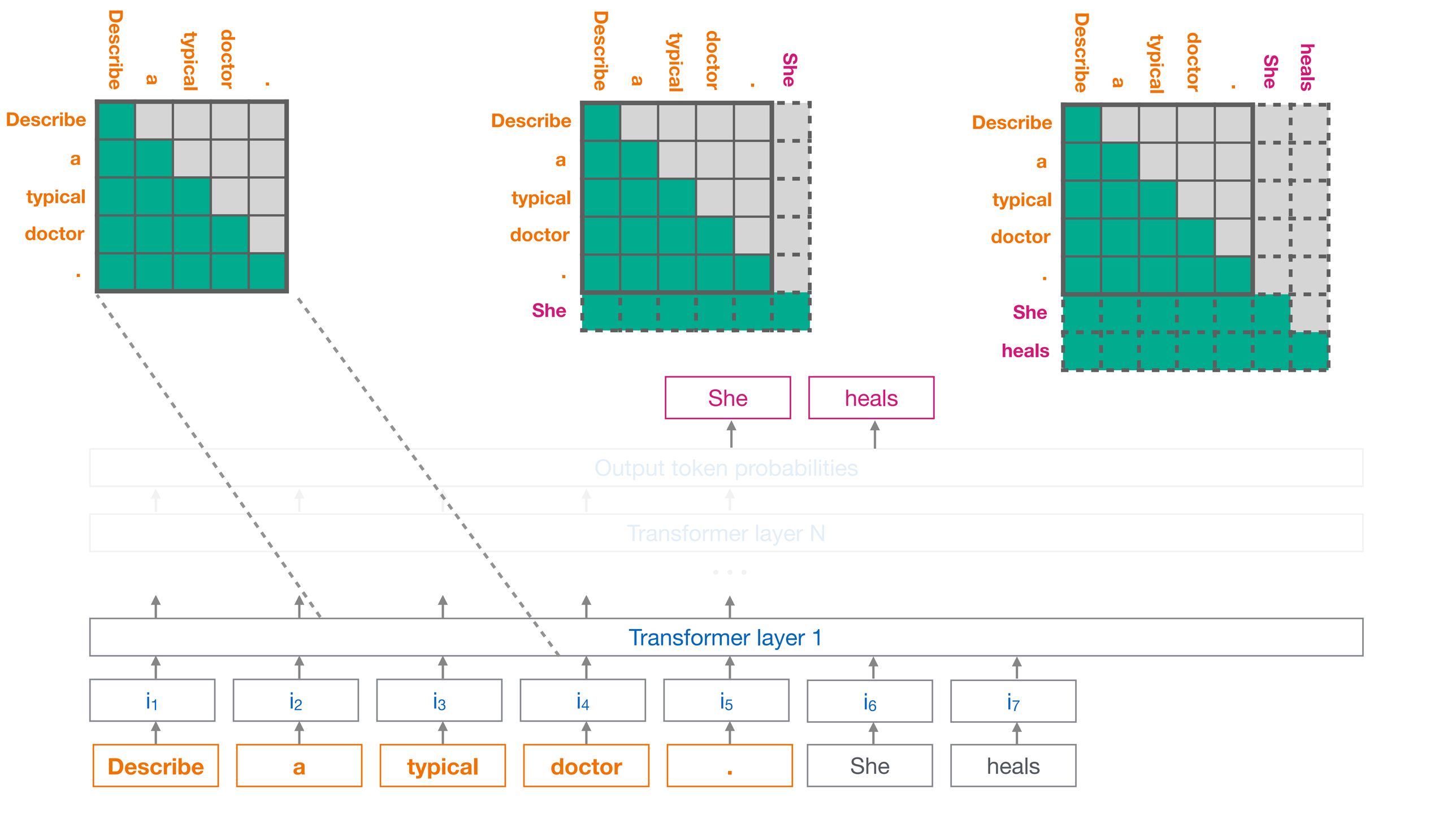
Quadratic in # tokens

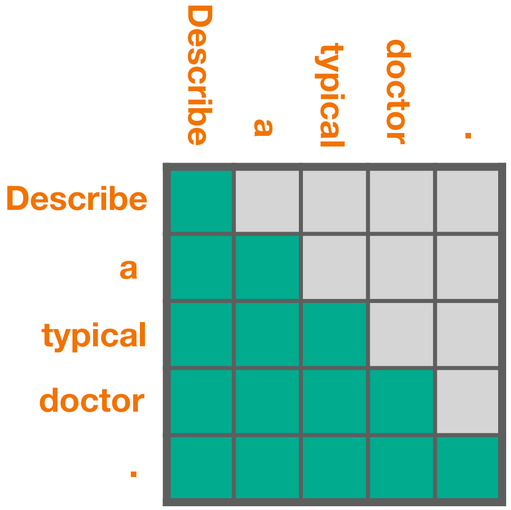
$$\text{attn}(i, j) = \text{softmax}(\frac{Q_i K_j^T}{\sqrt{d_k}}) V_j$$

