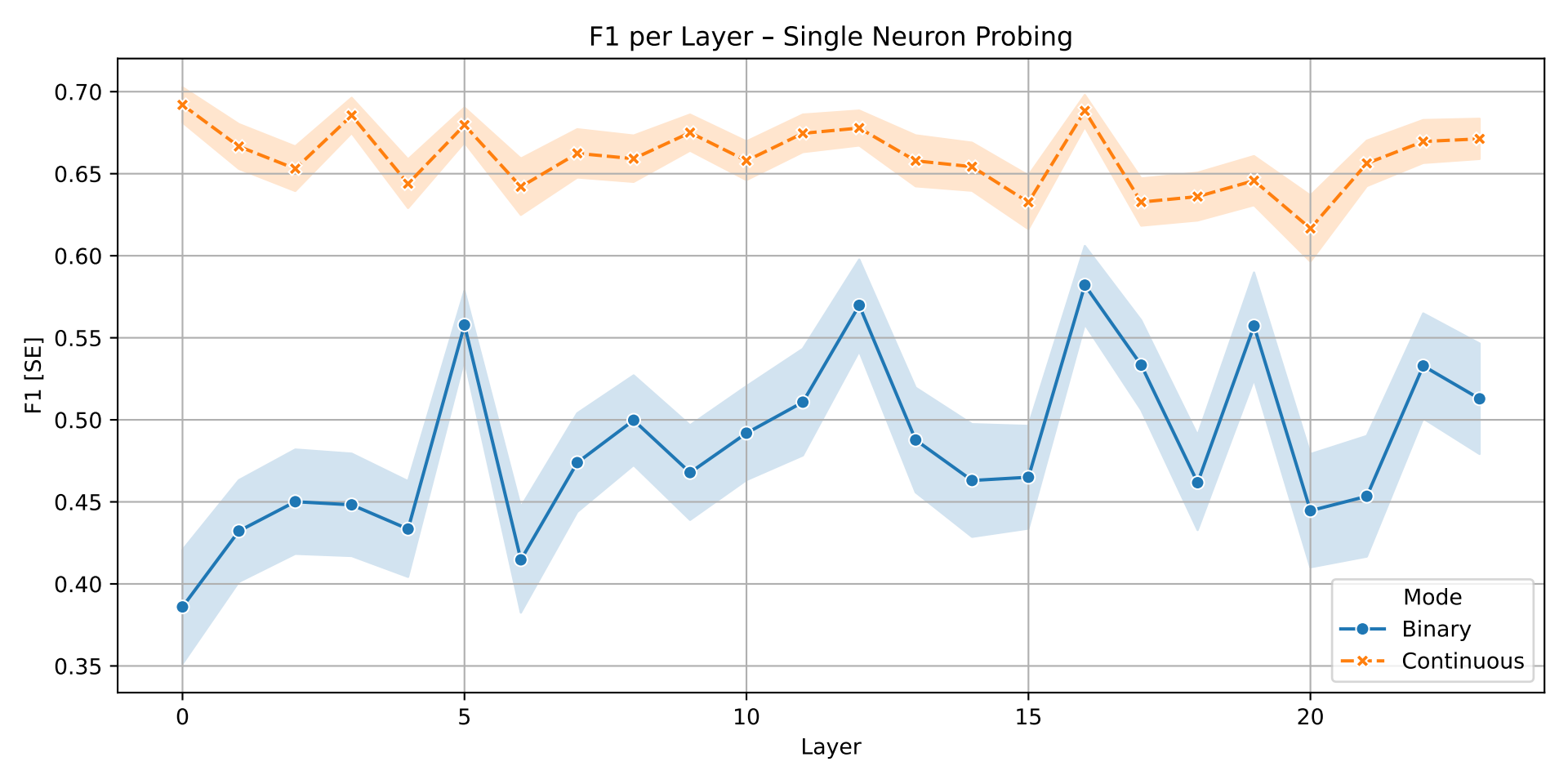
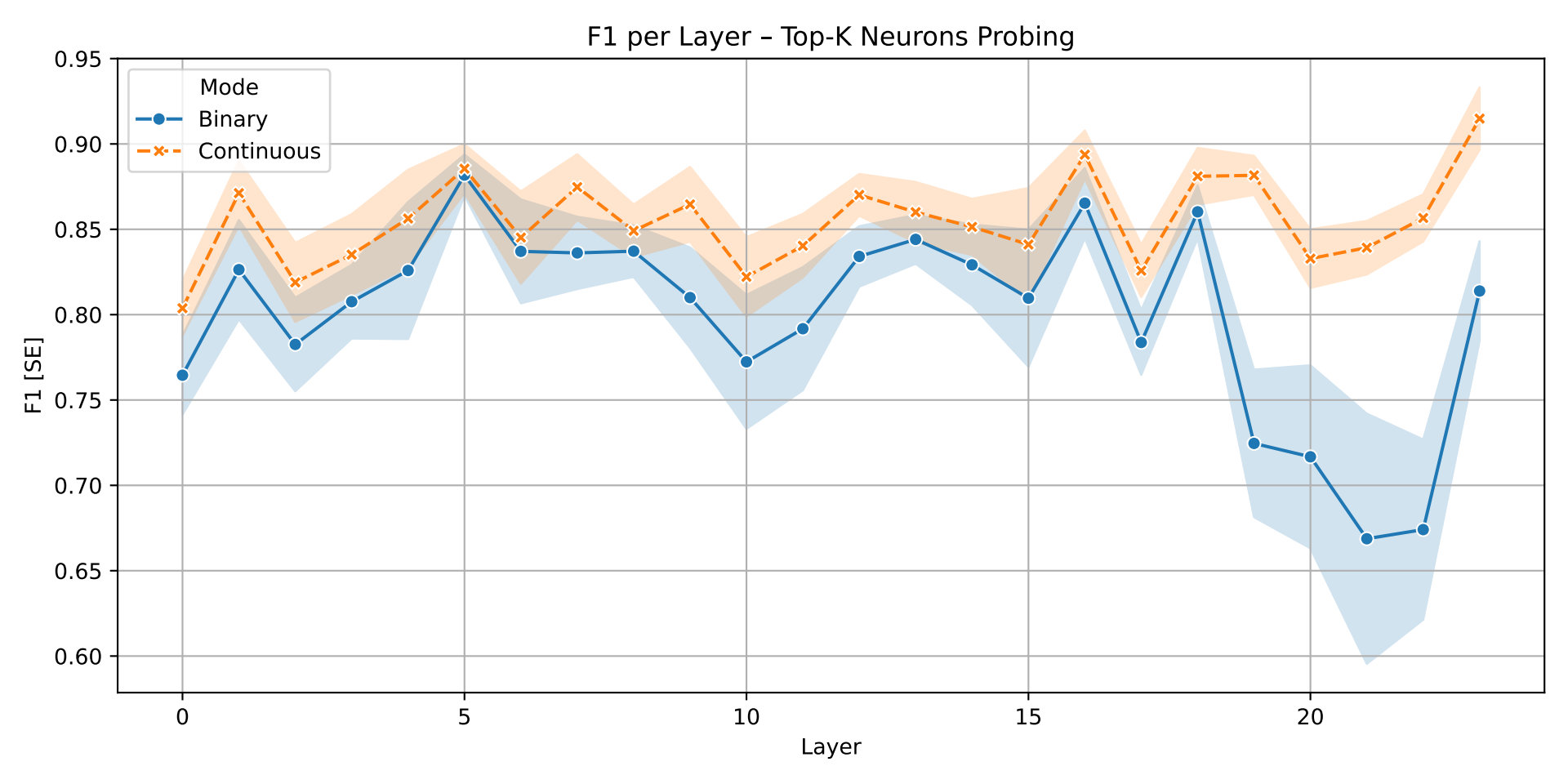
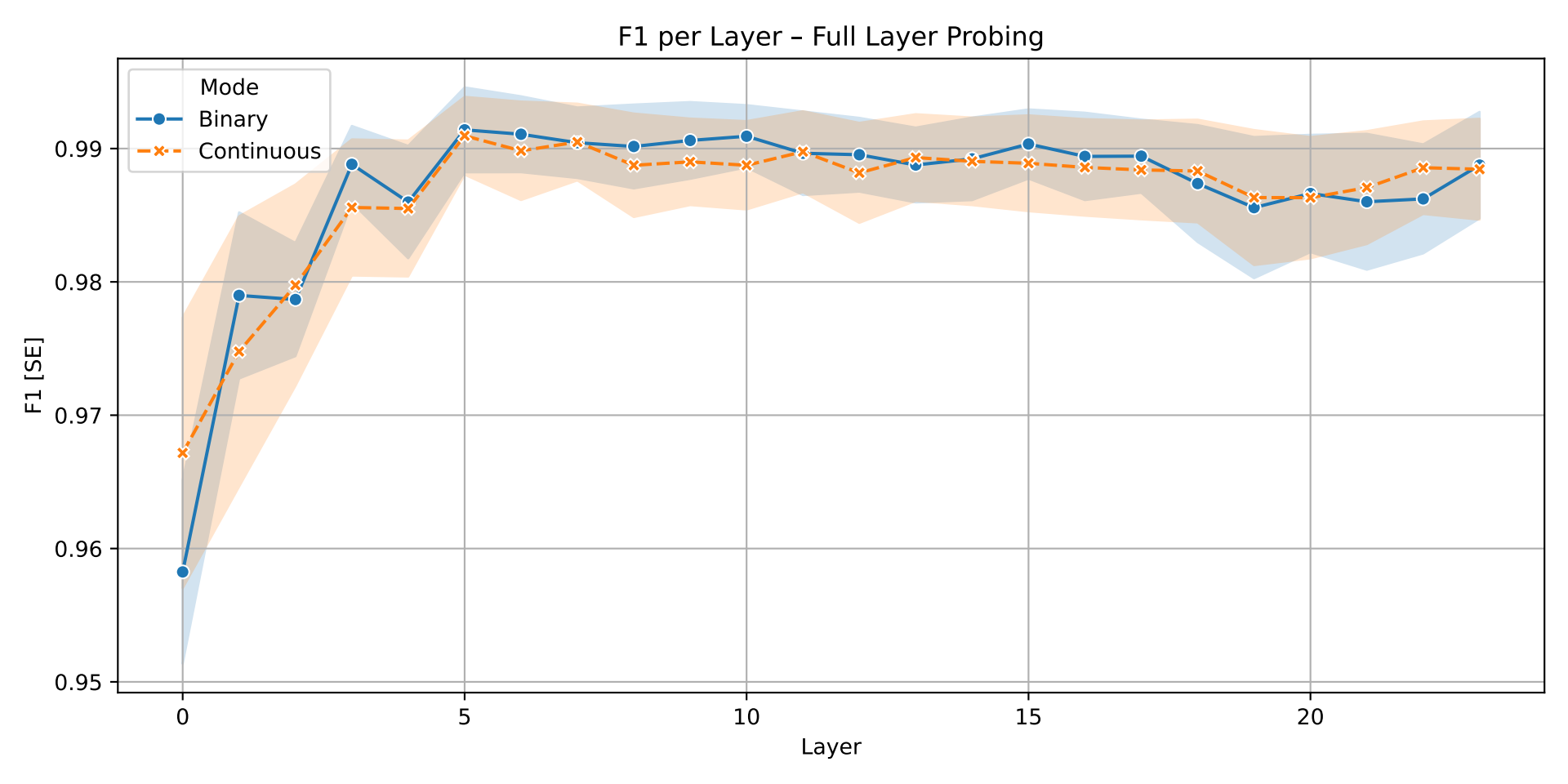
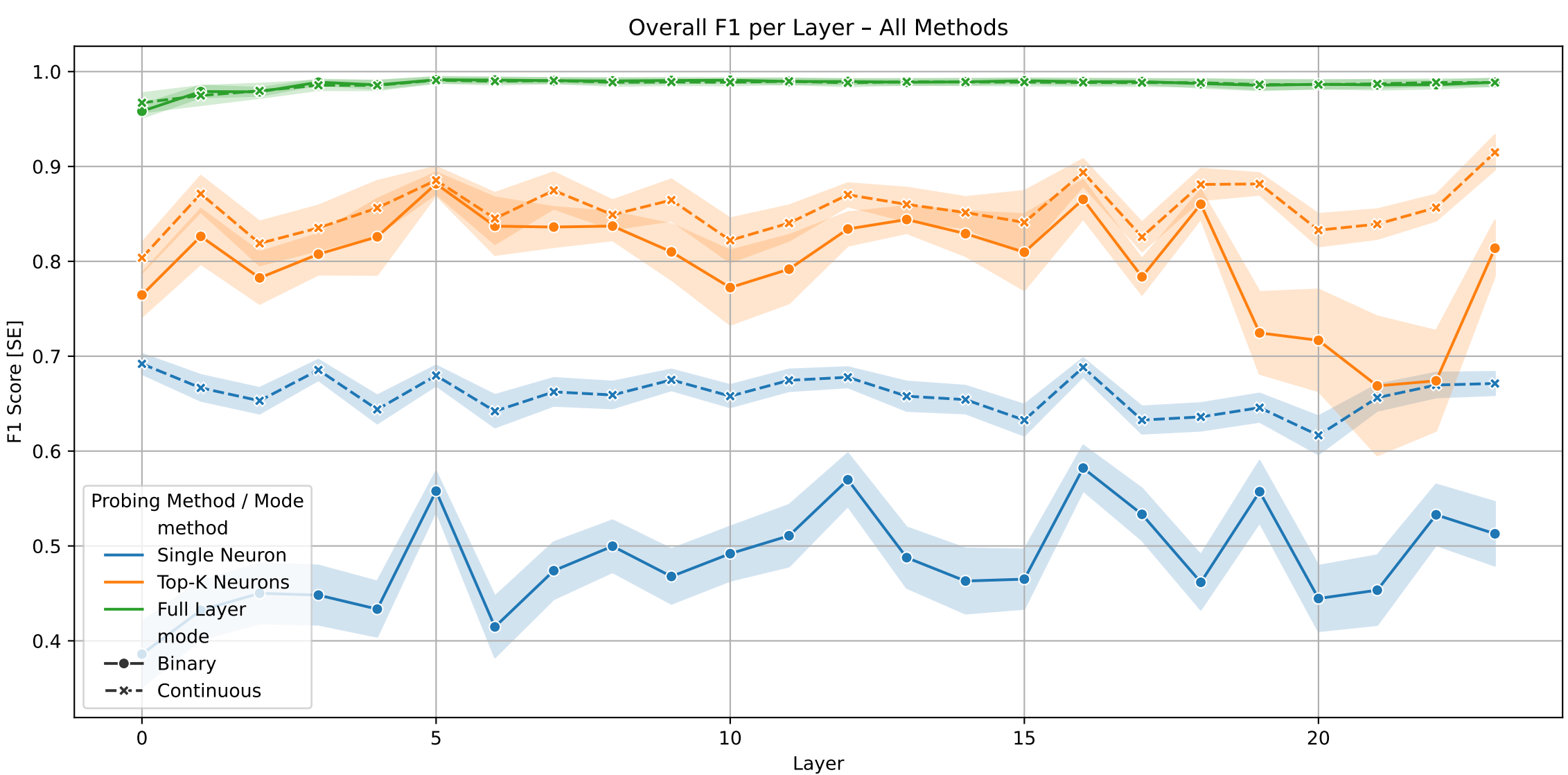

F1 per Layer – Single Neuron Probing
F1 [SE]
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0
5
10
15
