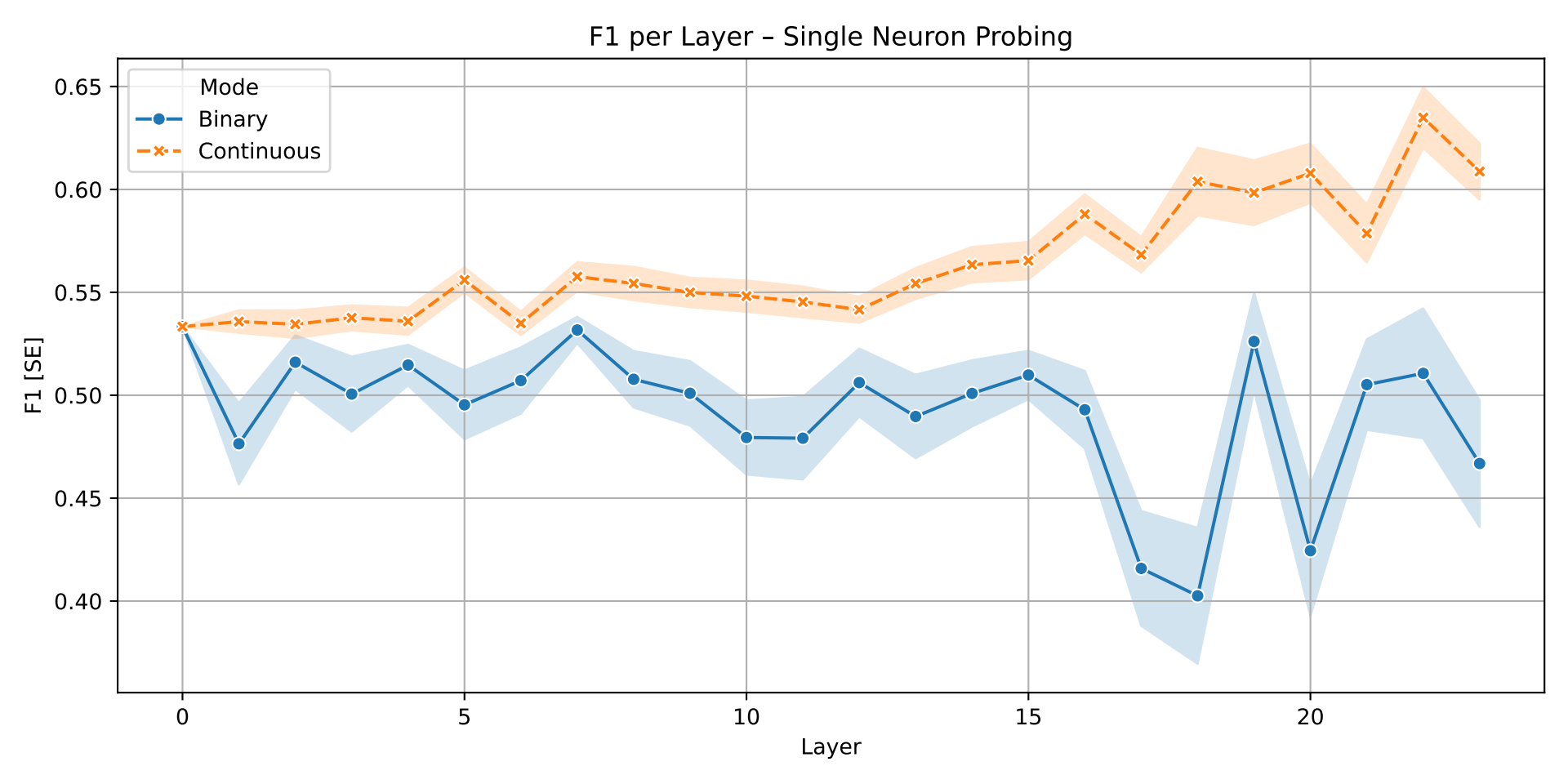
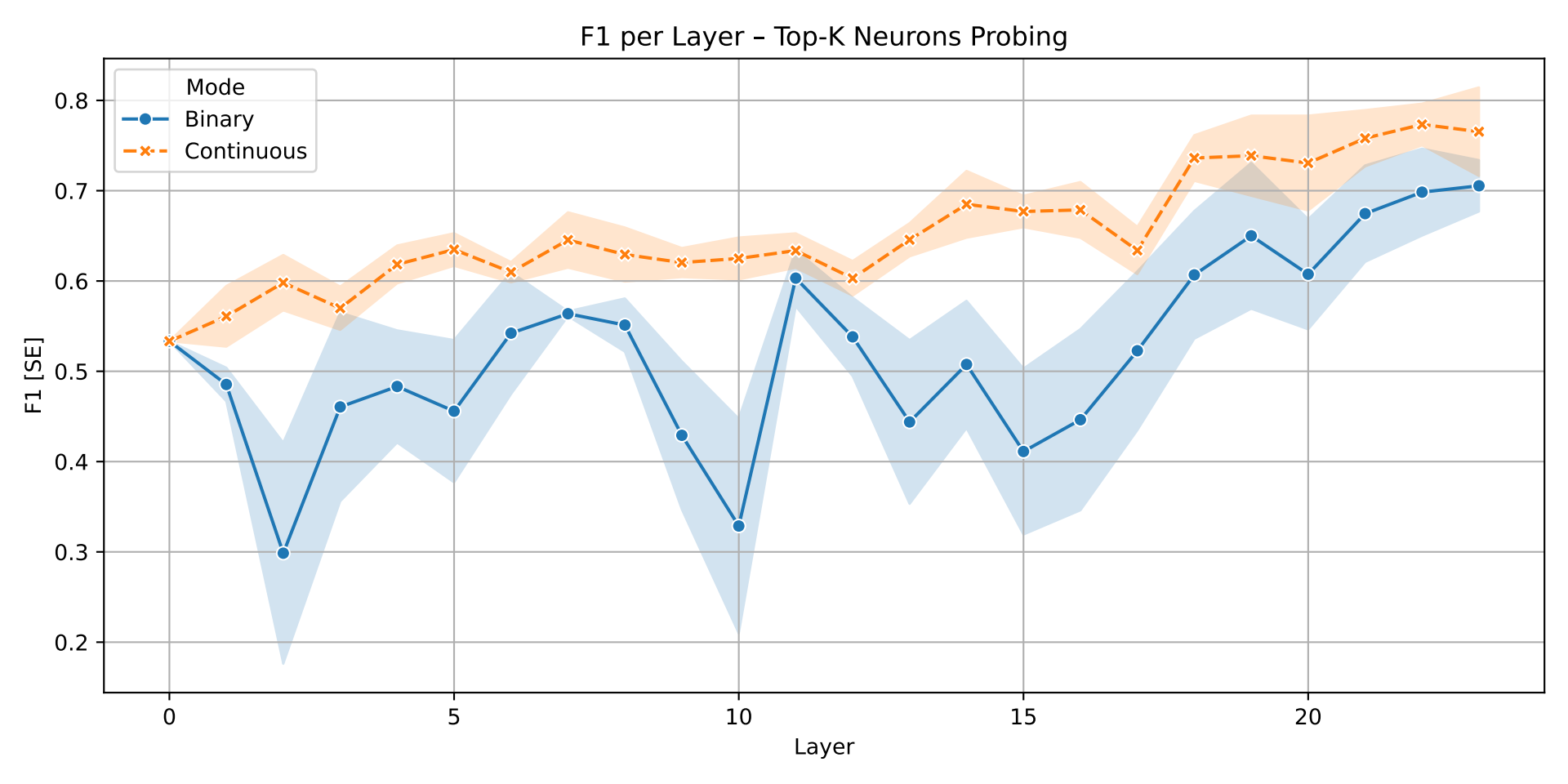
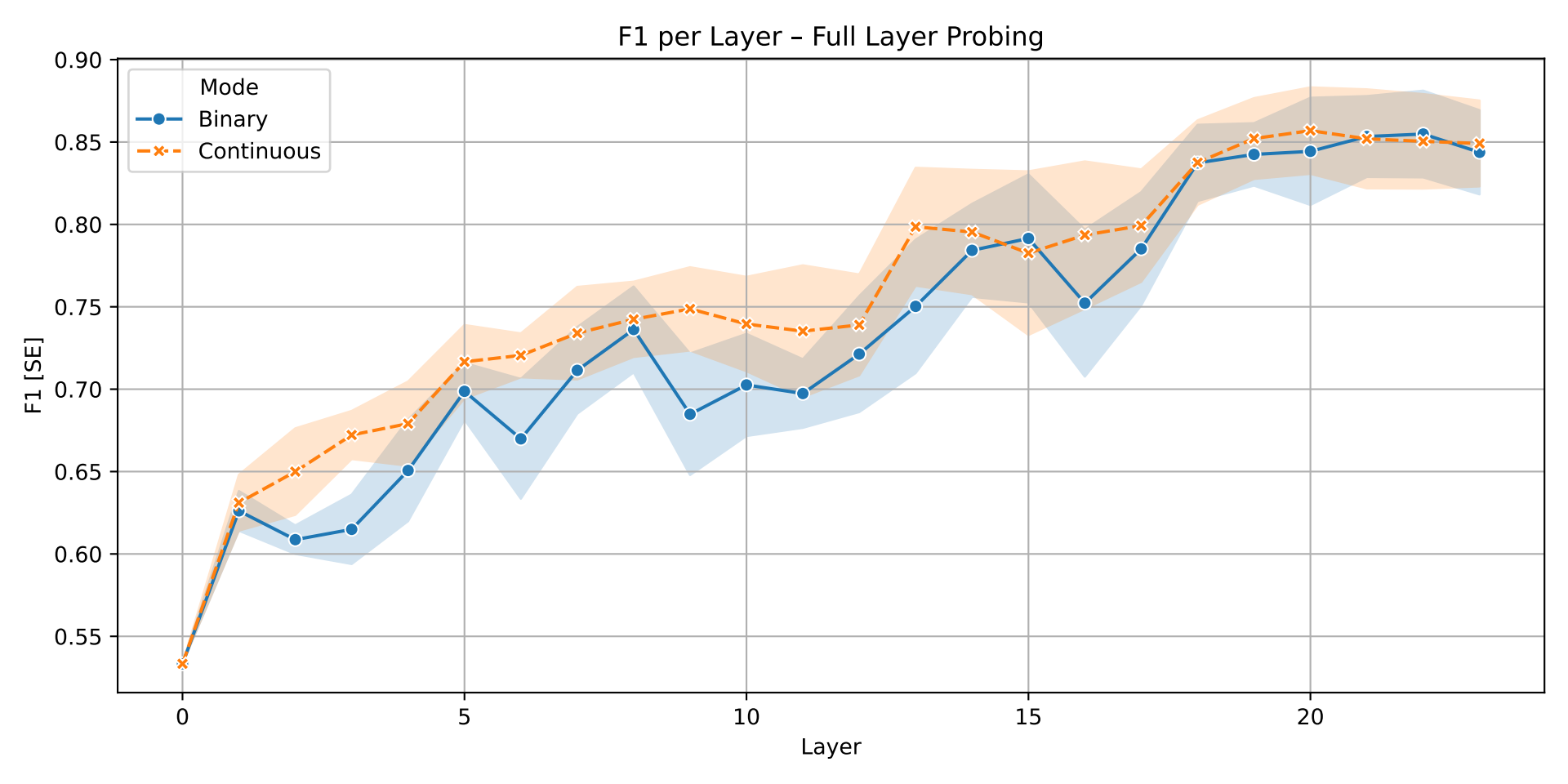
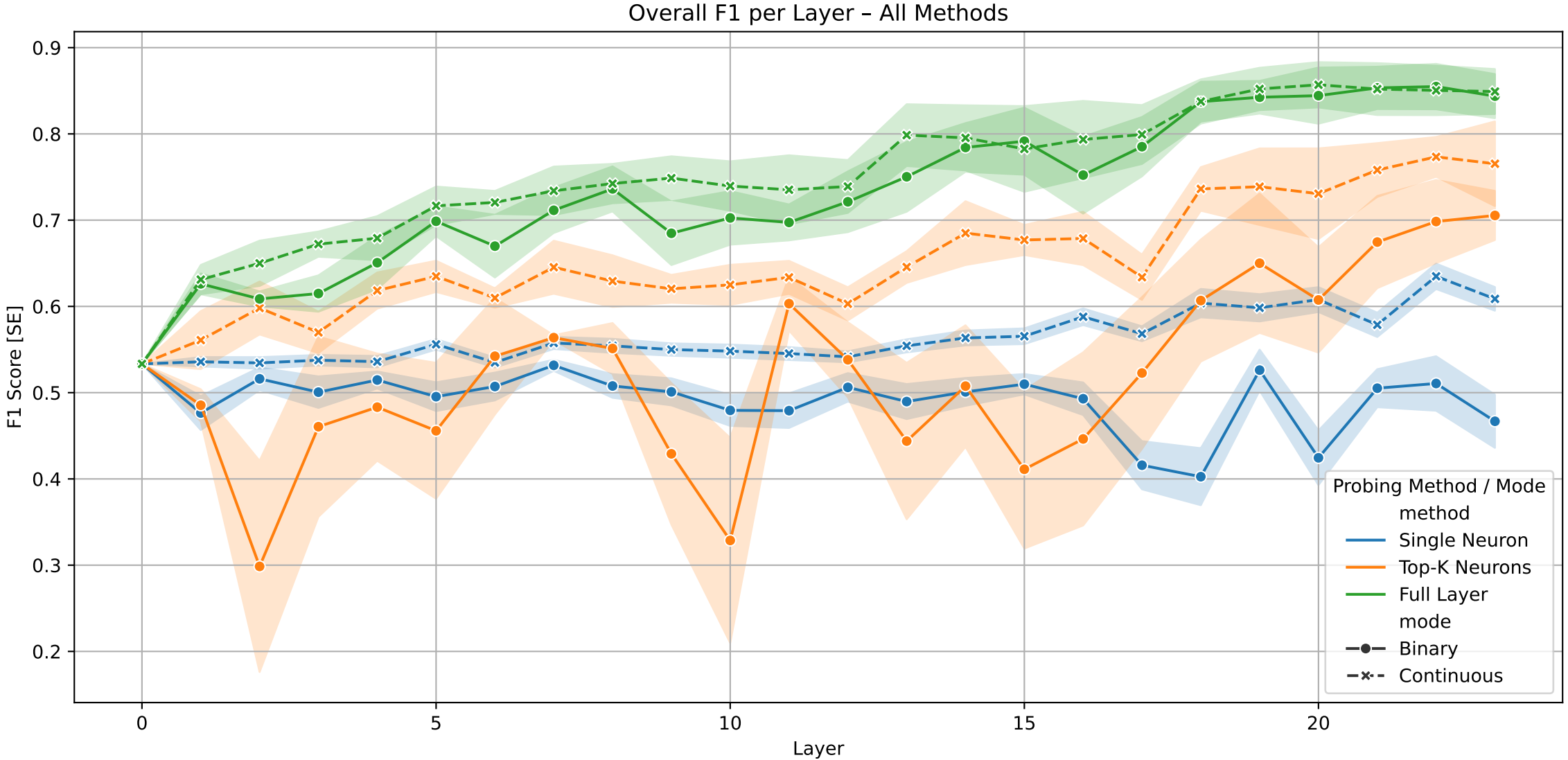

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.65
0.60
0.55
0.50
0.45
0.40
0
5
10
15
