Monosemanticity spectrum (decoder.layer4) 150 Neuron index