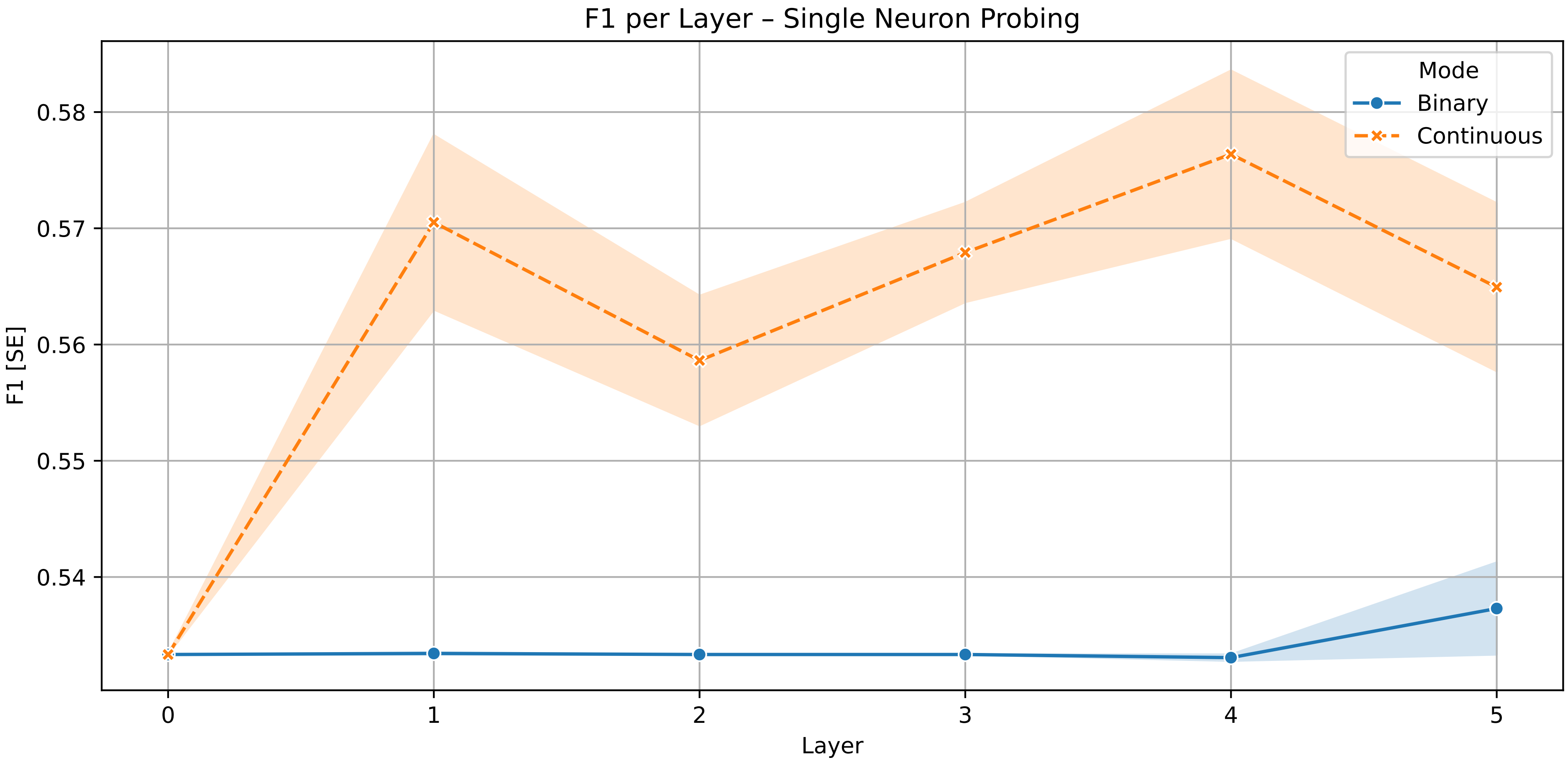
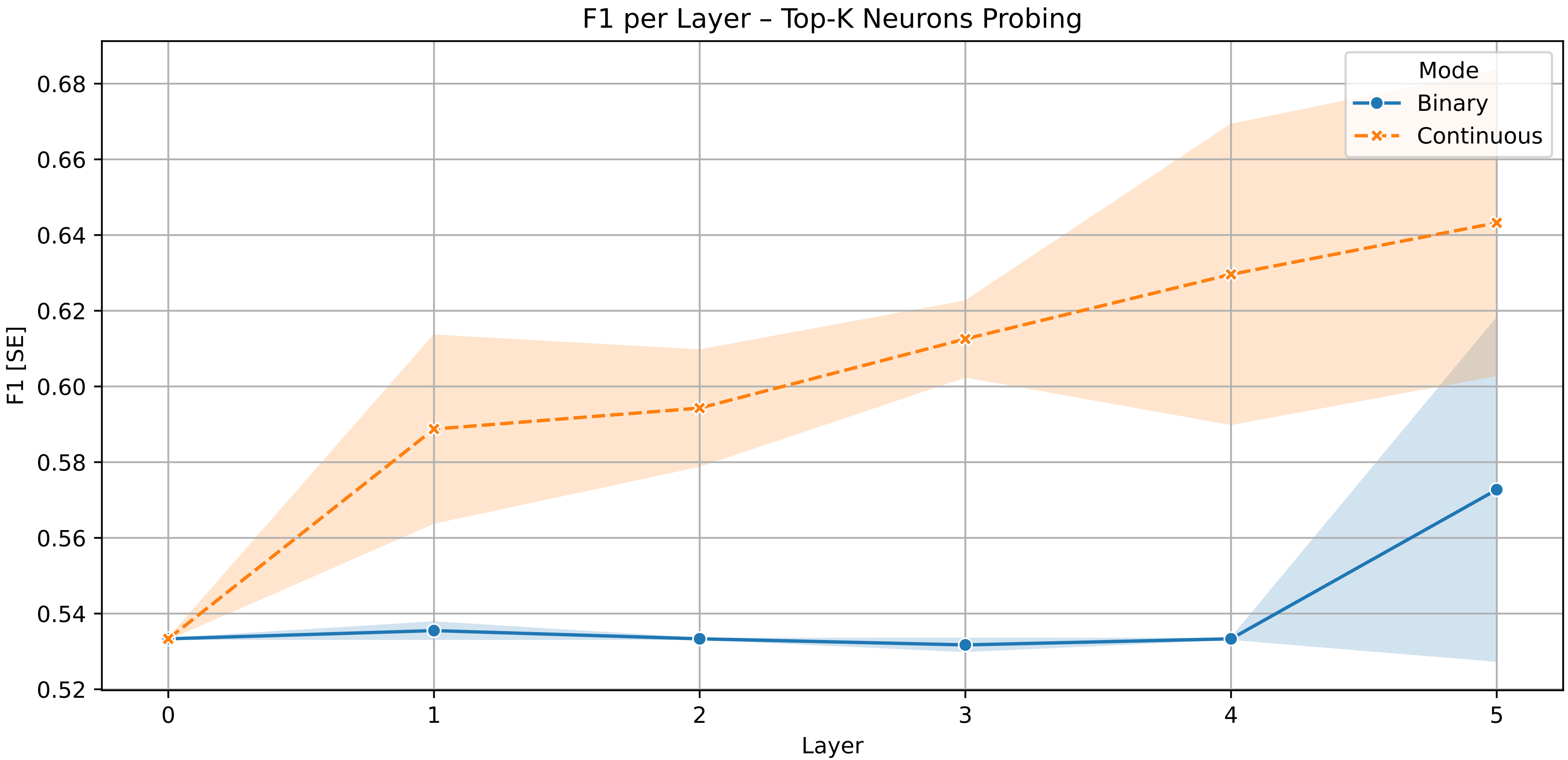
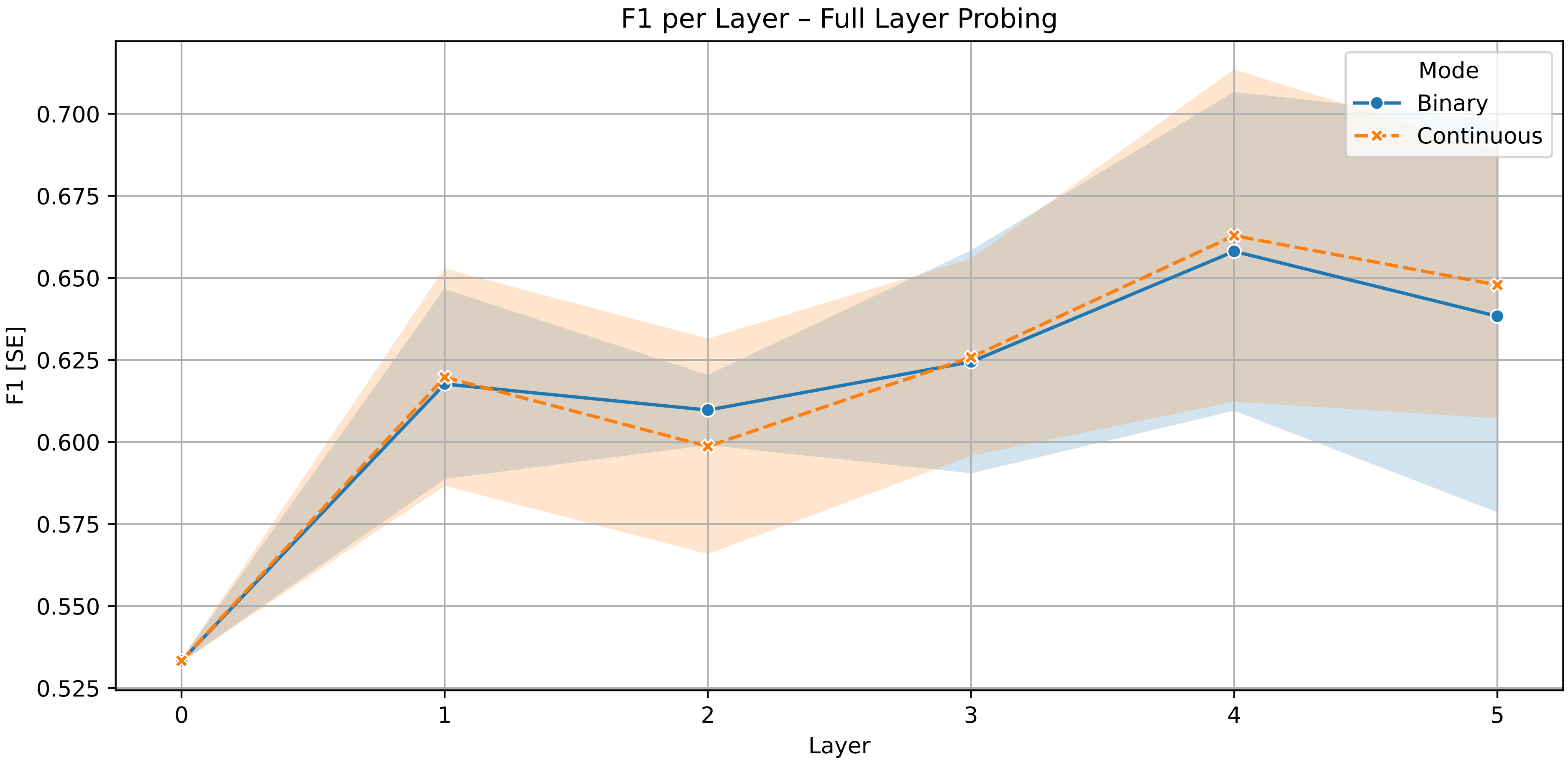
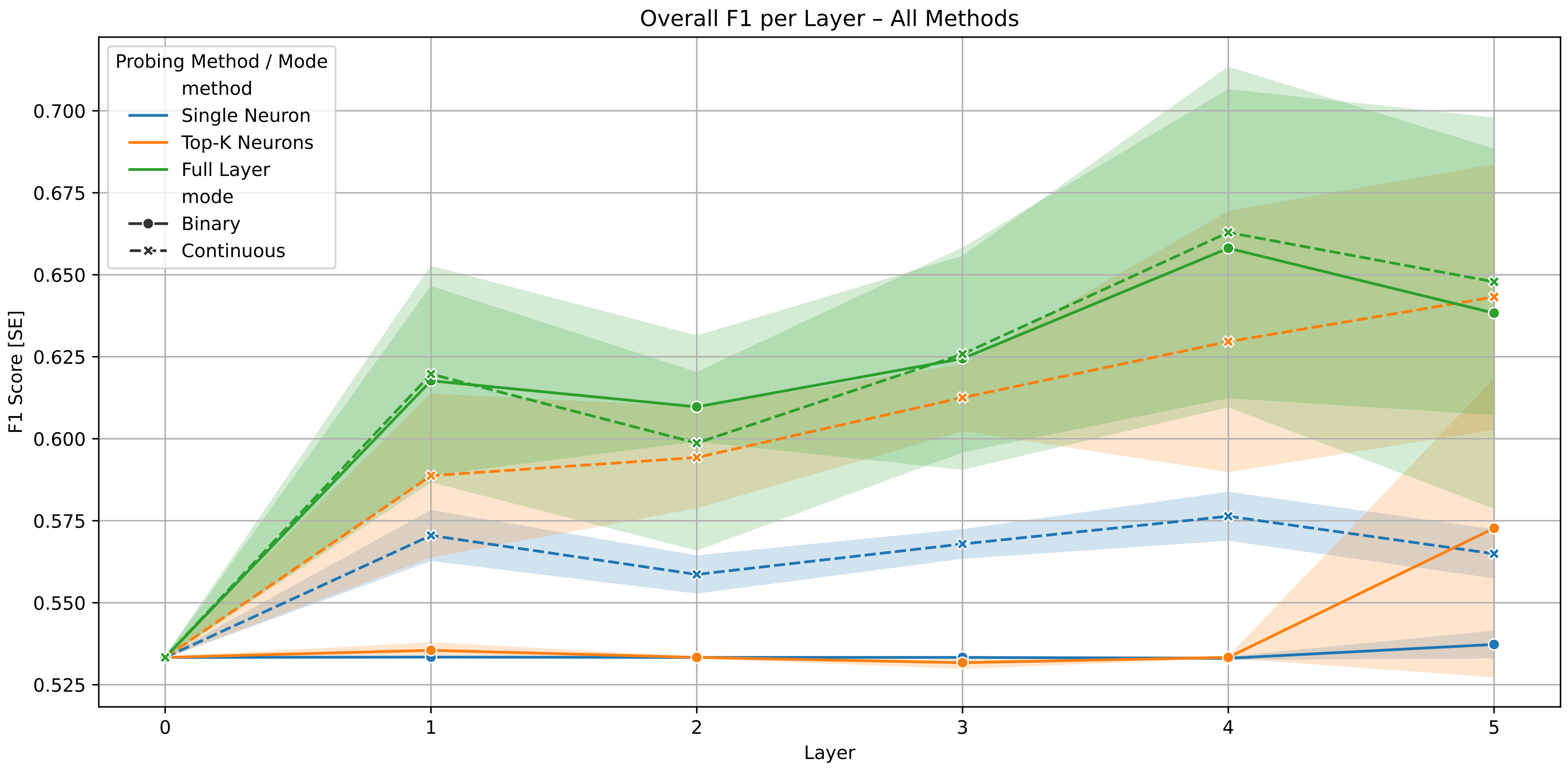

F1 per Layer – Single Neuron Probing
F1 [SE]
0.58
0.57
0.56
0.55
0.54
Layer
0
1
2
3
4
5
Mode
Binary
Continuous

