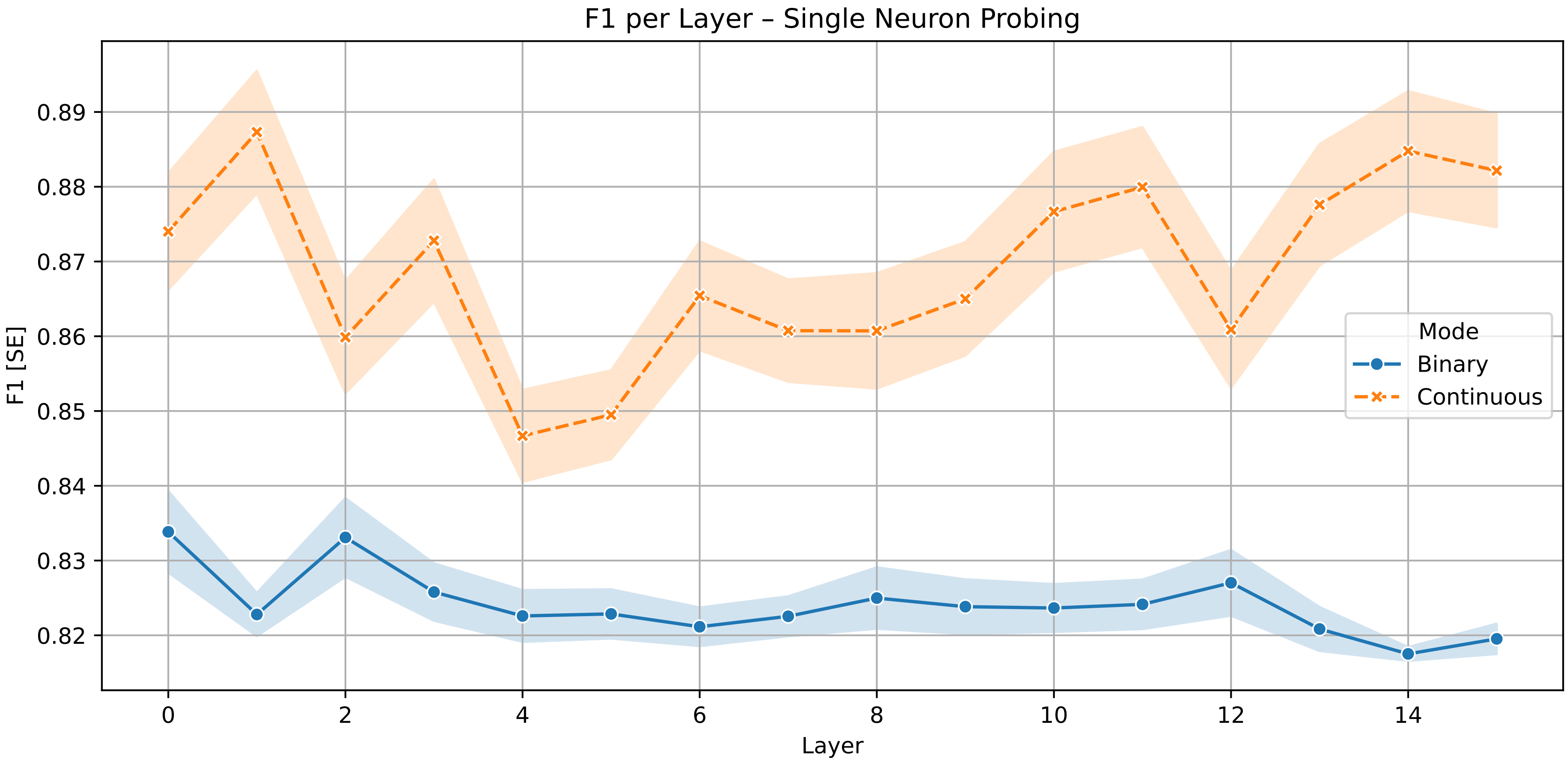
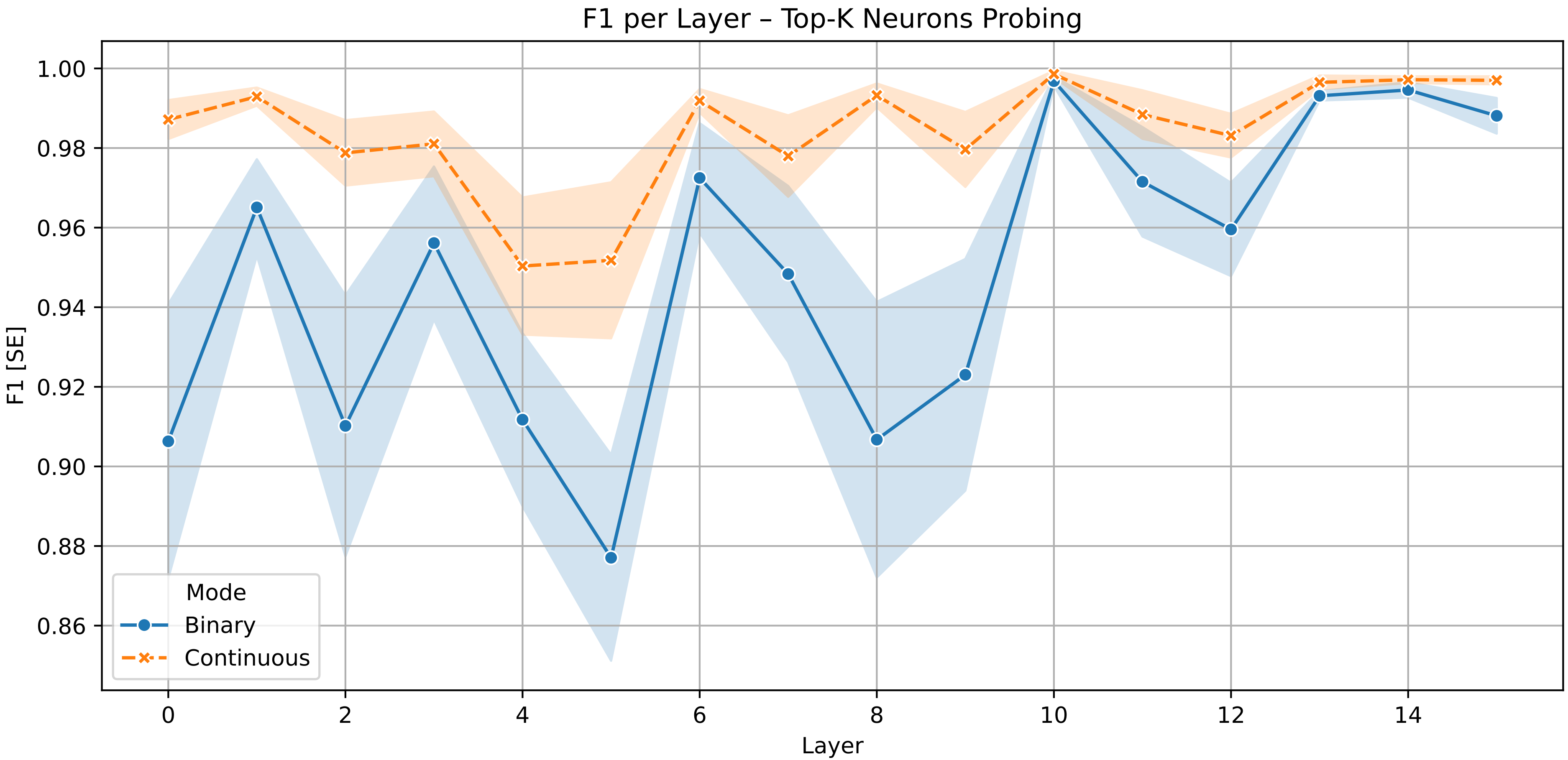
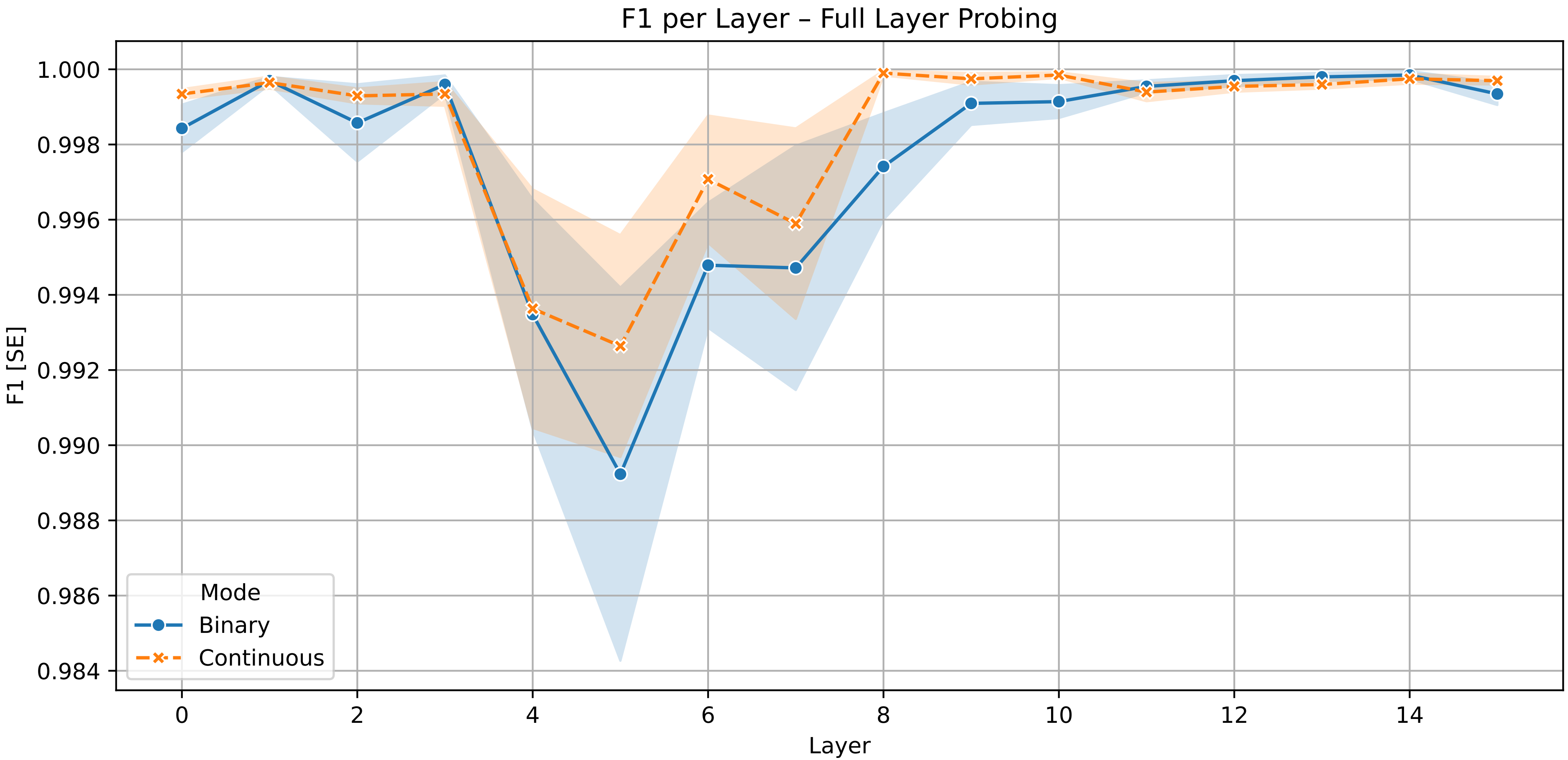
