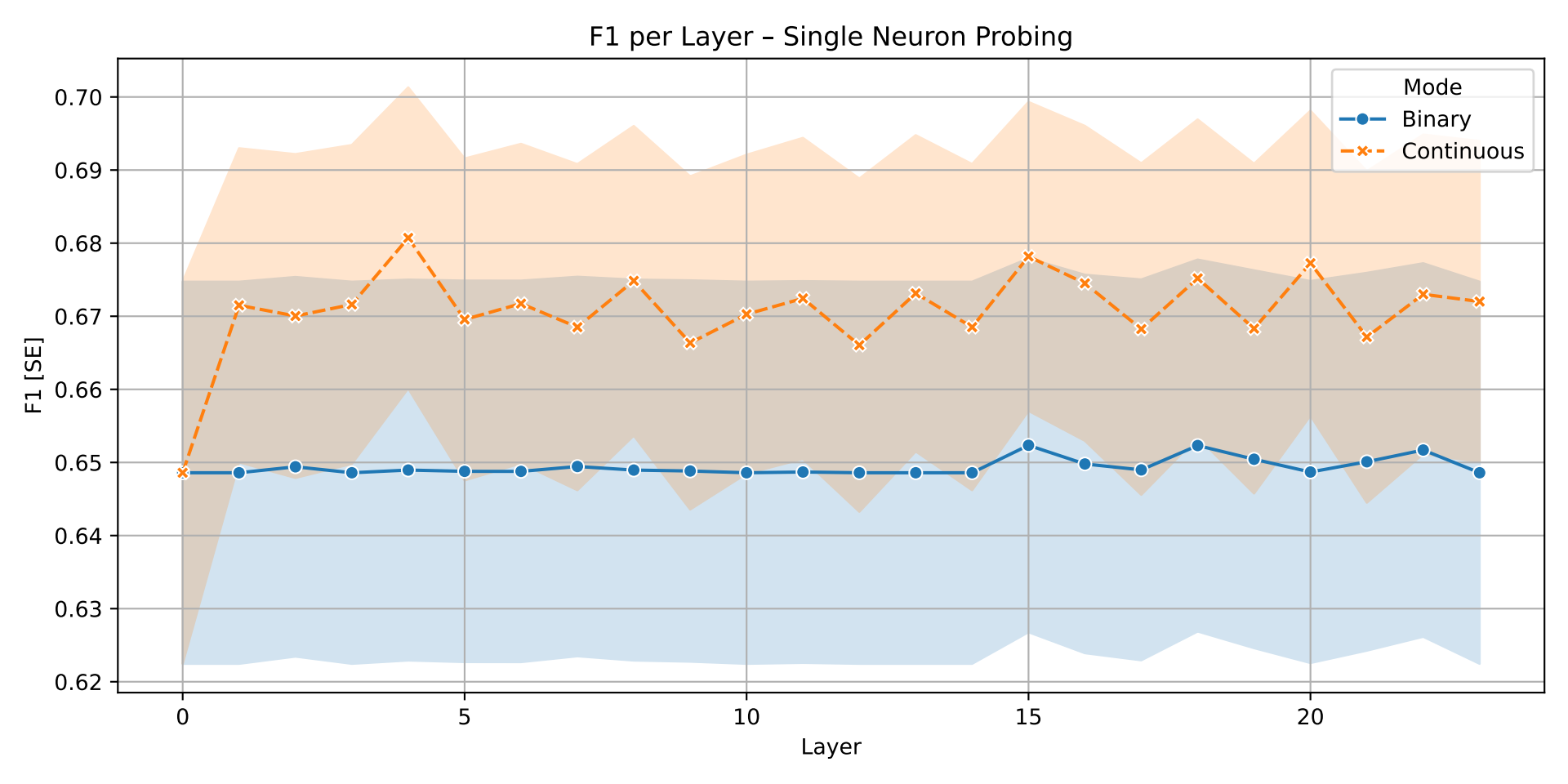
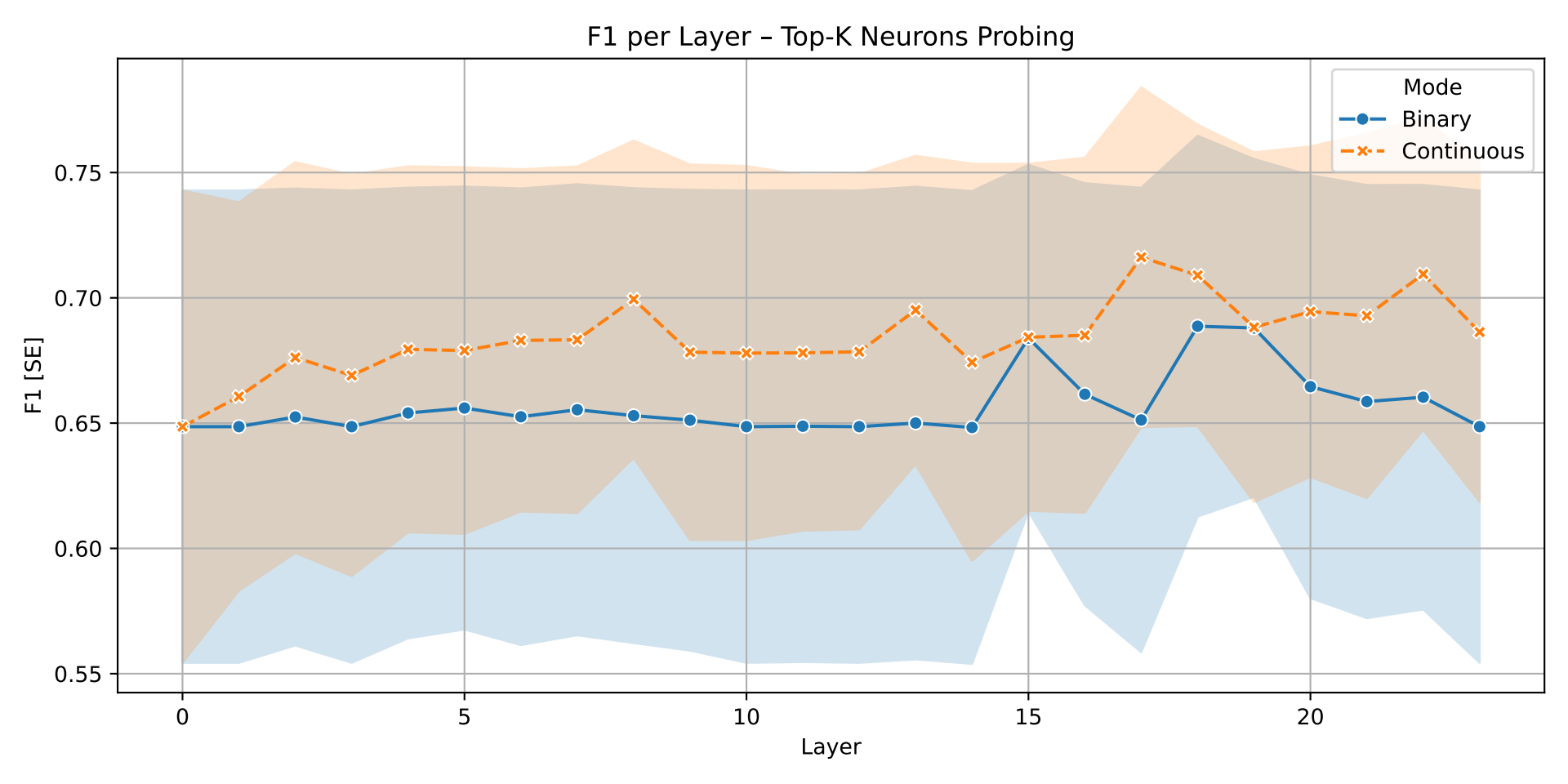
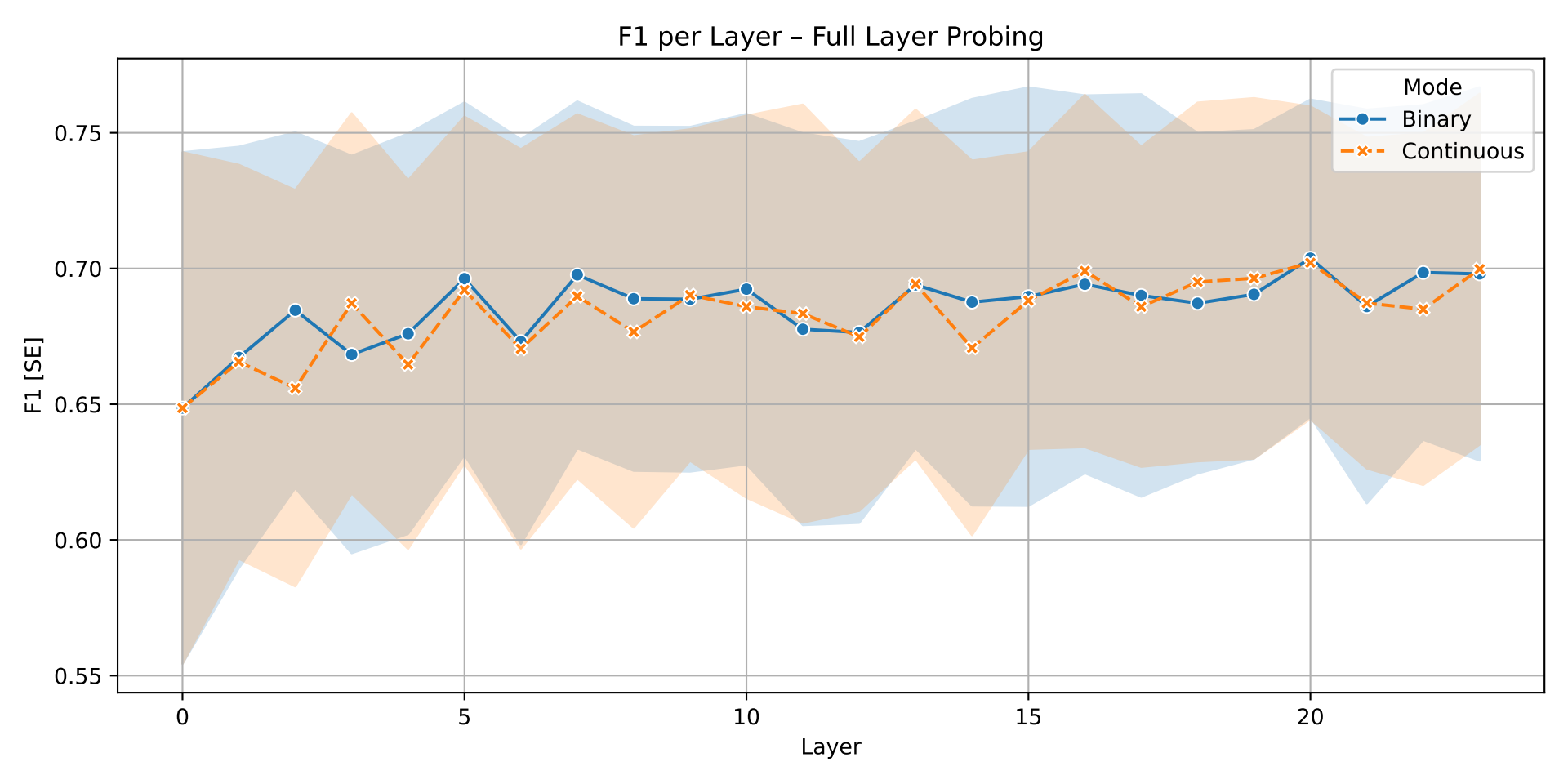
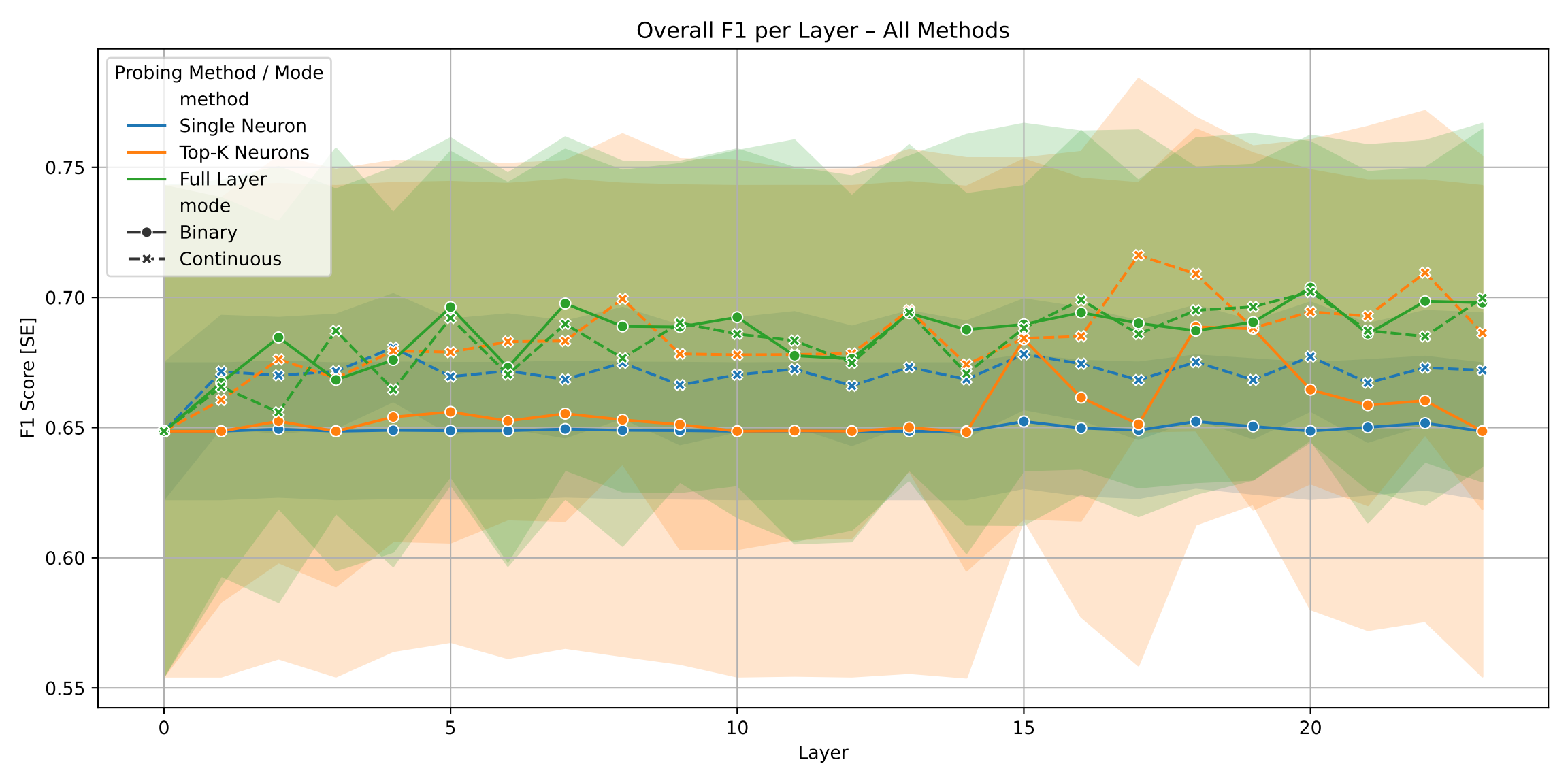

F1 per Layer – Single Neuron Probing
F1 [SE]
Layer
Mode
Binary
Continuous
0.70
0.69