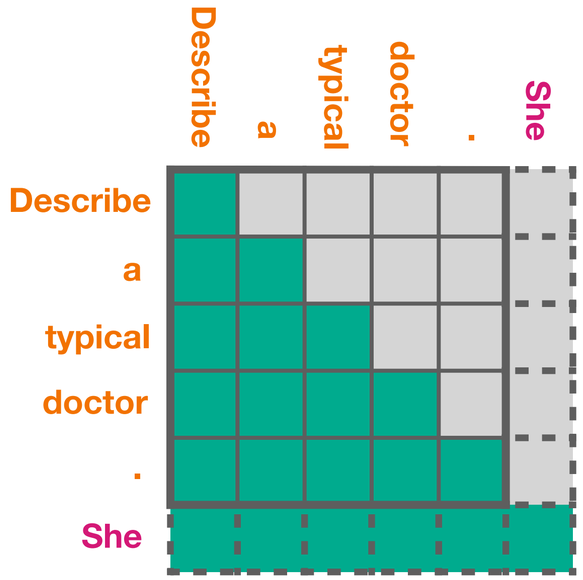
[Vaswani et al.]

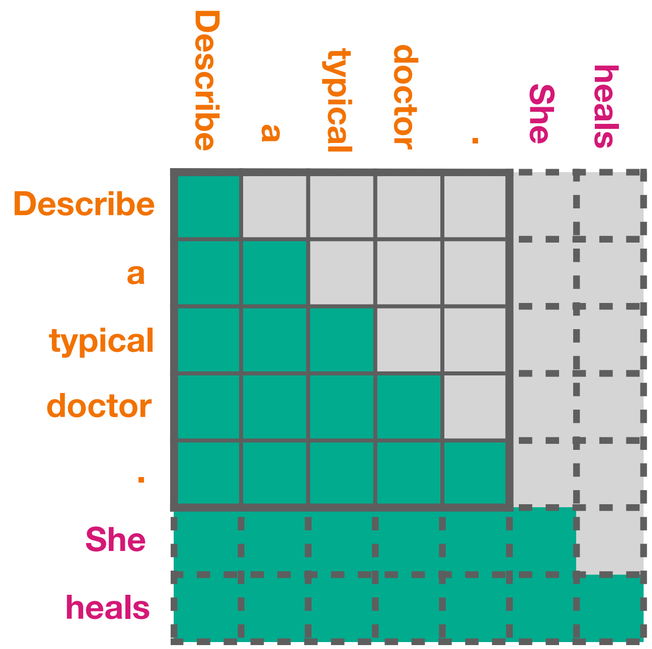

Describe
a
typical
doctor
.
Describe
a
typical
doctor
.
Output token probabilities
Transformer layer N
Transformer layer 1
i1
i2
i3
i4
i5
Describe
a
typical
doctor
.


Describe
a
typical
doctor
.
She
Describe
a
typical
doctor
.
She
Output token probabilities
Transformer layer N
Transformer layer 1
i1
i2
i3
i4
i5
i6
Describe
a
typical
doctor
.
She


Describe
a
typical
doctor
.
She
heals
Output token probabilities
Transformer layer N
Transformer layer 1
i1
i2
i3
i4
i5
i6
i7
Describe
a
typical
doctor
.
She
heals





