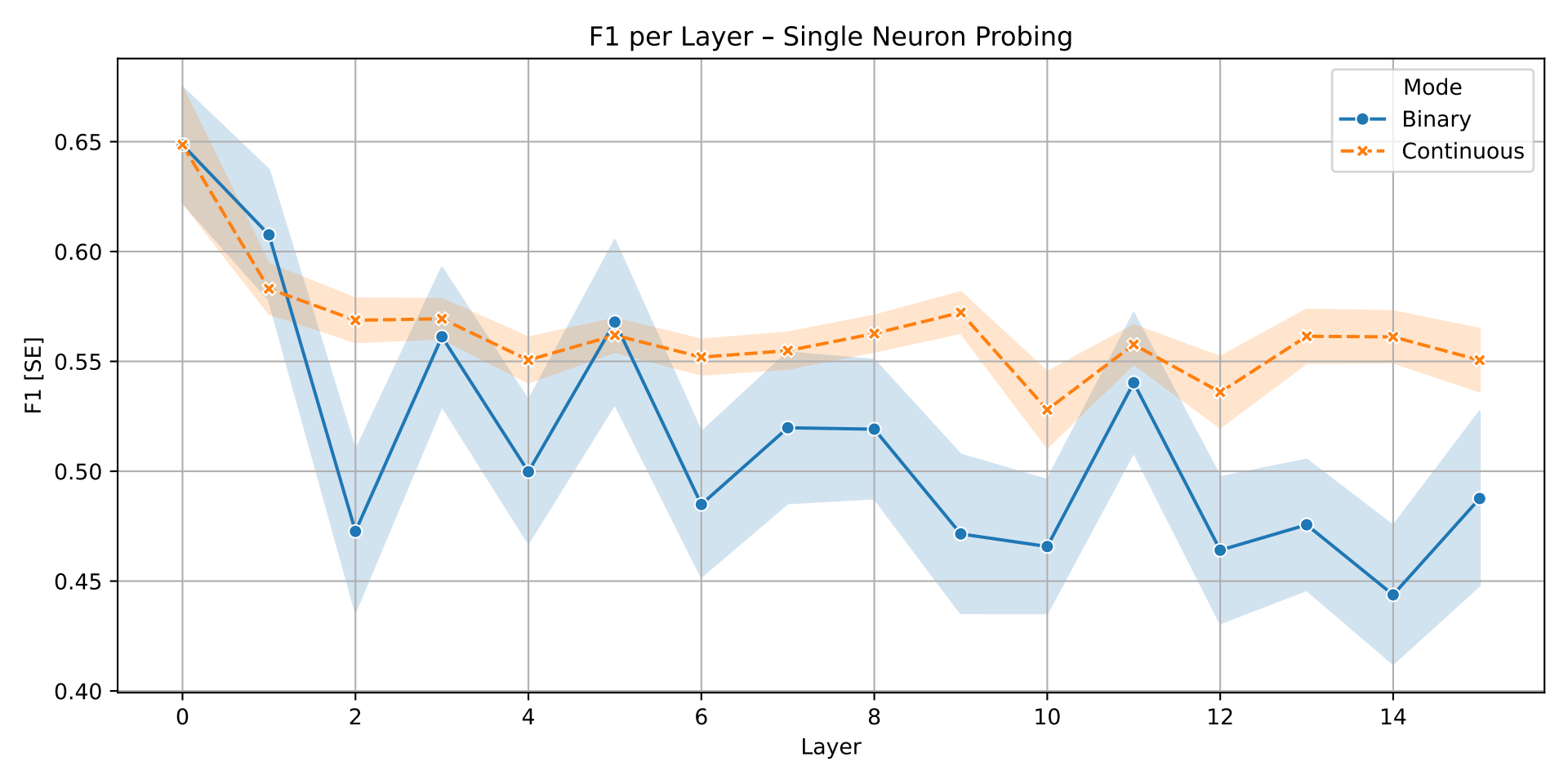
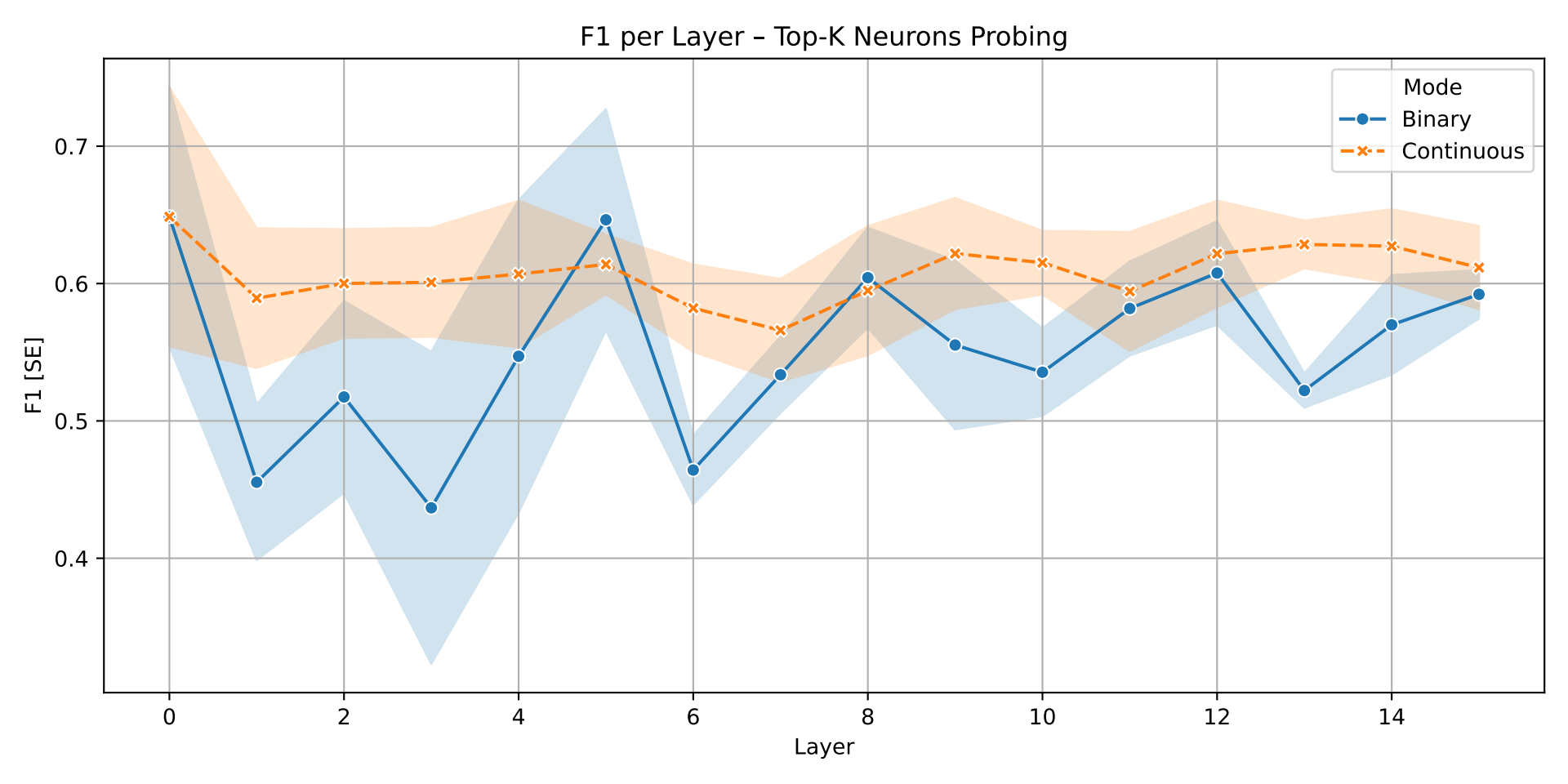
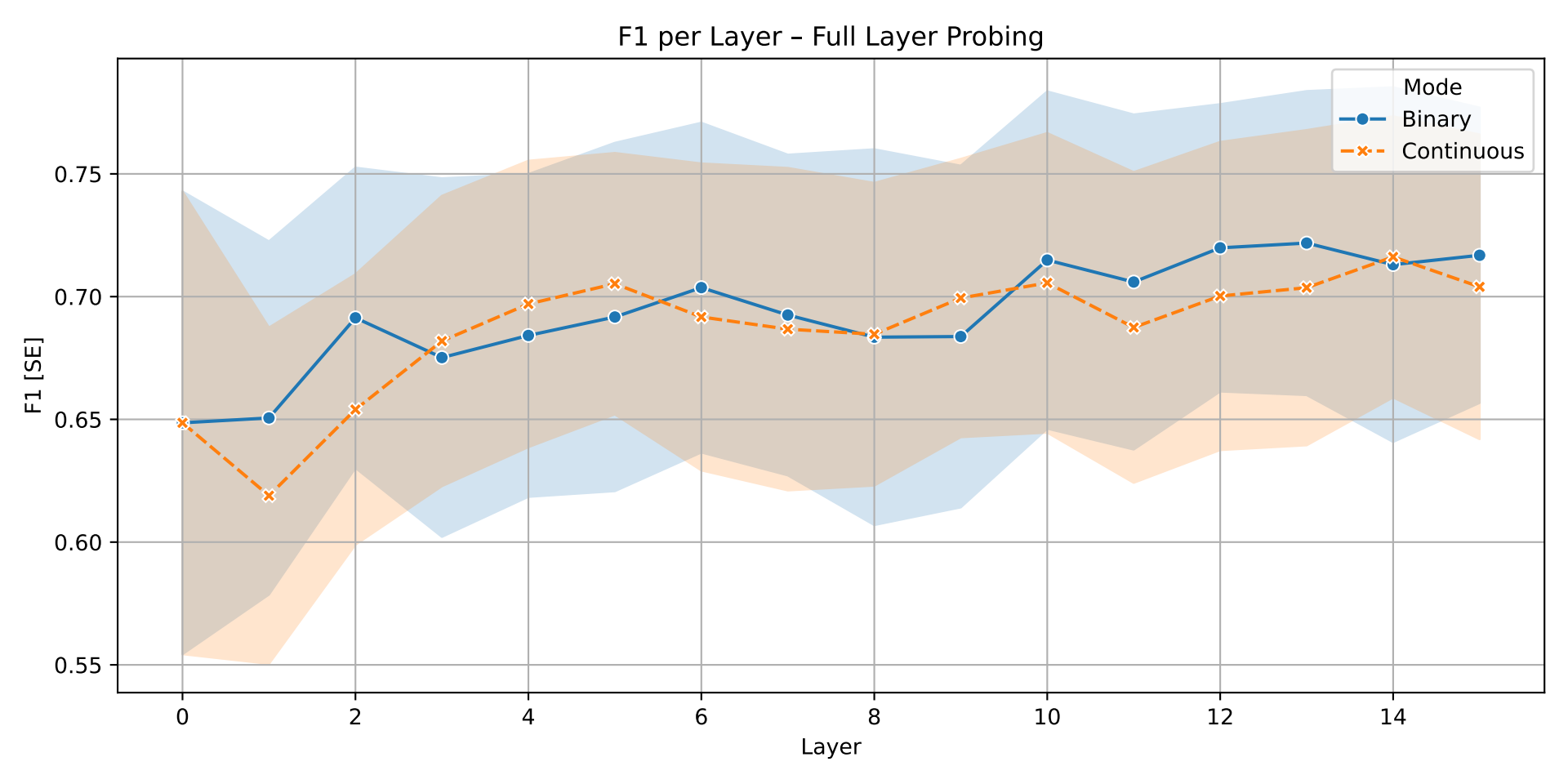
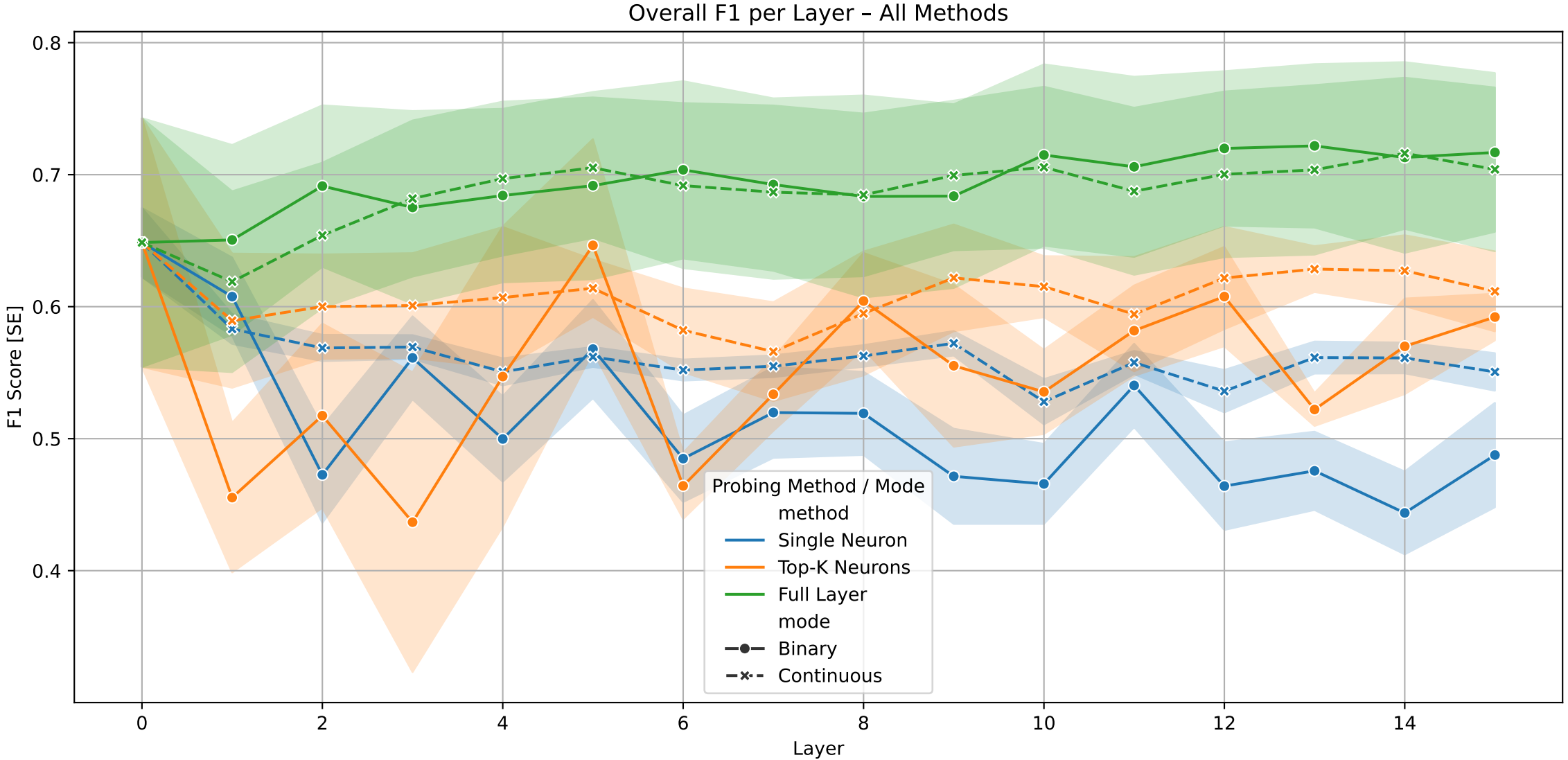

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.65
0.60
0.55
0.50
0.45
0.40
