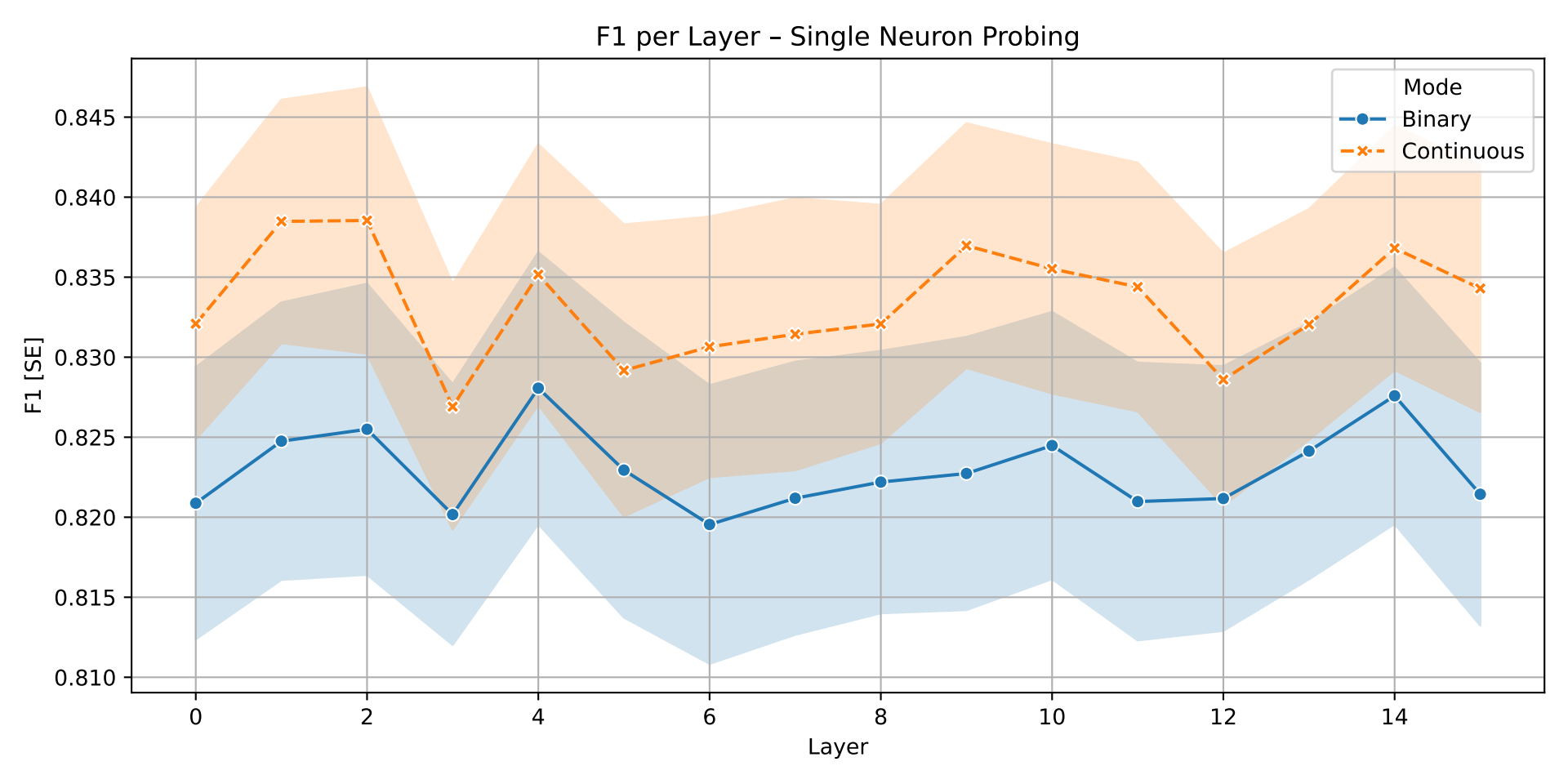
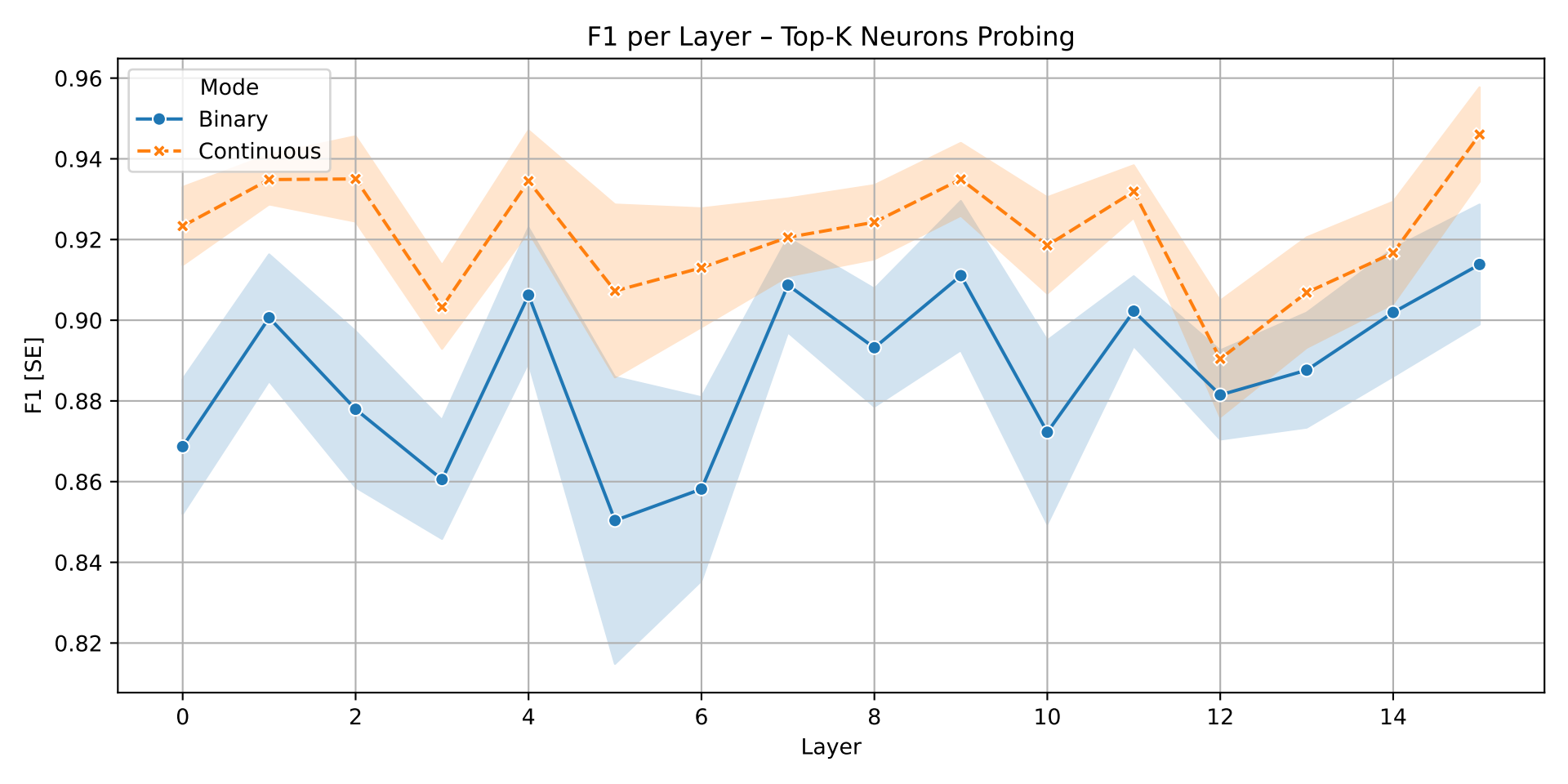
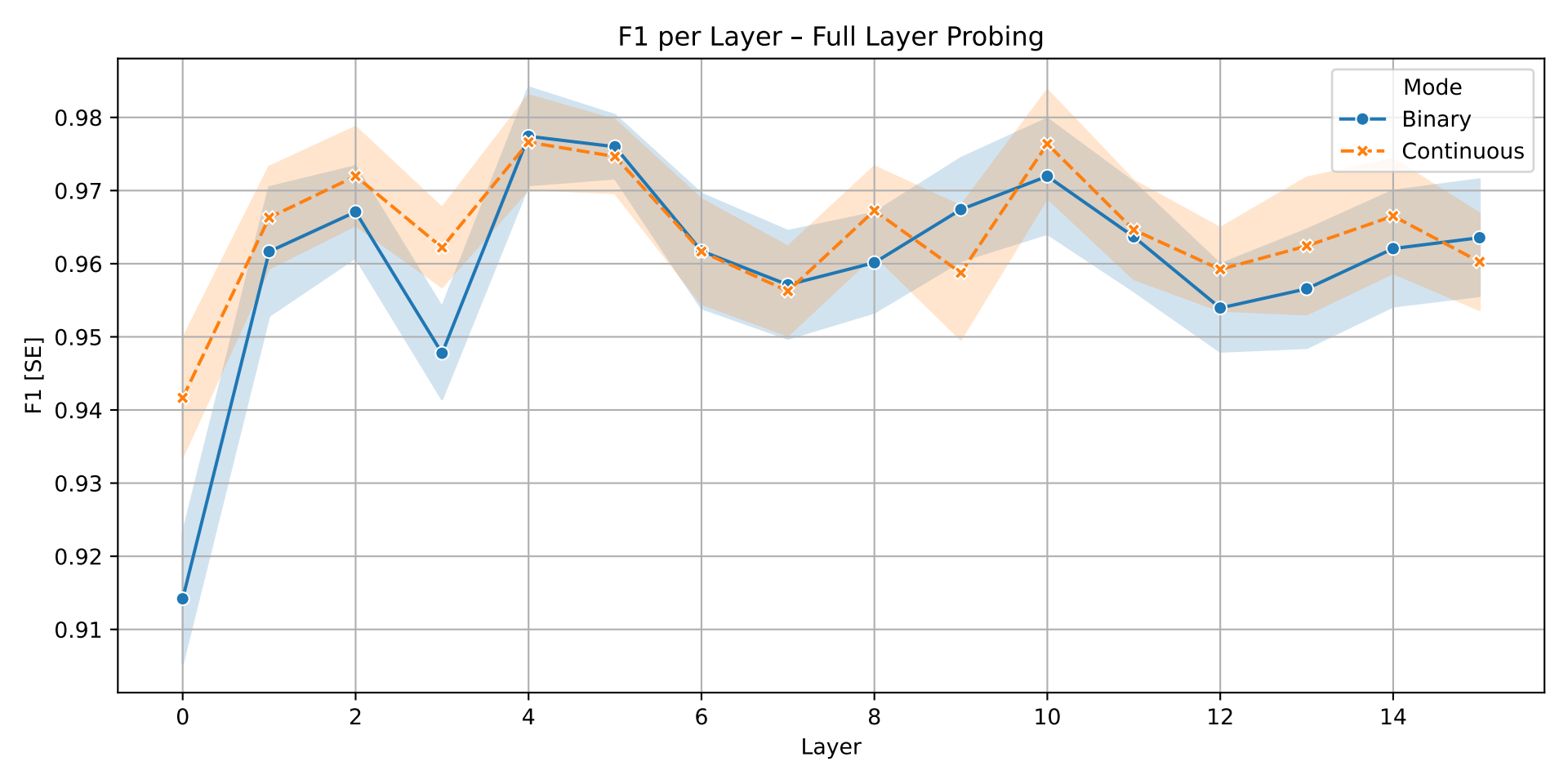
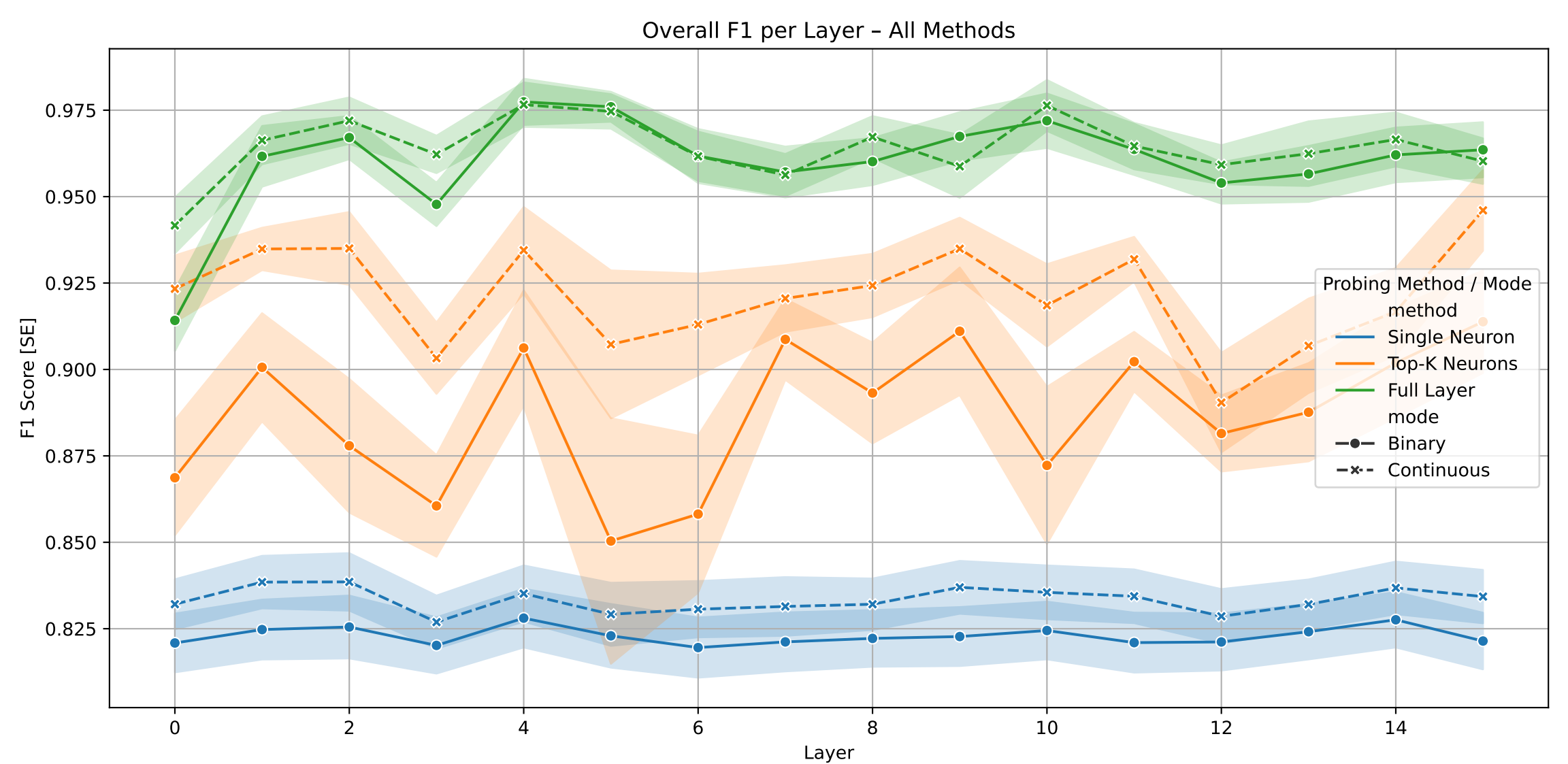

F1 per Layer – Single Neuron Probing
F1 [SE]
Layer
Mode
Binary
Continuous
0.845
0.840
0.835
0.830
0.825
