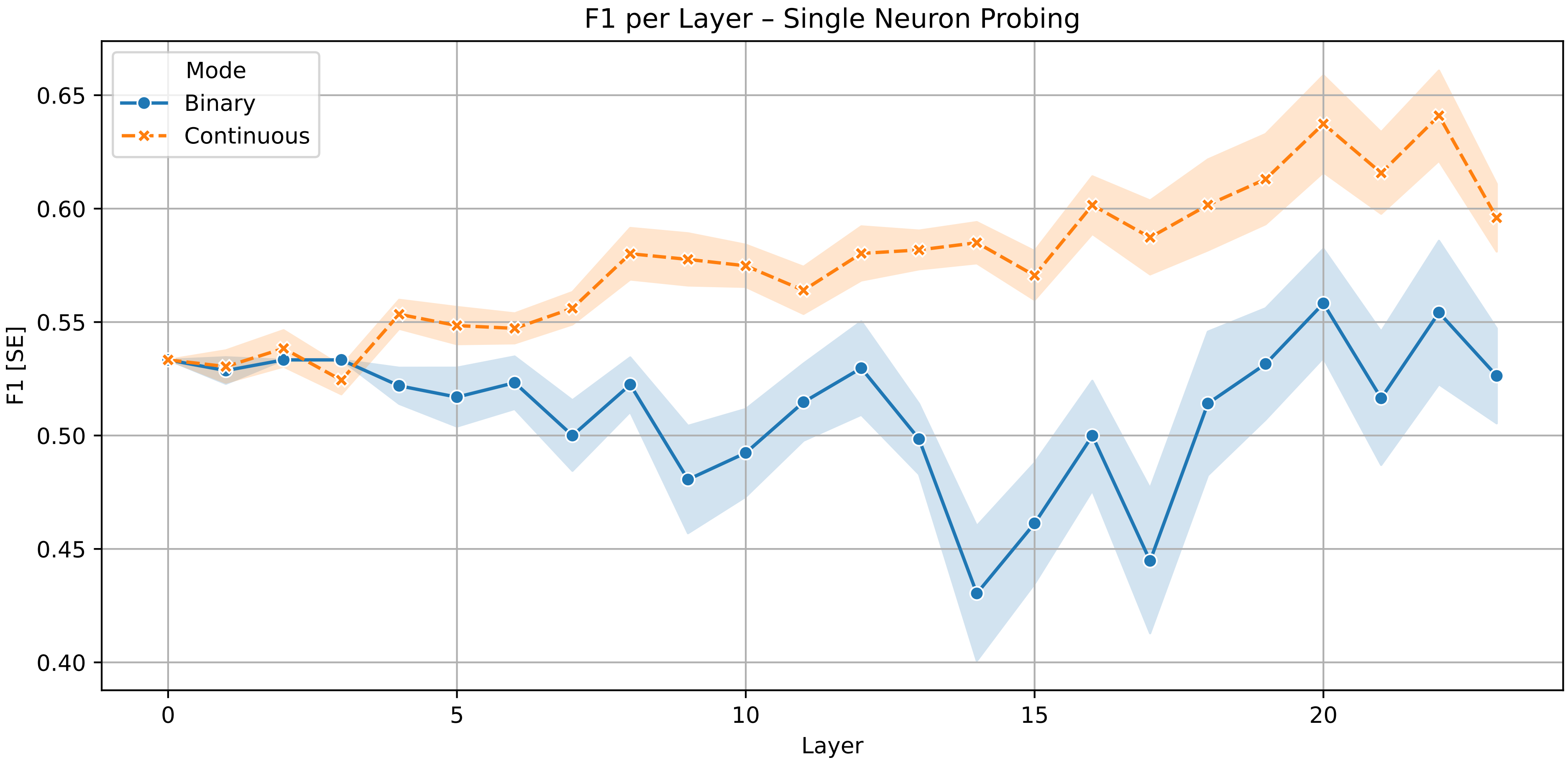
