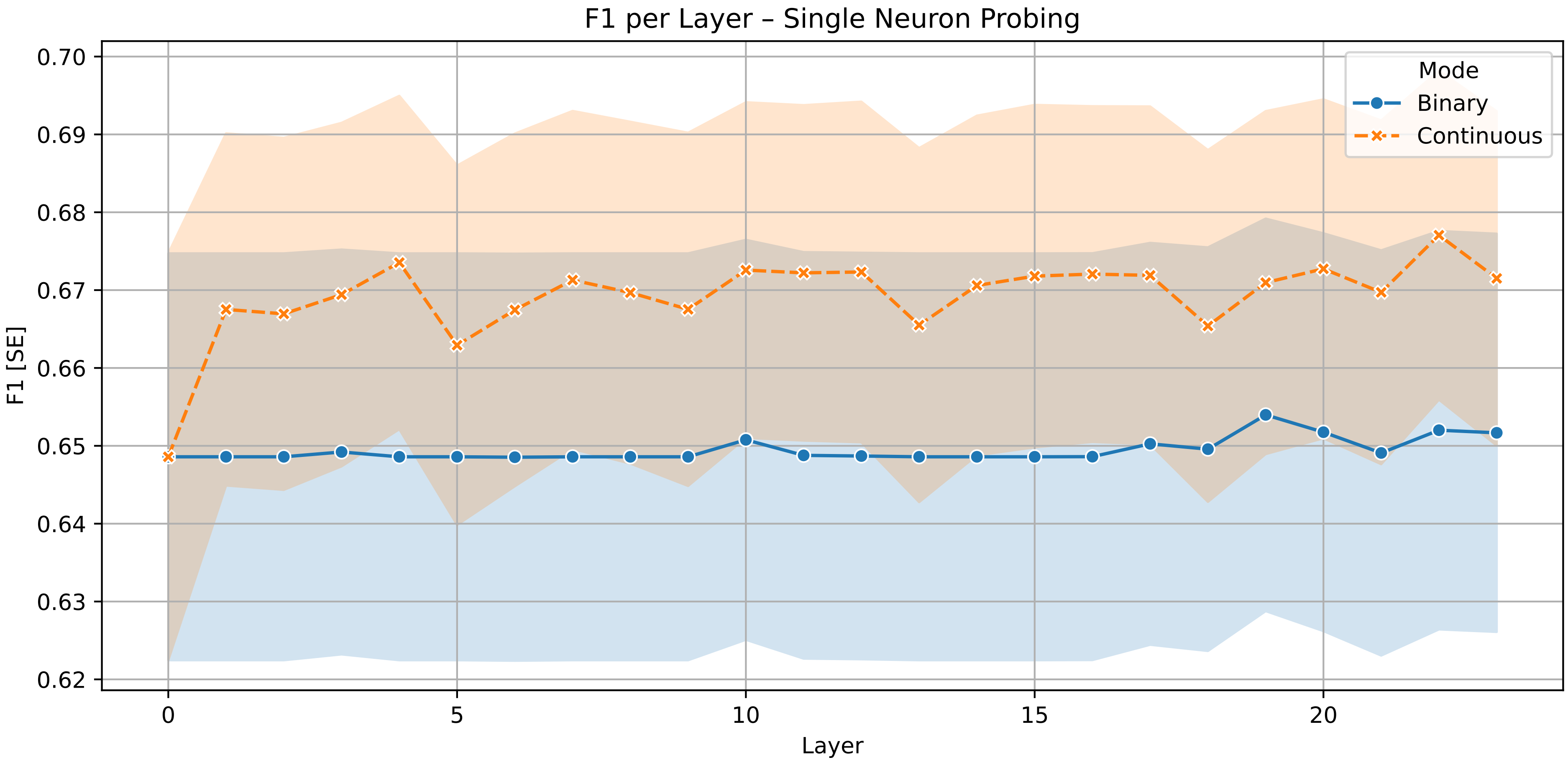
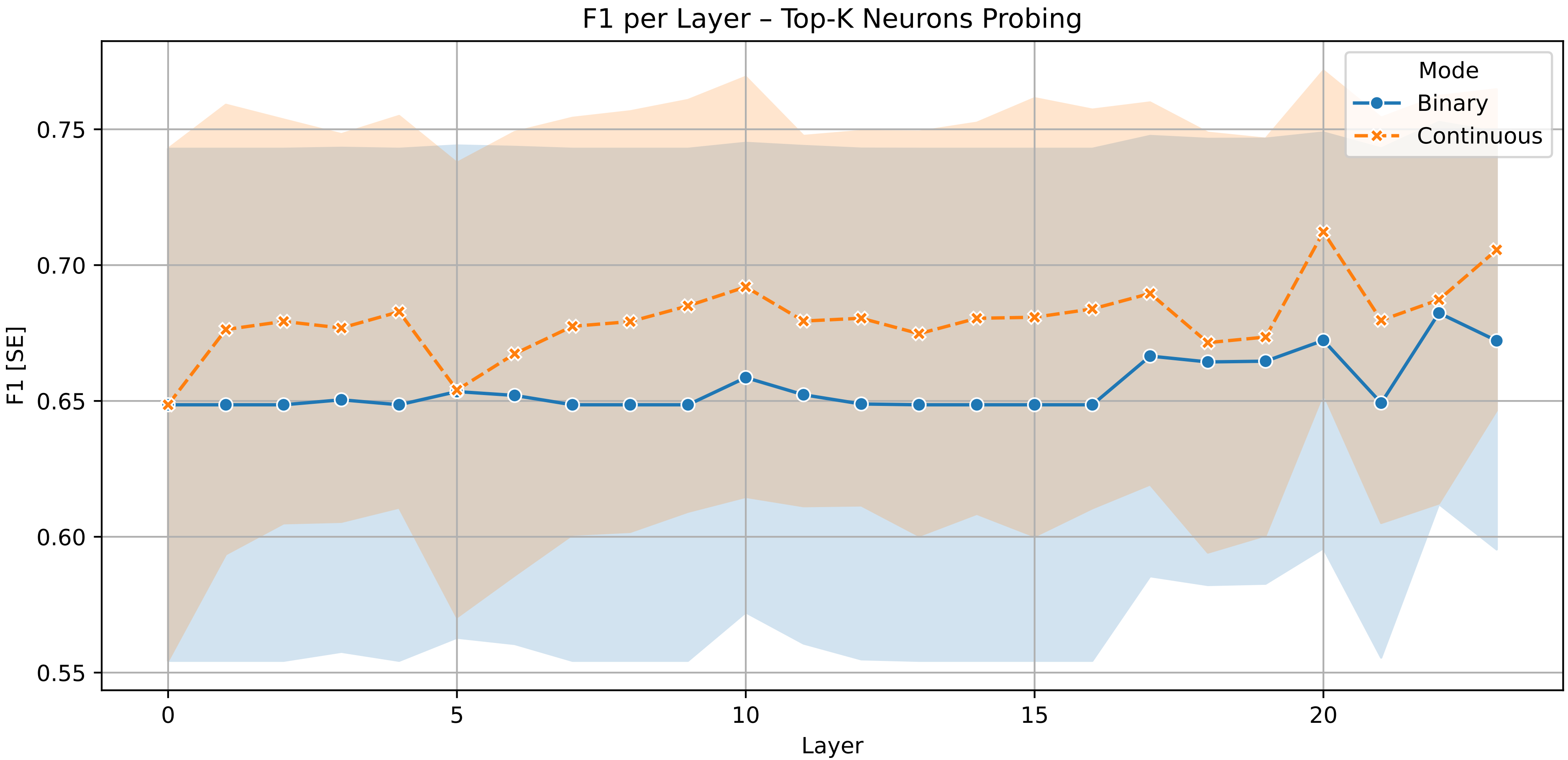