0.68
0.67
0.66
0.65
0.64
0.63
0.62
0
5
10
15
20

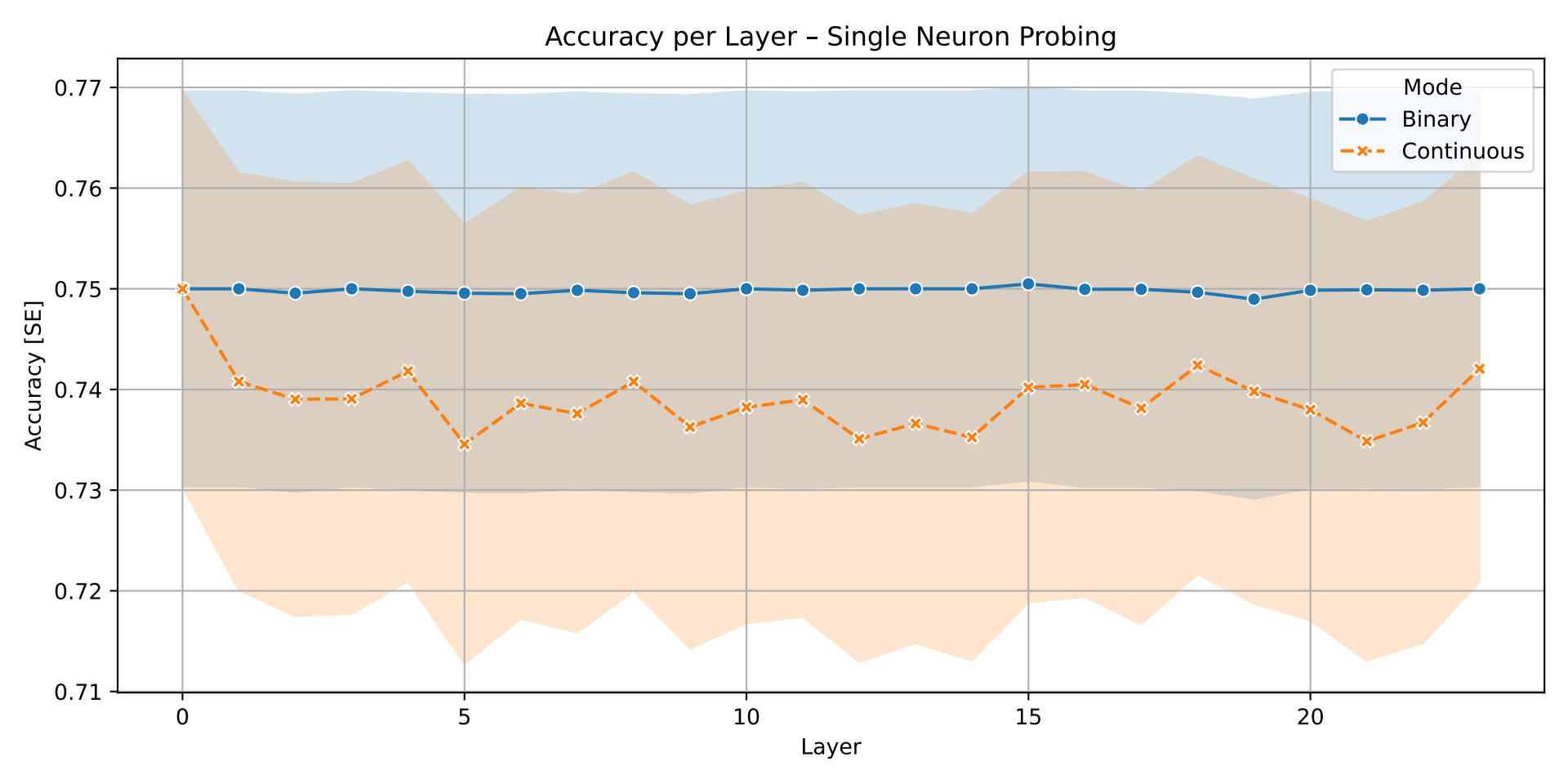

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
Layer

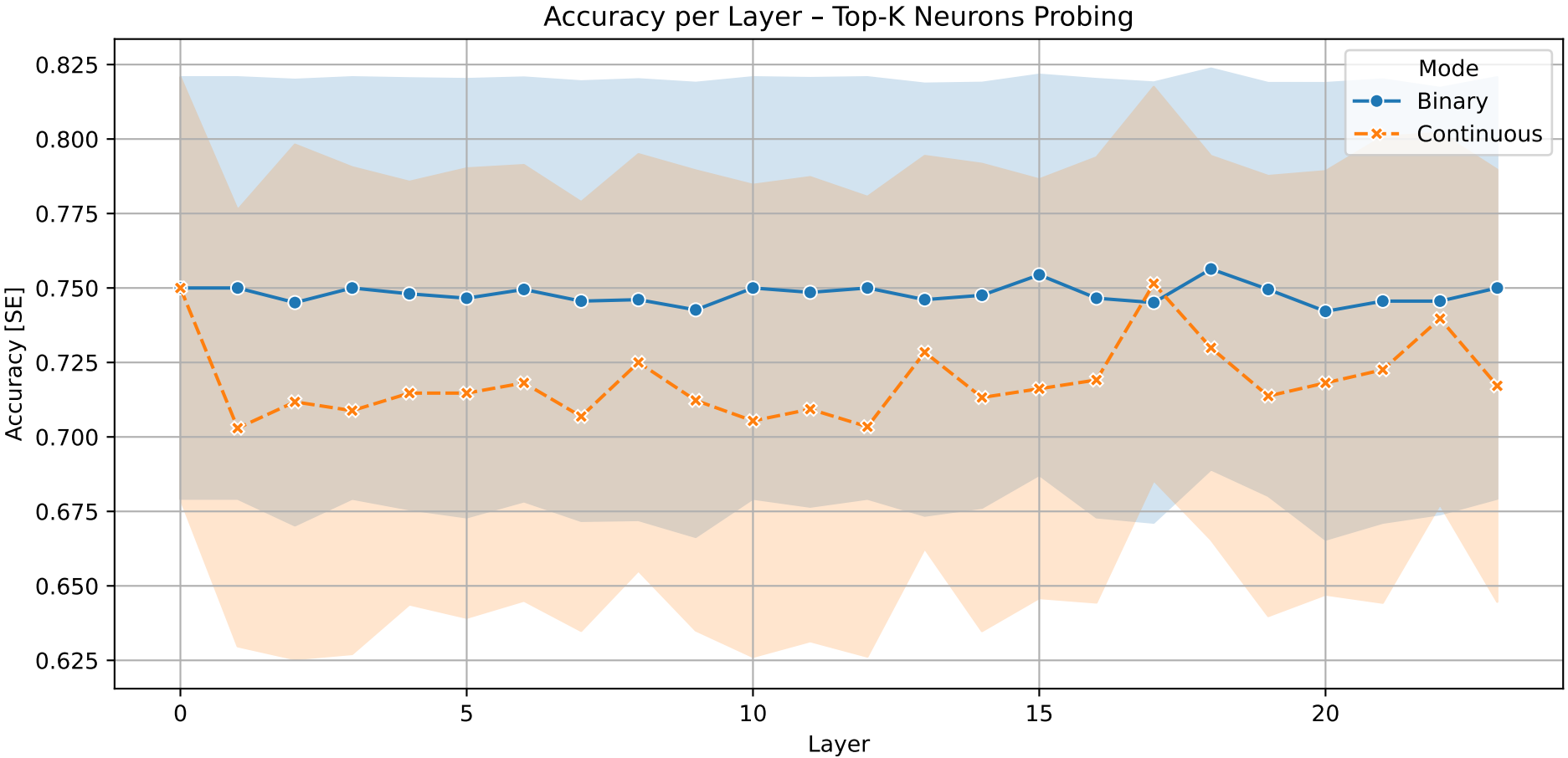

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
