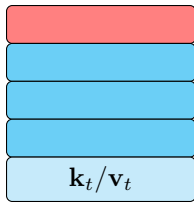




$$\left(\hat{\mathbf{q}}_t \begin{bmatrix} \mathbf{A}_{K,t-1} \\ \hat{\mathbf{k}}_t \end{bmatrix}^\top \right) \rightarrow \text{Light attention score}$$

The result is a sequence of 7 light blue boxes representing the light attention score.

$$\text{Attention} \left(\mathbf{q}_t, \begin{bmatrix} \mathbf{K}'_{\Omega,t-1} \\ \mathbf{k}_t \end{bmatrix}, \begin{bmatrix} \mathbf{V}'_{\Omega,t-1} \\ \mathbf{v}_t \end{bmatrix} \right)$$



miss
Fetch CPU

TopK

hit
Unchange

