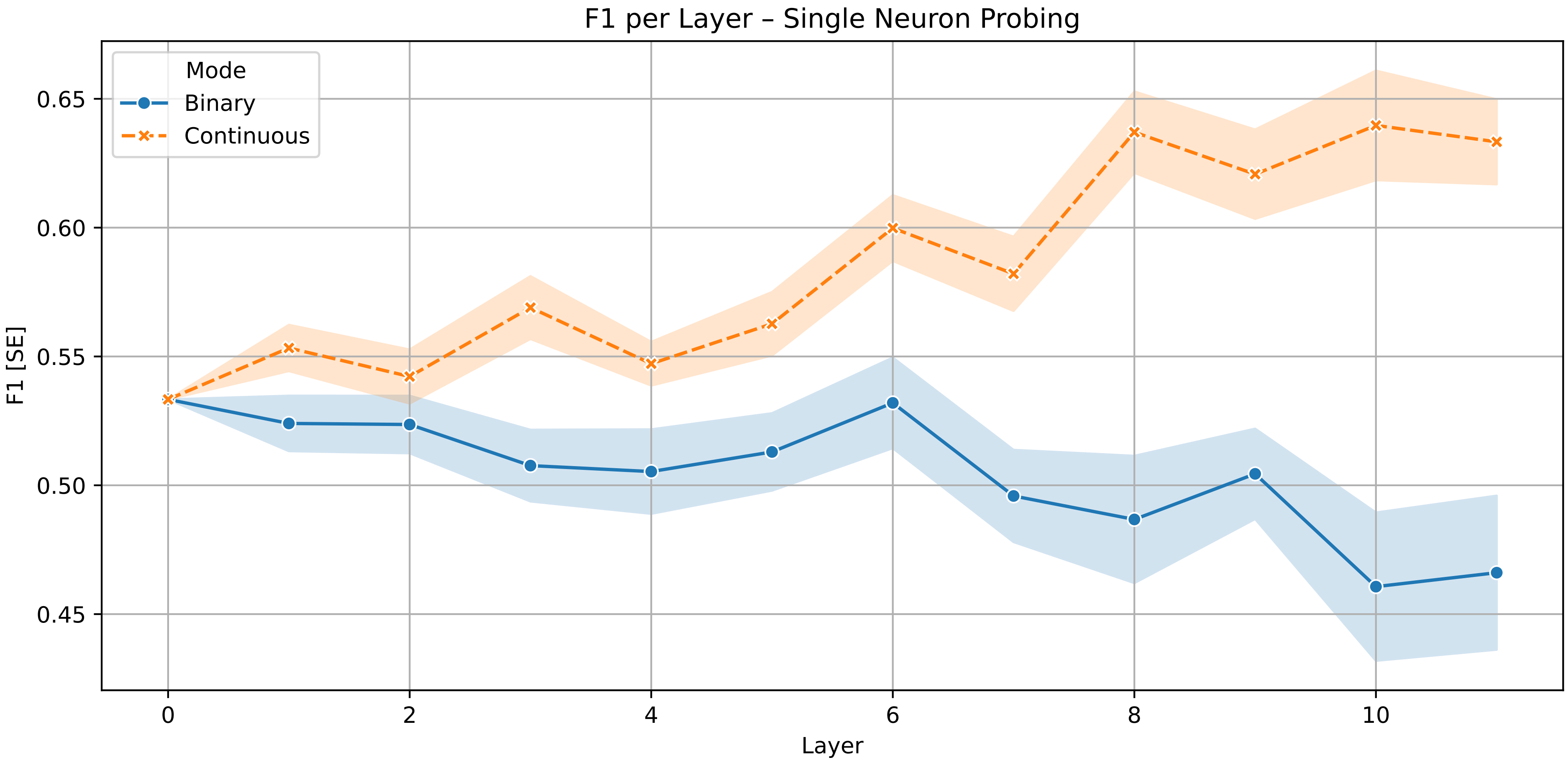
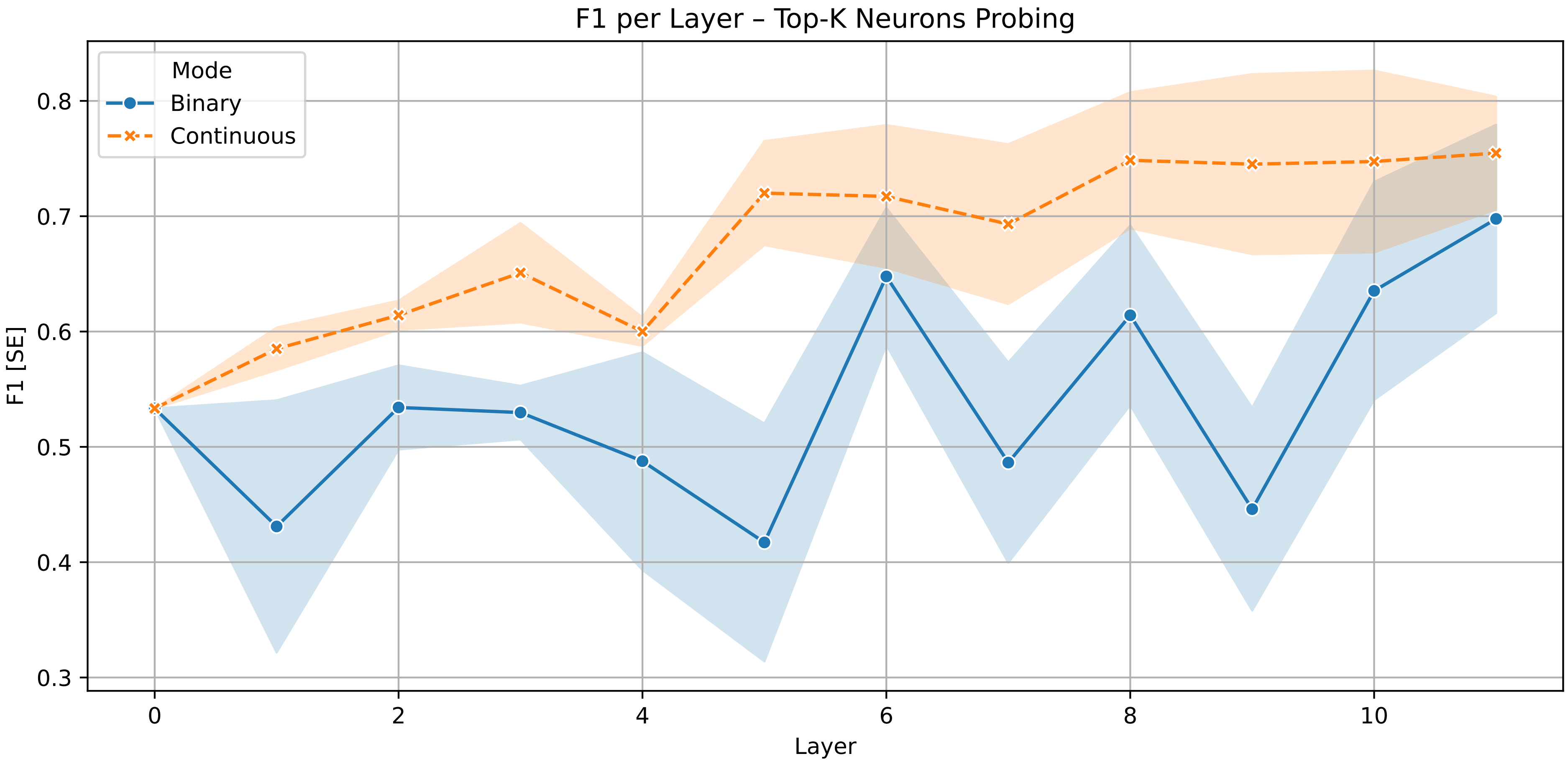
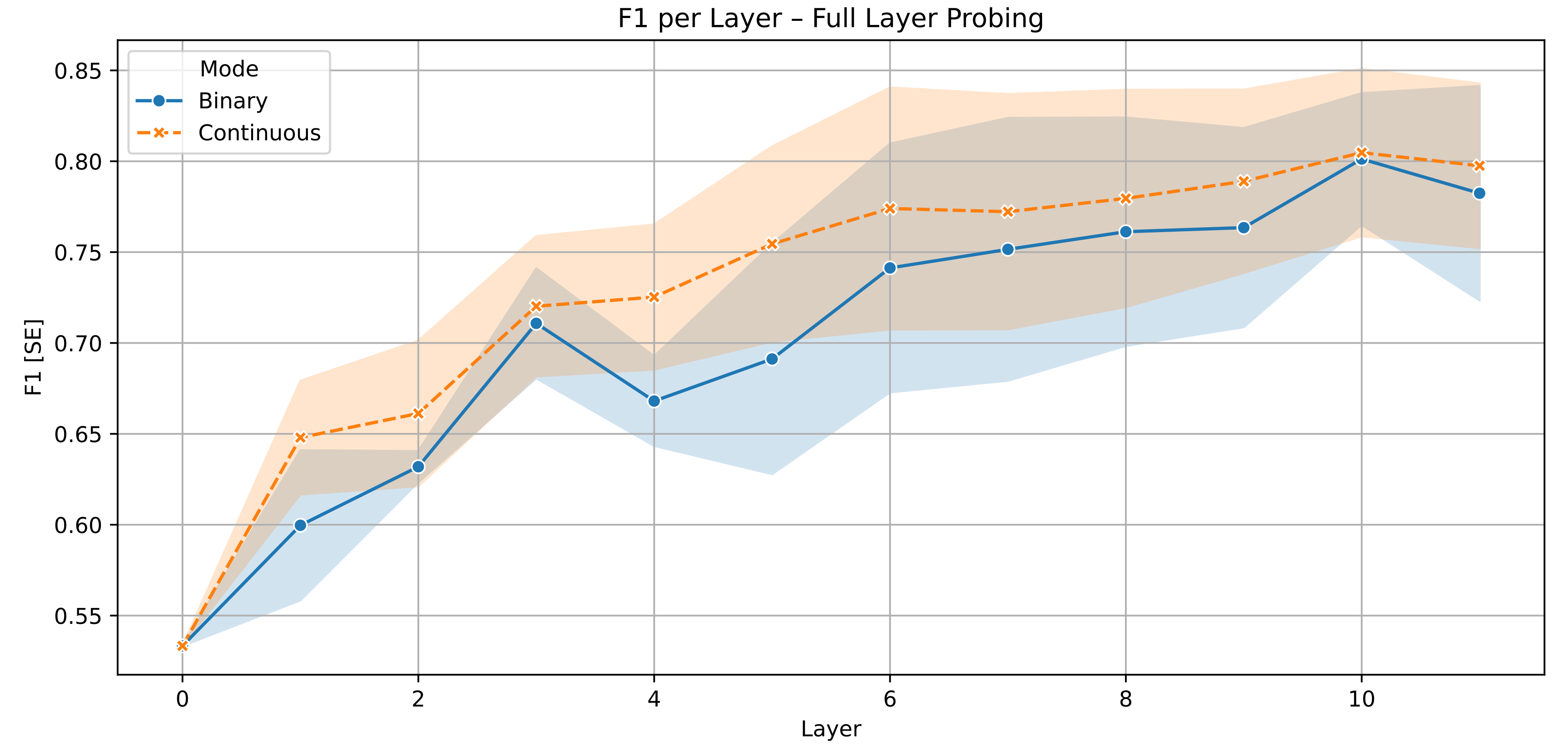
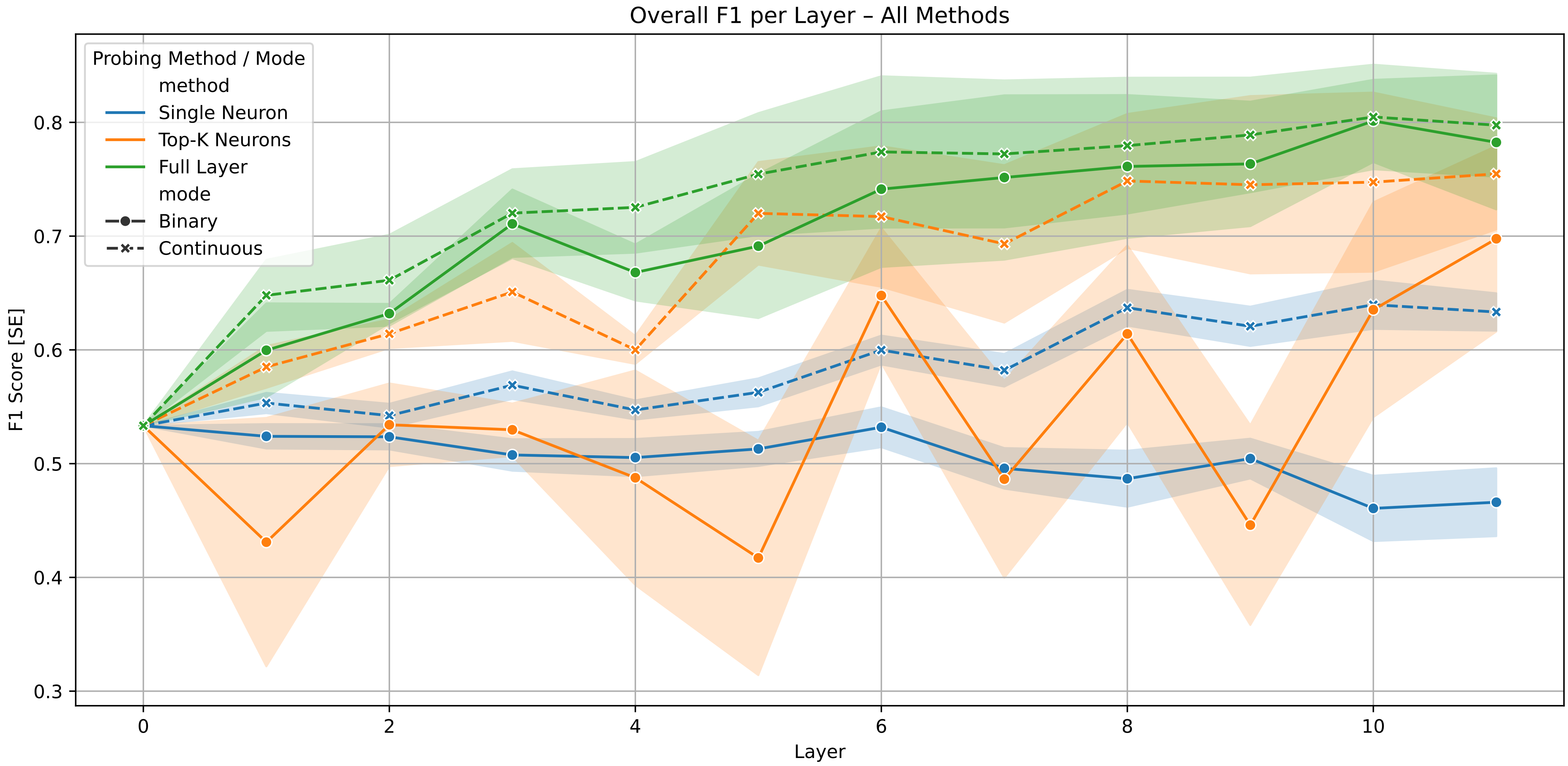
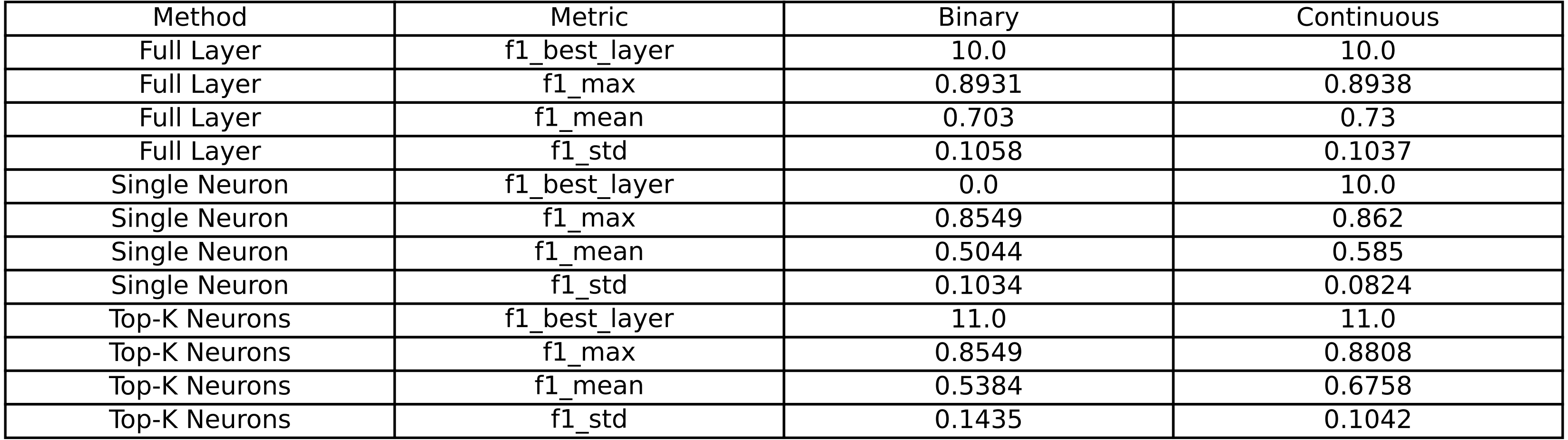

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.65
