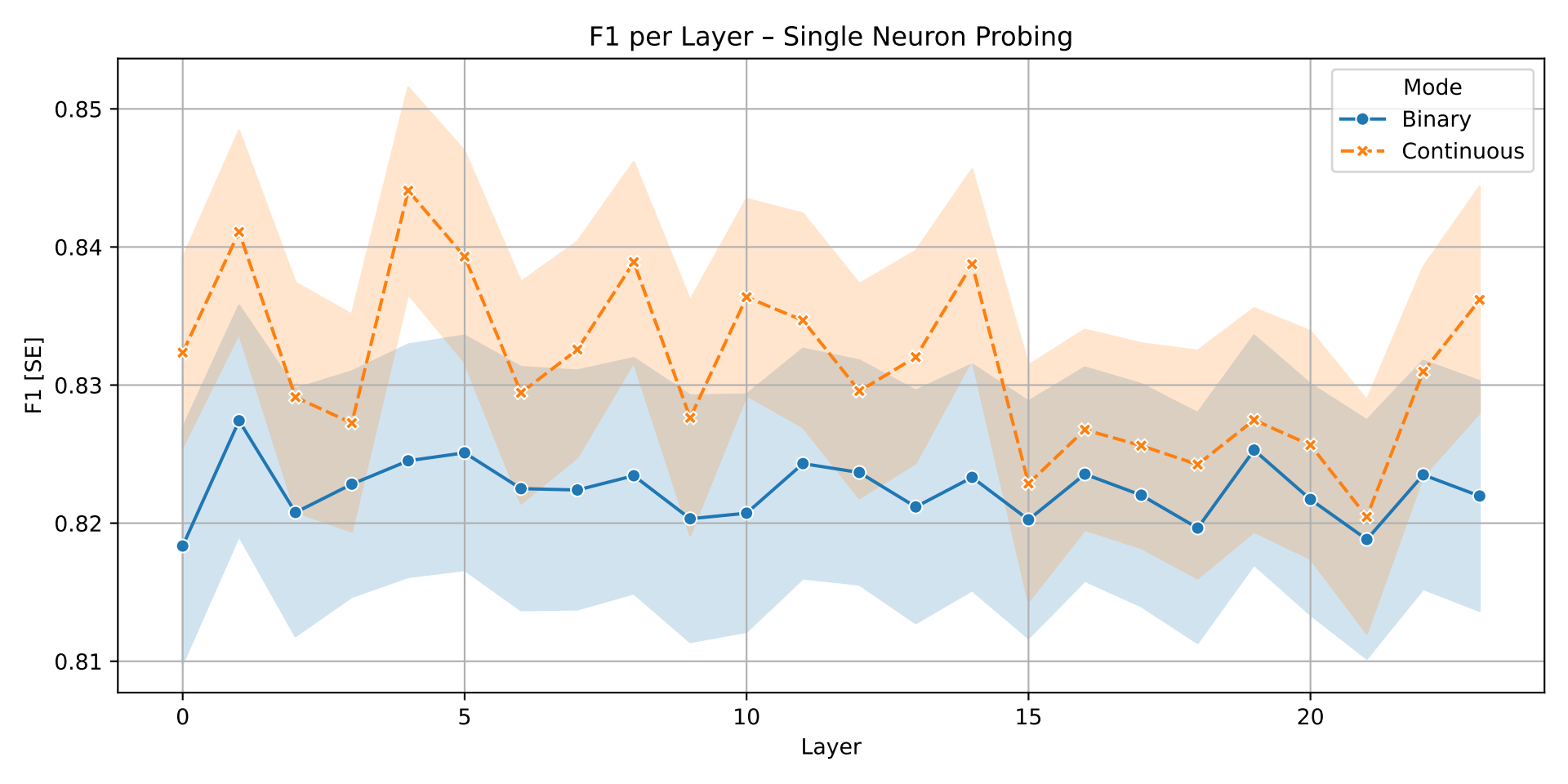
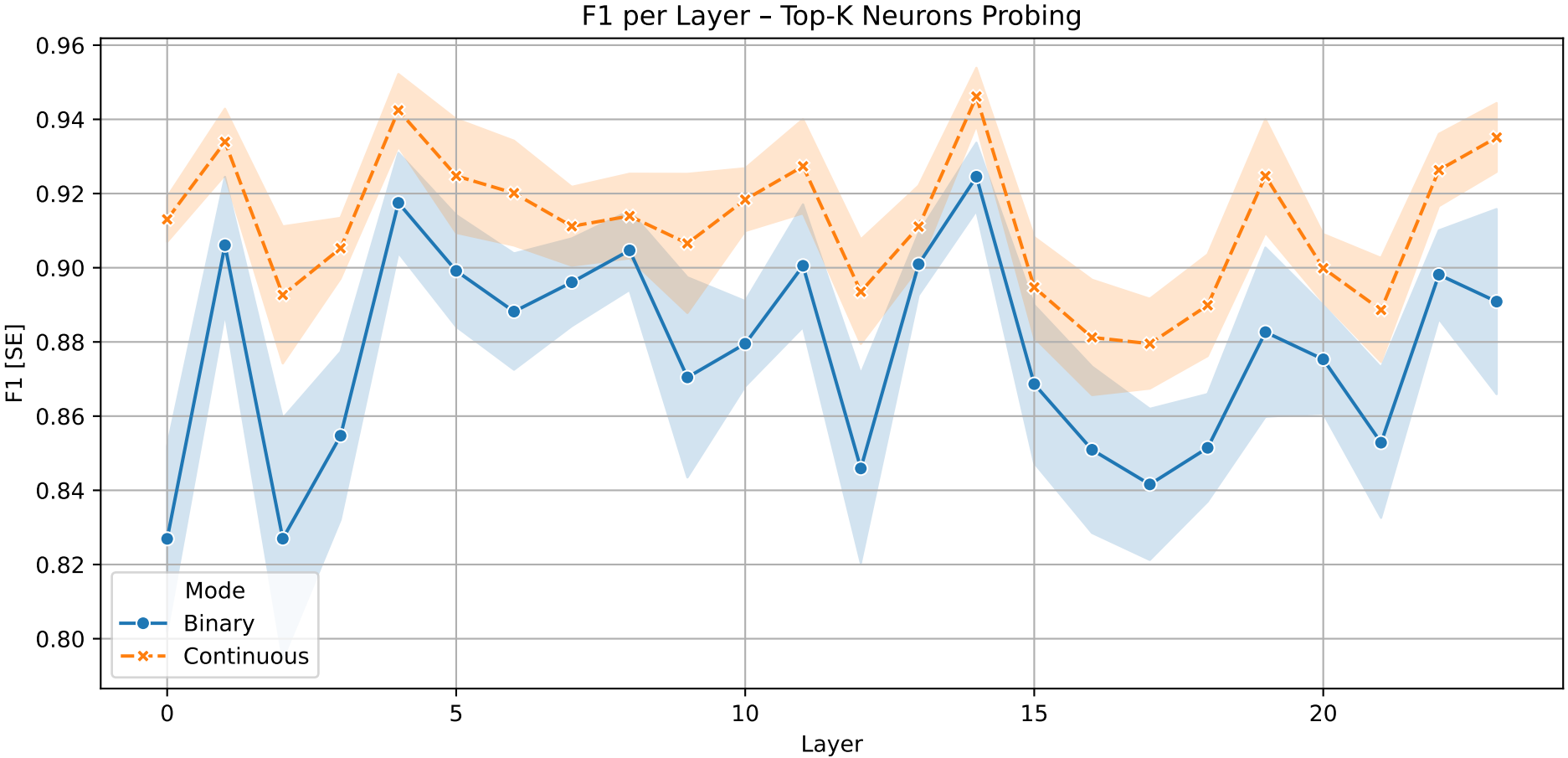
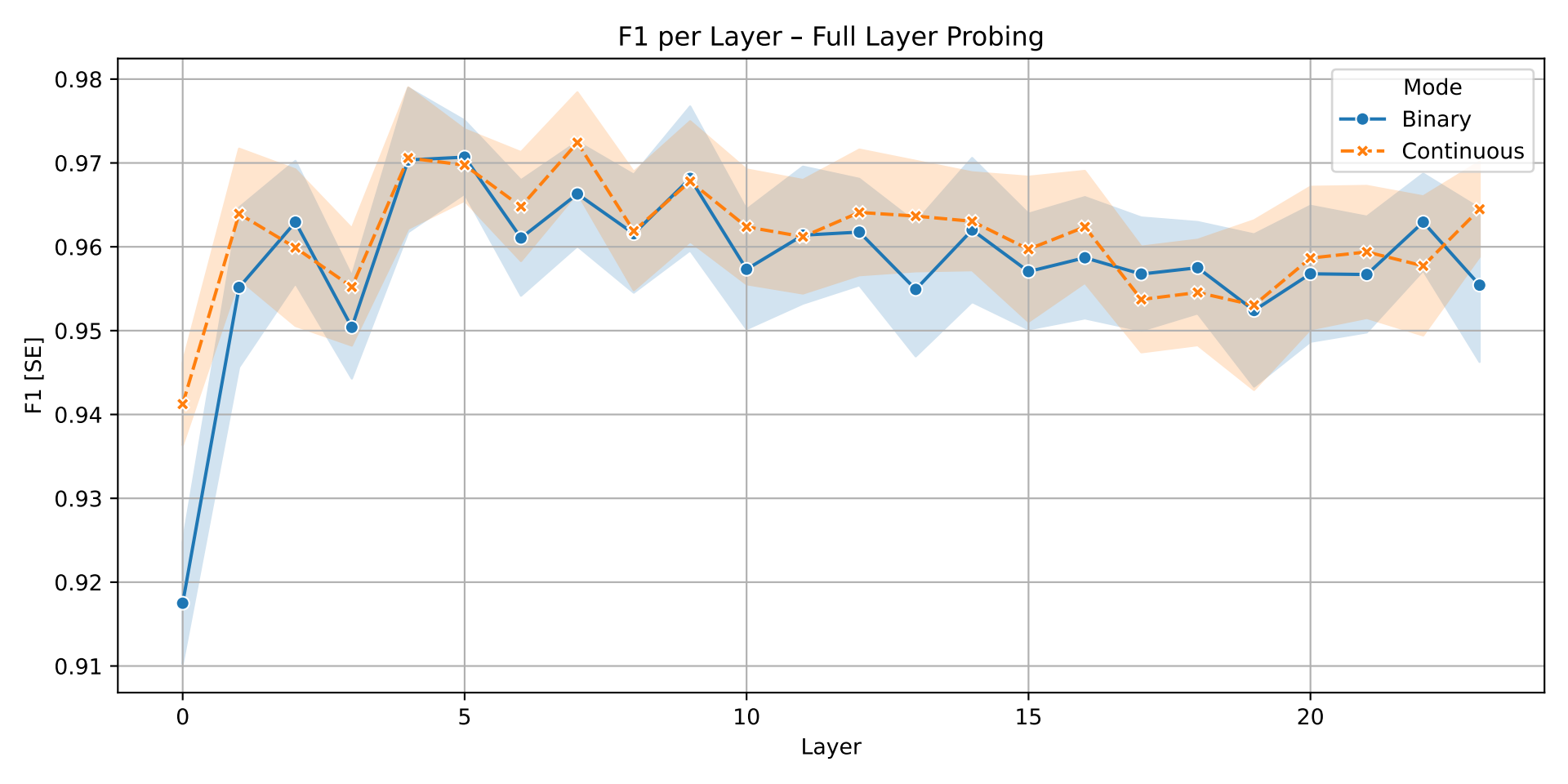
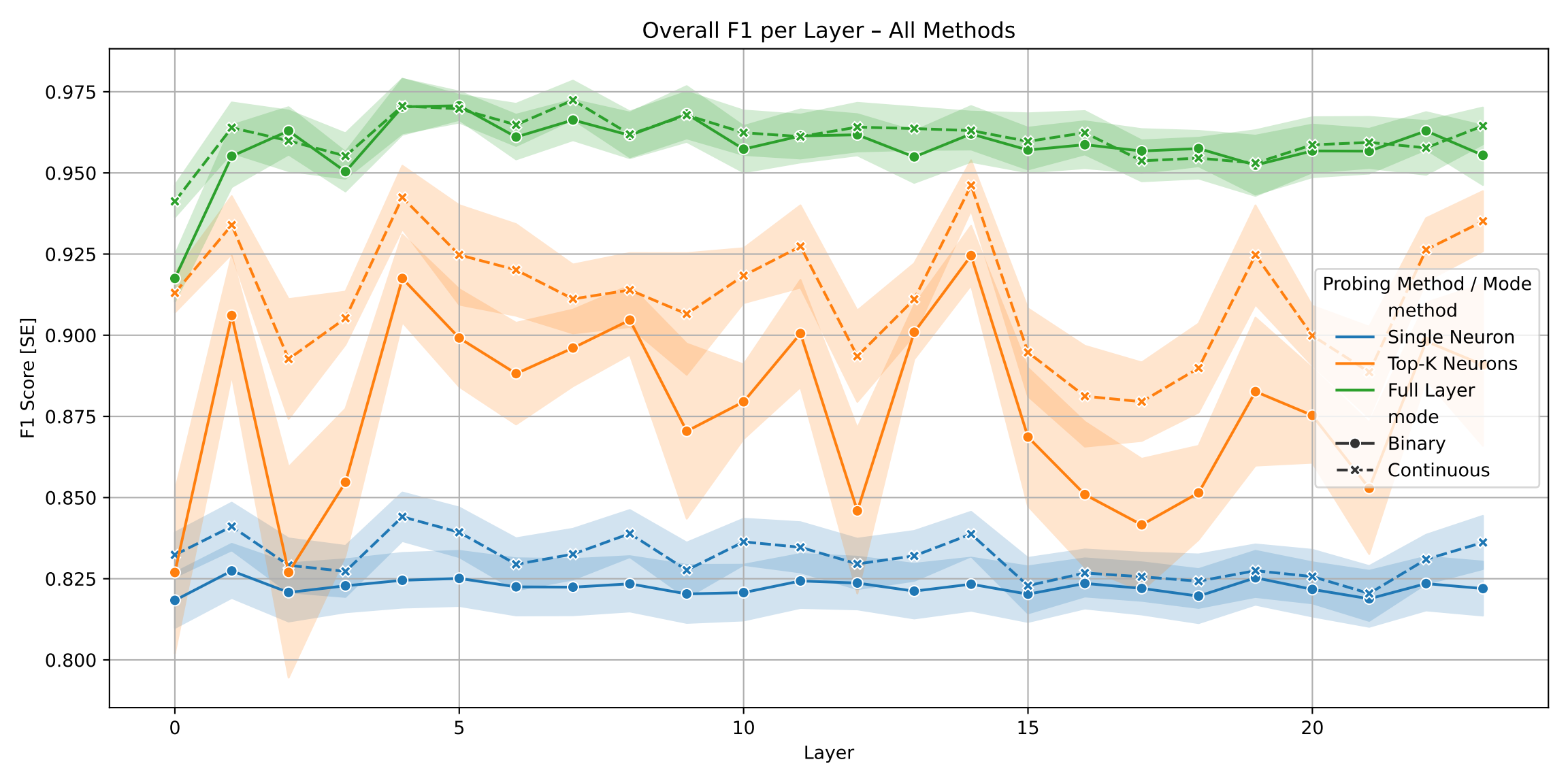

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.85
0.84
0.83
0.82
0.81
0
5
10
15
20
Layer

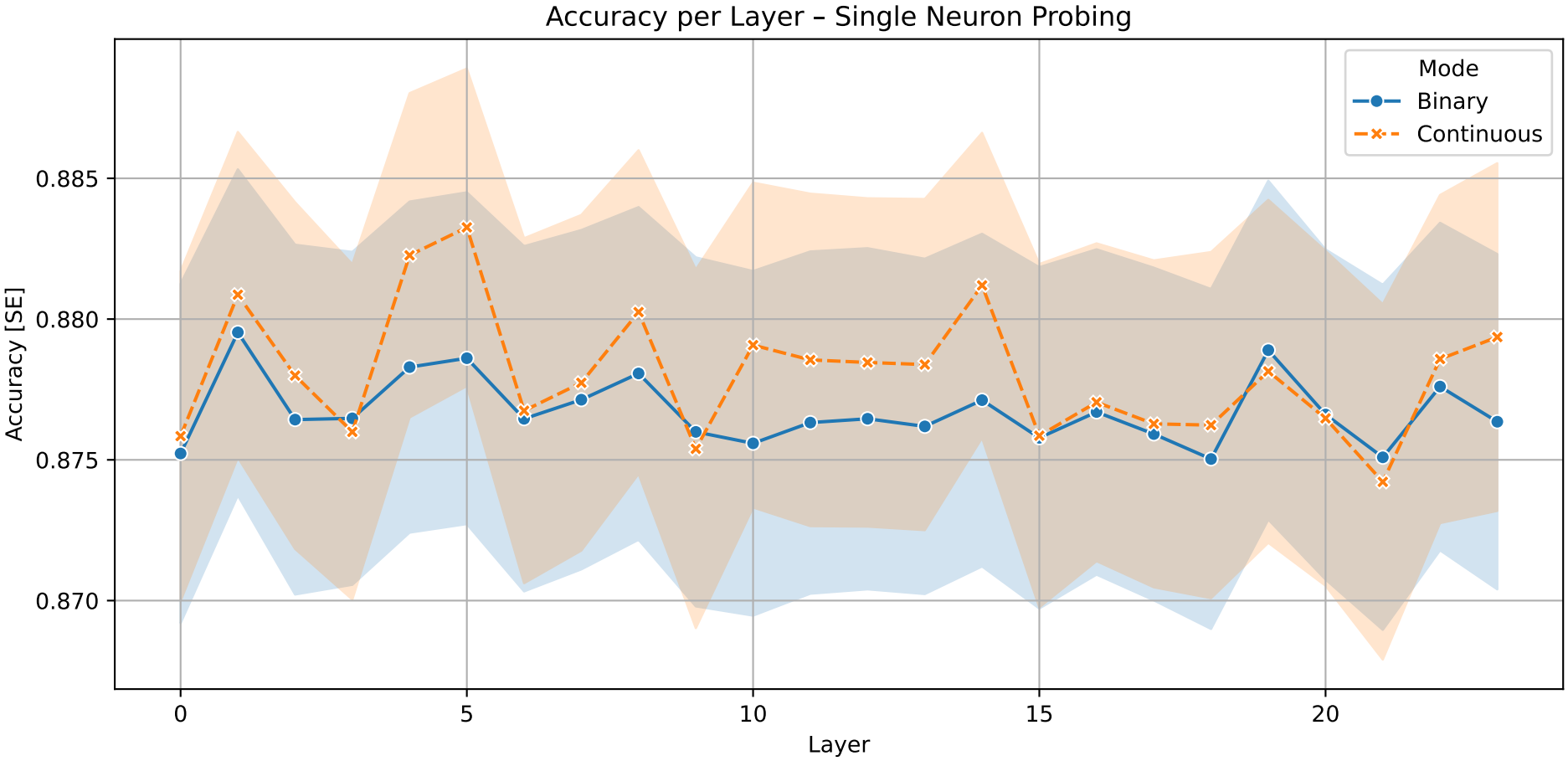

