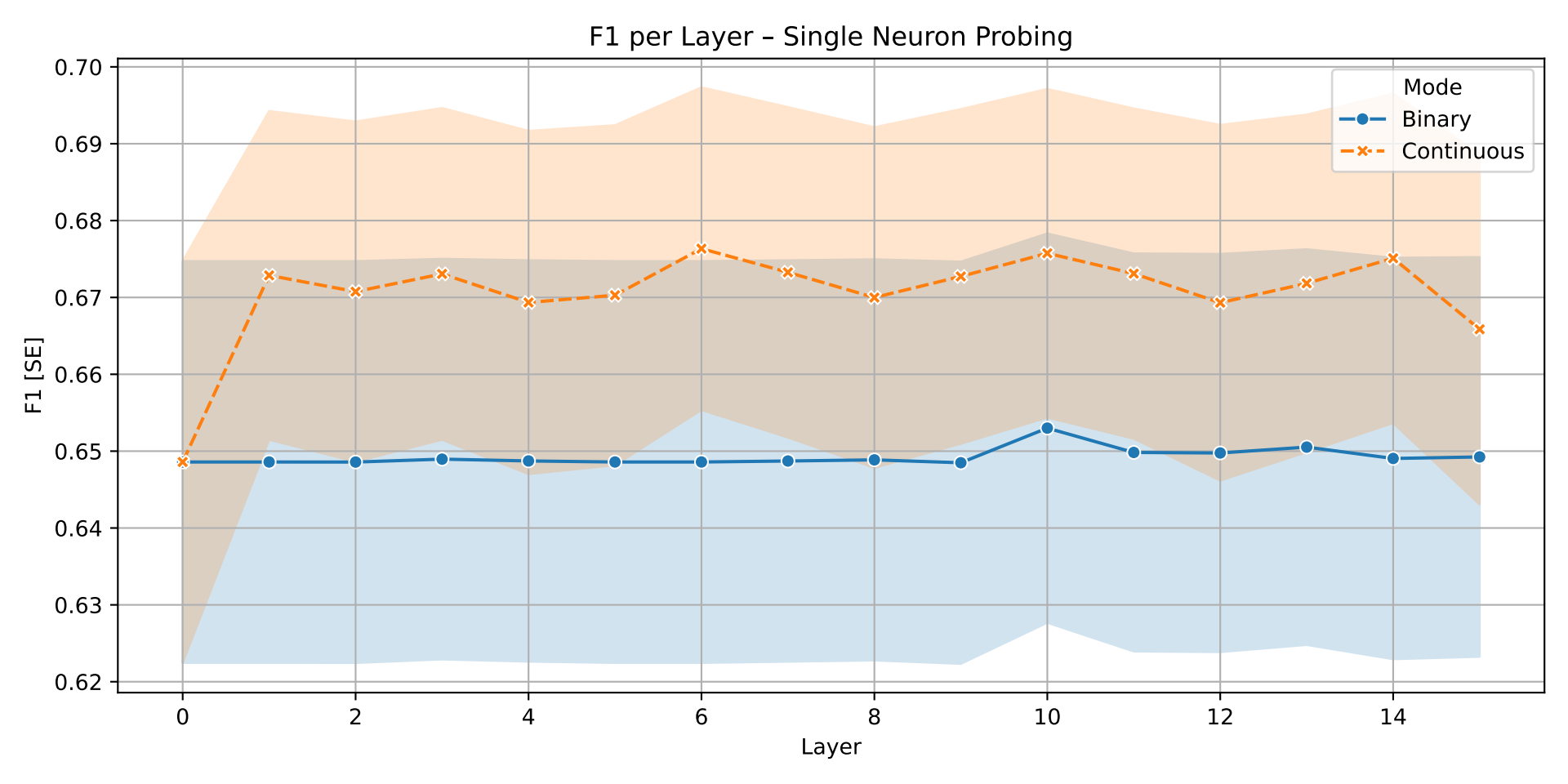
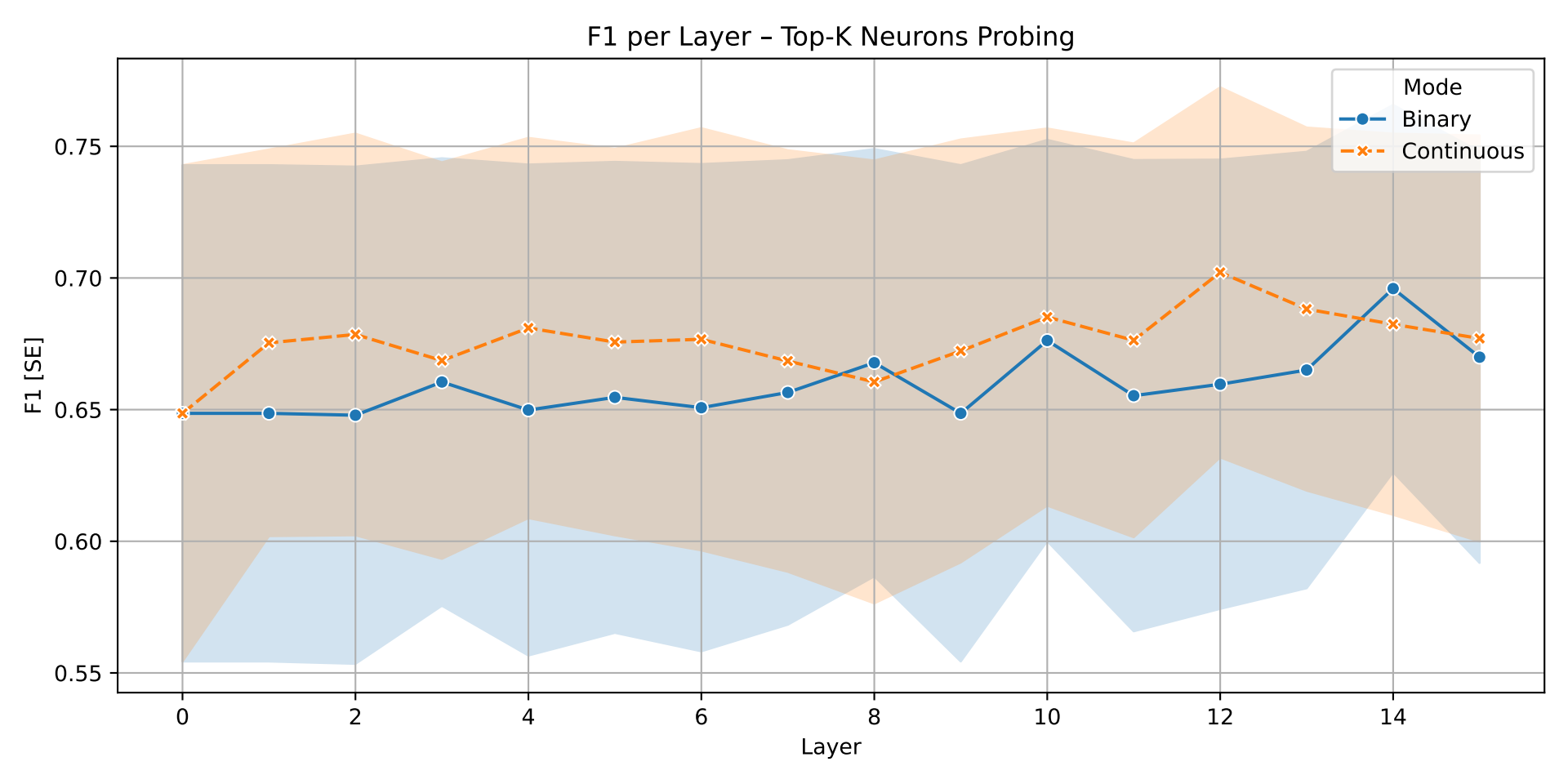
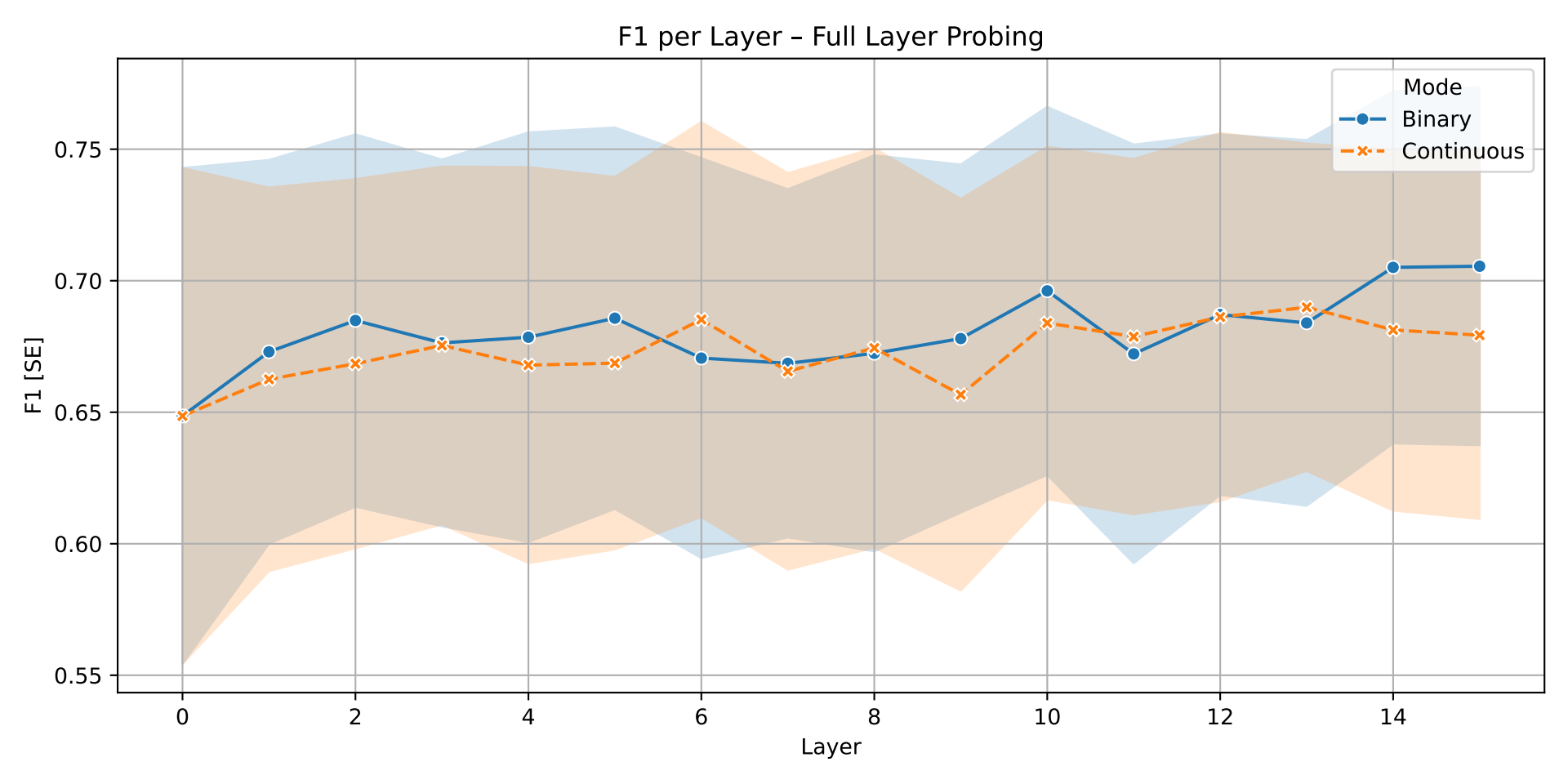
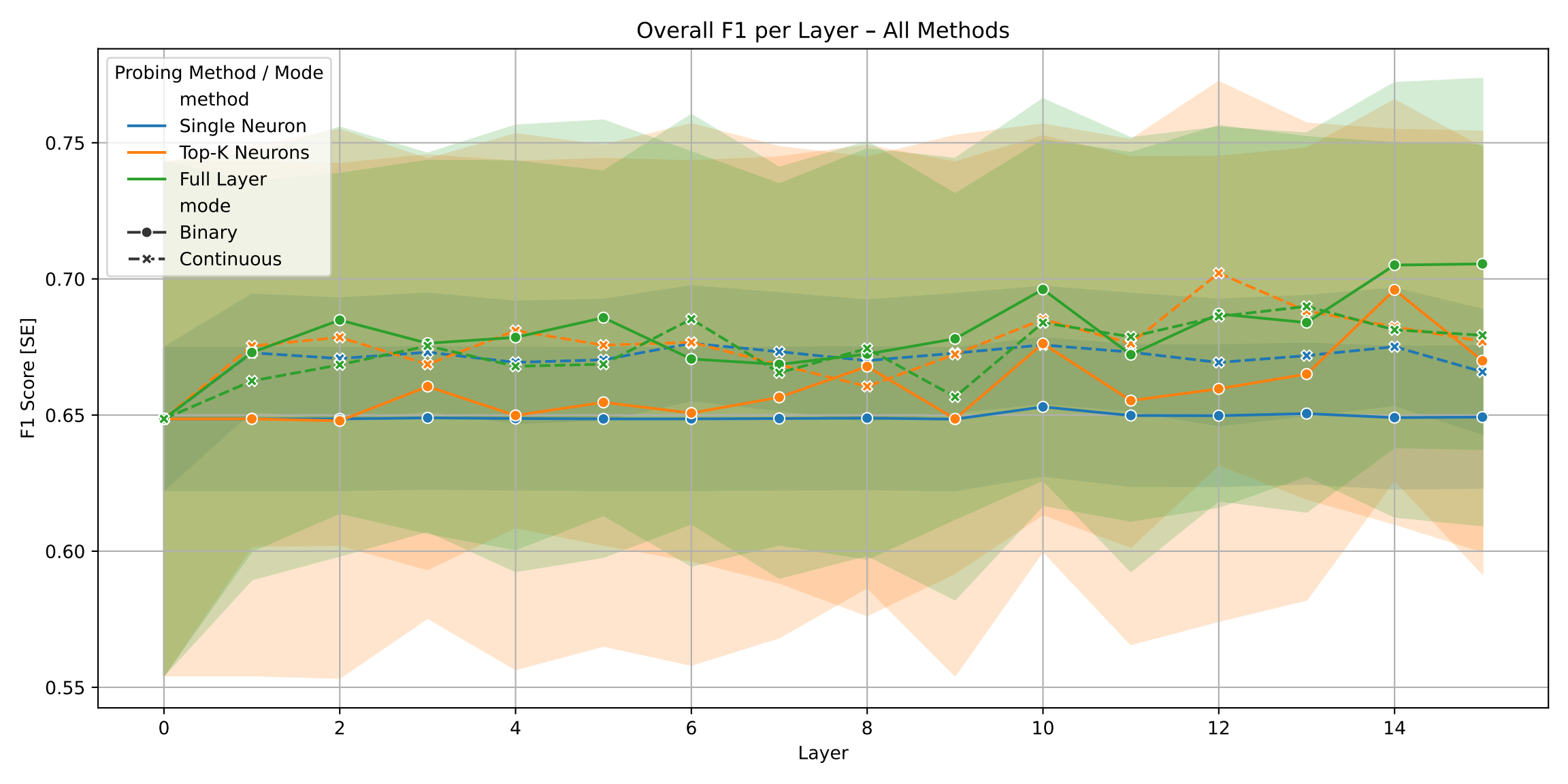

F1 per Layer – Single Neuron Probing
F1 [SE]
Layer
Mode
Binary
Continuous
0.70
