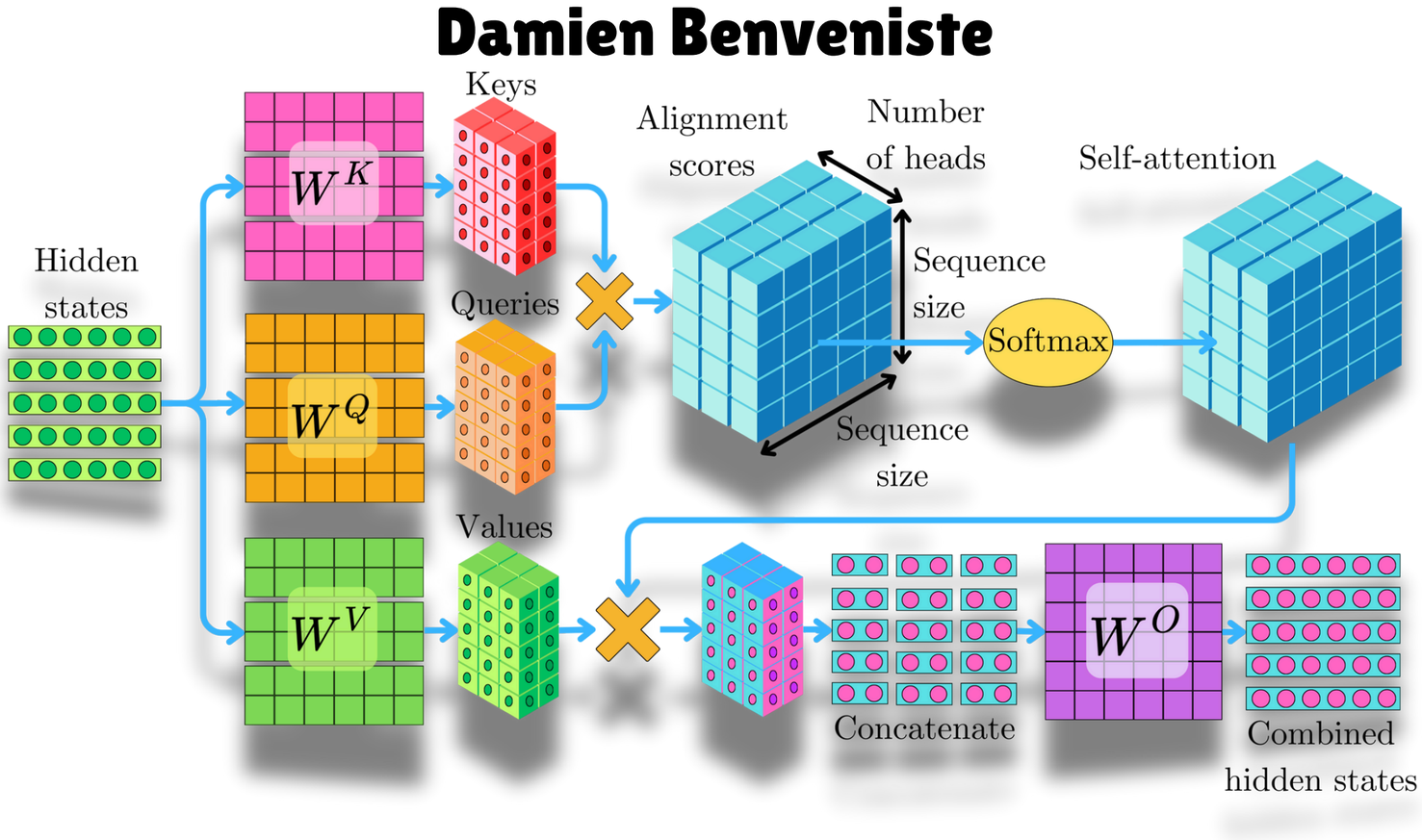
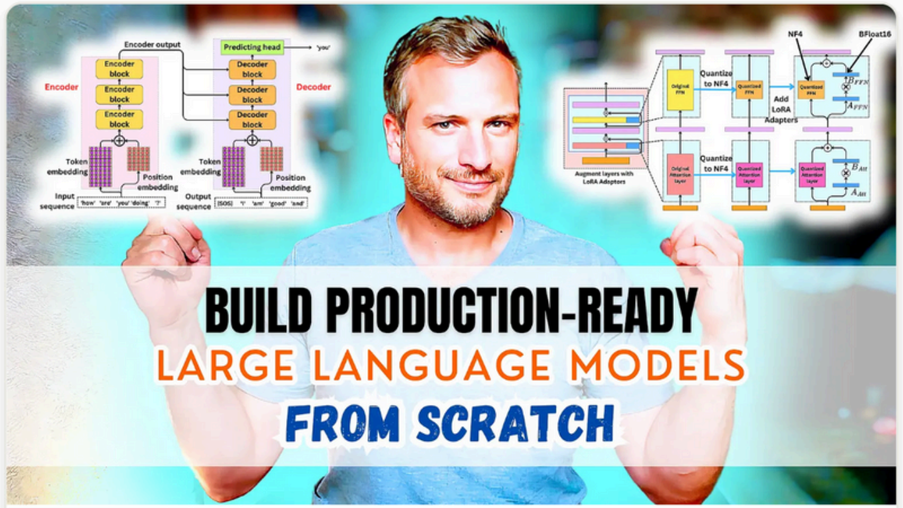
BUILD THE SELF-ATTENTION IN PYTORCH FROM SCRATCH





$1,500

NEXT COHORT

May 24—June 29, 2025

Apply

GET FUTURE COHORT DATES

damien.benveniste@gmail.com

Get reimbursed Bulk purchases

Build Production-Ready LLMs From Scratch

Starts May 24th

A 6-week, live bootcamp for ML engineers to architect, fine-tune, and deploy scalable LLM applications through six real-world projects:

- 6 weeks
- 12 live sessions
- 6 hands-on projects
- 64 recorded lectures

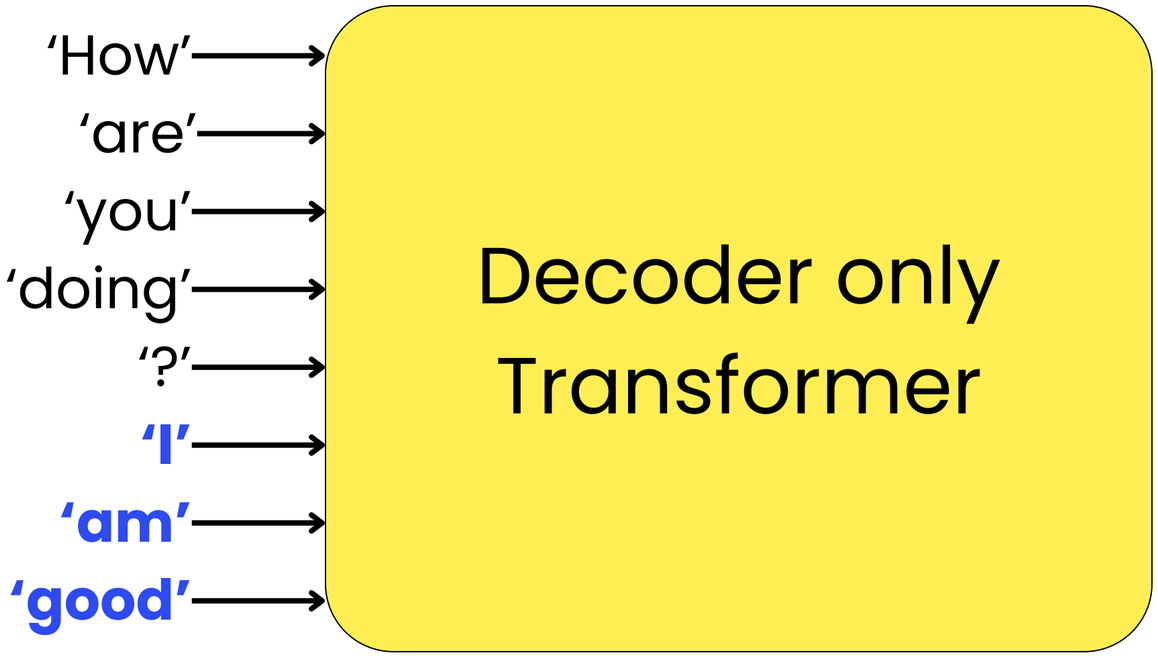
Special Offer:
20% off with the coupon code *20off*

Some Context: The Transformer Architecture

Decoder only
Transformer

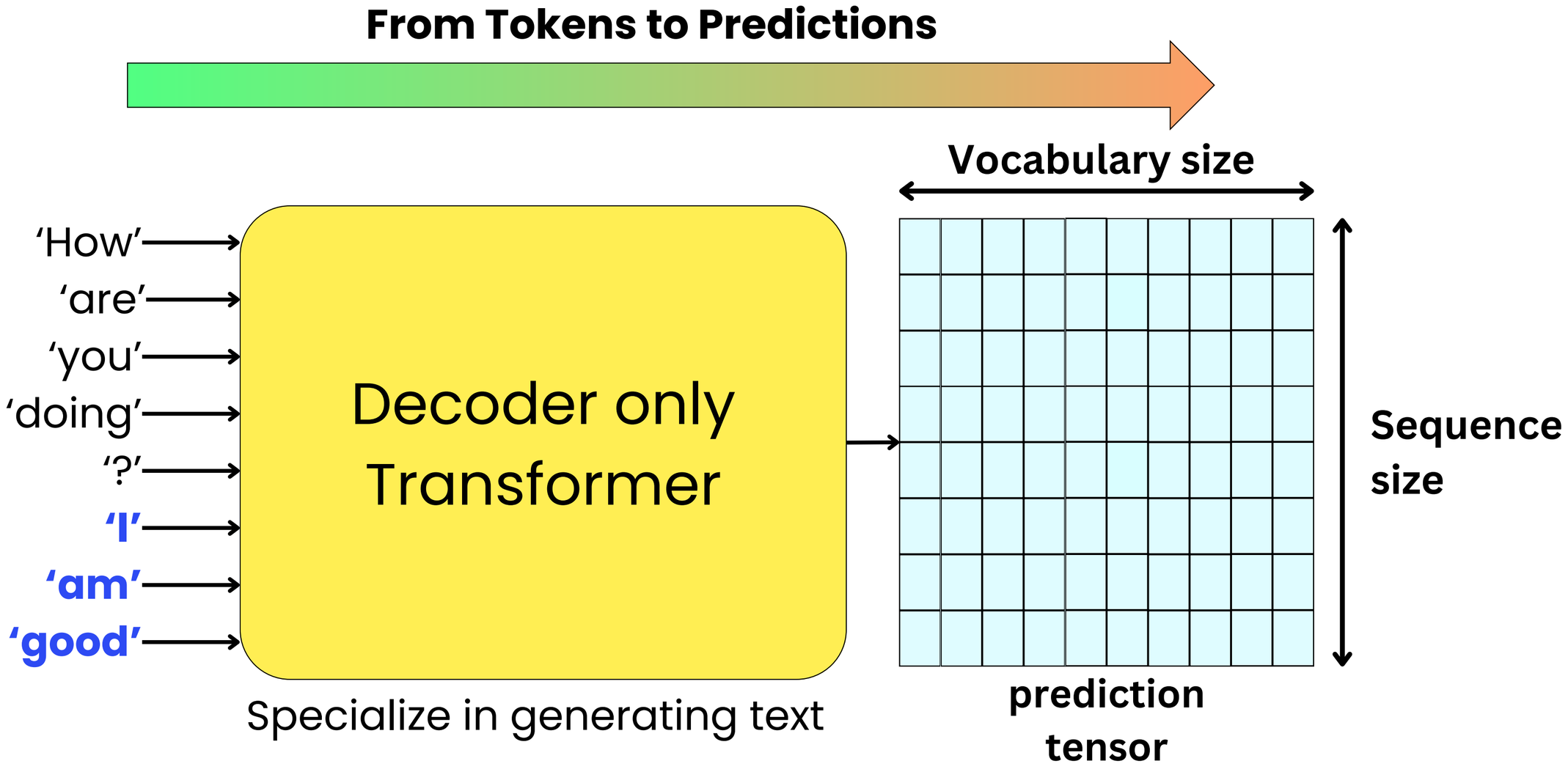
Specialize in generating text

Some Context: The Transformer Architecture

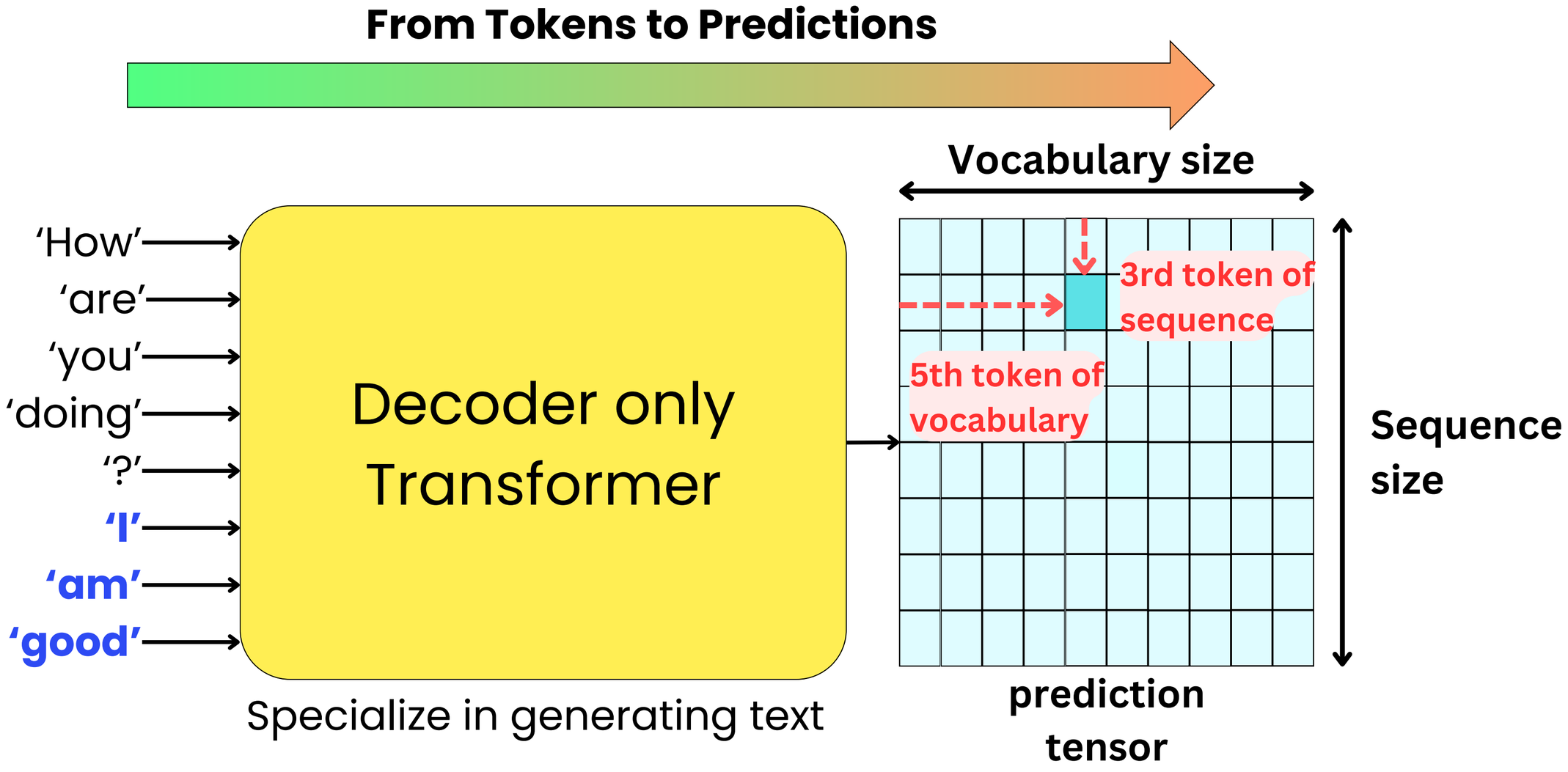


Specialize in generating text

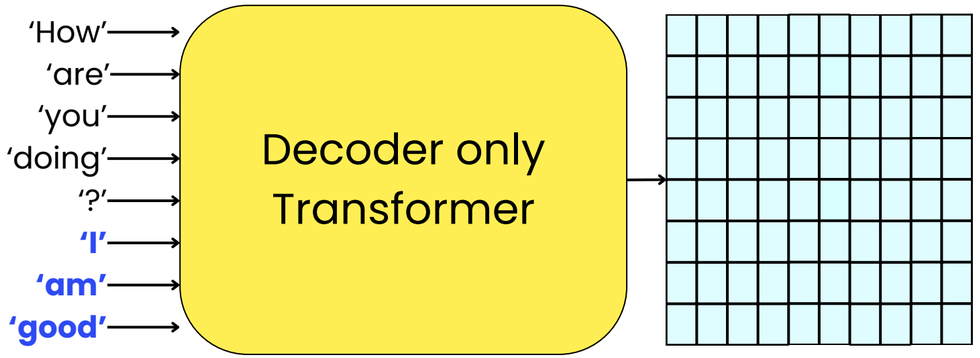
Some Context: The Transformer Architecture



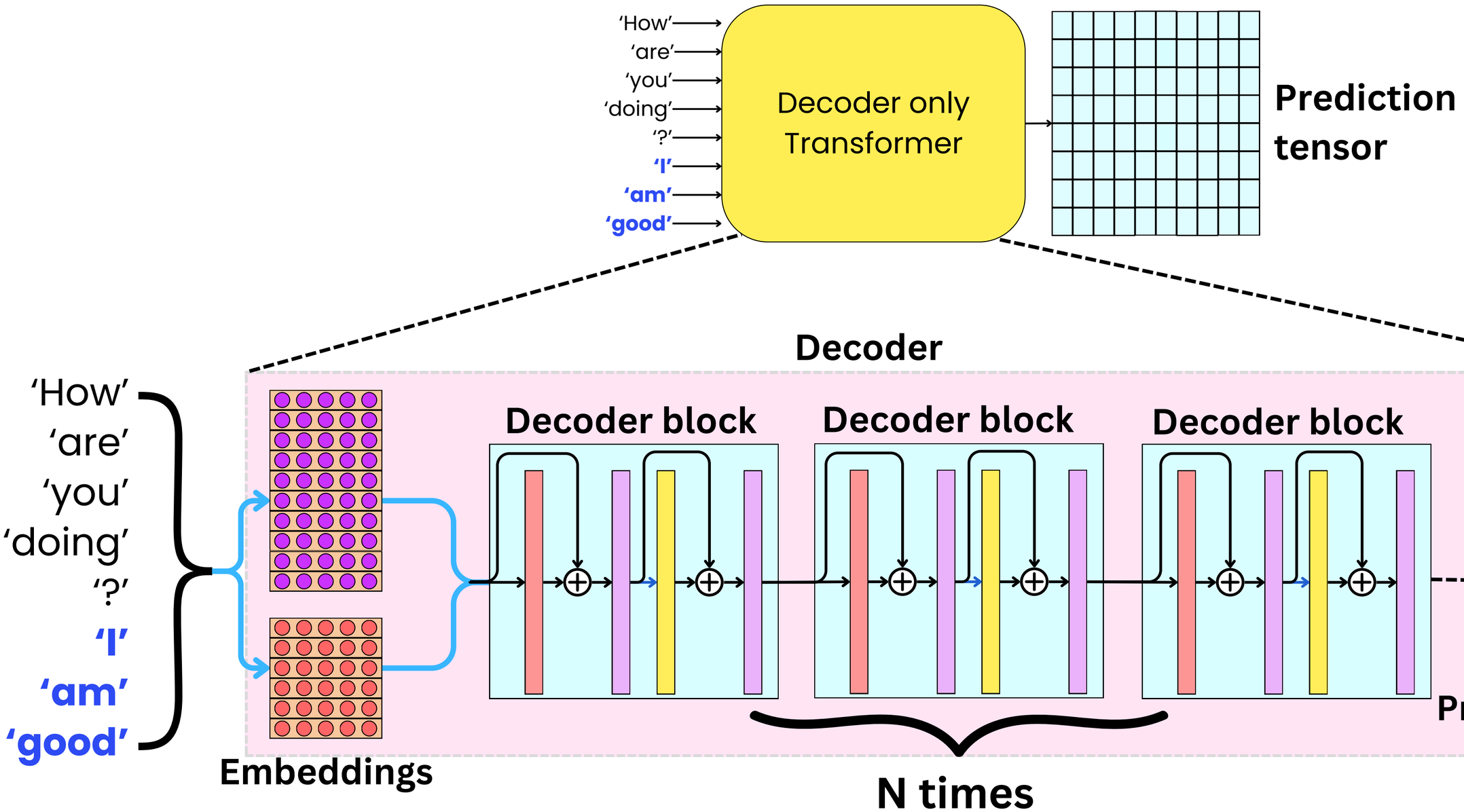
Some Context: The Transformer Architecture