0.820
0.815
0.810
0
2
4
6
8
10
12
14

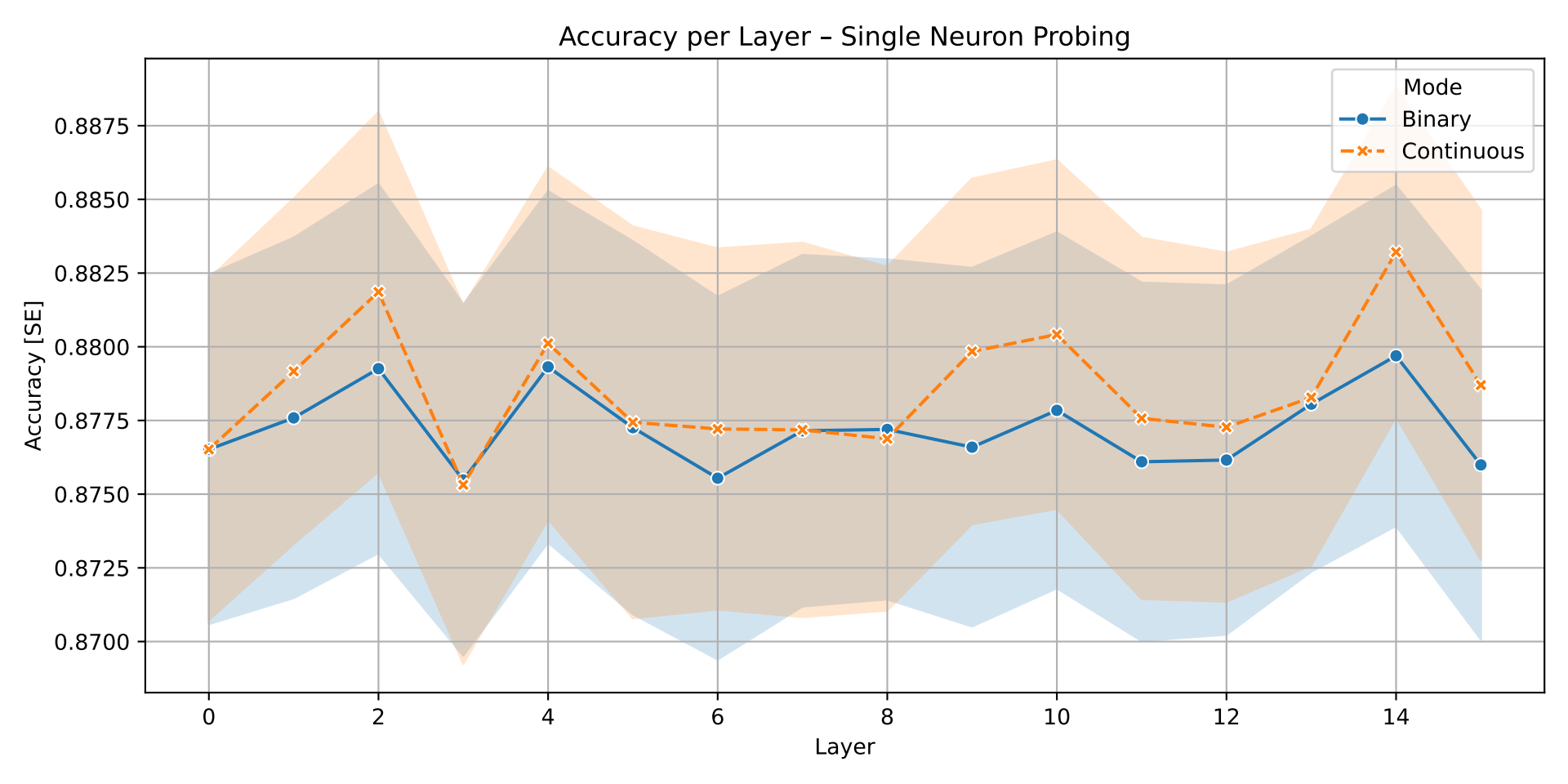

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
Layer
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0
2
4
6
8
10
12
14

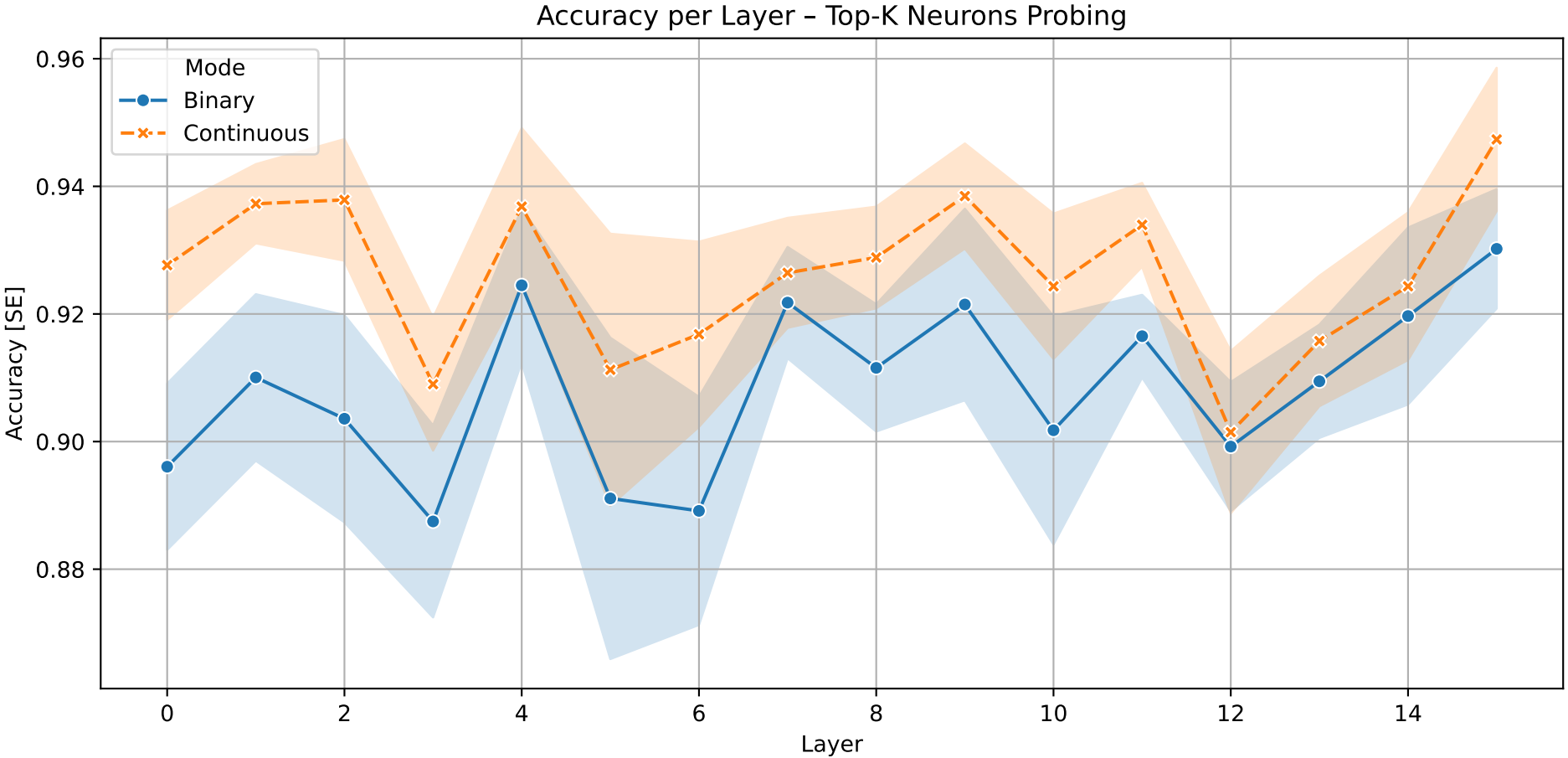

