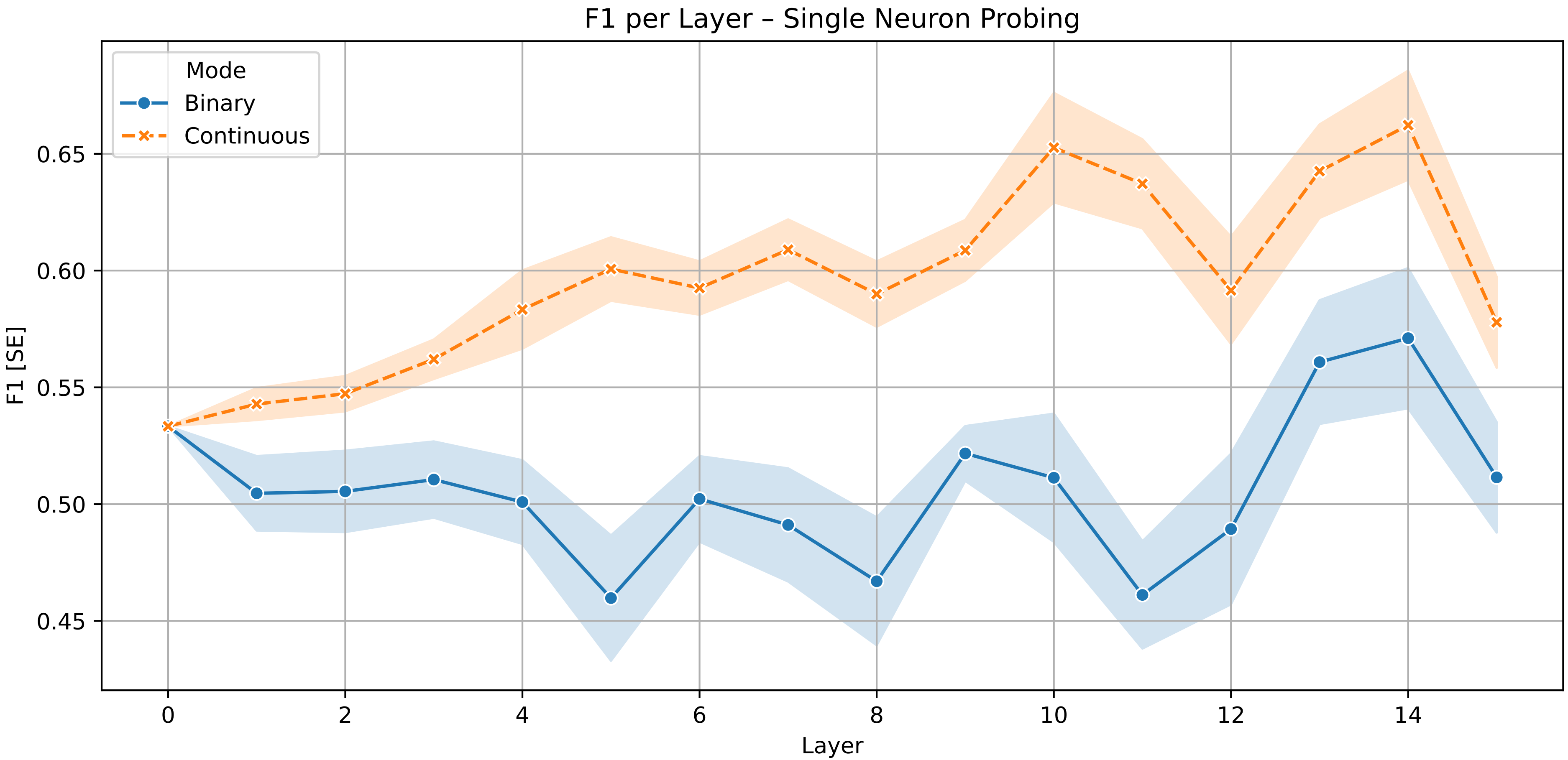