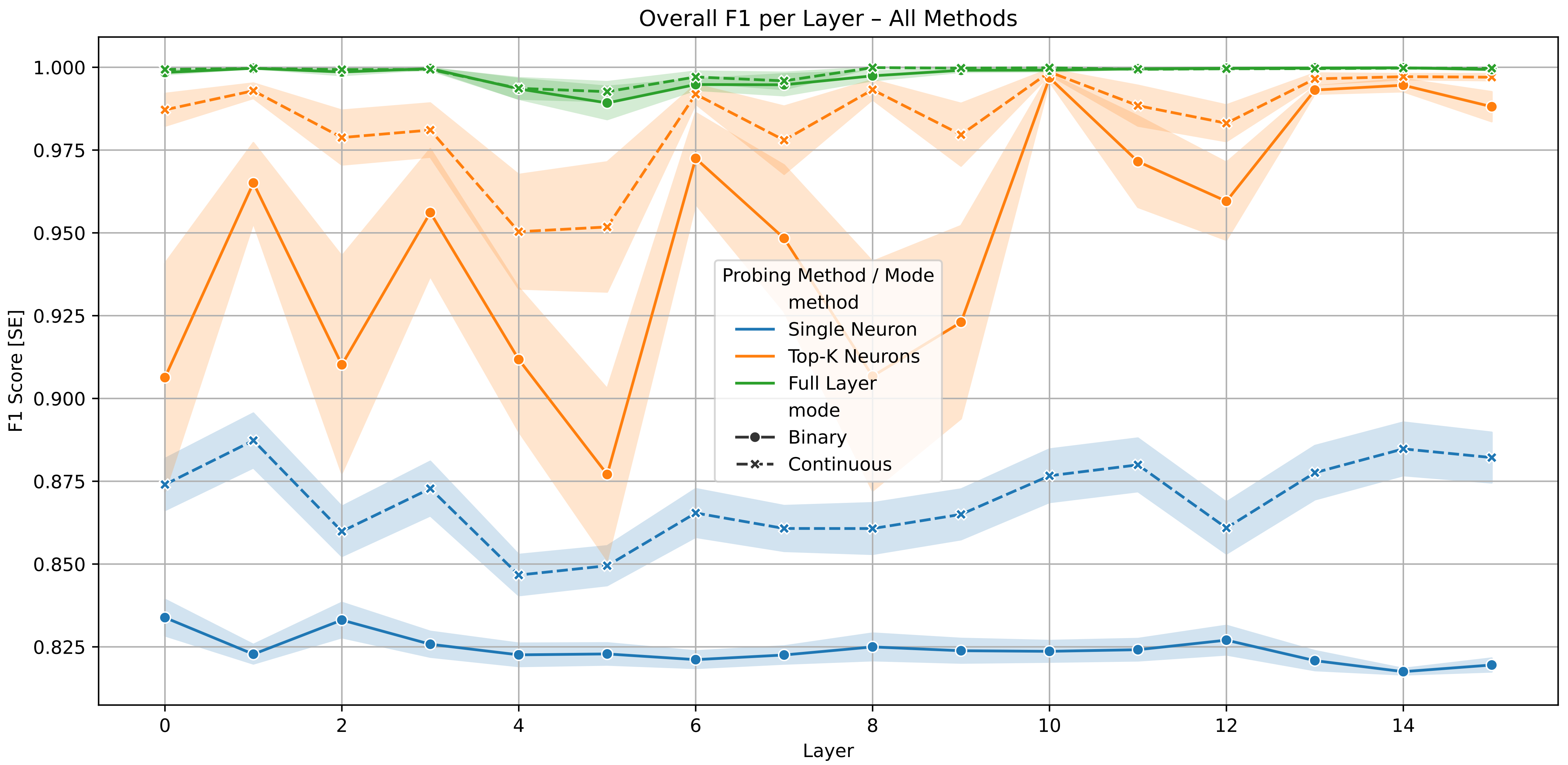
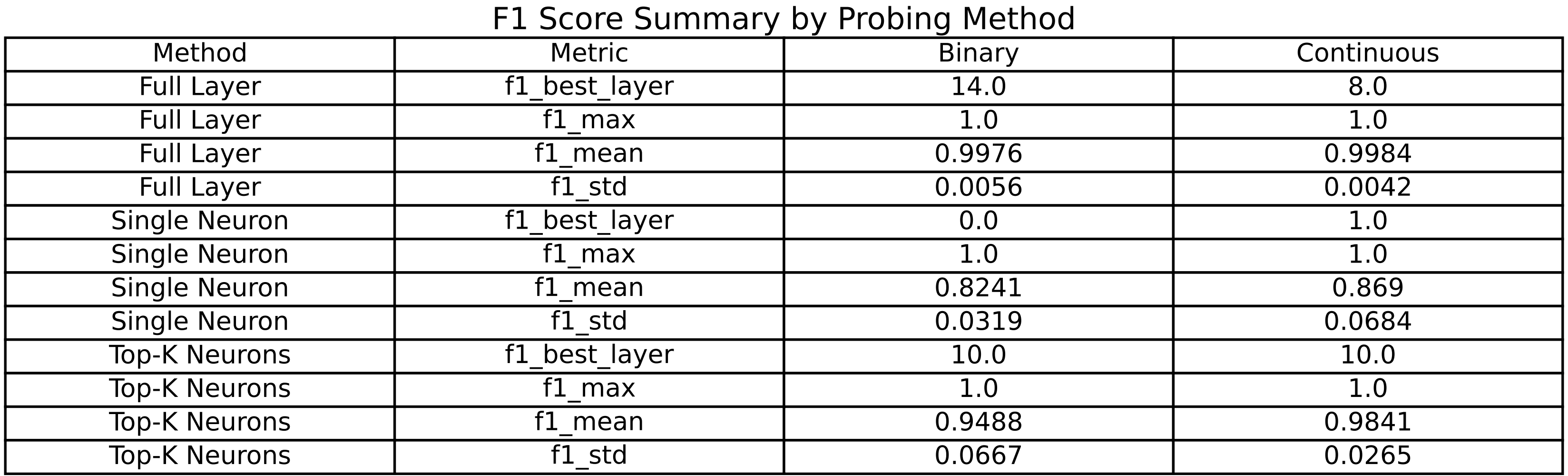

F1 per Layer – Single Neuron Probing
F1 [SE]
0.89
0.88
0.87
0.86