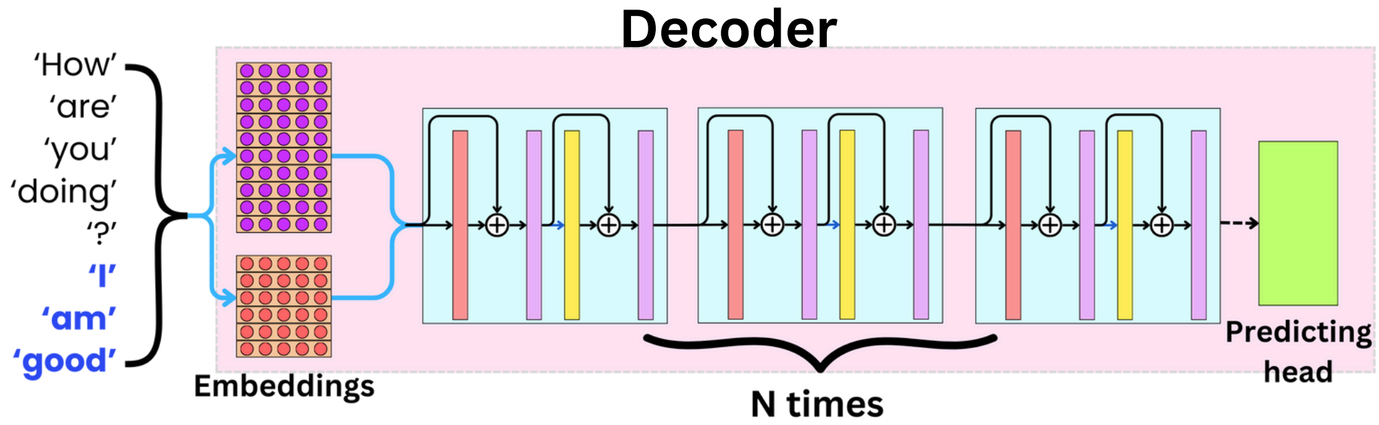
Some Context: The Transformer Architecture



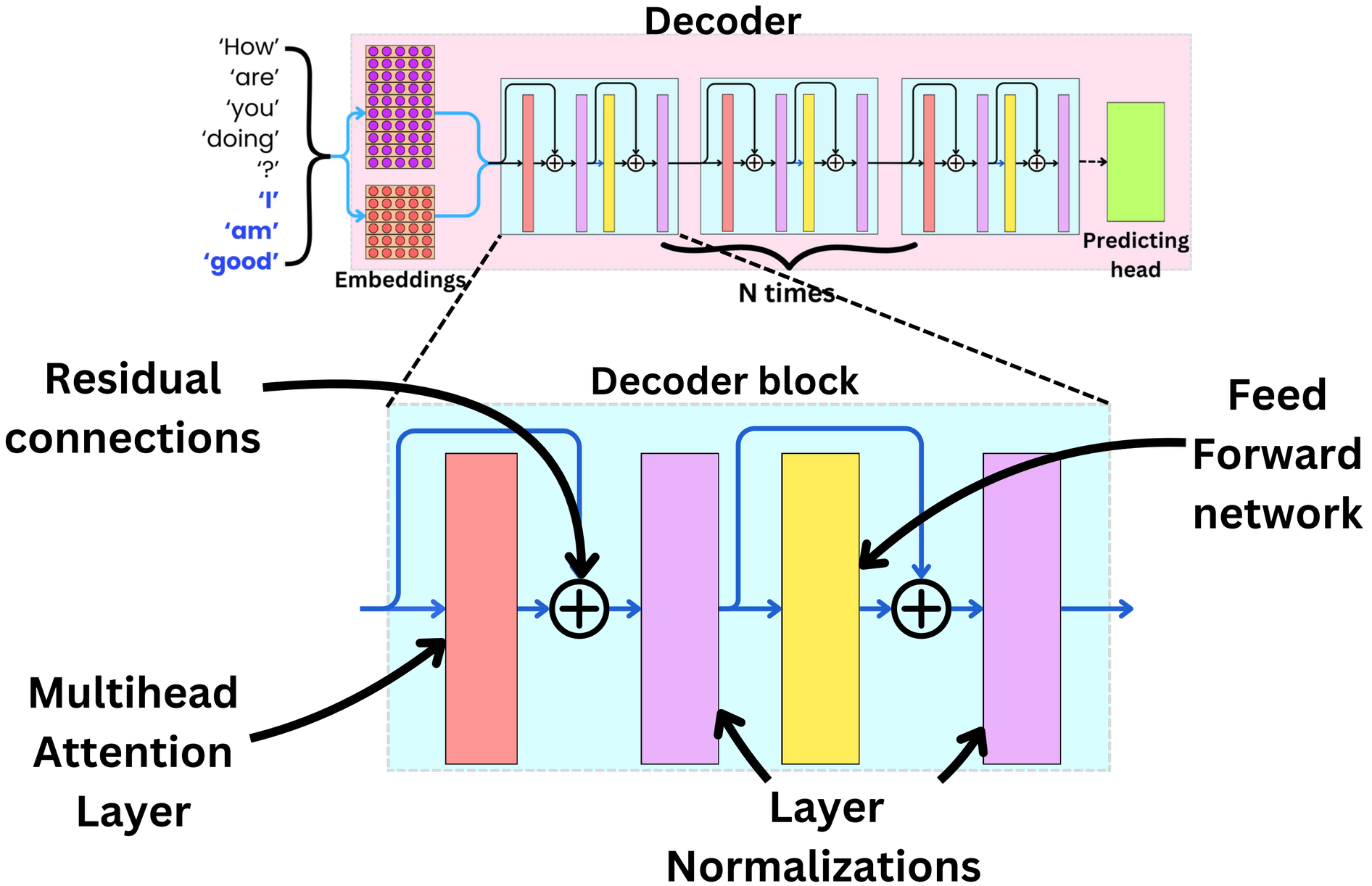
Some Context: The Transformer Architecture



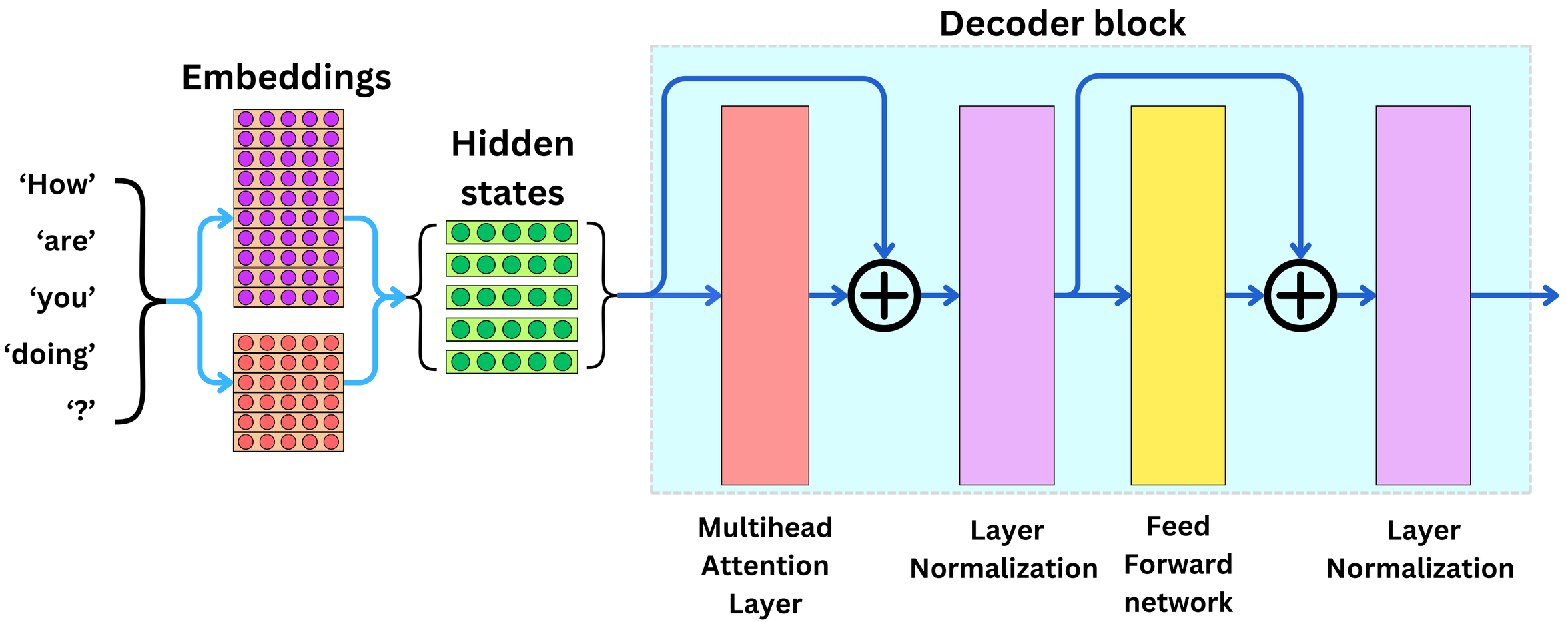
Some Context: The Transformer Architecture



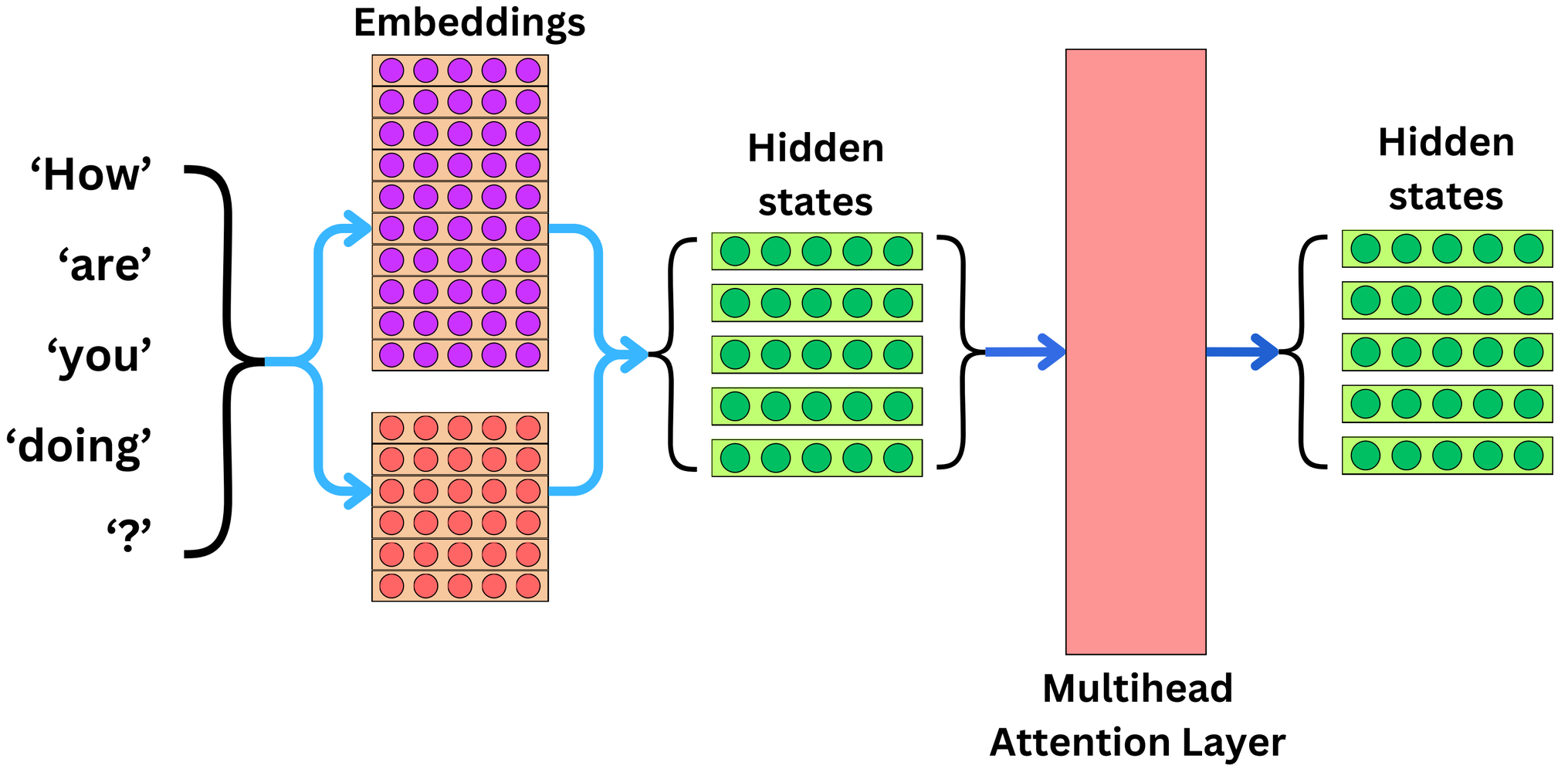
Some Context: The Transformer Architecture



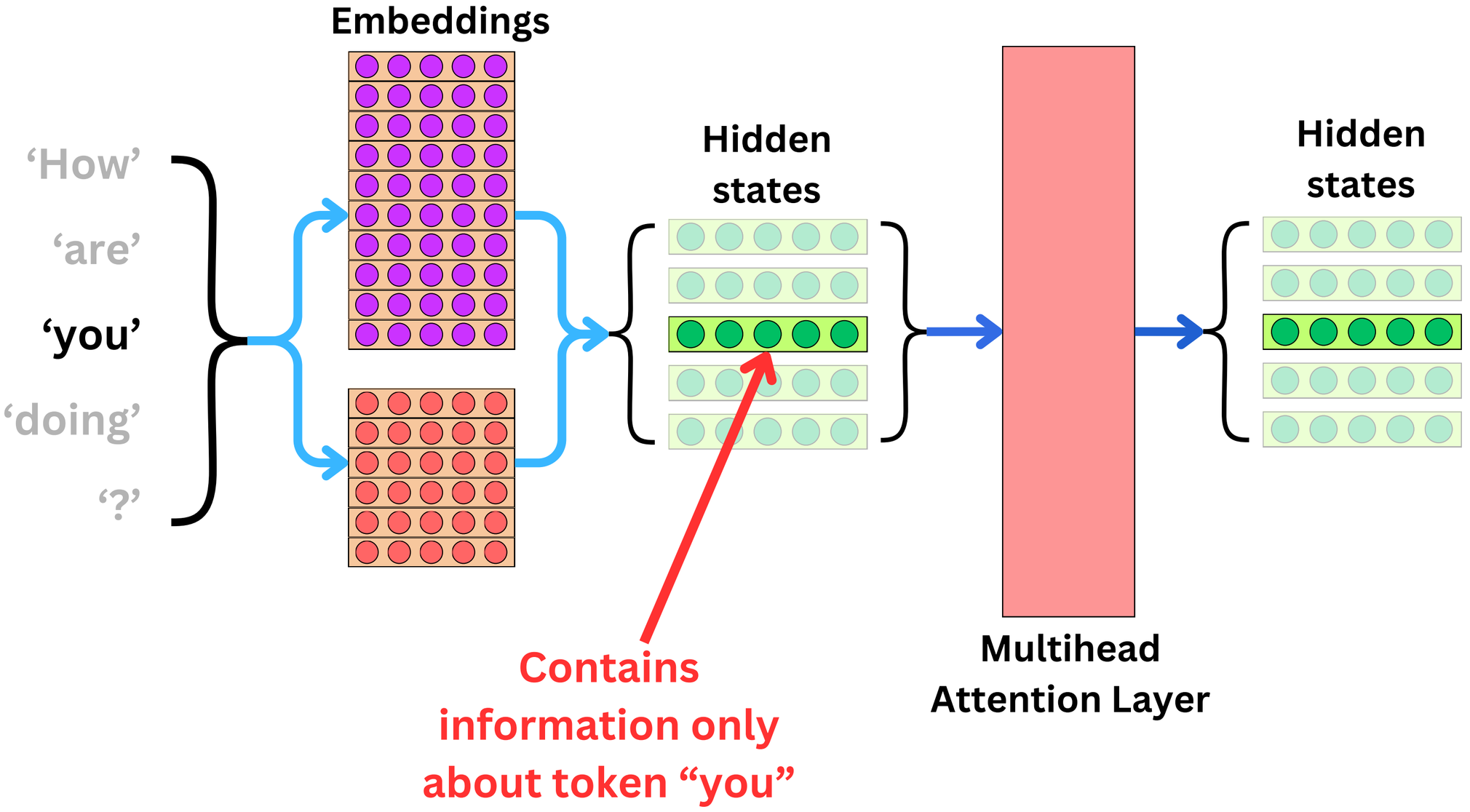
The Self-Attention Layer



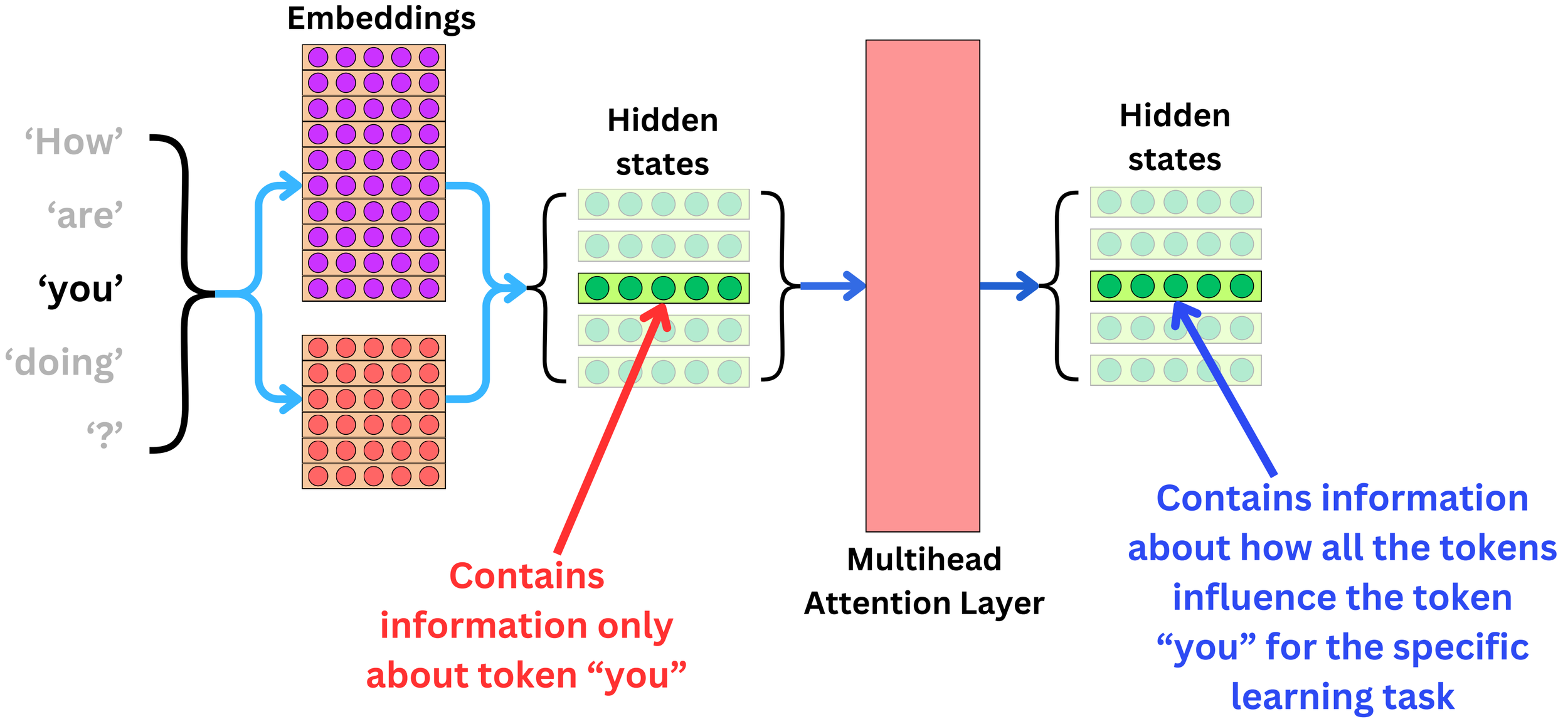
The Self-Attention Layer



The Self-Attention Layer



The Self-Attention Layer



The Self-Attention Layer

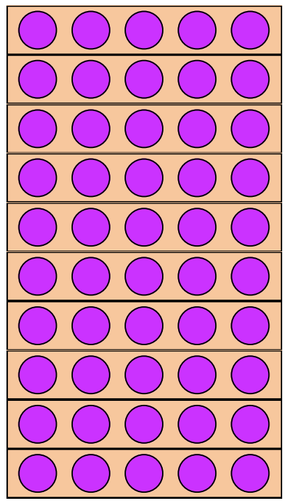
‘How’