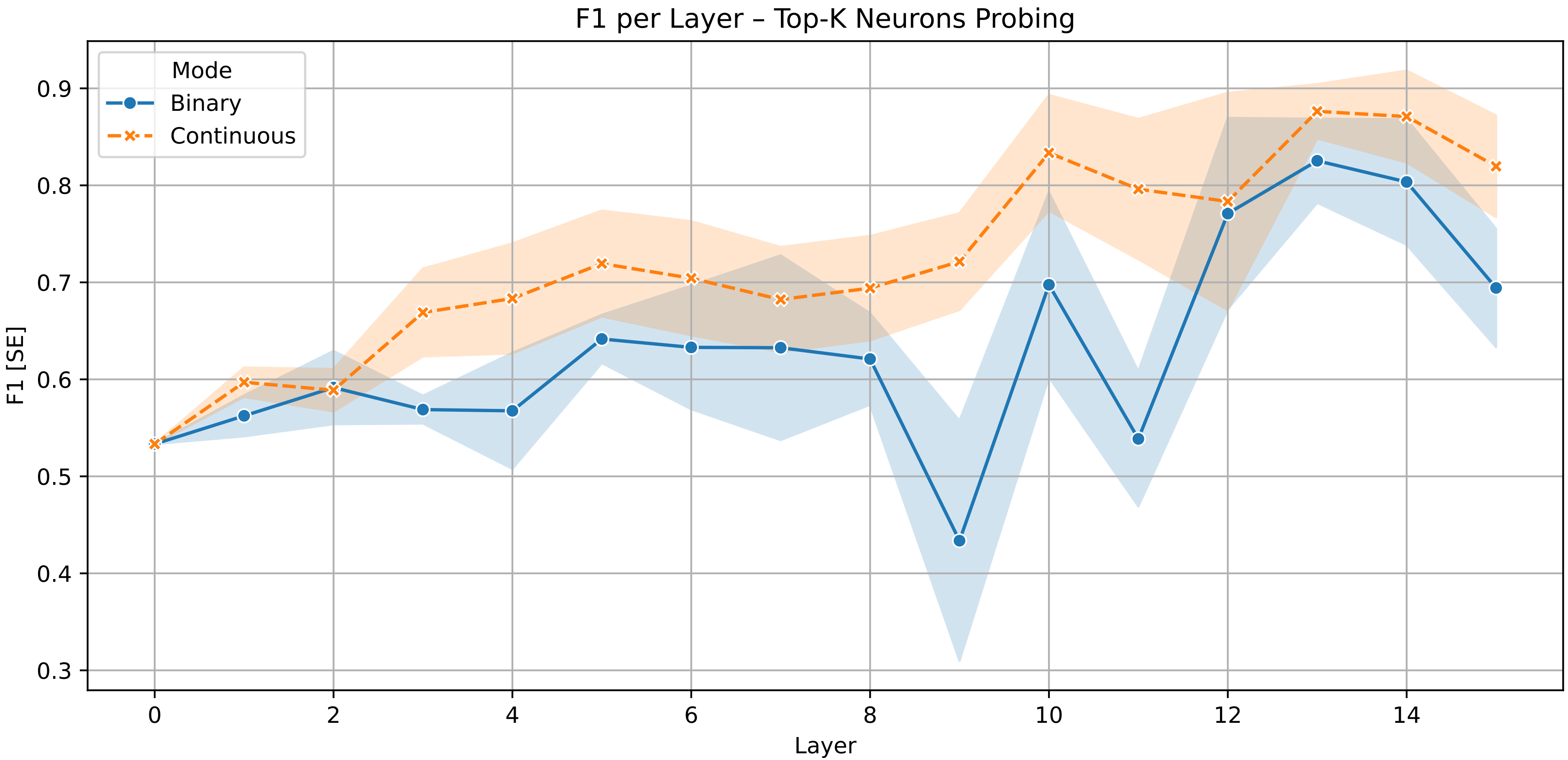
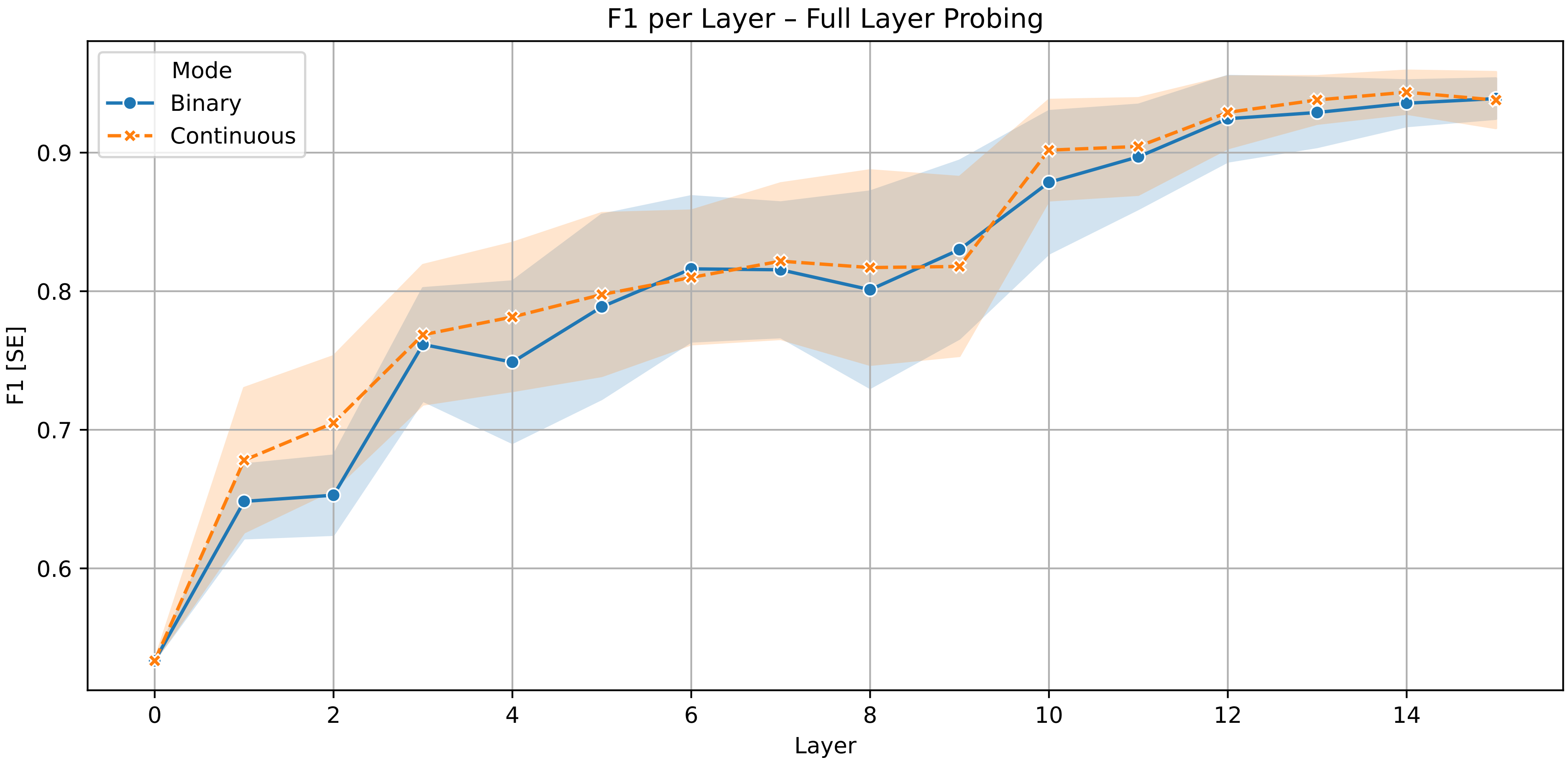