20
Layer
Mode
Binary
Continuous

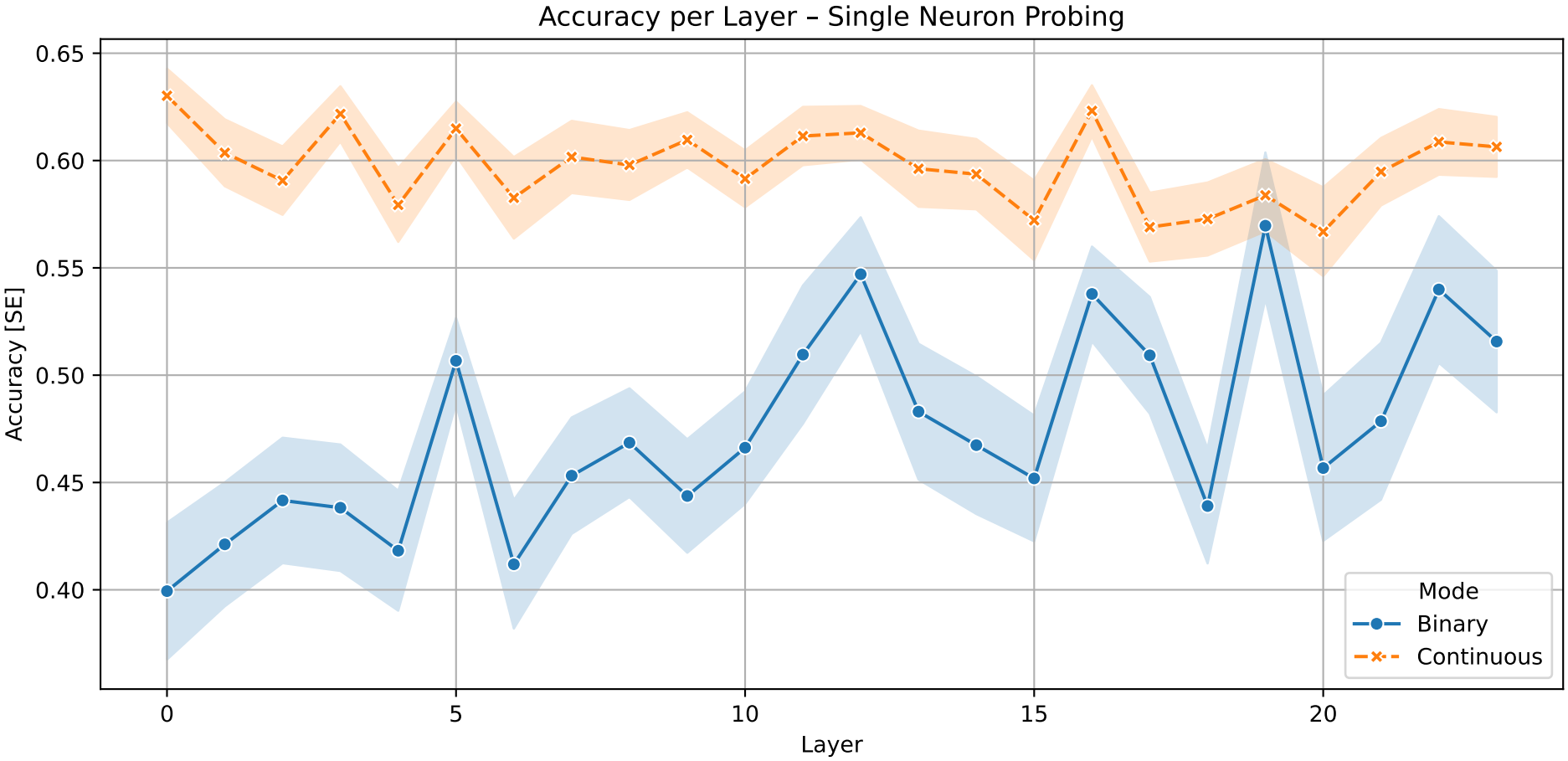

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
Layer
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0
5
10
15
20

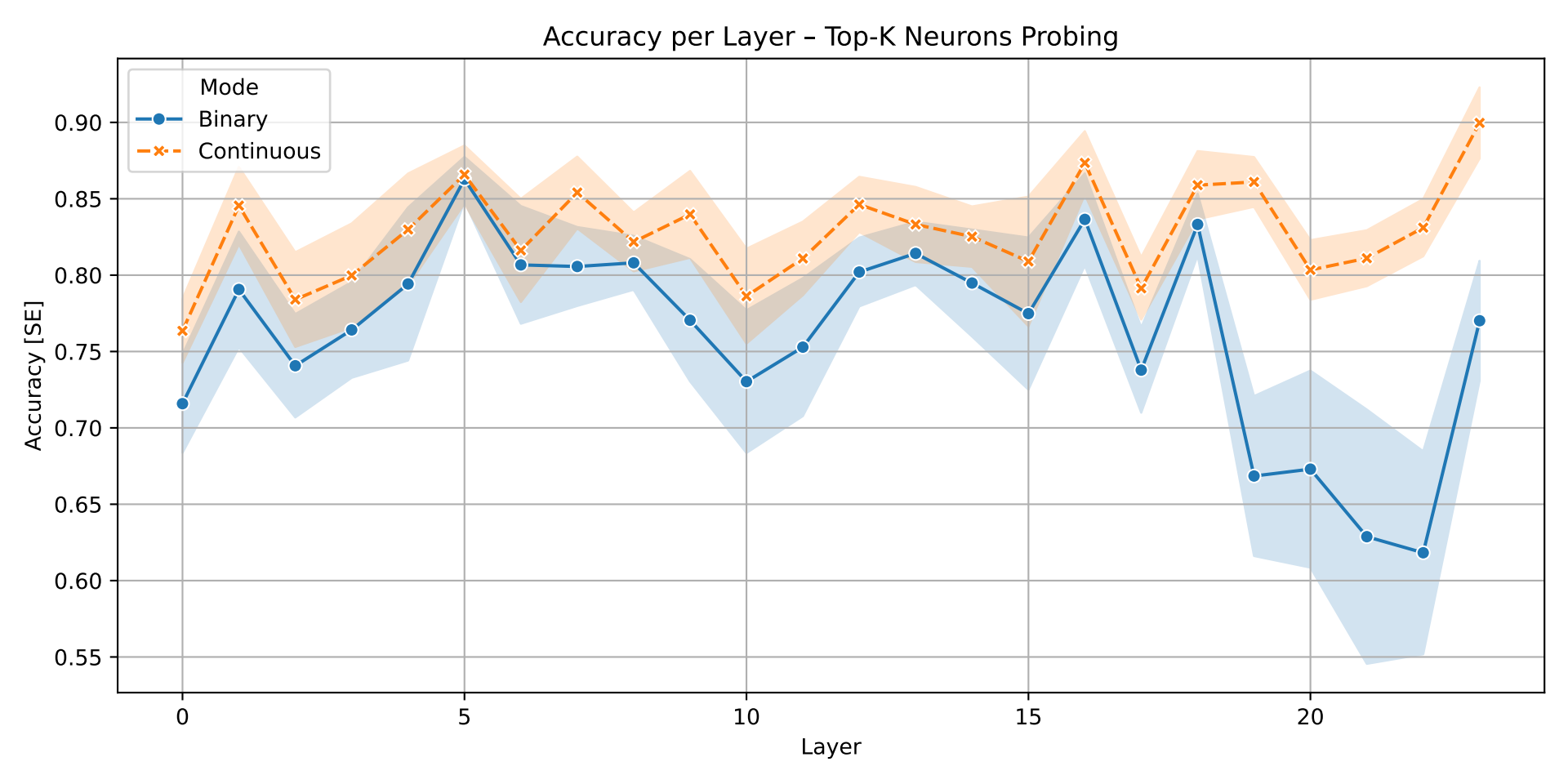

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.99