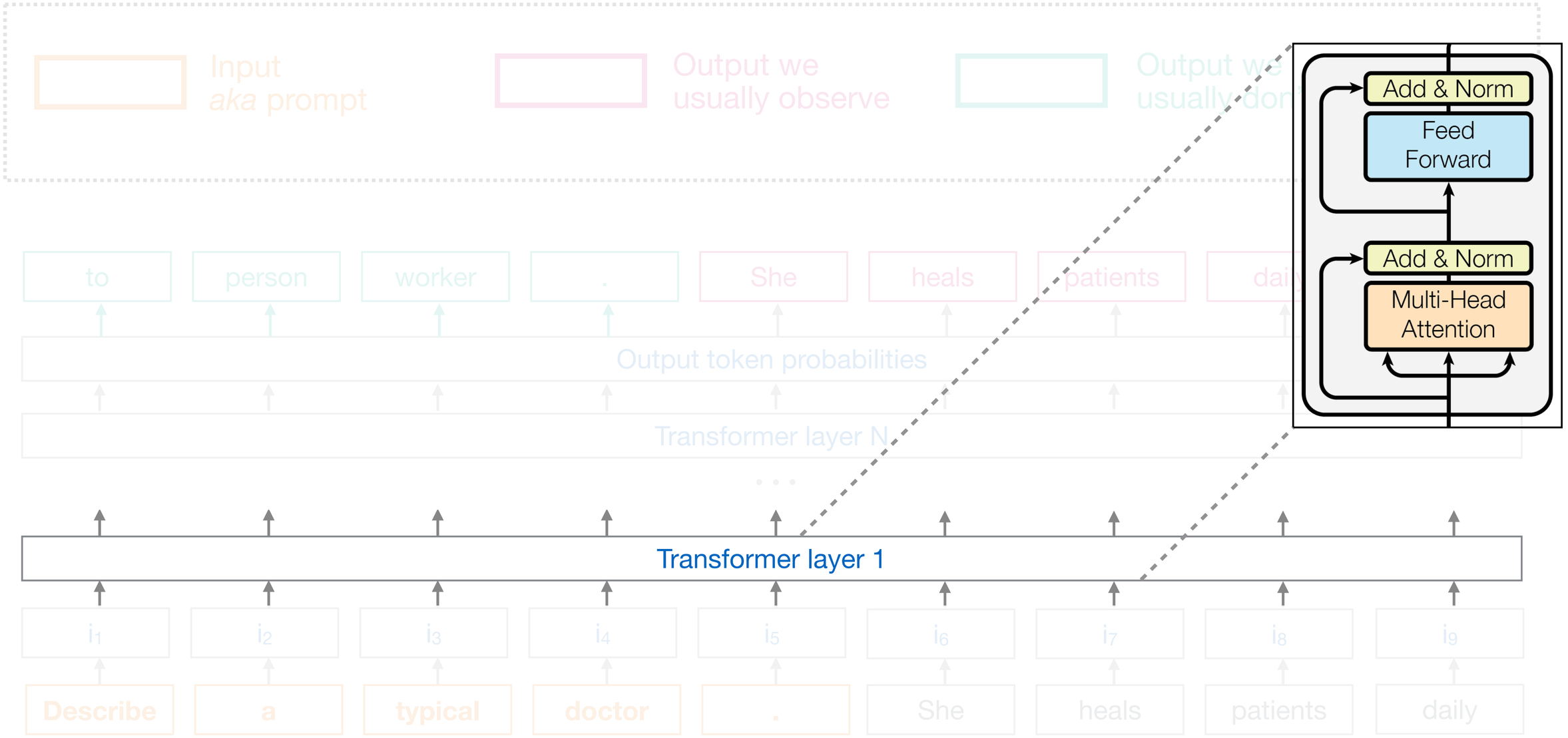


Repeated compute for every generated token

KV-Caching:

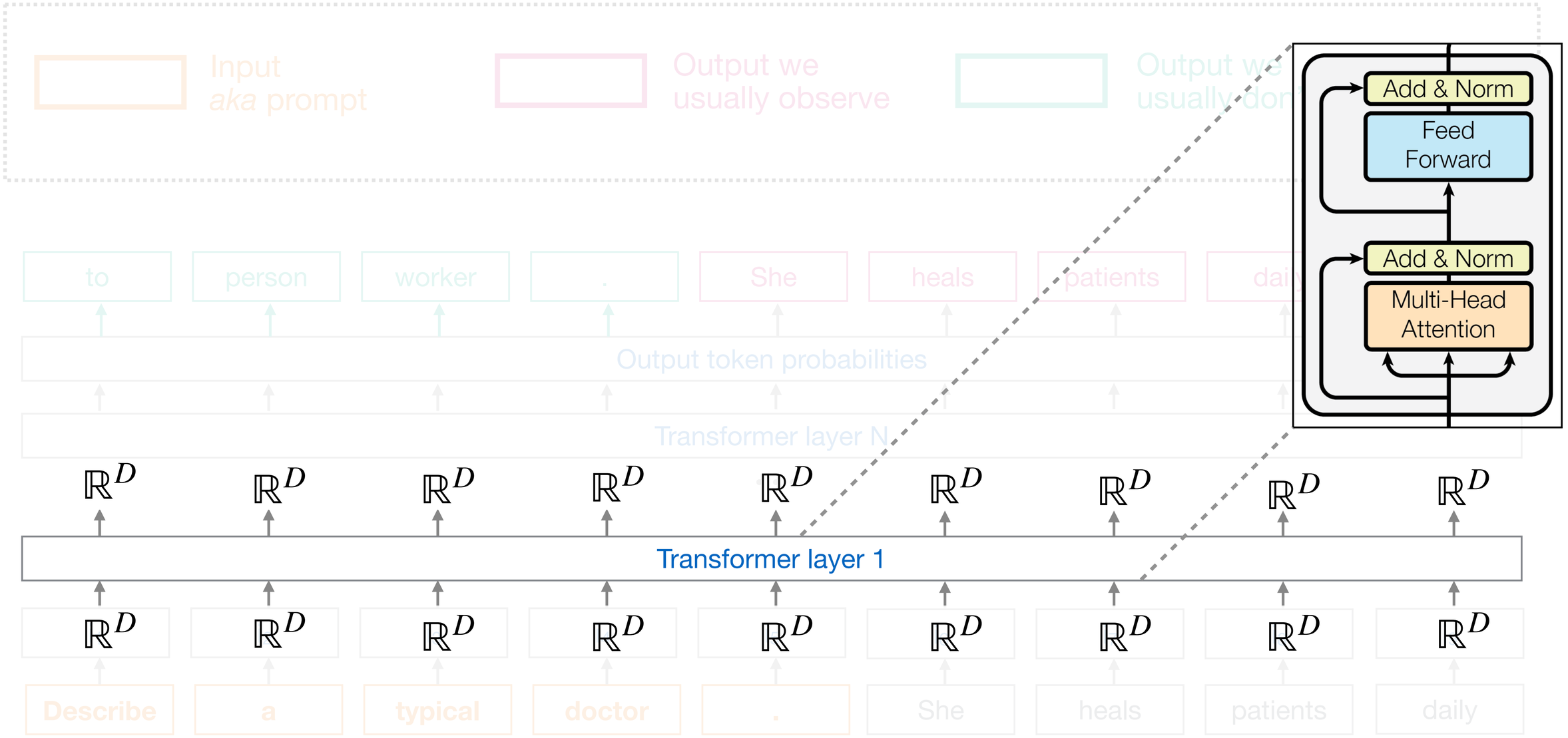
Cache the previous attention computation

