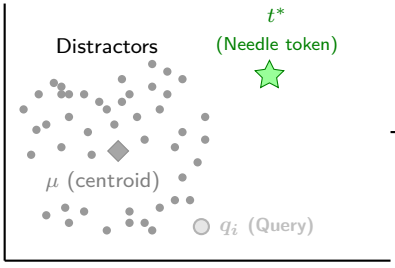
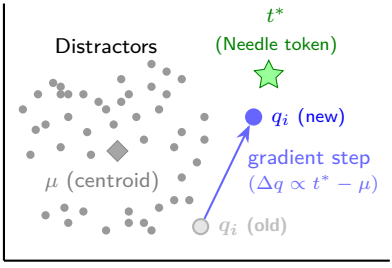


Gradient step



Small logit margin Δ
$\Rightarrow$ diffuse attention.

Δ grows to $\log(T) \Rightarrow$ attention concentrates on needle.

$\nabla_{q_i} \ell \propto \mu - t^* \Rightarrow$ gradient descent moves q toward t^* (needle) and away from μ.