0.98
0.97
0.96
0.95
0
5
10
15
20
Layer

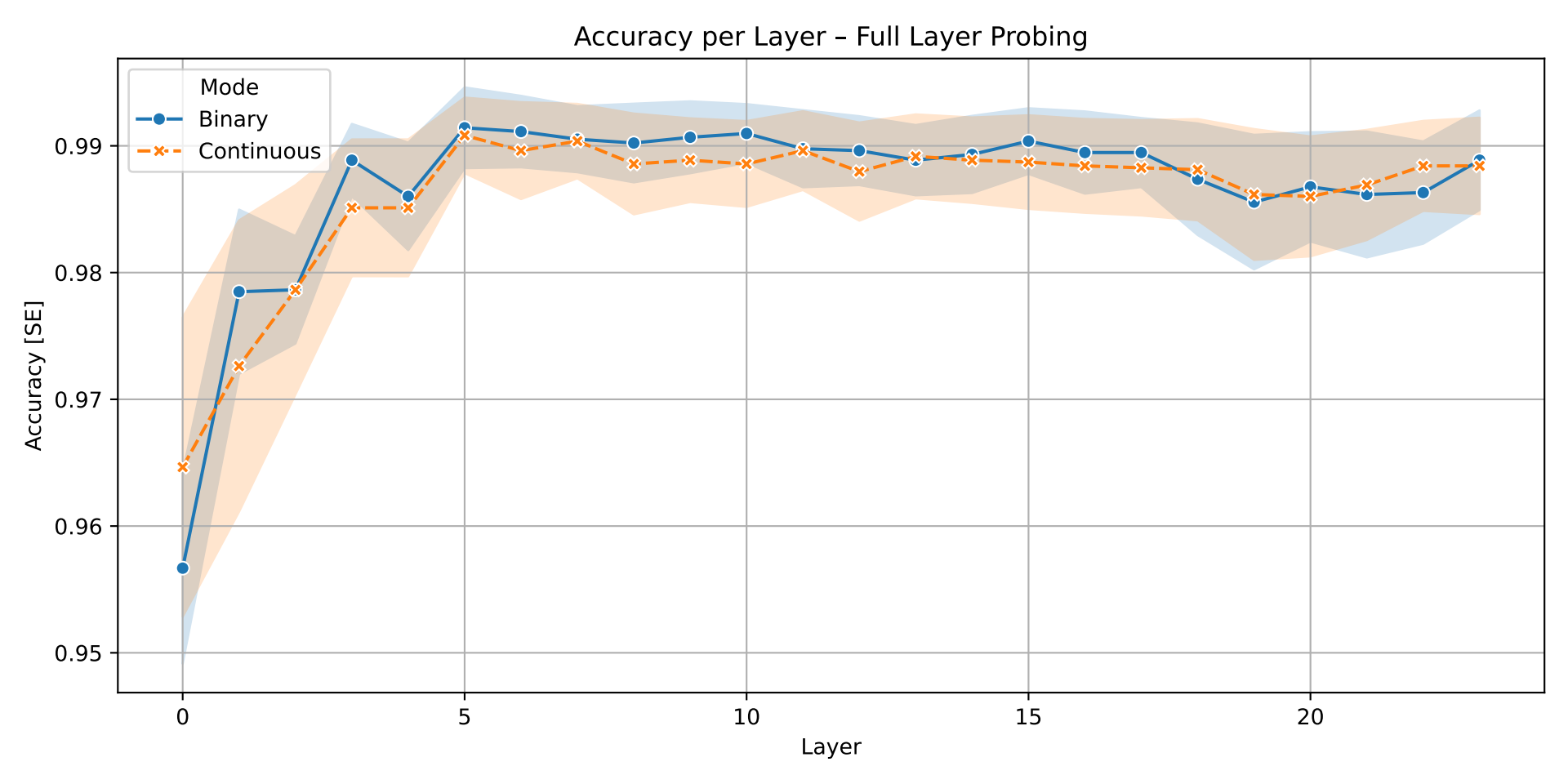

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
5
10
15
20
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

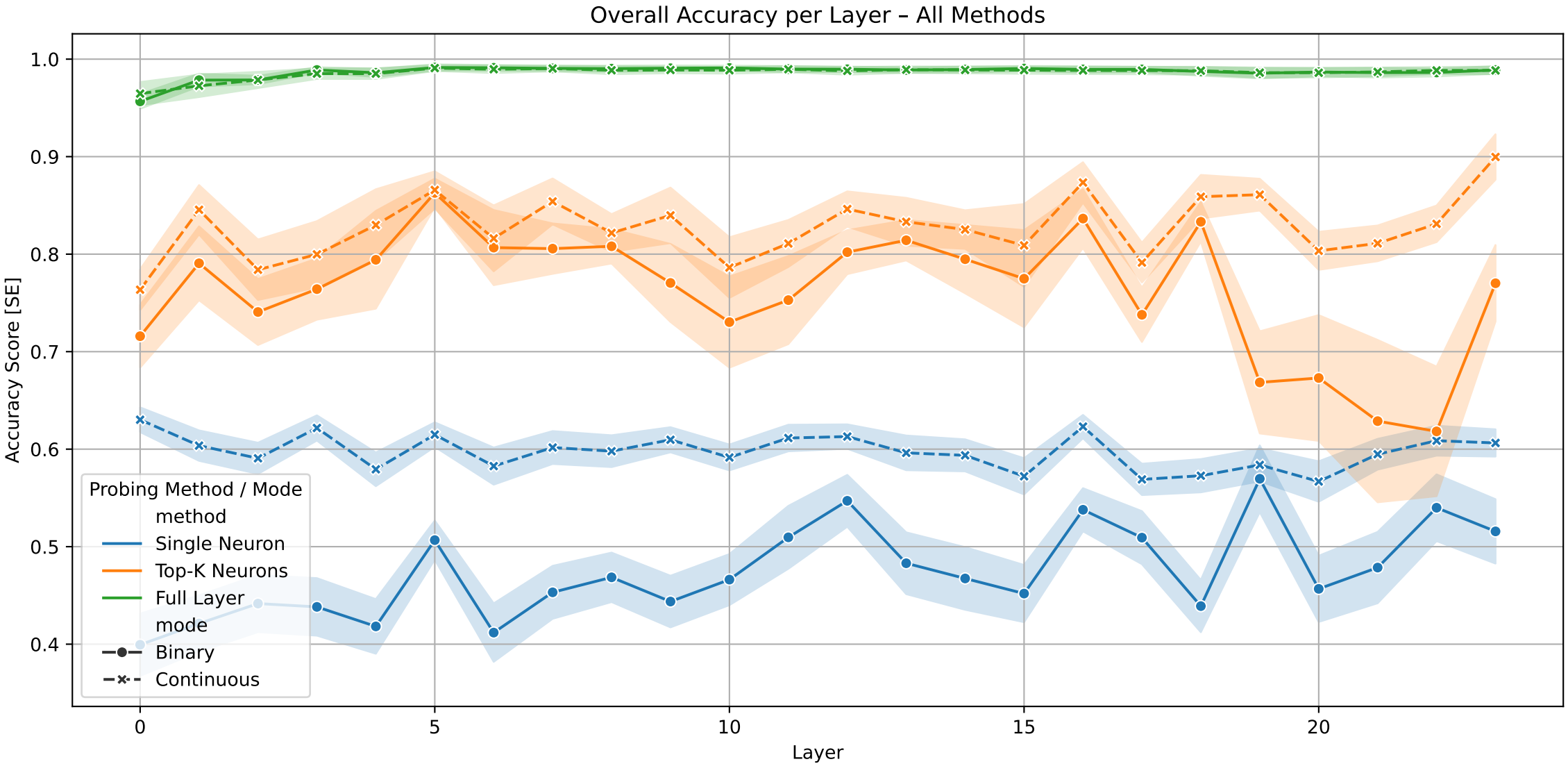