0
2
4
6
8
10
12
14
Layer

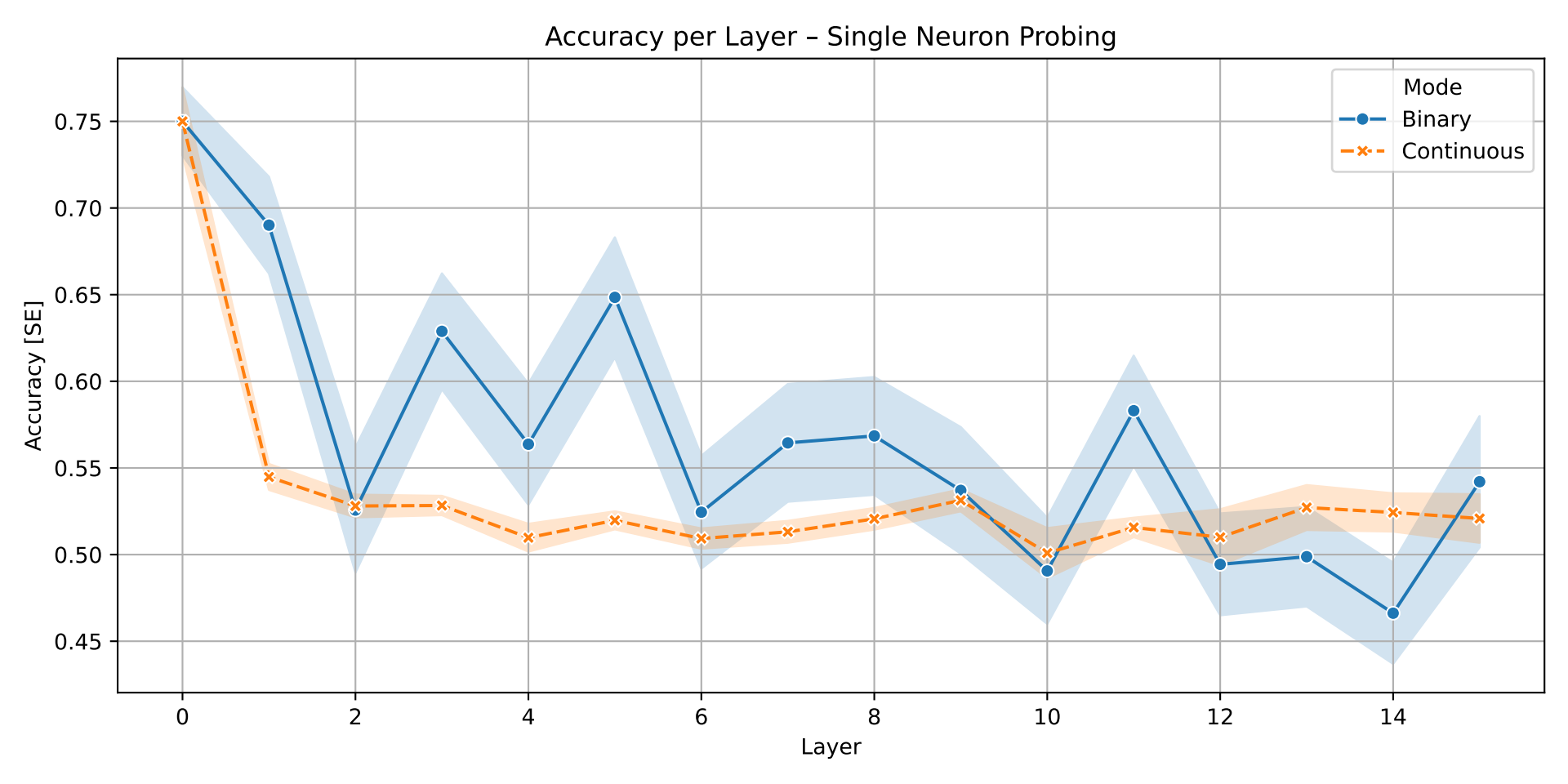

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.7
0.6
0.5
0.4
0
2
4
6
8
10
12
14
Layer

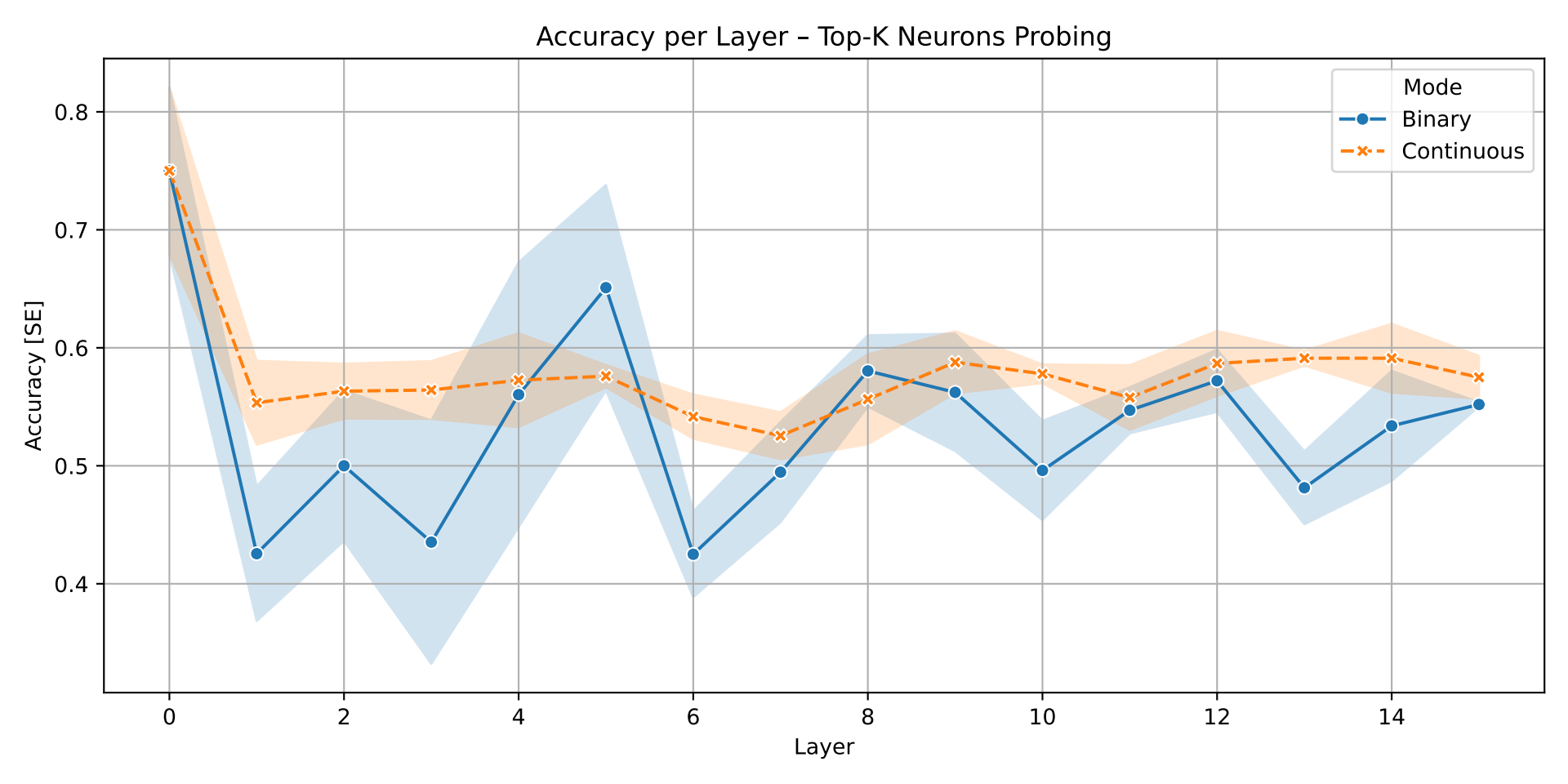

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