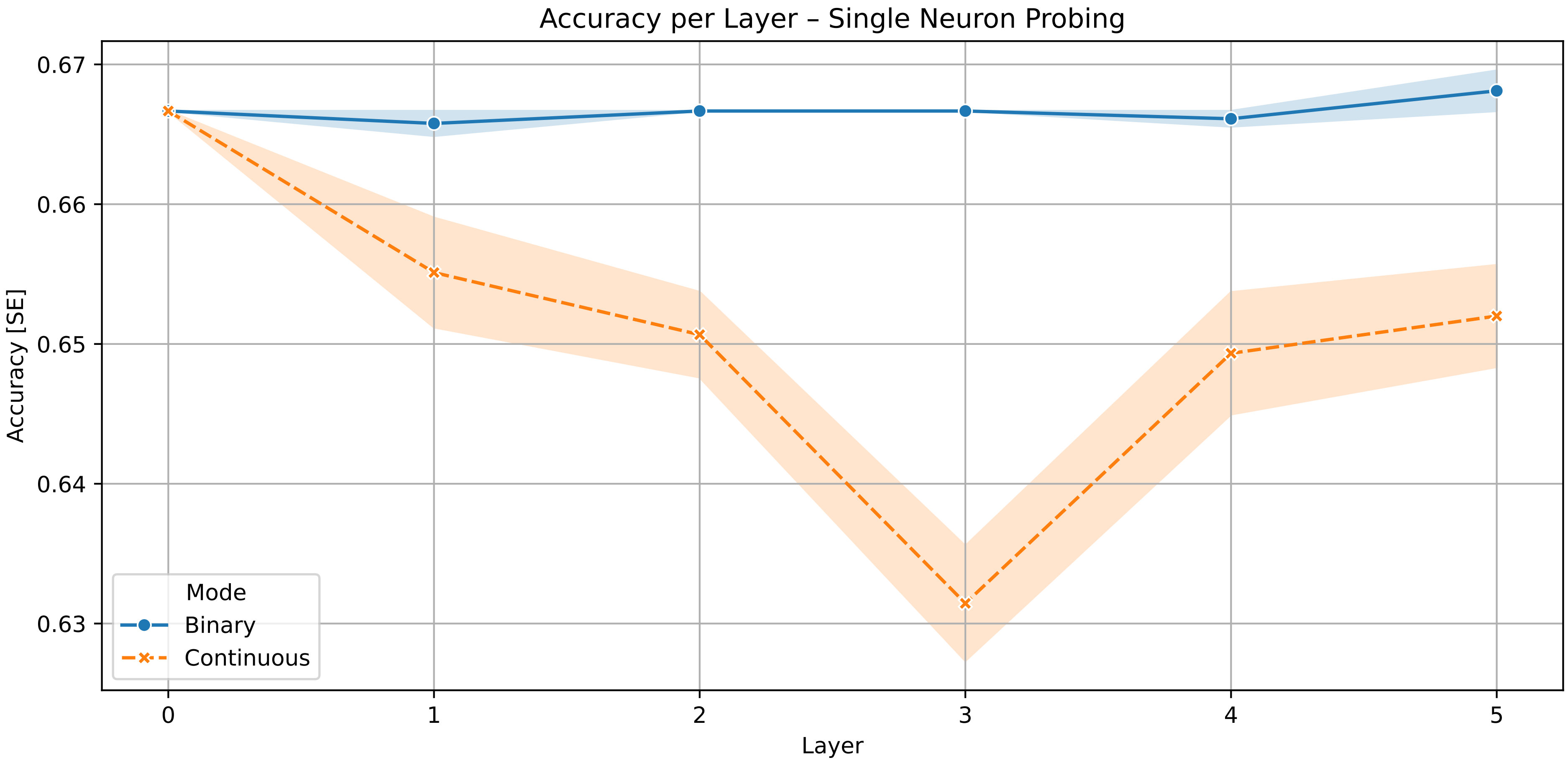

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.52
0.54
0.56
0.58
0.60
0.62
0.64
0.66
0.68
0
1
2
3
4
5
Layer

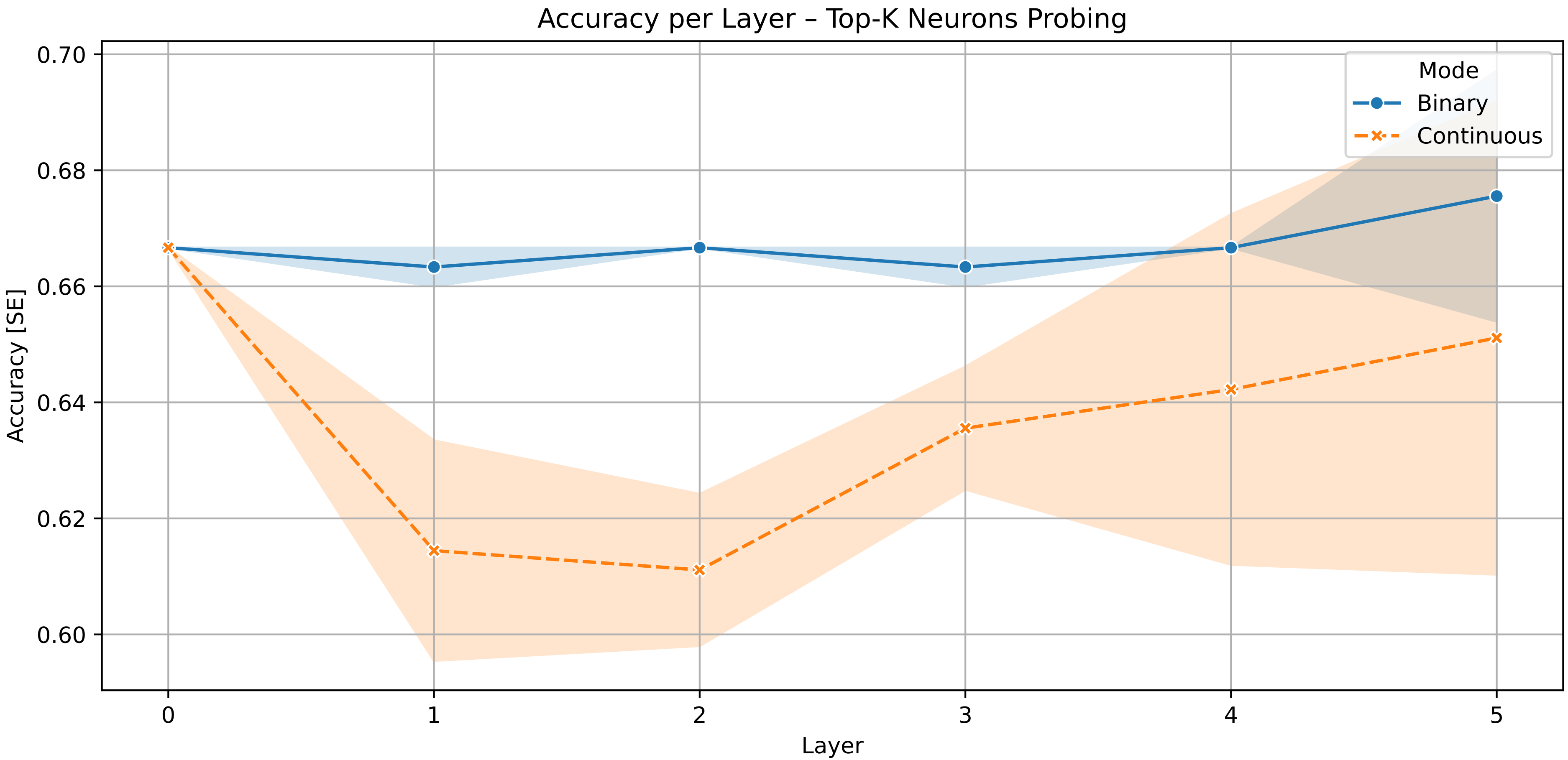

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0.525
Layer
0
1
2
3
4
5

