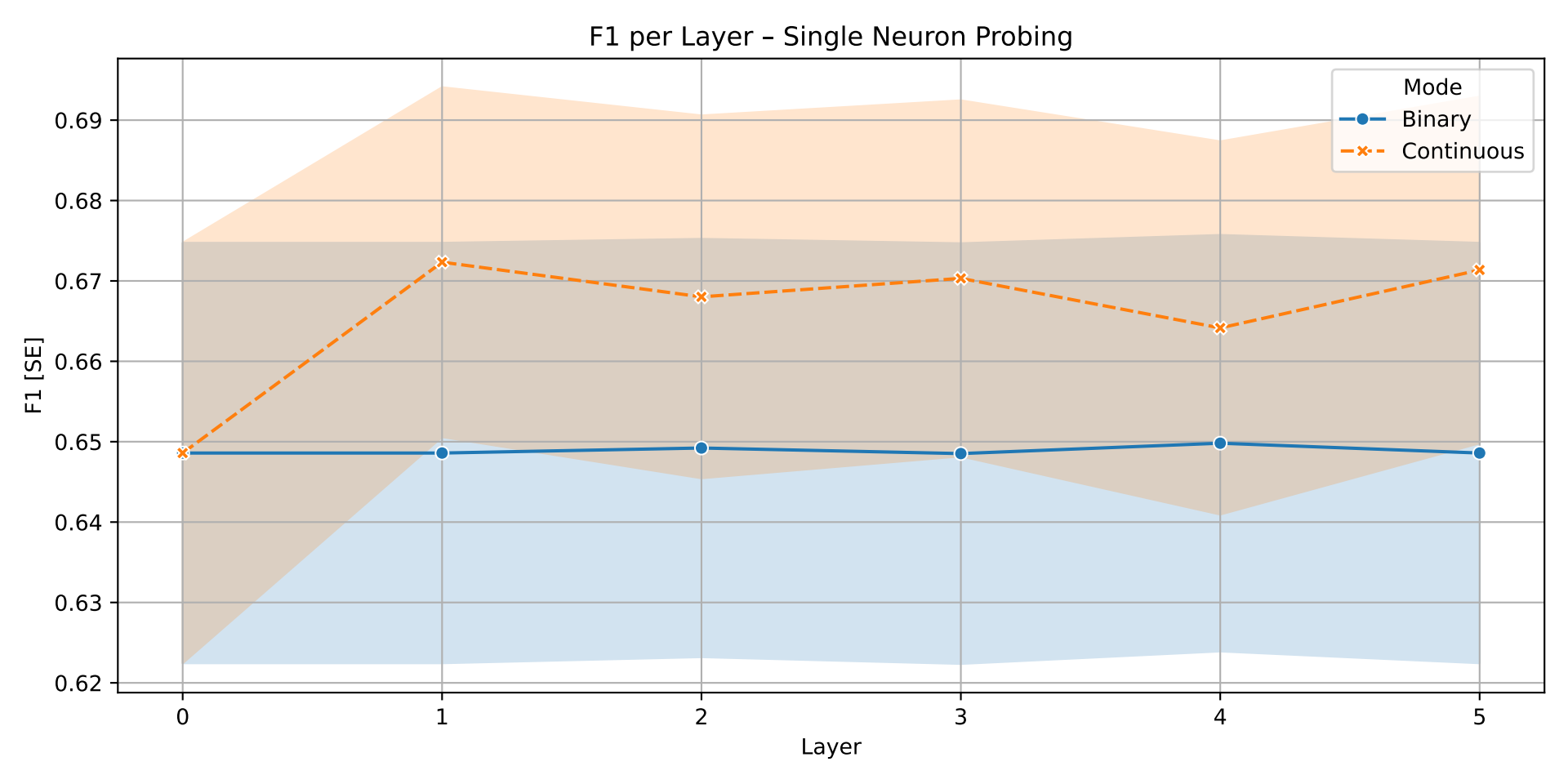
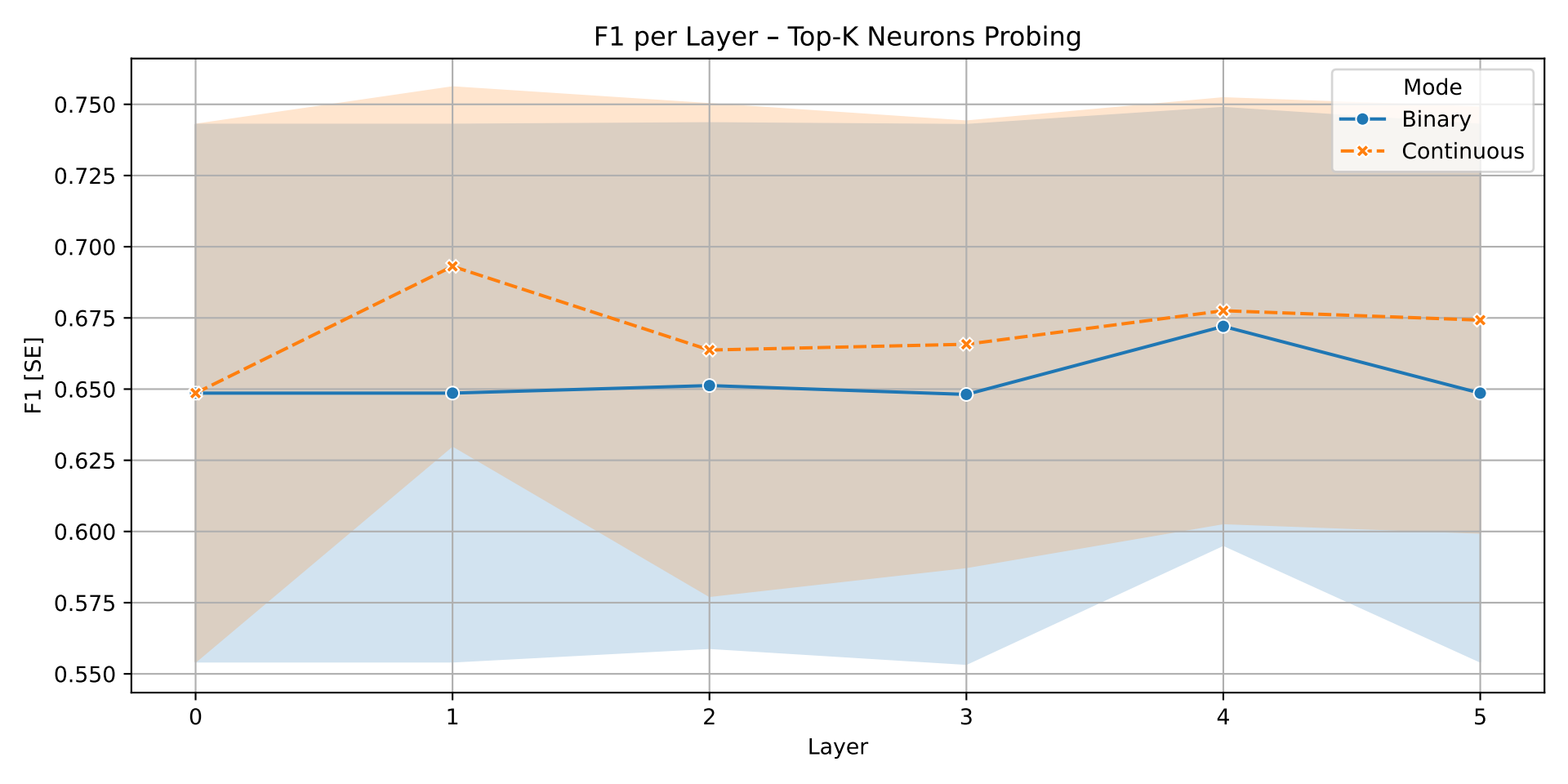
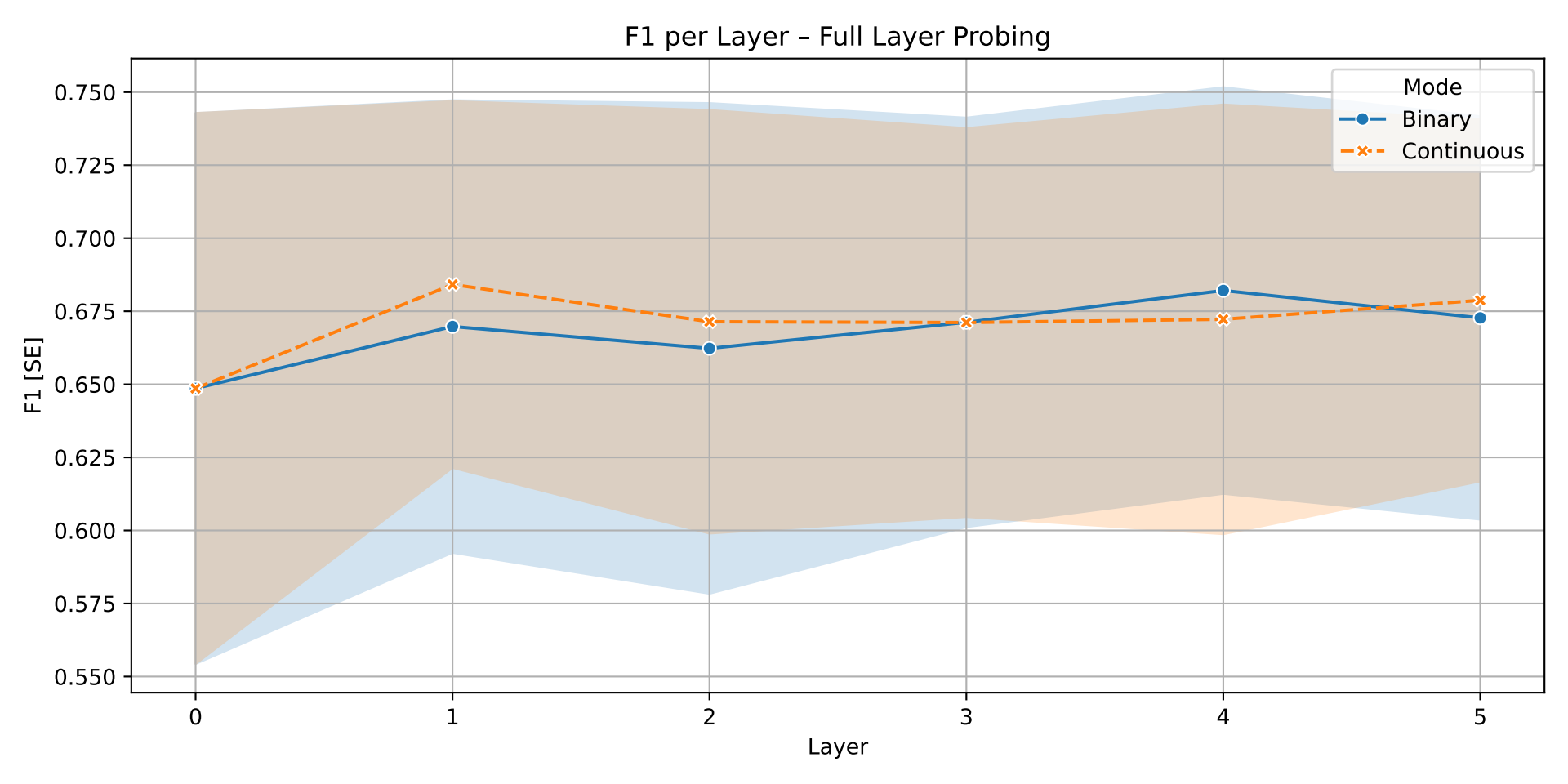
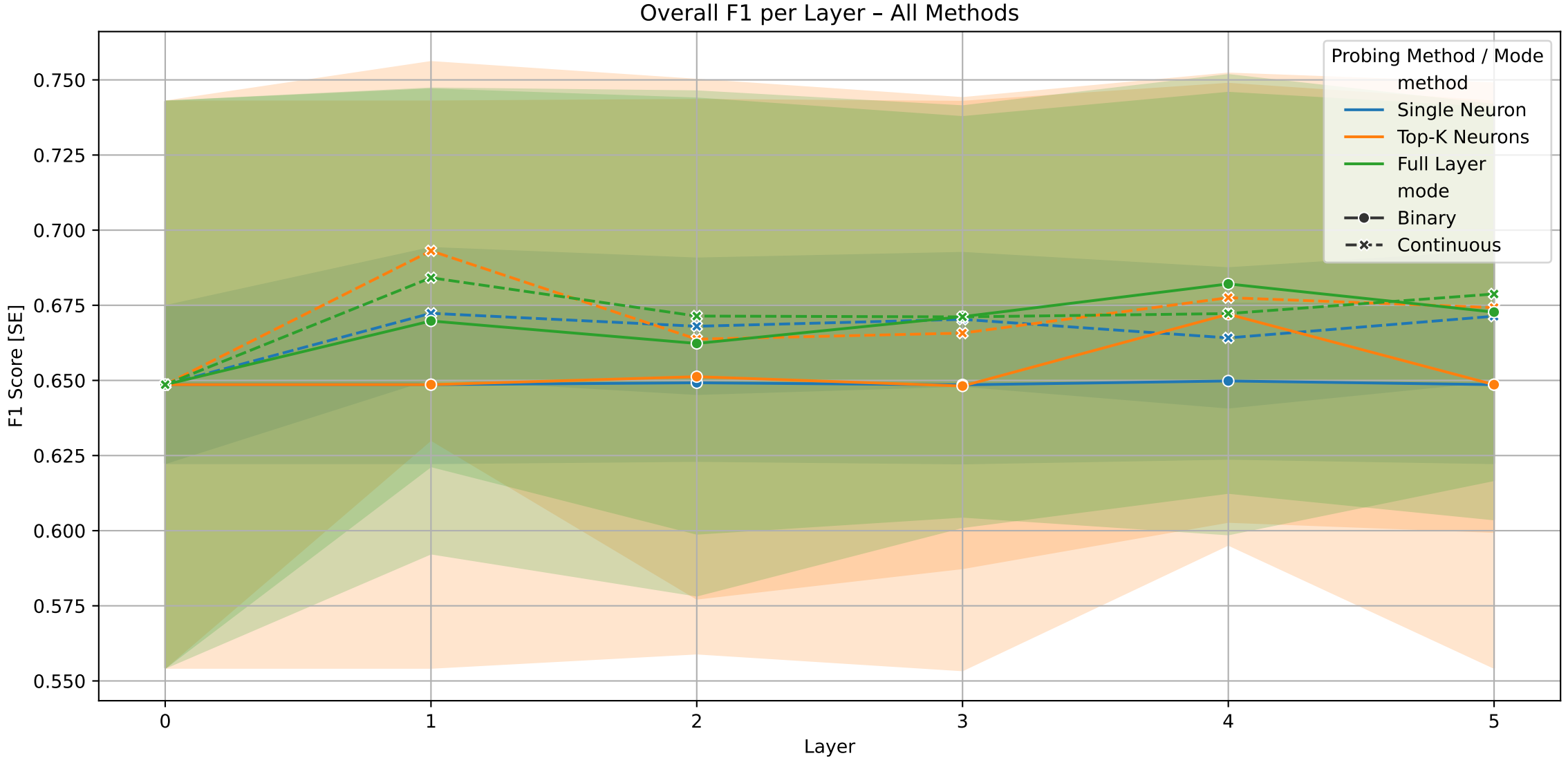

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.69
0.68
0.67
0.66
0.65
0.64
0.63
0.62
0
1
2
3
4
5
Layer