12
14
Layer

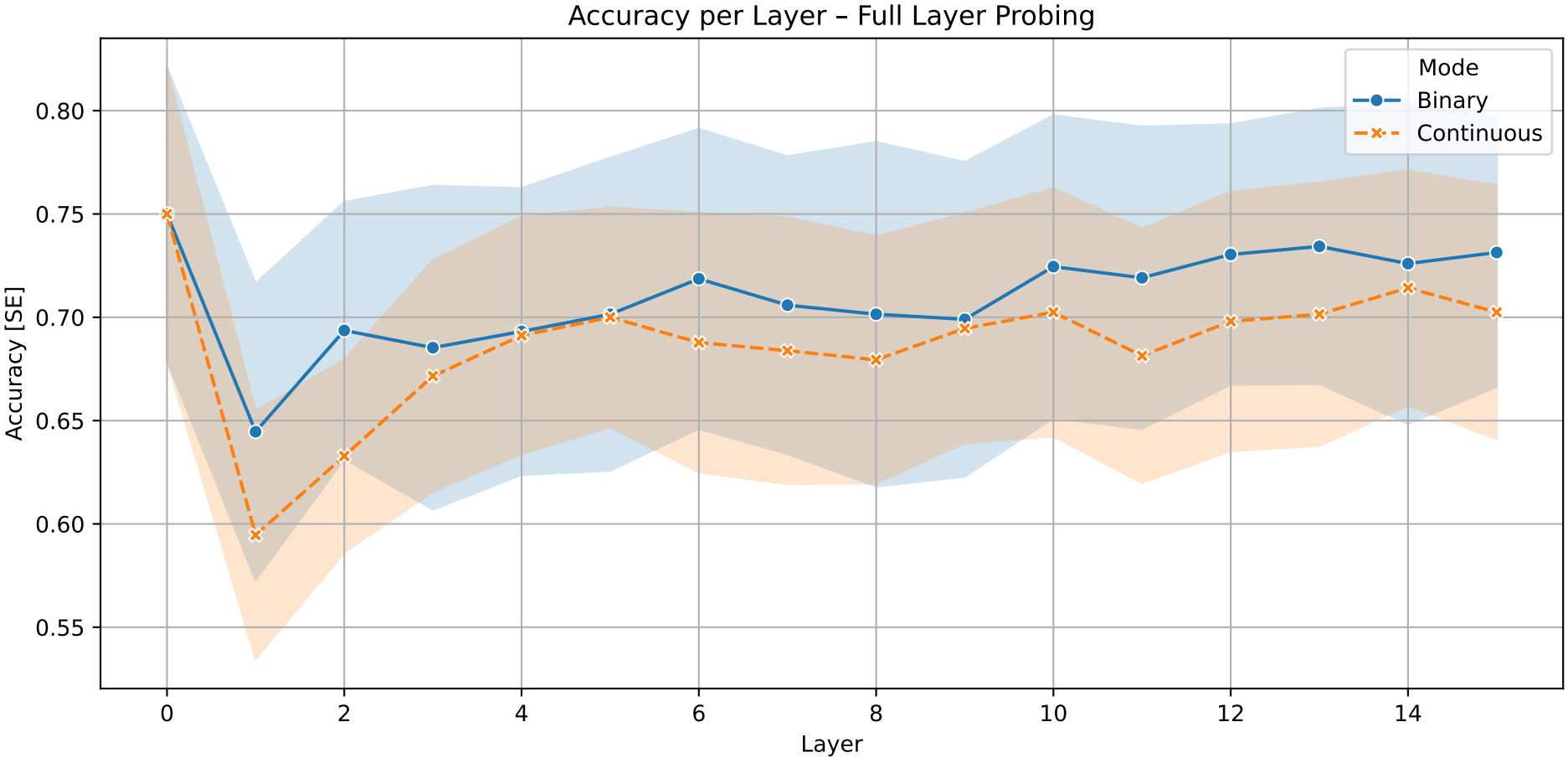

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
0.8
0.7
0.6
0.5
0.4
0
2
4
6
8
10
12
14
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

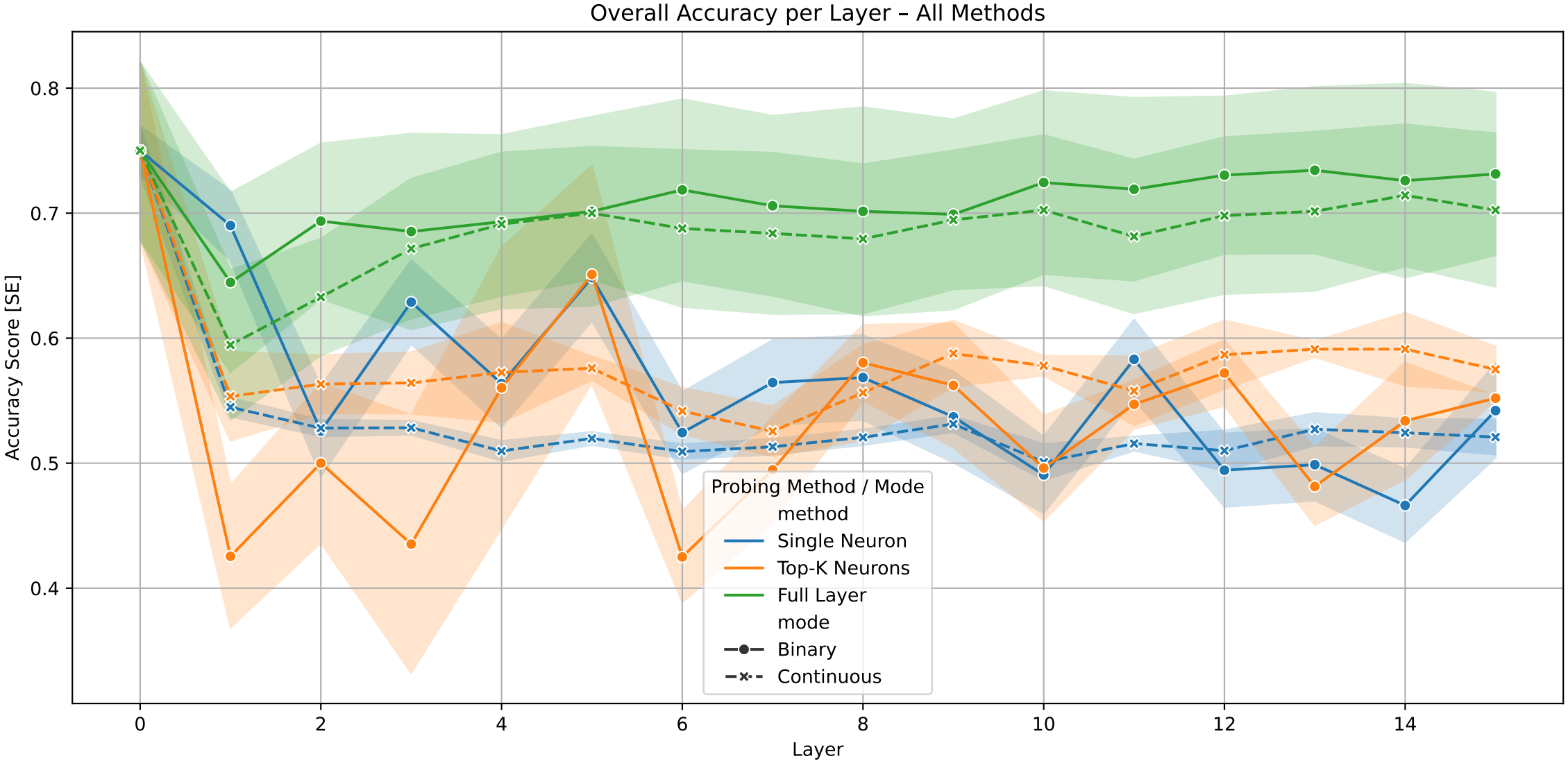