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0
2
4
6
8
10
12
14
Layer

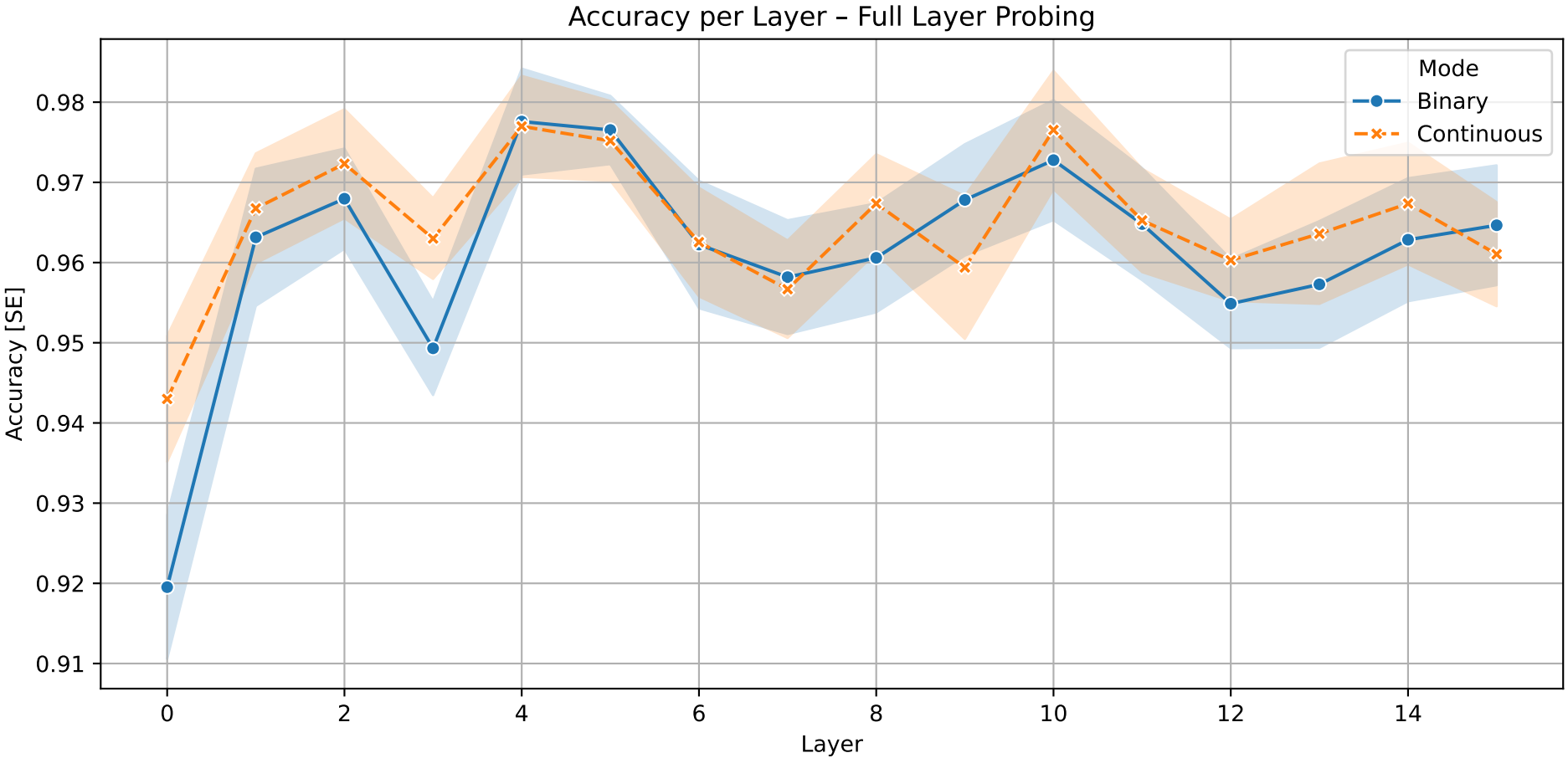

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0
2
4
6
8
10
12
14
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

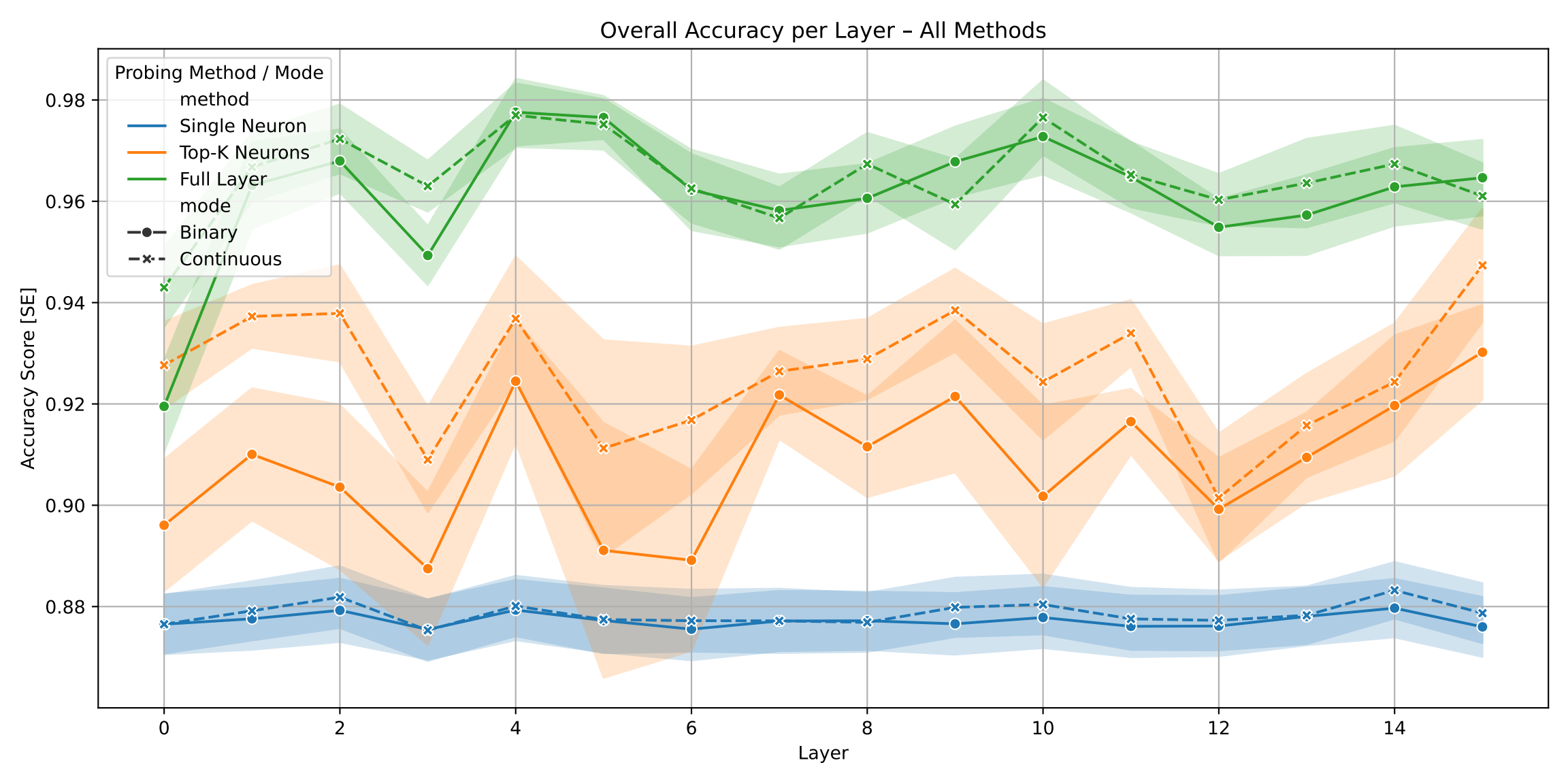
