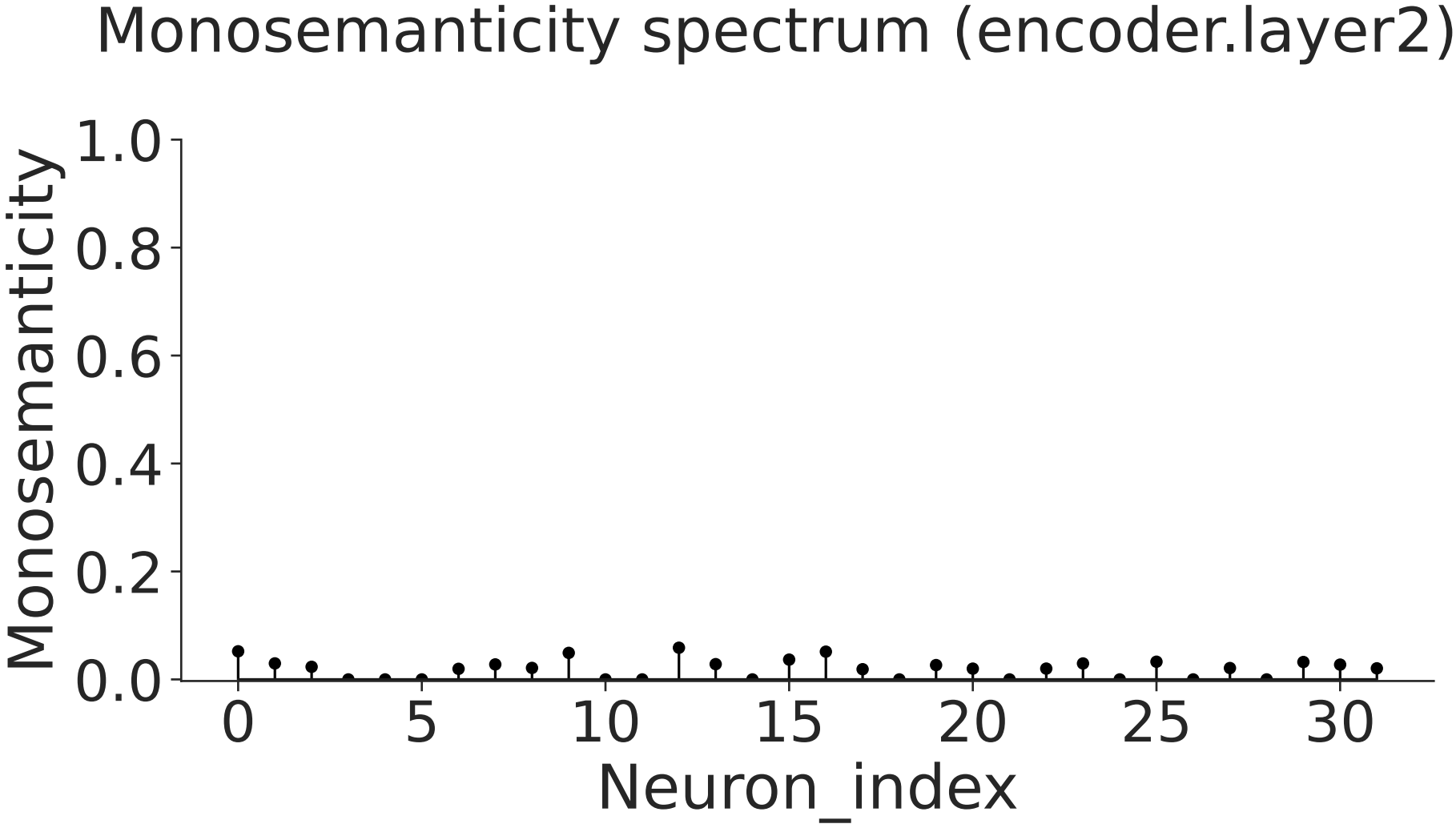
Monosemanticity spectrum (encoder.layer2)
Monosemanticity
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
Neuron_index