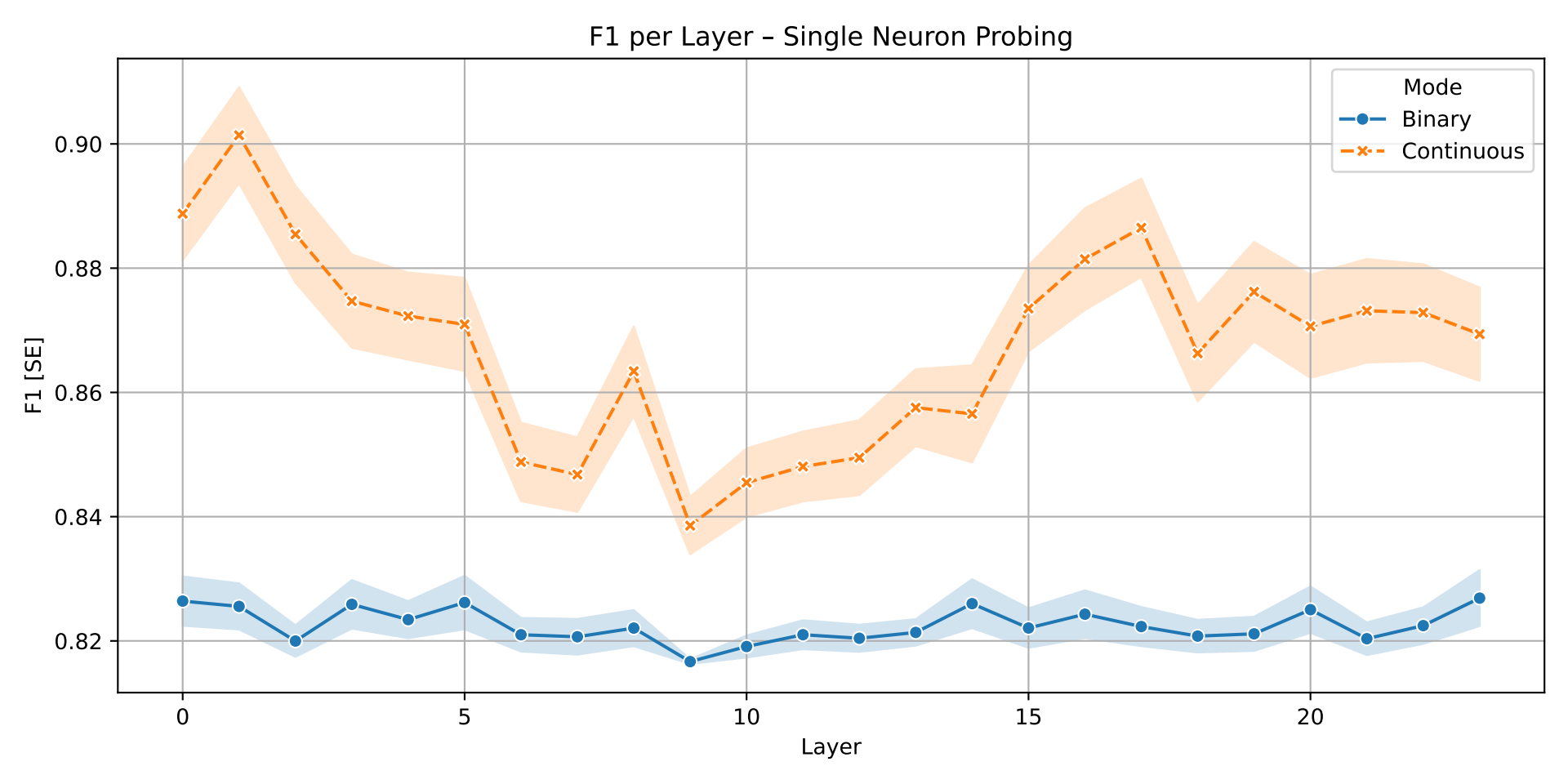
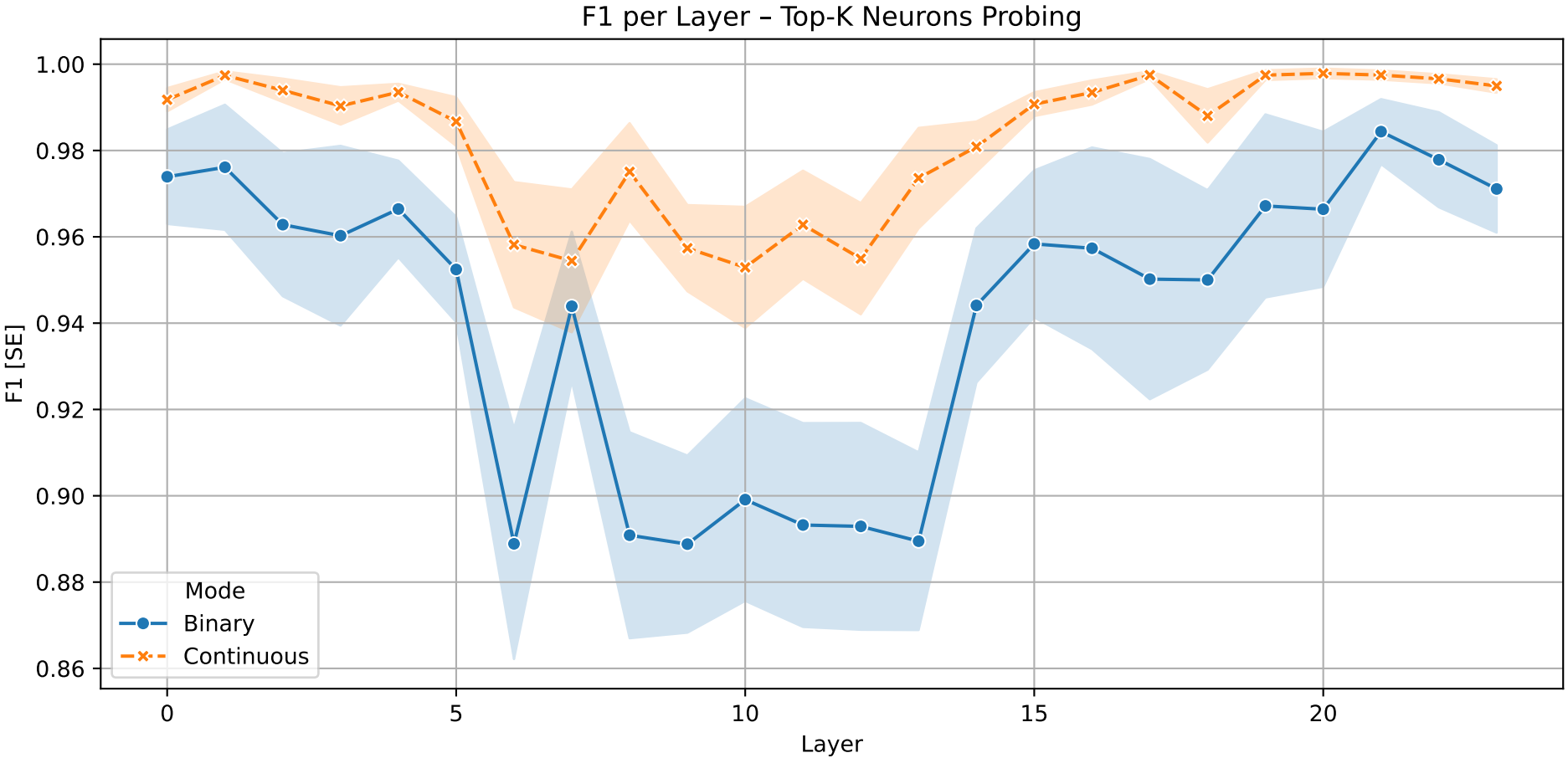
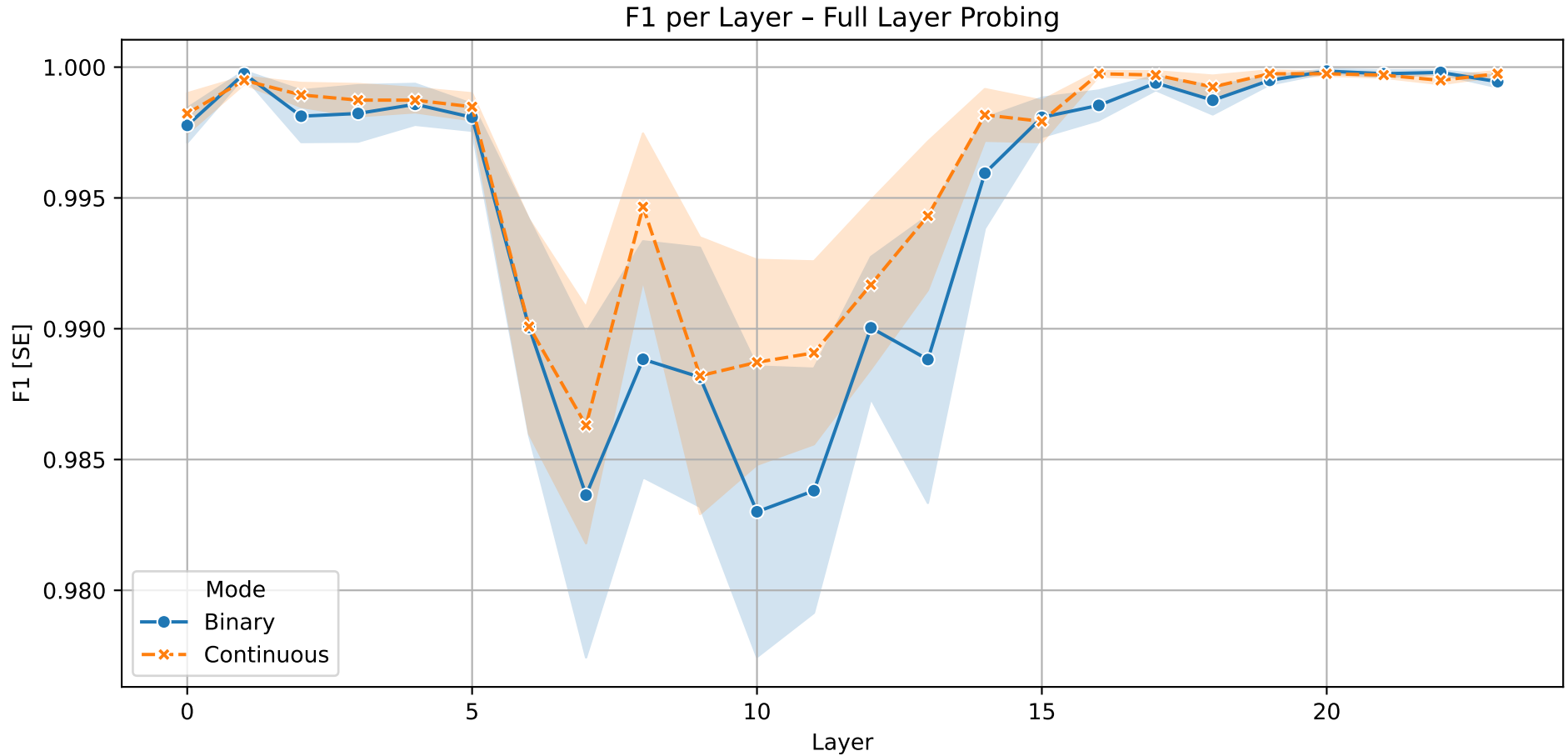
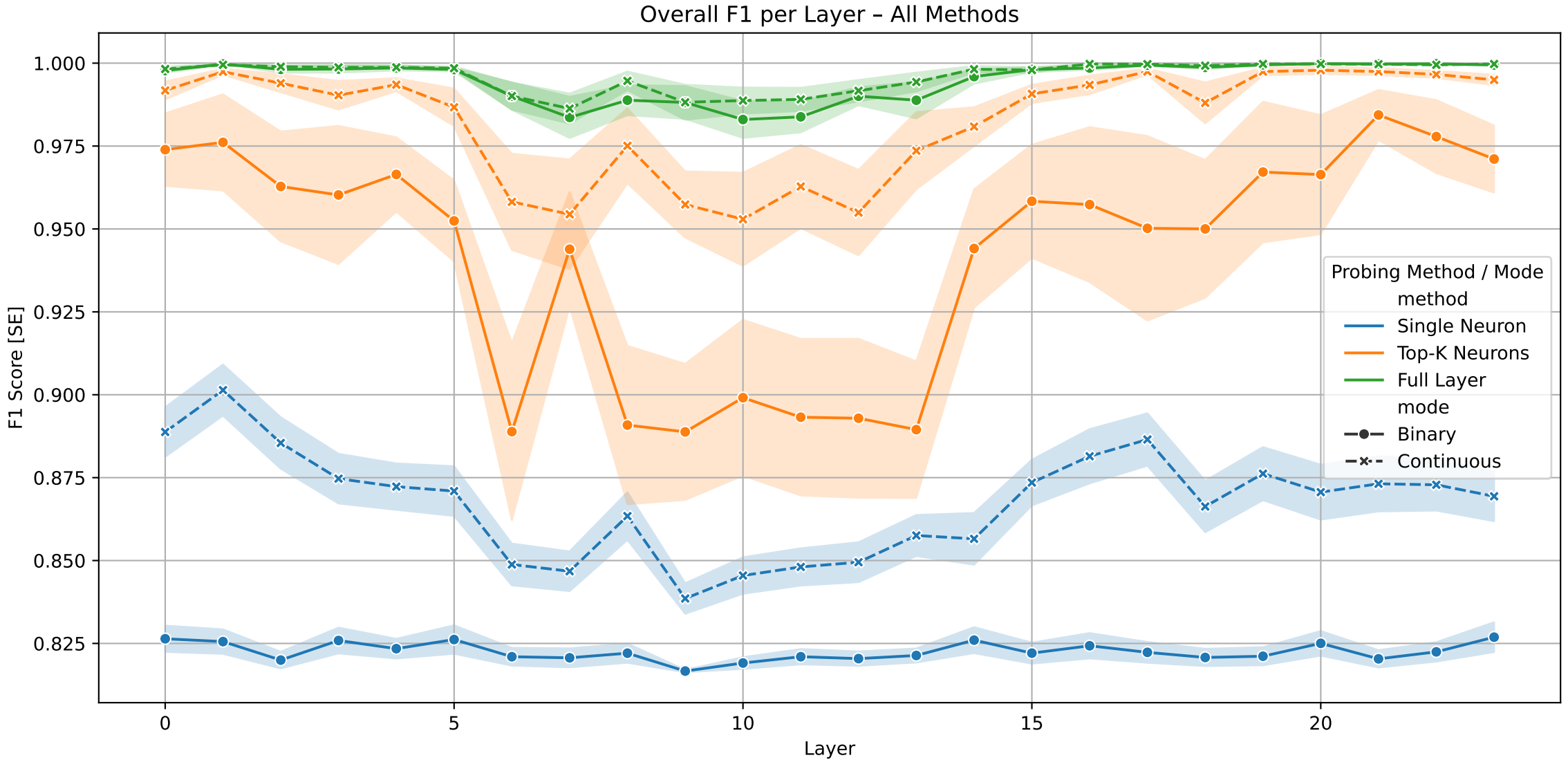

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.90
0.88
0.86
0.84
0.82
0
5
10