‘are’

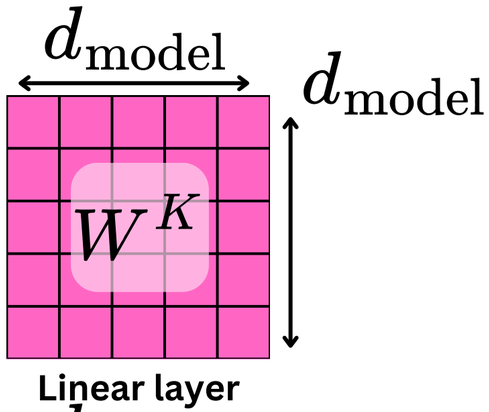
‘you’

‘doing’

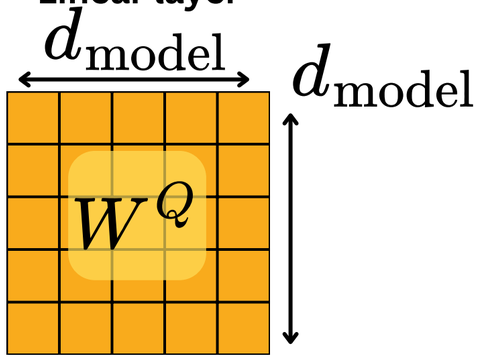
‘?’