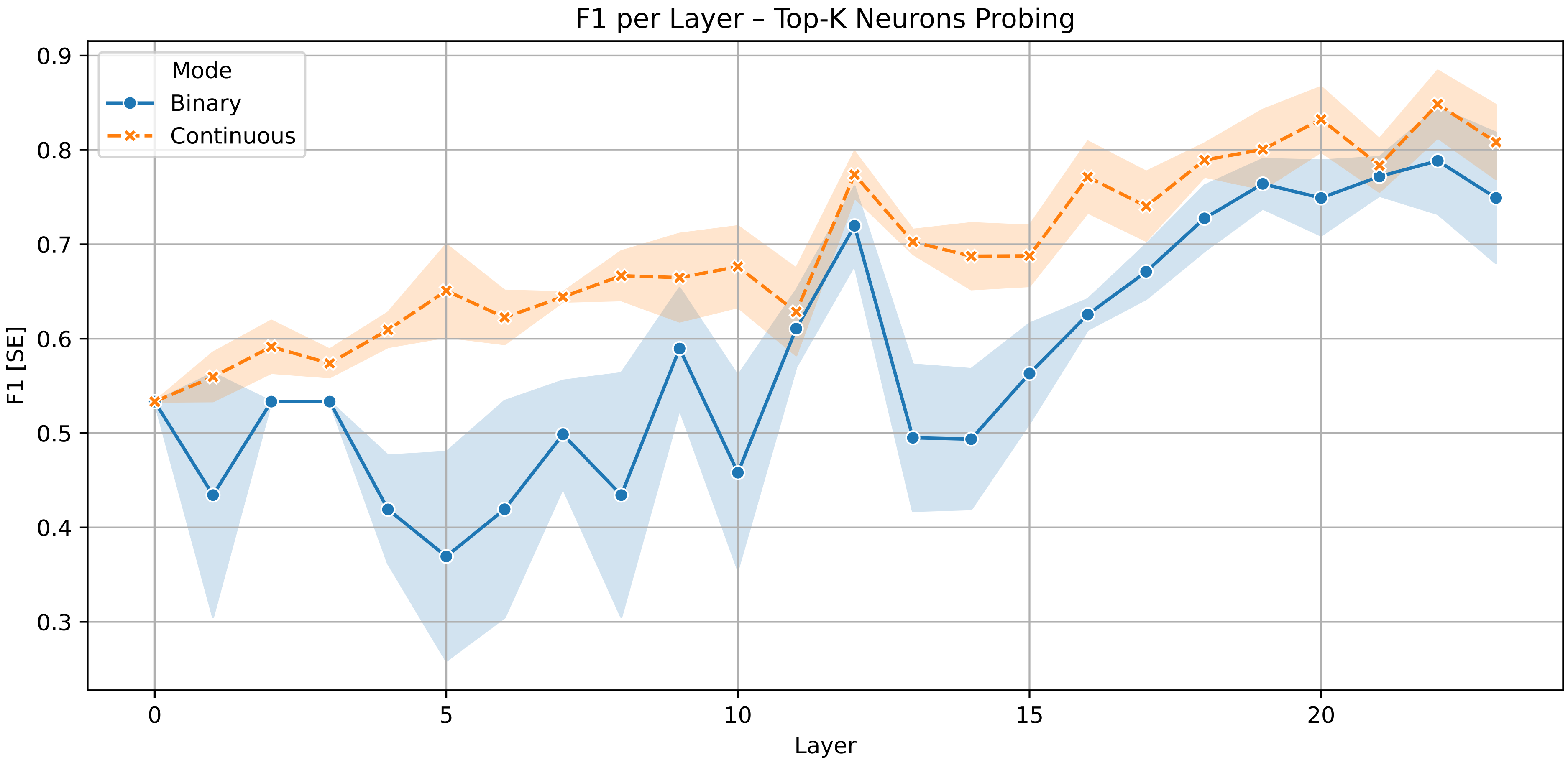
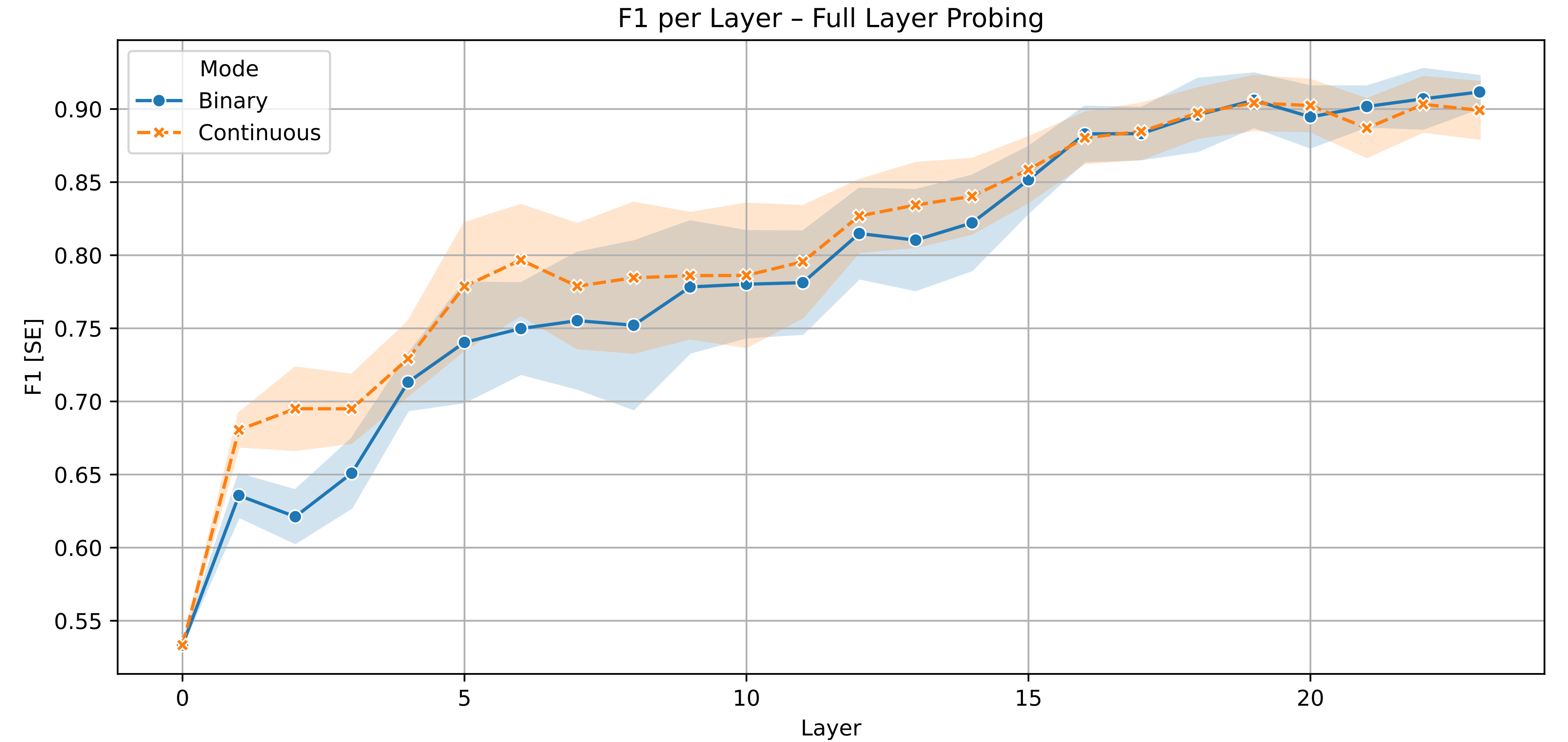
