Monosemanticity spectrum (decoder.layer3) Monosemanticity
0 0 0 0 0 0
0 0 0 0 0 1500 2000 2500

Neuron index