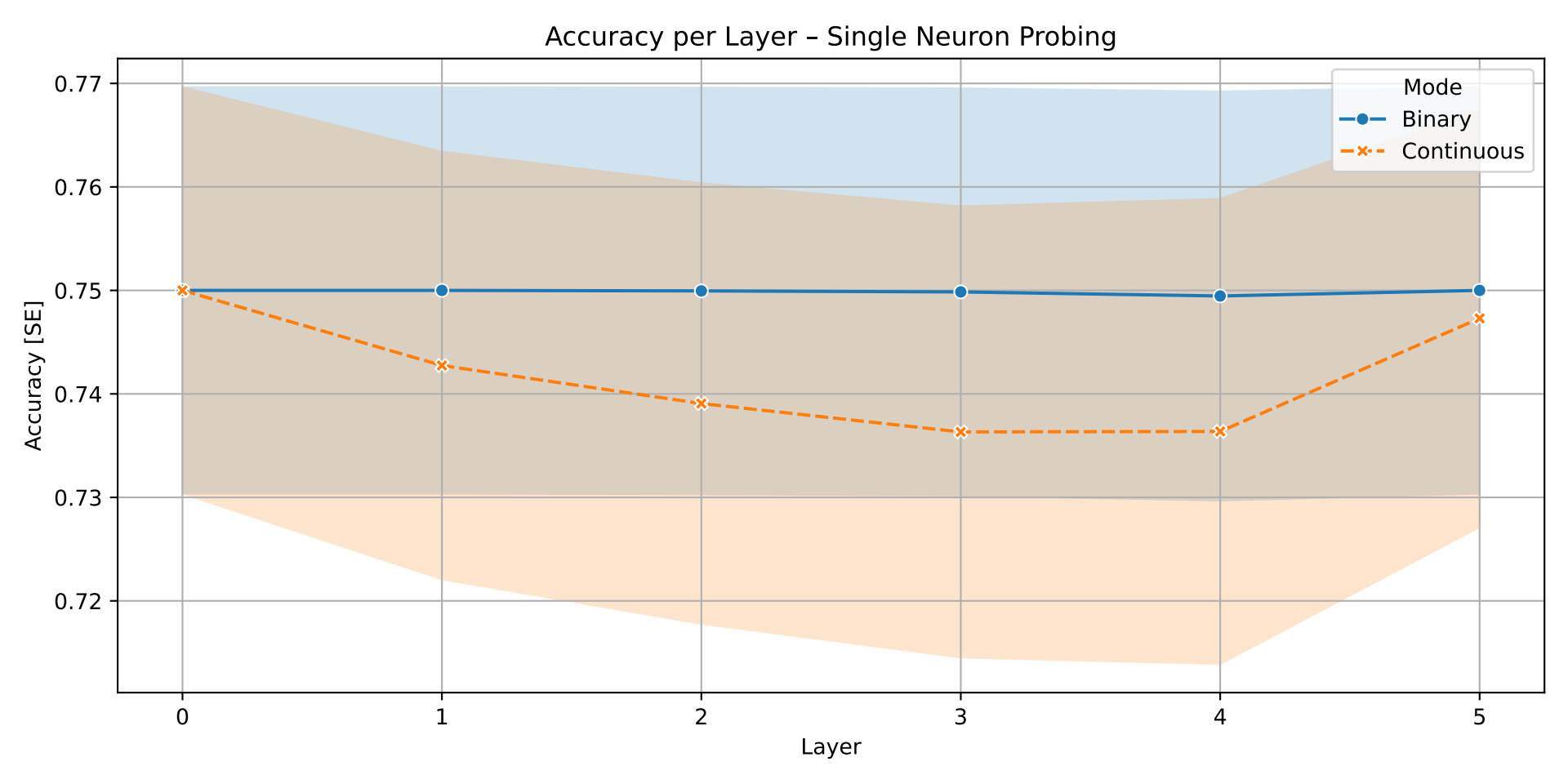

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.750
0.725
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0
1
2
3
4
5
Layer

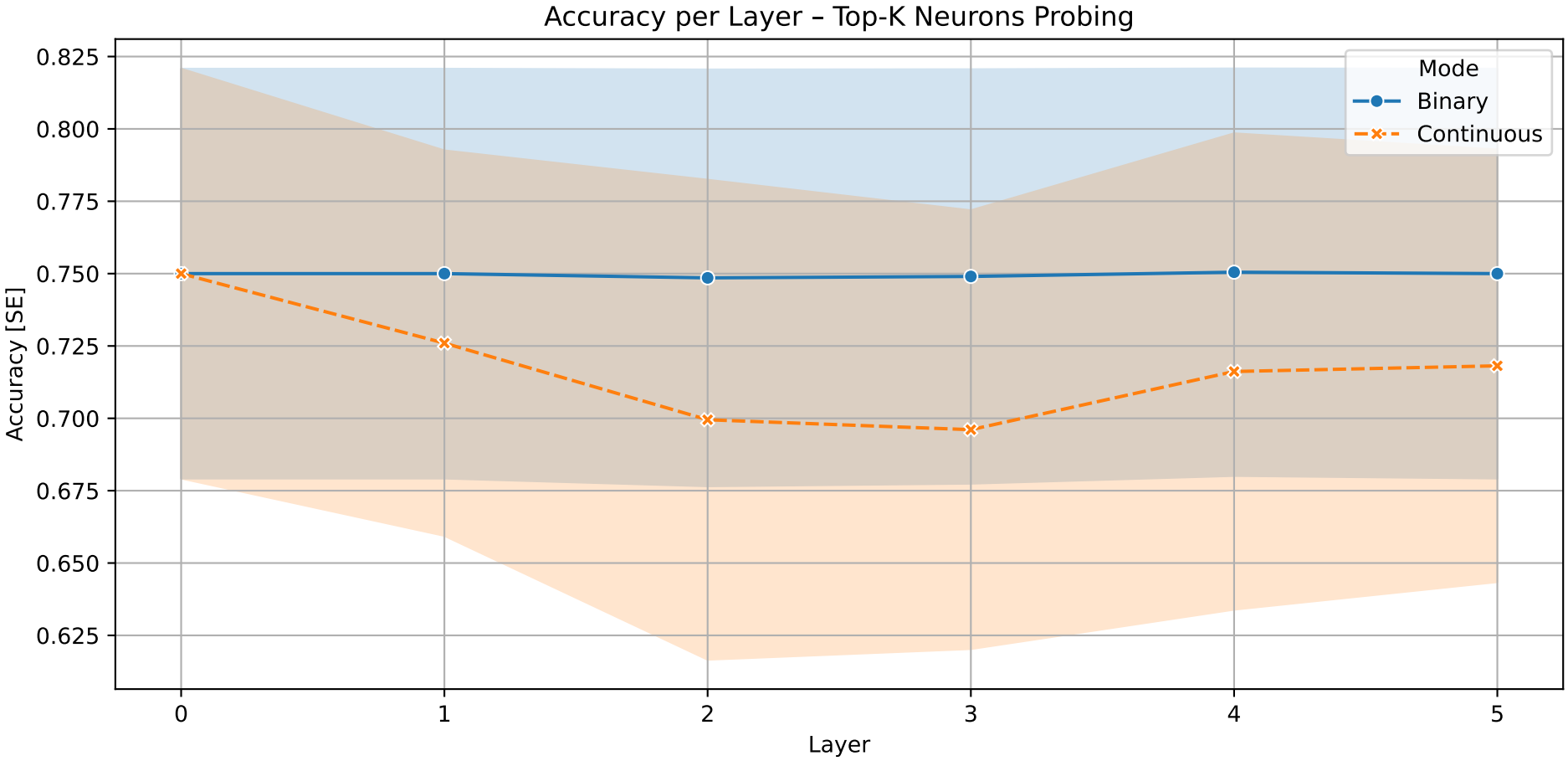

F1 per Layer – Full Layer Probing
F1 [SE]
0.750
0.725
