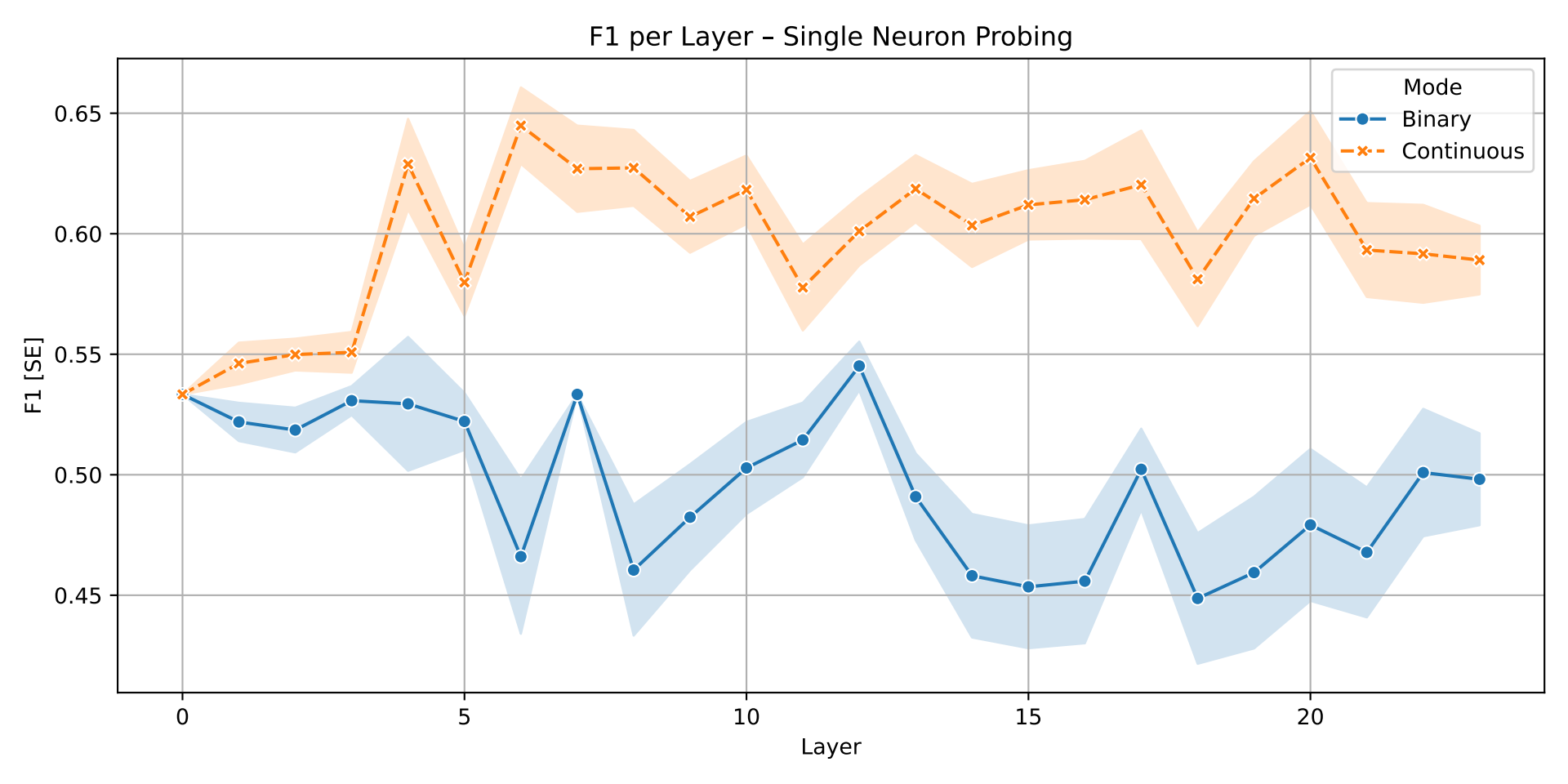
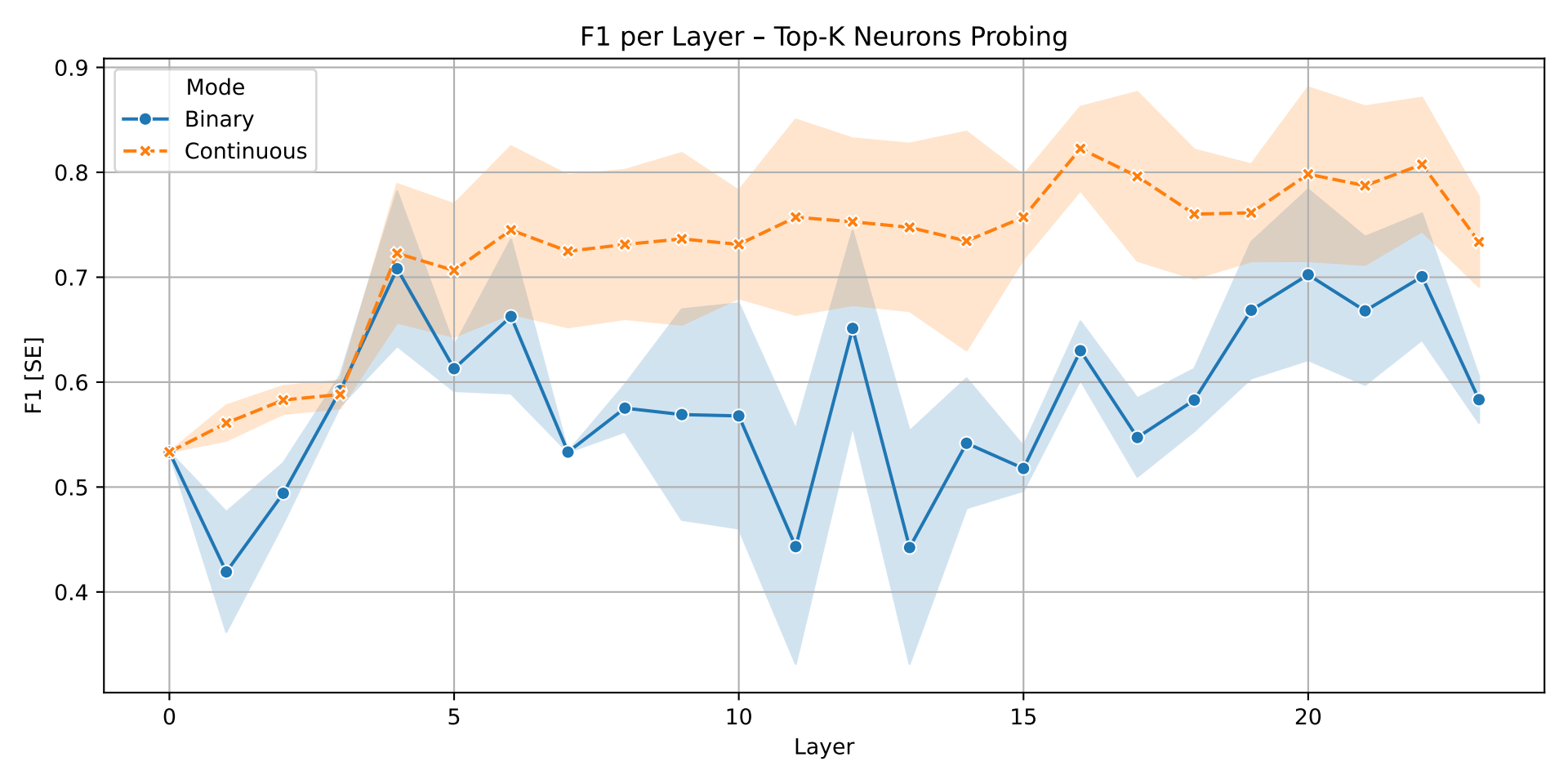
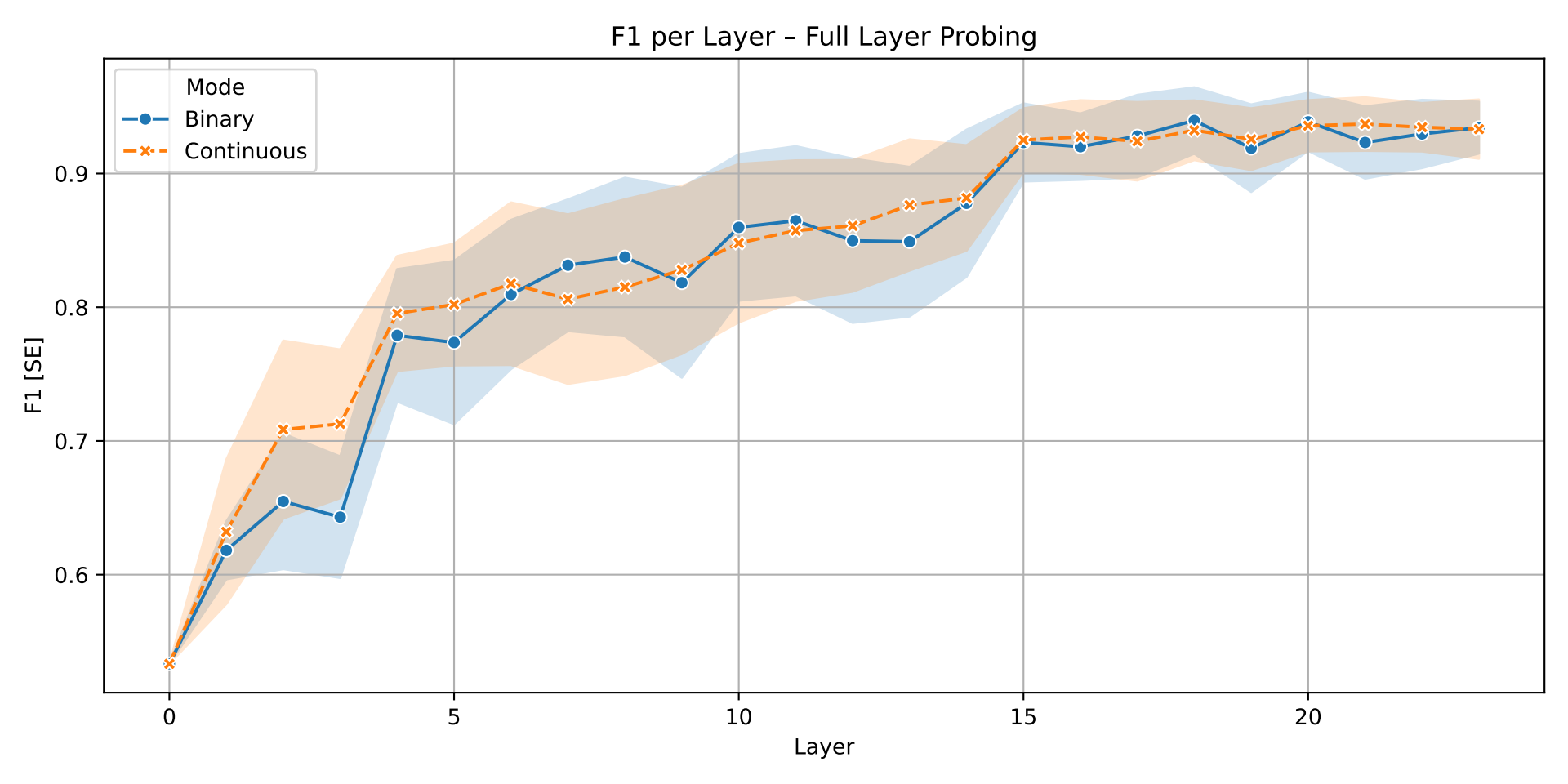
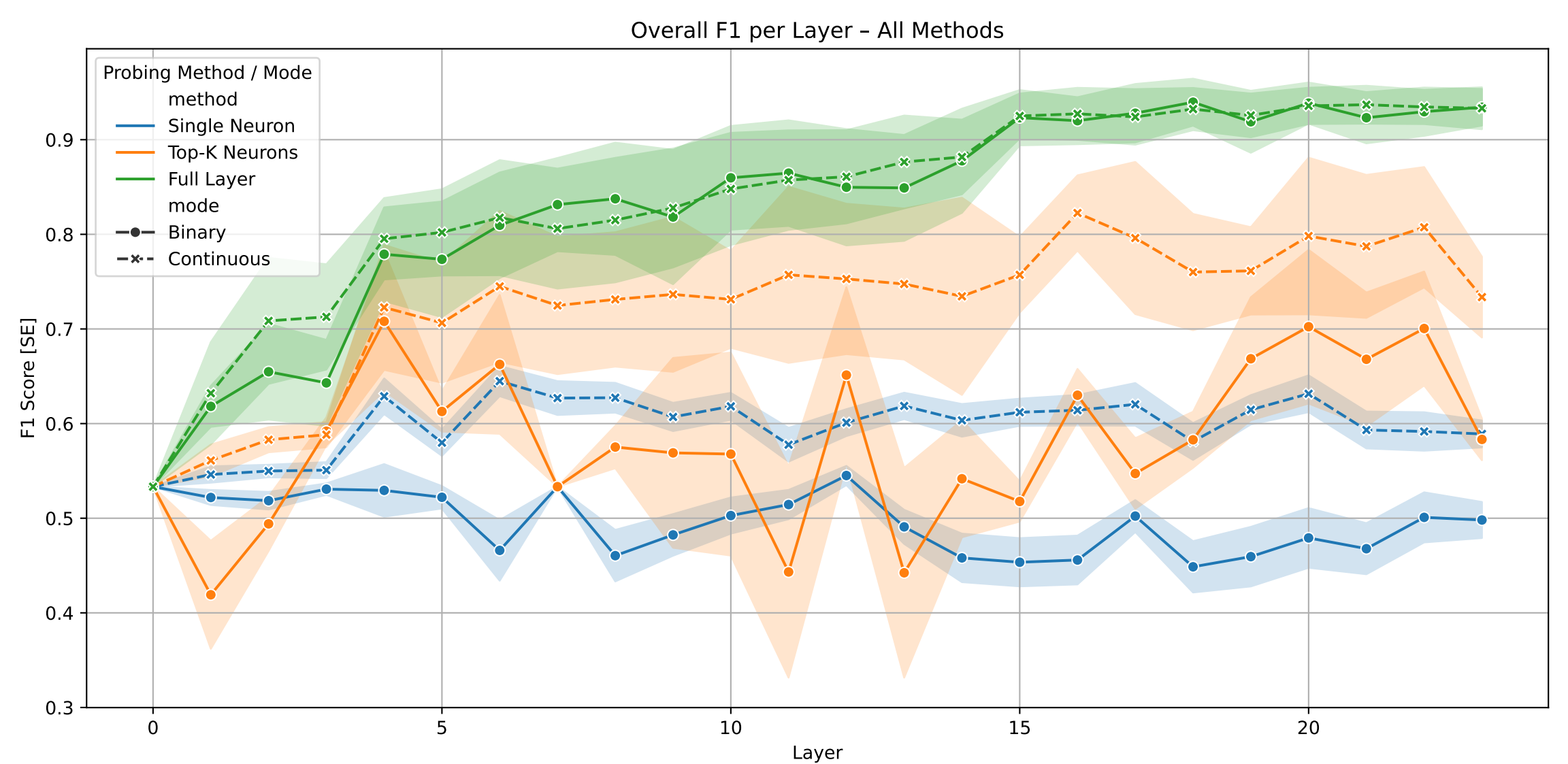

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.65
0.60
0.55
0.50
0.45
0
5
10