20
Layer

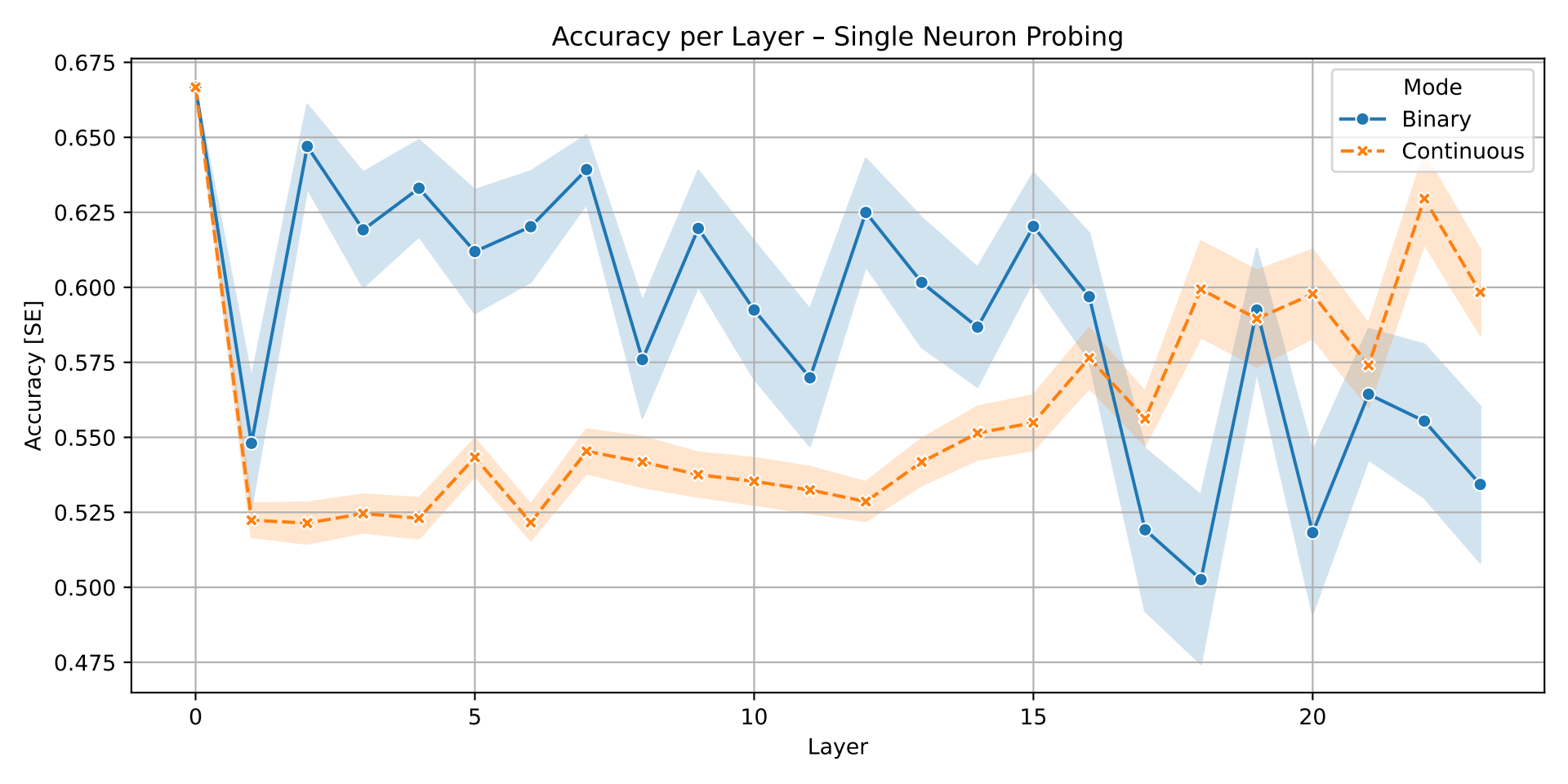

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
Layer
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
5
10
15
20

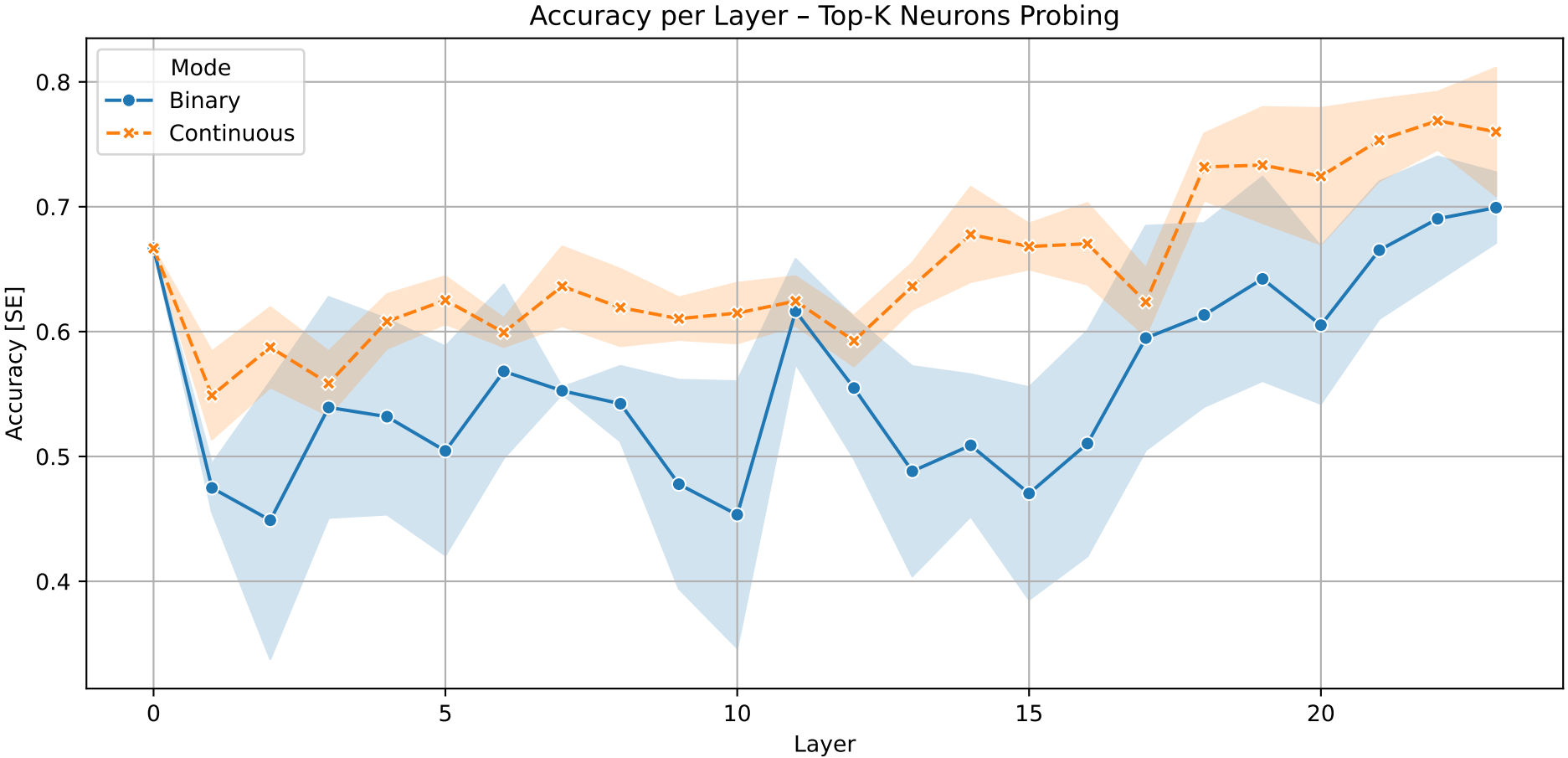

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
