cos similarity histogram - layer 3 (albert-xlarge-v2 uniform-tokens (Qx,Ky) trained-params dense-off 24-layer 3.5 3.0 2.5 2.0 1.5 1.0 0.5 0.0 -0.500.00 0.25 0.50 -1.00-0.75-0.250.75 1.00