15
20
Layer

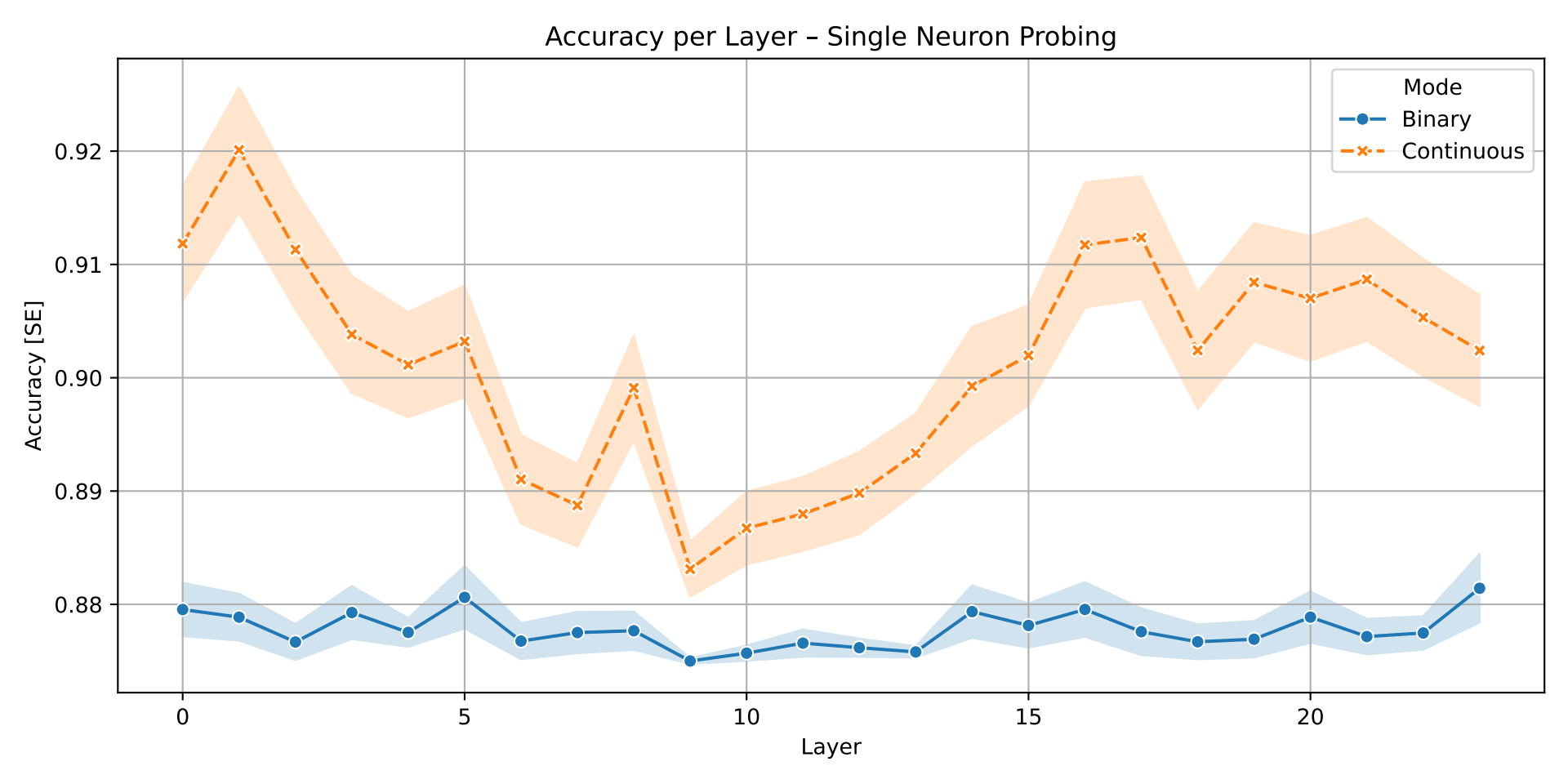

F1 per Layer – Top-K Neurons Probing
F1 [SE]
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0
5
10
15
20
Layer
Mode
Binary
Continuous

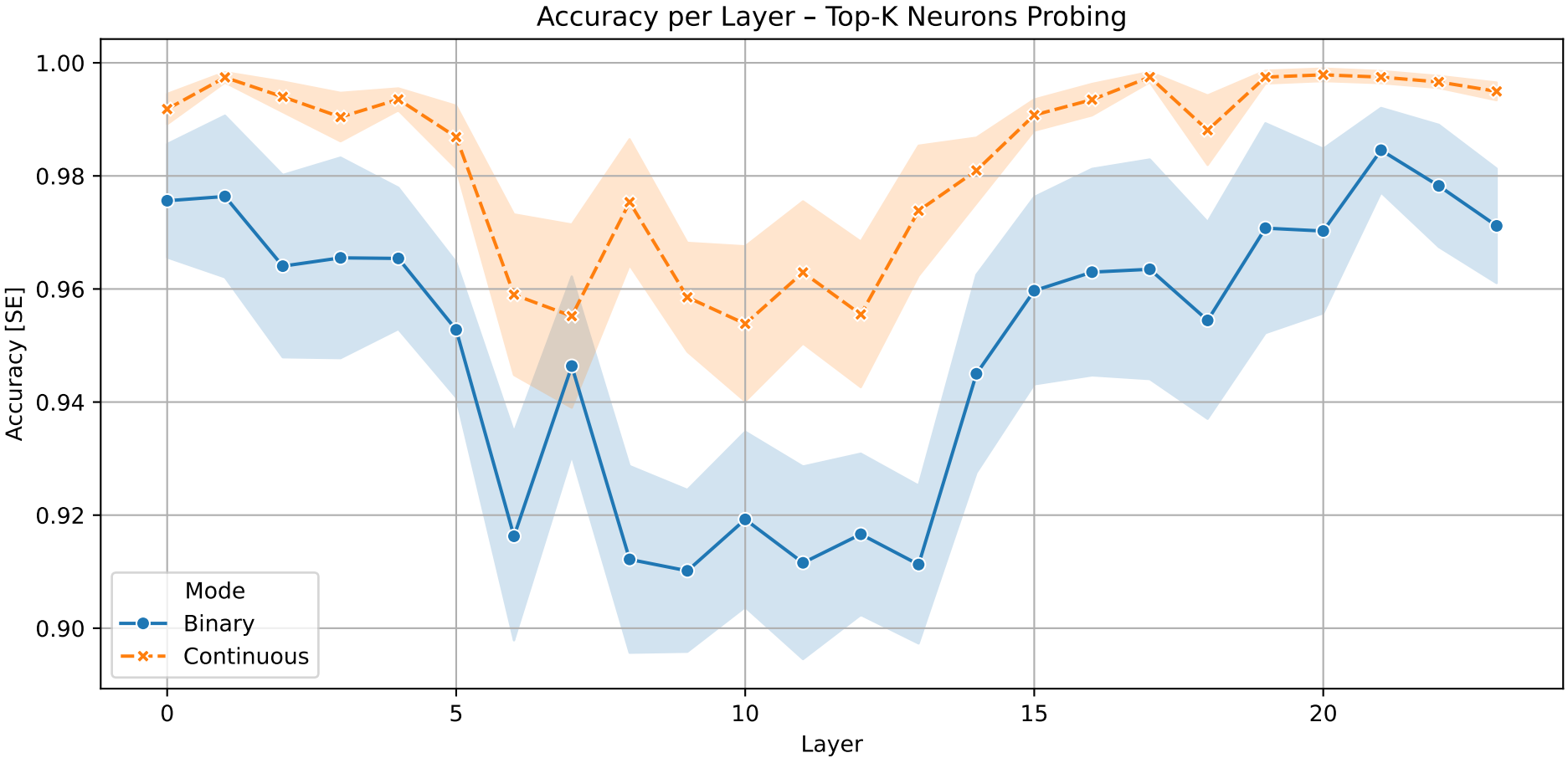

