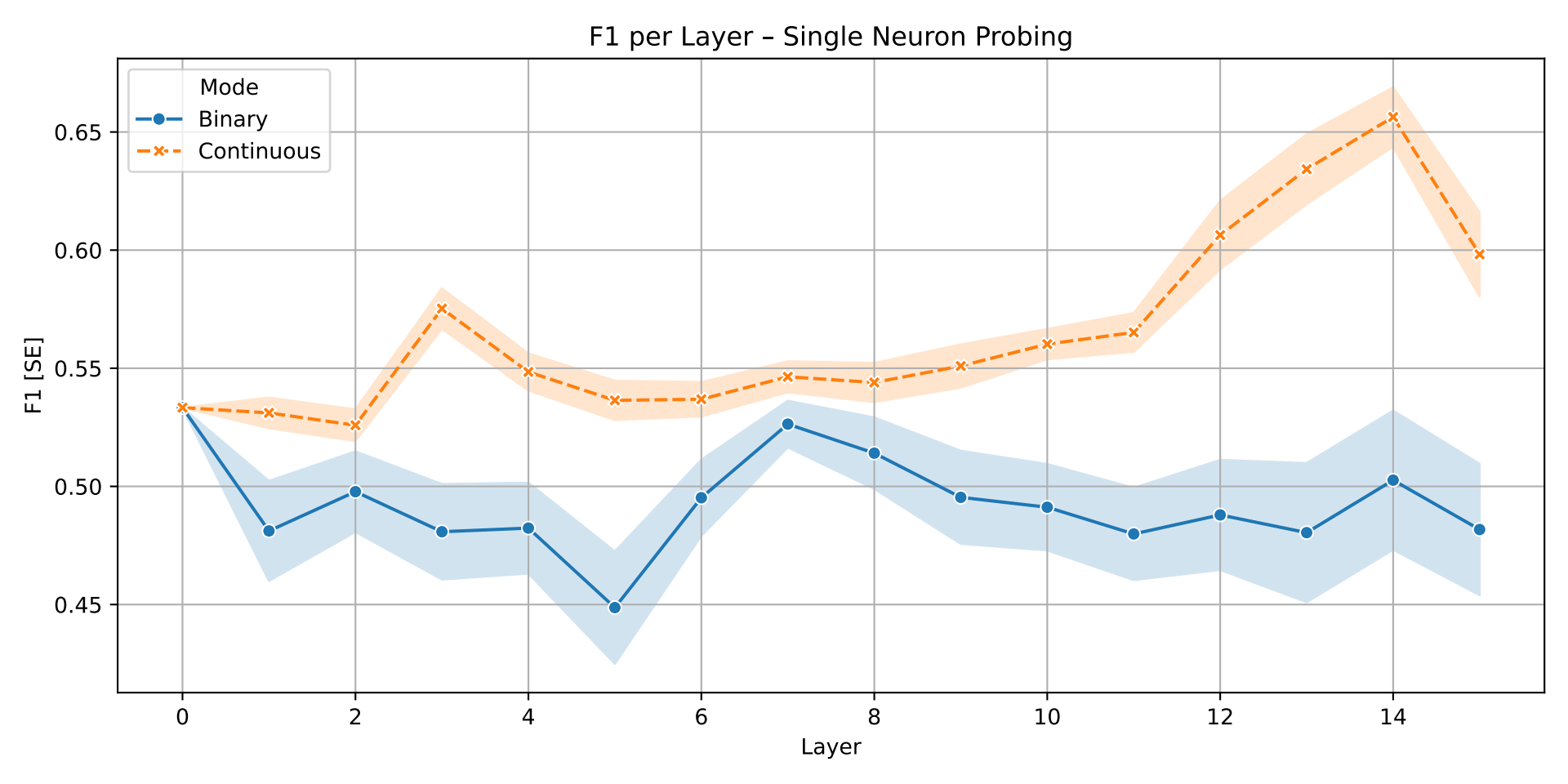
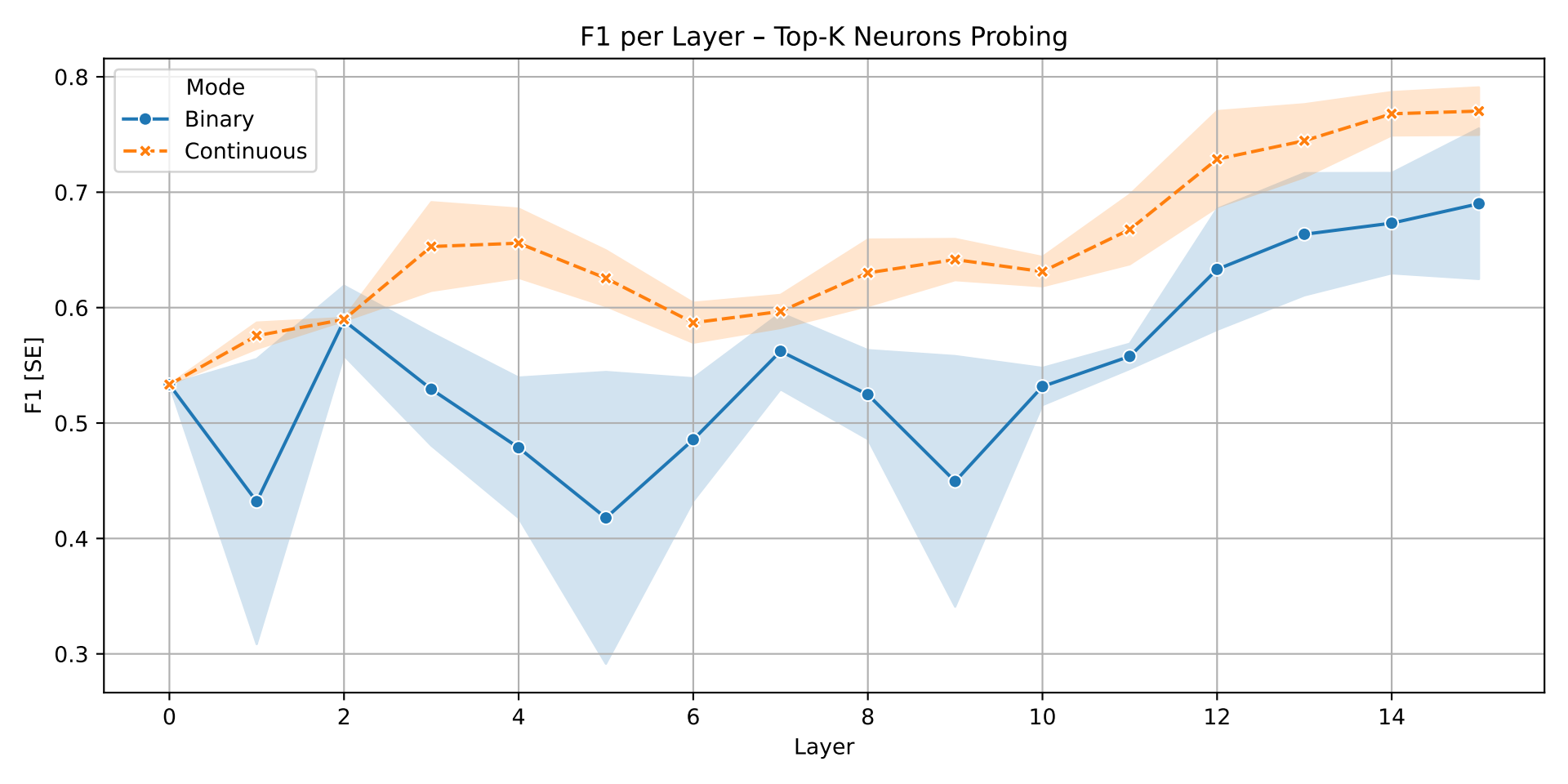
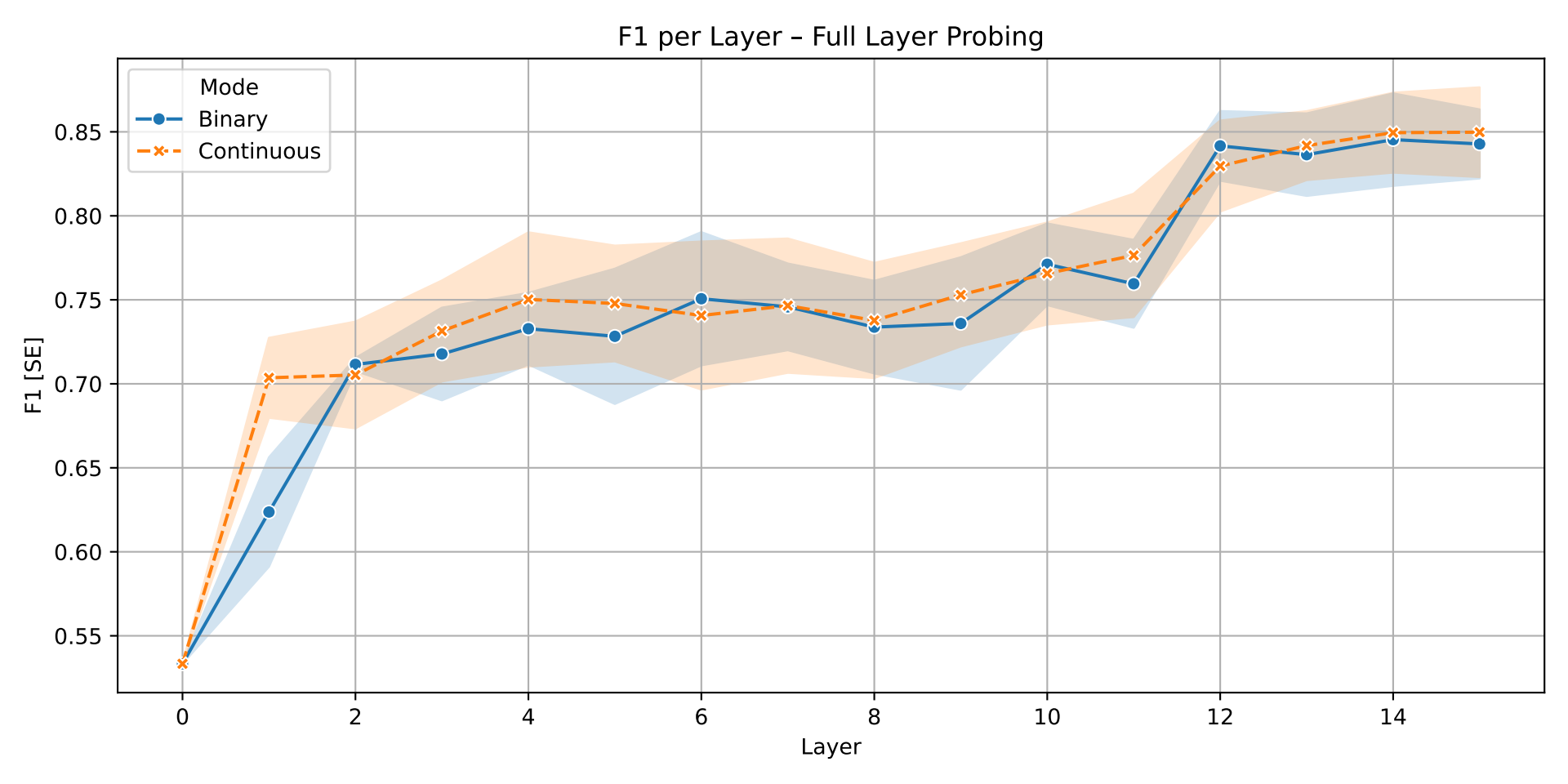
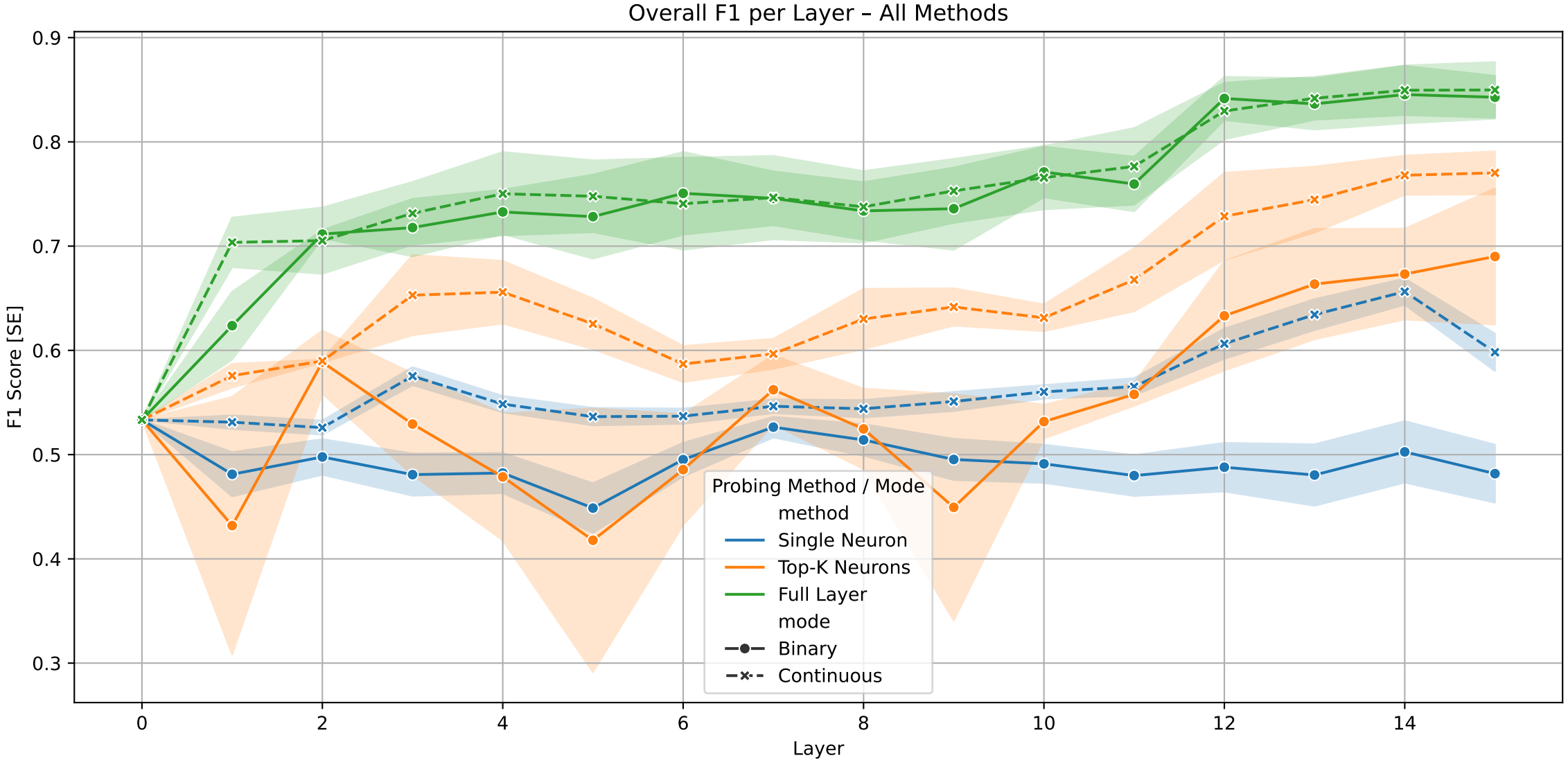

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.65
0.60
0.55
0.50
0.45
0
2
4
6
8
10
12
14
Layer

