Transformer Architecture
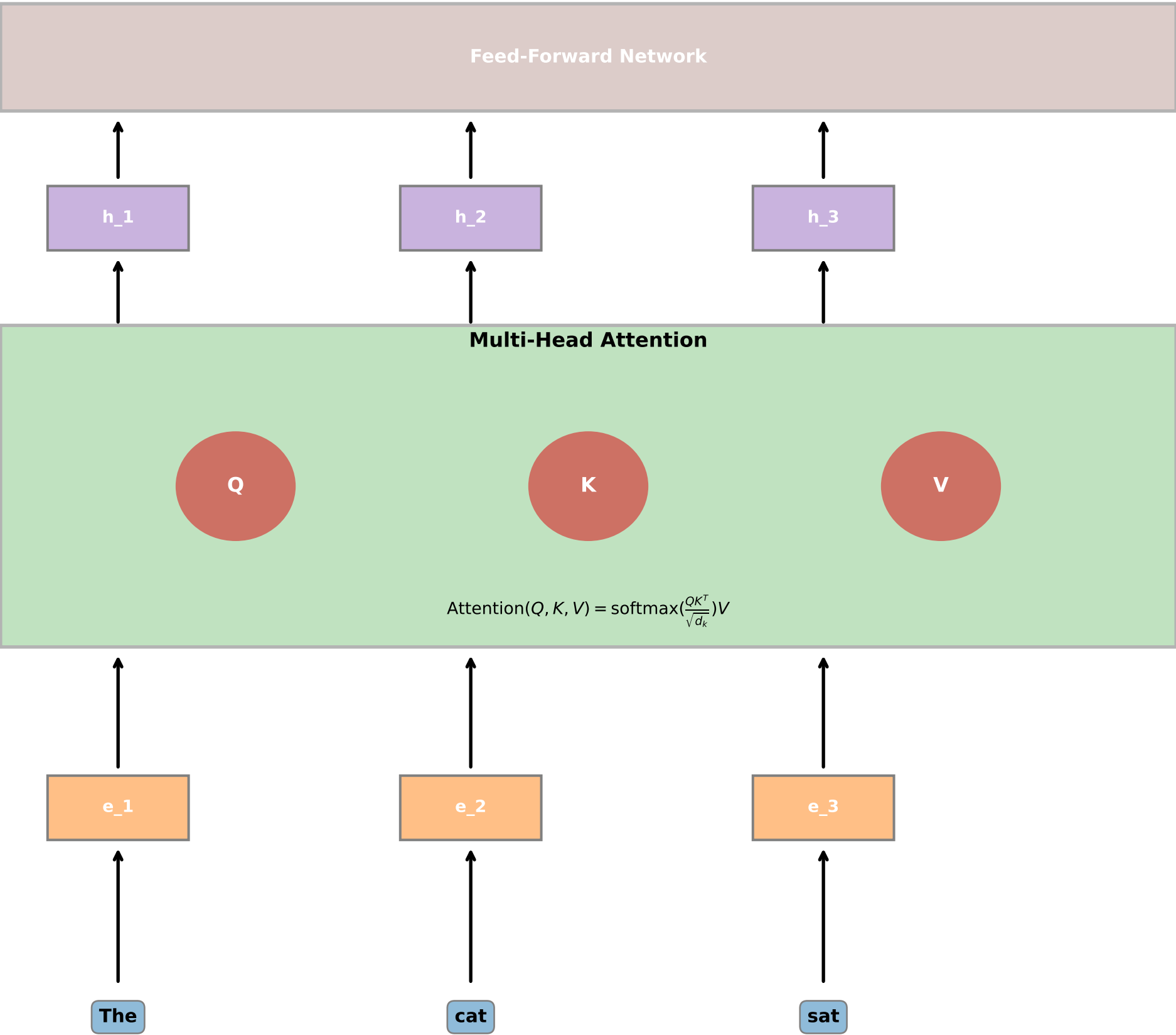
Feed-Forward Network

h_1

h_2

h_3

Multi-Head Attention

Q

K

V

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Embedding

e_1

e_2

e_3

The

cat

sat