F1 per Layer – Full Layer Probing
1.000
0.995
0.990
0.985
0.980
F1 [SE]
0
5
10
15
20
Layer
Mode
Binary
Continuous

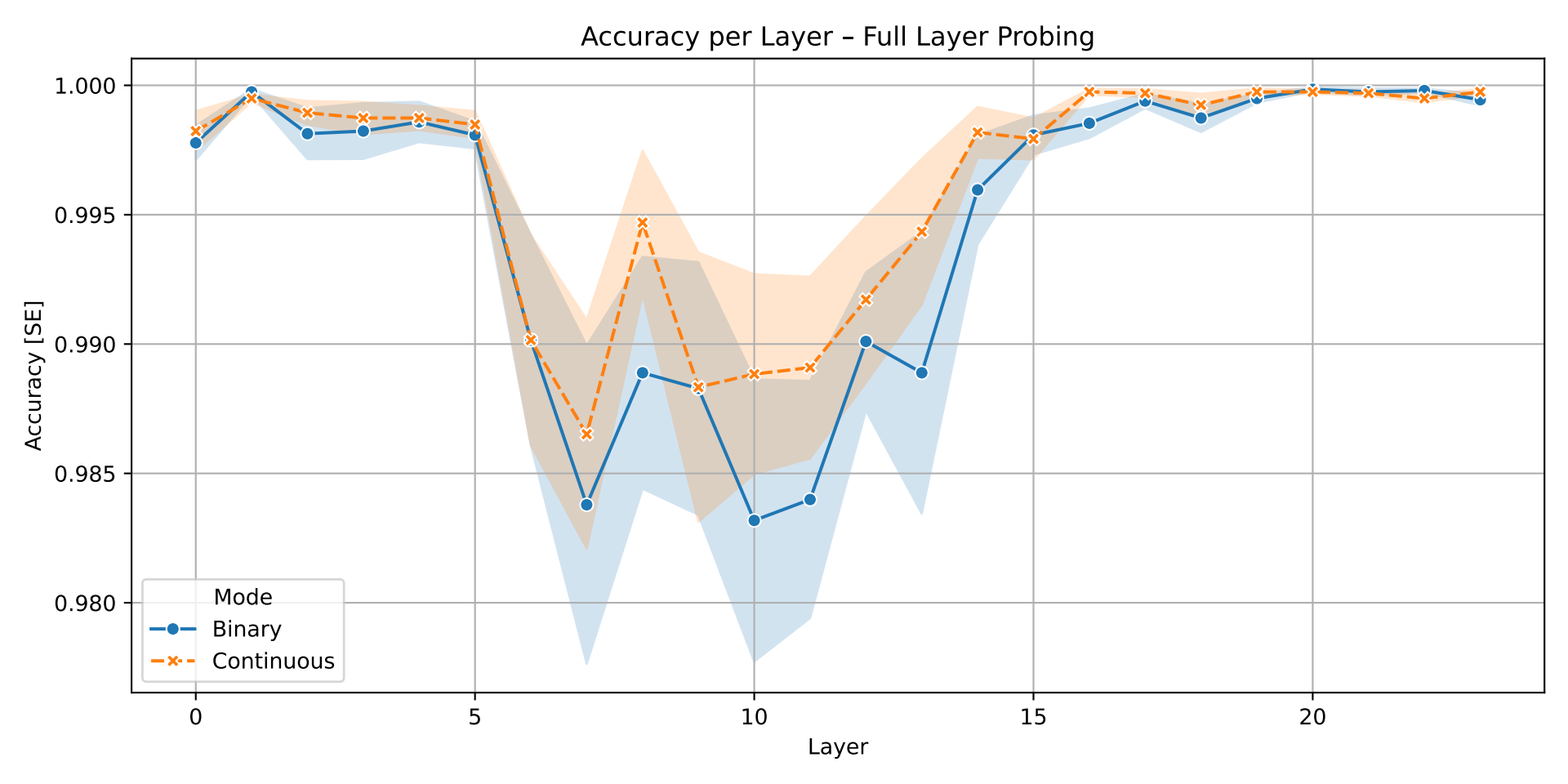

Overall F1 per Layer – All Methods
F1 Score [SE]
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0