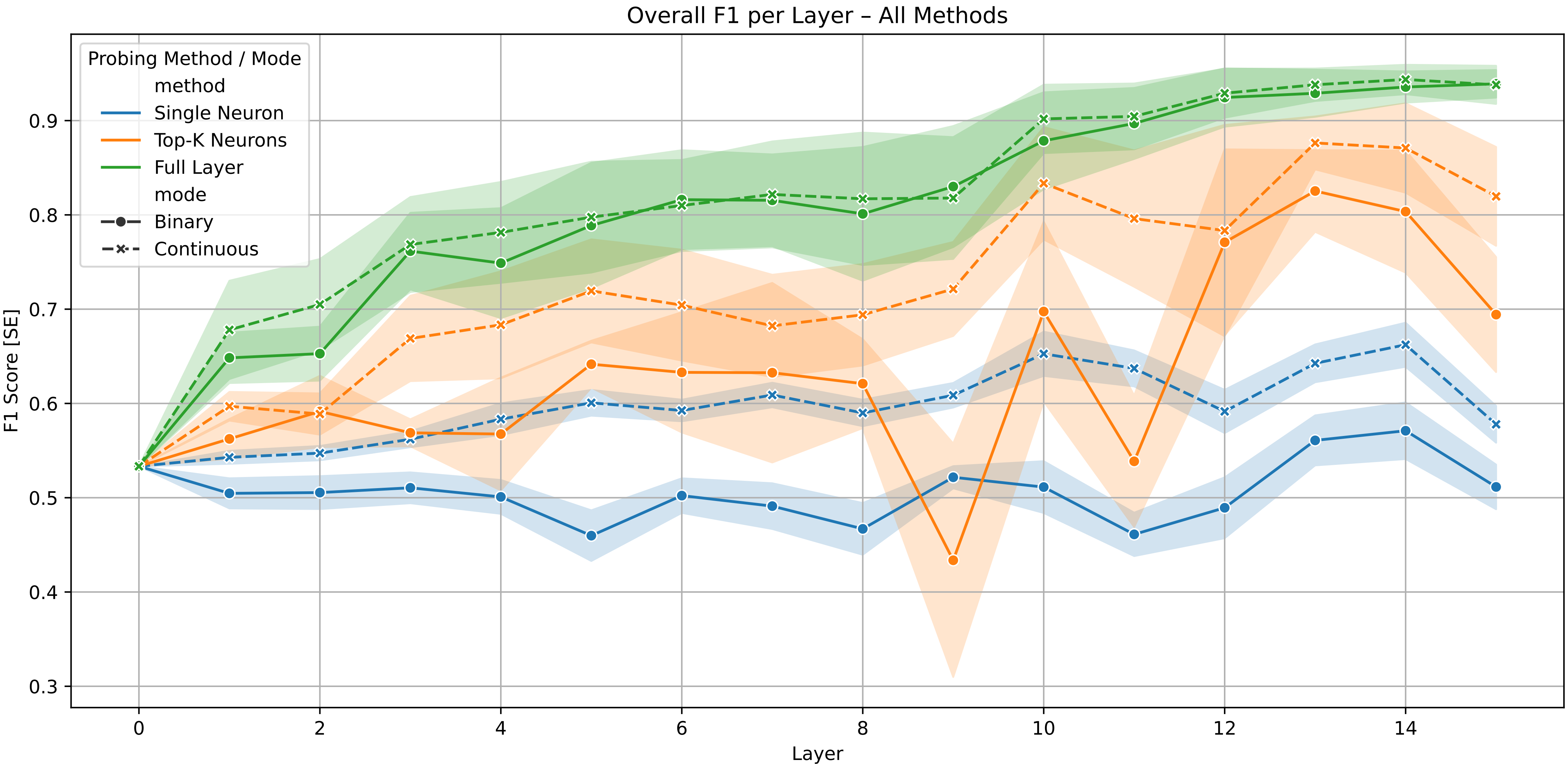