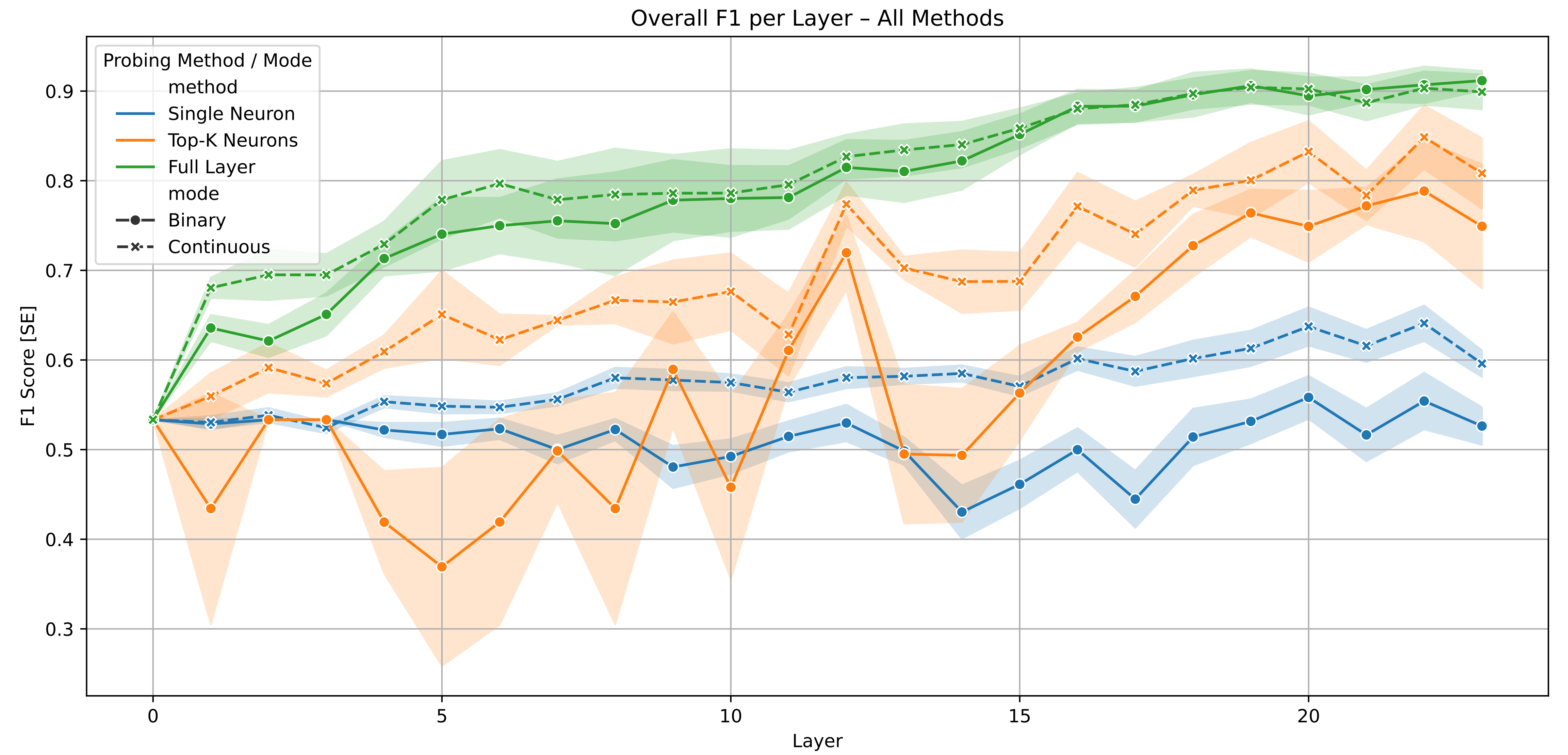