0.69
0.68
0.67
0.66
0.65
0.64
0.63
0.62
0
2
4
6
8
10
12
14

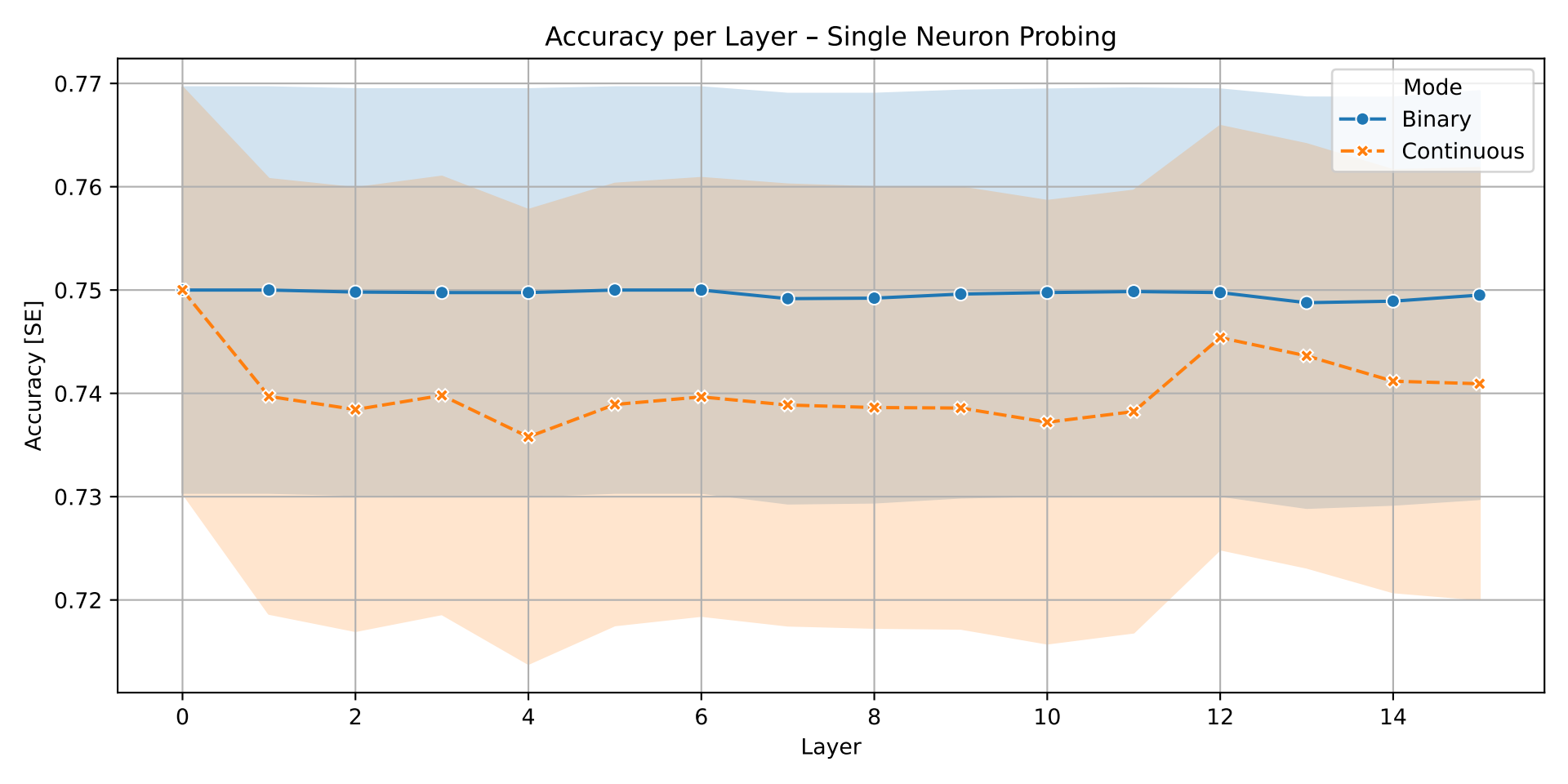

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
12
14
Layer

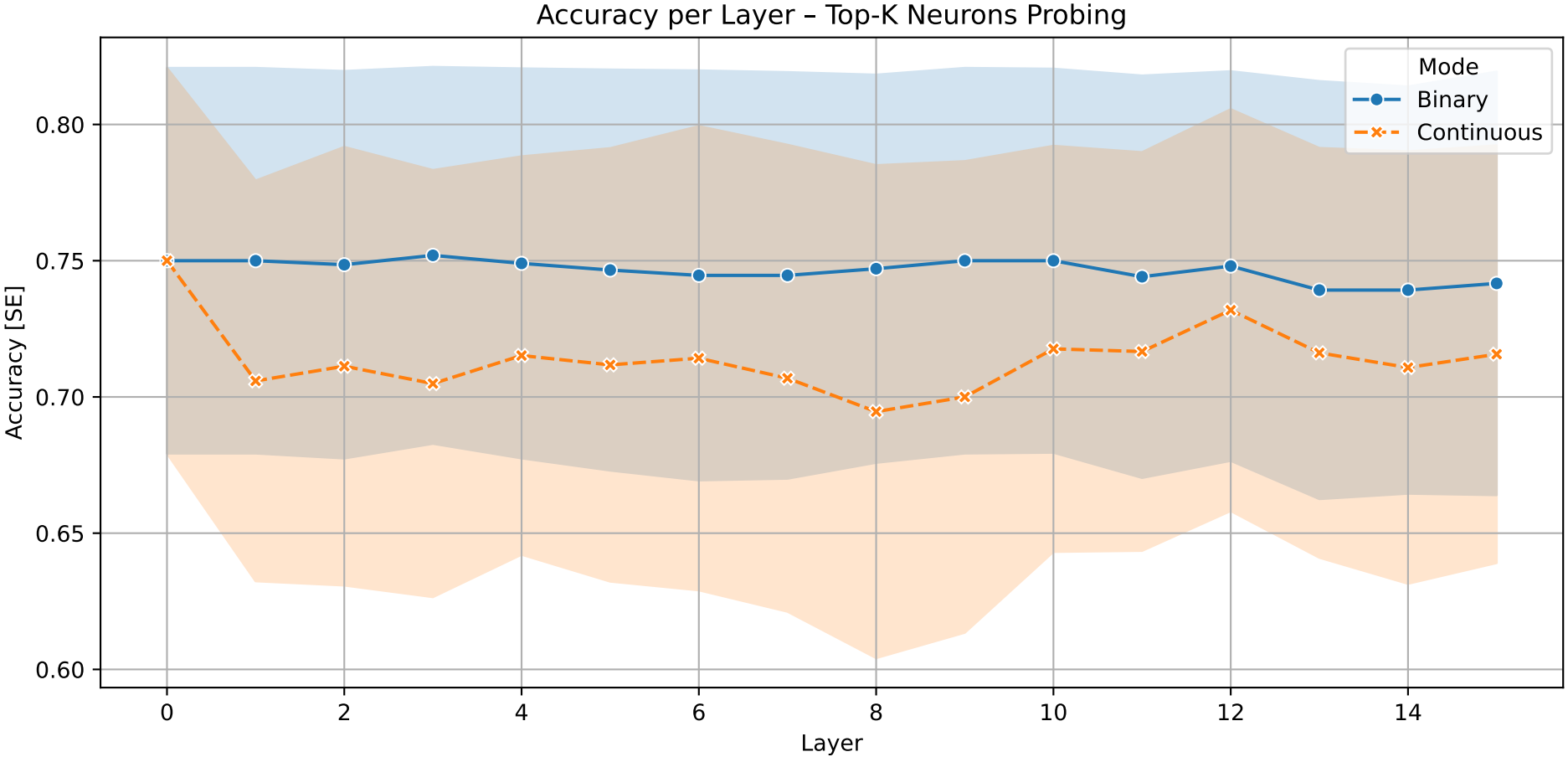

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
