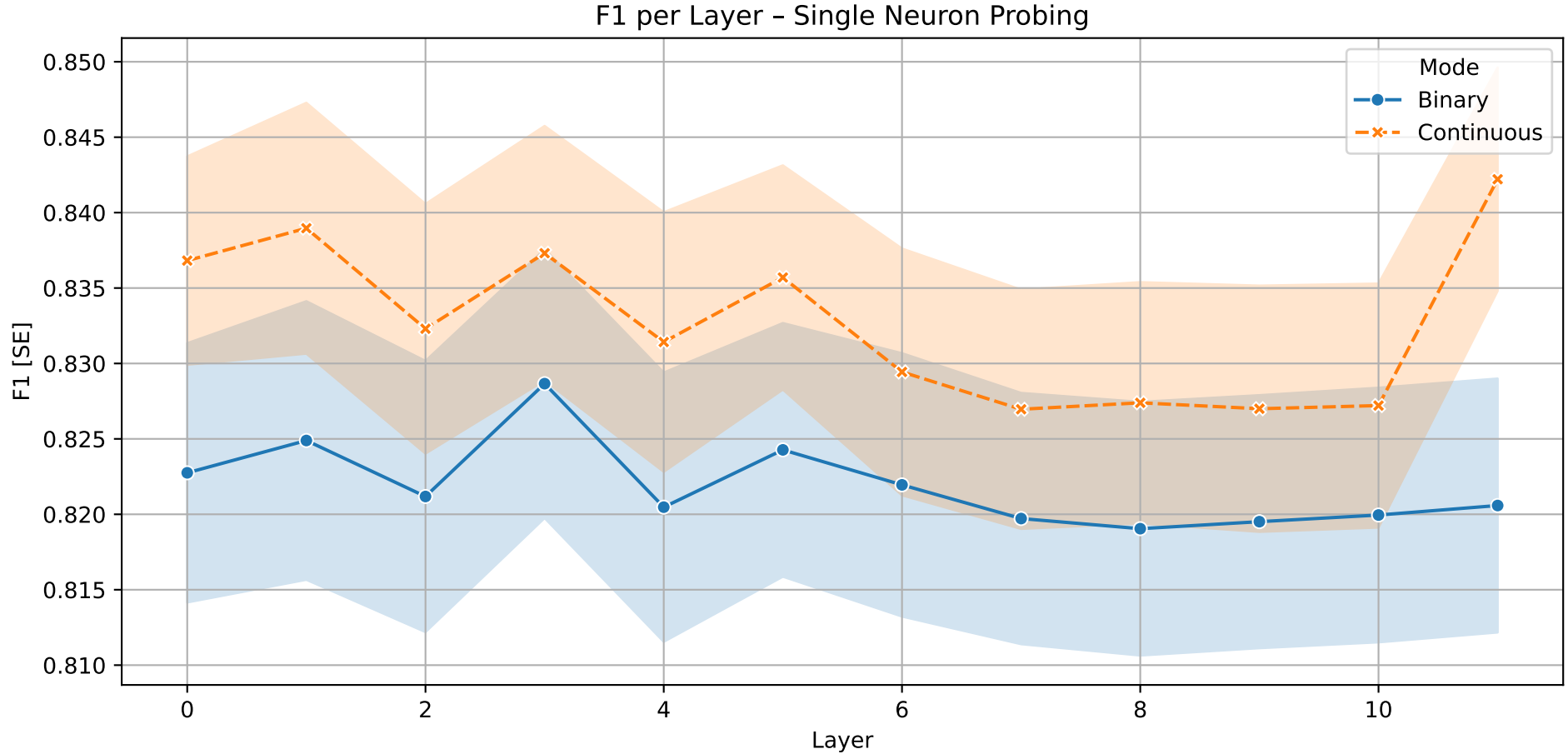
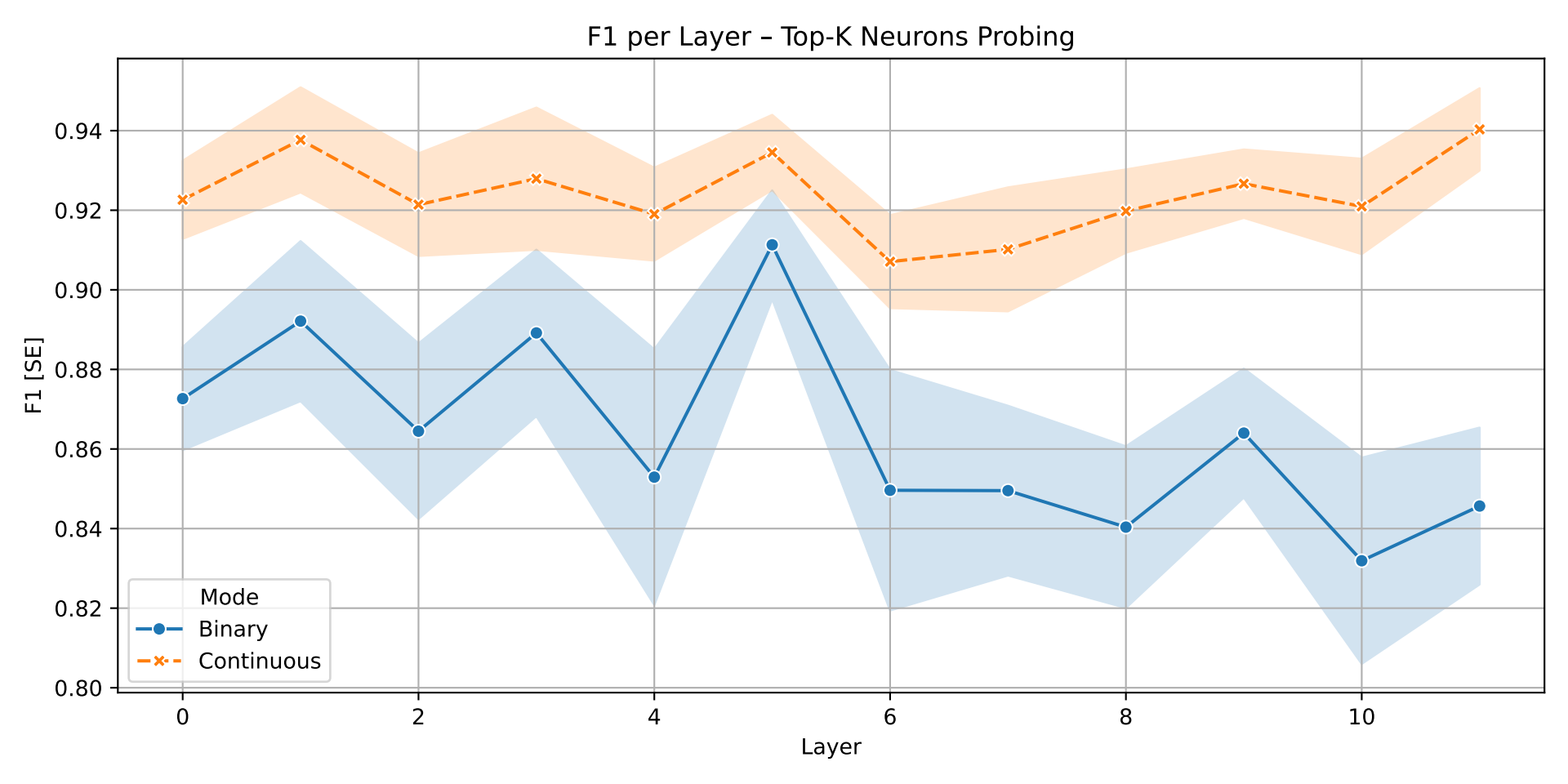
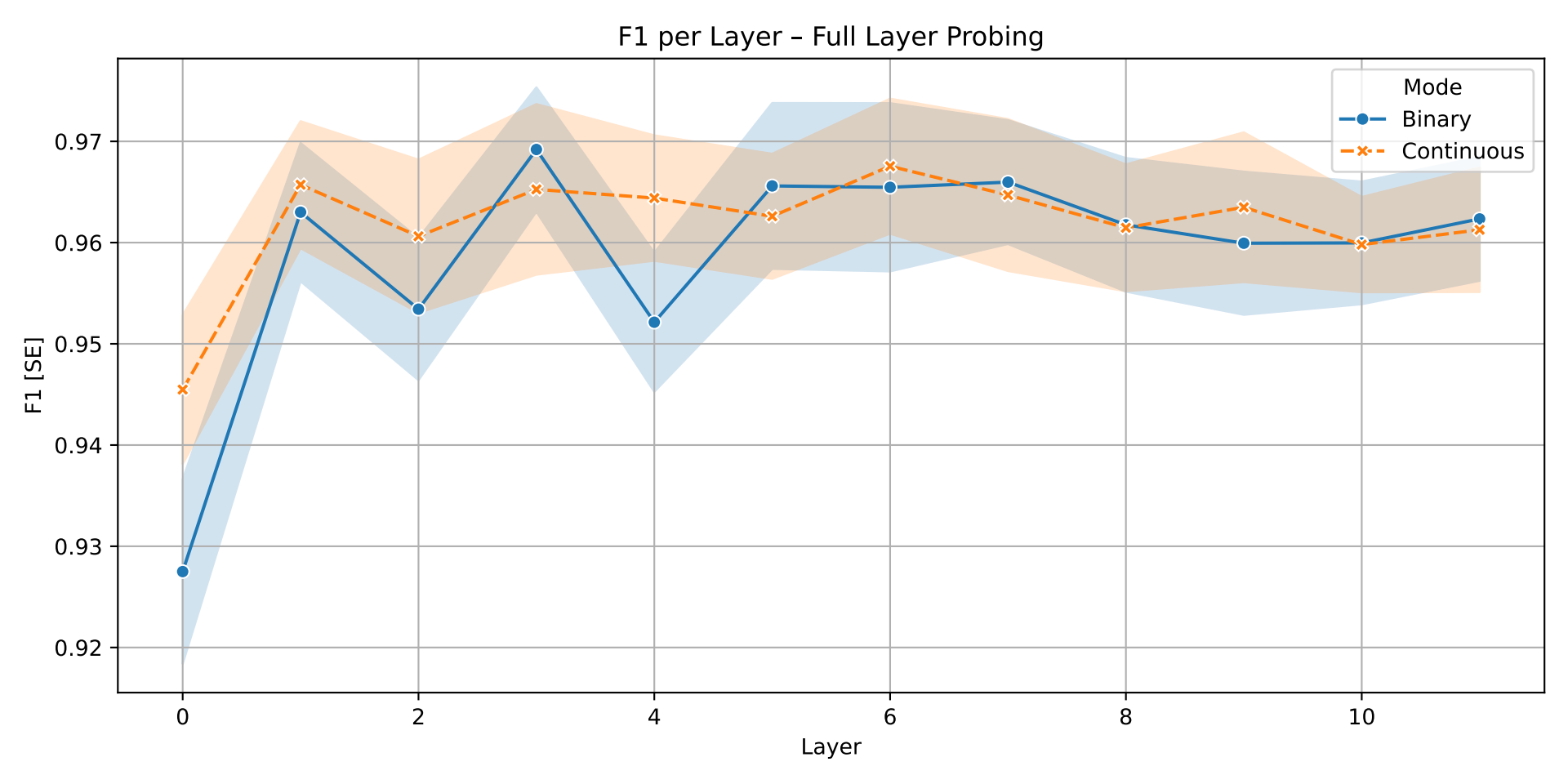
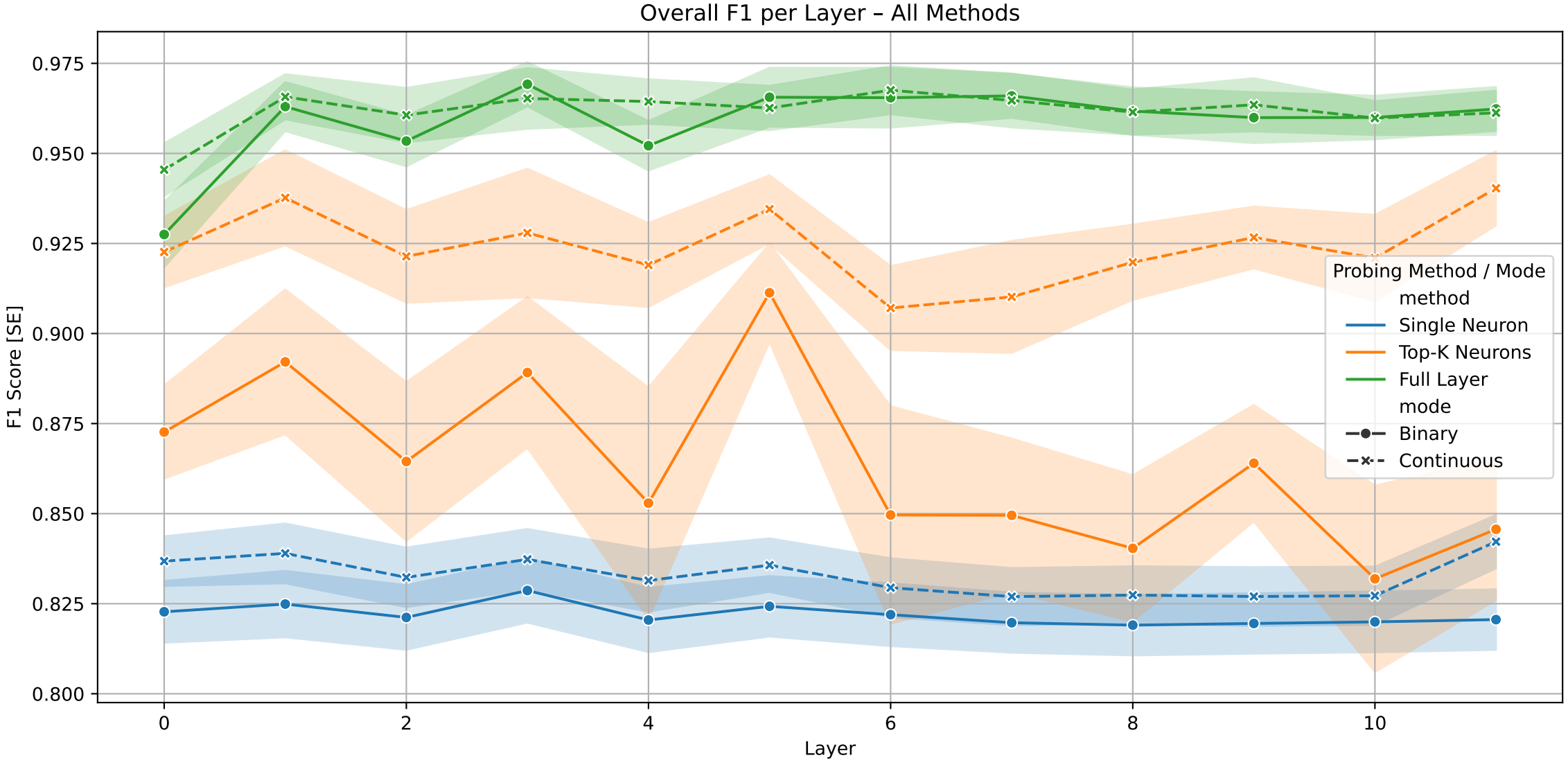

F1 per Layer – Single Neuron Probing
F1 [SE]
Layer
Mode
Binary
Continuous
0.850
0.845
0.840
0.835
0.830
0.825
0.820