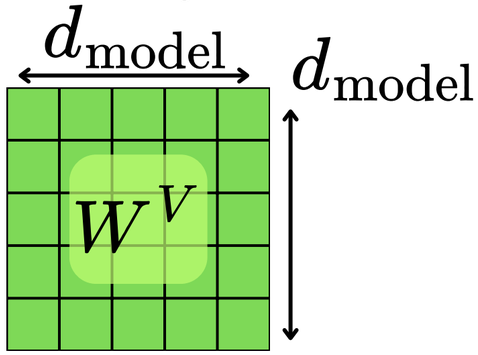
Word / token embedding



Linear layer

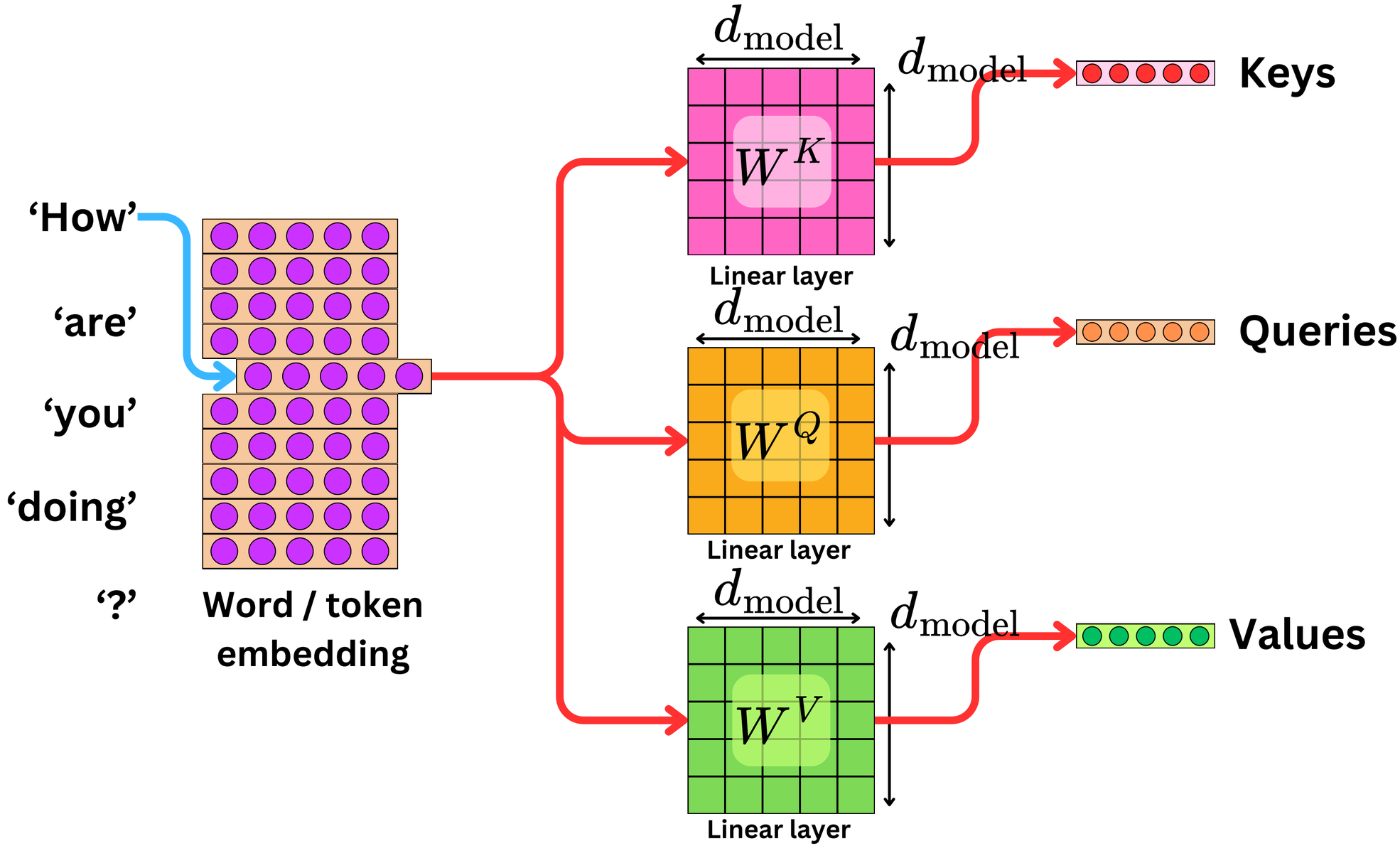


Linear layer

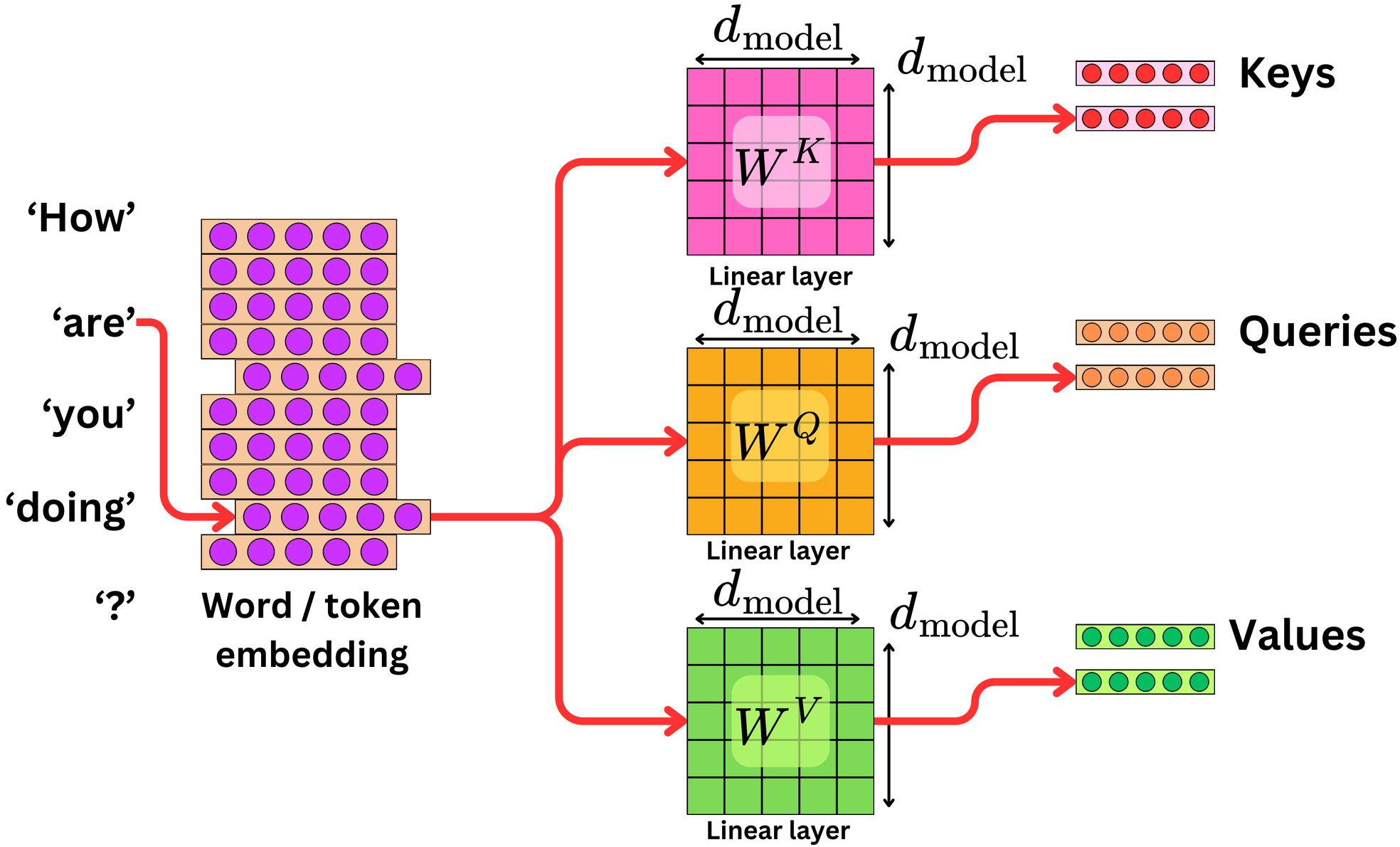


Linear layer

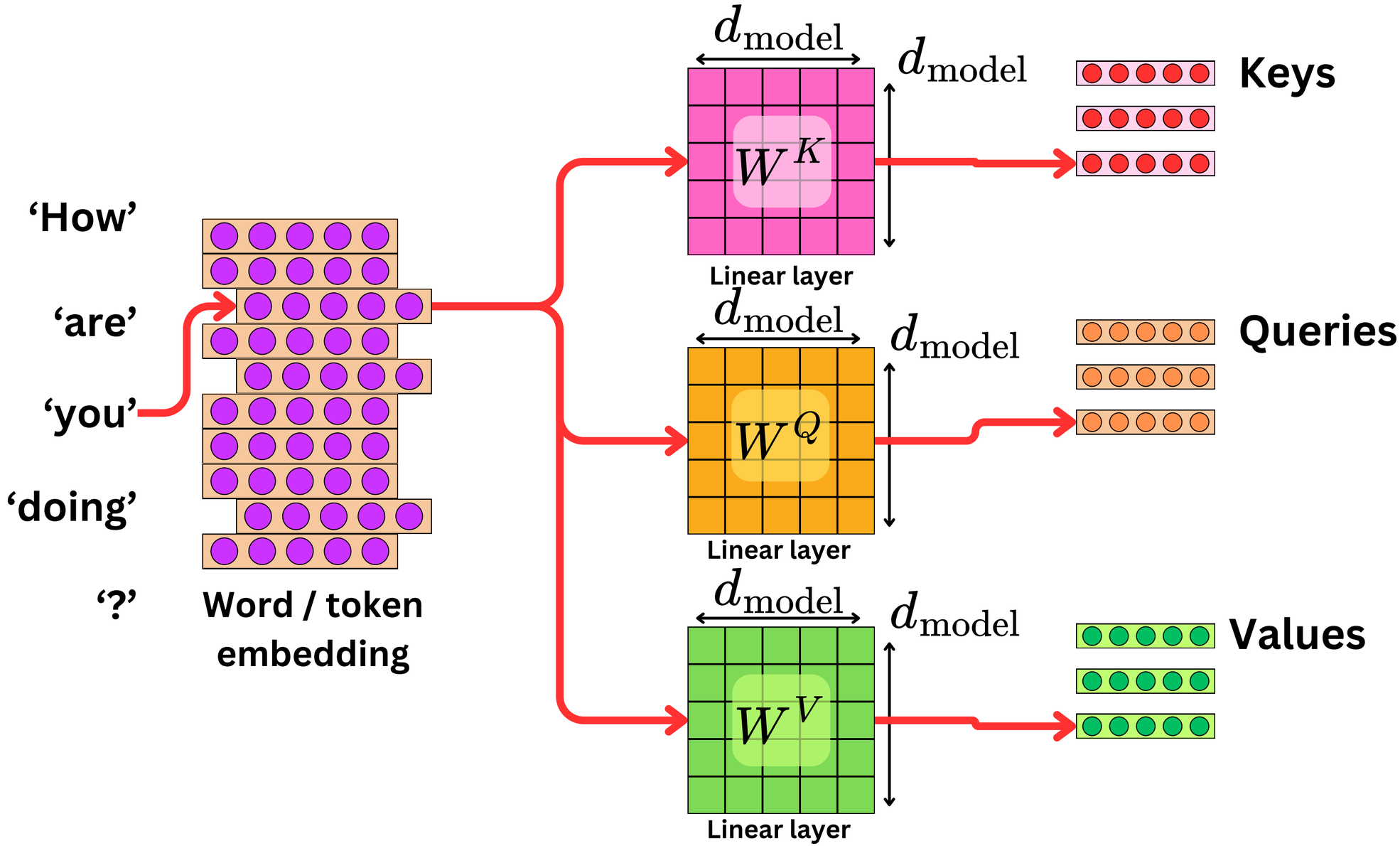
The Self-Attention Layer



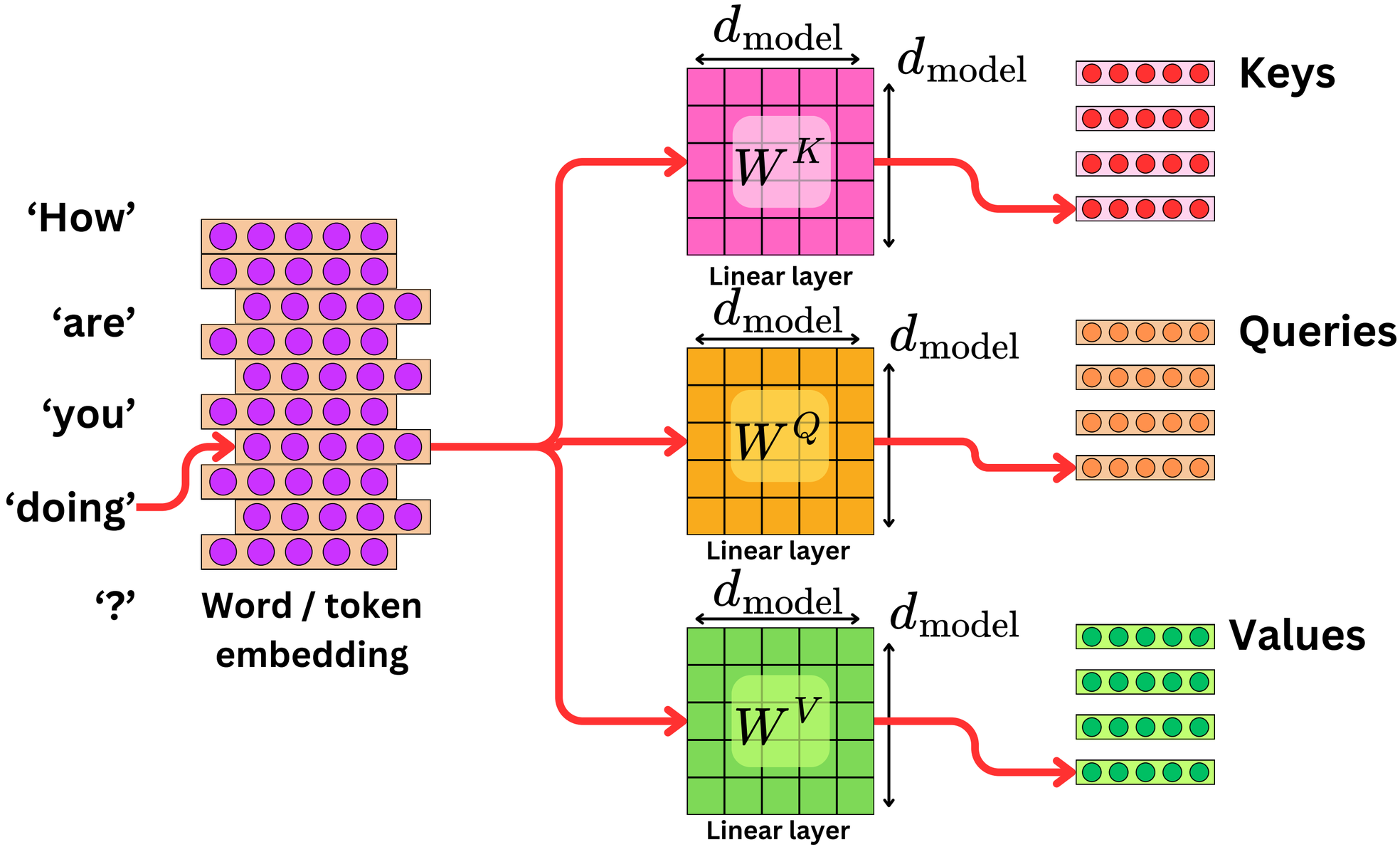
The Self-Attention Layer



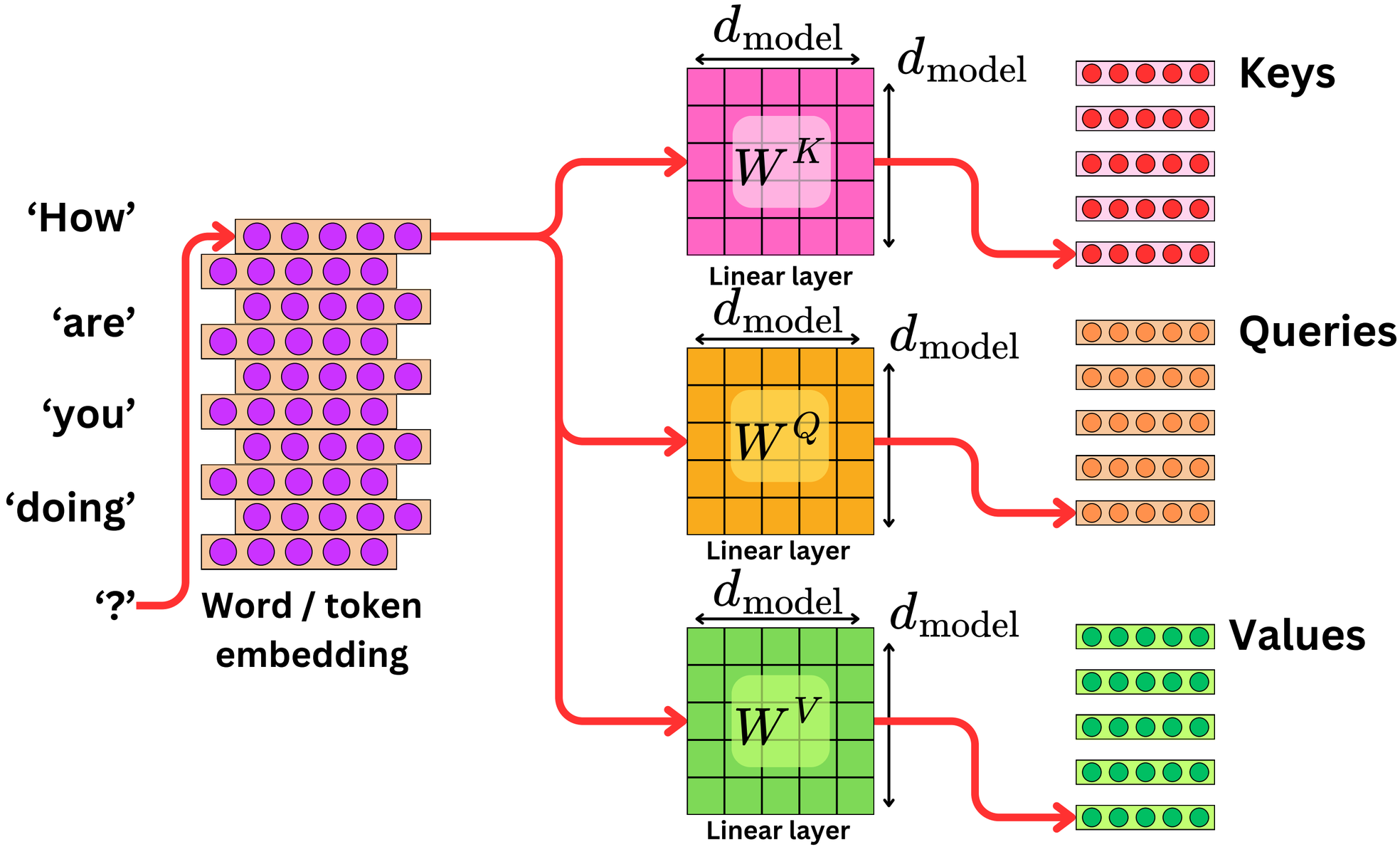
The Self-Attention Layer



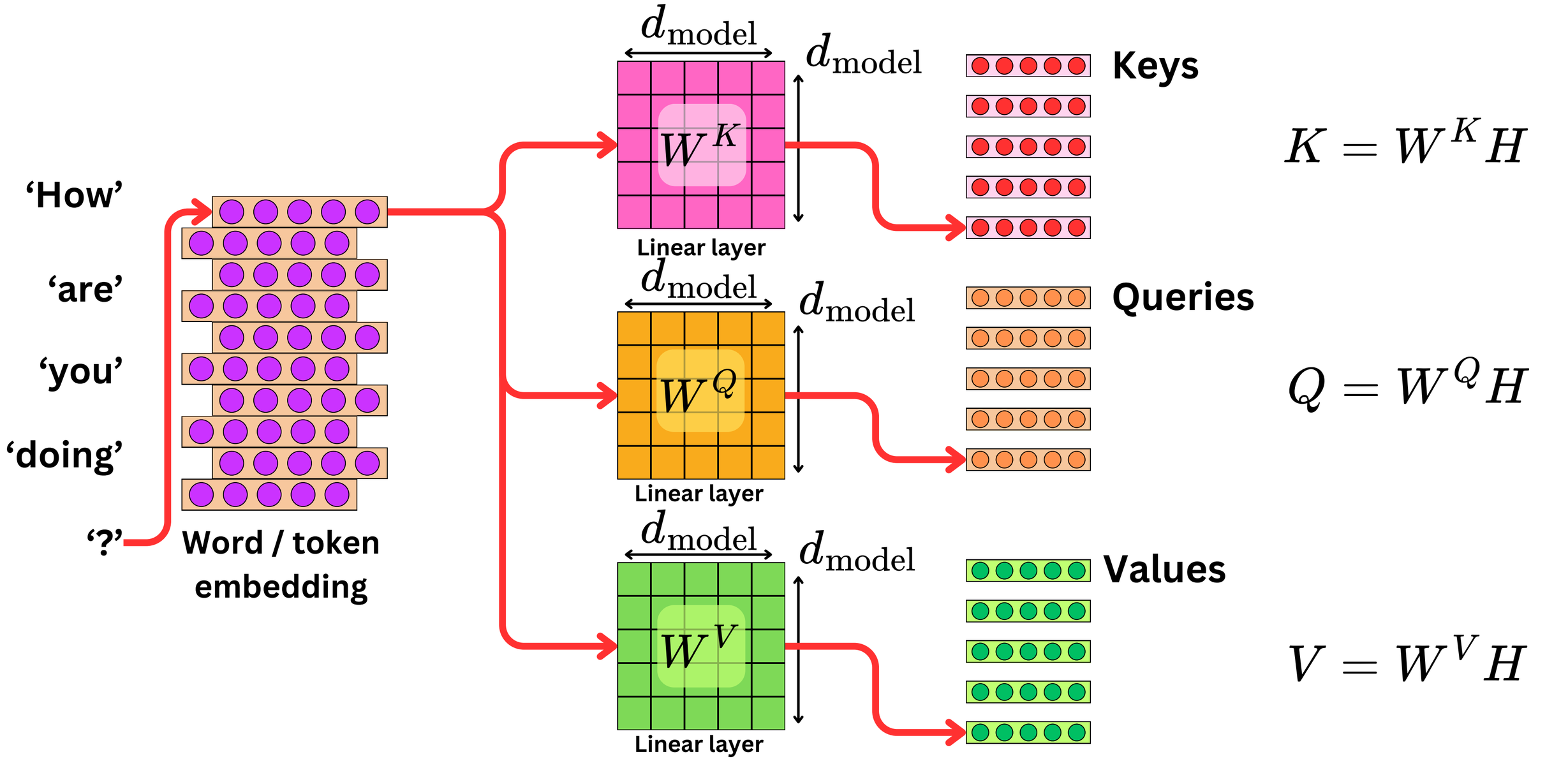
