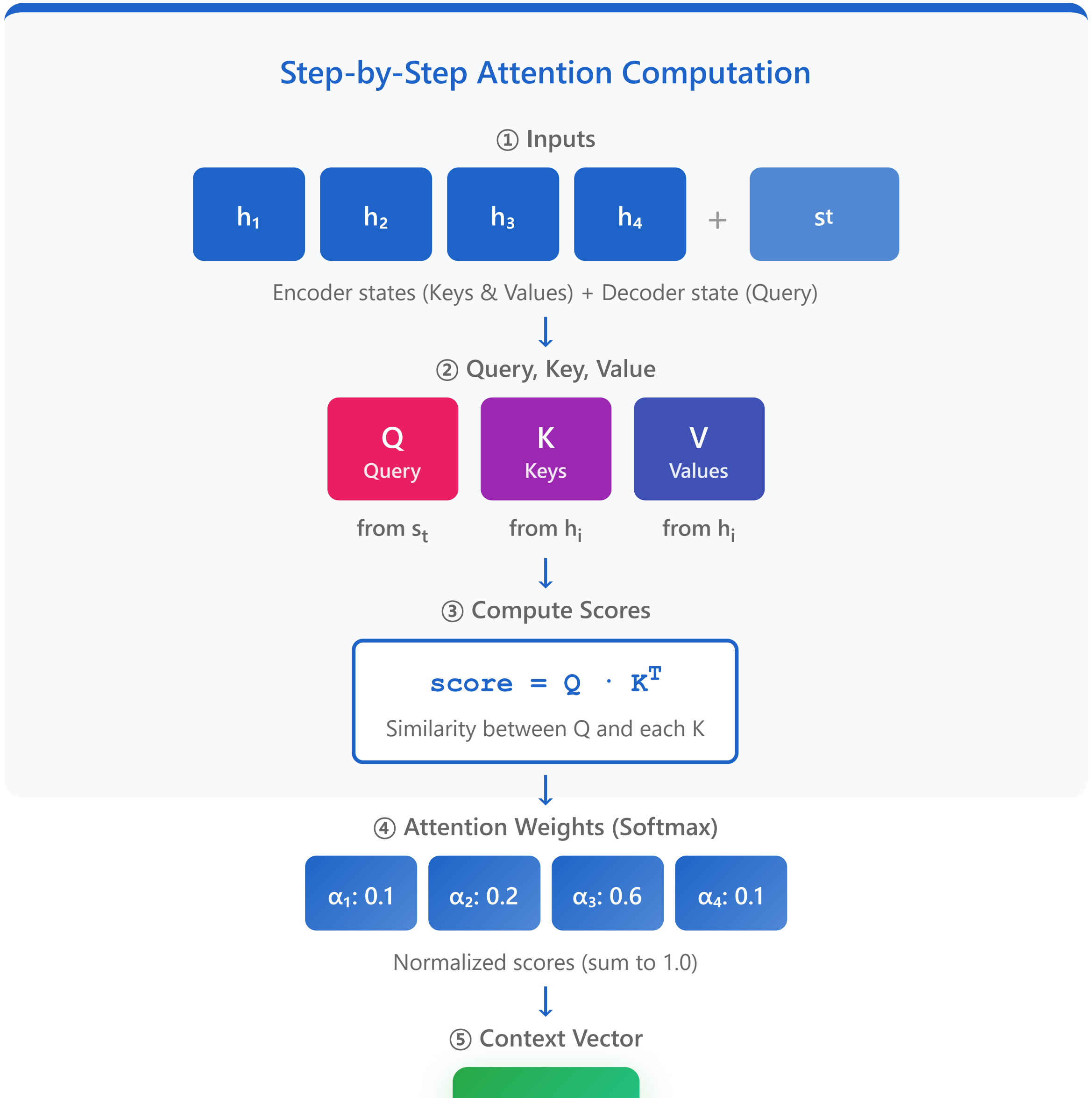
Attention Mechanism Structure



Step-by-Step Attention Computation

① Inputs

h_1 h_2 h_3 h_4 + st

Encoder states (Keys & Values) + Decoder state (Query)

↓

② Query, Key, Value

Q Query — from s_t

K Keys — from h_i

V Values — from h_i

↓

③ Compute Scores

$\text{score} = Q \cdot K^T$

Similarity between Q and each K

↓

④ Attention Weights (Softmax)

α_1: 0.1 α_2: 0.2 α_3: 0.6 α_4: 0.1

Normalized scores (sum to 1.0)

↓

⑤ Context Vector

Query

"What am I looking for?" Derived from current decoder state to find relevant encoder states.

$Q = W_Q \times s_t$

Key

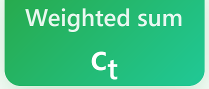
"What do I contain?" Encoder states transformed to be compared with query.

$K = W_K \times h_i$

Value

"What information to extract?" Actual content to be aggregated based on attention weights.

$V = W_V \times h_i$



Weighted sum

c_t

$c_t = \Sigma(\alpha_i \times V_i)$