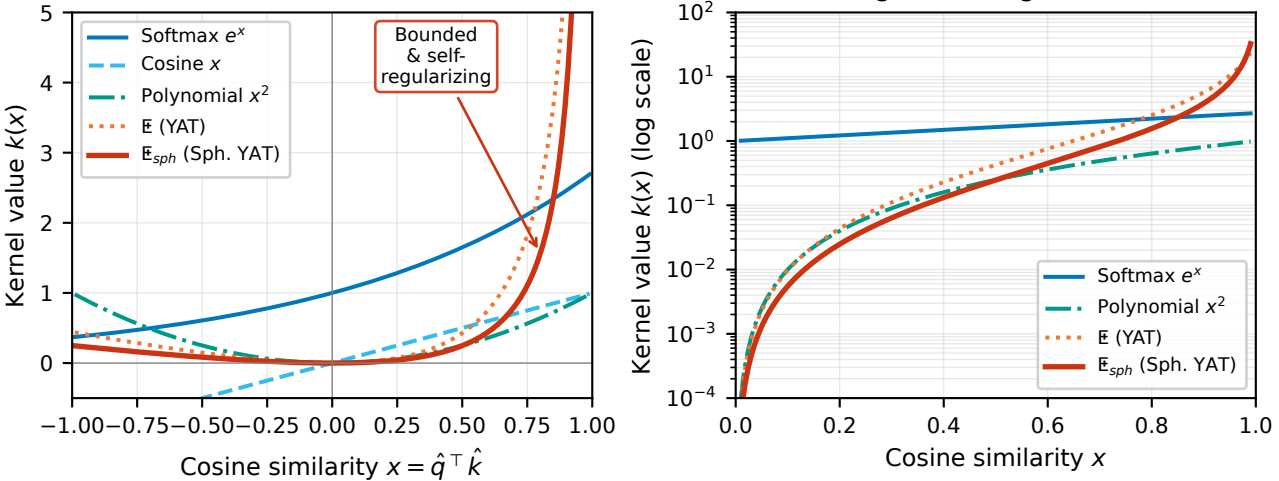
(a) Kernel functions (linear scale)
Softmax e^x
Cosine x
Polynomial x^2
𝔼 (YAT)
$𝔼_{sph}$ (Sph. YAT)
Bounded & self-regularizing
Kernel value $k(x)$
Cosine similarity $x = \hat{q}^\top \hat{k}$
(b) Log scale (aligned tokens)
Softmax e^x
Polynomial x^2
𝔼 (YAT)
$𝔼_{sph}$ (Sph. YAT)
Kernel value $k(x)$ (log scale)
Cosine similarity x