F1 per Layer – Top-K Neurons Probing
F1 [SE]
Layer
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0
5
10
15
20
Mode
Binary
Continuous

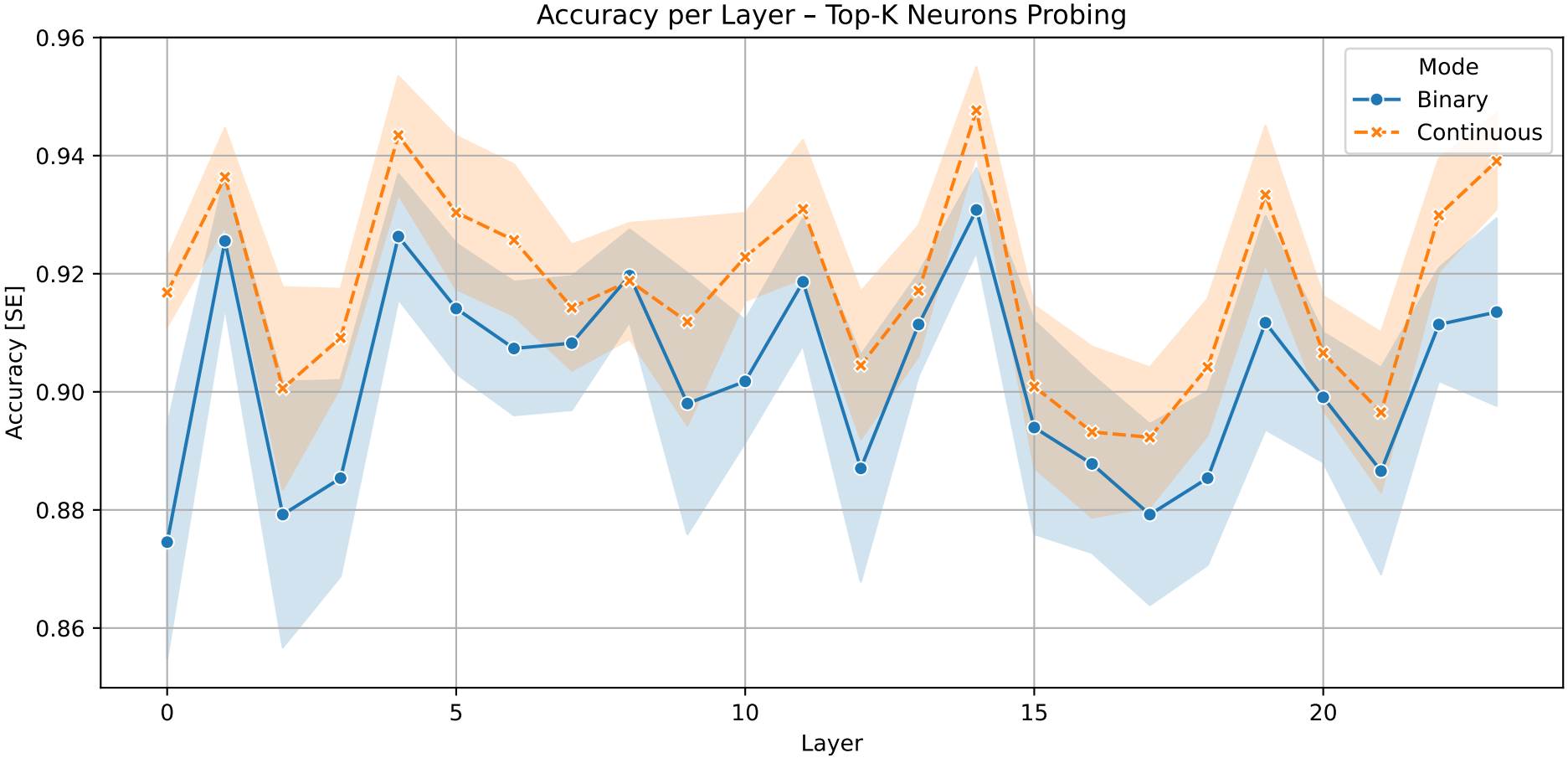

F1 per Layer – Full Layer Probing
F1 [SE]
Layer
Mode
Binary
Continuous
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0
