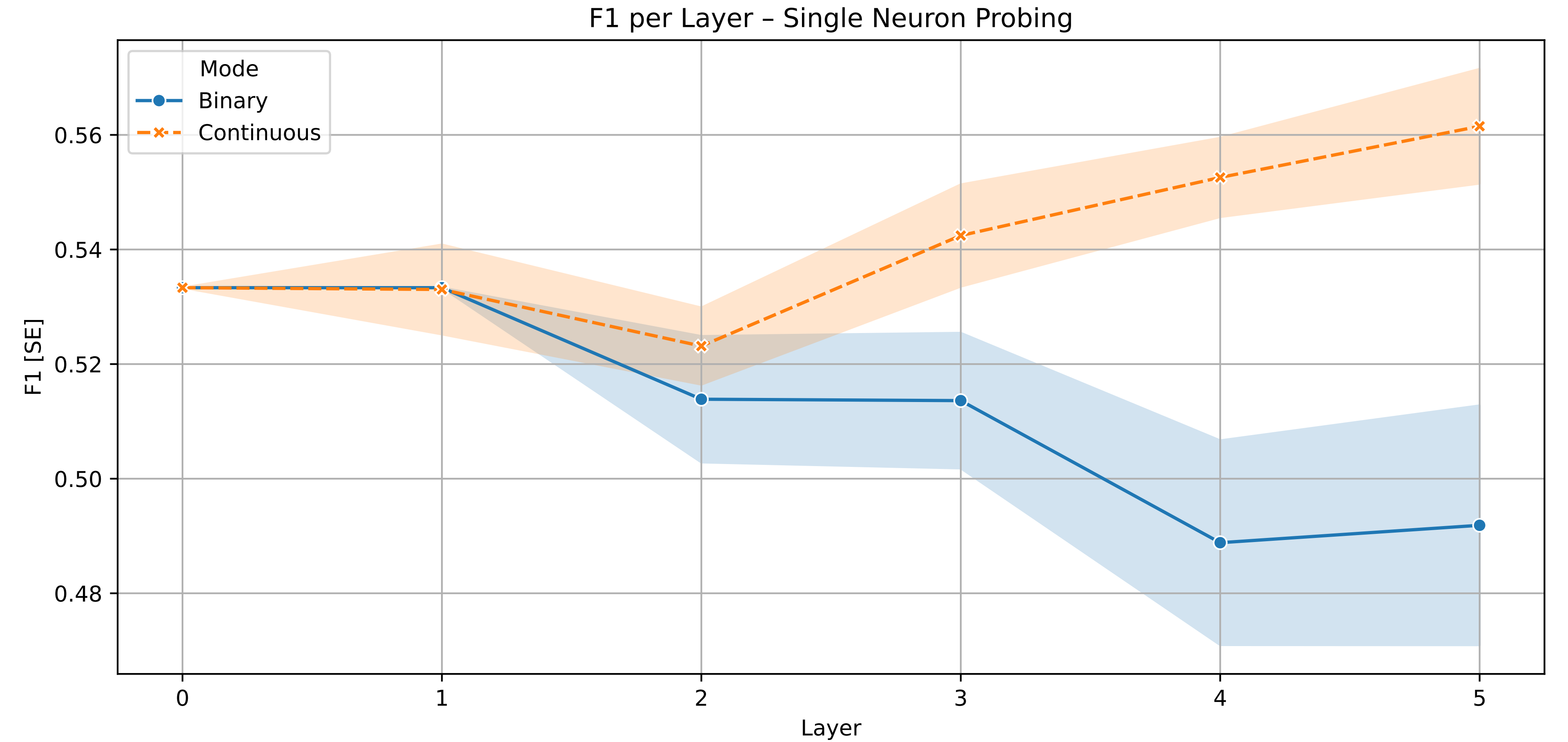
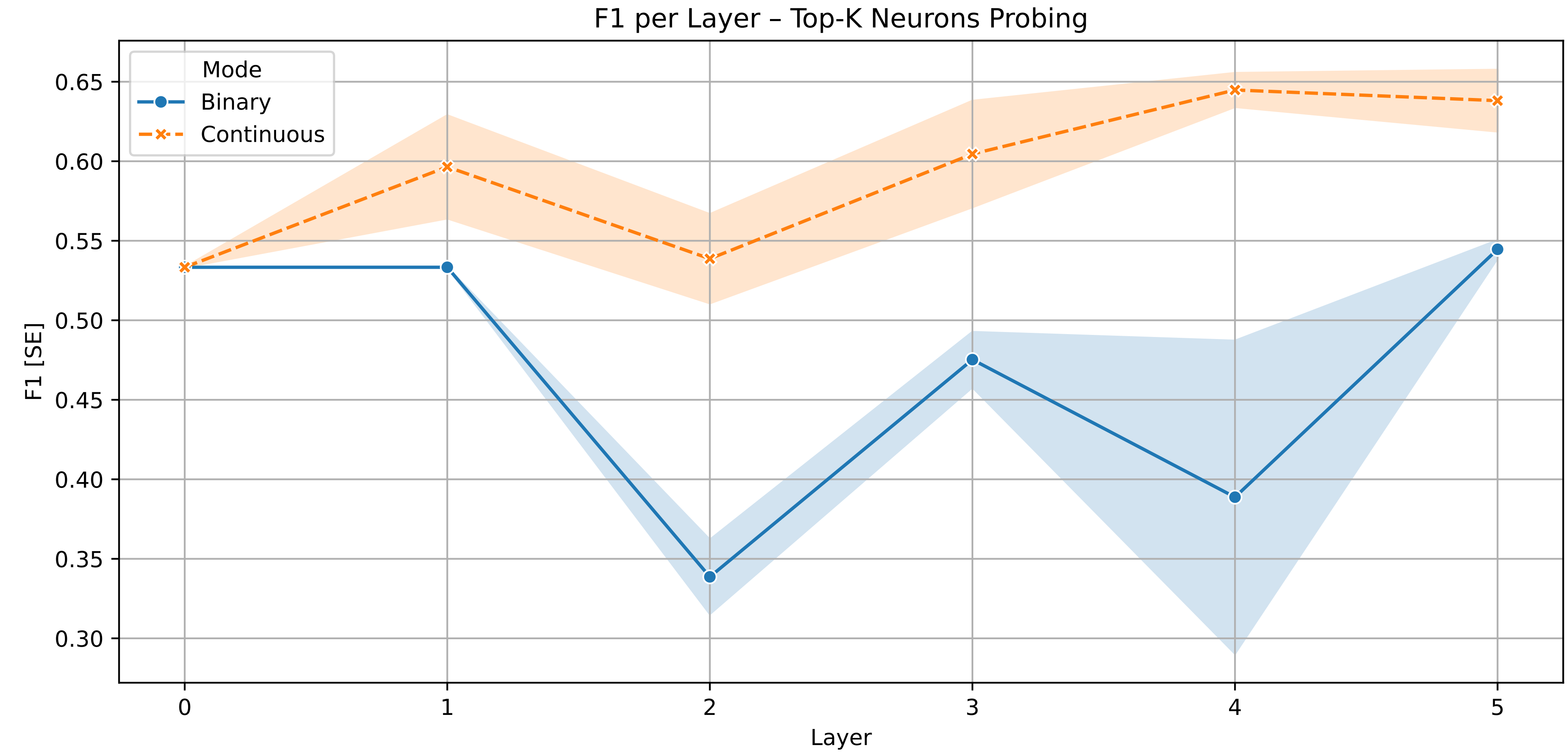
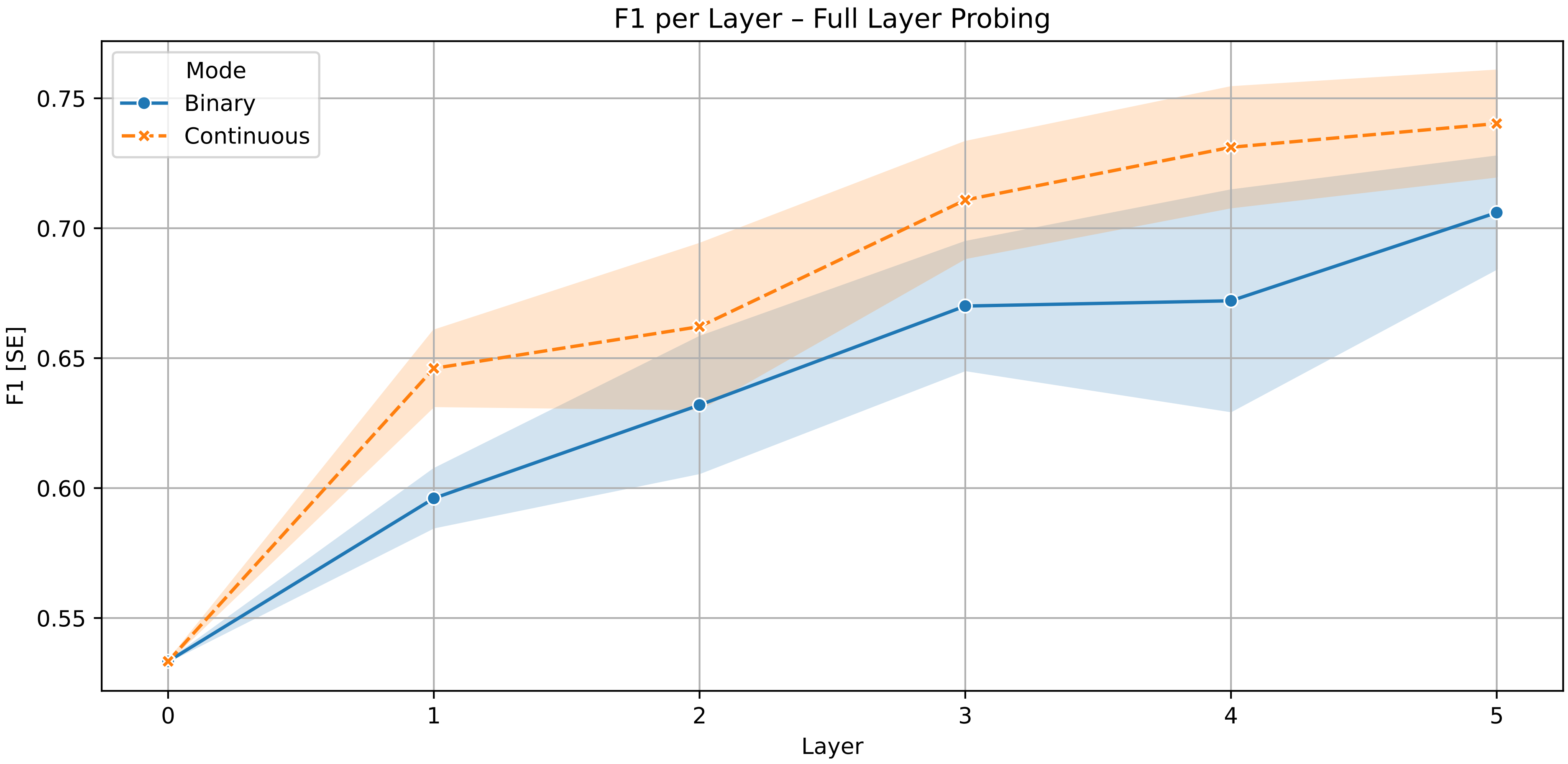
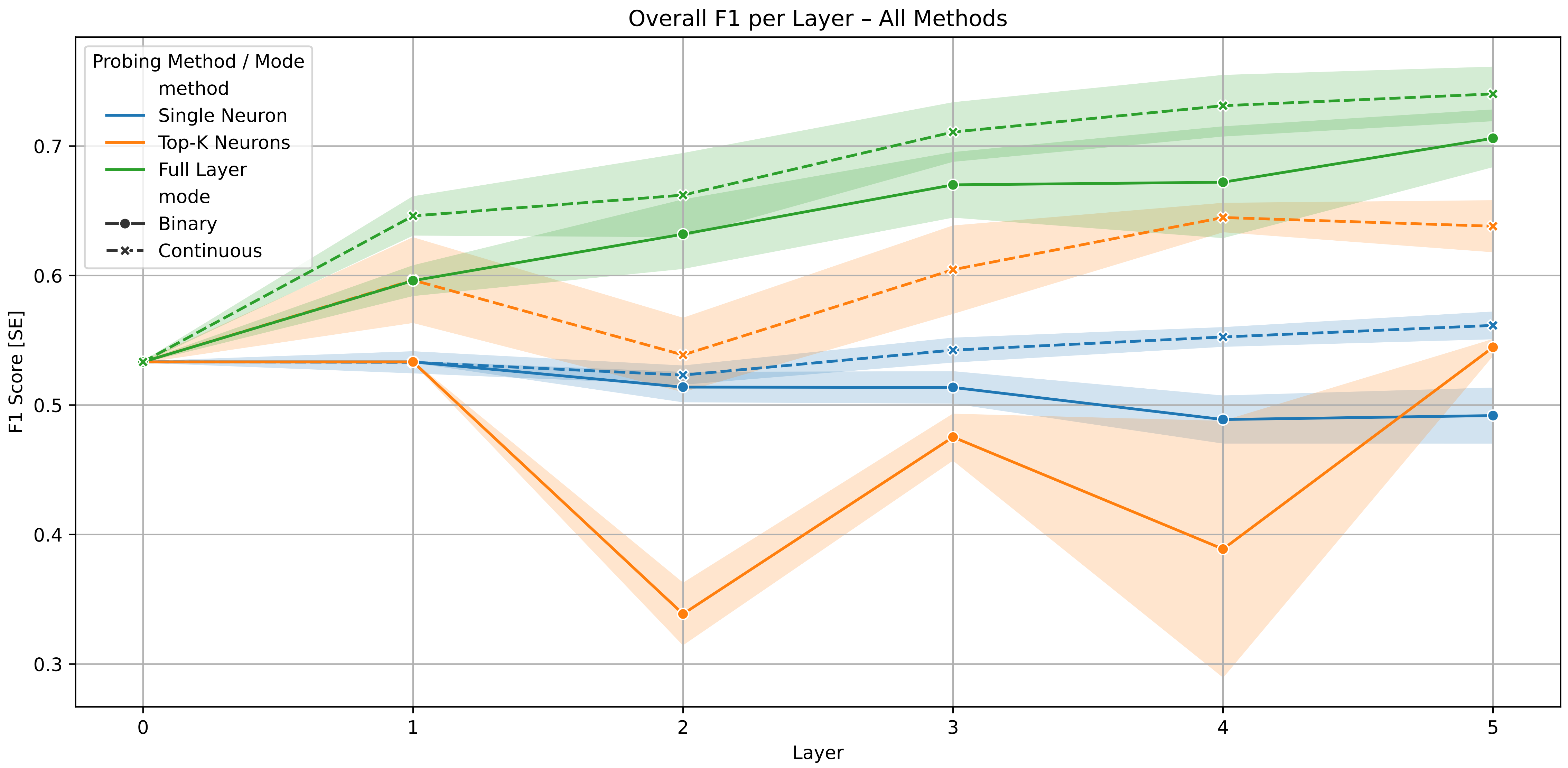
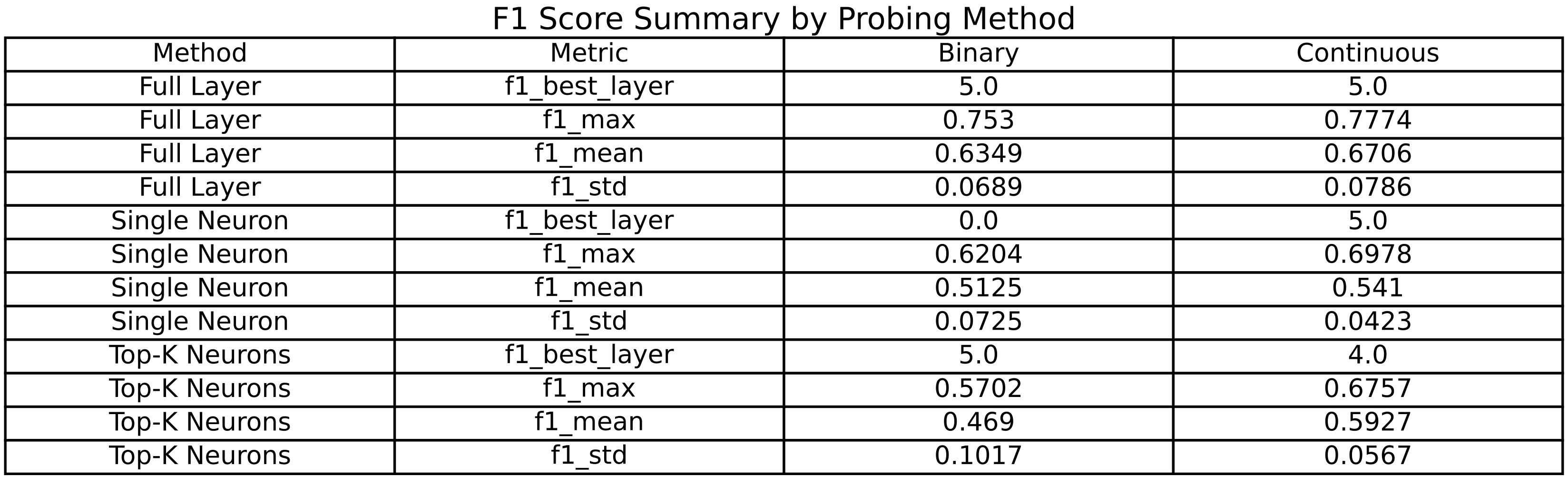

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.56
