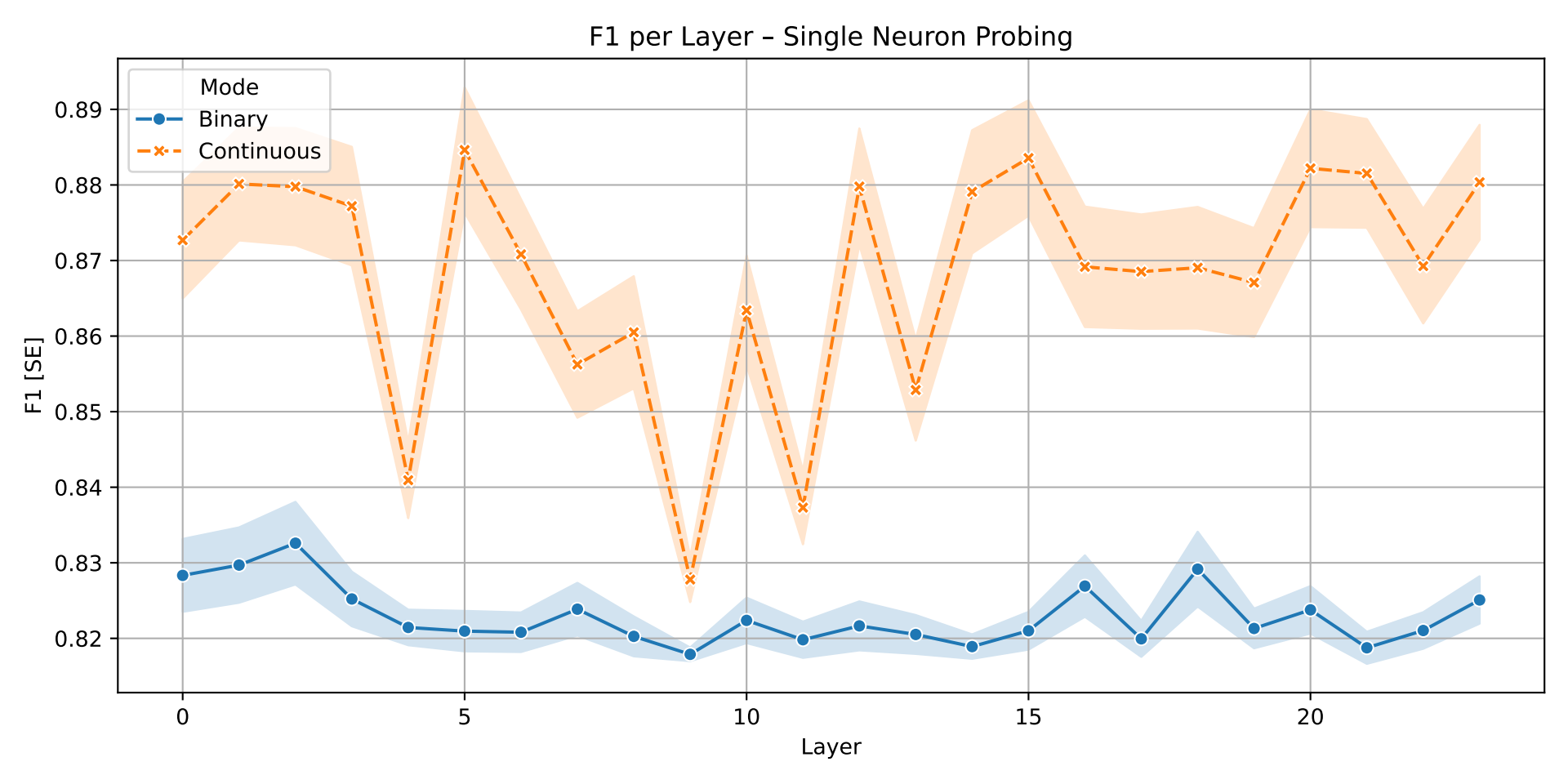
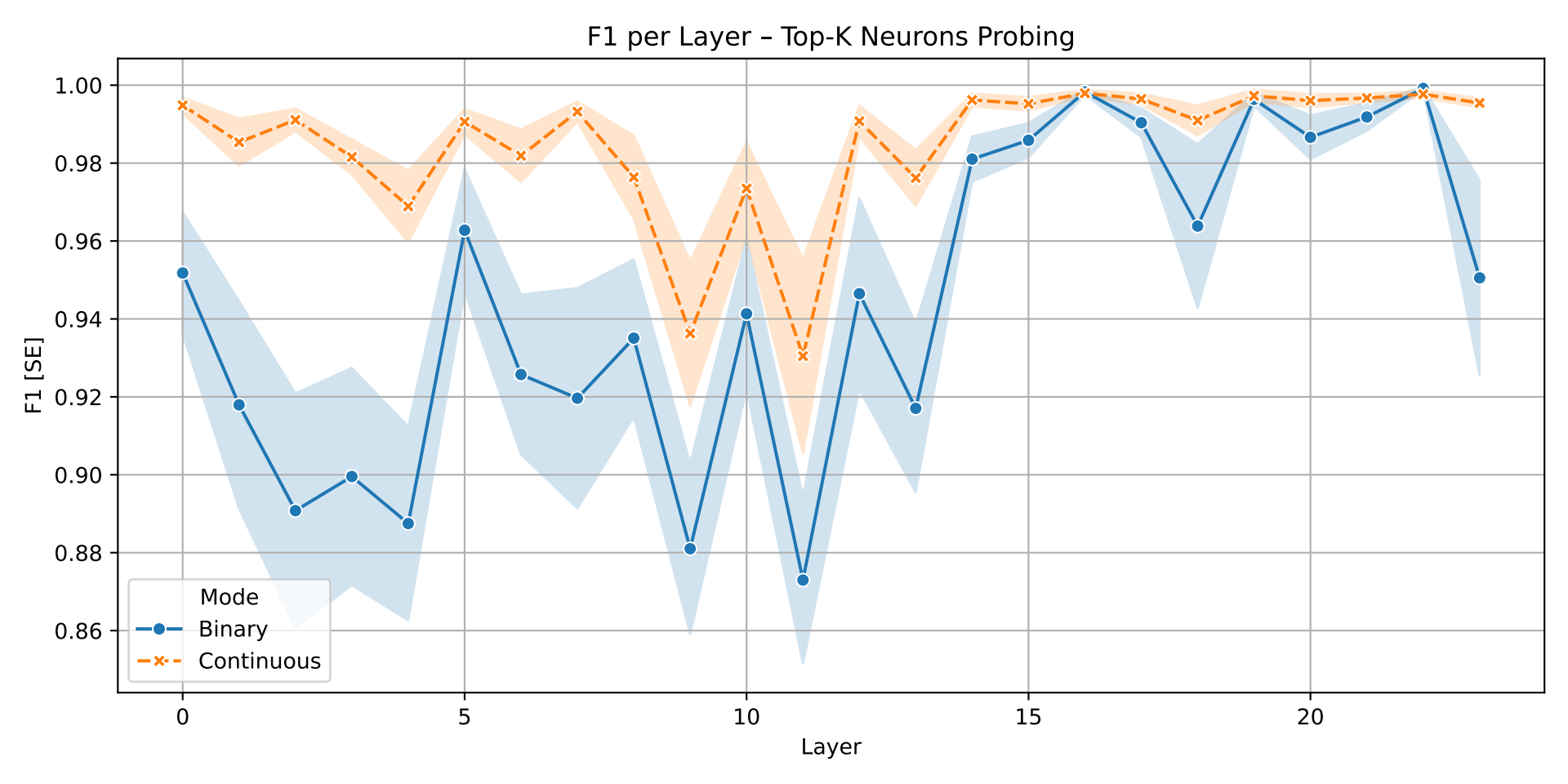
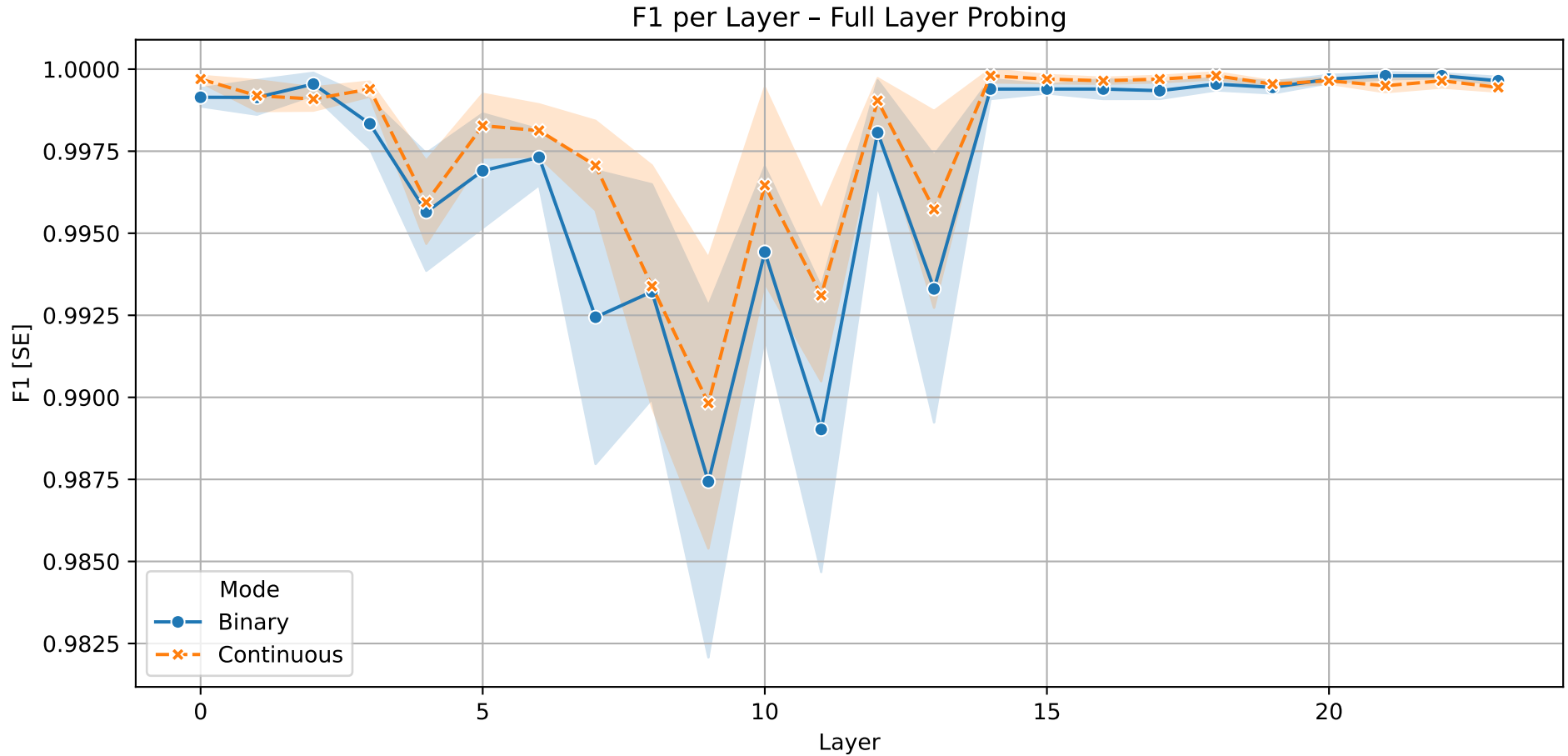
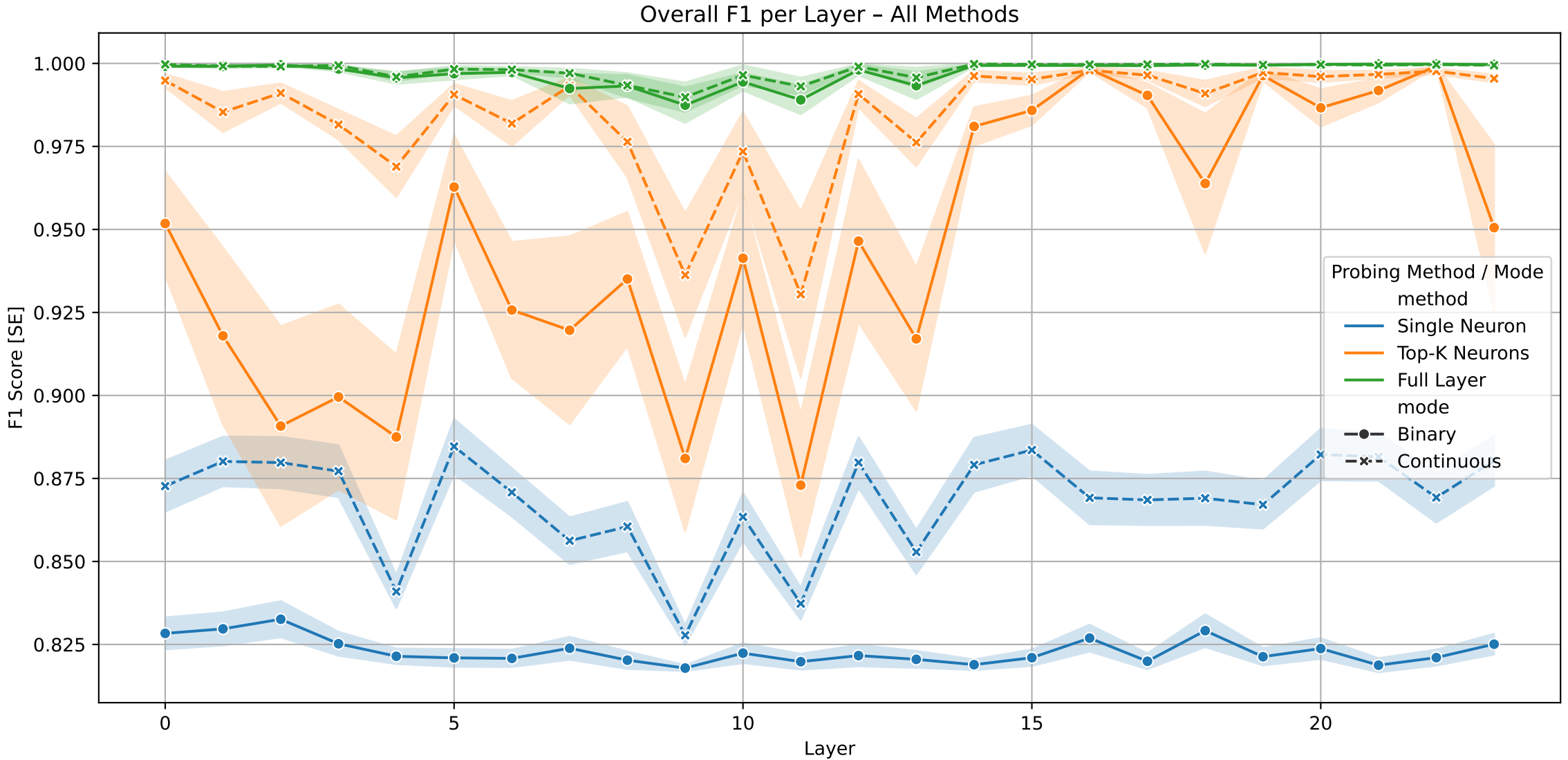

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.89
0.88
0.87
0.86
0.85
0.84
0.83
0.82
0
5
10
15
20
Layer