The Self-Attention Layer



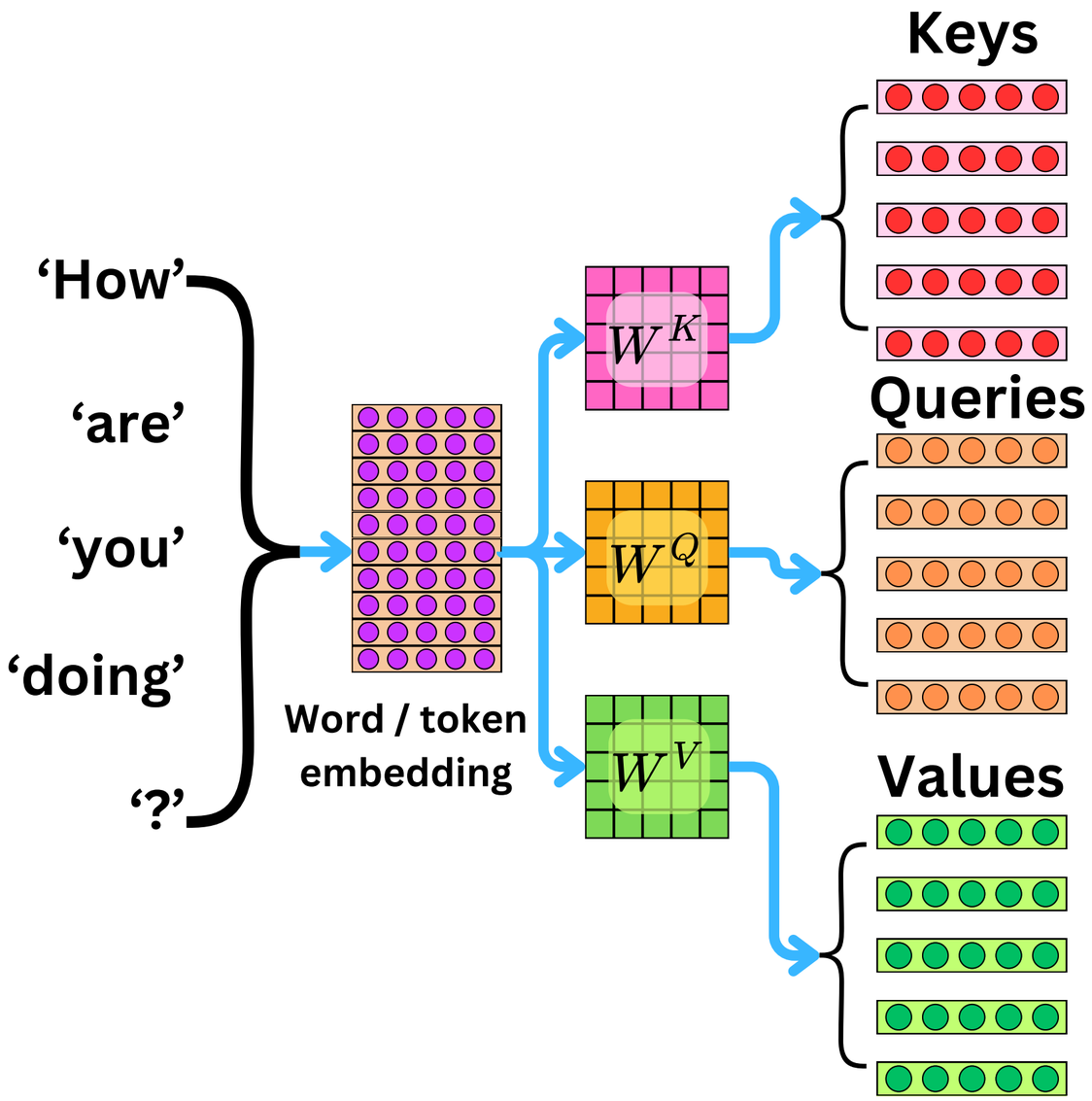
The Self-Attention Layer



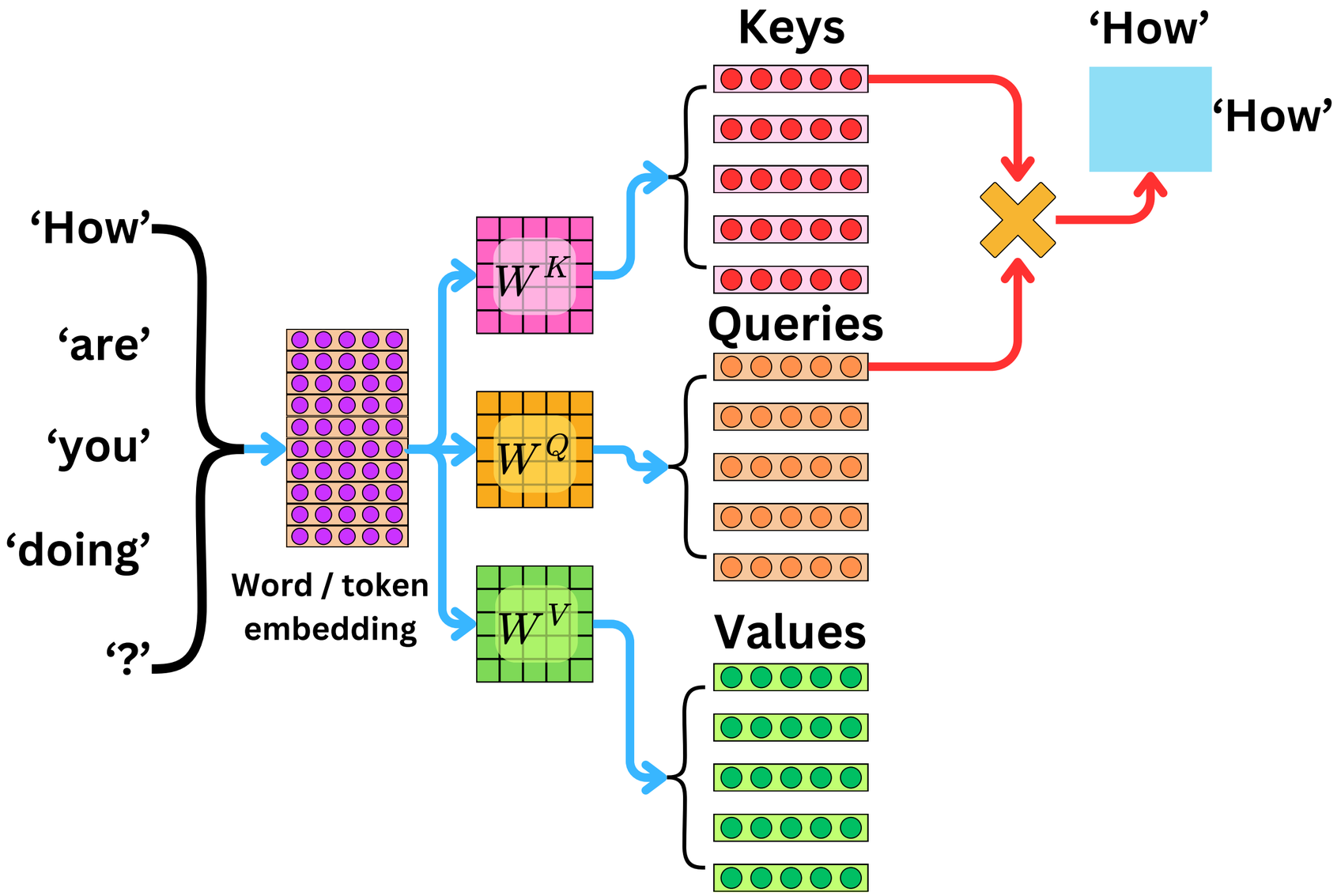
The Self-Attention Layer



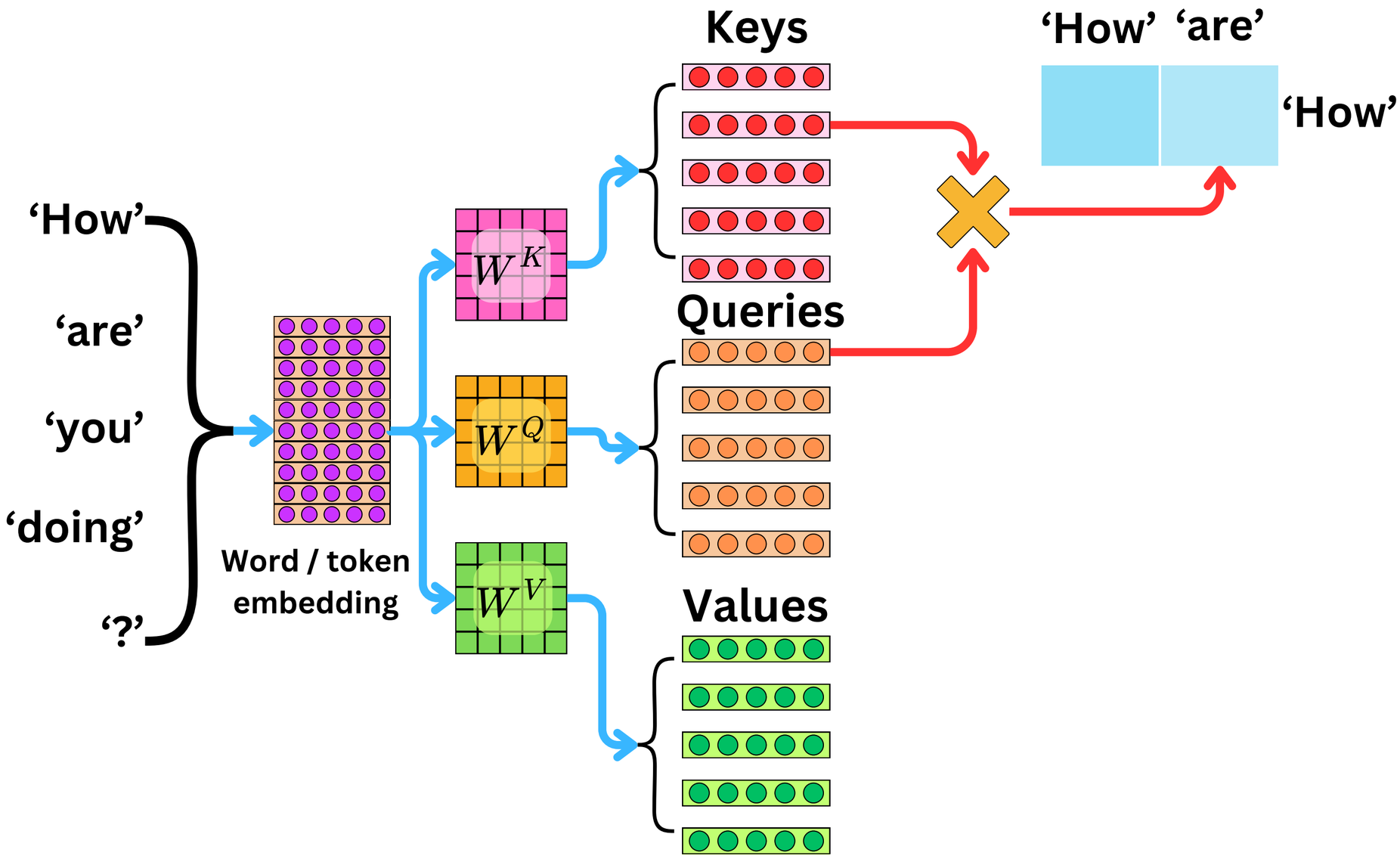
The Self-Attention Layer



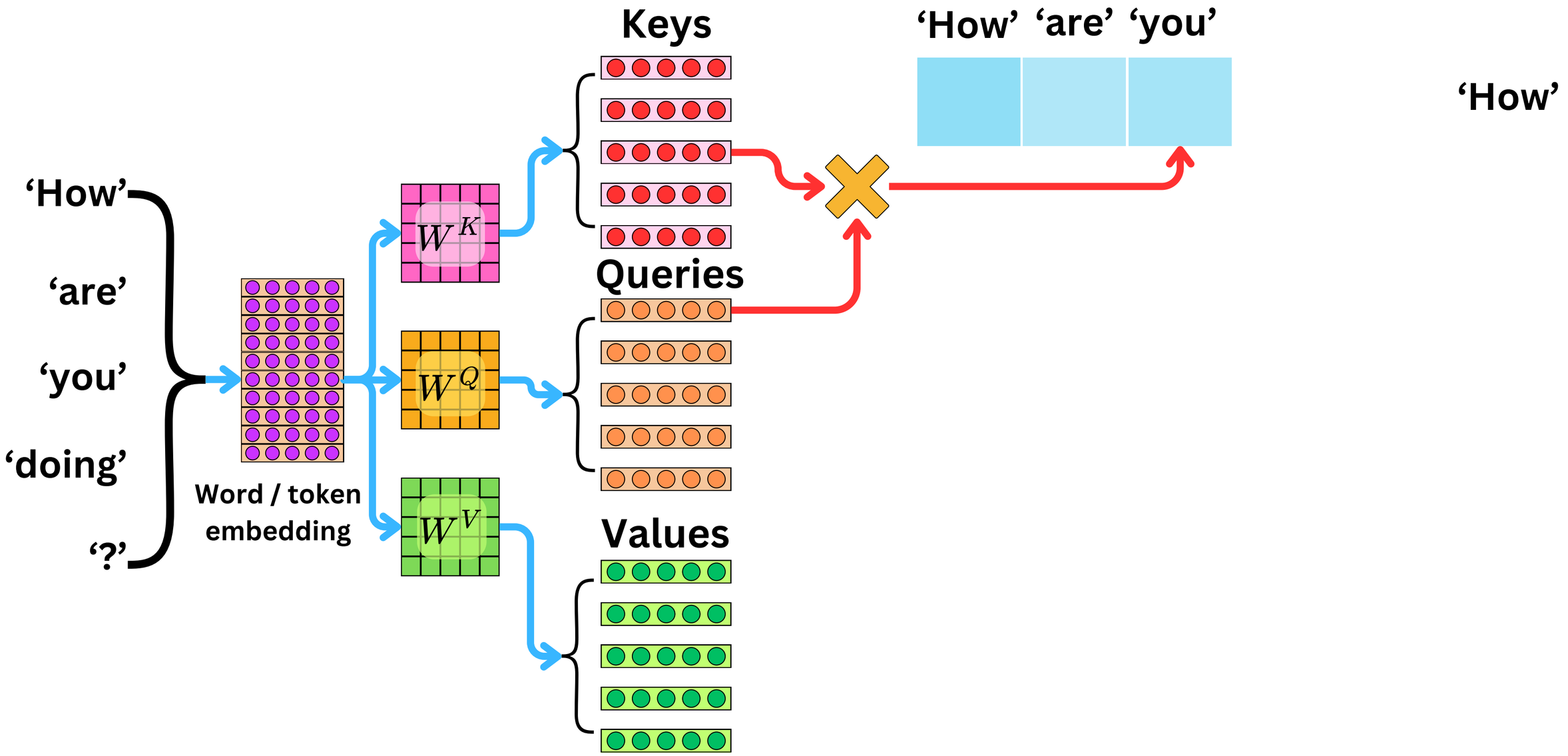
The Self-Attention Layer



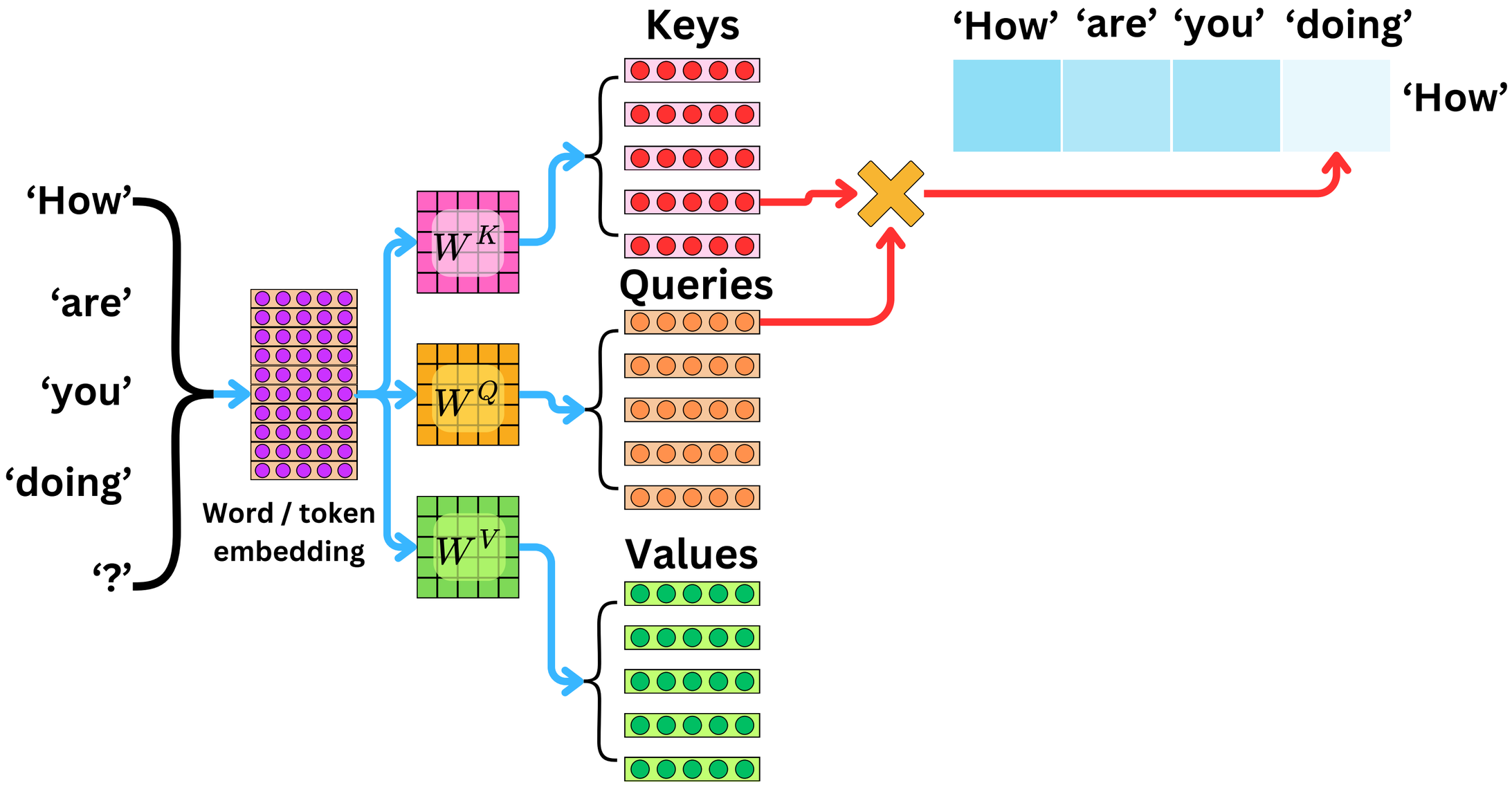
The Self-Attention Layer



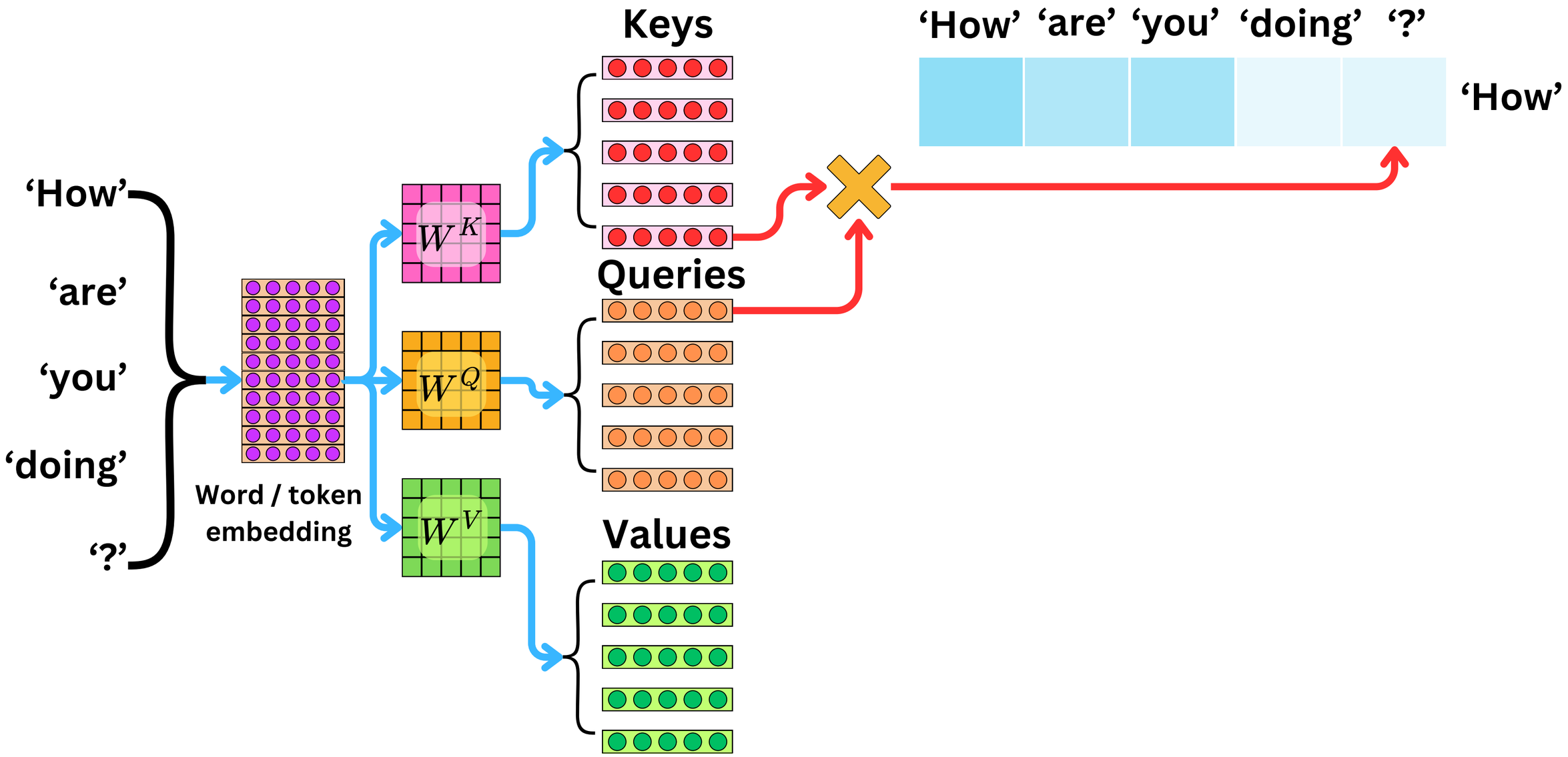
The Self-Attention Layer