0.54
0.52
0.50
0.48
0
1
2
3
4
5
Layer


F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
Layer
0
1
2
3
4
5

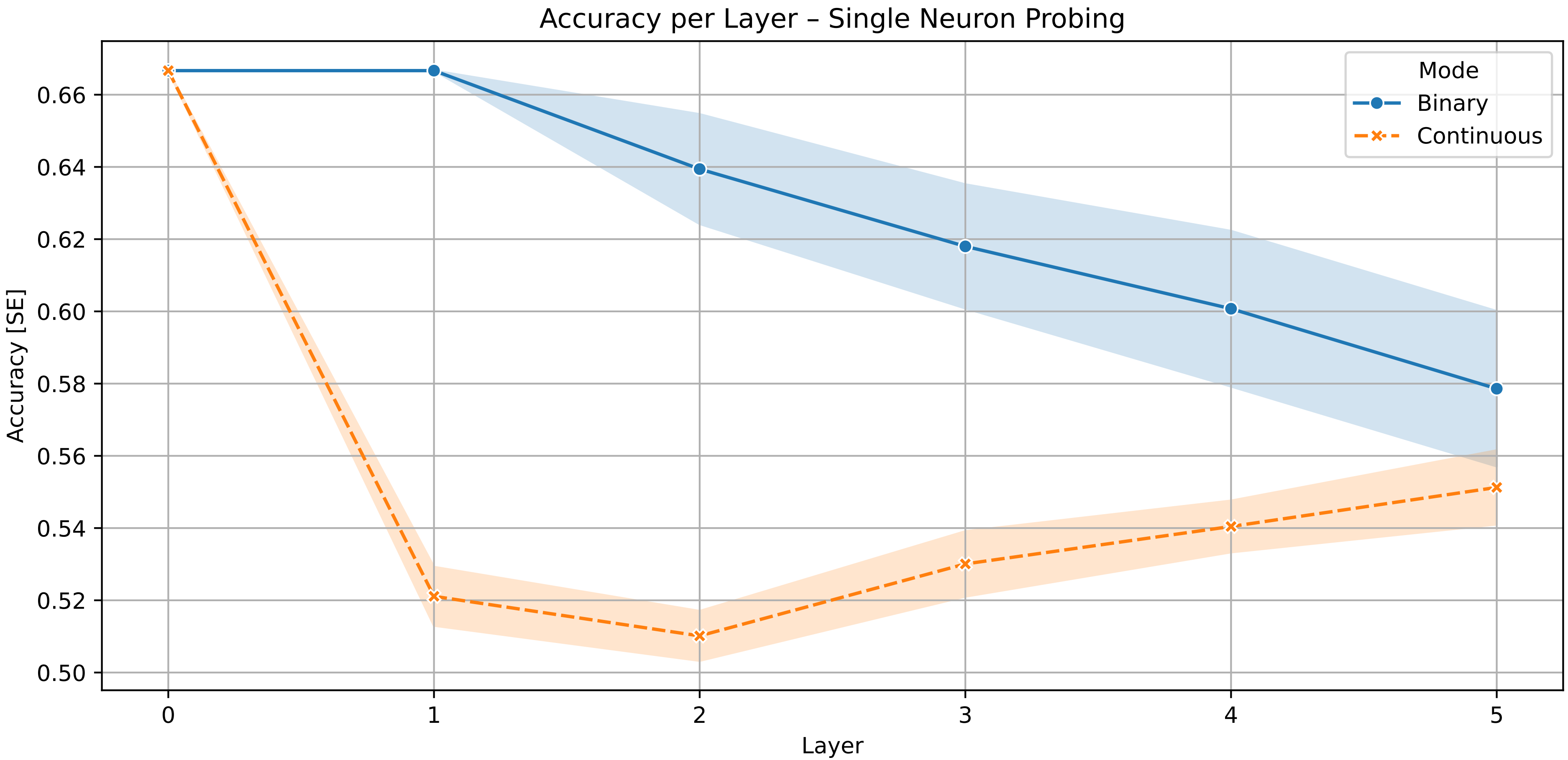

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.75
0.70
0.65
0.60
0.55
0
1
2
3
4
5
Layer