Layer

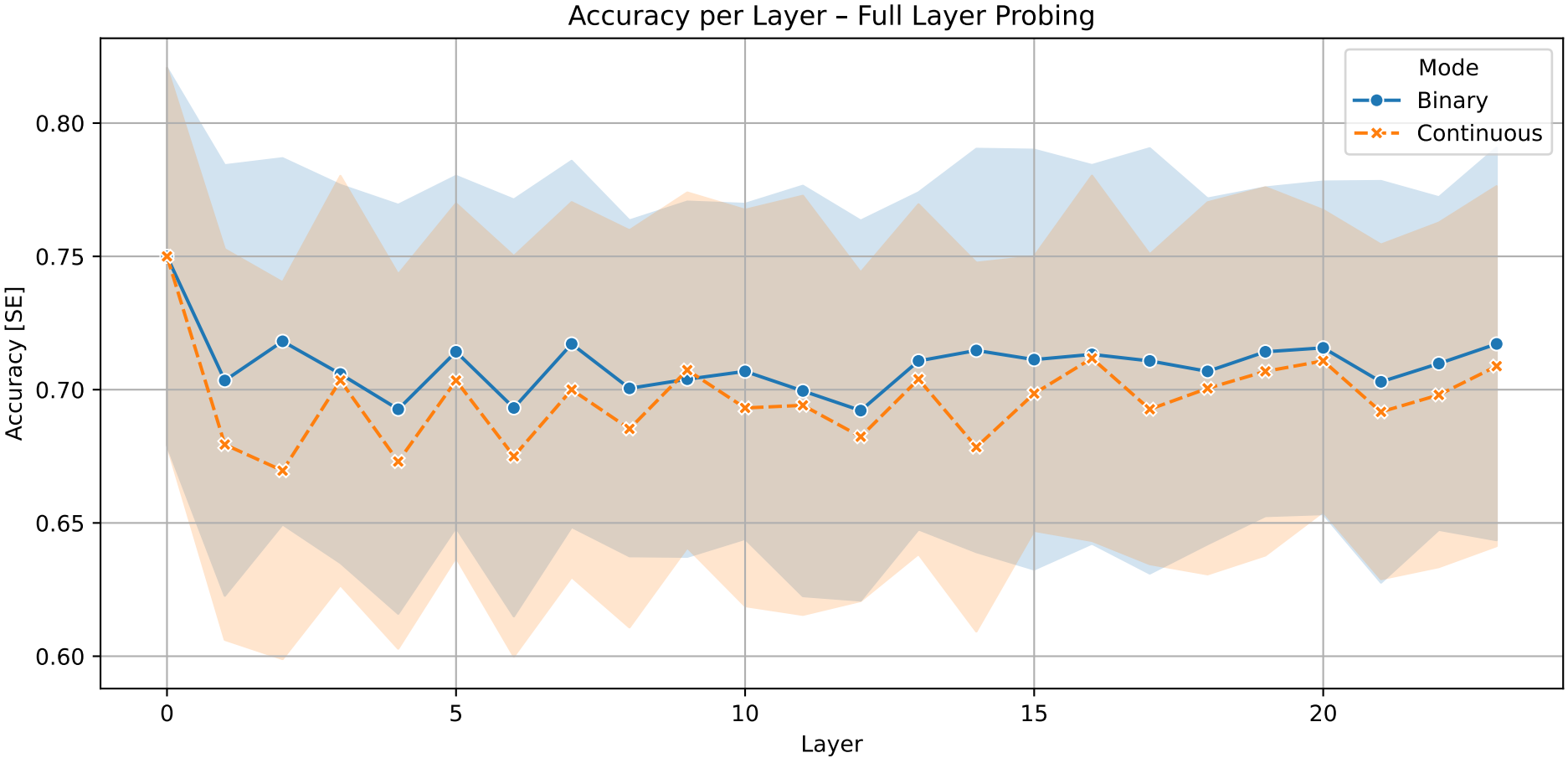

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
Layer

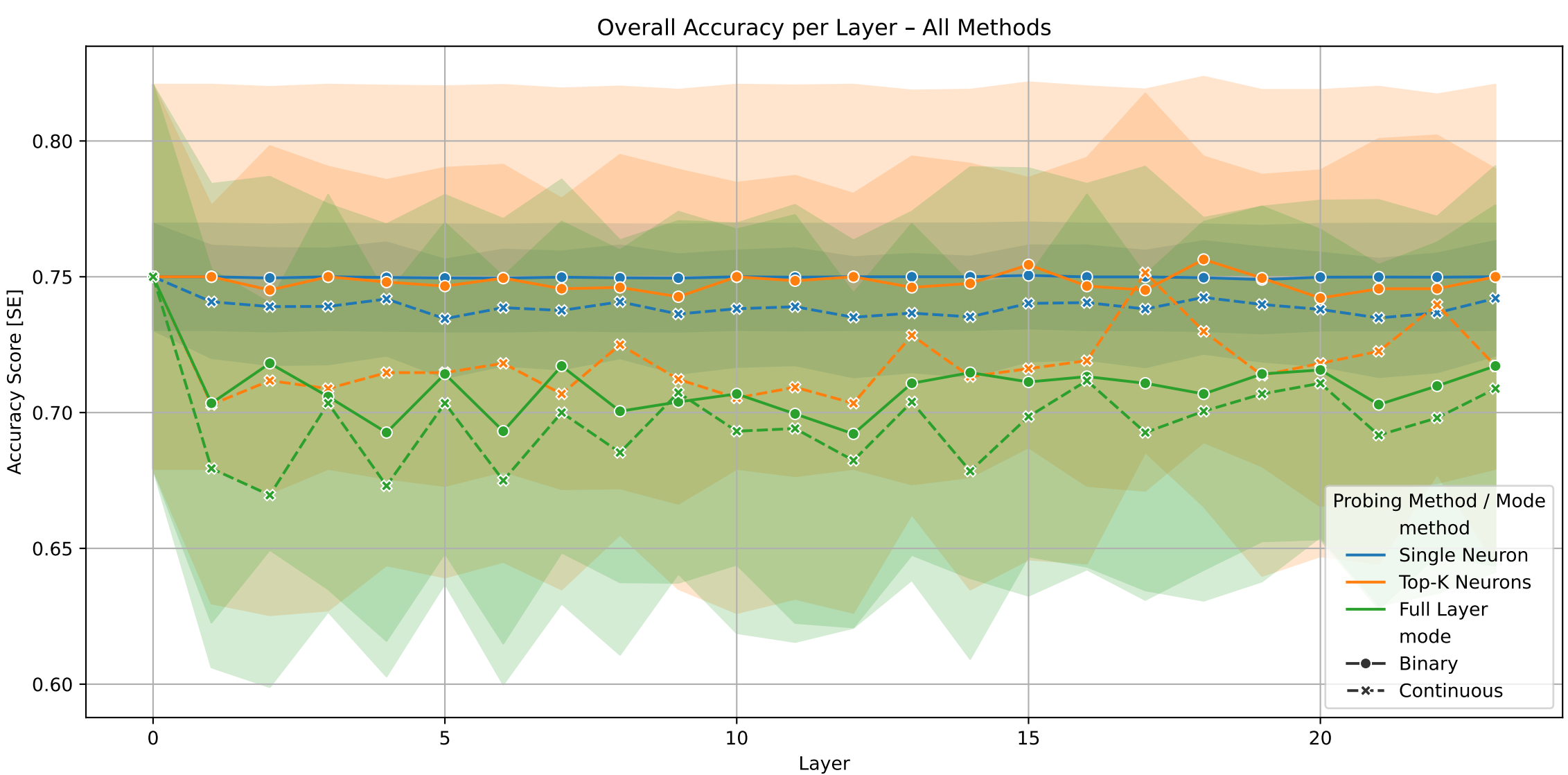
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	20.0	20.0
Full Layer	f1_max	0.8597	0.8586