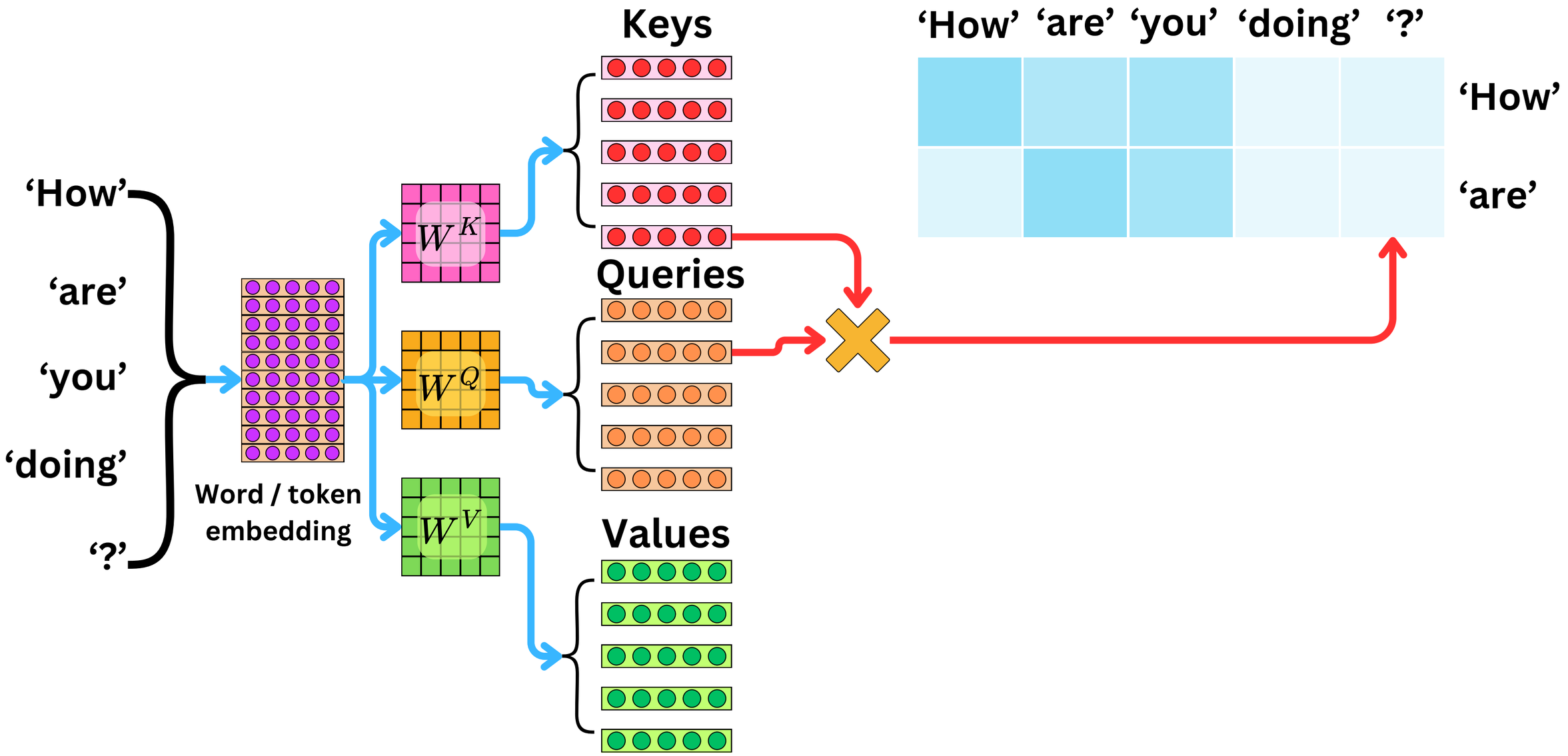
The Self-Attention Layer



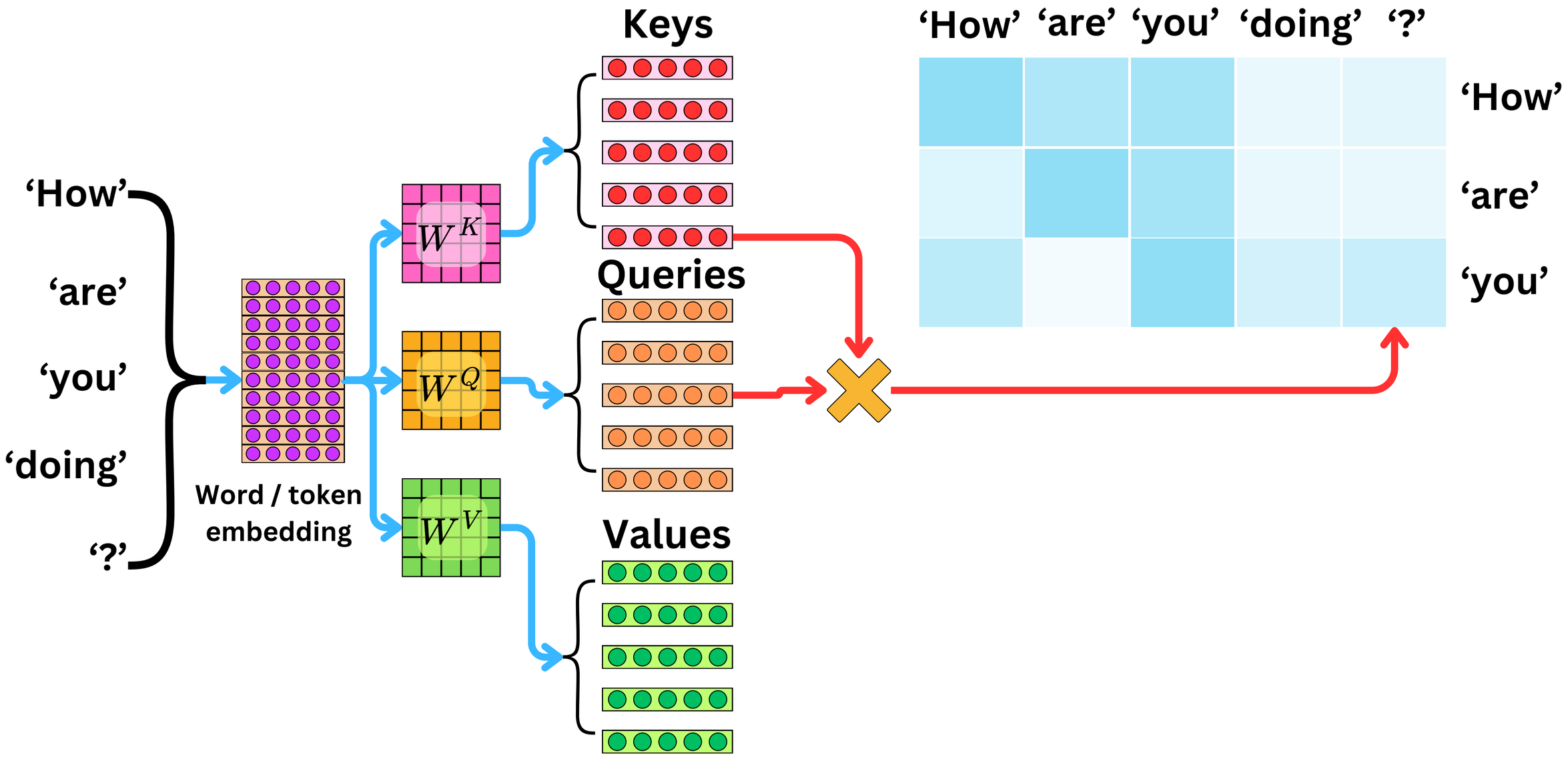
The Self-Attention Layer



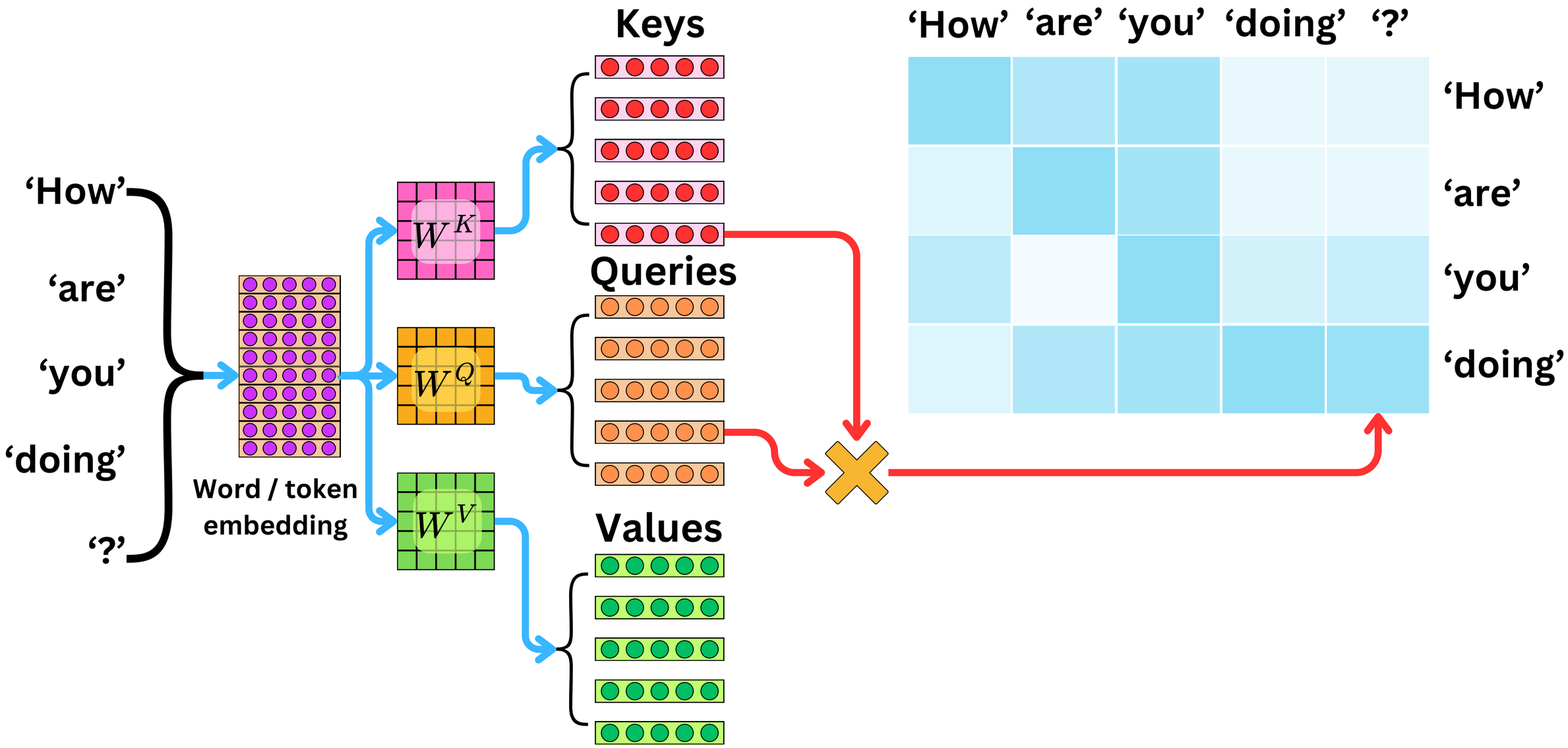
The Self-Attention Layer



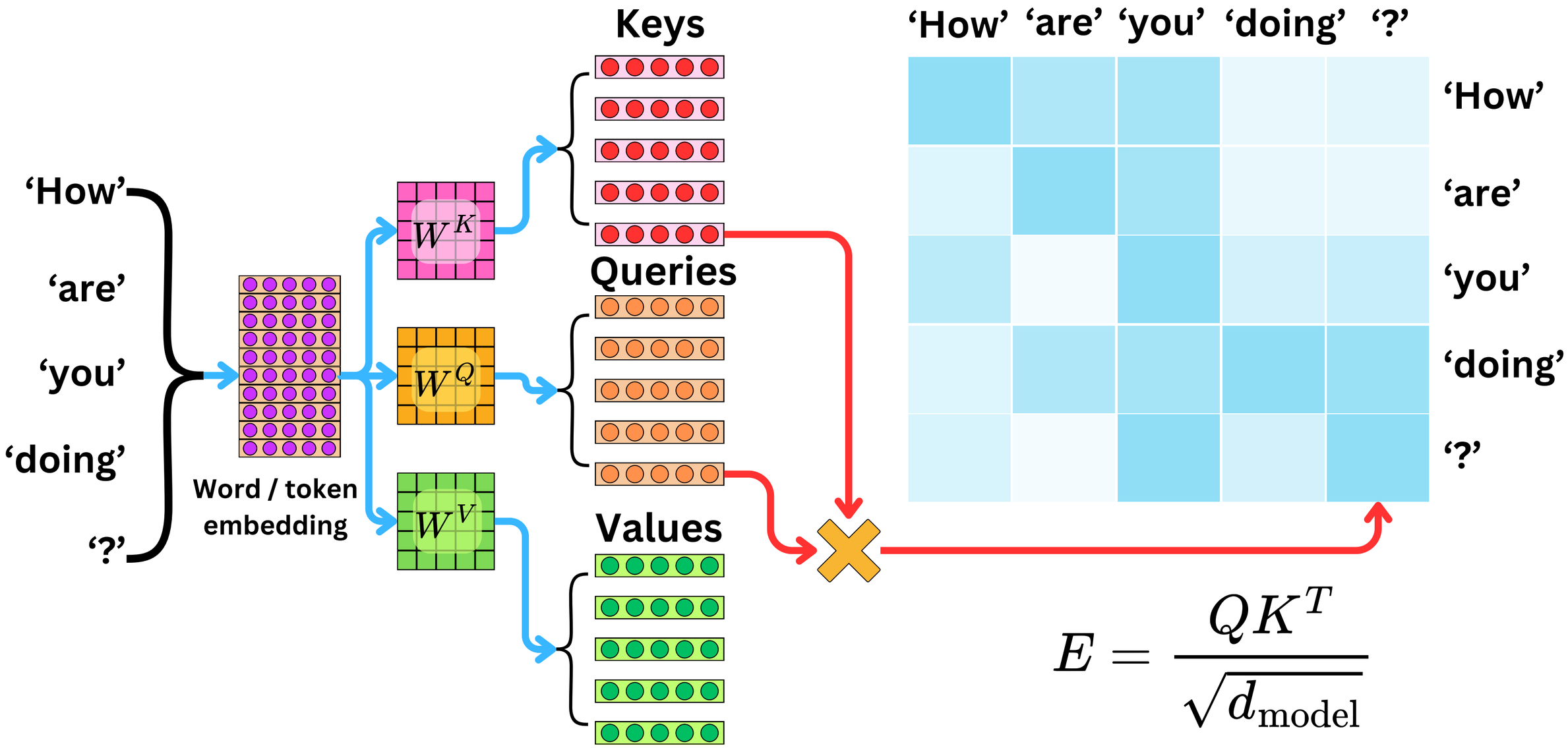
The Self-Attention Layer



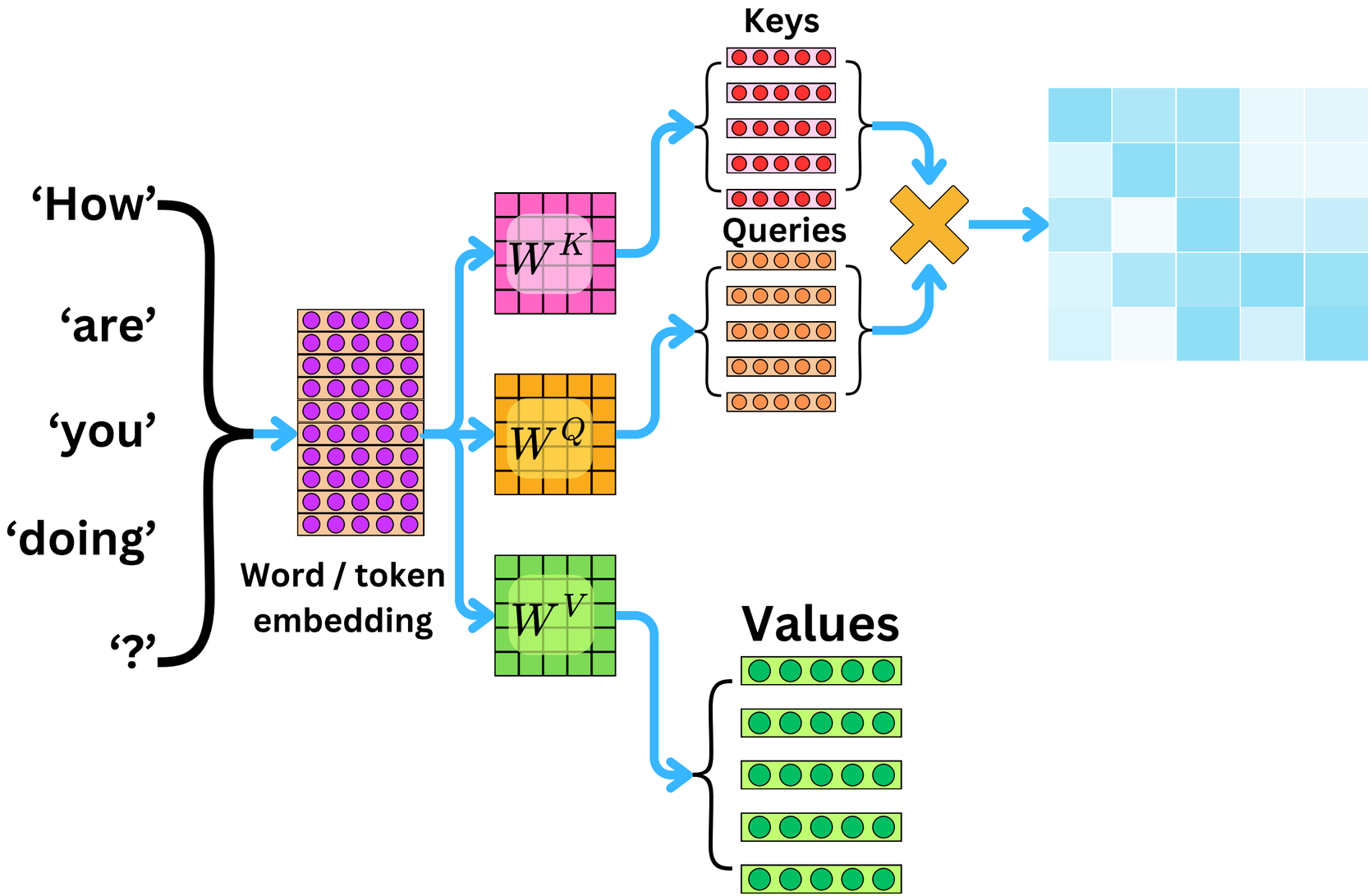
The Self-Attention Layer



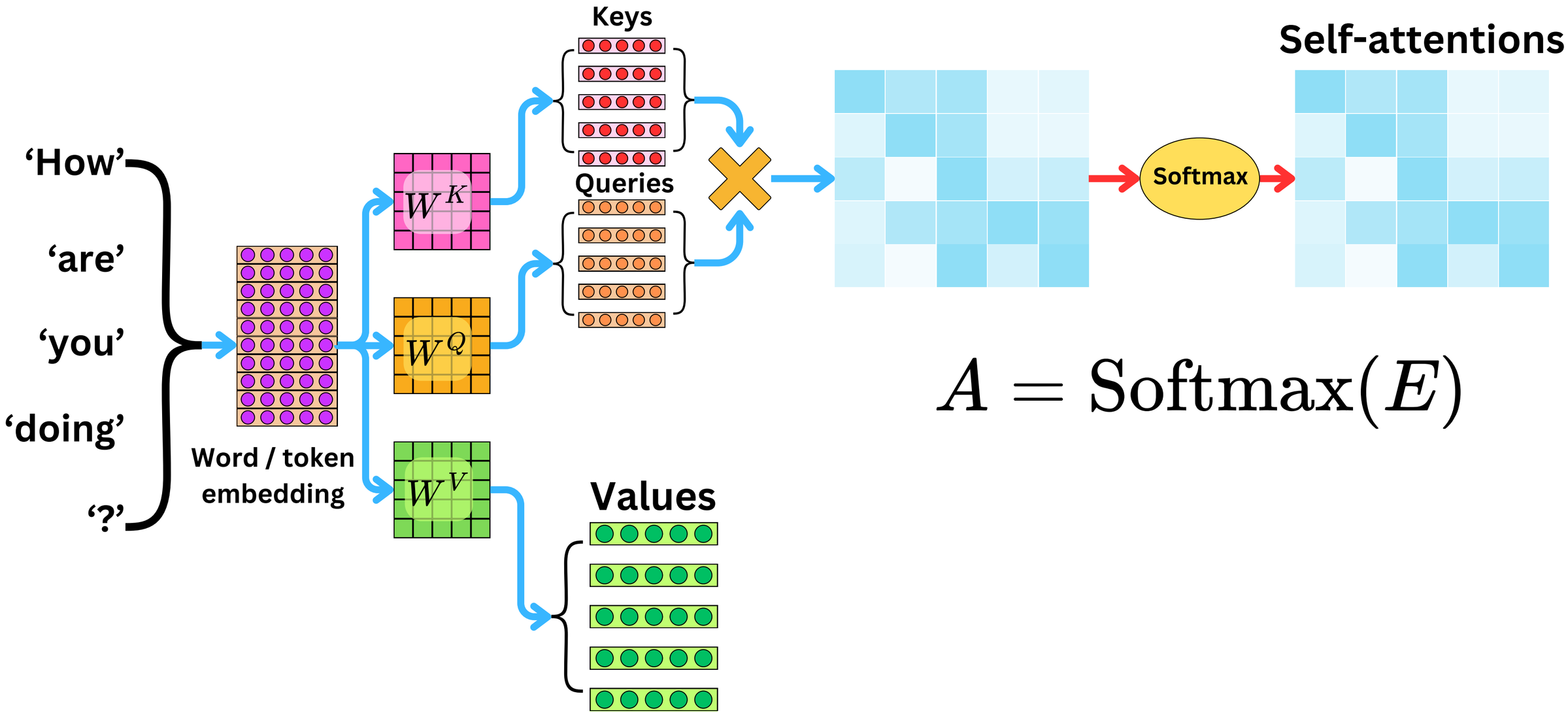
The Self-Attention Layer



The Self-Attention Layer

