cos similarity histogram - layer 1 ert-large-uncased uniform-tokens (Qx,Ky) trained-params dense-off 24-lay 7 6 5 4 3 2 1 -0.50-0.250.00 0.25 0.50 -1.00-0.750.75 1.00