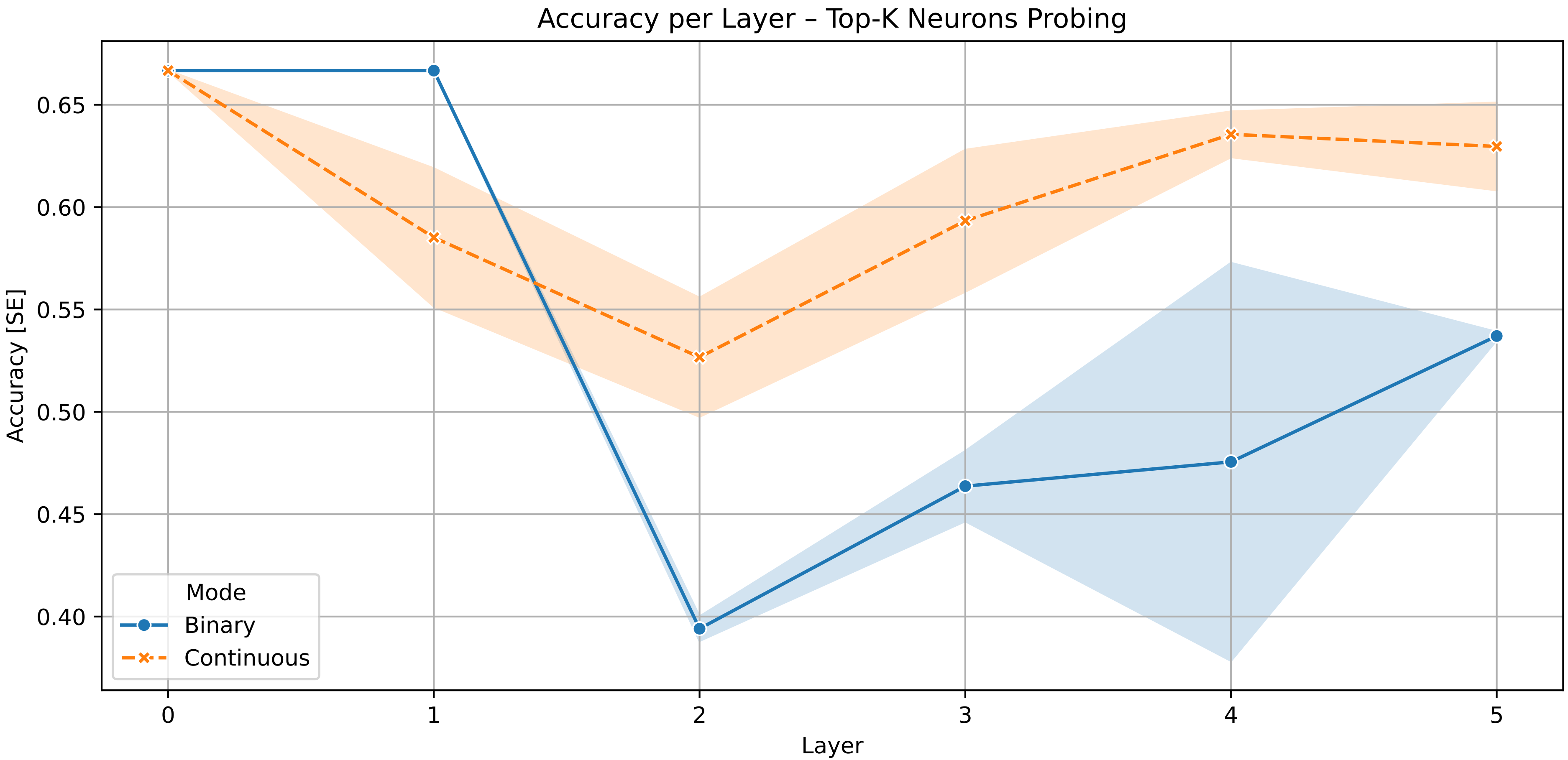

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.3
0.4
0.5
0.6
0.7
Layer
0
1
2
3
4
5

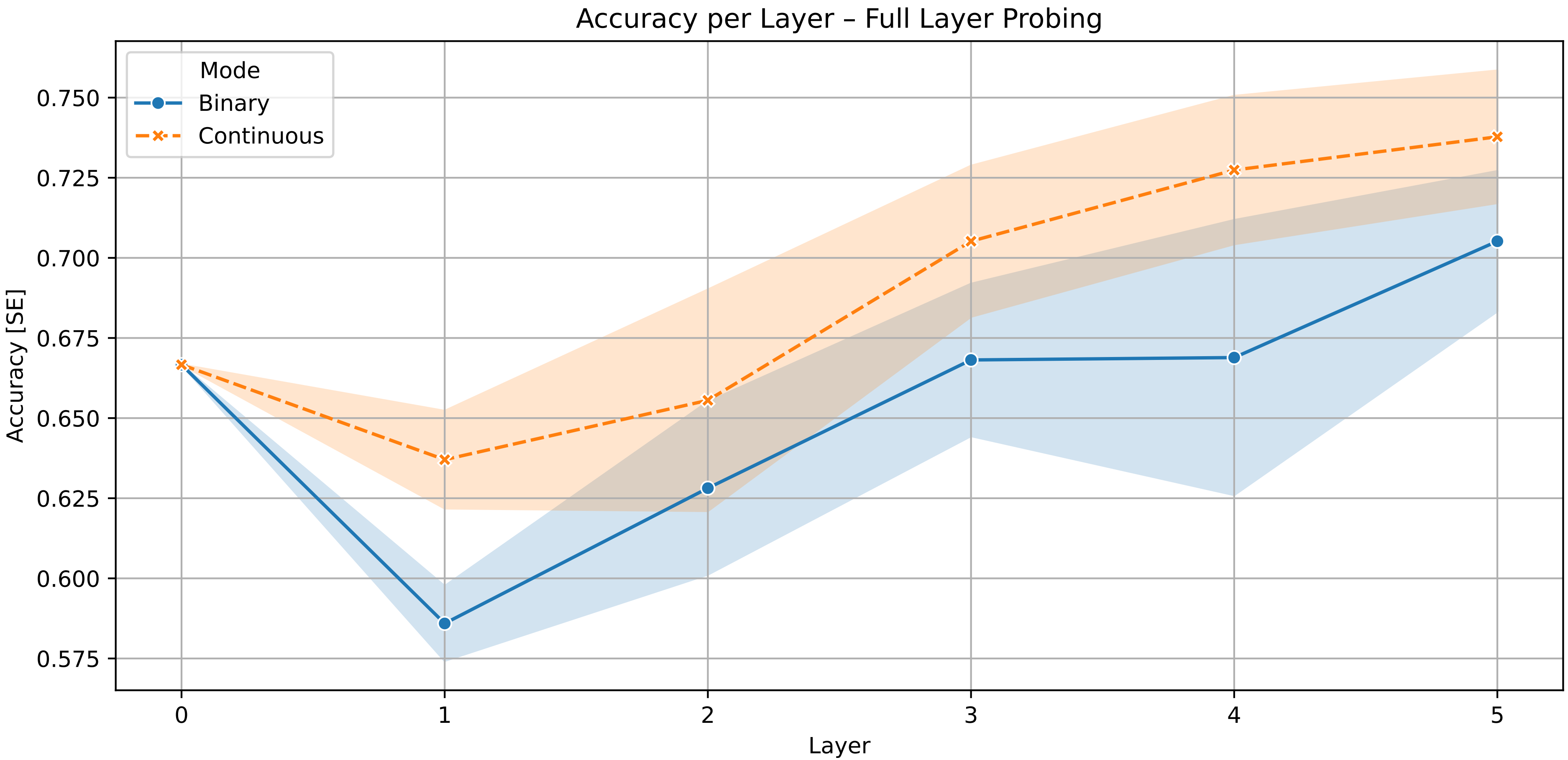


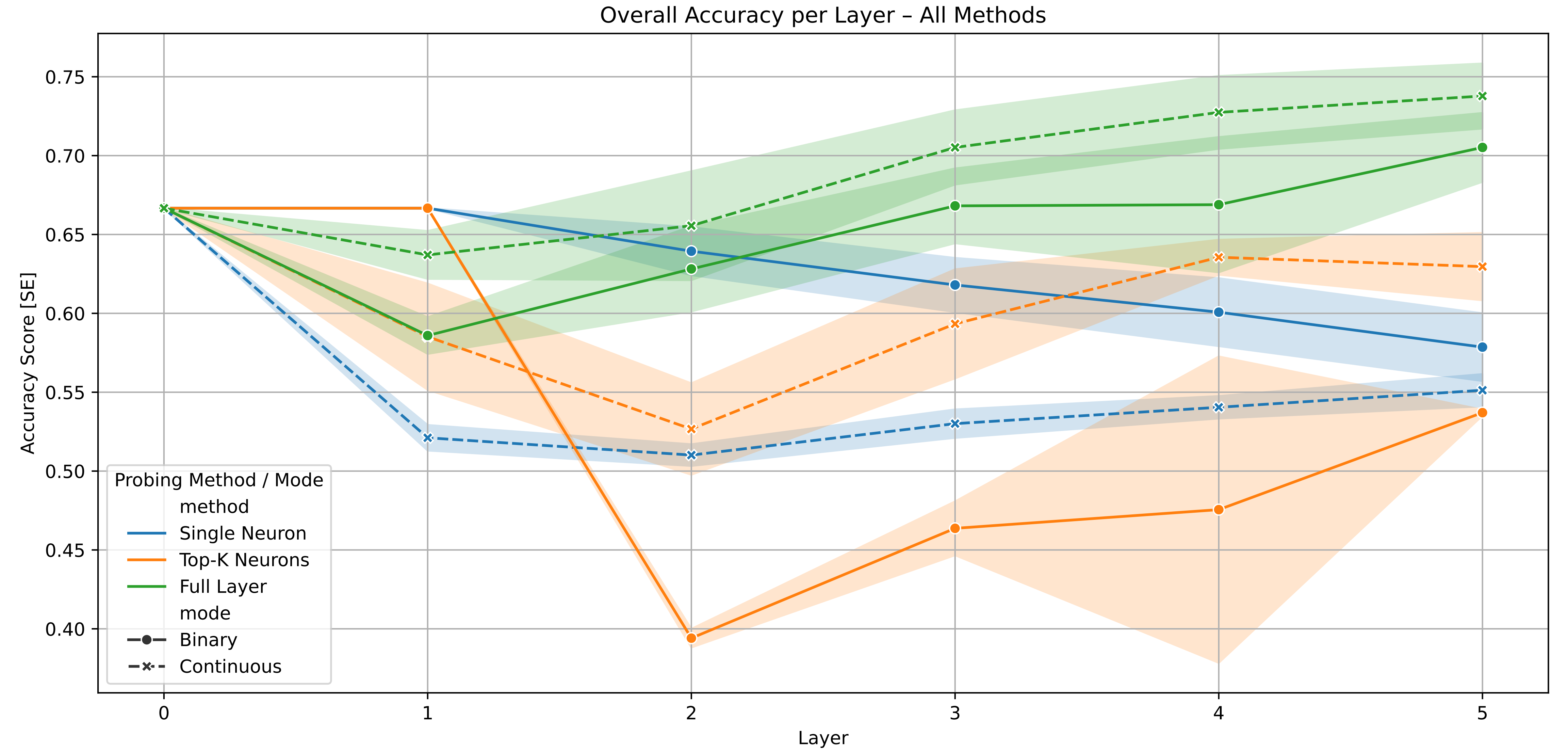
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	5.0	5.0