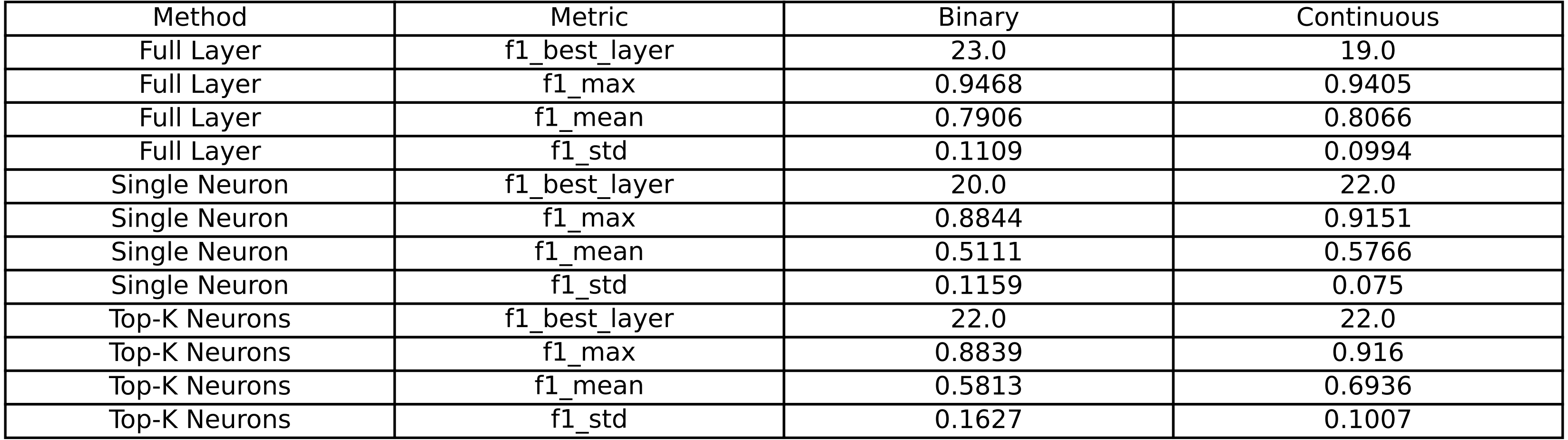