12
14
Layer

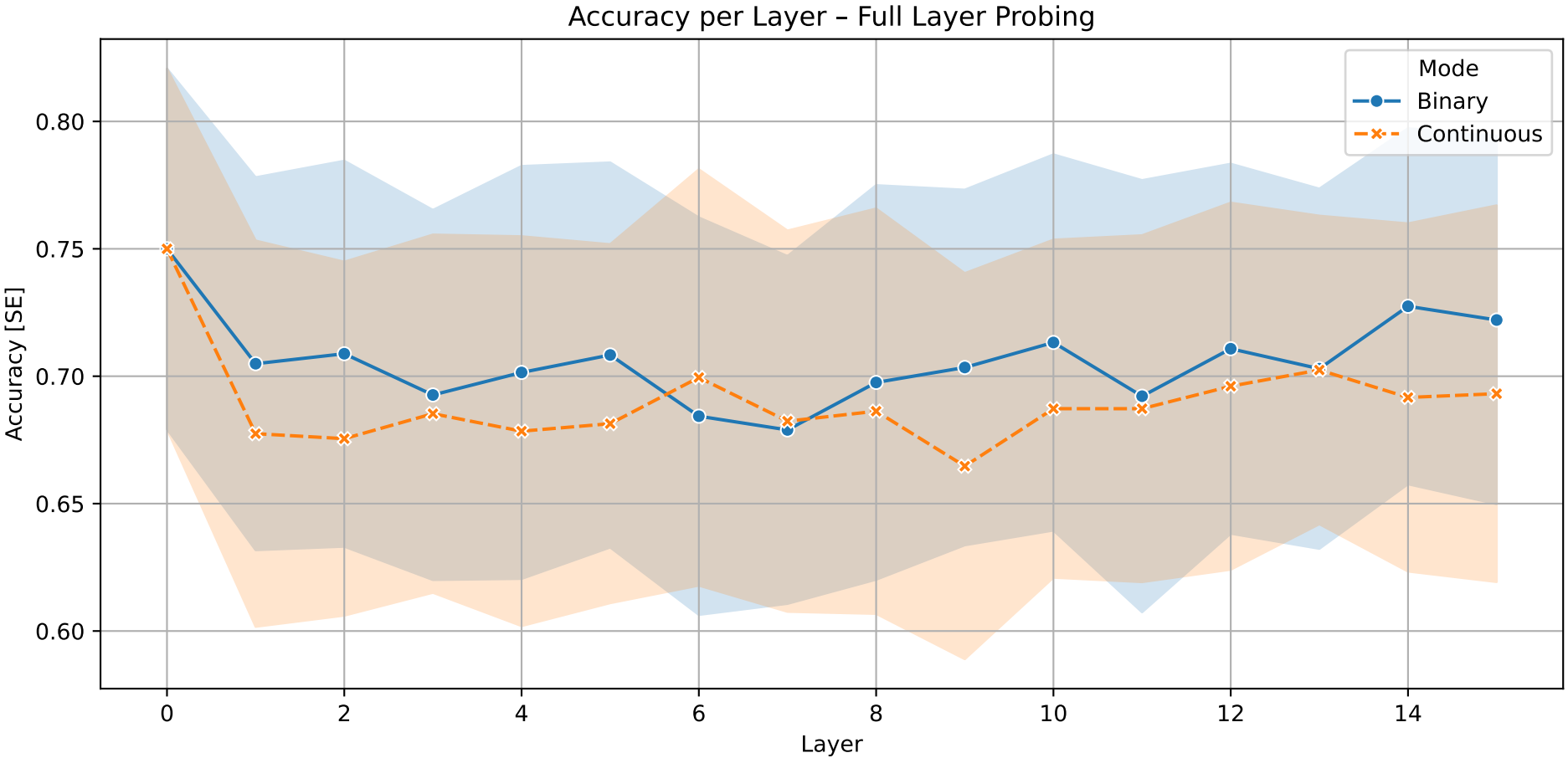

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
12
14
Layer

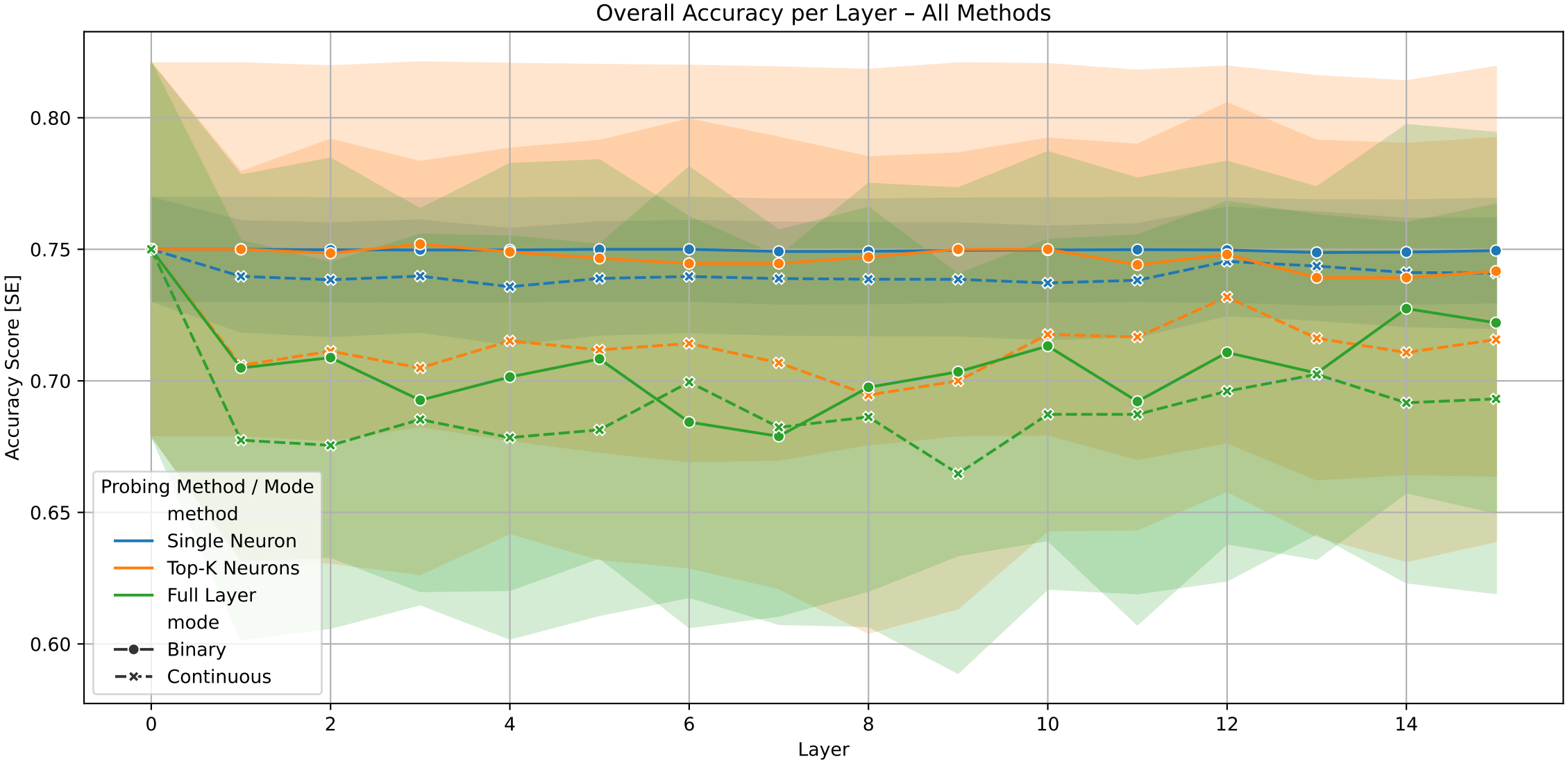
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	15.0	13.0