0.85
0.84
0.83
0.82
Layer
0
2
4
6
8
10
12
14
Mode
Binary
Continuous


F1 per Layer – Top-K Neurons Probing
F1 [SE]
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
Layer
0
2
4
6
8
10
12
14
Mode
Binary
Continuous

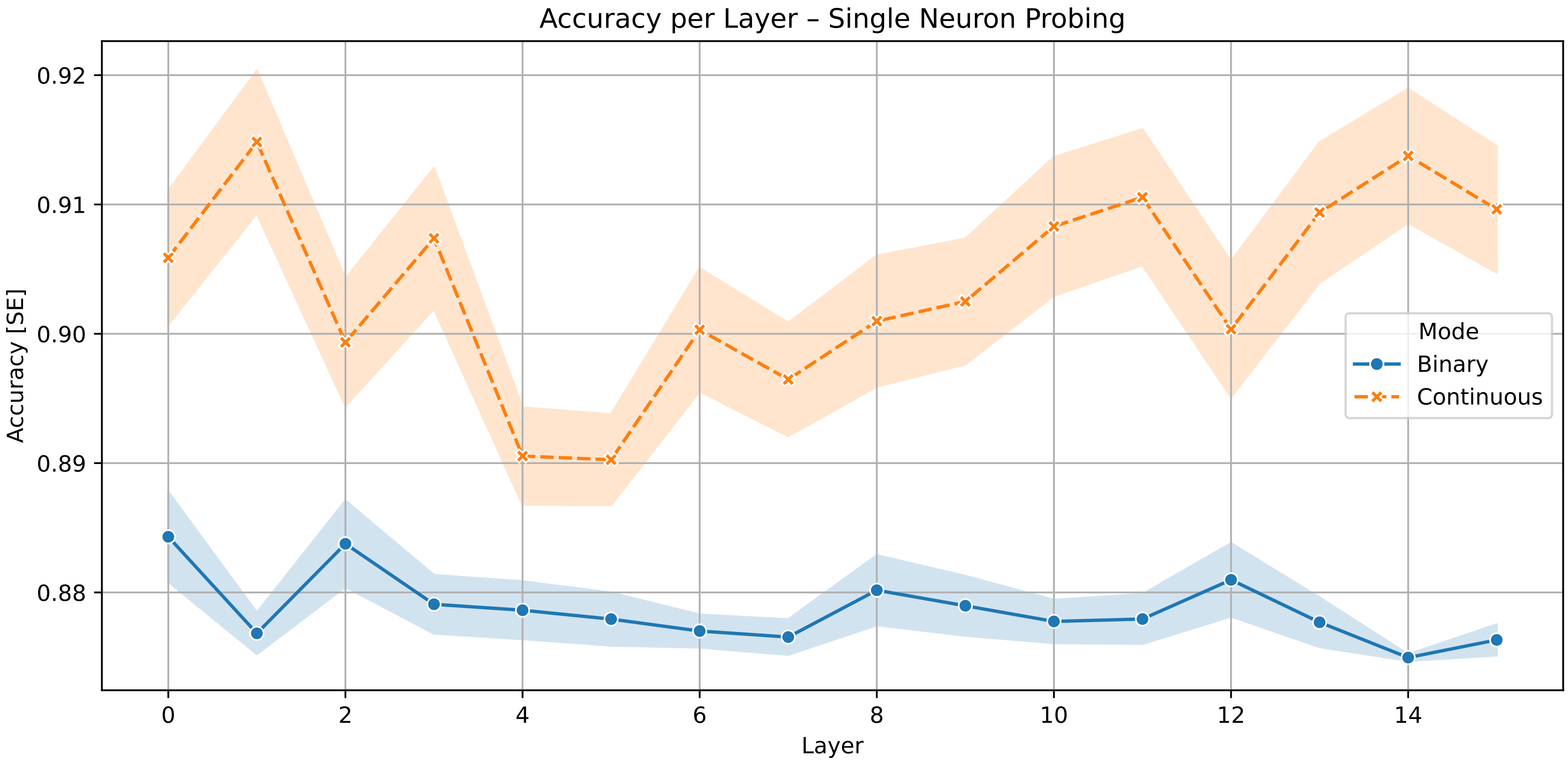