0.60
0.55
0.50
0.45
Layer
0
2
4
6
8
10


F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.3
0.4
0.5
0.6
0.7
0.8
Layer
0
2
4
6
8
10

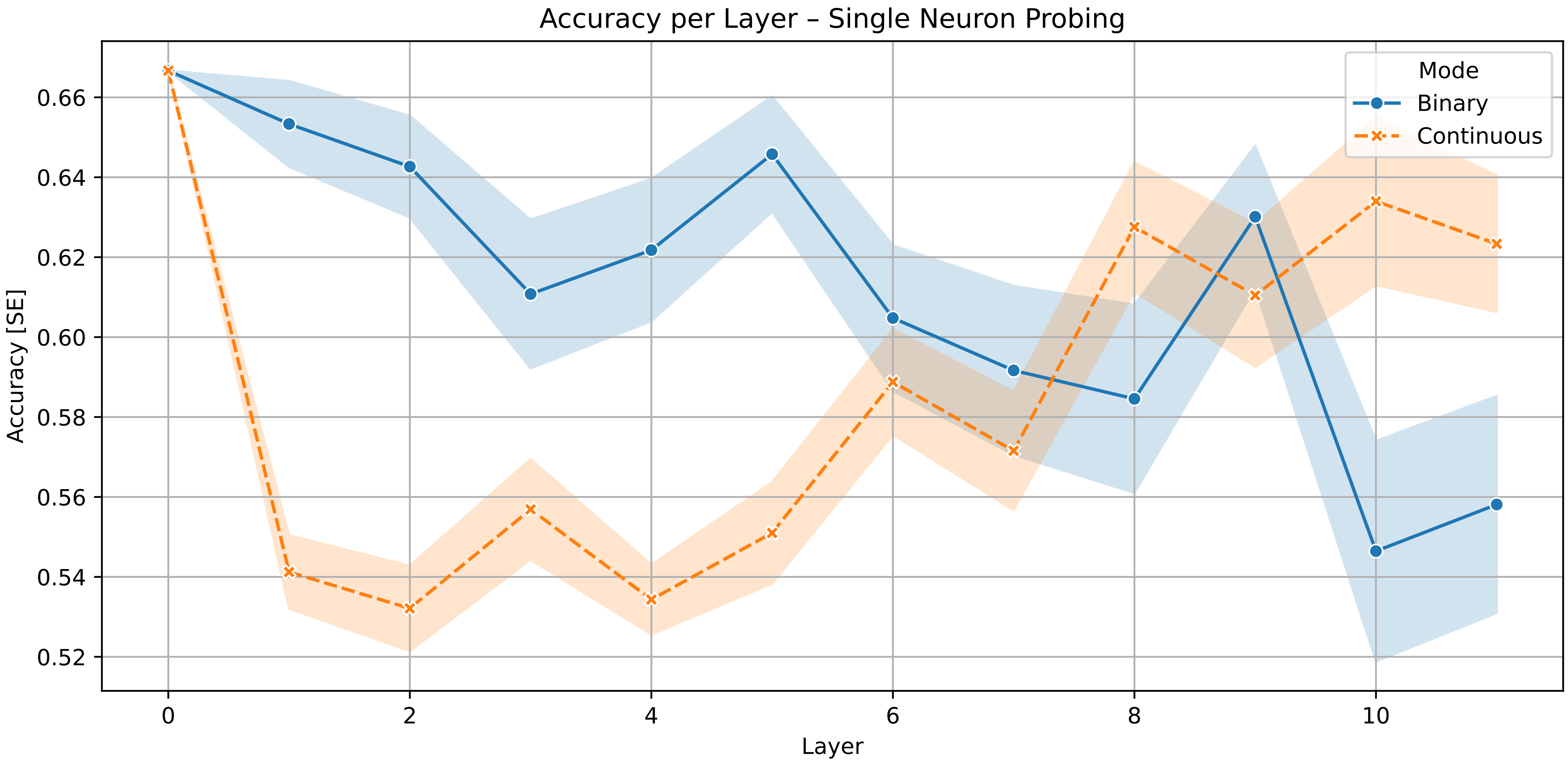

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.85
0.80
0.75
0.70
0.65
0.60