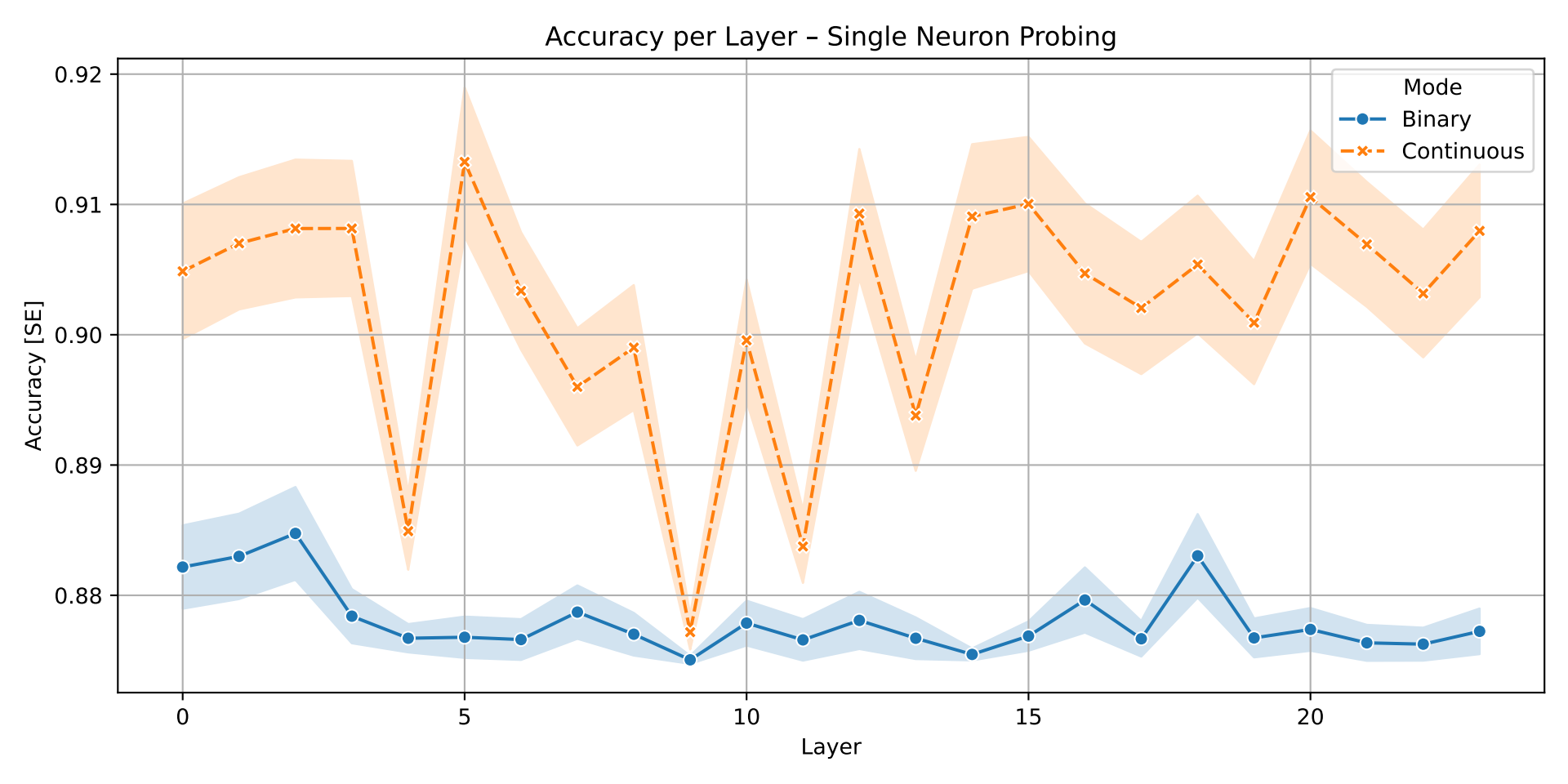

F1 per Layer – Top-K Neurons Probing
F1 [SE]
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0
5
10
15
20
Layer
Mode
Binary
Continuous

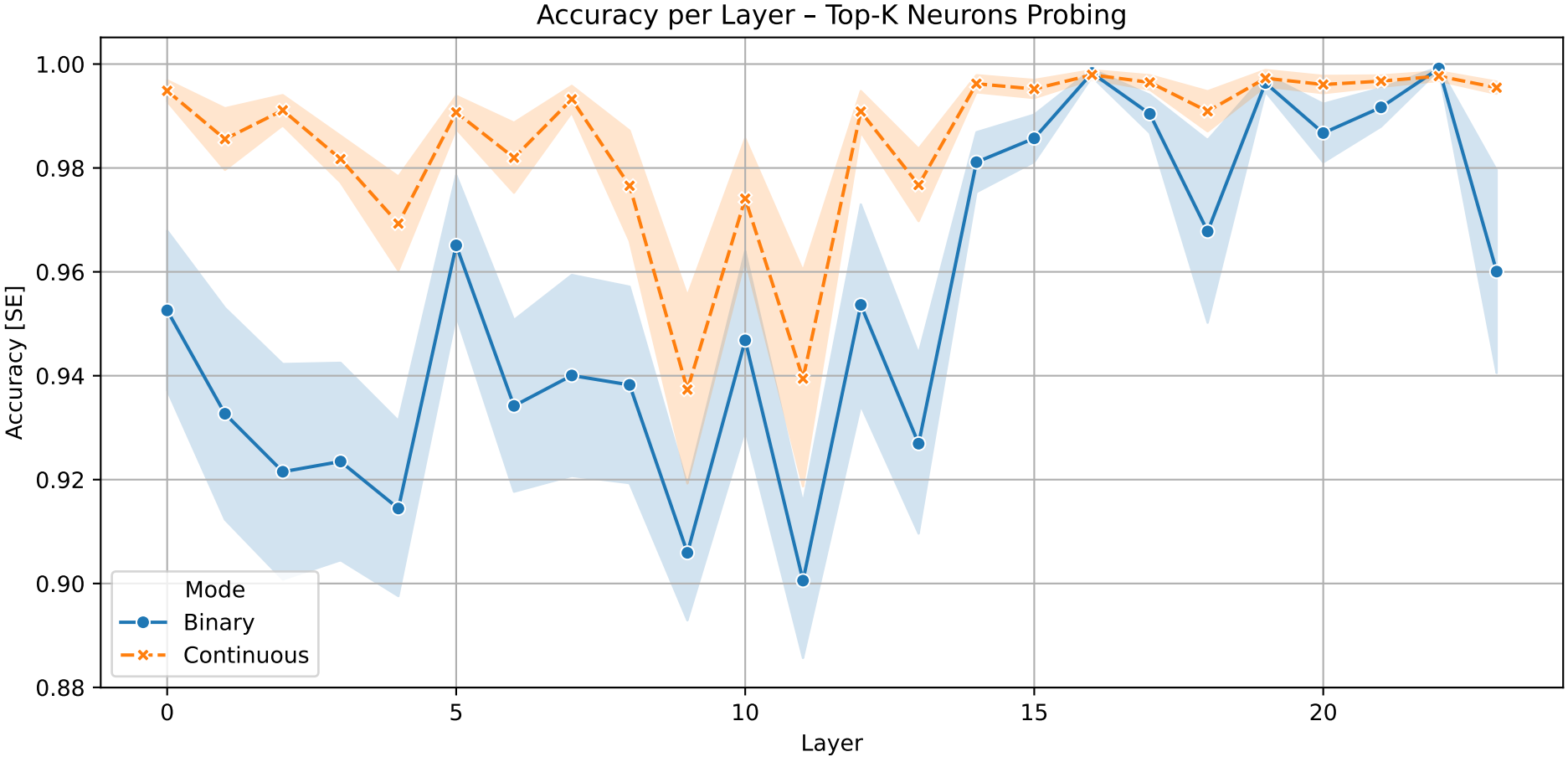

F1 per Layer – Full Layer Probing
F1 [SE]
1.0000
0.9975
0.9950
0.9925
0.9900