Layer
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20

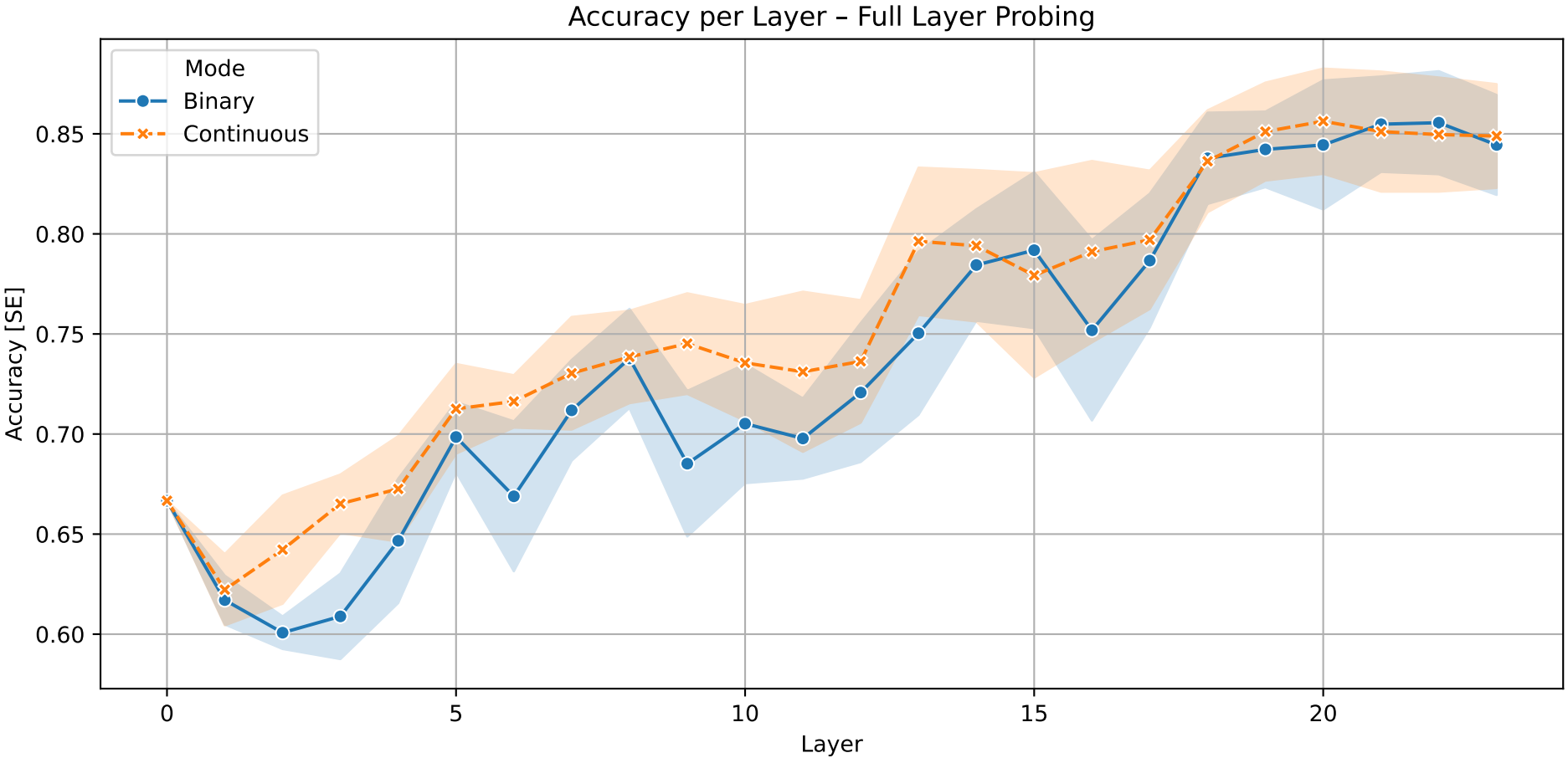

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
5
10
15
20
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