15
20
Layer

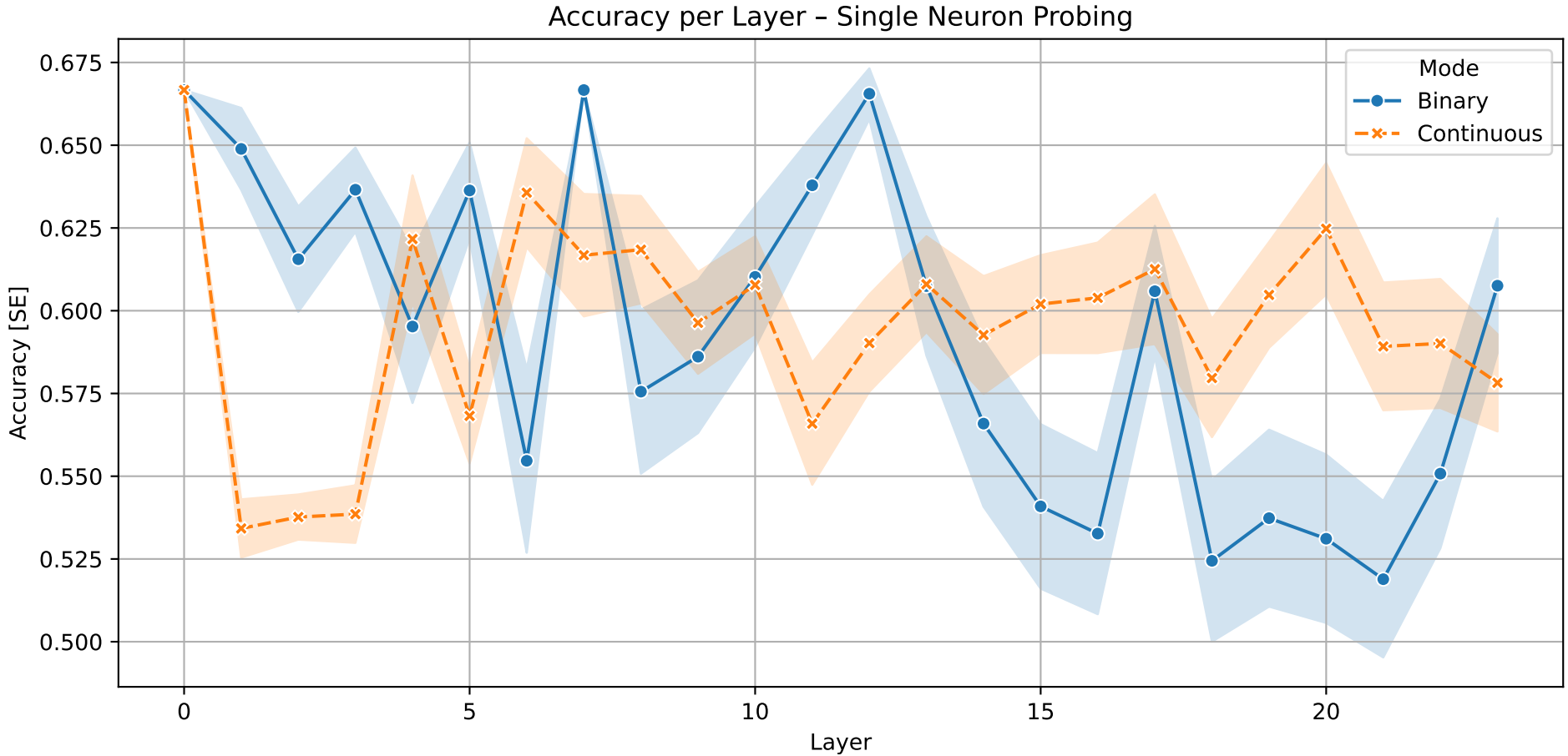

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
Layer
0.9
0.8
0.7
0.6
0.5
0.4
0
5
10
15
20

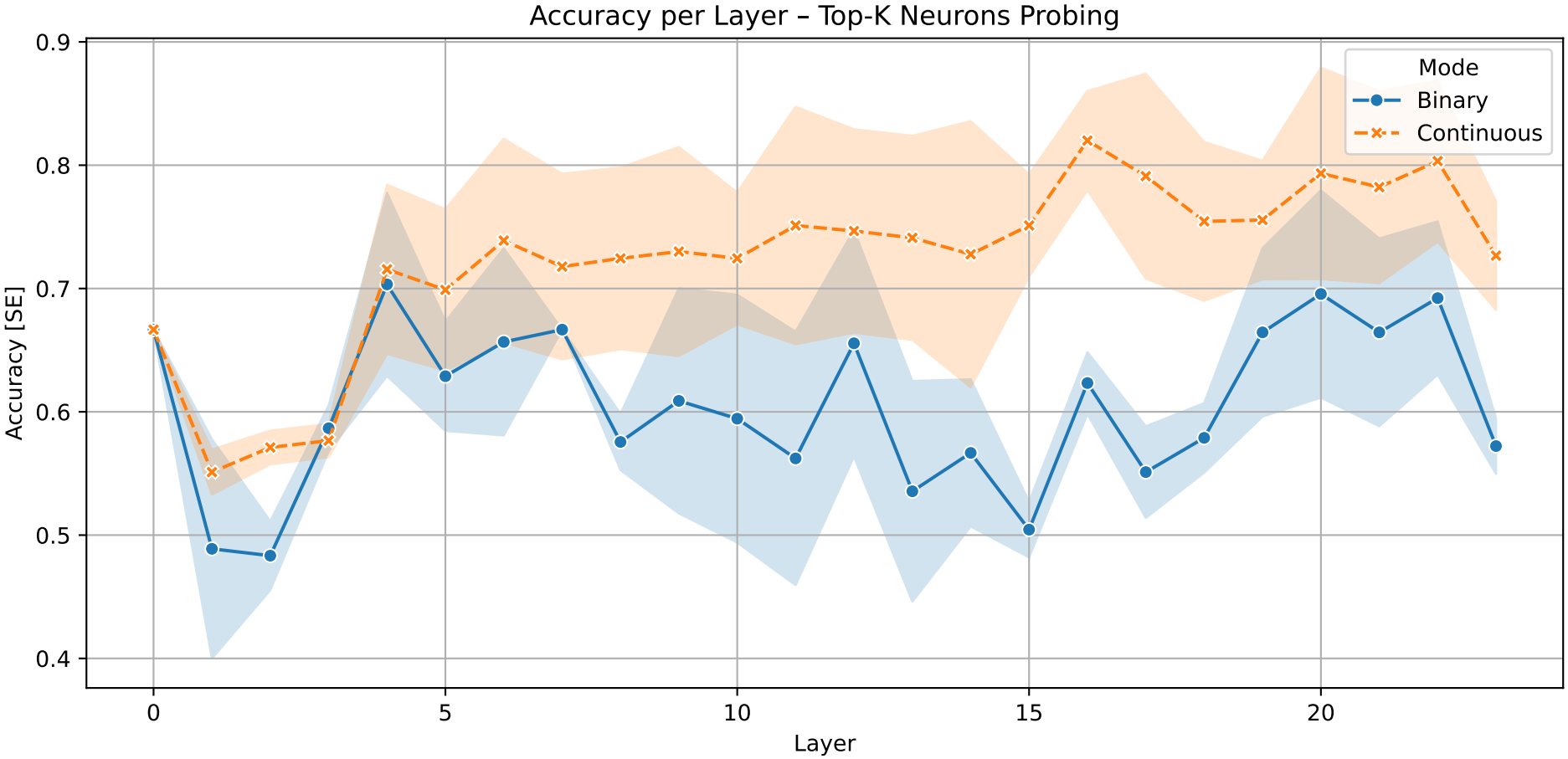

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.6