0.9875
0.9850
0.9825
0
5
10
15
20
Layer
Mode
Binary
Continuous

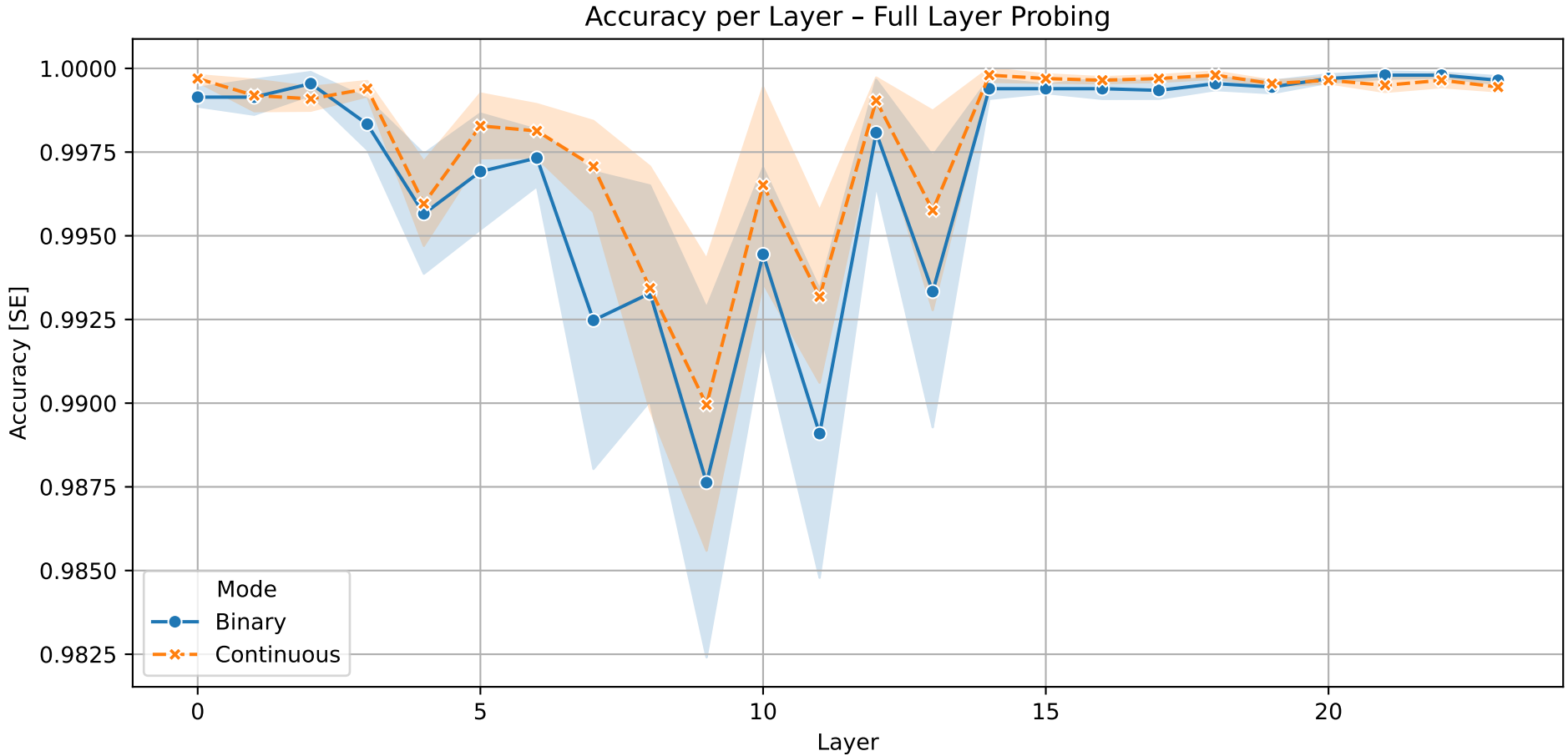

Overall F1 per Layer – All Methods
F1 Score [SE]
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0
5
10
15
20
Layer
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