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	5.0	5.0
Full Layer	f1_max	0.9988	0.9988
Full Layer	f1_mean	0.9868	0.9866
Full Layer	f1_std	0.0124	0.014
Single Neuron	f1_best_layer	16.0	0.0
Single Neuron	f1_max	0.9833	0.9866
Single Neuron	f1_mean	0.4846	0.6596
Single Neuron	f1_std	0.2778	0.1234
Top-K Neurons	f1_best_layer	5.0	23.0
Top-K Neurons	f1_max	0.9833	0.9904
Top-K Neurons	f1_mean	0.7999	0.8548
Top-K Neurons	f1_std	0.1049	0.058


Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
0.65
0.60
0.55
0.50
0.45
0.40
0
5
10
15
20
Layer
Mode
Binary
Continuous


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
Layer


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.99
0.98
0.97
0.96
0.95
0
5
10
15
20
Layer


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
Layer
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
5
10
15
20
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	5.0	5.0
Full Layer	accuracy_max	0.9988	0.9988
Full Layer	accuracy_mean	0.9867	0.9862
Full Layer	accuracy_std	0.0126	0.0152
Single Neuron	accuracy_best_layer	19.0	0.0
Single Neuron	accuracy_max	0.9832	0.9868
Single Neuron	accuracy_mean	0.4739	0.5973
Single Neuron	accuracy_std	0.2662	0.1401
Top-K Neurons	accuracy_best_layer	5.0	23.0
Top-K Neurons	accuracy_max	0.9832	0.9904
Top-K Neurons	accuracy_mean	0.7623	0.8275
Top-K Neurons	accuracy_std	0.1275	0.0747