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.40
0.45
0.50
0.55
0.60
0.65
Layer
0
5
10
15
20


F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Layer
0
5
10
15
20

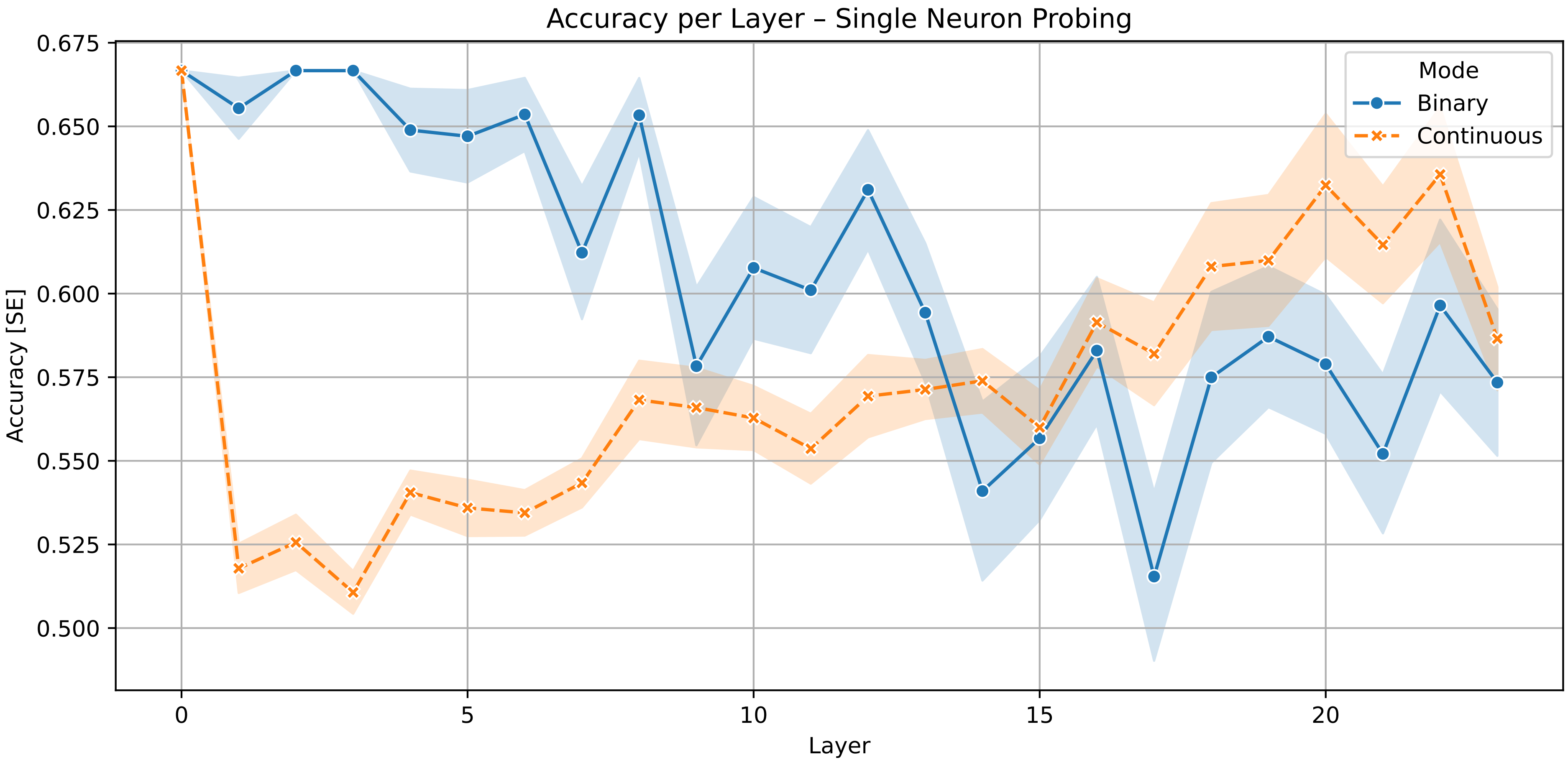

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
Layer
0
5
10
15
20

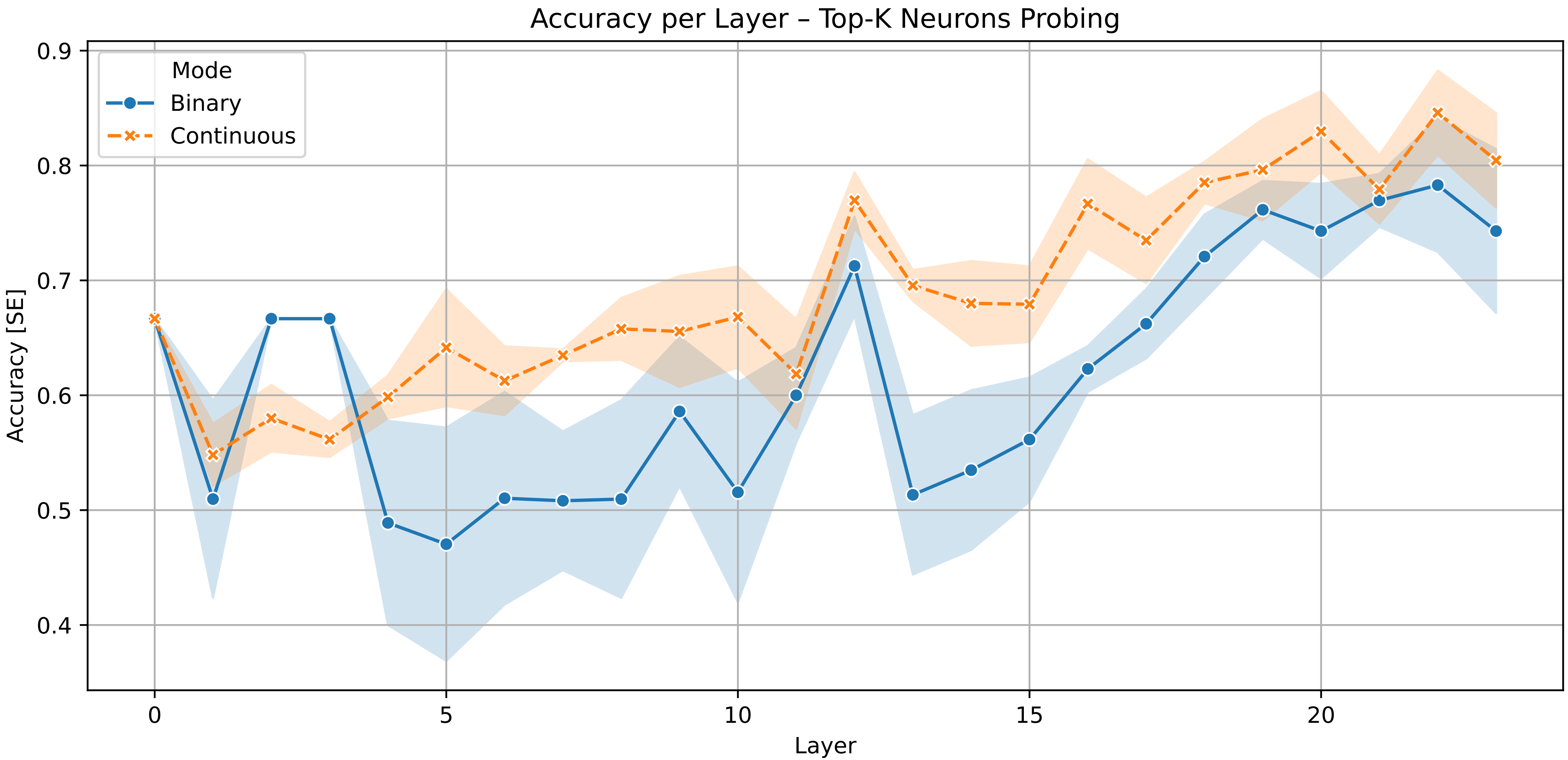