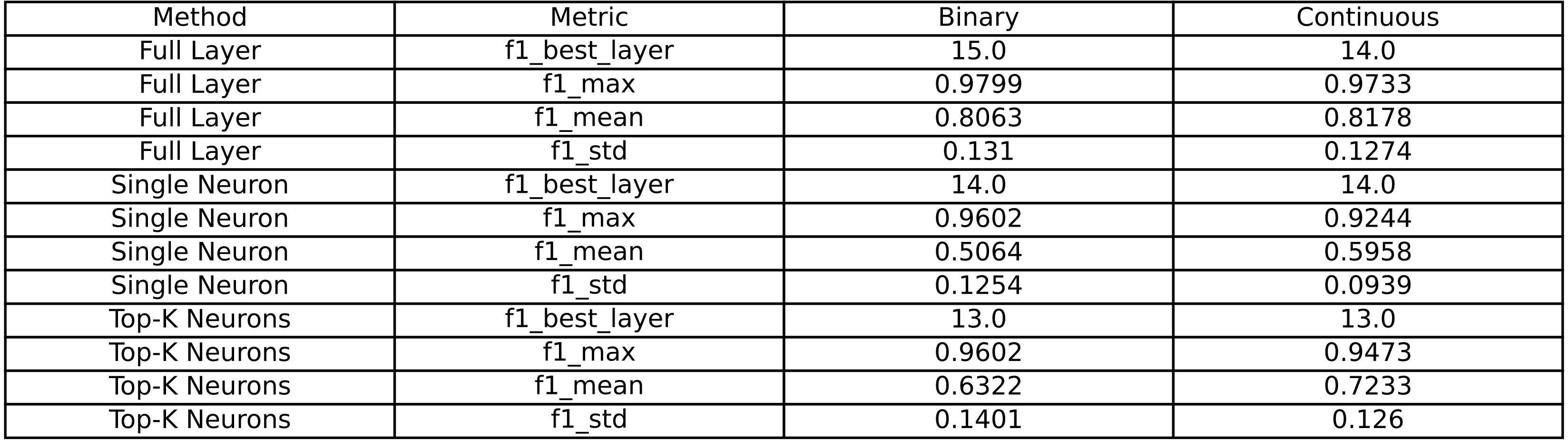