5
10
15
20

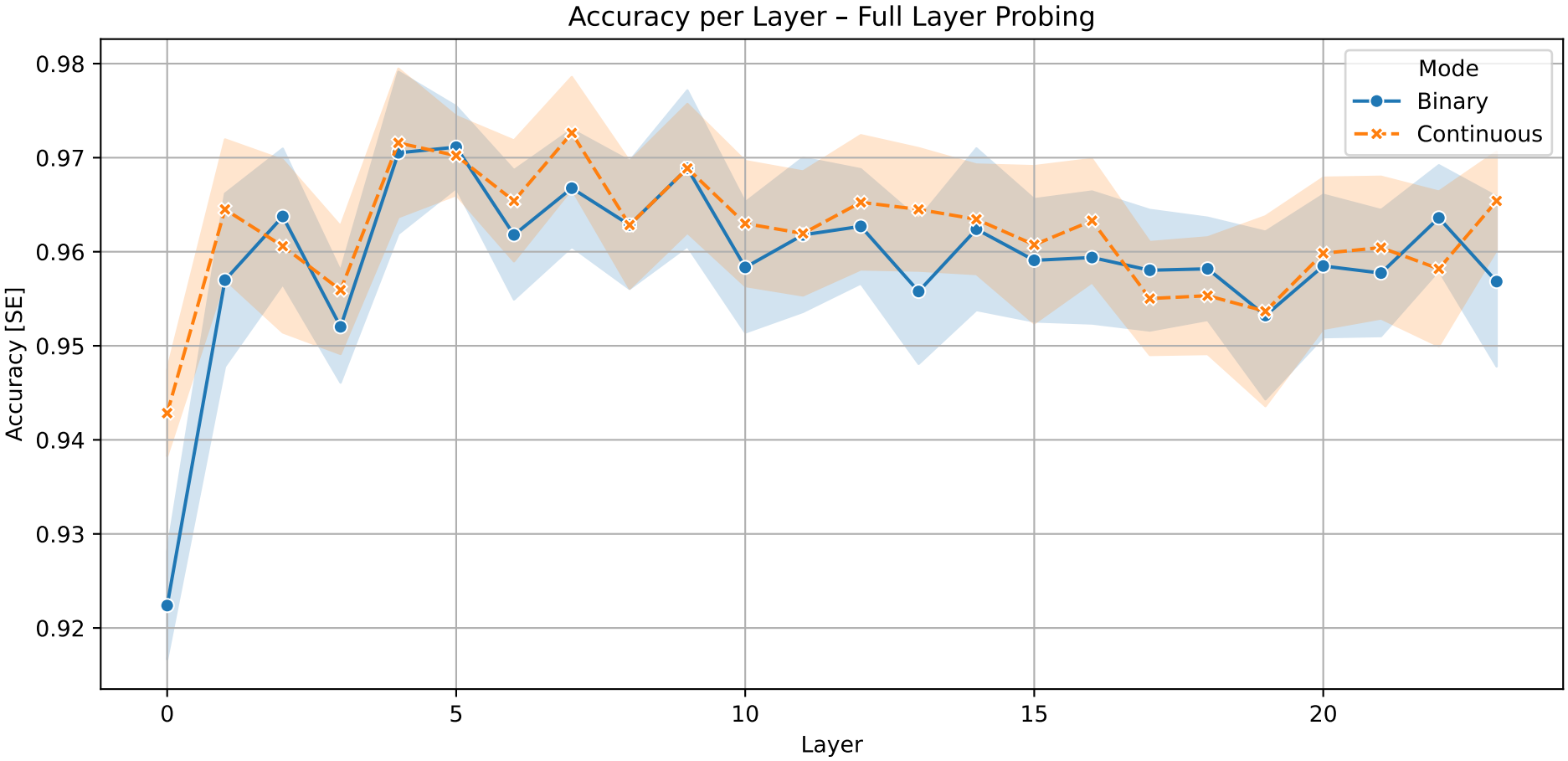

Overall F1 per Layer – All Methods
F1 Score [SE]
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
0
5
10
15
20
Layer
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

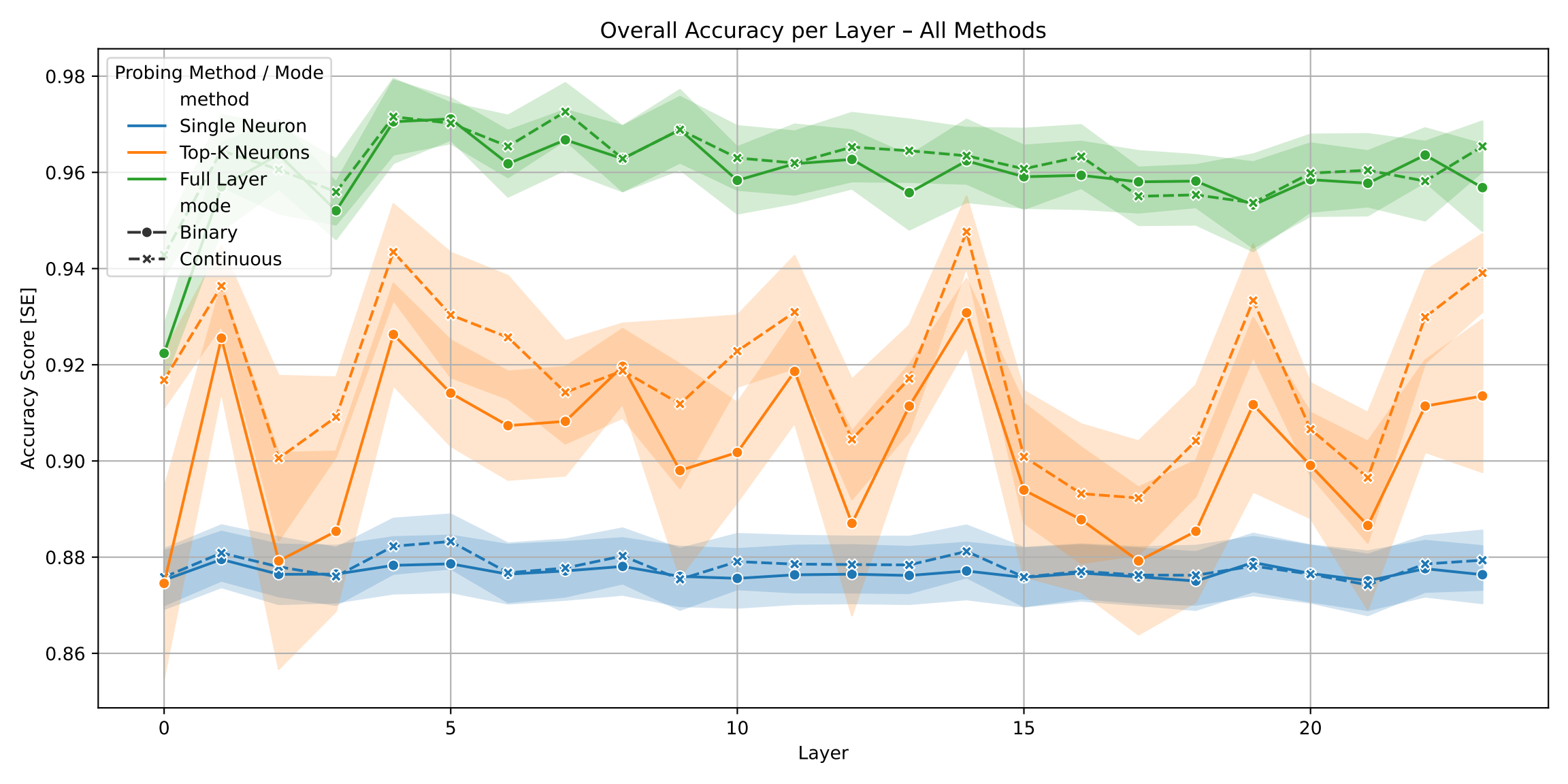