0.7
0.8
0.9
0
5
10
15
20
Layer

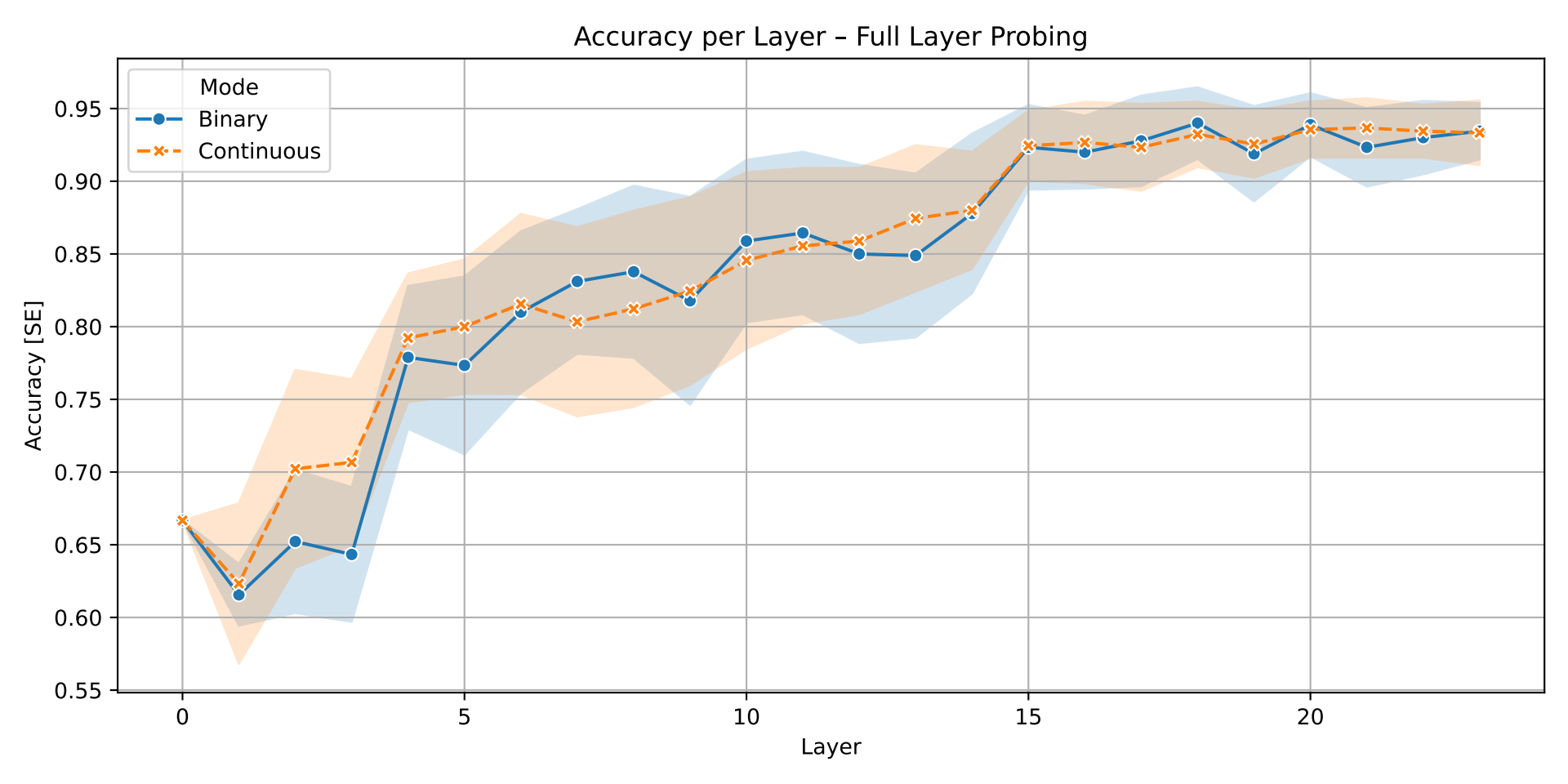

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
5
10
15
20
Layer

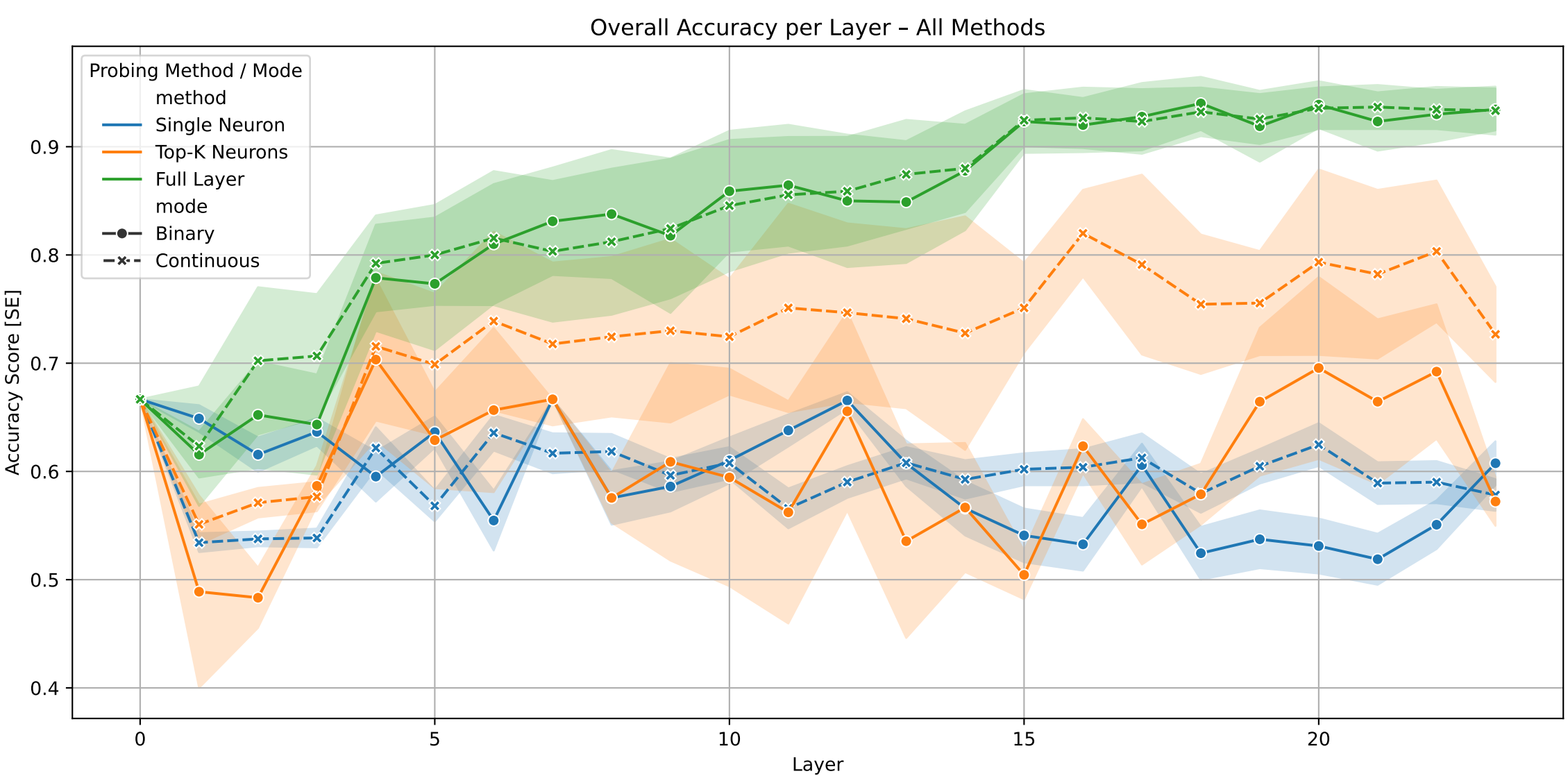
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	18.0	21.0