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0
1
2
3
4
5
Layer
Mode
Binary
Continuous

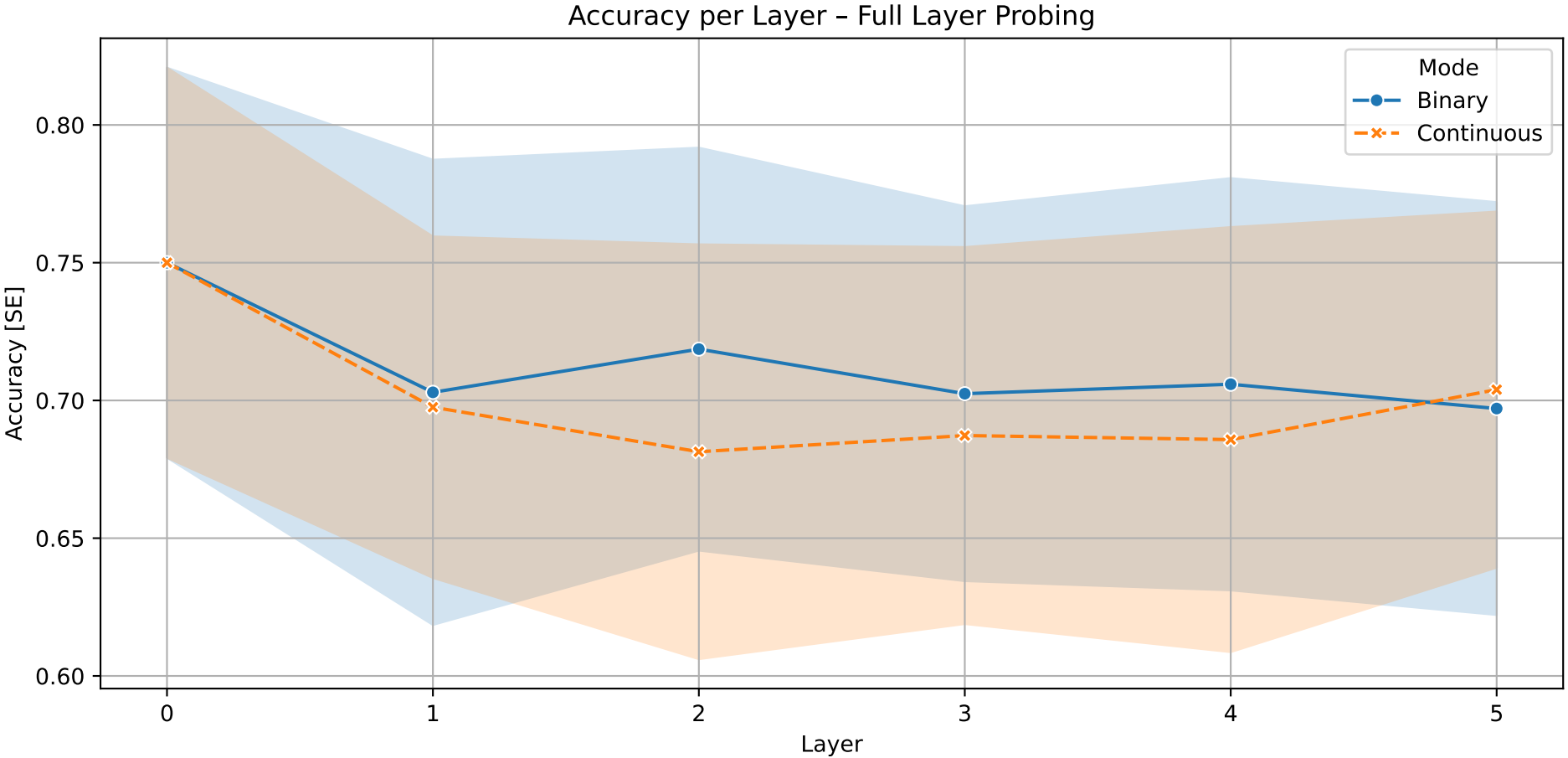

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
0.750
0.725
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0
1
2
3
4
5