F1 per Layer – Full Layer Probing
F1 [SE]
1.000
0.998
0.996
0.994
0.992
0.990
0.988
0.986
0.984
Layer
0
2
4
6
8
10
12
14
Mode
Binary
Continuous

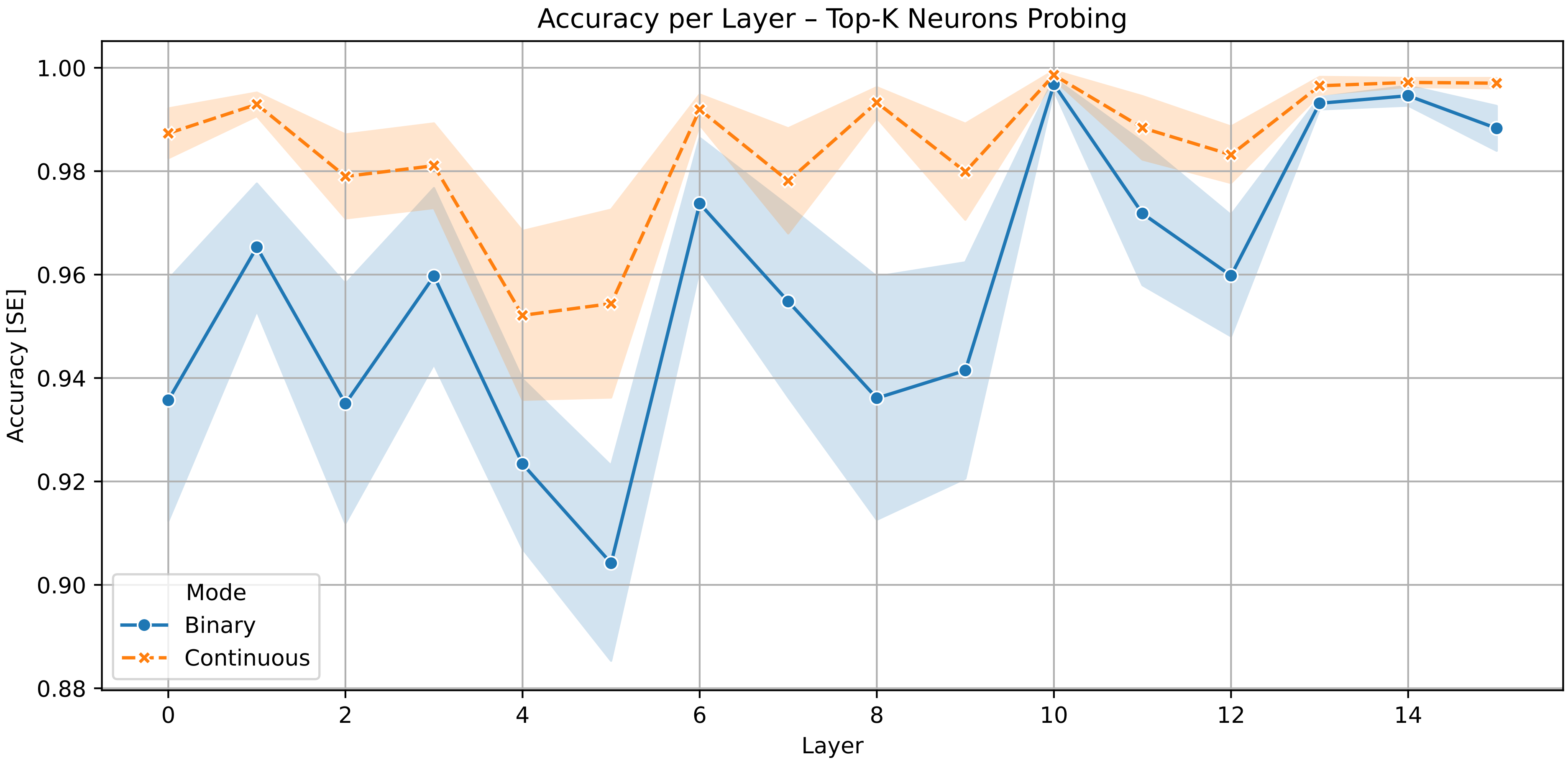

Overall F1 per Layer – All Methods
F1 Score [SE]
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