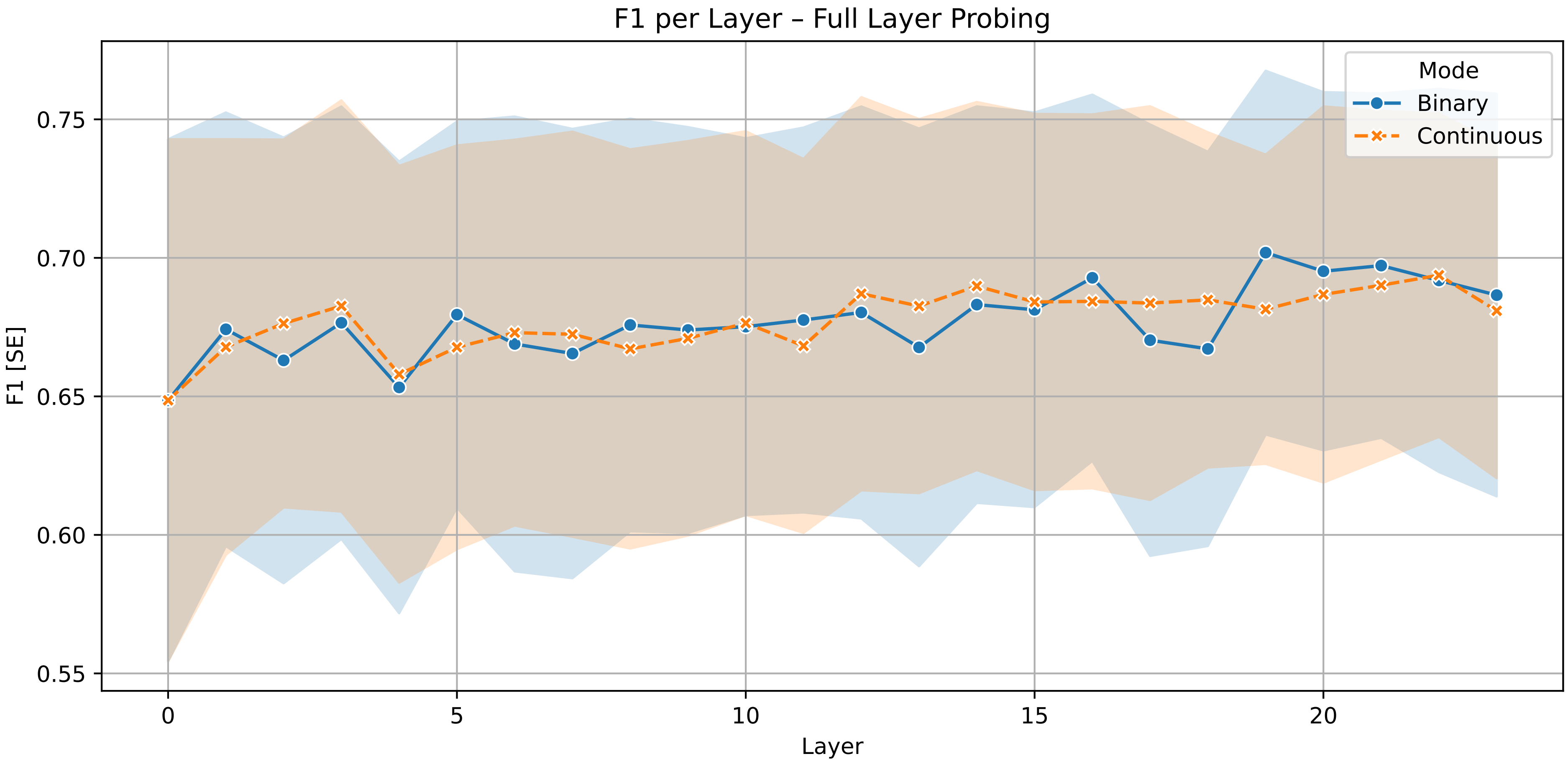