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

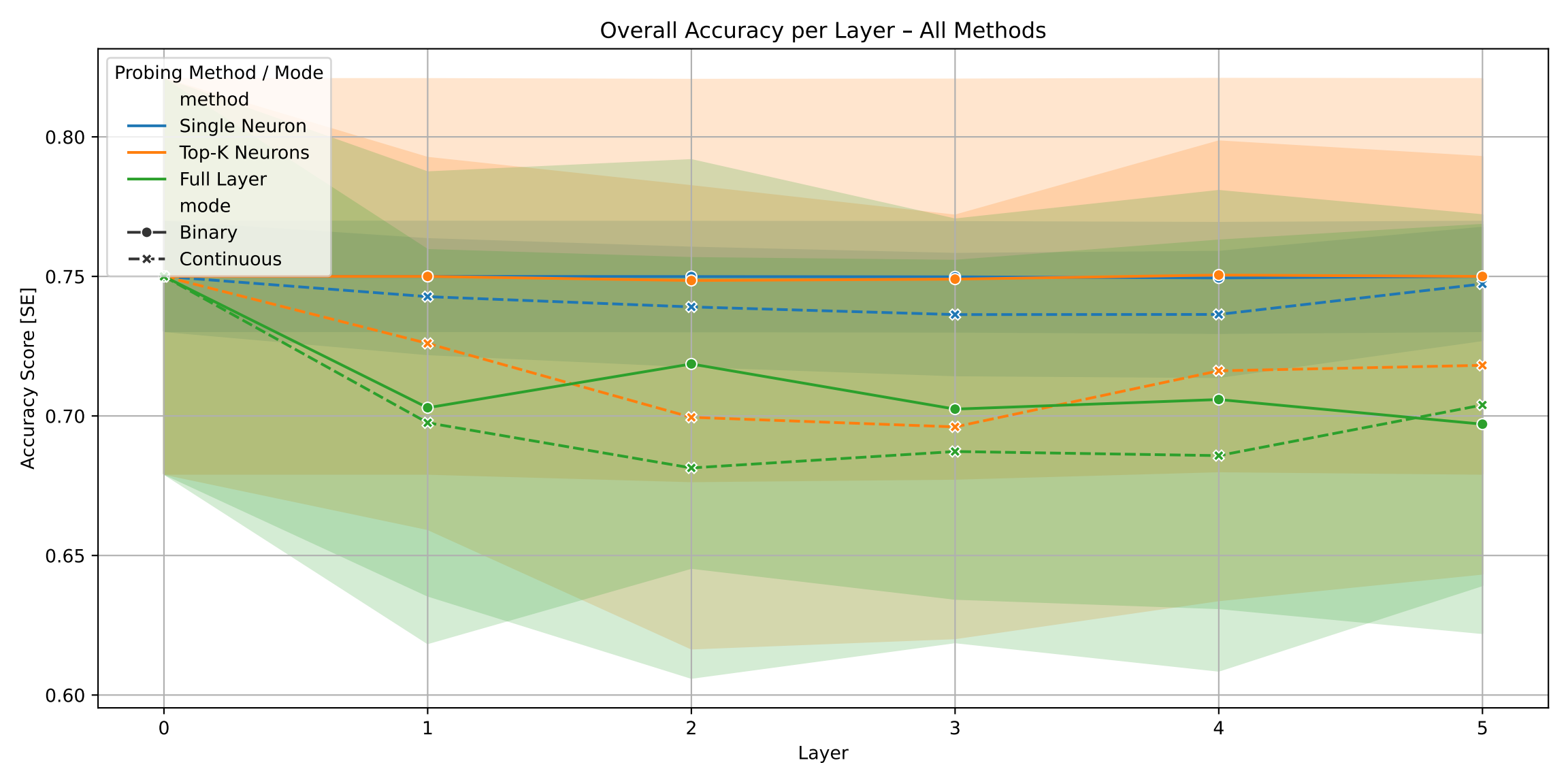
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	4.0	1.0
Full Layer	f1_max	0.8526	0.8526
Full Layer	f1_mean	0.6678	0.6711
Full Layer	f1_std	0.1382	0.1291
Single Neuron	f1_best_layer	4.0	1.0