Recap: What is the output dimensionality of each transformer component?



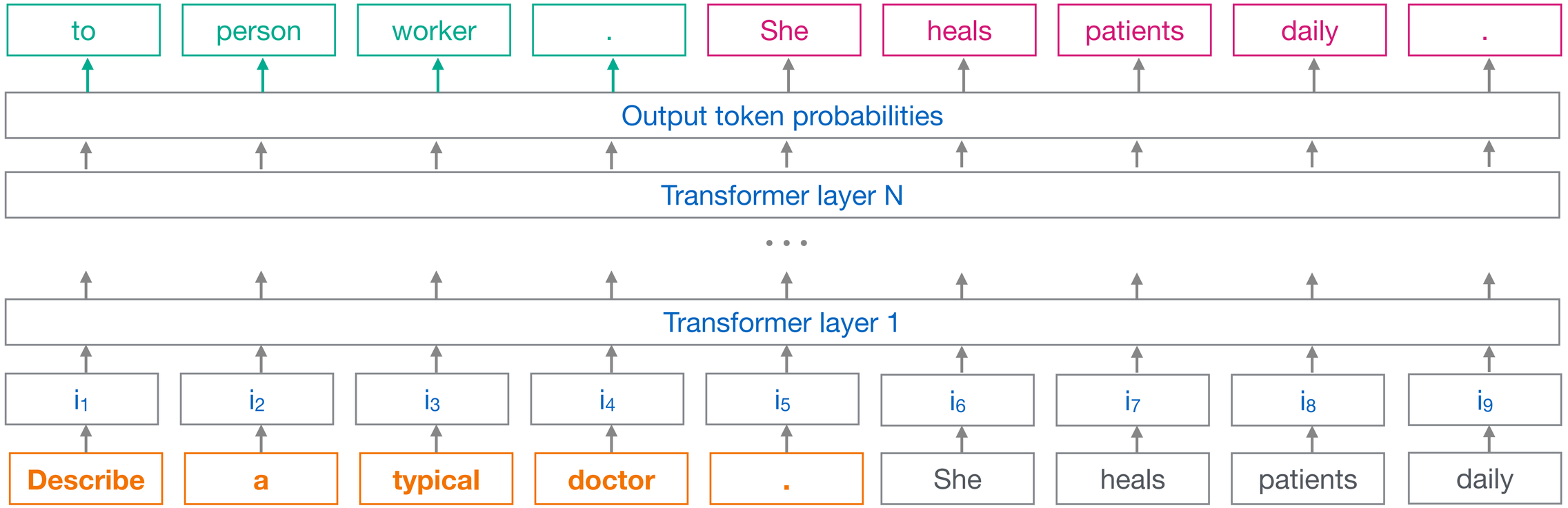
[Vaswani et al.]

Recap: What is the output dimensionality of each transformer component?



[Vaswani et al.]

Recap: What is the output dimensionality of each layer?



[Vaswani et al.]

Exercise

References

- Attention is all you need
- Mastering tensor dimensions in transformers
- The illustrated transformer