0.815
0.810
0
2
4
6
8
10

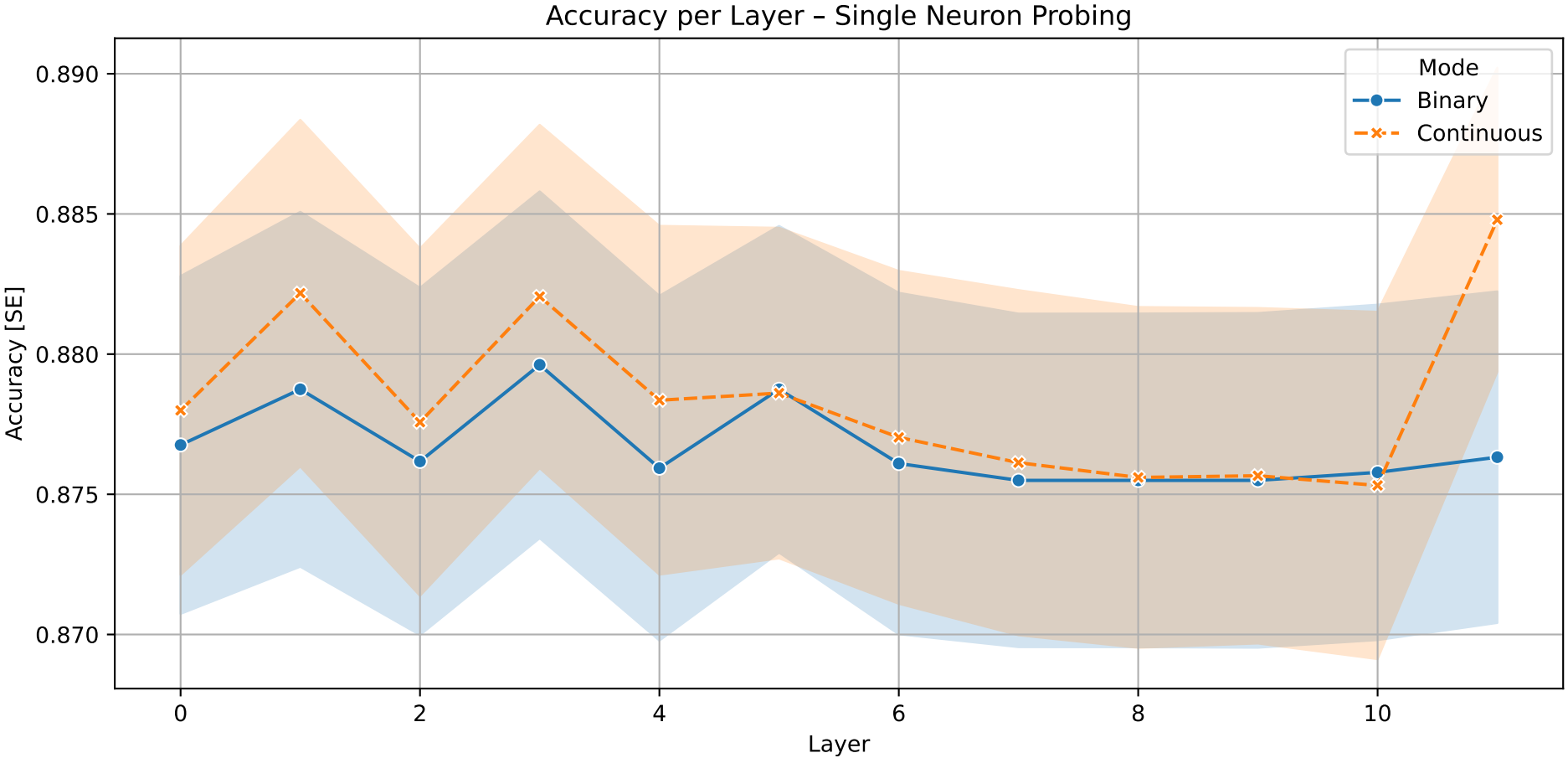

F1 per Layer – Top-K Neurons Probing
F1 [SE]
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0
2
4
6
8
10
Layer
Mode
Binary
Continuous

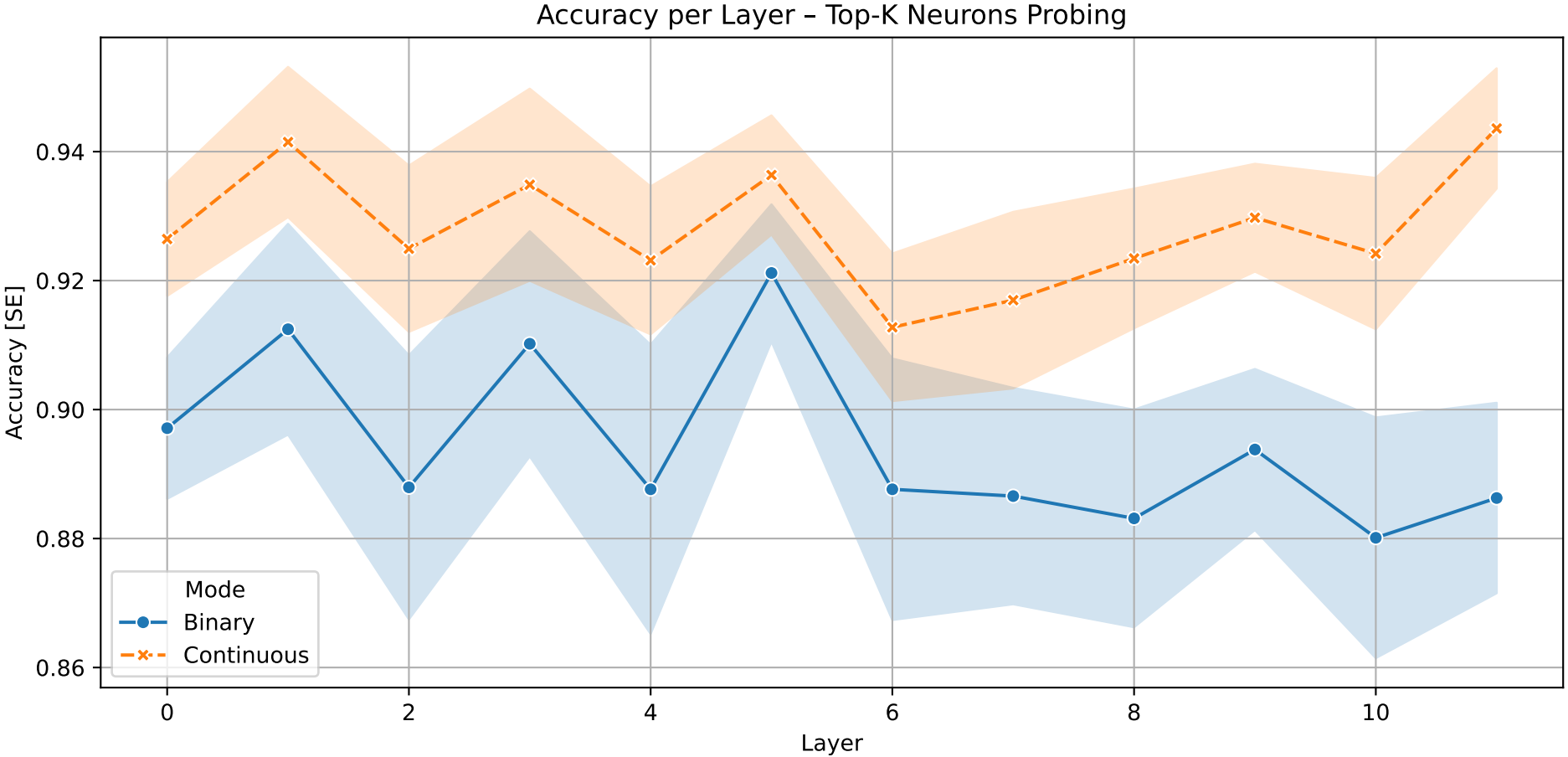