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	4.0	4.0
Full Layer	f1_max	0.9988	0.9988
Full Layer	f1_mean	0.9601	0.9642
Full Layer	f1_std	0.024	0.0204
Single Neuron	f1_best_layer	4.0	2.0
Single Neuron	f1_max	0.9976	0.9976
Single Neuron	f1_mean	0.823	0.8333
Single Neuron	f1_std	0.0754	0.0701
Top-K Neurons	f1_best_layer	15.0	15.0
Top-K Neurons	f1_max	0.9976	0.9988
Top-K Neurons	f1_mean	0.8872	0.9213
Top-K Neurons	f1_std	0.0521	0.035


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.8875
0.8850
0.8825
0.8800
0.8775
0.8750
0.8725
0.8700
0
2
4
6
8
10
12
14
Layer


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.96
0.94
0.92
0.90
0.88
0
2
4
6
8
10
12
14
Layer


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
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]
0.98
0.96
0.94
0.92
0.90
0.88
0
2
4
6
8
10
12
14
Layer

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	4.0	4.0
Full Layer	accuracy_max	0.9988	0.9988
Full Layer	accuracy_mean	0.9613	0.9648
Full Layer	accuracy_std	0.023	0.0199
Single Neuron	accuracy_best_layer	14.0	14.0
Single Neuron	accuracy_max	0.9976	0.9976
Single Neuron	accuracy_mean	0.8772	0.8786
Single Neuron	accuracy_std	0.0533	0.0533
Top-K Neurons	accuracy_best_layer	15.0	15.0
Top-K Neurons	accuracy_max	0.9976	0.9988
Top-K Neurons	accuracy_mean	0.9083	0.9261
Top-K Neurons	accuracy_std	0.0395	0.0323