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	5.0	7.0
Full Layer	f1_max	0.9988	0.9988
Full Layer	f1_mean	0.9582	0.9611
Full Layer	f1_std	0.022	0.0202
Single Neuron	f1_best_layer	1.0	4.0
Single Neuron	f1_max	0.9976	0.9988
Single Neuron	f1_mean	0.8224	0.8314
Single Neuron	f1_std	0.0749	0.0694
Top-K Neurons	f1_best_layer	14.0	14.0
Top-K Neurons	f1_max	0.9976	0.9988
Top-K Neurons	f1_mean	0.8773	0.9117
Top-K Neurons	f1_std	0.0579	0.0379


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.885
0.880
0.875
0.870
0
5
10
15
20
Layer


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
Layer
Mode
Binary
Continuous
0.96
0.94
0.92
0.90
0.88
0.86
0
5
10
15
20


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
Layer
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0
5
10
15
20


Overall Accuracy per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
Accuracy Score [SE]
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0
5
10
15
20
Layer

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	5.0	7.0
Full Layer	accuracy_max	0.9988	0.9988
Full Layer	accuracy_mean	0.9593	0.9619
Full Layer	accuracy_std	0.0211	0.0196
Single Neuron	accuracy_best_layer	1.0	5.0
Single Neuron	accuracy_max	0.9976	0.9988
Single Neuron	accuracy_mean	0.8767	0.8781
Single Neuron	accuracy_std	0.0532	0.0528
Top-K Neurons	accuracy_best_layer	14.0	14.0
Top-K Neurons	accuracy_max	0.9976	0.9988
Top-K Neurons	accuracy_mean	0.9024	0.9178
Top-K Neurons	accuracy_std	0.0421	0.0344