Full Layer	f1_max	0.8598	0.8562
Full Layer	f1_mean	0.6804	0.6733
Full Layer	f1_std	0.1283	0.1273
Single Neuron	f1_best_layer	10.0	6.0
Single Neuron	f1_max	0.8526	0.8597
Single Neuron	f1_mean	0.6493	0.6705
Single Neuron	f1_std	0.1626	0.1387
Top-K Neurons	f1_best_layer	14.0	12.0
Top-K Neurons	f1_max	0.8526	0.8609
Top-K Neurons	f1_mean	0.6597	0.6761
Top-K Neurons	f1_std	0.1525	0.1345


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
Layer
0.77
0.76
0.75
0.74
0.73
0.72
0
2
4
6
8
10
12
14


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.80
0.75
0.70
0.65
0.60
0
2
4
6
8
10
12
14
Layer


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.80
0.75
0.70
0.65
0.60
0
2
4
6
8
10
12
14
Layer


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
Layer
0.80
0.75
0.70
0.65
0.60
0
2
4
6
8
10
12
14
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	0.7062	0.6899
Full Layer	accuracy_std	0.1307	0.1275
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.9	0.9
Single Neuron	accuracy_mean	0.7496	0.7403
Single Neuron	accuracy_std	0.123	0.1317
Top-K Neurons	accuracy_best_layer	3.0	0.0
Top-K Neurons	accuracy_max	0.9	0.9
Top-K Neurons	accuracy_mean	0.7465	0.714
Top-K Neurons	accuracy_std	0.1271	0.1379