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Layer
0
5
10
15
20

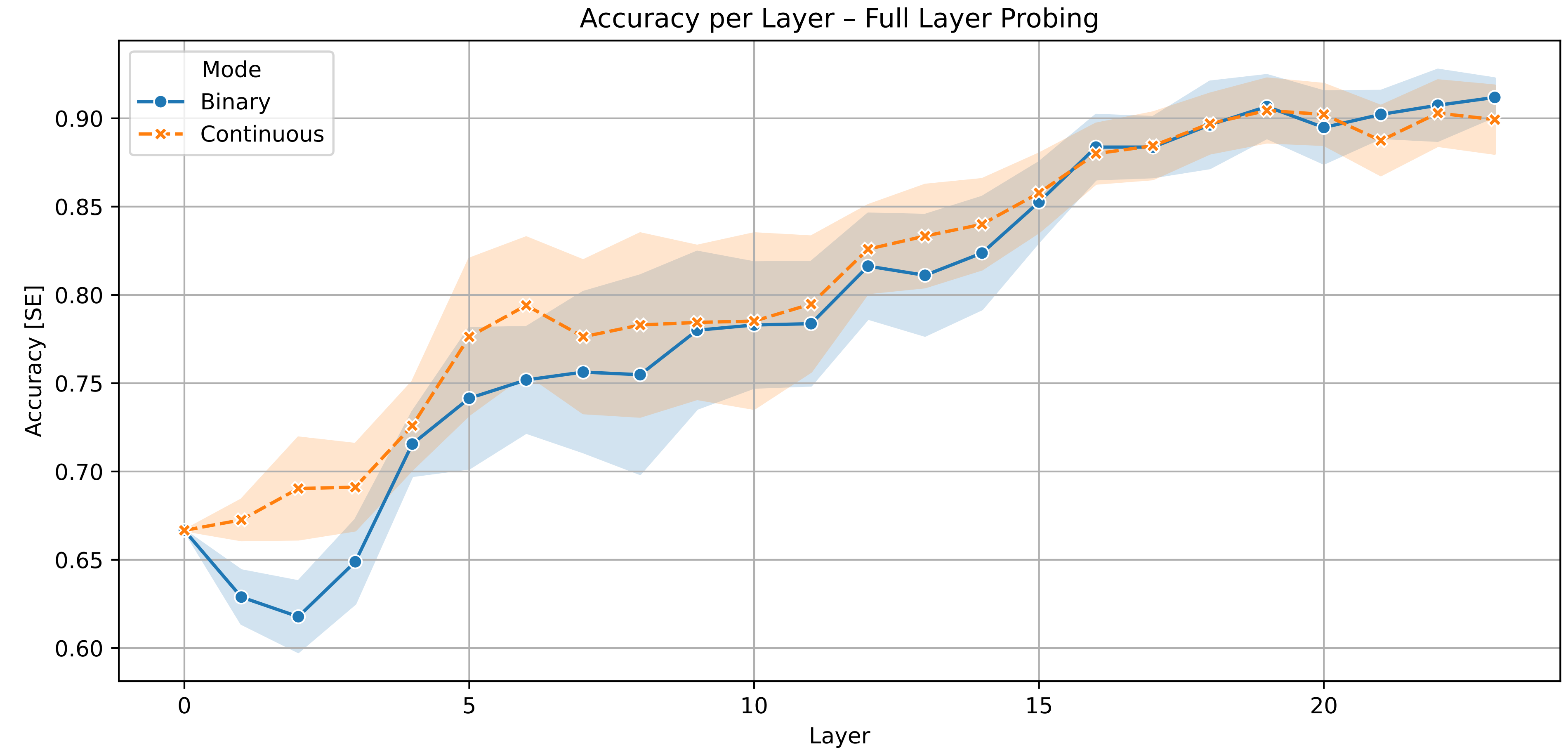
F1 Score Summary by Probing Method



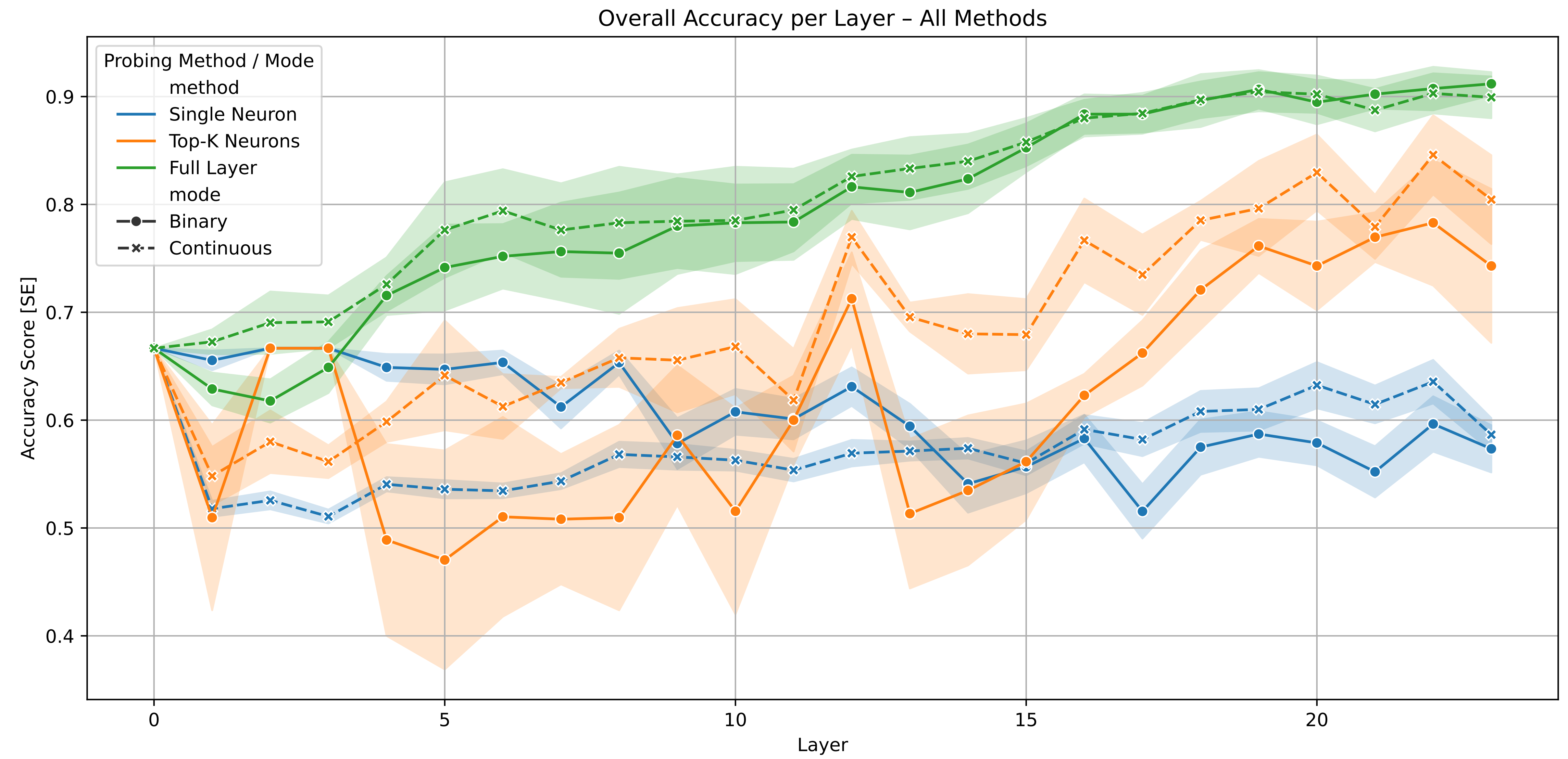
Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	23.0	19.0
Full Layer	f1_max	0.9468	0.9405
Full Layer	f1_mean	0.7906	0.8066
Full Layer	f1_std	0.1109	0.0994
Single Neuron	f1_best_layer	20.0	22.0
Single Neuron	f1_max	0.8844	0.9151
Single Neuron	f1_mean	0.5111	0.5766
Single Neuron	f1_std	0.1159	0.075
Top-K Neurons	f1_best_layer	22.0	22.0
Top-K Neurons	f1_max	0.8839	0.916
Top-K Neurons	f1_mean	0.5813	0.6936
Top-K Neurons	f1_std	0.1627	0.1007