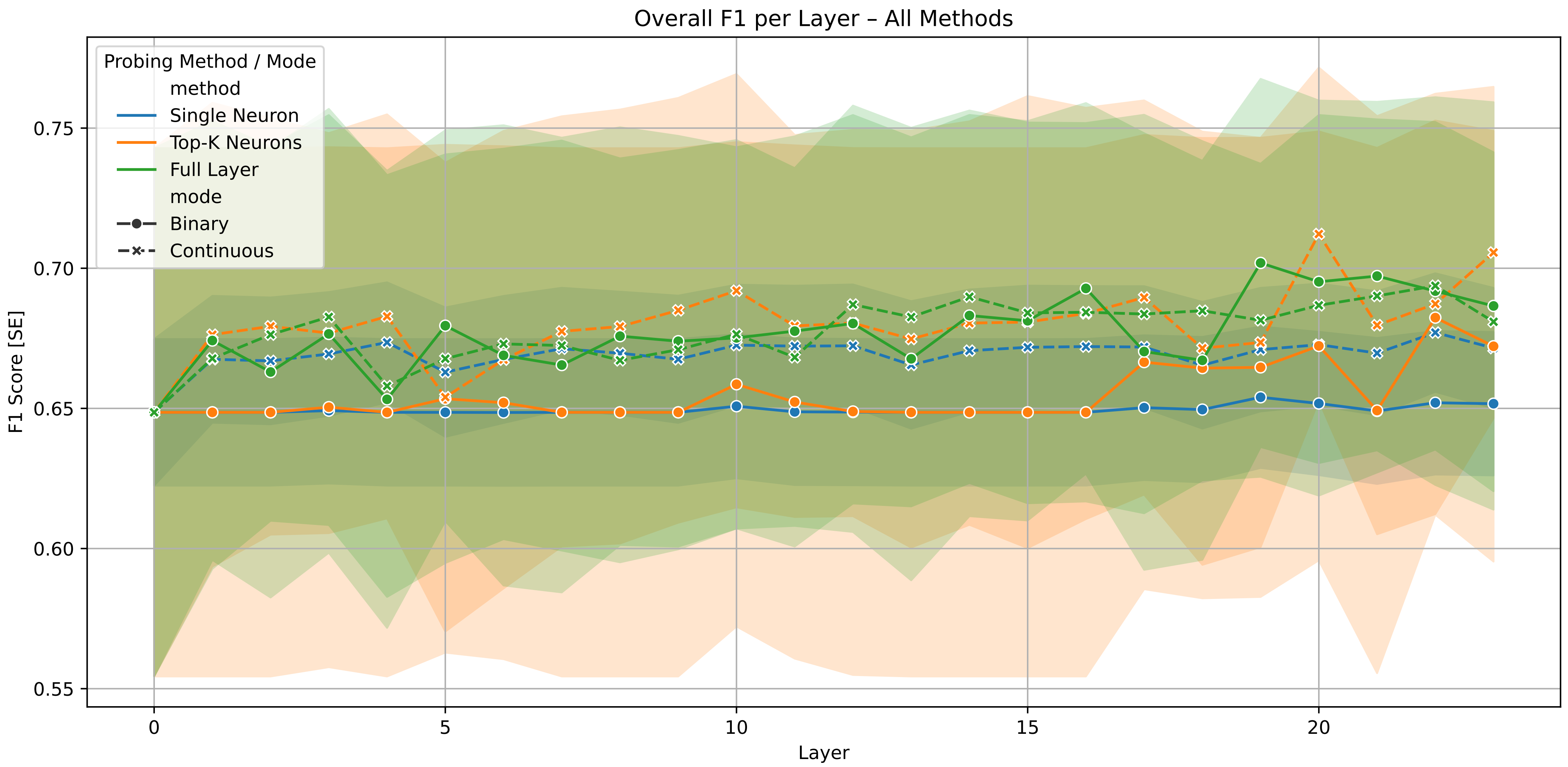
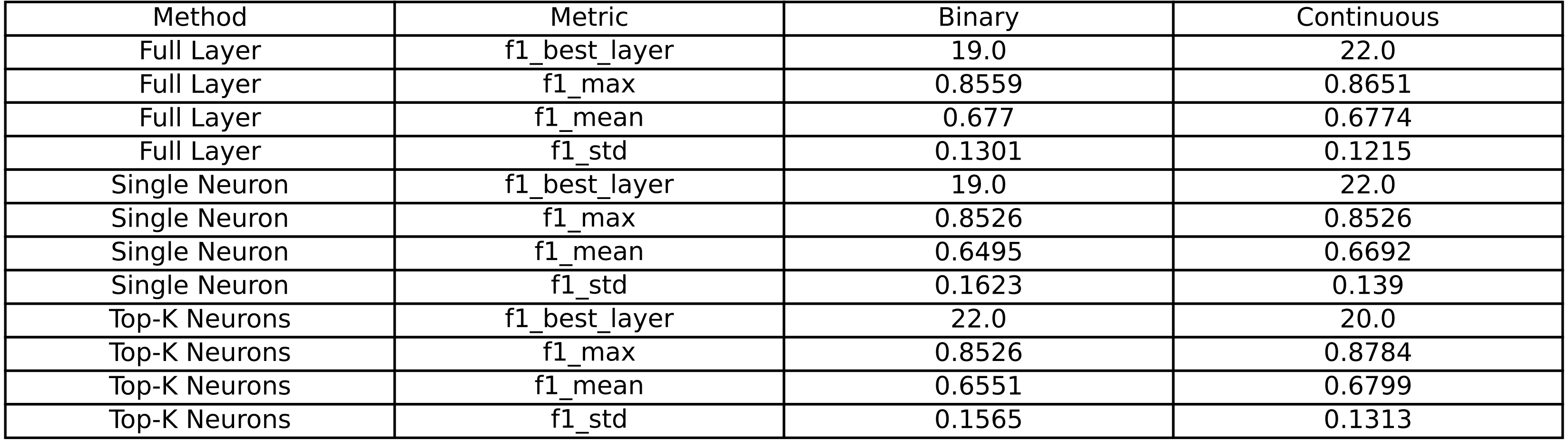