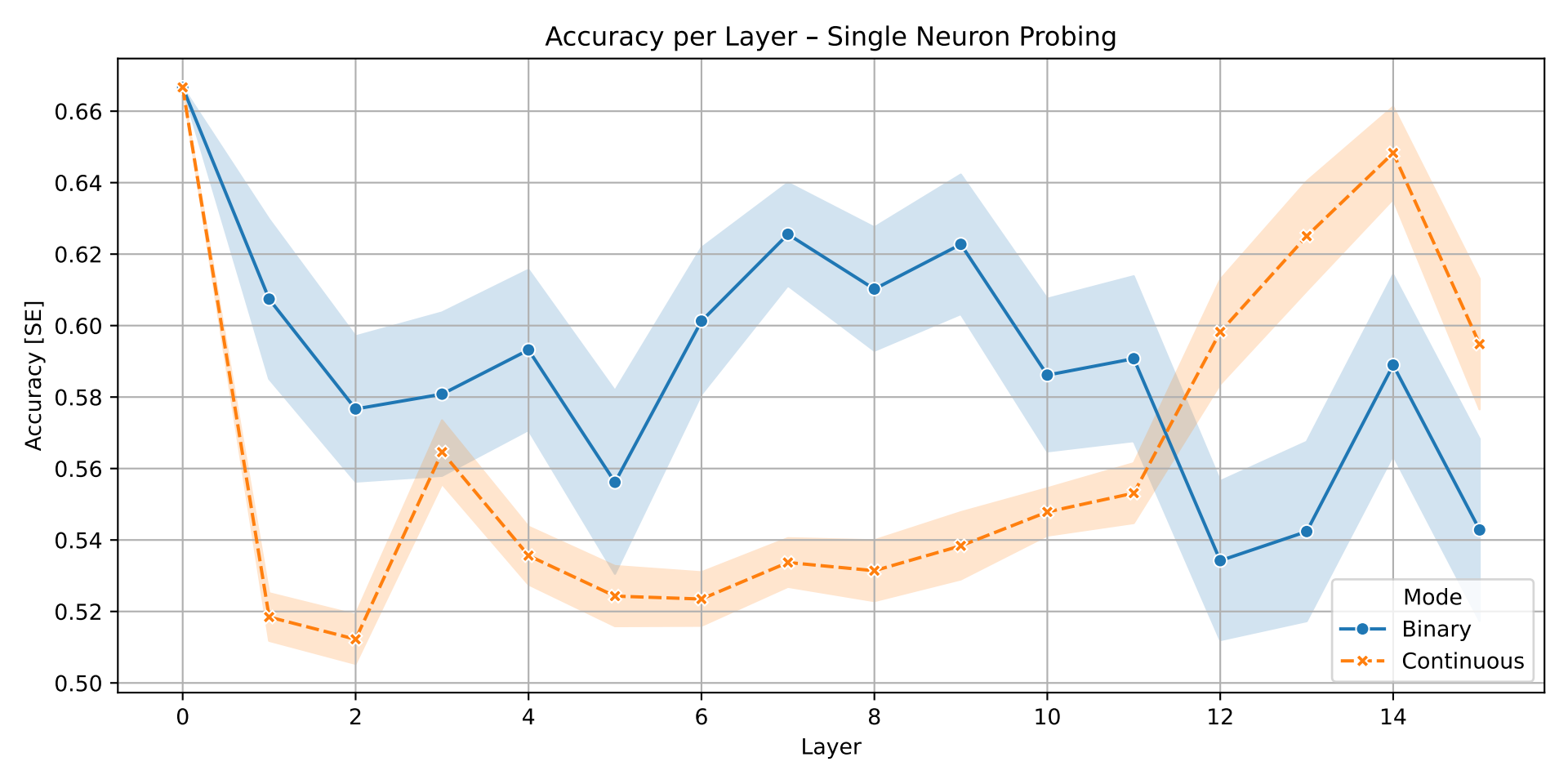

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
Layer
0.8
0.7
0.6
0.5
0.4
0.3
0
2
4
6
8
10
12
14

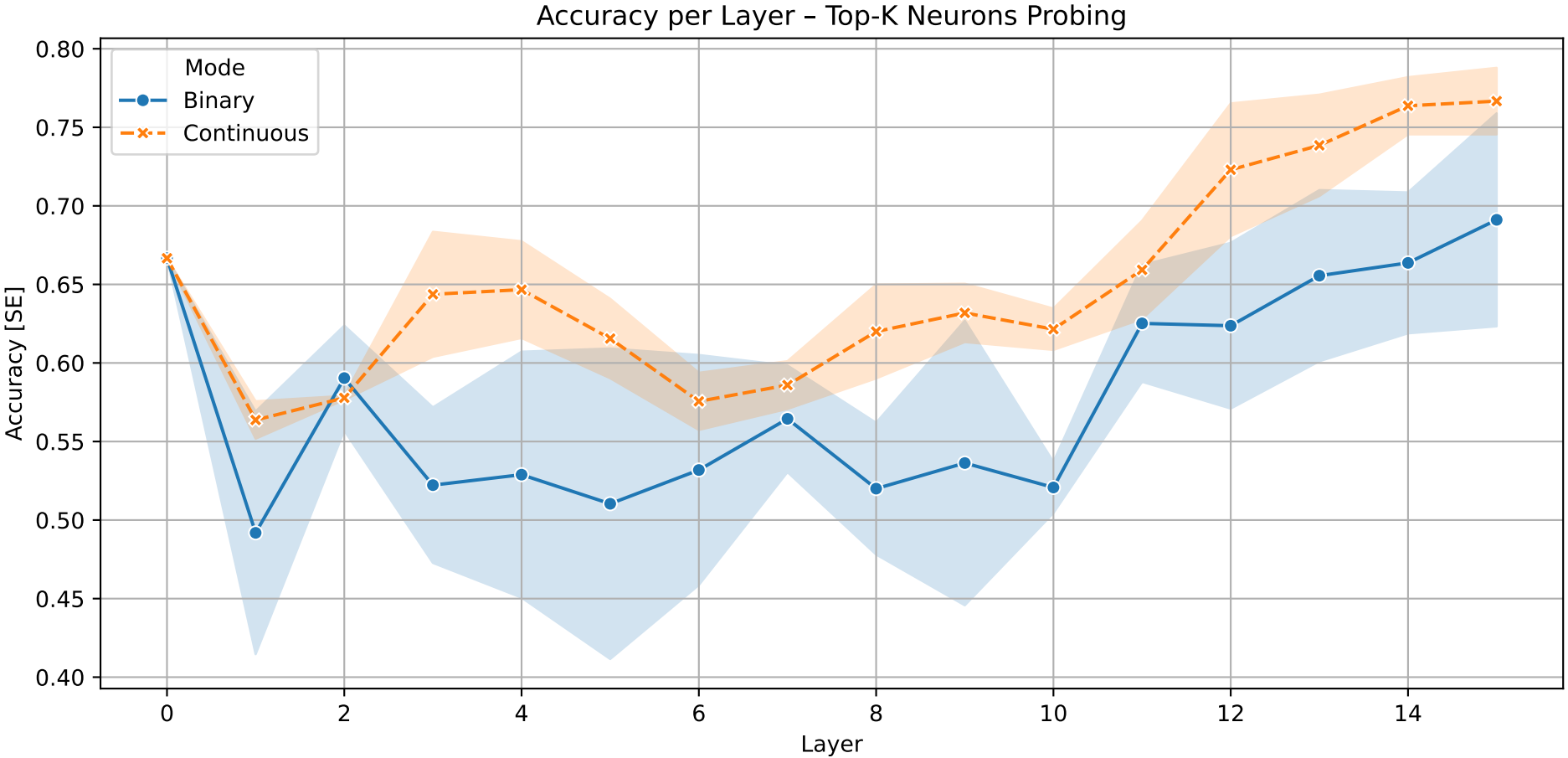

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
12
14
Layer