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.97
0.96
0.95
0.94
0.93
0.92
0
2
4
6
8
10
Layer

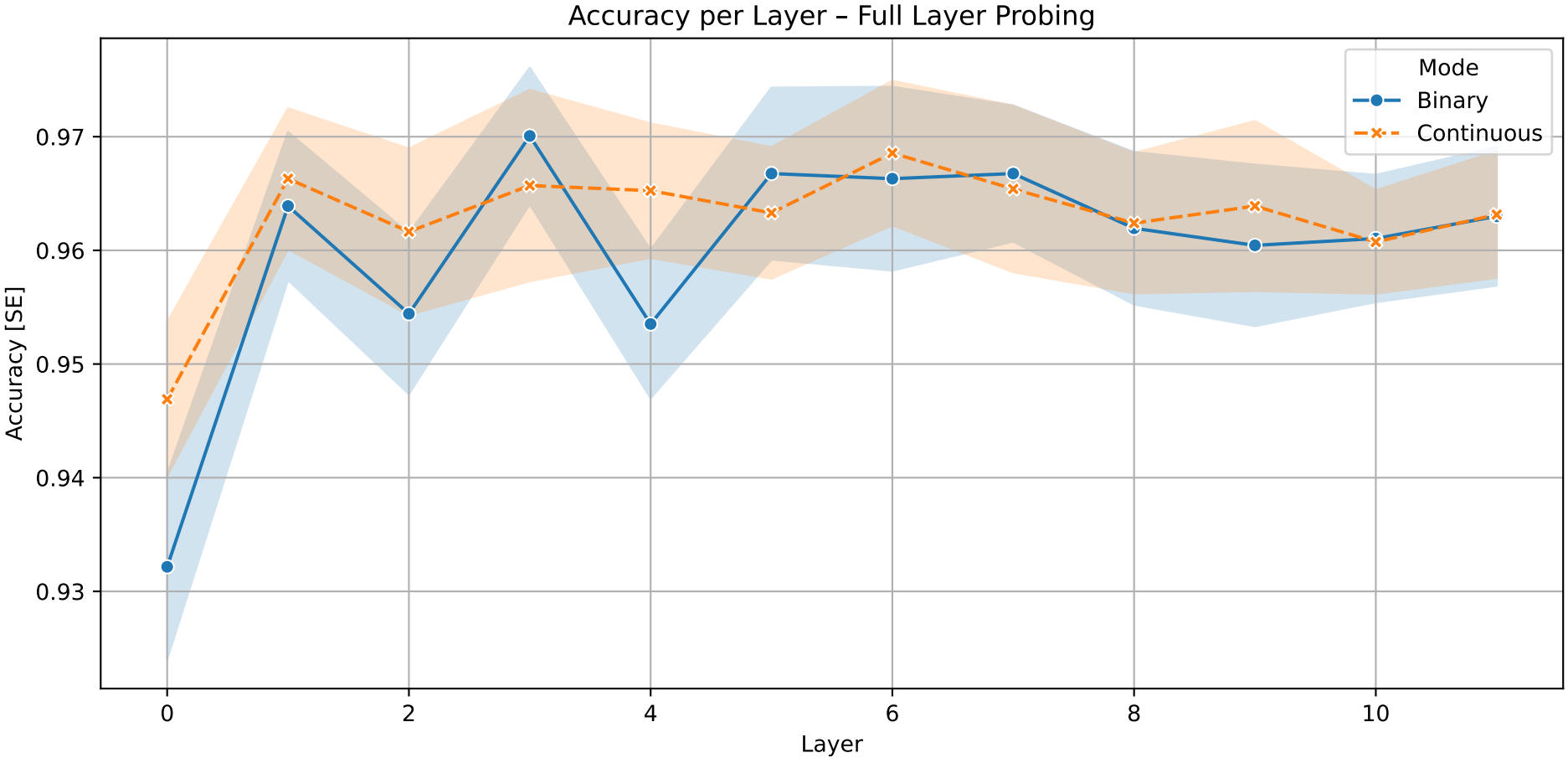

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
0
2
4
6
8
10
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

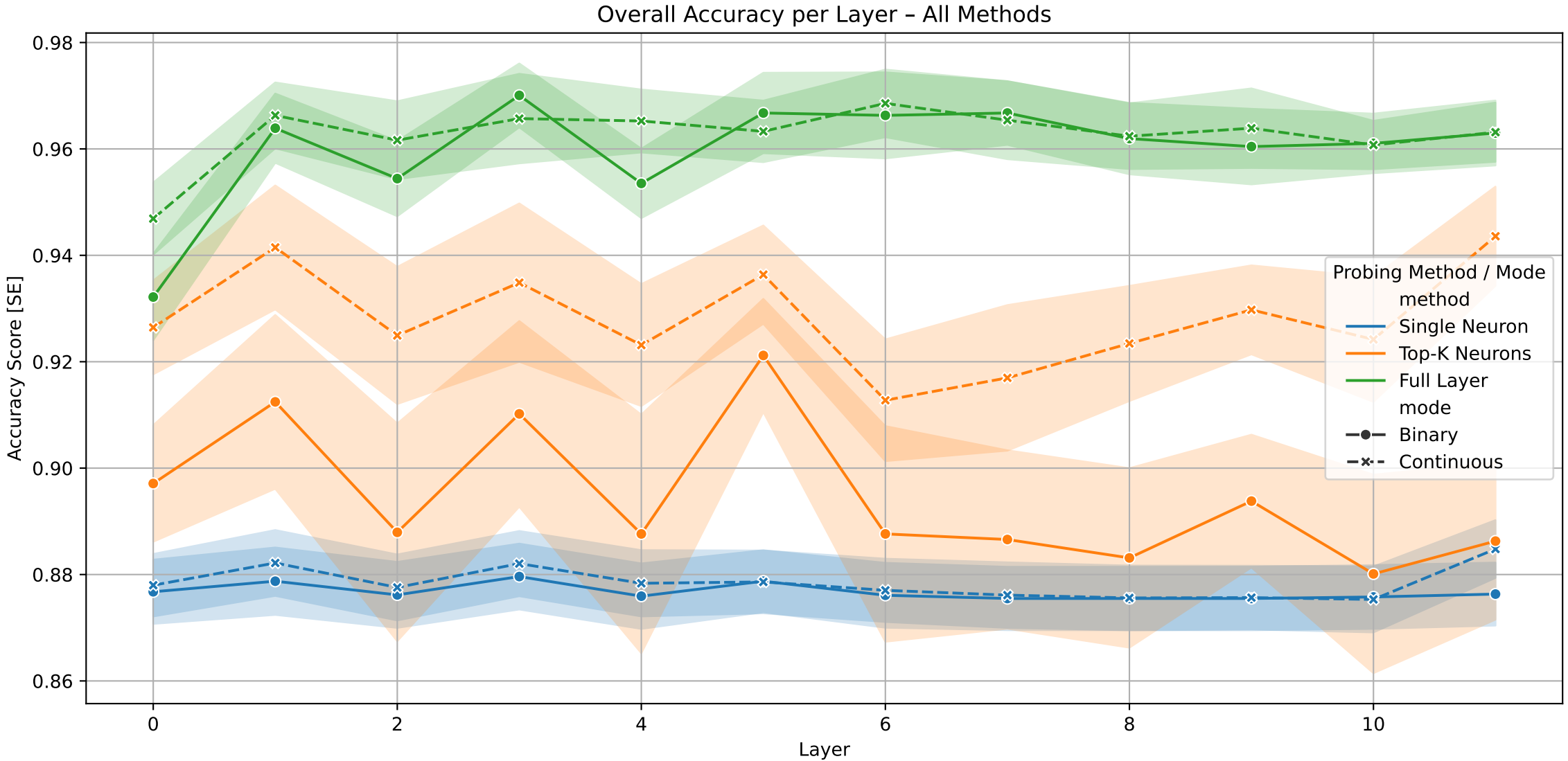
