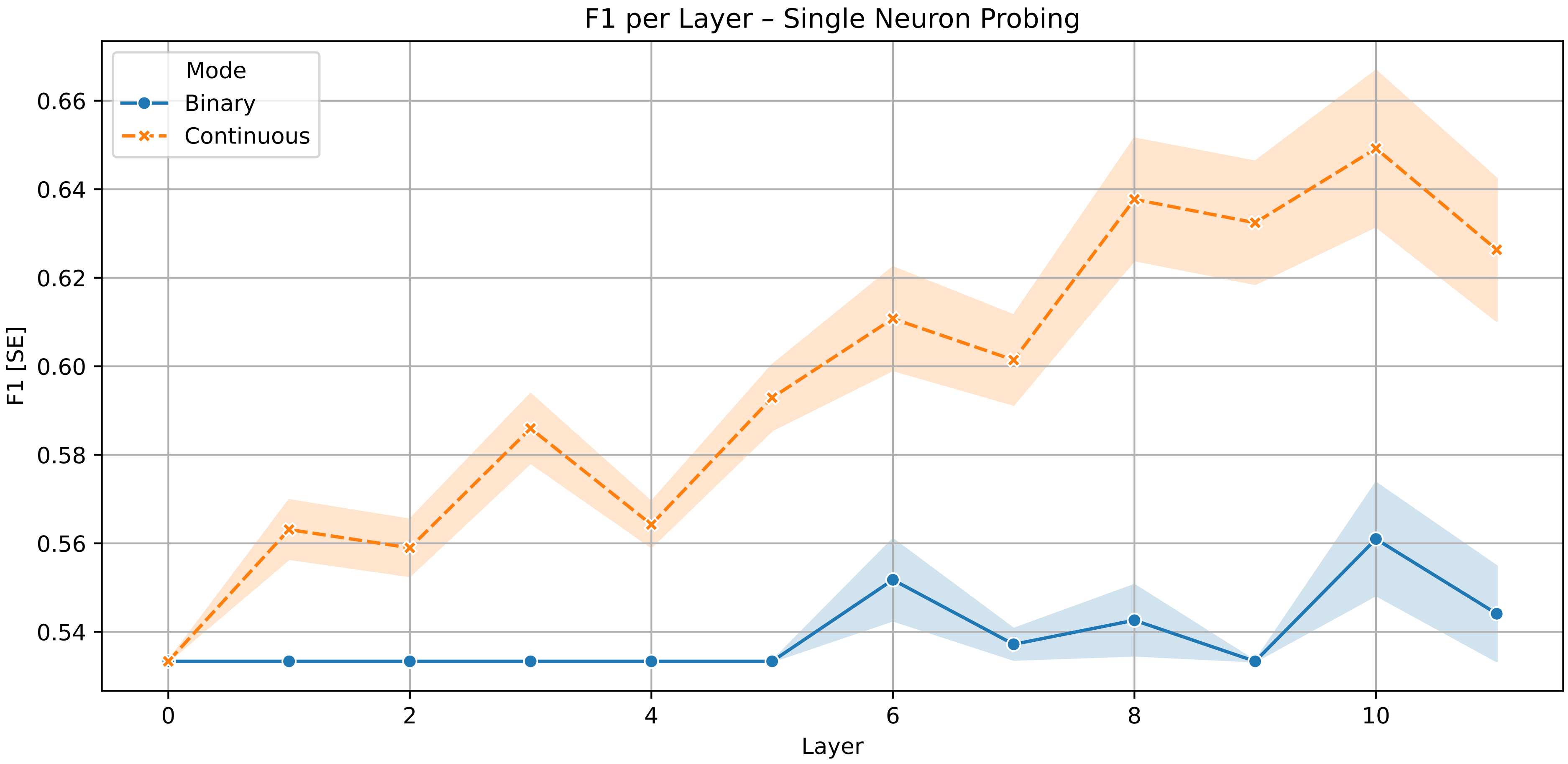
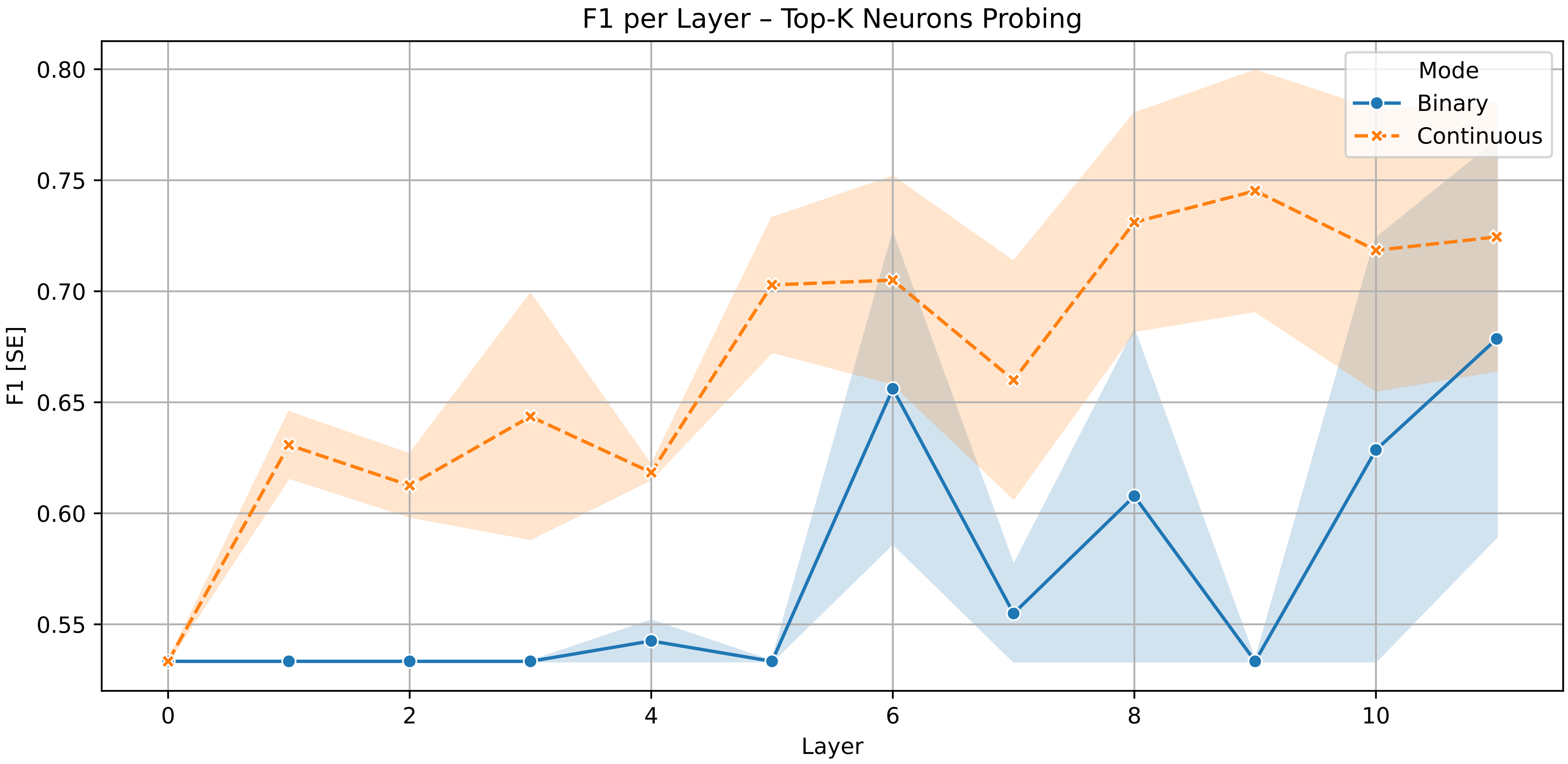
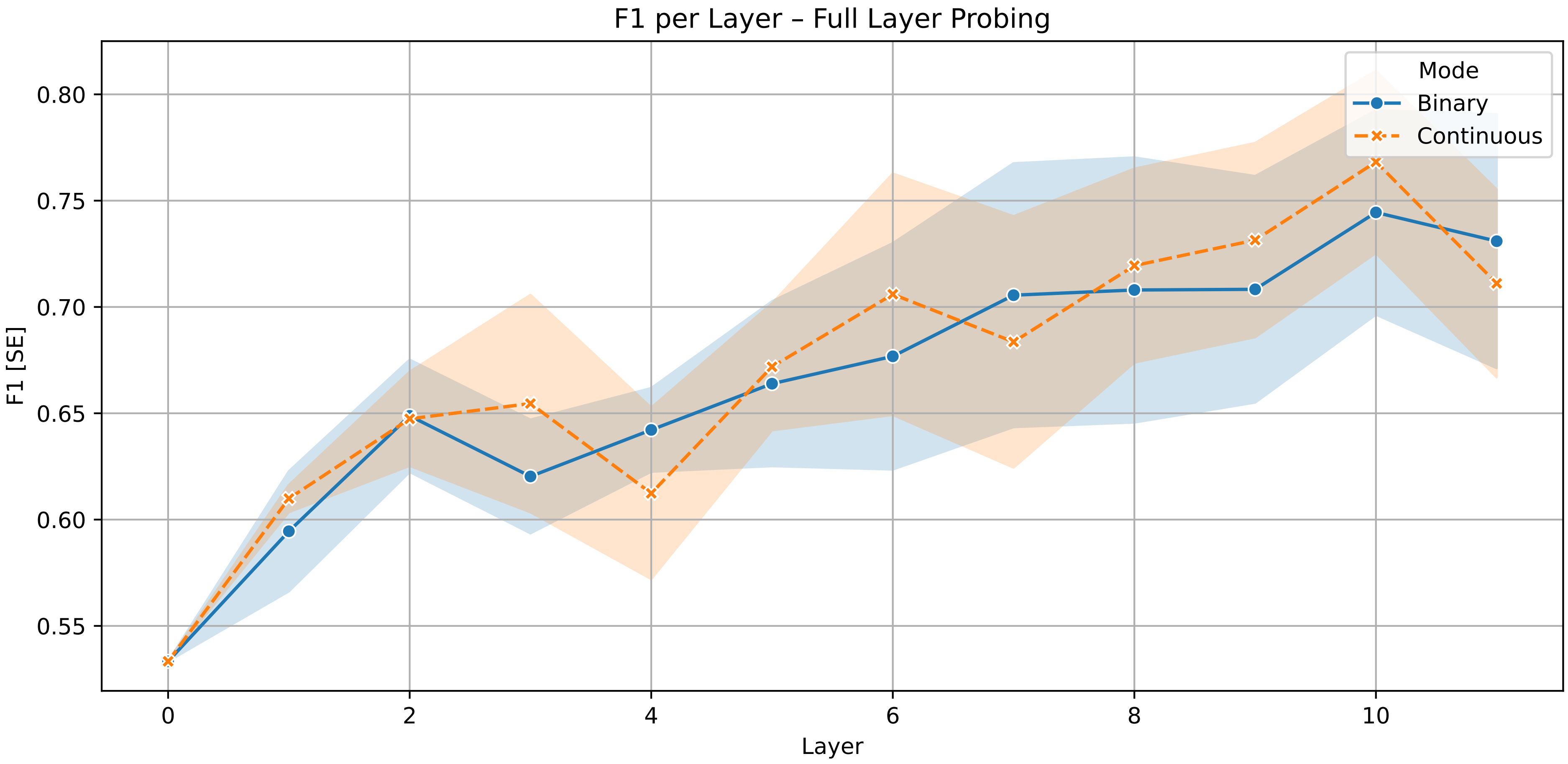
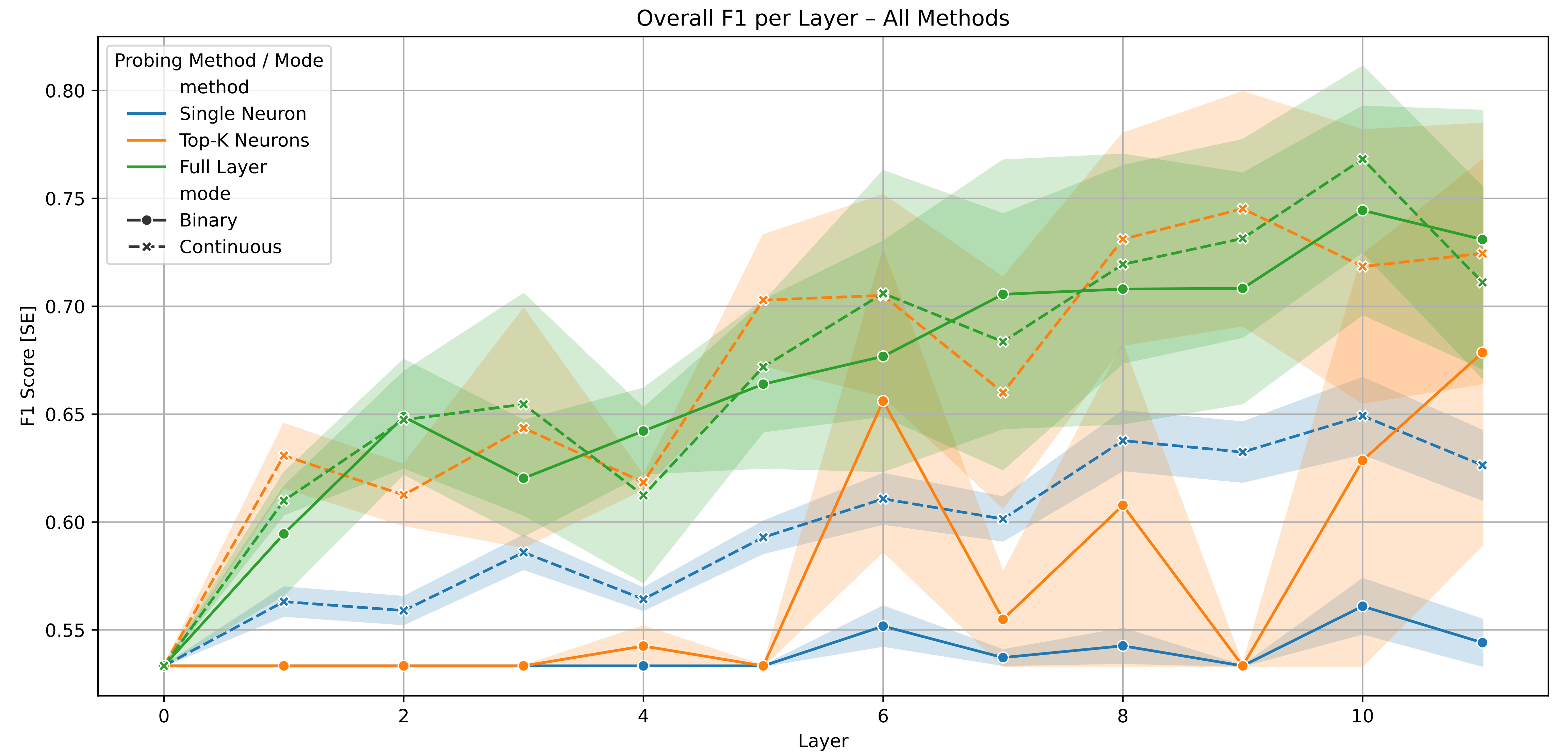

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.66
0.64
0.62
0.60
0.58
0.56
0.54
Layer
0
2
4
6
8
10

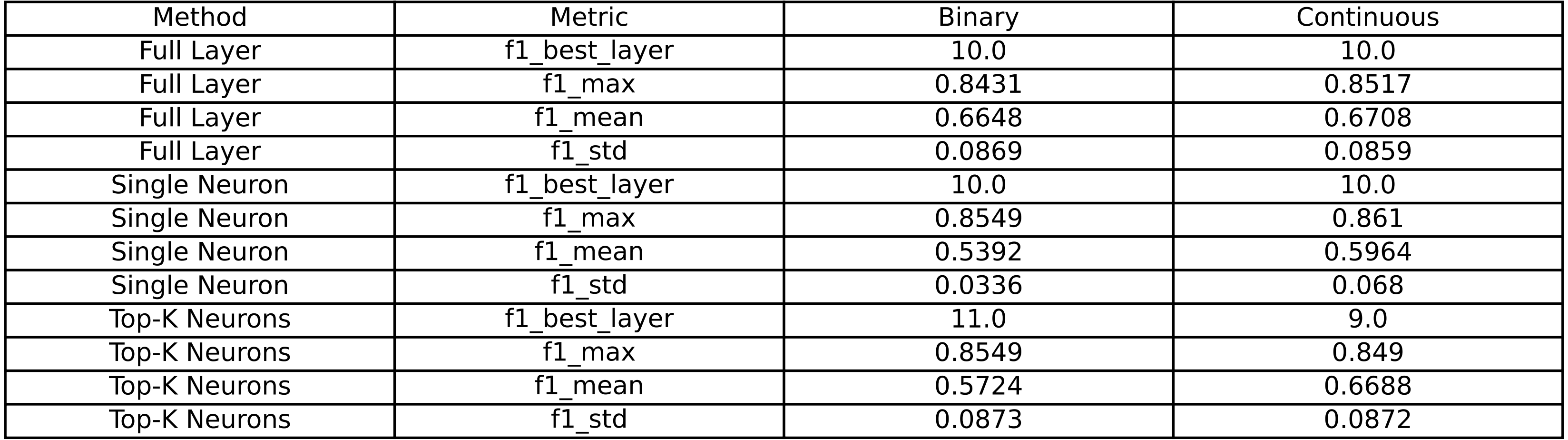