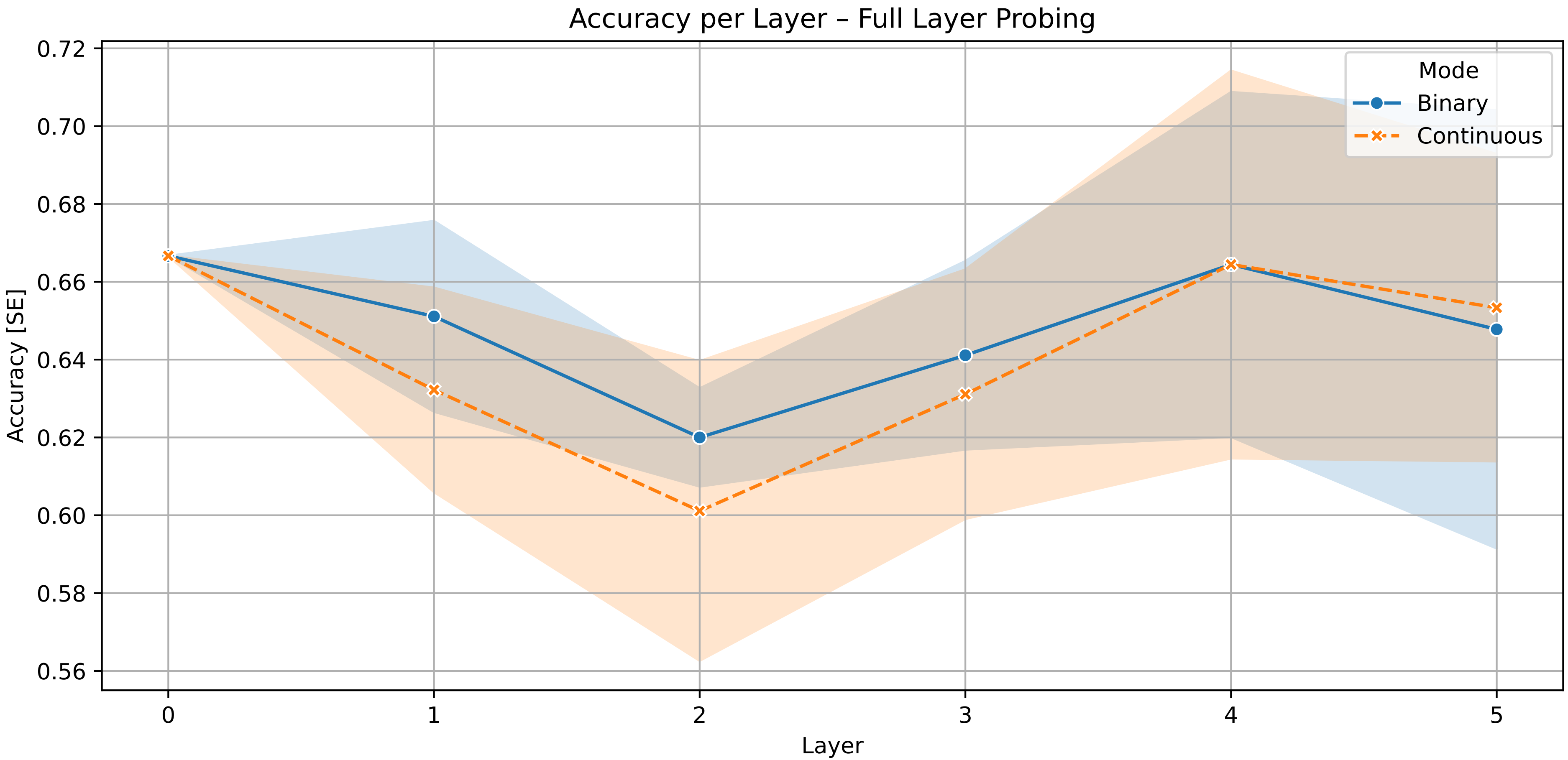

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.525
0.550
0.575
0.600
0.625
0.650
0.675
0.700
Layer
0
1
2
3
4
5

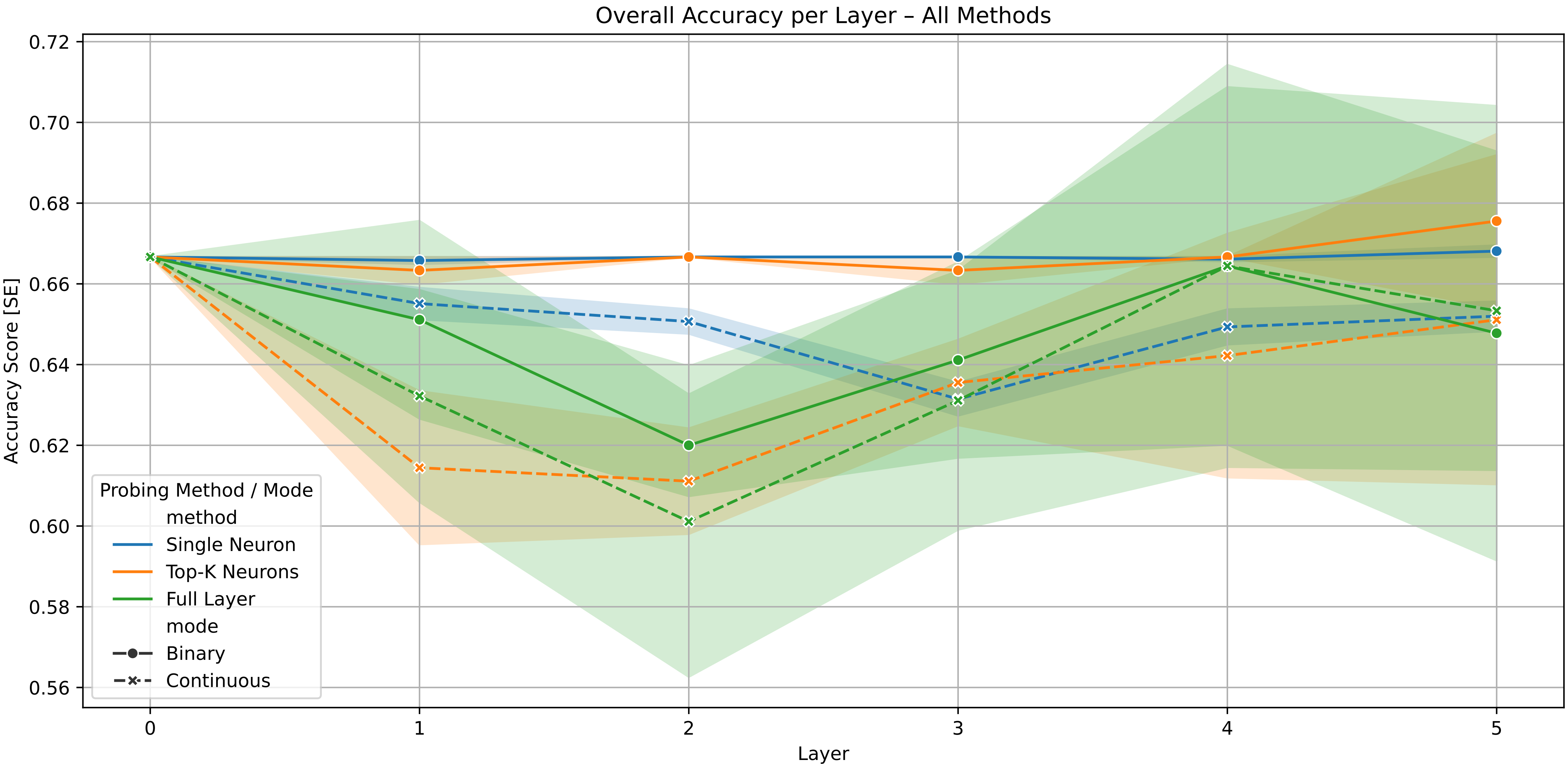
F1 Score Summary by Probing Method

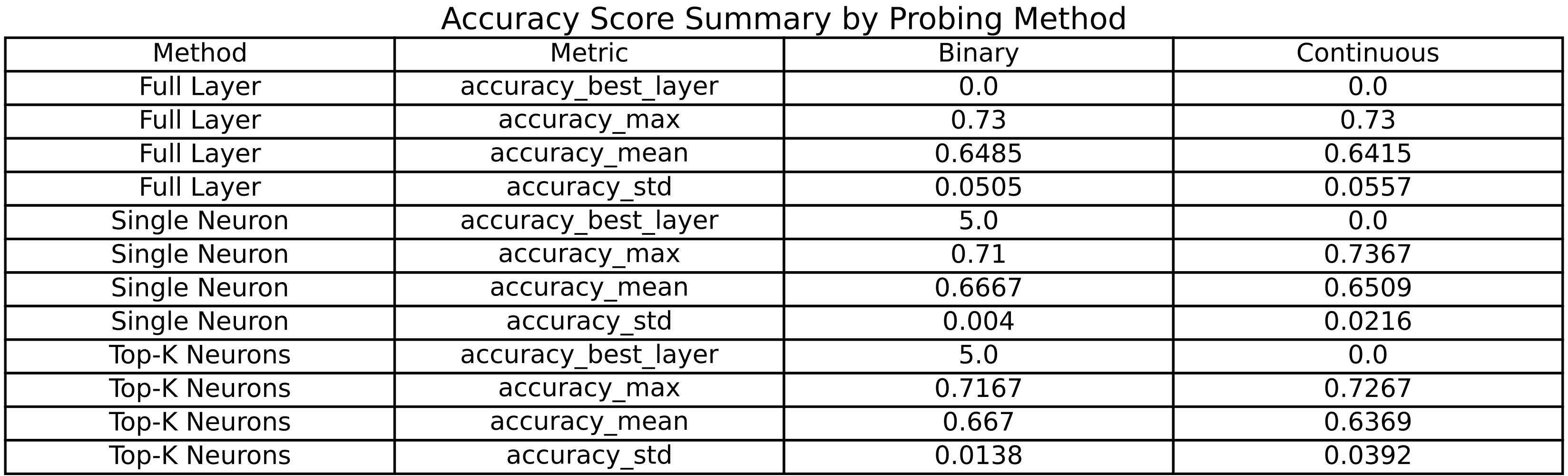
Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	4.0	4.0
Full Layer	f1_max	0.7303	0.7297
Full Layer	f1_mean	0.6136	0.6147
Full Layer	f1_std	0.0664	0.0661
Single Neuron	f1_best_layer	5.0	4.0
Single Neuron	f1_max	0.6521	0.7118
Single Neuron	f1_mean	0.534	0.562
Single Neuron	f1_std	0.0089	0.0348
Top-K Neurons	f1_best_layer	5.0	5.0
Top-K Neurons	f1_max	0.6629	0.719
Top-K Neurons	f1_mean	0.54	0.6003
Top-K Neurons	f1_std	0.0309	0.0527


Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
0.67
0.66
0.65
0.64
0.63
Layer
0
1
2
3
4
5
Mode
Binary
Continuous


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.70
0.68
0.66
0.64
0.62
0.60
Layer
0
1
2
3
4
5


Accuracy per Layer – Full Layer Probing
Accuracy [SE]
0.72
0.70
0.68
0.66
0.64
0.62
0.60
0.58
0.56
Layer
0
1
2
3
4
5
Mode
Binary
Continuous


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
0.72
0.70
0.68
0.66
0.64
0.62
0.60
0.58
0.56
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	0.0	0.0
Full Layer	accuracy_max	0.73	0.73
Full Layer	accuracy_mean	0.6485	0.6415
Full Layer	accuracy_std	0.0505	0.0557
Single Neuron	accuracy_best_layer	5.0	0.0
Single Neuron	accuracy_max	0.71	0.7367
Single Neuron	accuracy_mean	0.6667	0.6509
Single Neuron	accuracy_std	0.004	0.0216
Top-K Neurons	accuracy_best_layer	5.0	0.0
Top-K Neurons	accuracy_max	0.7167	0.7267
Top-K Neurons	accuracy_mean	0.667	0.6369
Top-K Neurons	accuracy_std	0.0138	0.0392