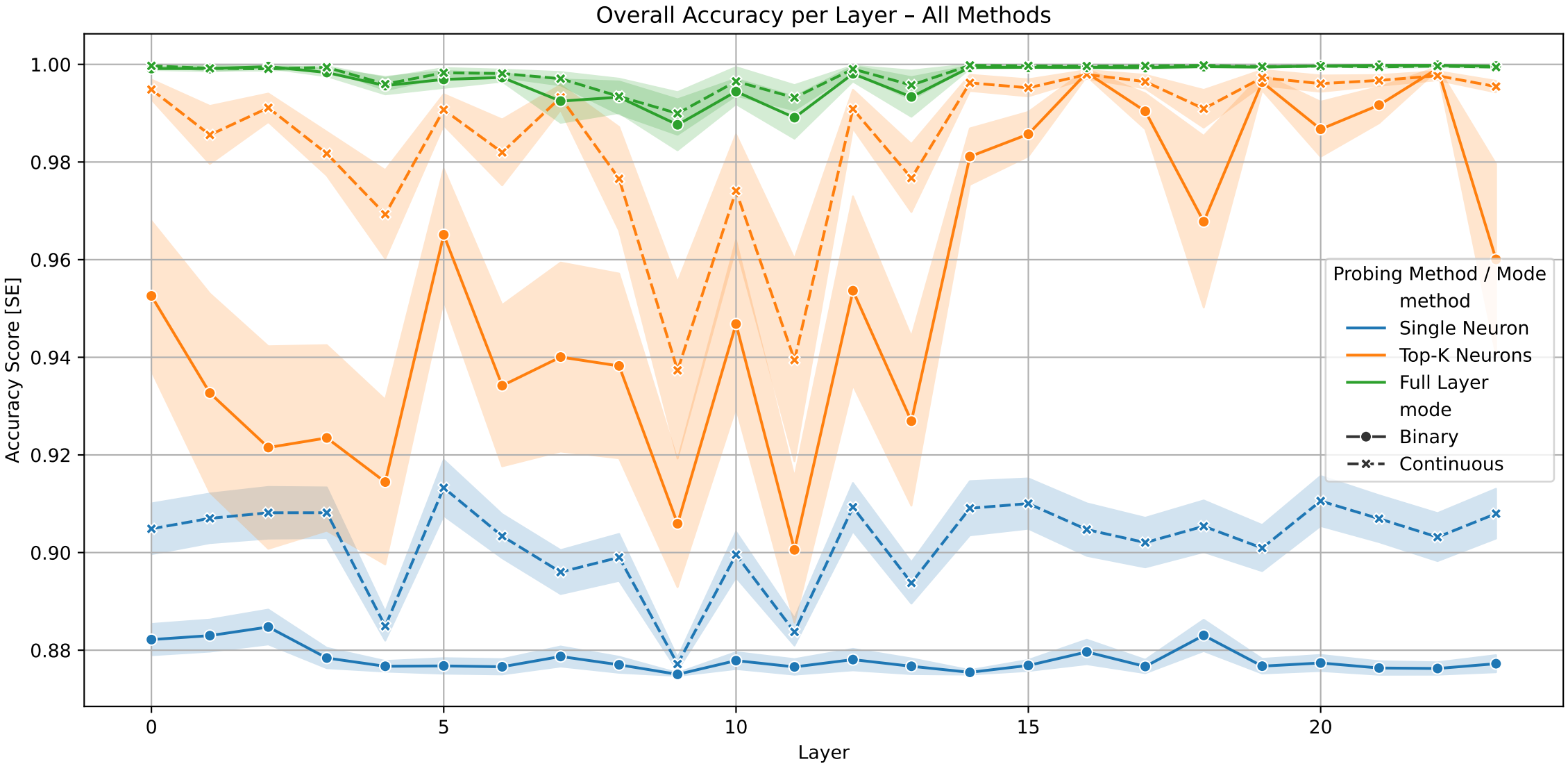
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	21.0	18.0
Full Layer	f1_max	1.0	1.0
Full Layer	f1_mean	0.9971	0.9979
Full Layer	f1_std	0.0067	0.005
Single Neuron	f1_best_layer	2.0	5.0
Single Neuron	f1_max	1.0	1.0
Single Neuron	f1_mean	0.823	0.8681