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.45
0.50
0.55
0.60
0.65
Layer
0
2
4
6
8
10
12
14


F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Layer
0
2
4
6
8
10
12
14

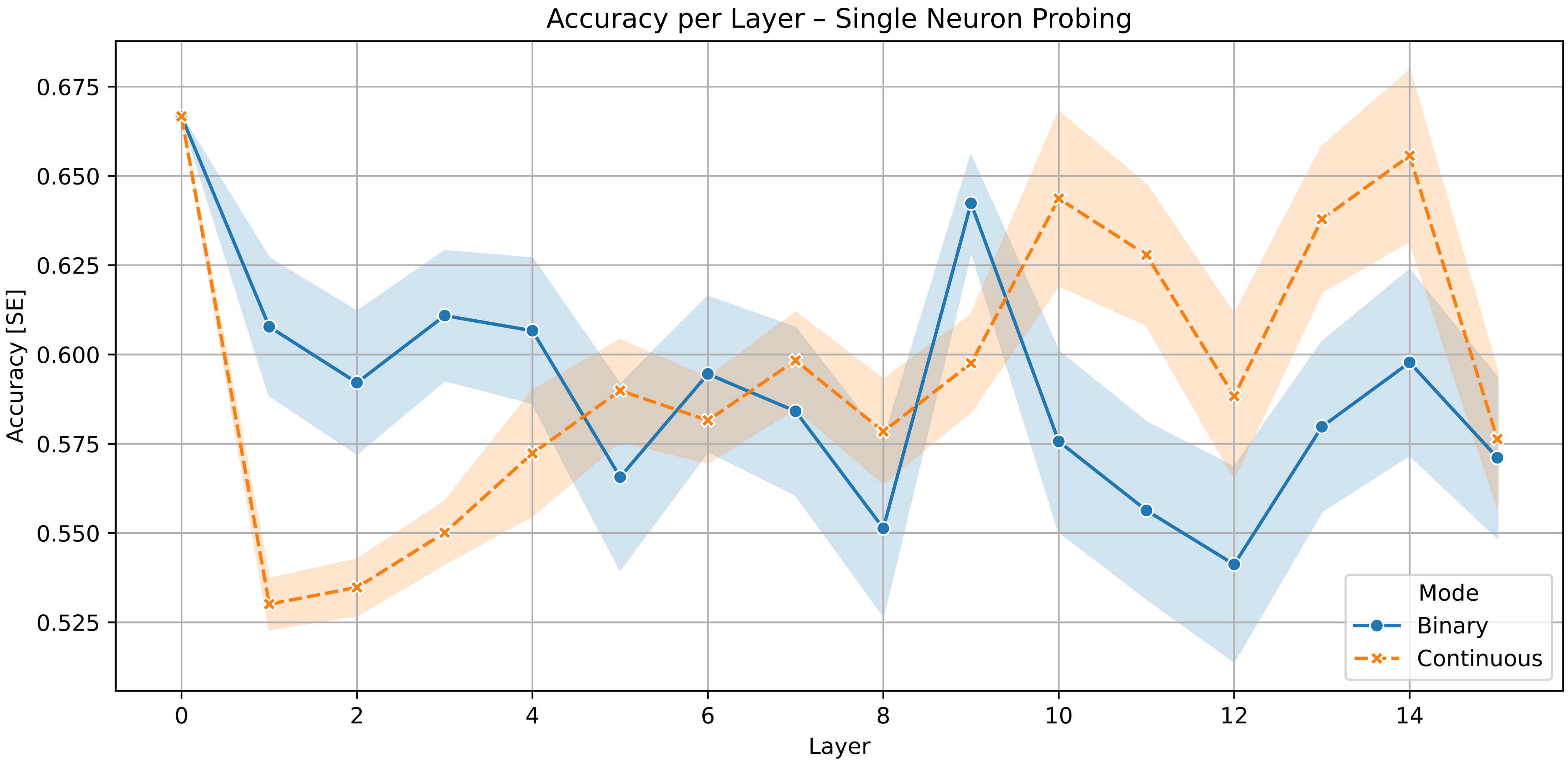

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.6
0.7
0.8
0.9
Layer
0
2
4
6
8
10
12
14