F1 per Layer – Single Neuron Probing
F1 [SE]
0.70
0.69
0.68
0.67
0.66
0.65
0.64
0.63
0.62
Layer
0
5
10
15
20
Mode
Binary
Continuous


F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
Layer

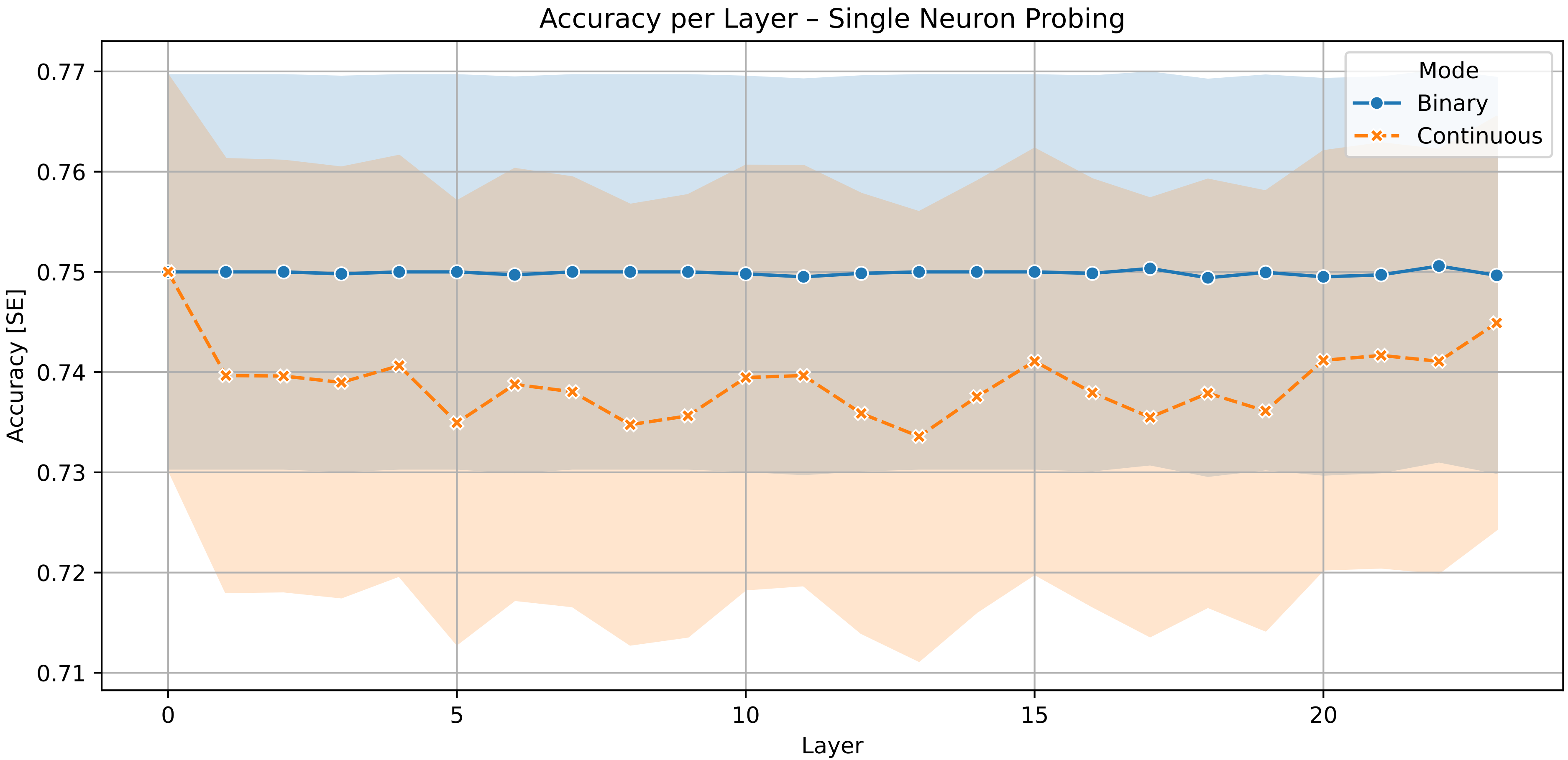

F1 per Layer – Full Layer Probing
F1 [SE]
0.55
0.60
0.65
0.70
0.75
Layer
0
5
10
15
20
Mode
Binary
Continuous