Single Neuron	f1_std	0.0292	0.0663
Top-K Neurons	f1_best_layer	22.0	16.0
Top-K Neurons	f1_max	1.0	1.0
Top-K Neurons	f1_mean	0.9456	0.9846
Top-K Neurons	f1_std	0.0645	0.0276


Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
Layer
Mode
Binary
Continuous
0.92
0.91
0.90
0.89
0.88
0
5
10
15
20


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
Layer
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0
5
10
15
20
Mode
Binary
Continuous


Accuracy per Layer – Full Layer Probing
Accuracy [SE]
1.0000
0.9975
0.9950
0.9925
0.9900
0.9875
0.9850
0.9825
0
5
10
15
20
Layer
Mode
Binary
Continuous


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0
5
10
15
20
Layer
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	21.0	14.0
Full Layer	accuracy_max	1.0	1.0
Full Layer	accuracy_mean	0.9971	0.998
Full Layer	accuracy_std	0.0067	0.005
Single Neuron	accuracy_best_layer	2.0	5.0
Single Neuron	accuracy_max	1.0	1.0
Single Neuron	accuracy_mean	0.8781	0.902
Single Neuron	accuracy_std	0.0179	0.0434
Top-K Neurons	accuracy_best_layer	22.0	16.0
Top-K Neurons	accuracy_max	1.0	1.0
Top-K Neurons	accuracy_mean	0.9547	0.9851
Top-K Neurons	accuracy_std	0.0493	0.0255