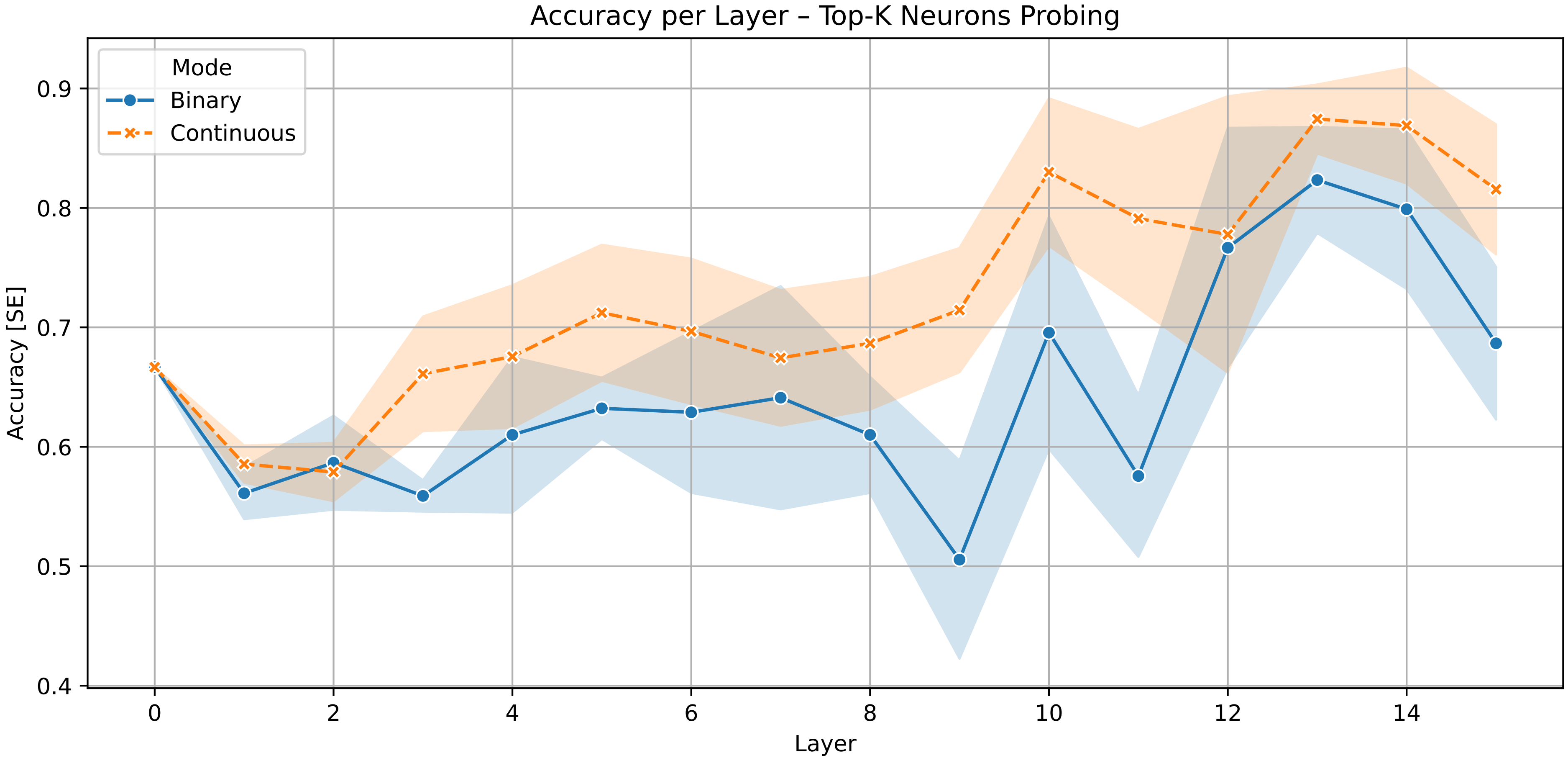

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
2
4
6
8
10
12
14
Layer

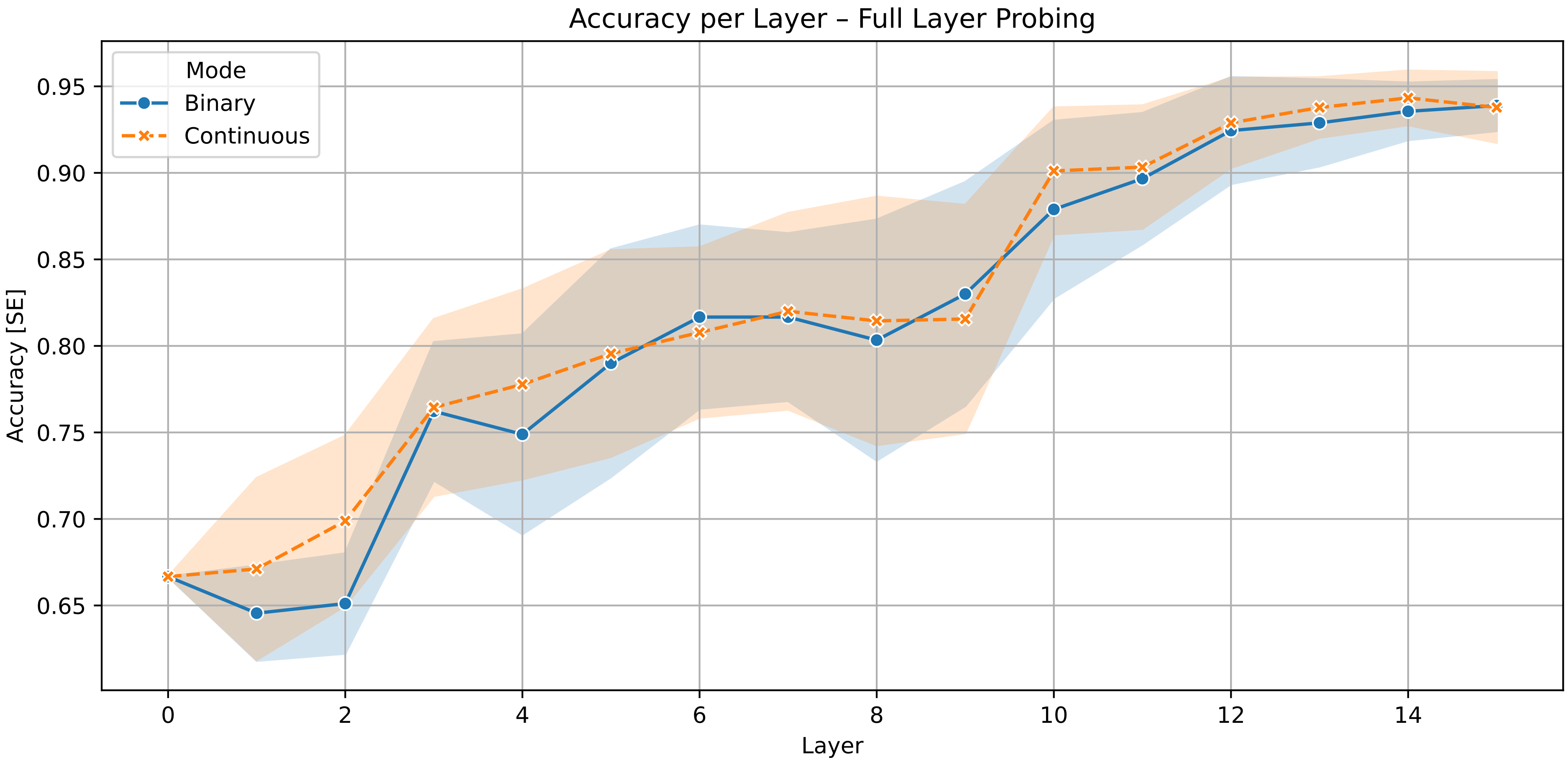
F1 Score Summary by Probing Method

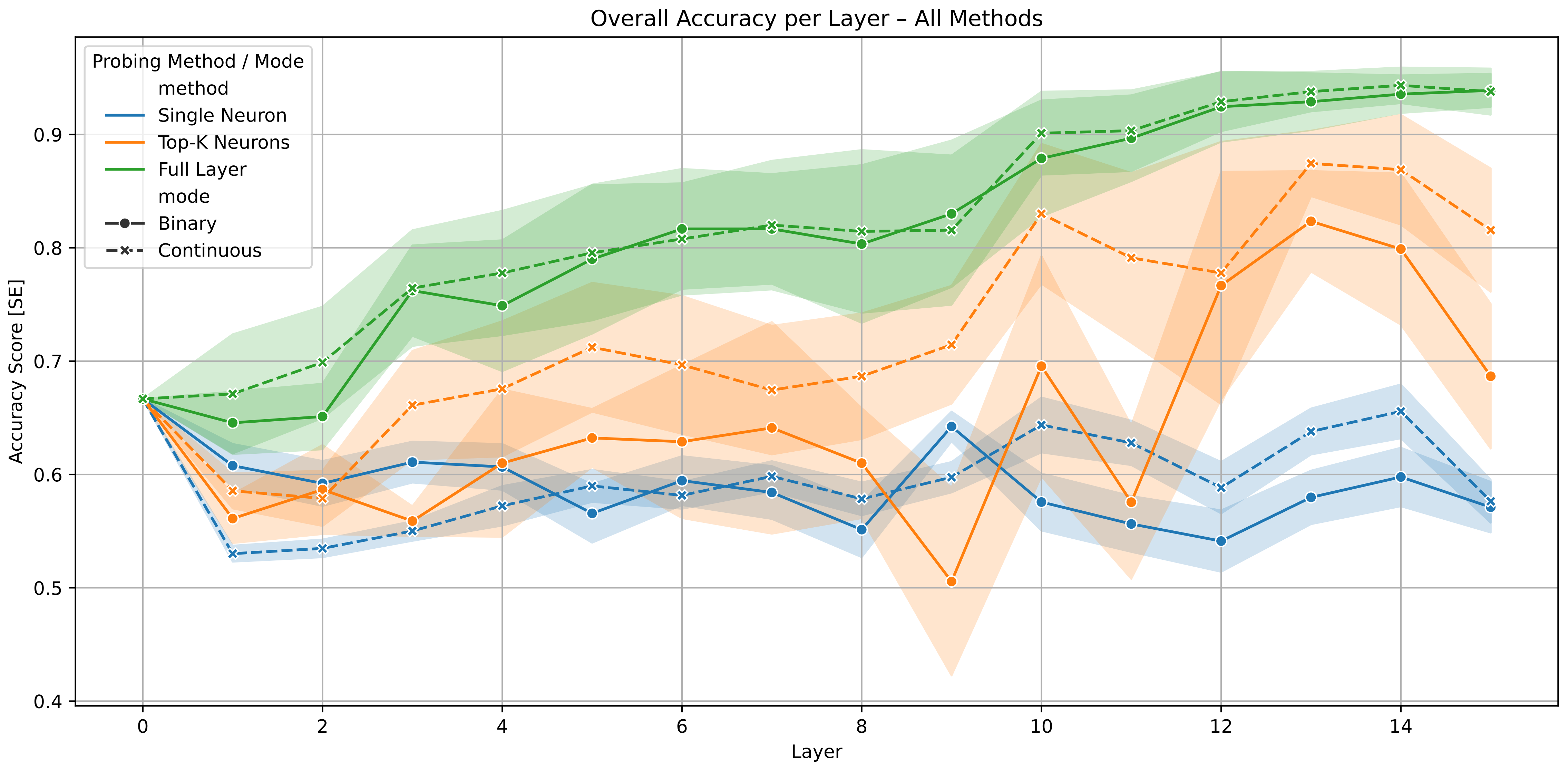


Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	15.0	14.0
Full Layer	f1_max	0.9799	0.9733
Full Layer	f1_mean	0.8063	0.8178
Full Layer	f1_std	0.131	0.1274
Single Neuron	f1_best_layer	14.0	14.0
Single Neuron	f1_max	0.9602	0.9244
Single Neuron	f1_mean	0.5064	0.5958
Single Neuron	f1_std	0.1254	0.0939