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.80
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
Layer



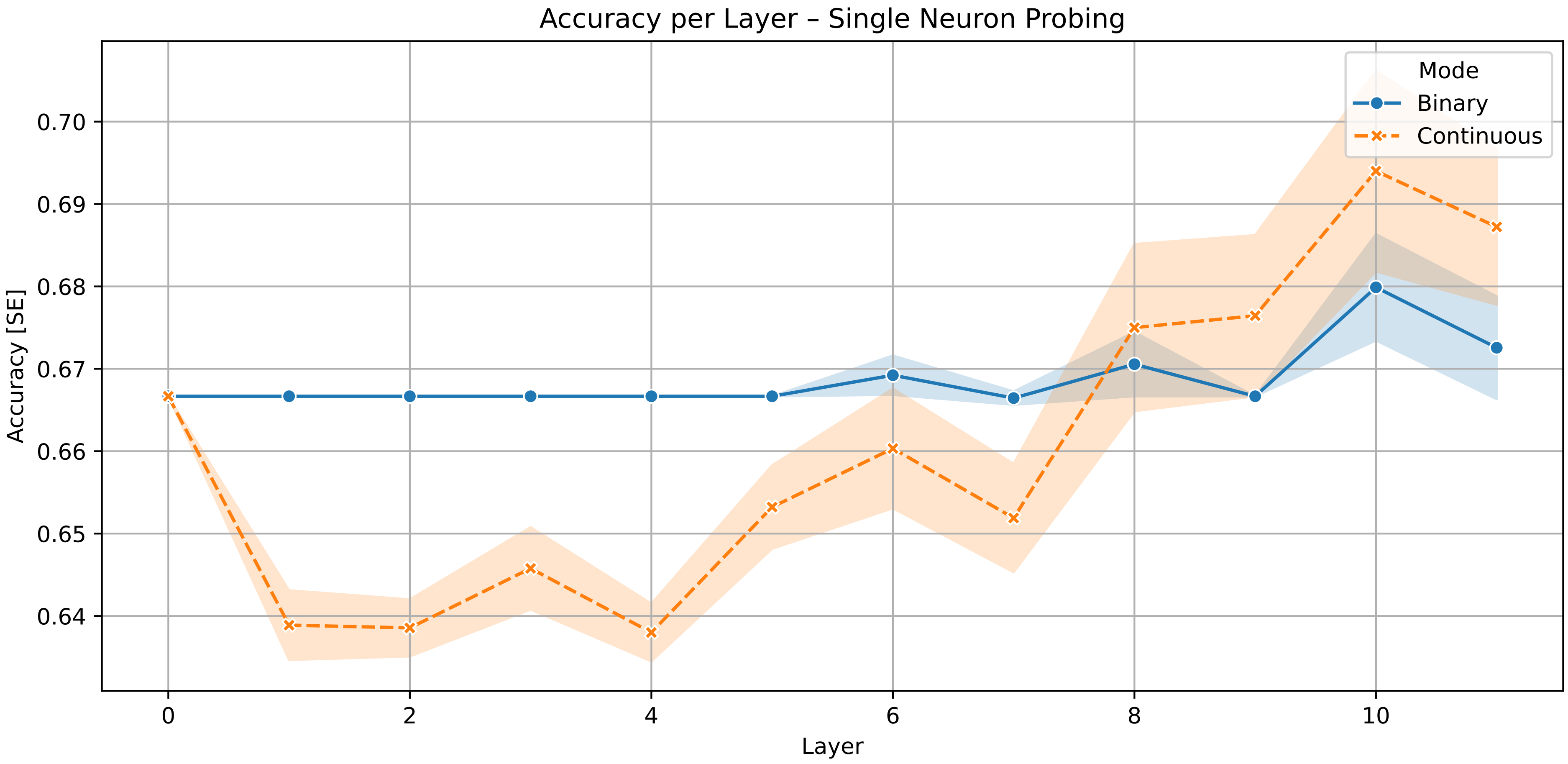
F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.55
0.60
0.65
0.70
0.75
0.80
Layer
0
2
4
6
8
10


Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.55
0.60
0.65
0.70
0.75
0.80
Layer
0
2
4
6
8
10

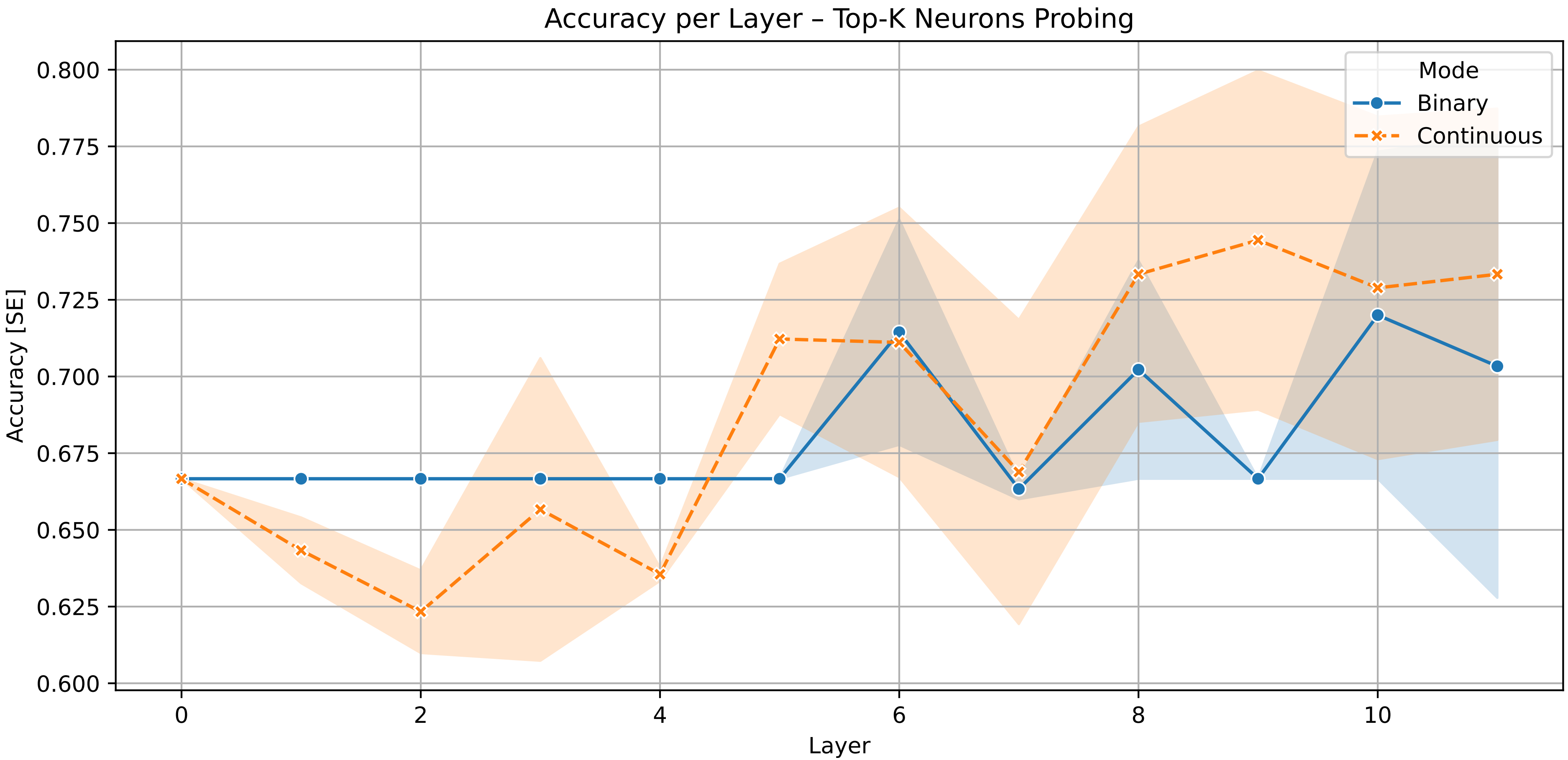
F1 Score Summary by Probing Method

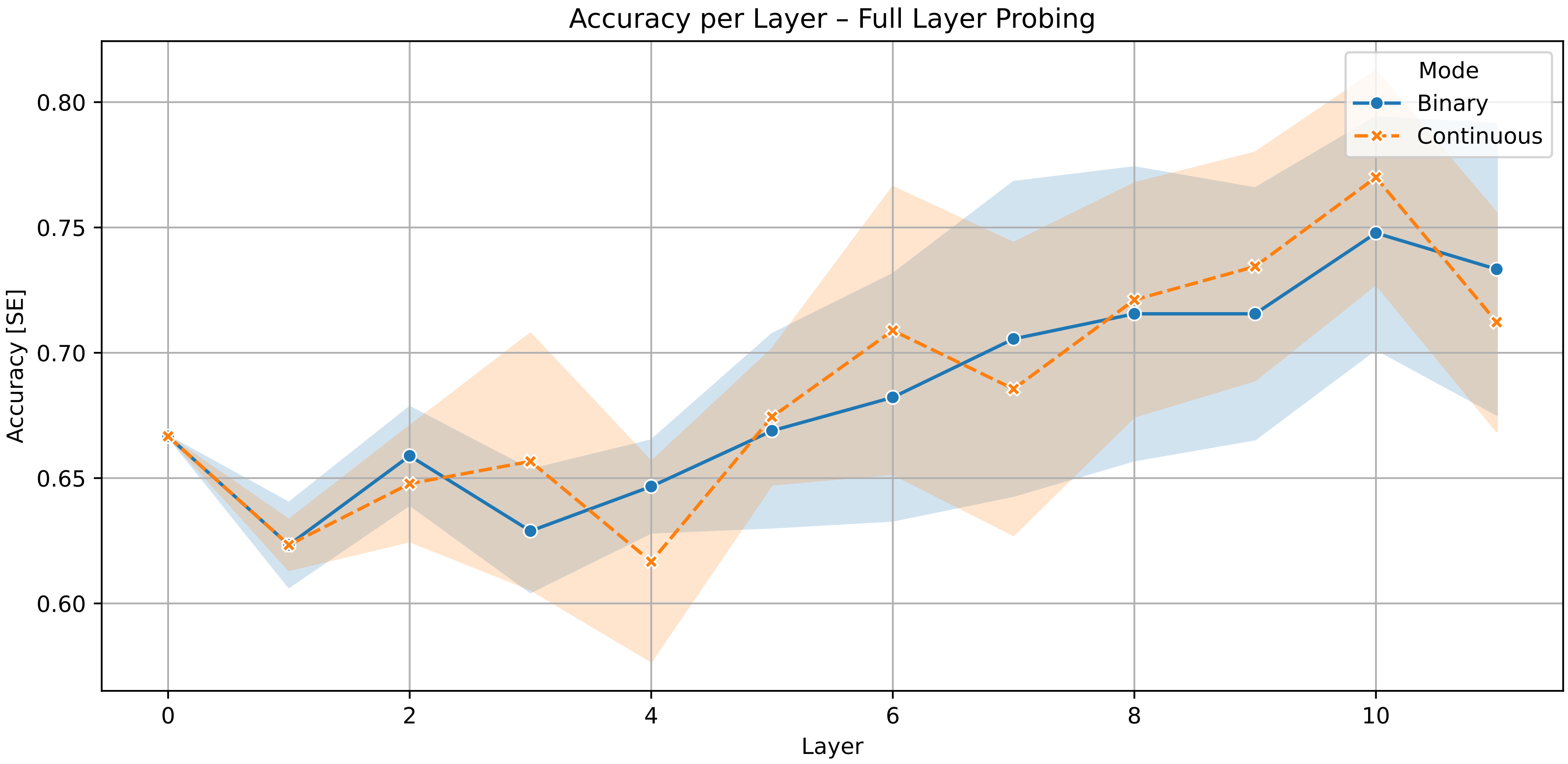


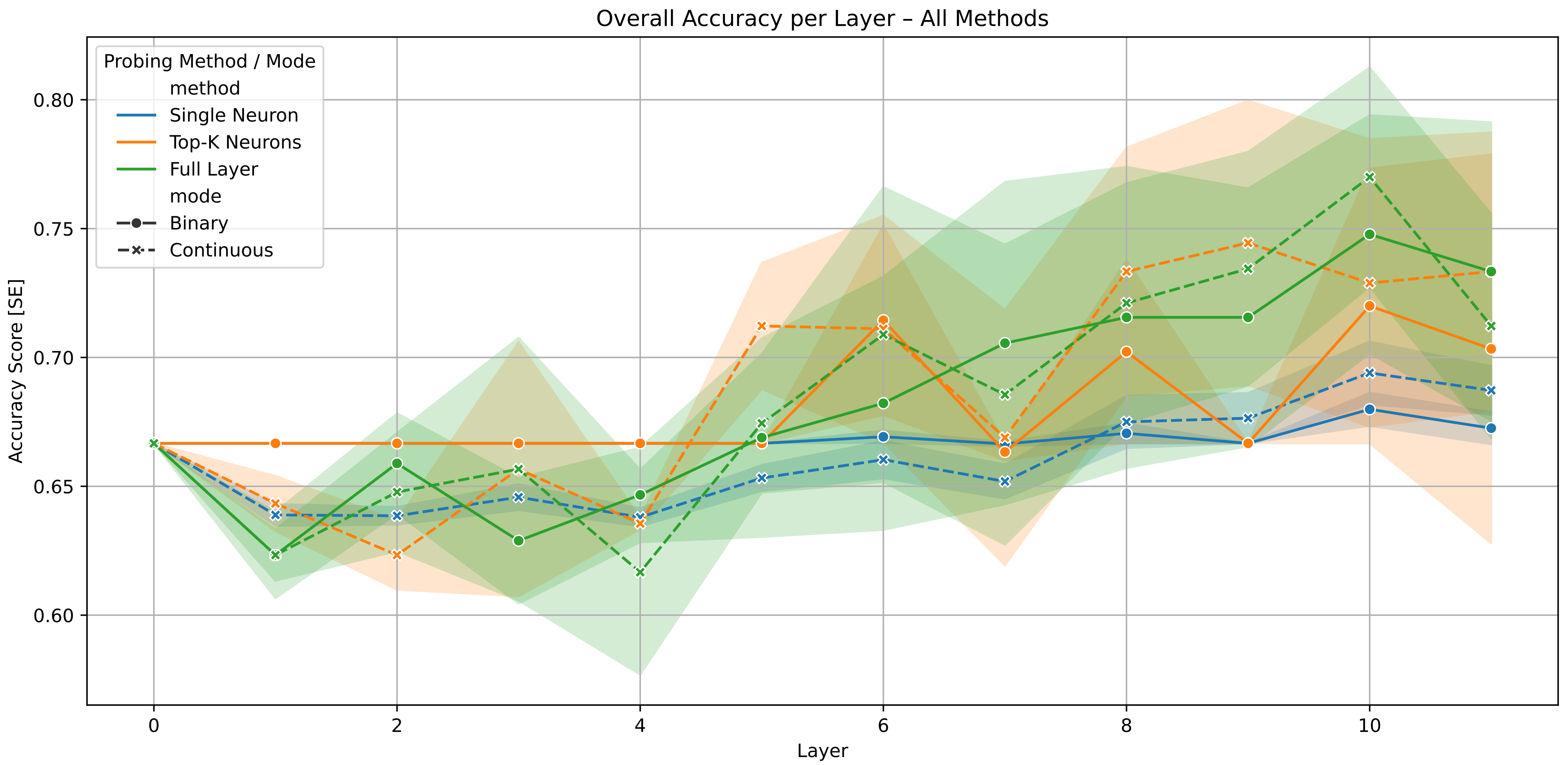
Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	10.0	10.0