Full Layer	f1_mean	0.6856	0.6829
Full Layer	f1_std	0.1219	0.119
Single Neuron	f1_best_layer	15.0	4.0
Single Neuron	f1_max	0.8526	0.8553
Single Neuron	f1_mean	0.6494	0.6707
Single Neuron	f1_std	0.1624	0.1382
Top-K Neurons	f1_best_layer	18.0	17.0
Top-K Neurons	f1_max	0.8526	0.8582
Top-K Neurons	f1_mean	0.6571	0.6844
Top-K Neurons	f1_std	0.1553	0.1259


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
Layer
0.77
0.76
0.75
0.74
0.73
0.72
0.71
0
5
10
15
20


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
Layer
Mode
Binary
Continuous
0.825
0.800
0.775
0.750
0.725
0.700
0.675
0.650
0.625
0
5
10
15
20


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


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
0.80
0.75
0.70
0.65
0.60
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	0.0	0.0
Full Layer	accuracy_max	0.9	0.9
Full Layer	accuracy_mean	0.7094	0.6966
Full Layer	accuracy_std	0.1236	0.1201
Single Neuron	accuracy_best_layer	15.0	0.0
Single Neuron	accuracy_max	0.9	0.9
Single Neuron	accuracy_mean	0.7498	0.739
Single Neuron	accuracy_std	0.1228	0.1334
Top-K Neurons	accuracy_best_layer	18.0	17.0
Top-K Neurons	accuracy_max	0.9	0.9
Top-K Neurons	accuracy_mean	0.748	0.7189
Top-K Neurons	accuracy_std	0.1256	0.1289