Top-K Neurons	f1_best_layer	13.0	13.0
Top-K Neurons	f1_max	0.9602	0.9473
Top-K Neurons	f1_mean	0.6322	0.7233
Top-K Neurons	f1_std	0.1401	0.126

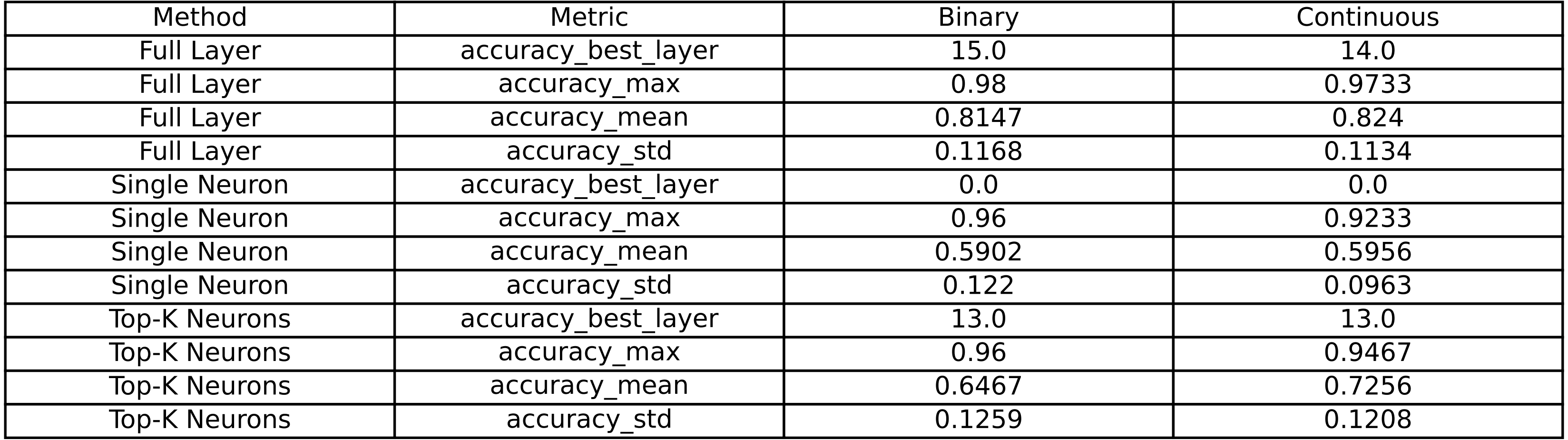

Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
0.675
0.650
0.625
0.600
0.575
0.550
0.525
Layer
0
2
4
6
8
10
12
14
Mode
Binary
Continuous


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
2
4
6
8
10
12
14


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.95
0.90
0.85
0.80
0.75
0.70
0.65
Layer
0
2
4
6
8
10
12
14


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
Layer
0
2
4
6
8
10
12
14

Accuracy Score Summary by Probing Method



Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	15.0	14.0
Full Layer	accuracy_max	0.98	0.9733
Full Layer	accuracy_mean	0.8147	0.824
Full Layer	accuracy_std	0.1168	0.1134
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.96	0.9233
Single Neuron	accuracy_mean	0.5902	0.5956
Single Neuron	accuracy_std	0.122	0.0963
Top-K Neurons	accuracy_best_layer	13.0	13.0
Top-K Neurons	accuracy_max	0.96	0.9467
Top-K Neurons	accuracy_mean	0.6467	0.7256
Top-K Neurons	accuracy_std	0.1259	0.1208