5
10
15
20
Layer
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

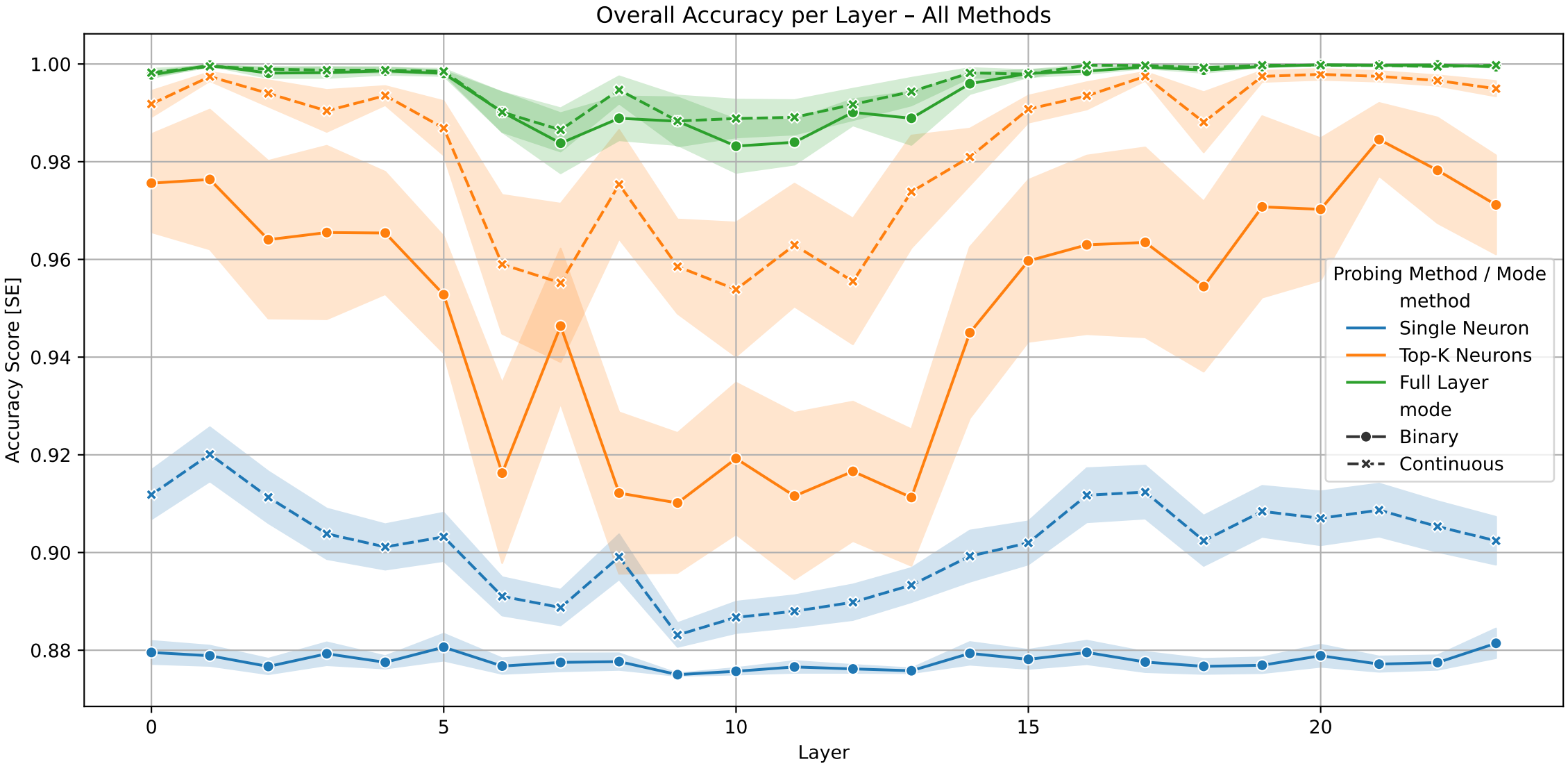
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	20.0	23.0
Full Layer	f1_max	1.0	1.0
Full Layer	f1_mean	0.9948	0.9962