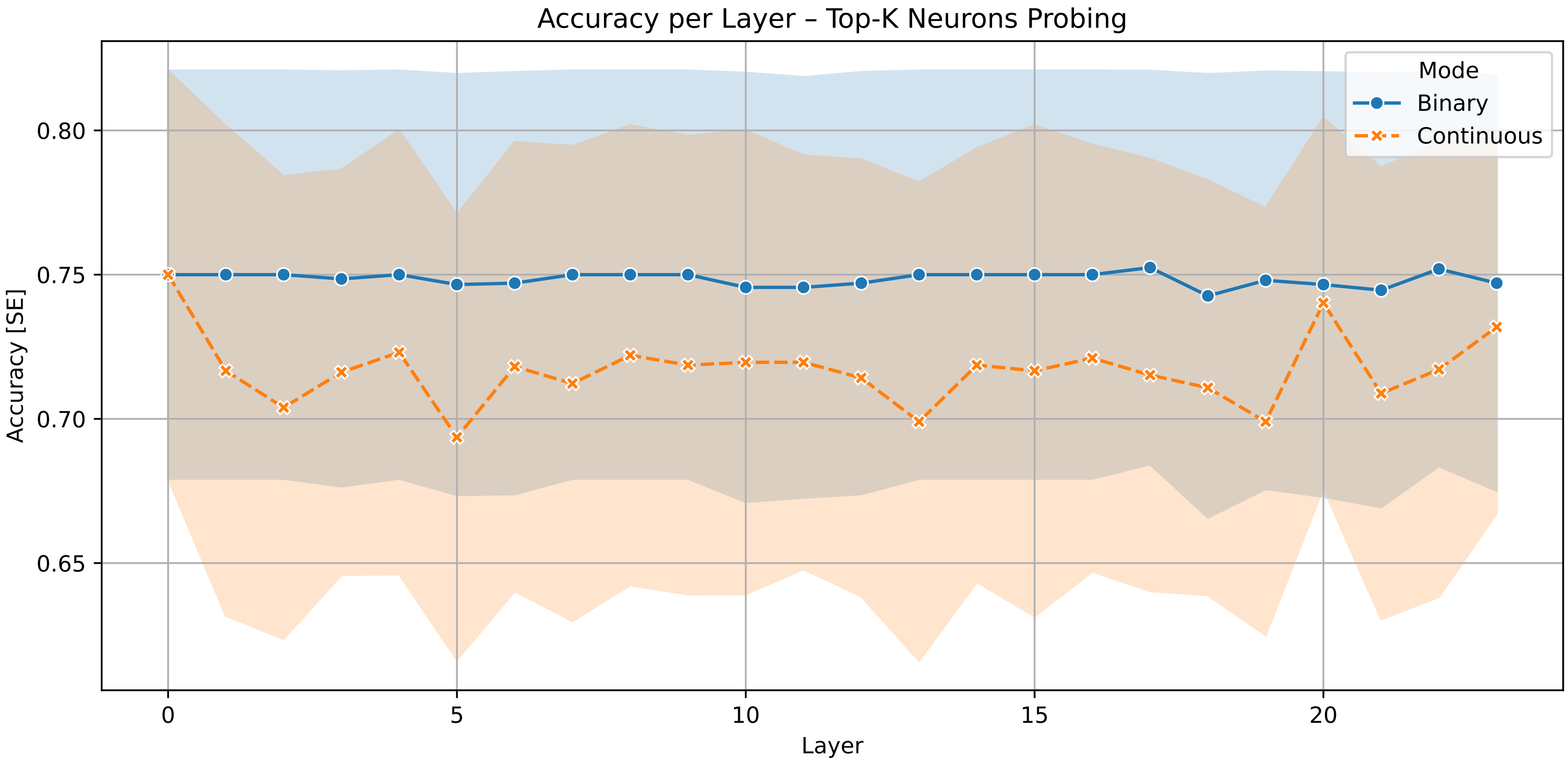

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.55
0.60
0.65
0.70
0.75
Layer
0
5
10
15
20