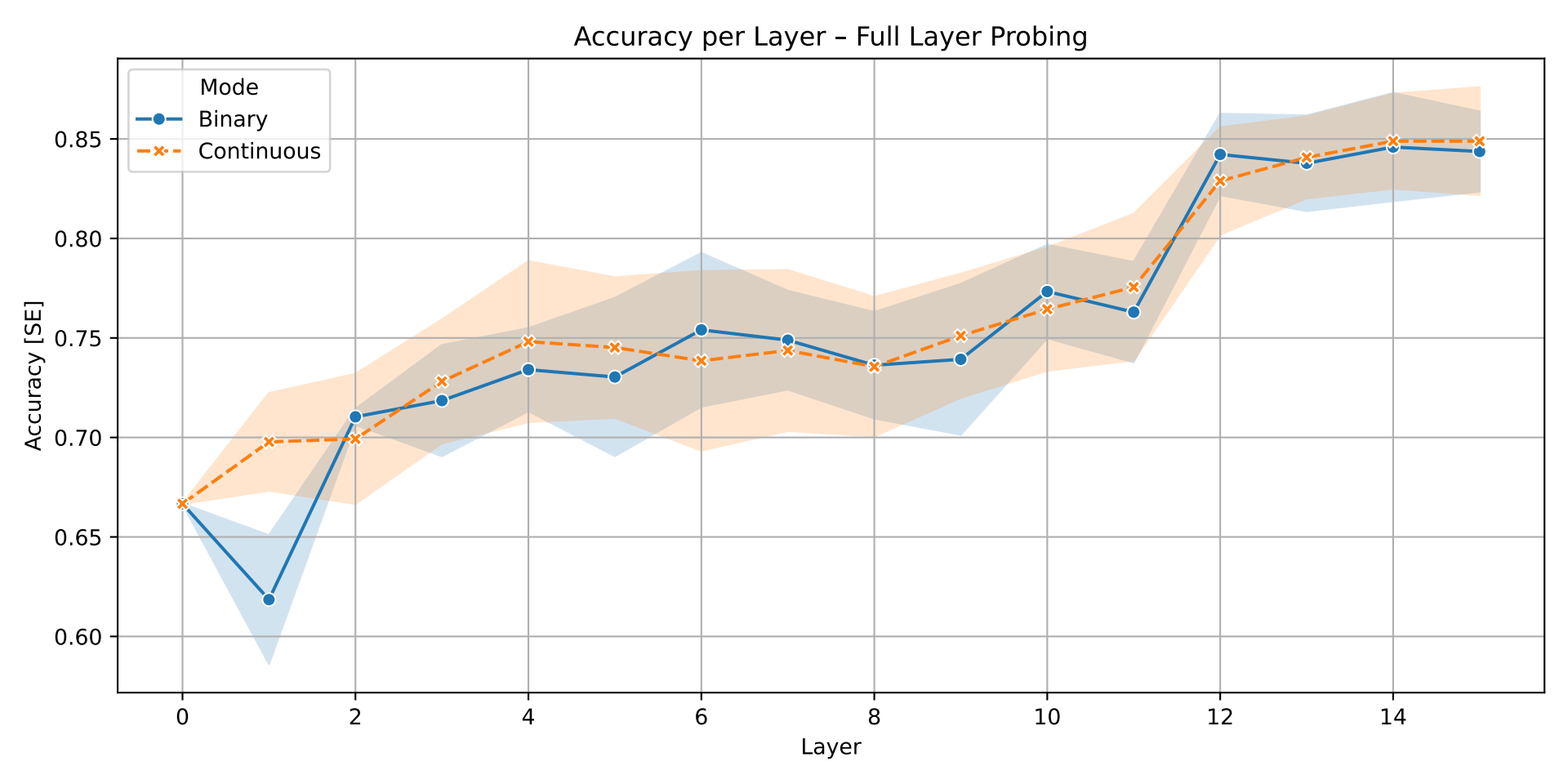

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
2
4
6
8
10
12
14
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

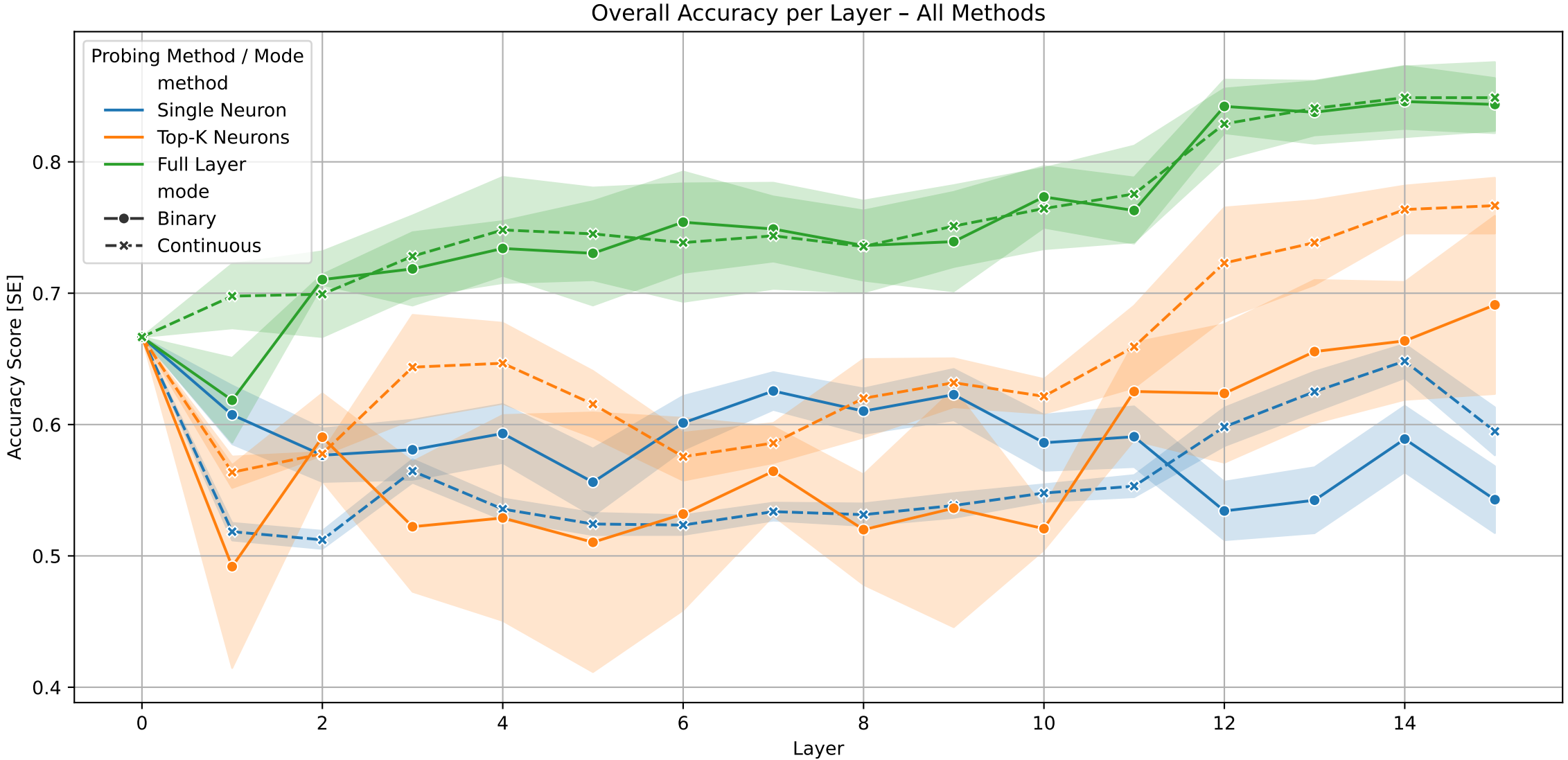
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	14.0	15.0