Full Layer	f1_std	0.0096	0.0075
Single Neuron	f1_best_layer	23.0	1.0
Single Neuron	f1_max	1.0	1.0
Single Neuron	f1_mean	0.8225	0.8674
Single Neuron	f1_std	0.0278	0.0657
Top-K Neurons	f1_best_layer	21.0	20.0
Top-K Neurons	f1_max	1.0	1.0
Top-K Neurons	f1_mean	0.9419	0.9824
Top-K Neurons	f1_std	0.061	0.0267


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


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


Accuracy per Layer – Full Layer Probing
Accuracy [SE]
1.000
0.995
0.990
0.985
0.980
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	20.0	16.0
Full Layer	accuracy_max	1.0	1.0
Full Layer	accuracy_mean	0.9949	0.9962
Full Layer	accuracy_std	0.0095	0.0074
Single Neuron	accuracy_best_layer	23.0	1.0
Single Neuron	accuracy_max	1.0	1.0
Single Neuron	accuracy_mean	0.8778	0.9013
Single Neuron	accuracy_std	0.0164	0.0428
Top-K Neurons	accuracy_best_layer	21.0	20.0
Top-K Neurons	accuracy_max	1.0	1.0
Top-K Neurons	accuracy_mean	0.9502	0.9826
Top-K Neurons	accuracy_std	0.0477	0.0263