Single Neuron	f1_max	0.8526	0.8526
Single Neuron	f1_mean	0.6489	0.6658
Single Neuron	f1_std	0.1632	0.1438
Top-K Neurons	f1_best_layer	4.0	1.0
Top-K Neurons	f1_max	0.8526	0.8526
Top-K Neurons	f1_mean	0.6529	0.6705
Top-K Neurons	f1_std	0.1617	0.1406


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.77
0.76
0.75
0.74
0.73
0.72
0
1
2
3
4
5
Layer


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
0.825
0.800
0.775
0.750
0.725
0.700
0.675
0.650
0.625
0
1
2
3
4
5
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
0
1
2
3
4
5
Layer


Overall Accuracy per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
Accuracy Score [SE]
0.80
0.75
0.70
0.65
0.60
0
1
2
3
4
5
Layer

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	0.7128	0.701
Full Layer	accuracy_std	0.133	0.1259
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.9	0.9
Single Neuron	accuracy_mean	0.7499	0.742
Single Neuron	accuracy_std	0.123	0.1316
Top-K Neurons	accuracy_best_layer	4.0	0.0
Top-K Neurons	accuracy_max	0.9	0.9
Top-K Neurons	accuracy_mean	0.7497	0.7176
Top-K Neurons	accuracy_std	0.1256	0.1351