Full Layer	f1_max	0.8956	0.894
Full Layer	f1_mean	0.7444	0.7539
Full Layer	f1_std	0.089	0.0869
Single Neuron	f1_best_layer	0.0	14.0
Single Neuron	f1_max	0.7776	0.8091
Single Neuron	f1_mean	0.4924	0.5656
Single Neuron	f1_std	0.1129	0.0651
Top-K Neurons	f1_best_layer	15.0	15.0
Top-K Neurons	f1_max	0.8085	0.8111
Top-K Neurons	f1_mean	0.5469	0.6499
Top-K Neurons	f1_std	0.1247	0.0773


Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
0.50
0
2
4
6
8
10
12
14
Layer
Mode
Binary
Continuous


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
Layer
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0
2
4
6
8
10
12
14


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.85
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
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
Accuracy Score [SE]
0.8
0.7
0.6
0.5
0.4
0
2
4
6
8
10
12
14
Layer

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	14.0	14.0
Full Layer	accuracy_max	0.8956	0.8933
Full Layer	accuracy_mean	0.7539	0.7601
Full Layer	accuracy_std	0.0737	0.0708
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.7911	0.8044
Single Neuron	accuracy_mean	0.5891	0.5635
Single Neuron	accuracy_std	0.1199	0.0719
Top-K Neurons	accuracy_best_layer	15.0	15.0
Top-K Neurons	accuracy_max	0.8156	0.8067
Top-K Neurons	accuracy_mean	0.5777	0.65
Top-K Neurons	accuracy_std	0.1067	0.0738