Full Layer	f1_max	0.9799	0.9765
Full Layer	f1_mean	0.8315	0.8396
Full Layer	f1_std	0.1289	0.1216
Single Neuron	f1_best_layer	12.0	6.0
Single Neuron	f1_max	0.831	0.9131
Single Neuron	f1_mean	0.4948	0.5984
Single Neuron	f1_std	0.1199	0.0898
Top-K Neurons	f1_best_layer	4.0	16.0
Top-K Neurons	f1_max	0.831	0.9297
Top-K Neurons	f1_mean	0.5811	0.7242
Top-K Neurons	f1_std	0.1226	0.1203


Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
Layer
Mode
Binary
Continuous
0.675
0.650
0.625
0.600
0.575
0.550
0.525
0.500
0
5
10
15
20


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
Layer
0.9
0.8
0.7
0.6
0.5
0.4
0
5
10
15
20


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
Layer


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

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	18.0	21.0
Full Layer	accuracy_max	0.98	0.9767
Full Layer	accuracy_mean	0.8368	0.8431
Full Layer	accuracy_std	0.1186	0.1121
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.8333	0.9133
Single Neuron	accuracy_mean	0.5924	0.5952
Single Neuron	accuracy_std	0.1202	0.0914
Top-K Neurons	accuracy_best_layer	4.0	16.0
Top-K Neurons	accuracy_max	0.8333	0.93
Top-K Neurons	accuracy_mean	0.6053	0.7233
Top-K Neurons	accuracy_std	0.1112	0.1179