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

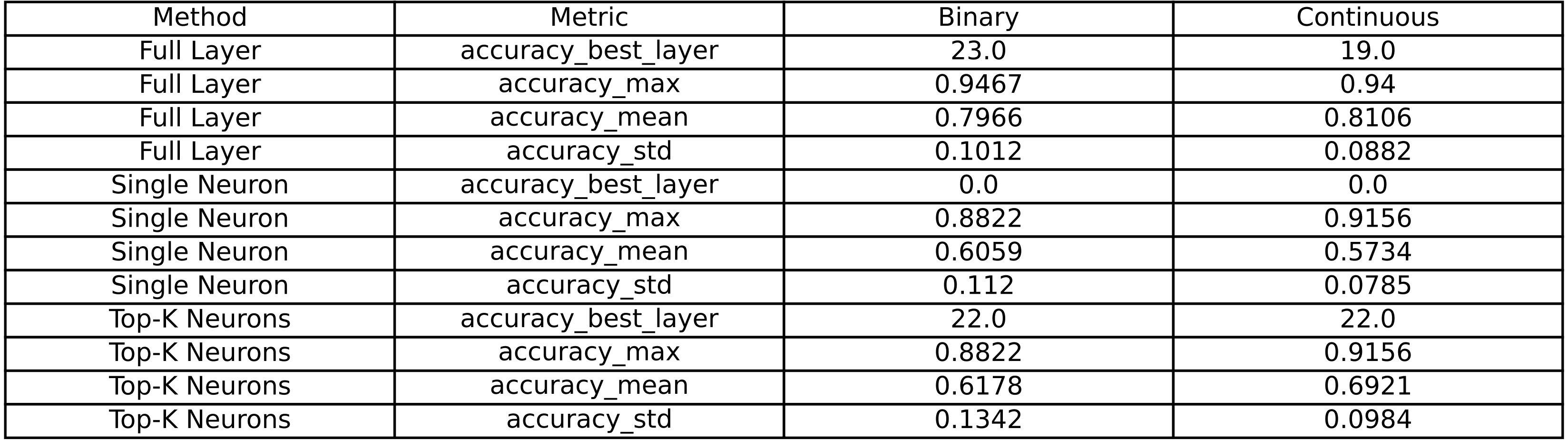

Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.4
0.5
0.6
0.7
0.8
0.9
Layer
0
5
10
15
20



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


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

Accuracy Score Summary by Probing Method



Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	23.0	19.0
Full Layer	accuracy_max	0.9467	0.94
Full Layer	accuracy_mean	0.7966	0.8106
Full Layer	accuracy_std	0.1012	0.0882
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.8822	0.9156
Single Neuron	accuracy_mean	0.6059	0.5734
Single Neuron	accuracy_std	0.112	0.0785
Top-K Neurons	accuracy_best_layer	22.0	22.0
Top-K Neurons	accuracy_max	0.8822	0.9156
Top-K Neurons	accuracy_mean	0.6178	0.6921
Top-K Neurons	accuracy_std	0.1342	0.0984