0.55
Layer
0
2
4
6
8
10

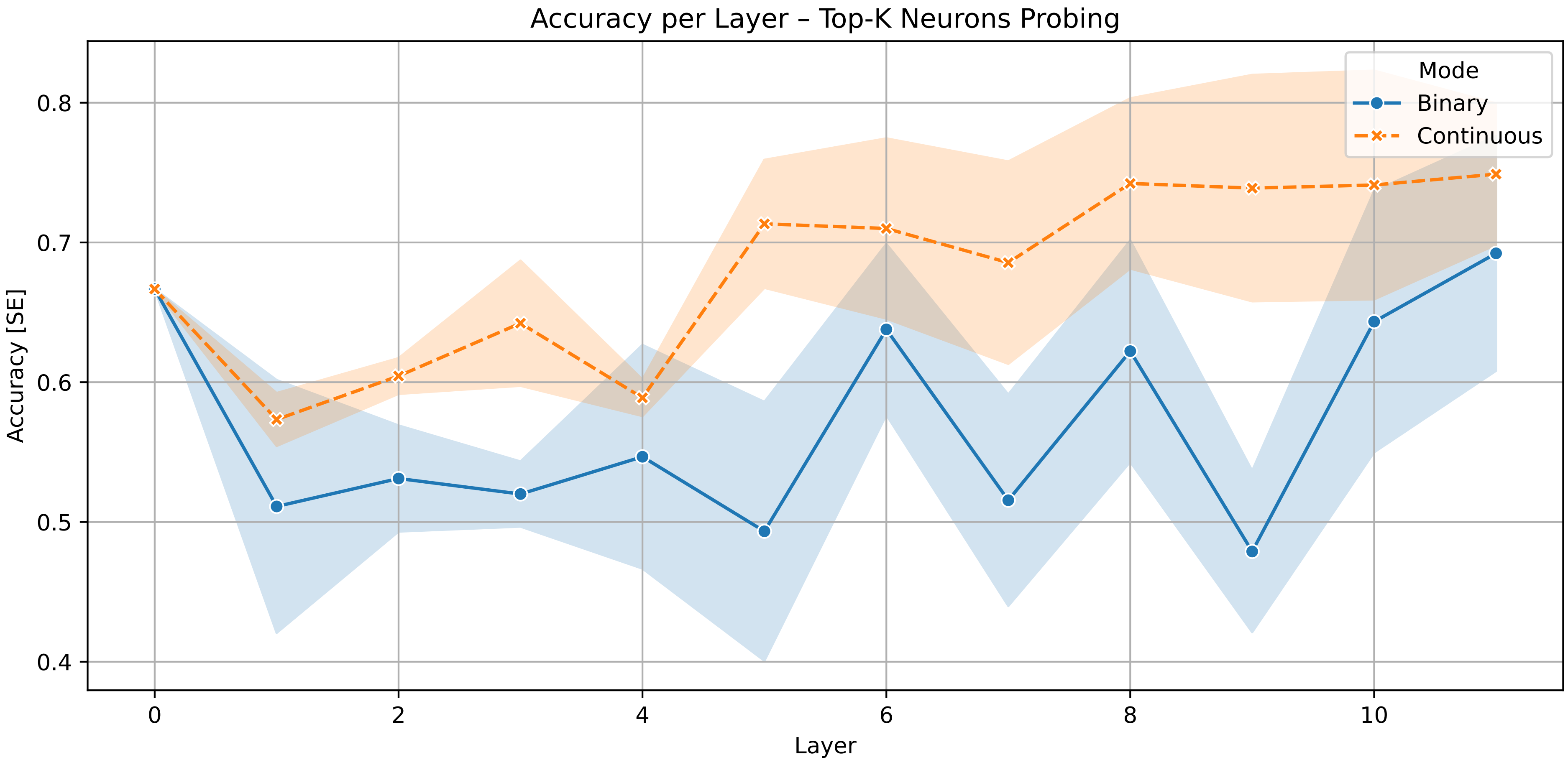

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.3
0.4
0.5
0.6
0.7
0.8
Layer
0
2
4
6
8
10

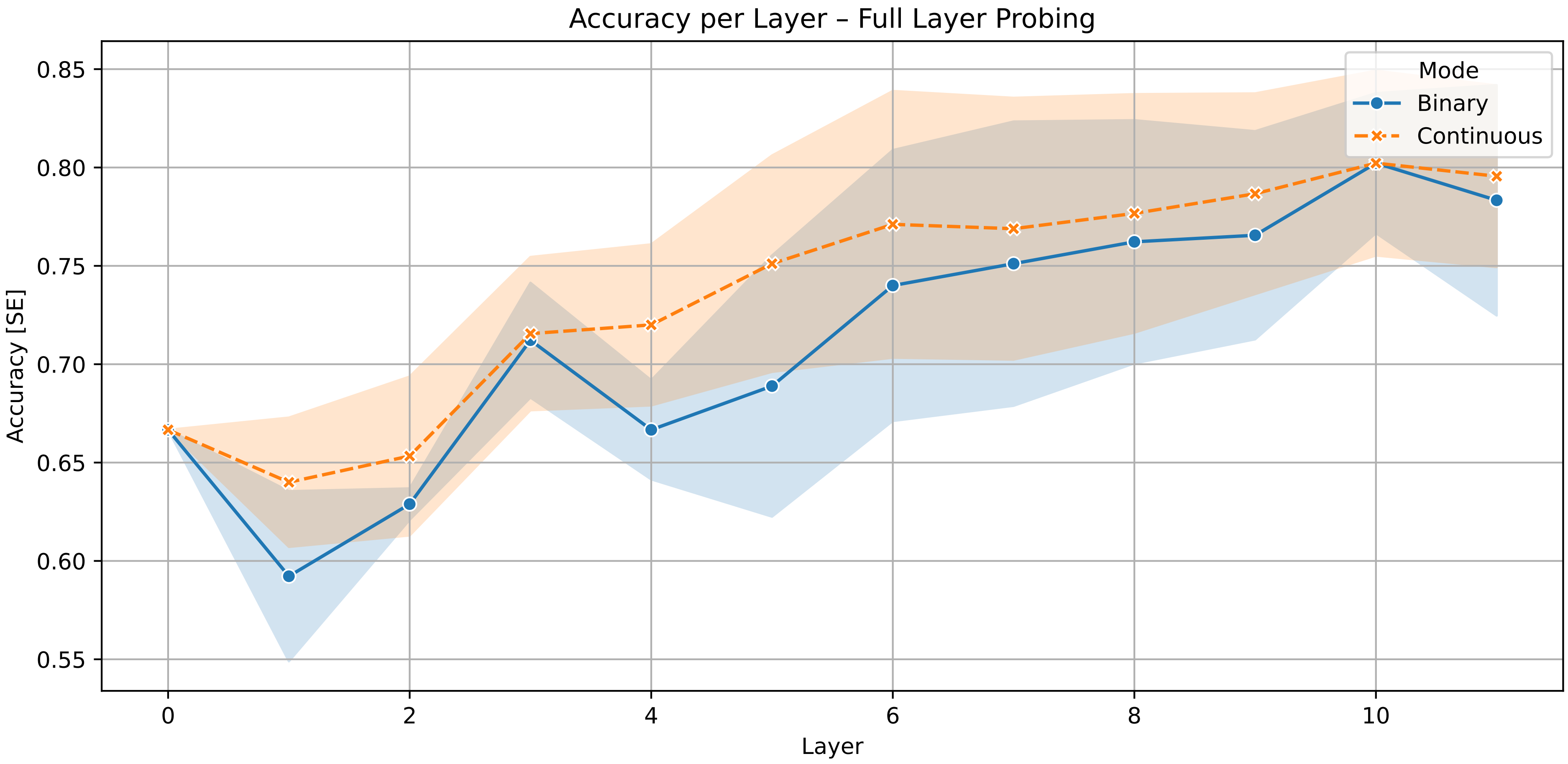
F1 Score Summary by Probing Method

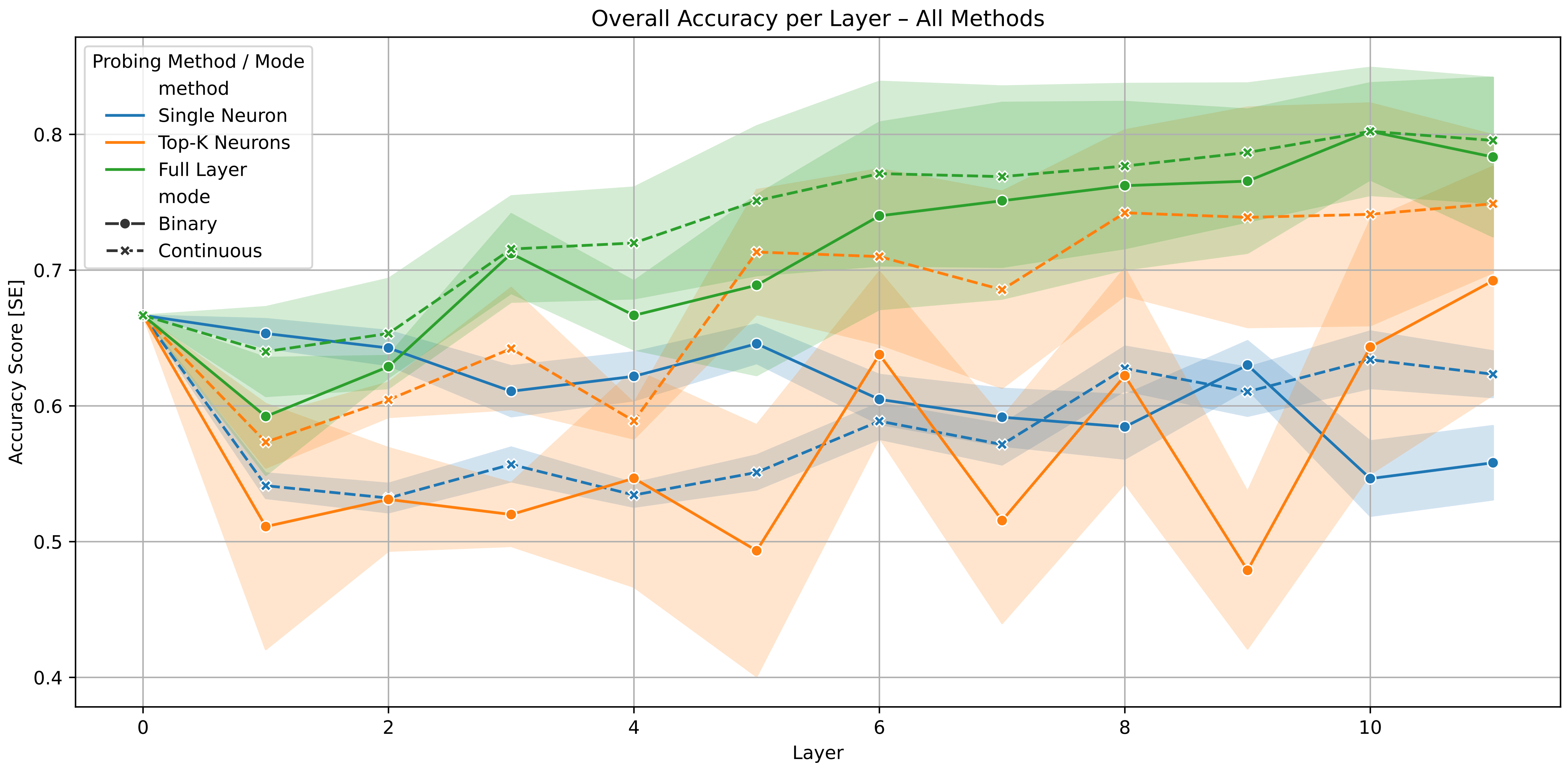


Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	10.0	10.0
Full Layer	f1_max	0.8931	0.8938
Full Layer	f1_mean	0.703	0.73
Full Layer	f1_std	0.1058	0.1037
Single Neuron	f1_best_layer	0.0	10.0
Single Neuron	f1_max	0.8549	0.862
Single Neuron	f1_mean	0.5044	0.585
Single Neuron	f1_std	0.1034	0.0824
Top-K Neurons	f1_best_layer	11.0	11.0
Top-K Neurons	f1_max	0.8549	0.8808
Top-K Neurons	f1_mean	0.5384	0.6758
Top-K Neurons	f1_std	0.1435	0.1042

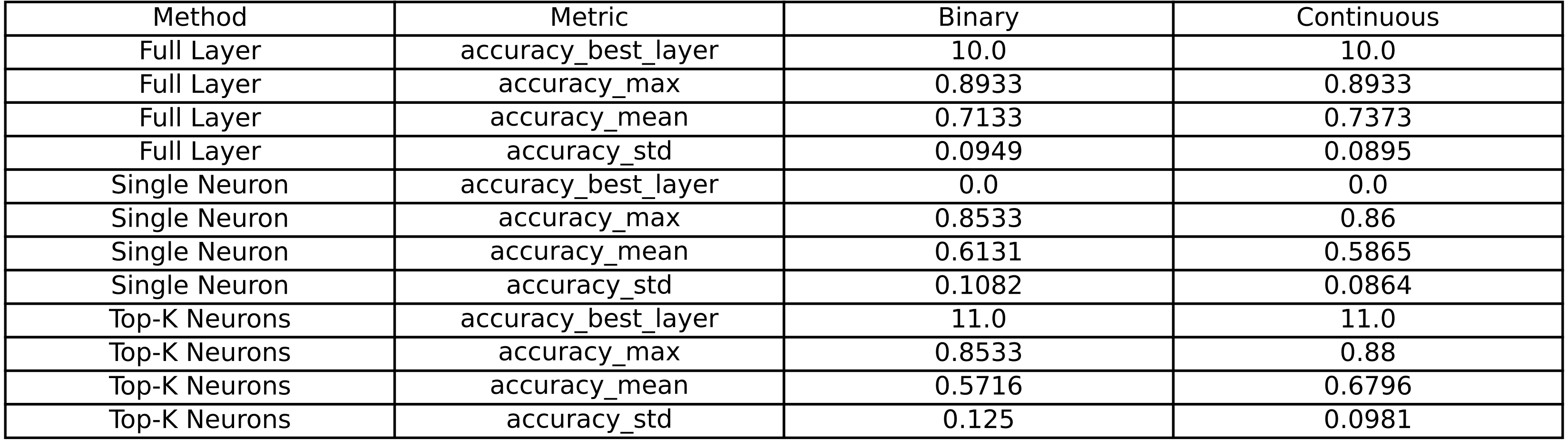

Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
Layer
0
2
4
6
8
10
Mode
Binary
Continuous


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.4
0.5
0.6
0.7
0.8
Layer
0
2
4
6
8
10


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.55
0.60
0.65
0.70
0.75
0.80
0.85
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]
0.4
0.5
0.6
0.7
0.8
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	0.8933	0.8933
Full Layer	accuracy_mean	0.7133	0.7373
Full Layer	accuracy_std	0.0949	0.0895
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.8533	0.86
Single Neuron	accuracy_mean	0.6131	0.5865
Single Neuron	accuracy_std	0.1082	0.0864
Top-K Neurons	accuracy_best_layer	11.0	11.0
Top-K Neurons	accuracy_max	0.8533	0.88
Top-K Neurons	accuracy_mean	0.5716	0.6796
Top-K Neurons	accuracy_std	0.125	0.0981