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	13.0	14.0
Full Layer	f1_max	0.8778	0.8625
Full Layer	f1_mean	0.6936	0.6866
Full Layer	f1_std	0.1234	0.1136
Single Neuron	f1_best_layer	0.0	0.0
Single Neuron	f1_max	0.8608	0.8526
Single Neuron	f1_mean	0.5144	0.5637
Single Neuron	f1_std	0.2129	0.0829
Top-K Neurons	f1_best_layer	0.0	0.0
Top-K Neurons	f1_max	0.8526	0.8526
Top-K Neurons	f1_mean	0.5511	0.6077
Top-K Neurons	f1_std	0.1243	0.0782


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0
2
4
6
8
10
12
14
Layer


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.8
0.7
0.6
0.5
0.4
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
0.55
0
2
4
6
8
10
12
14
Layer


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
Layer
0.8
0.7
0.6
0.5
0.4
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	0.902	0.9
Full Layer	accuracy_mean	0.7099	0.6866
Full Layer	accuracy_std	0.1282	0.1095
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.902	0.9
Single Neuron	accuracy_mean	0.5673	0.5346
Single Neuron	accuracy_std	0.215	0.0863
Top-K Neurons	accuracy_best_layer	0.0	0.0
Top-K Neurons	accuracy_max	0.9	0.9
Top-K Neurons	accuracy_mean	0.5354	0.5794
Top-K Neurons	accuracy_std	0.1317	0.0714