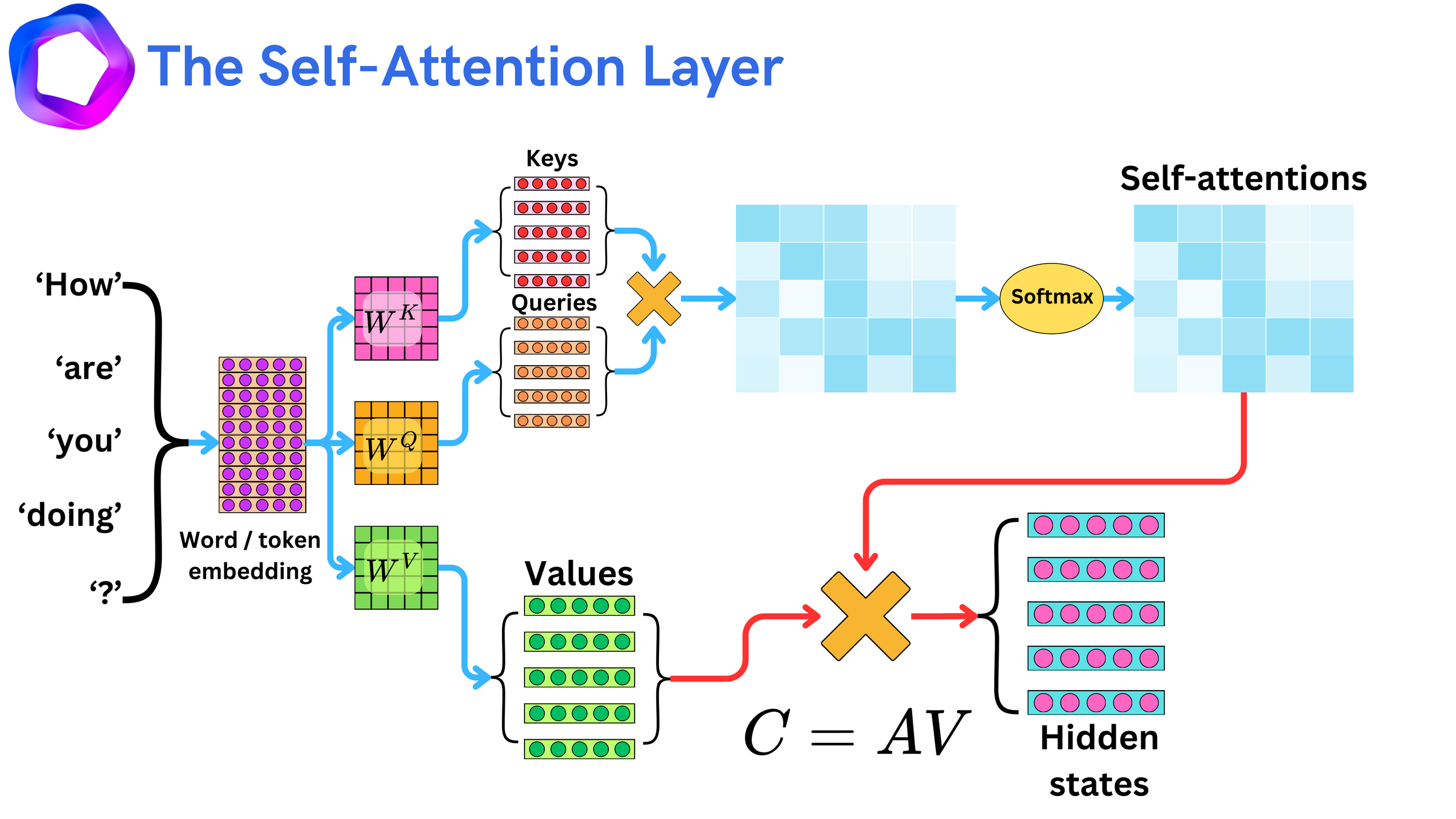


The Self-Attention Layer

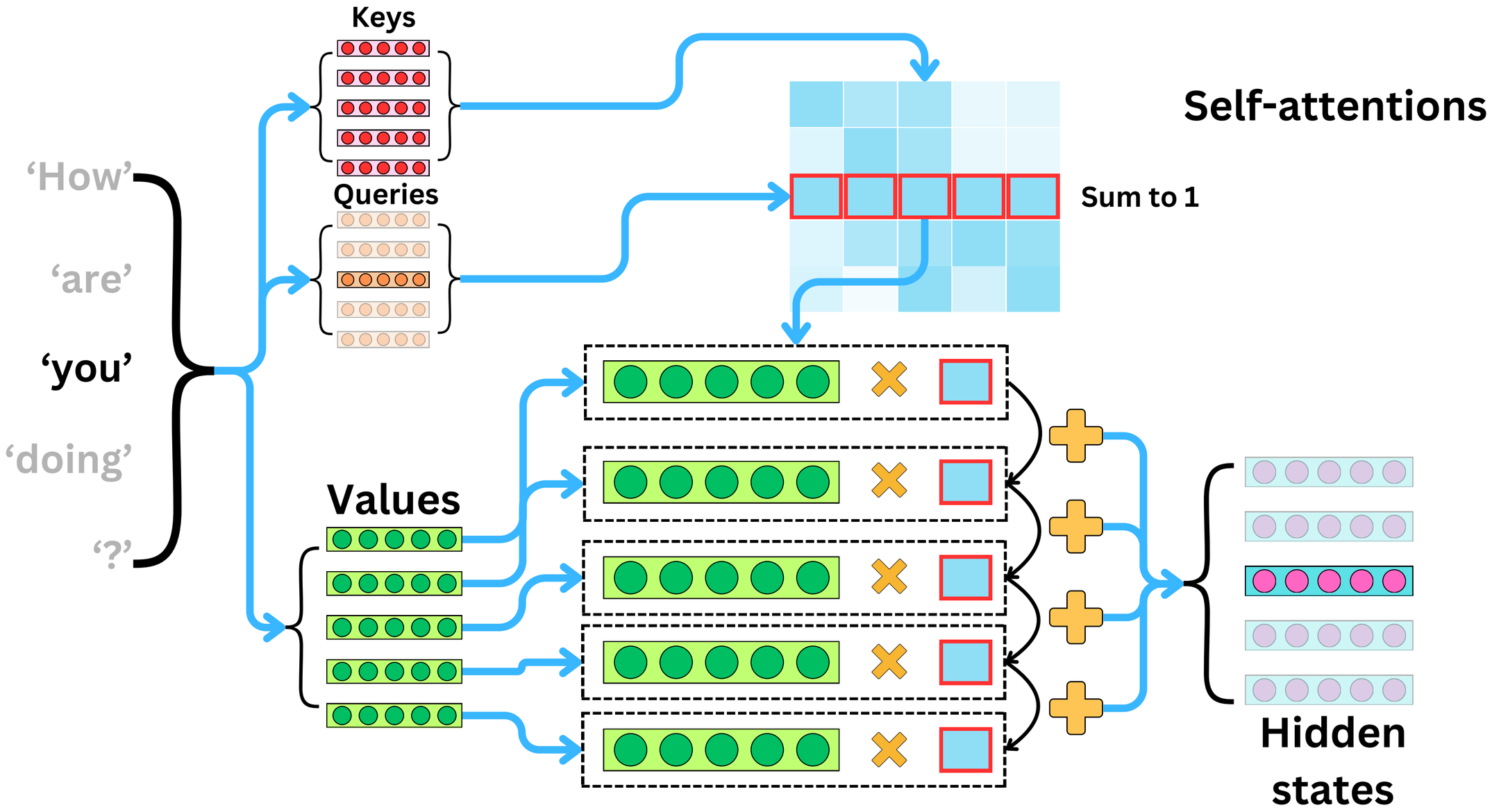


$$A = \text{Softmax}(E)$$

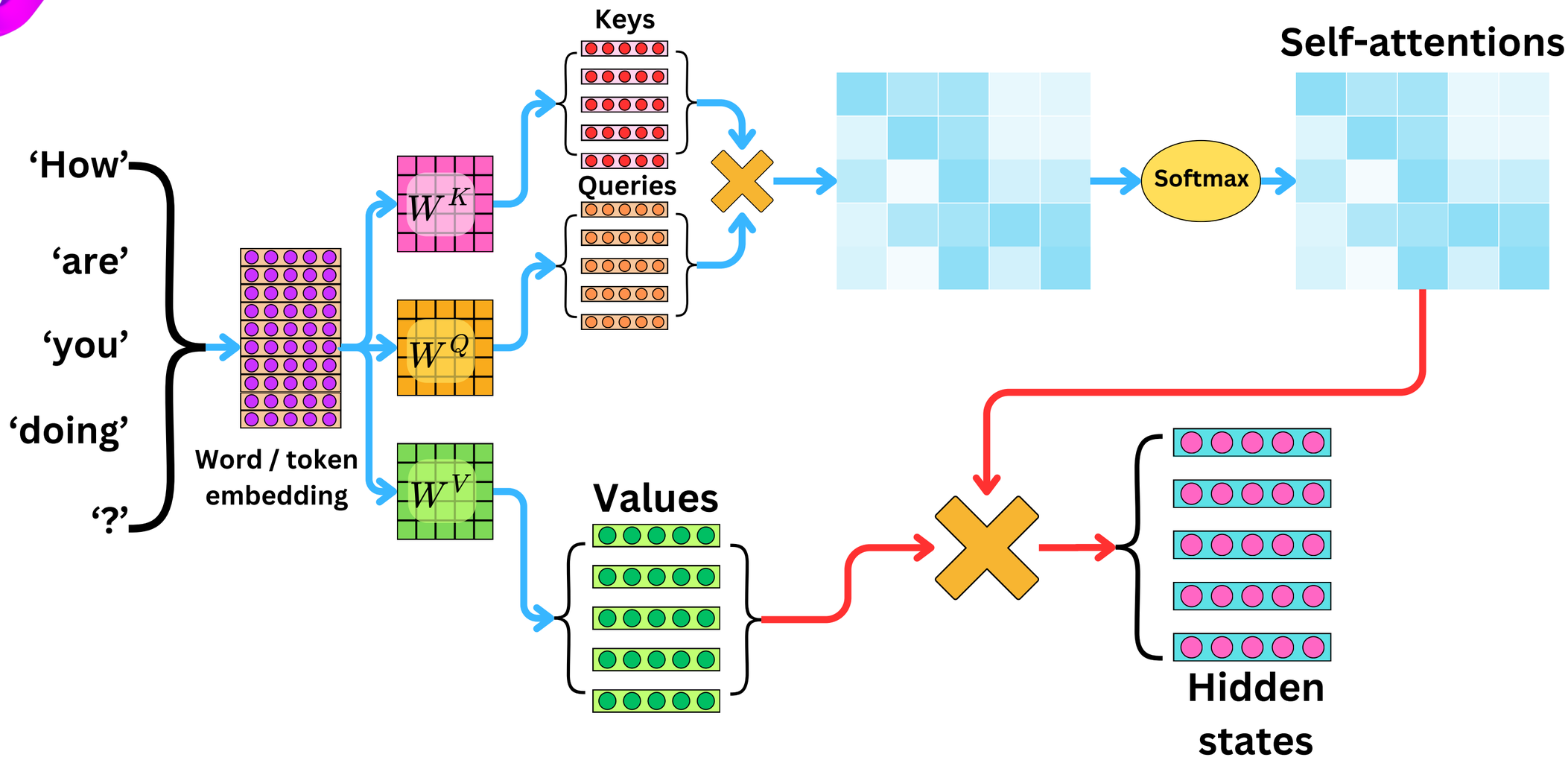
The Self-Attention Layer



The Self-Attention Layer

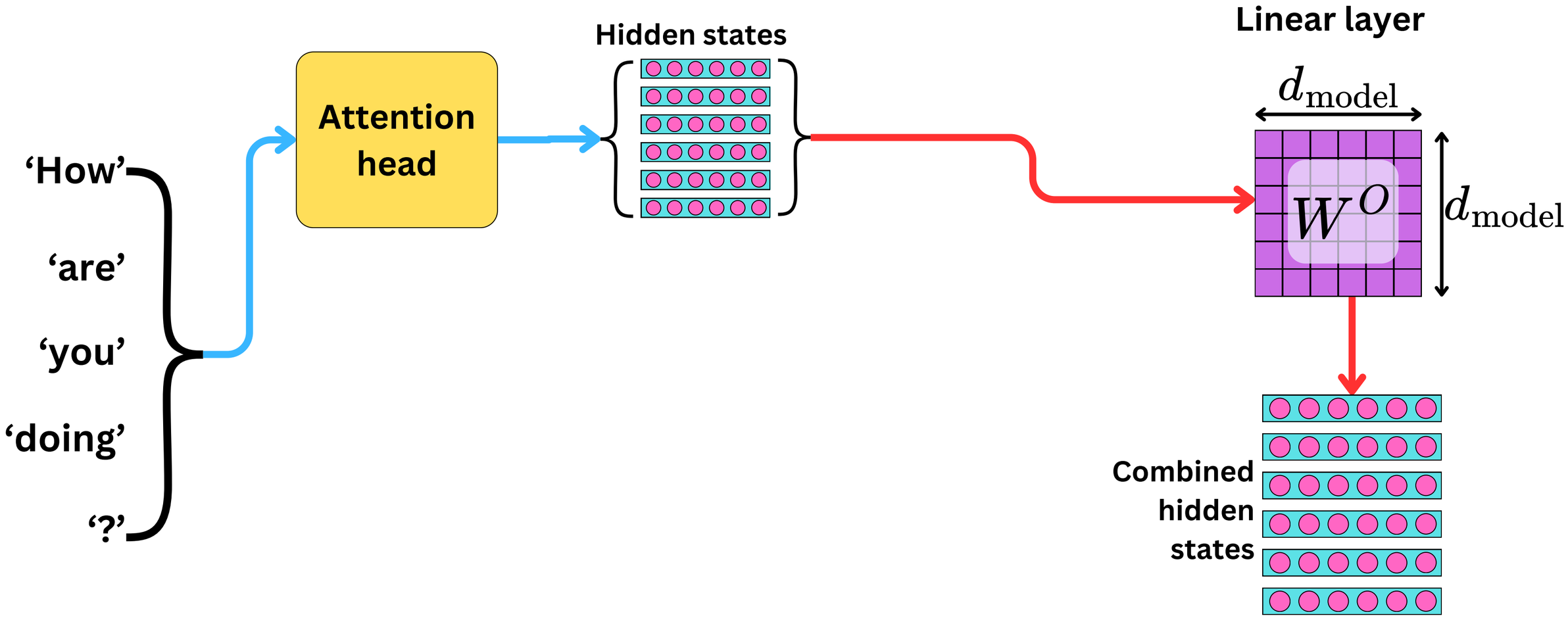


The Self-Attention Layer

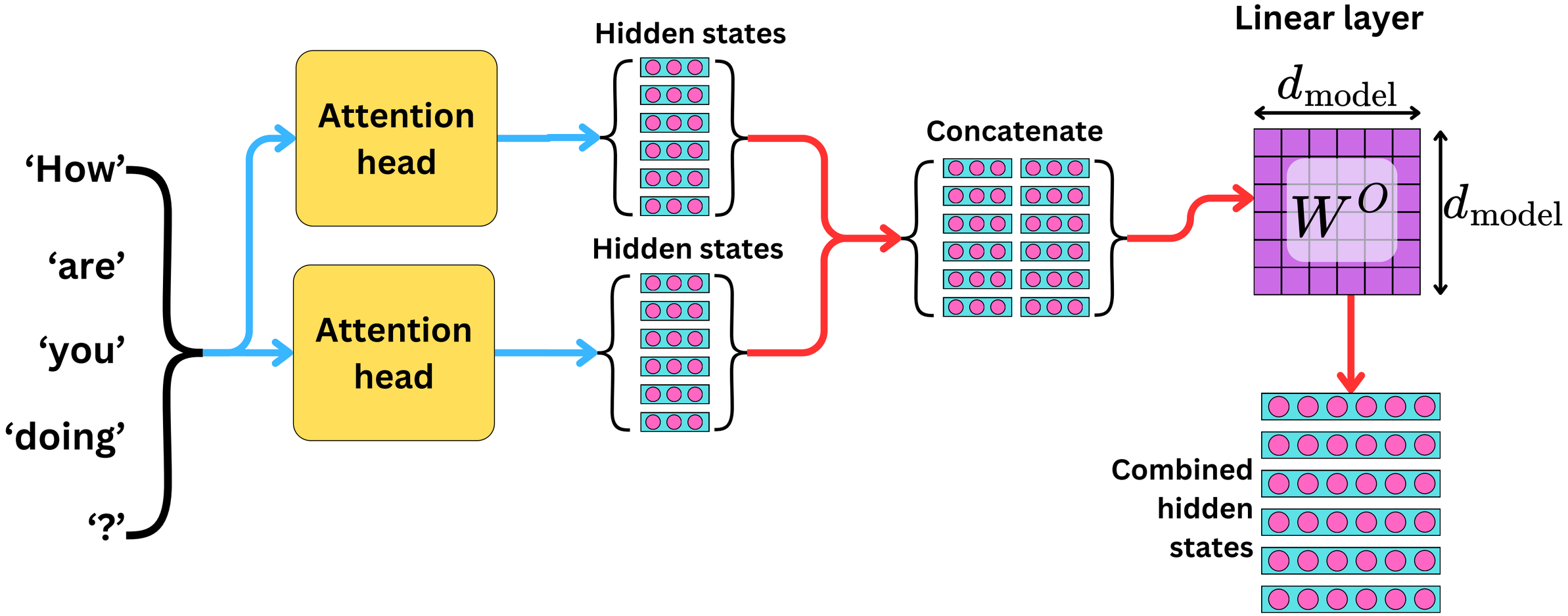


$$C = \text{Softmax}\left(\frac{W^Q H (W^K H)^T}{\sqrt{d_{\text{model}}}}\right) W^V H$$