F1 Score Summary by Probing Method

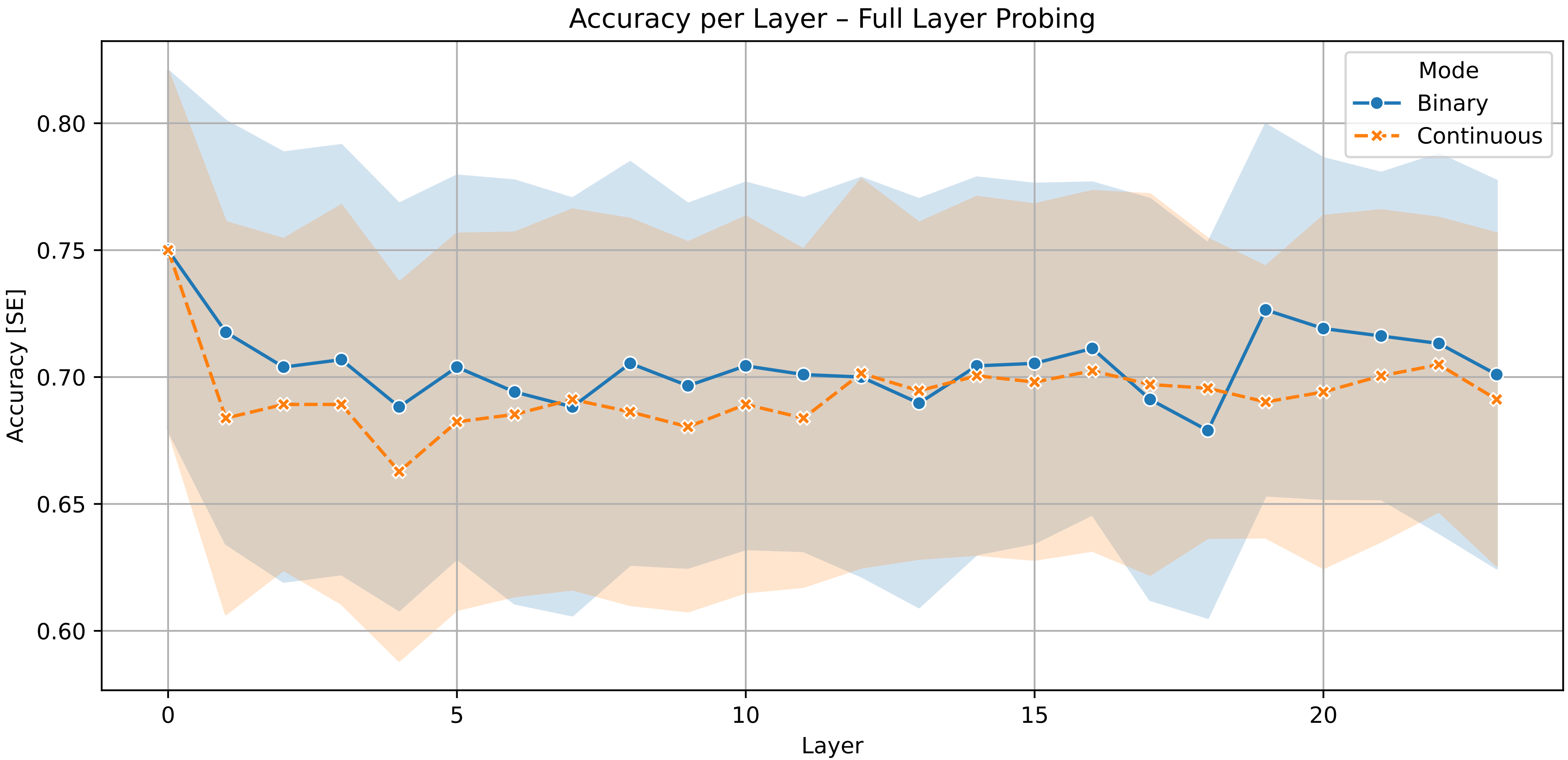


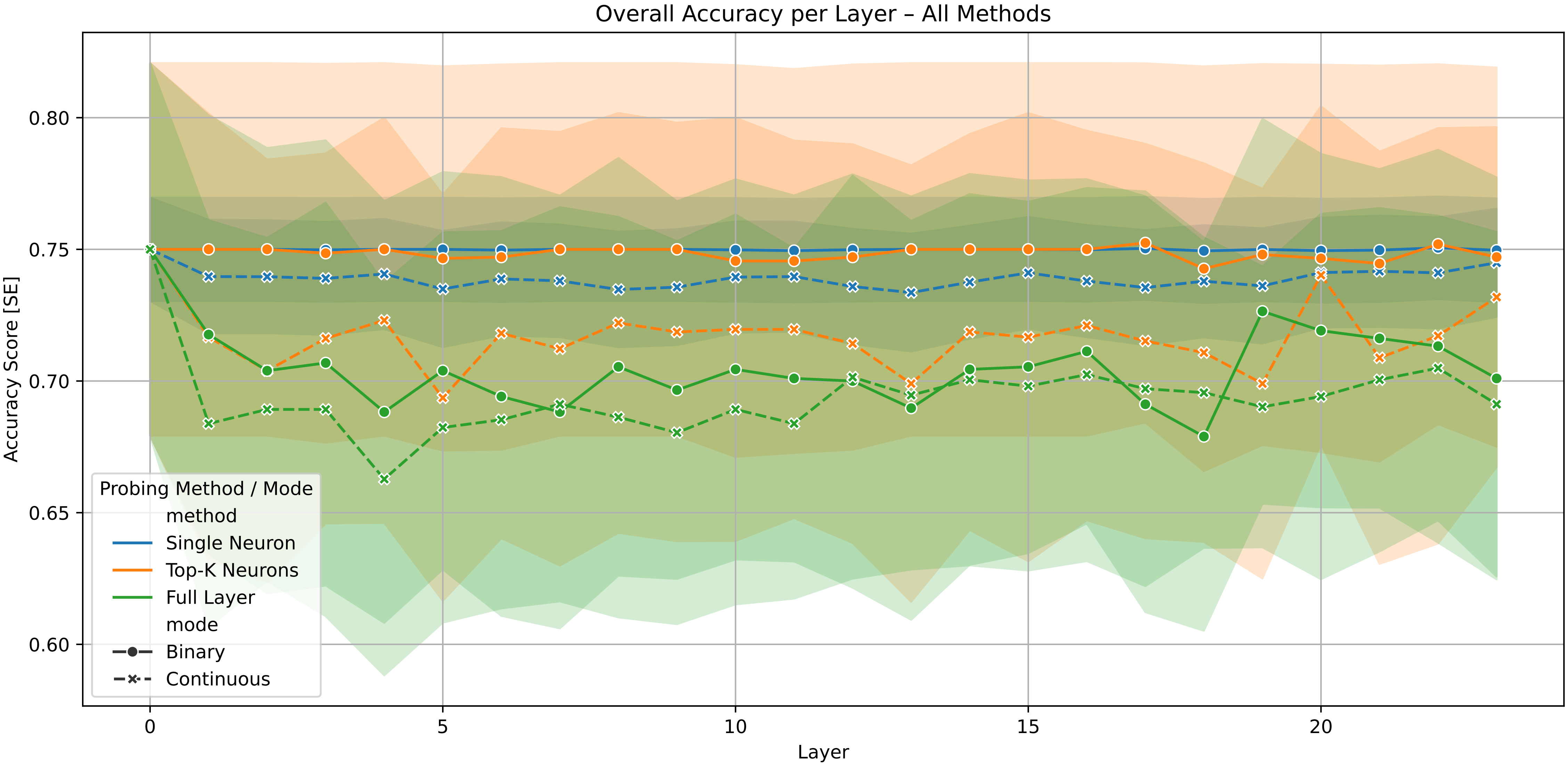
Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	19.0	22.0
Full Layer	f1_max	0.8559	0.8651
Full Layer	f1_mean	0.677	0.6774
Full Layer	f1_std	0.1301	0.1215
Single Neuron	f1_best_layer	19.0	22.0
Single Neuron	f1_max	0.8526	0.8526
Single Neuron	f1_mean	0.6495	0.6692
Single Neuron	f1_std	0.1623	0.139
Top-K Neurons	f1_best_layer	22.0	20.0
Top-K Neurons	f1_max	0.8526	0.8784
Top-K Neurons	f1_mean	0.6551	0.6799
Top-K Neurons	f1_std	0.1565	0.1313