Full Layer	f1_max	0.753	0.7774
Full Layer	f1_mean	0.6349	0.6706
Full Layer	f1_std	0.0689	0.0786
Single Neuron	f1_best_layer	0.0	5.0
Single Neuron	f1_max	0.6204	0.6978
Single Neuron	f1_mean	0.5125	0.541
Single Neuron	f1_std	0.0725	0.0423
Top-K Neurons	f1_best_layer	5.0	4.0
Top-K Neurons	f1_max	0.5702	0.6757
Top-K Neurons	f1_mean	0.469	0.5927
Top-K Neurons	f1_std	0.1017	0.0567

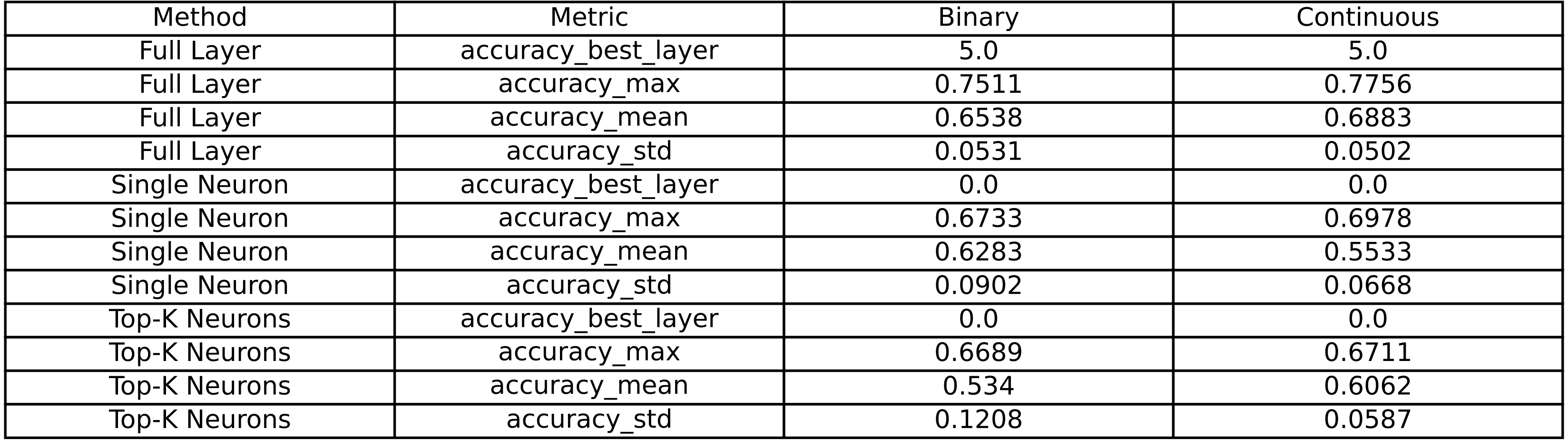

Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
0.50
0
1
2
3
4
5
Layer


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
0.65
0.60
0.55
0.50
0.45
0.40
Layer
0
1
2
3
4
5
Mode
Binary
Continuous


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.575
0.600
0.625
0.650
0.675
0.700
0.725
0.750
Layer
0
1
2
3
4
5


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
Layer
0
1
2
3
4
5
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	5.0	5.0
Full Layer	accuracy_max	0.7511	0.7756
Full Layer	accuracy_mean	0.6538	0.6883
Full Layer	accuracy_std	0.0531	0.0502
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.6733	0.6978
Single Neuron	accuracy_mean	0.6283	0.5533
Single Neuron	accuracy_std	0.0902	0.0668
Top-K Neurons	accuracy_best_layer	0.0	0.0
Top-K Neurons	accuracy_max	0.6689	0.6711
Top-K Neurons	accuracy_mean	0.534	0.6062
Top-K Neurons	accuracy_std	0.1208	0.0587