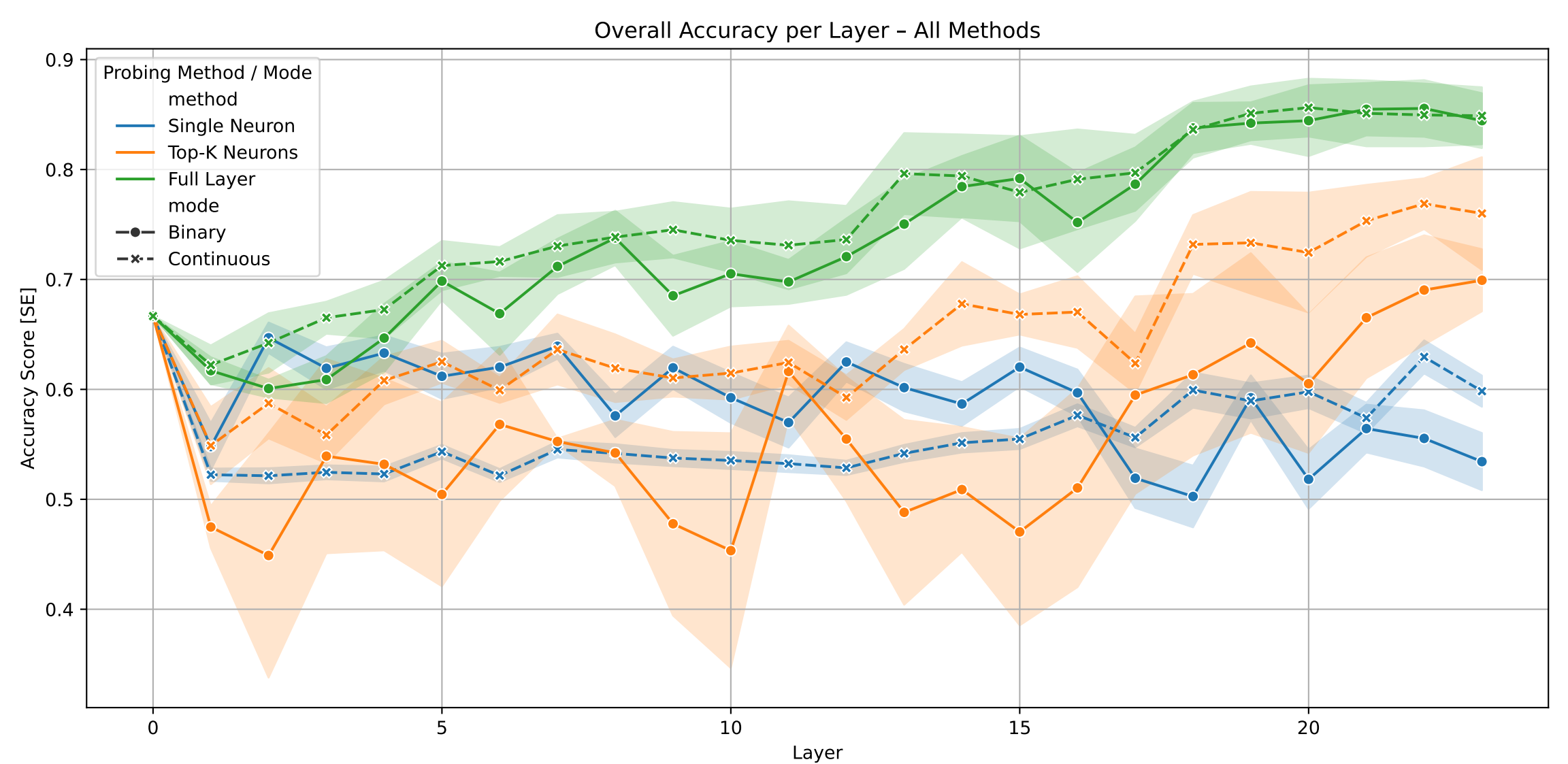
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	22.0	20.0
Full Layer	f1_max	0.9094	0.912
Full Layer	f1_mean	0.7331	0.7545
Full Layer	f1_std	0.0971	0.0908
Single Neuron	f1_best_layer	0.0	22.0
Single Neuron	f1_max	0.7915	0.8589
Single Neuron	f1_mean	0.4914	0.564
Single Neuron	f1_std	0.1152	0.0594
Top-K Neurons	f1_best_layer	23.0	22.0
Top-K Neurons	f1_max	0.7869	0.8633
Top-K Neurons	f1_mean	0.5228	0.6544
Top-K Neurons	f1_std	0.1472	0.0775


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
0.475
0
5
10
15
20


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


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.85
0.80
0.75
0.70
0.65
0.60
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]
Layer
0.9
0.8
0.7
0.6
0.5
0.4
0
5
10
15
20

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	22.0	20.0
Full Layer	accuracy_max	0.9089	0.9111
Full Layer	accuracy_mean	0.7379	0.7569
Full Layer	accuracy_std	0.0905	0.0823
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.7889	0.8578
Single Neuron	accuracy_mean	0.59	0.5589
Single Neuron	accuracy_std	0.1199	0.0648
Top-K Neurons	accuracy_best_layer	23.0	22.0
Top-K Neurons	accuracy_max	0.7822	0.8622
Top-K Neurons	accuracy_mean	0.5591	0.6517
Top-K Neurons	accuracy_std	0.1256	0.0761