Pairwise Correlation between Alignment of Neurons and Zero-shot Performance on XNLI Through Pre-training

bloom-560m: $r = 0.57, \rho = 0.54$

bloom-1b1: $r = 0.72, \rho = 0.63$

bloom-1b7: $r = 0.57, \rho = 0.46$

Legend: ar, es, fr, hi, ur, vi, zh

Axes: Acc. (y-axis), Neuron Overlap Rate (x-axis)