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.77
0.76
0.75
0.74
0.73
0.72
0.71
Layer
0
5
10
15
20

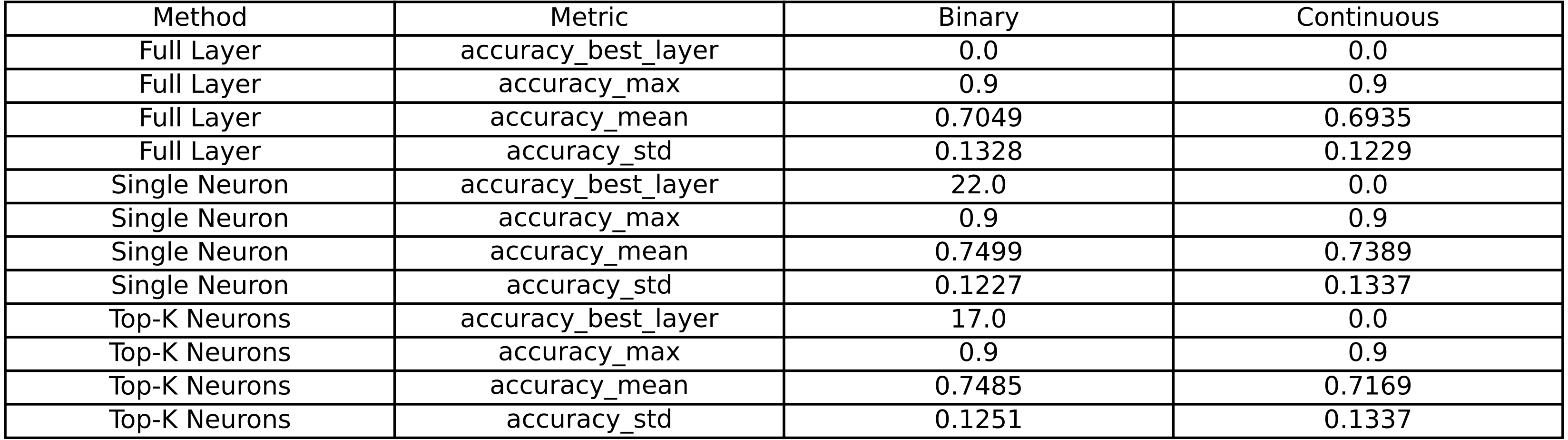


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
0.80
0.75
0.70
0.65
0
5
10
15
20
Layer
Mode
Binary
Continuous


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.80
0.75
0.70
0.65
0.60
Layer
0
5
10
15
20


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
0.60
0.65
0.70
0.75
0.80
Layer
0
5
10
15
20
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

Accuracy Score Summary by Probing Method



Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	0.0	0.0
Full Layer	accuracy_max	0.9	0.9
Full Layer	accuracy_mean	0.7049	0.6935
Full Layer	accuracy_std	0.1328	0.1229
Single Neuron	accuracy_best_layer	22.0	0.0
Single Neuron	accuracy_max	0.9	0.9
Single Neuron	accuracy_mean	0.7499	0.7389
Single Neuron	accuracy_std	0.1227	0.1337
Top-K Neurons	accuracy_best_layer	17.0	0.0
Top-K Neurons	accuracy_max	0.9	0.9
Top-K Neurons	accuracy_mean	0.7485	0.7169
Top-K Neurons	accuracy_std	0.1251	0.1337