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	3.0	6.0
Full Layer	f1_max	0.9988	0.9976
Full Layer	f1_mean	0.9589	0.9619
Full Layer	f1_std	0.0217	0.0188
Single Neuron	f1_best_layer	3.0	11.0
Single Neuron	f1_max	0.9761	0.9814
Single Neuron	f1_mean	0.8219	0.8327
Single Neuron	f1_std	0.077	0.0713
Top-K Neurons	f1_best_layer	5.0	11.0
Top-K Neurons	f1_max	0.9738	0.9802
Top-K Neurons	f1_mean	0.8636	0.924
Top-K Neurons	f1_std	0.0628	0.0339


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.890
0.885
0.880
0.875
0.870
0
2
4
6
8
10
Layer


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
0.94
0.92
0.90
0.88
0.86
0
2
4
6
8
10
Layer
Mode
Binary
Continuous


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.97
0.96
0.95
0.94
0.93
0
2
4
6
8
10
Layer


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
Layer
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0
2
4
6
8
10
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	3.0	6.0
Full Layer	accuracy_max	0.9988	0.9976
Full Layer	accuracy_mean	0.96	0.9628
Full Layer	accuracy_std	0.0206	0.0182
Single Neuron	accuracy_best_layer	3.0	11.0
Single Neuron	accuracy_max	0.9771	0.9819
Single Neuron	accuracy_mean	0.8767	0.8784
Single Neuron	accuracy_std	0.0538	0.0536
Top-K Neurons	accuracy_best_layer	5.0	11.0
Top-K Neurons	accuracy_max	0.9747	0.9807
Top-K Neurons	accuracy_mean	0.8945	0.9282
Top-K Neurons	accuracy_std	0.0466	0.0315