Full Layer	f1_max	0.8431	0.8517
Full Layer	f1_mean	0.6648	0.6708
Full Layer	f1_std	0.0869	0.0859
Single Neuron	f1_best_layer	10.0	10.0
Single Neuron	f1_max	0.8549	0.861
Single Neuron	f1_mean	0.5392	0.5964
Single Neuron	f1_std	0.0336	0.068
Top-K Neurons	f1_best_layer	11.0	9.0
Top-K Neurons	f1_max	0.8549	0.849
Top-K Neurons	f1_mean	0.5724	0.6688
Top-K Neurons	f1_std	0.0873	0.0872



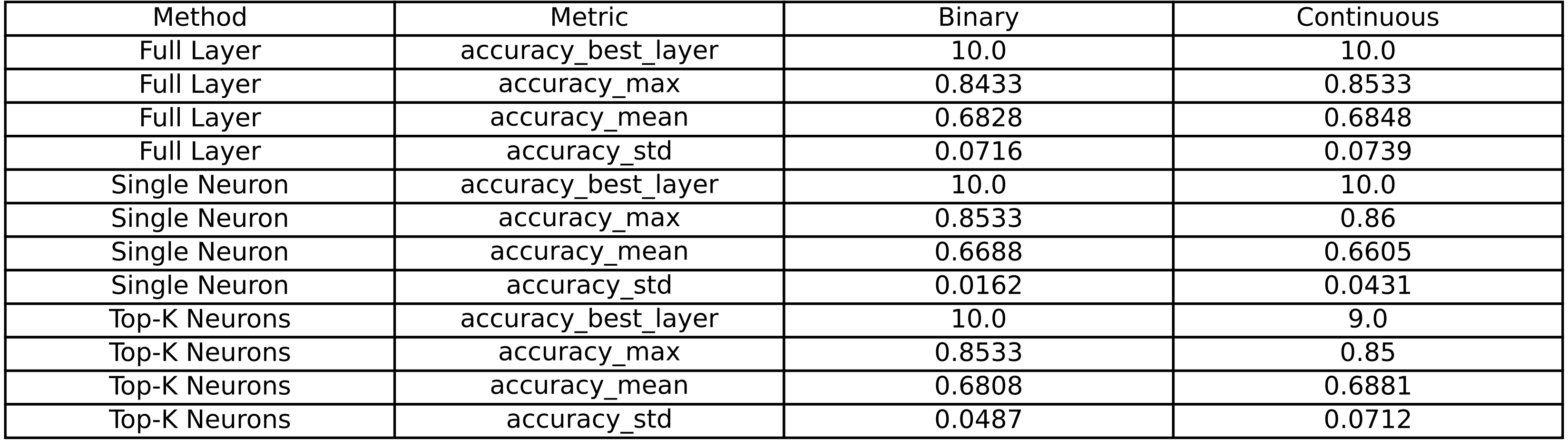
Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.64
0.65
0.66
0.67
0.68
0.69
0.70
Layer
0
2
4
6
8
10


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.600
0.625
0.650
0.675
0.700
0.725
0.750
0.775
0.800
Layer
0
2
4
6
8
10


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.60
0.65
0.70
0.75
0.80
Layer
0
2
4
6
8
10


Overall Accuracy per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
Accuracy Score [SE]
0.80
0.75
0.70
0.65
0.60
Layer
0
2
4
6
8
10

Accuracy Score Summary by Probing Method



Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	10.0	10.0
Full Layer	accuracy_max	0.8433	0.8533
Full Layer	accuracy_mean	0.6828	0.6848
Full Layer	accuracy_std	0.0716	0.0739
Single Neuron	accuracy_best_layer	10.0	10.0
Single Neuron	accuracy_max	0.8533	0.86
Single Neuron	accuracy_mean	0.6688	0.6605
Single Neuron	accuracy_std	0.0162	0.0431
Top-K Neurons	accuracy_best_layer	10.0	9.0
Top-K Neurons	accuracy_max	0.8533	0.85
Top-K Neurons	accuracy_mean	0.6808	0.6881
Top-K Neurons	accuracy_std	0.0487	0.0712