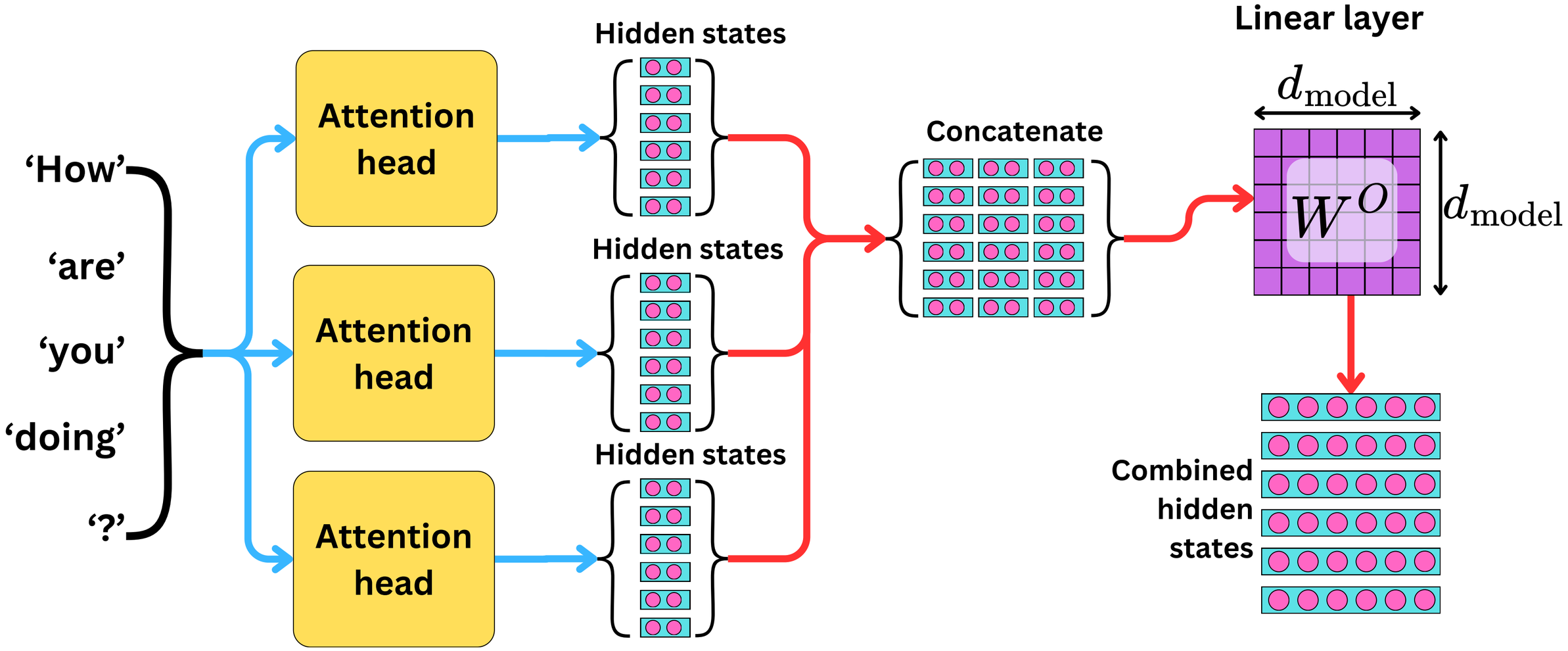
Multihead Attention Layer



Multihead Attention Layer

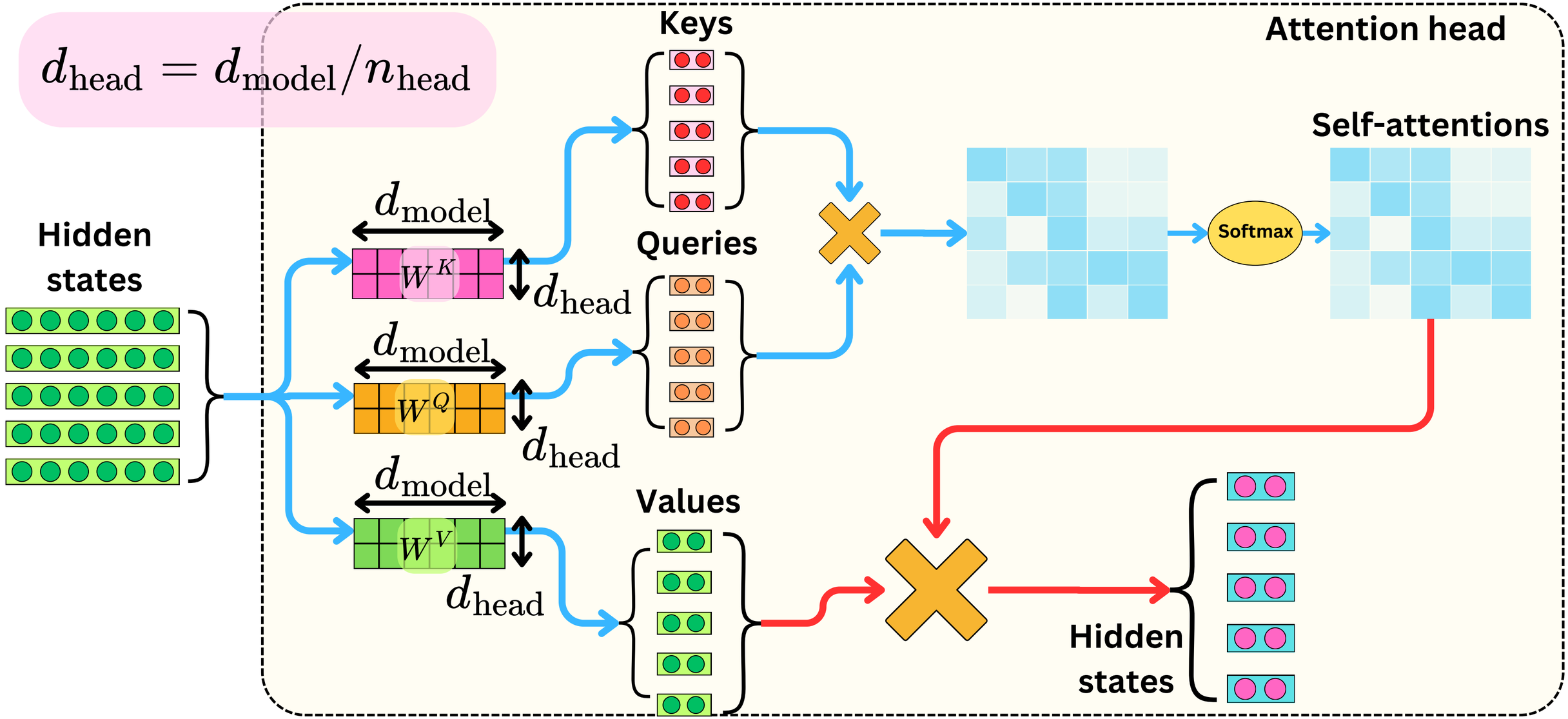


Multihead Attention Layer

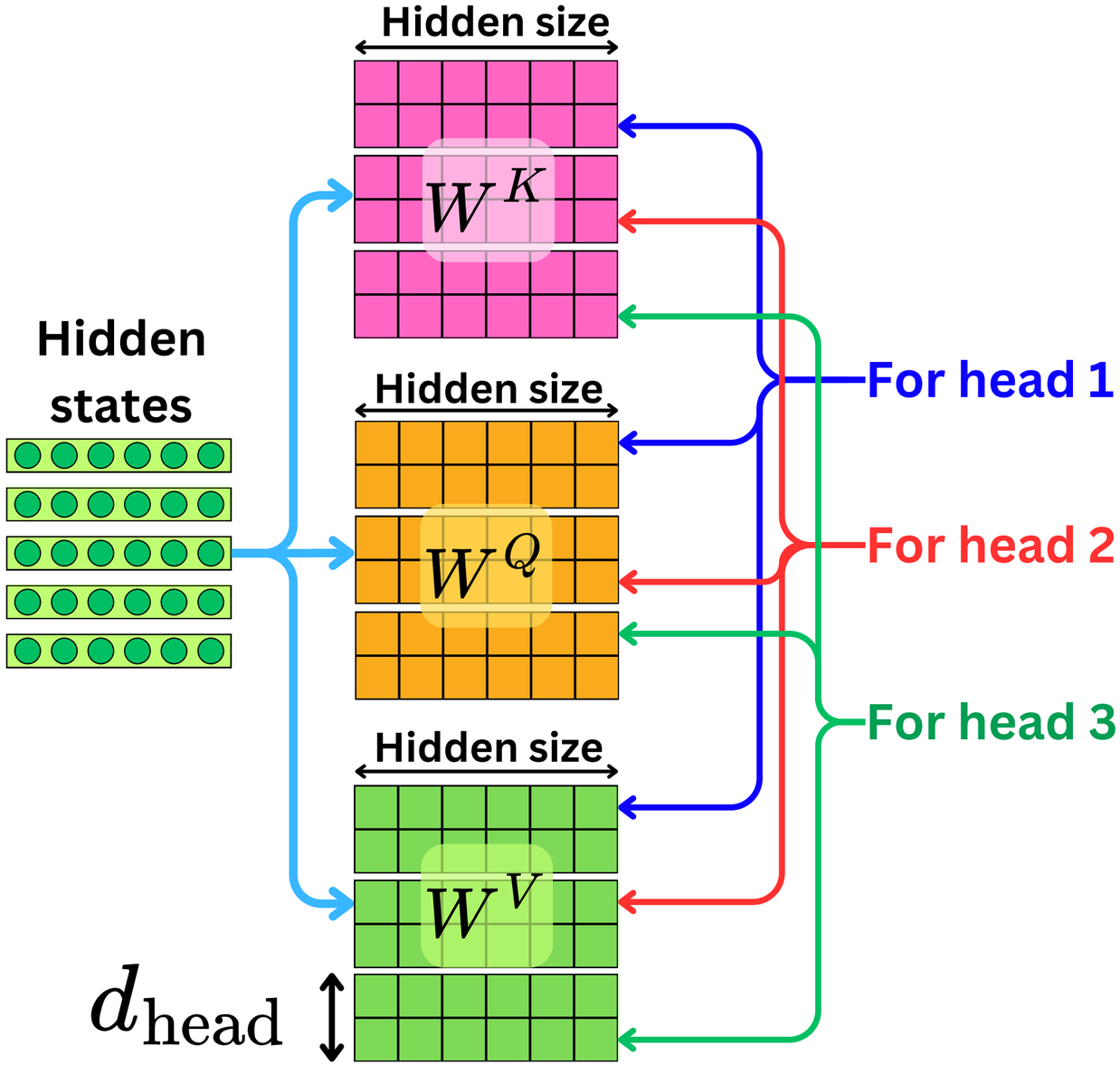


Multihead Attention Layer

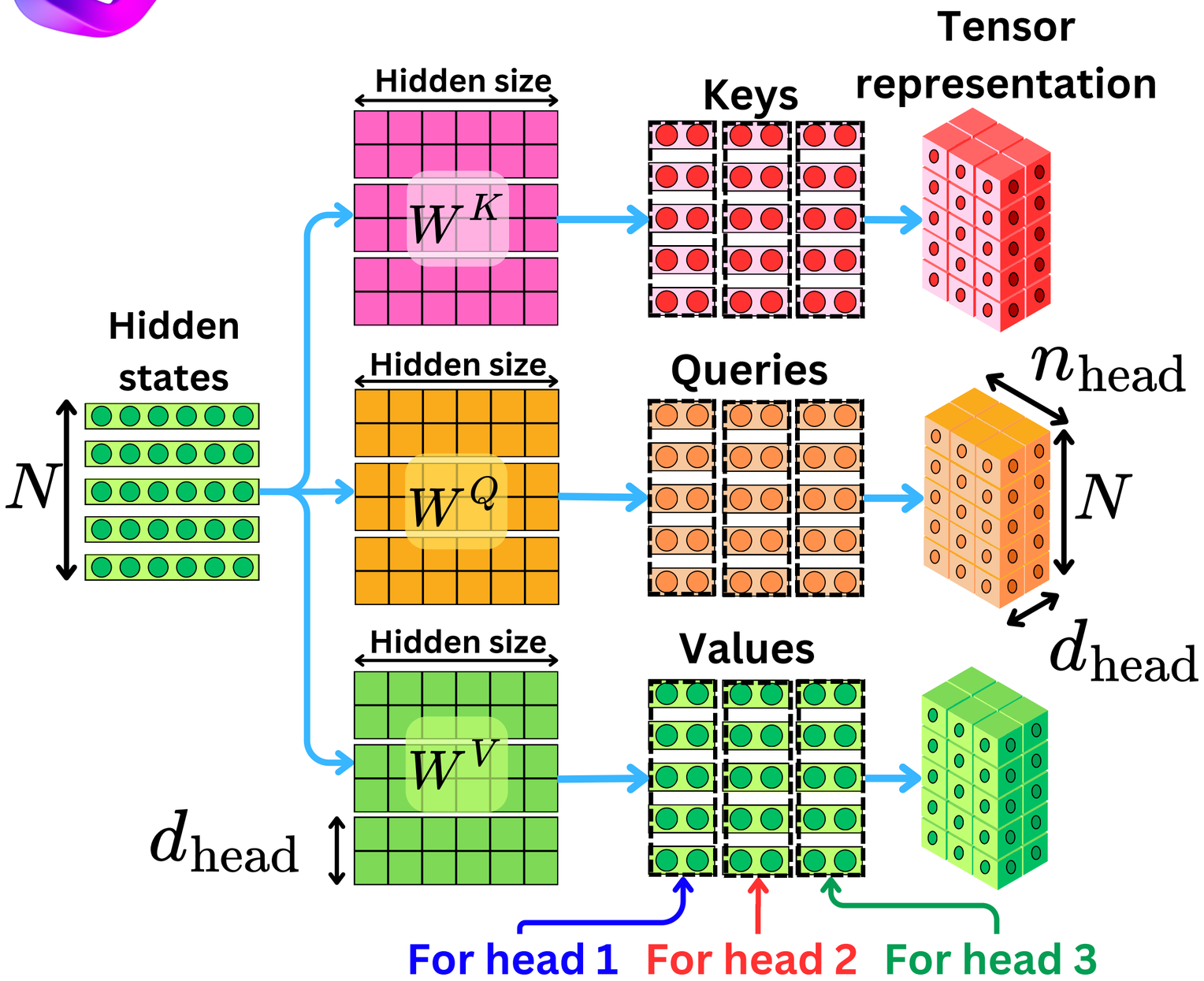
$$d_{\text{head}} = d_{\text{model}} / n_{\text{head}}$$



Multihead Attention Layer: In Practice



Multihead Attention Layer: In Practice



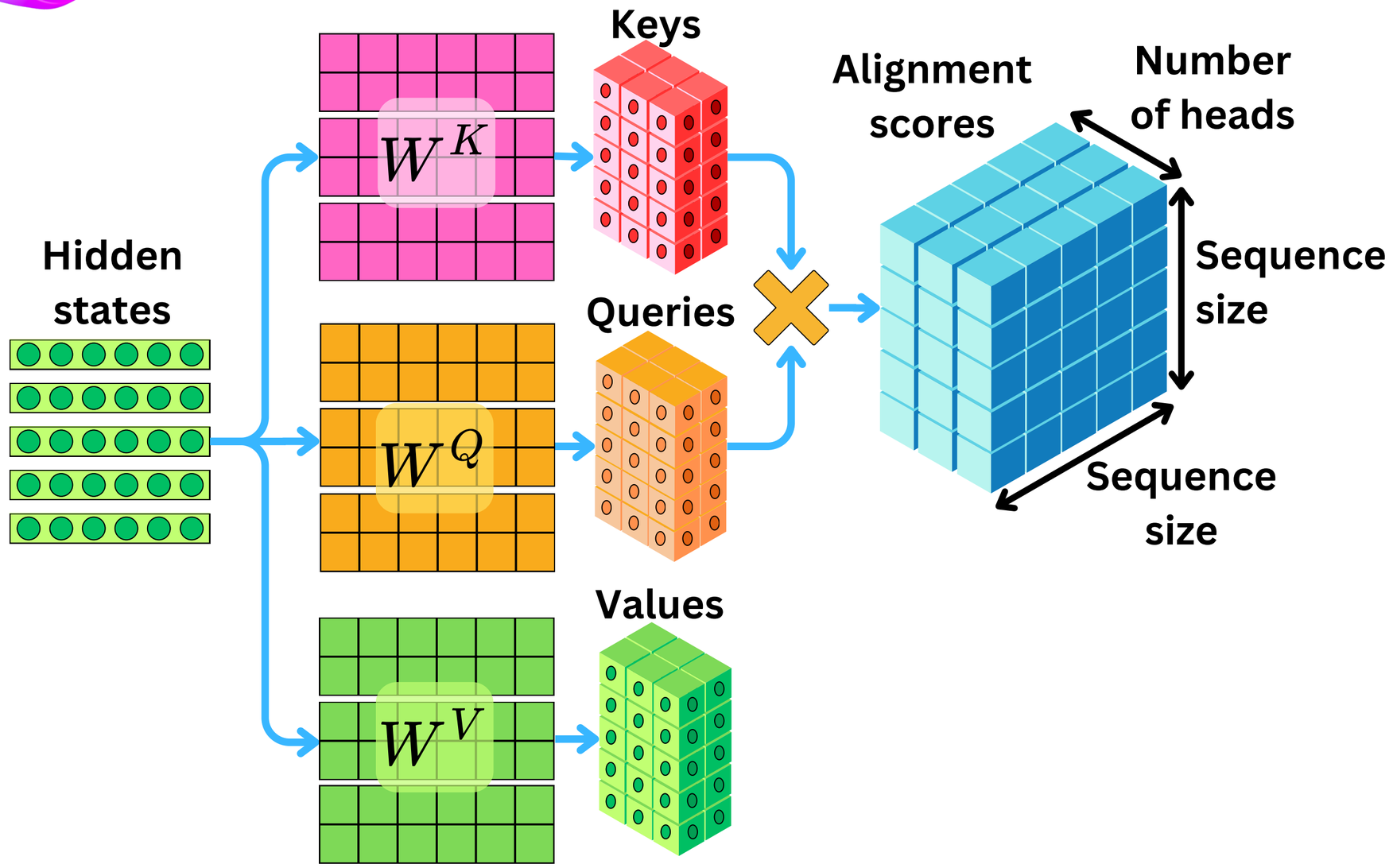
Reshape from

$$(d_{\text{model}} \times N)$$

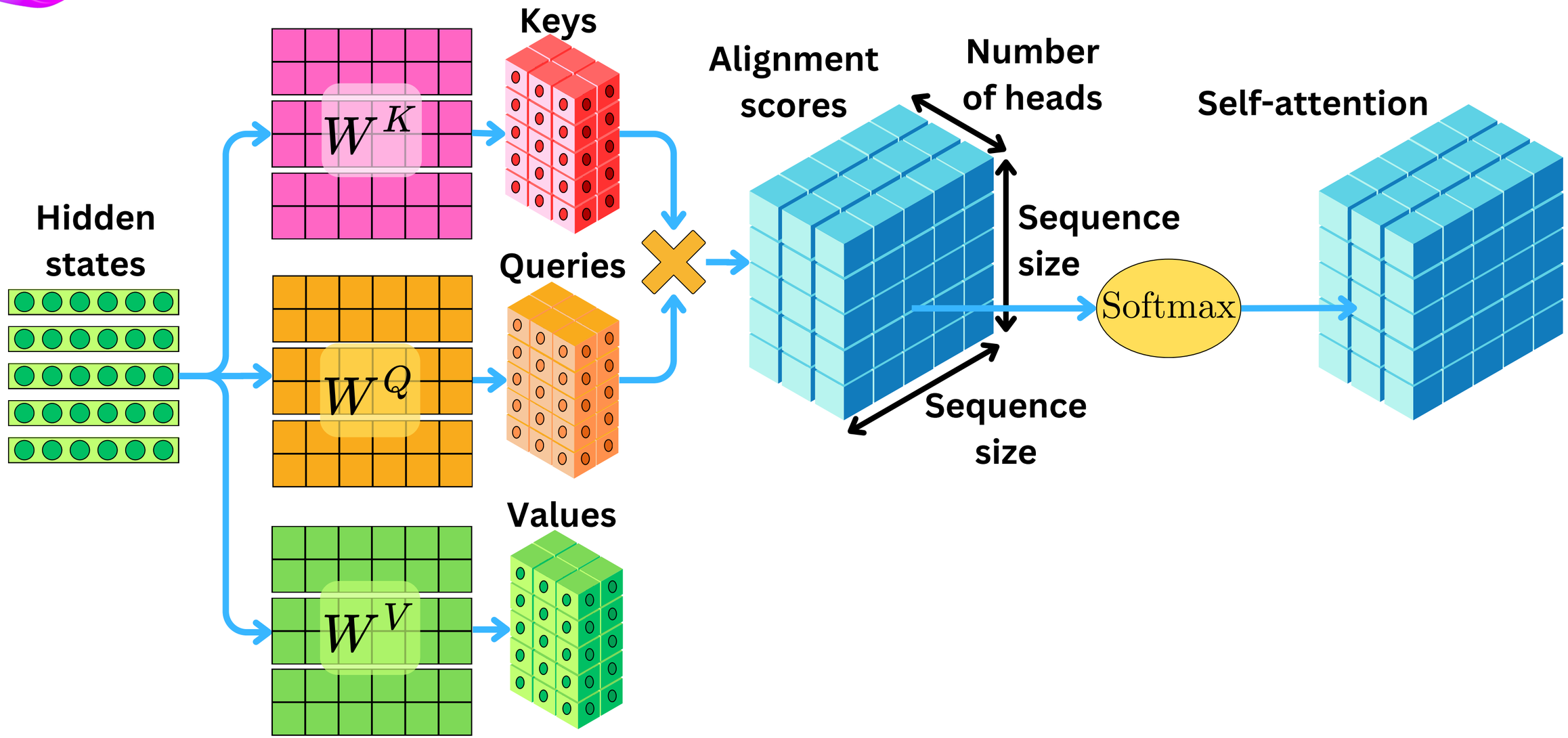
to

$$(d_{\text{head}} \times n_{\text{head}} \times N)$$

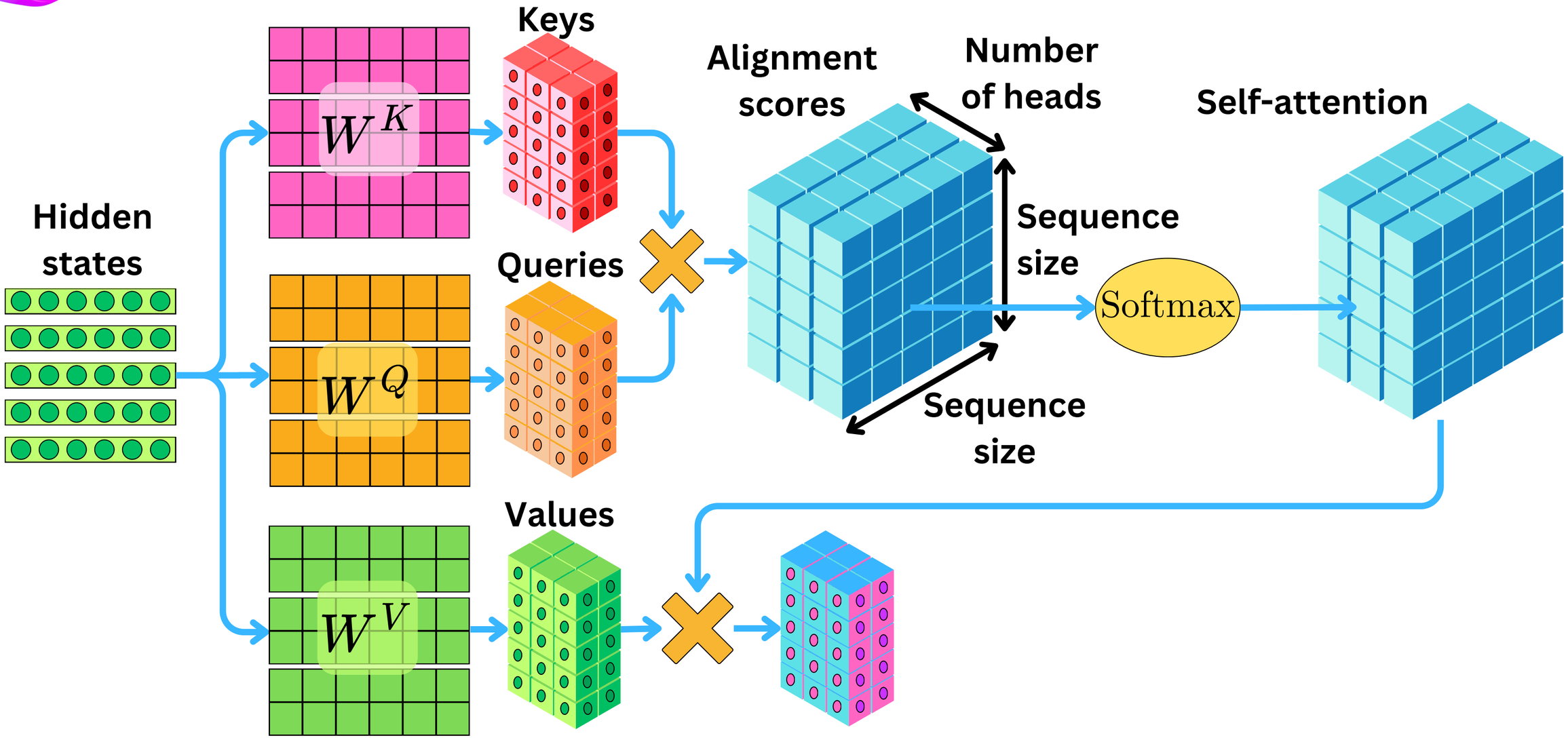
Multihead Attention Layer: In Practice



Multihead Attention Layer: In Practice



Multihead Attention Layer: In Practice



Multihead Attention Layer: In Practice