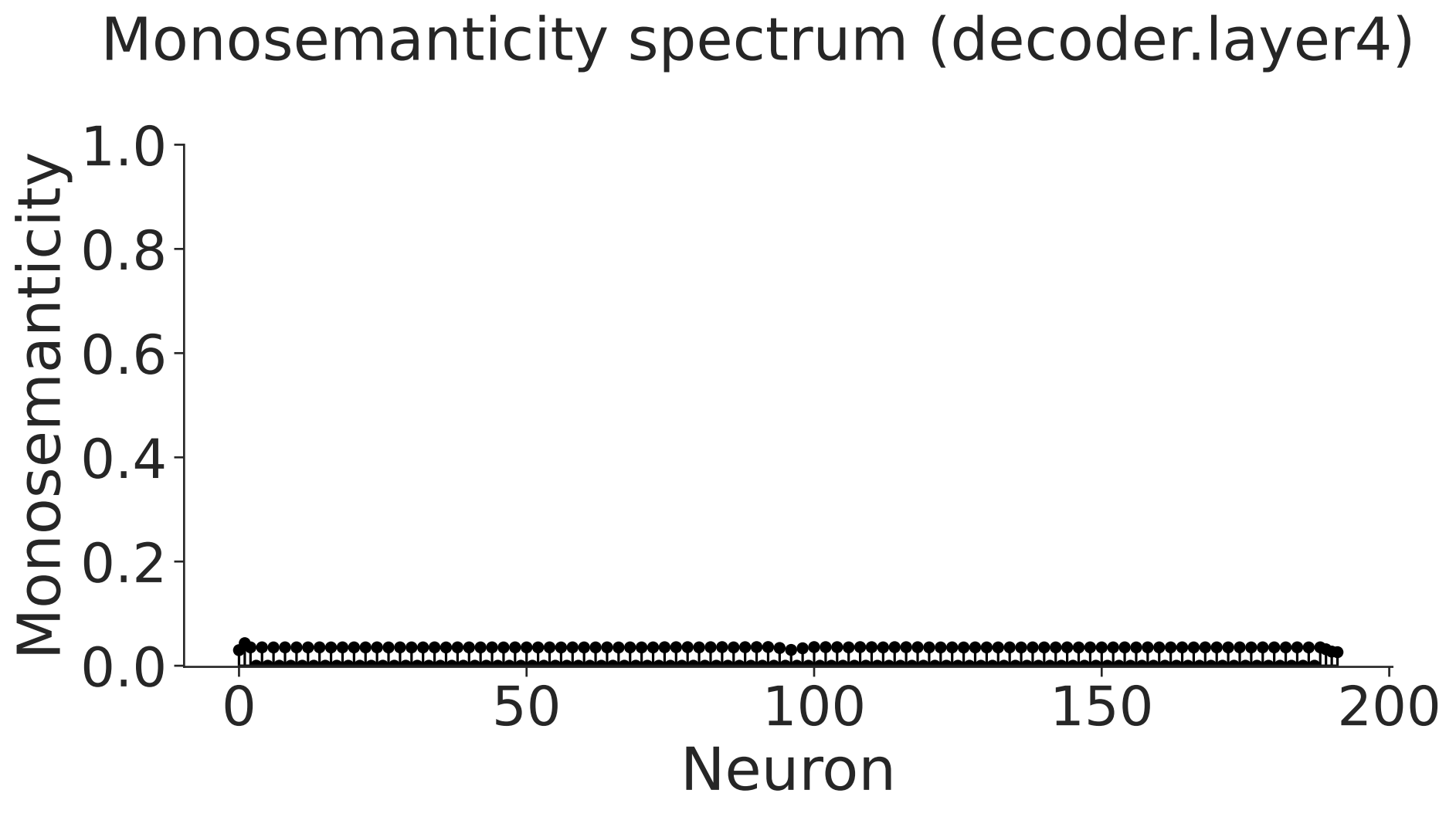
Monosemanticity spectrum (decoder.layer4)
Monosemanticity
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
Neuron