Layer
0
2
4
6
8
10
12
14
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

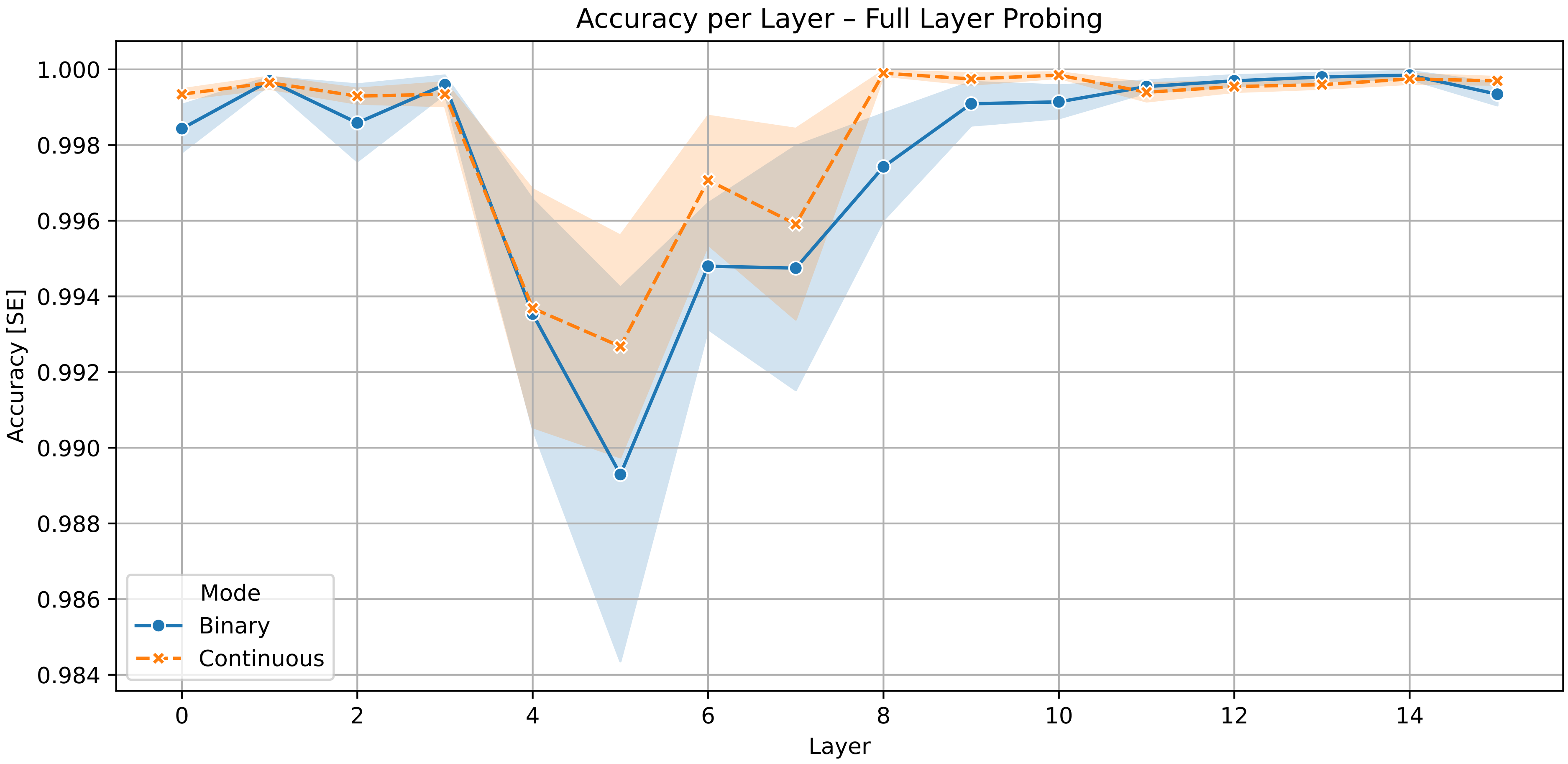


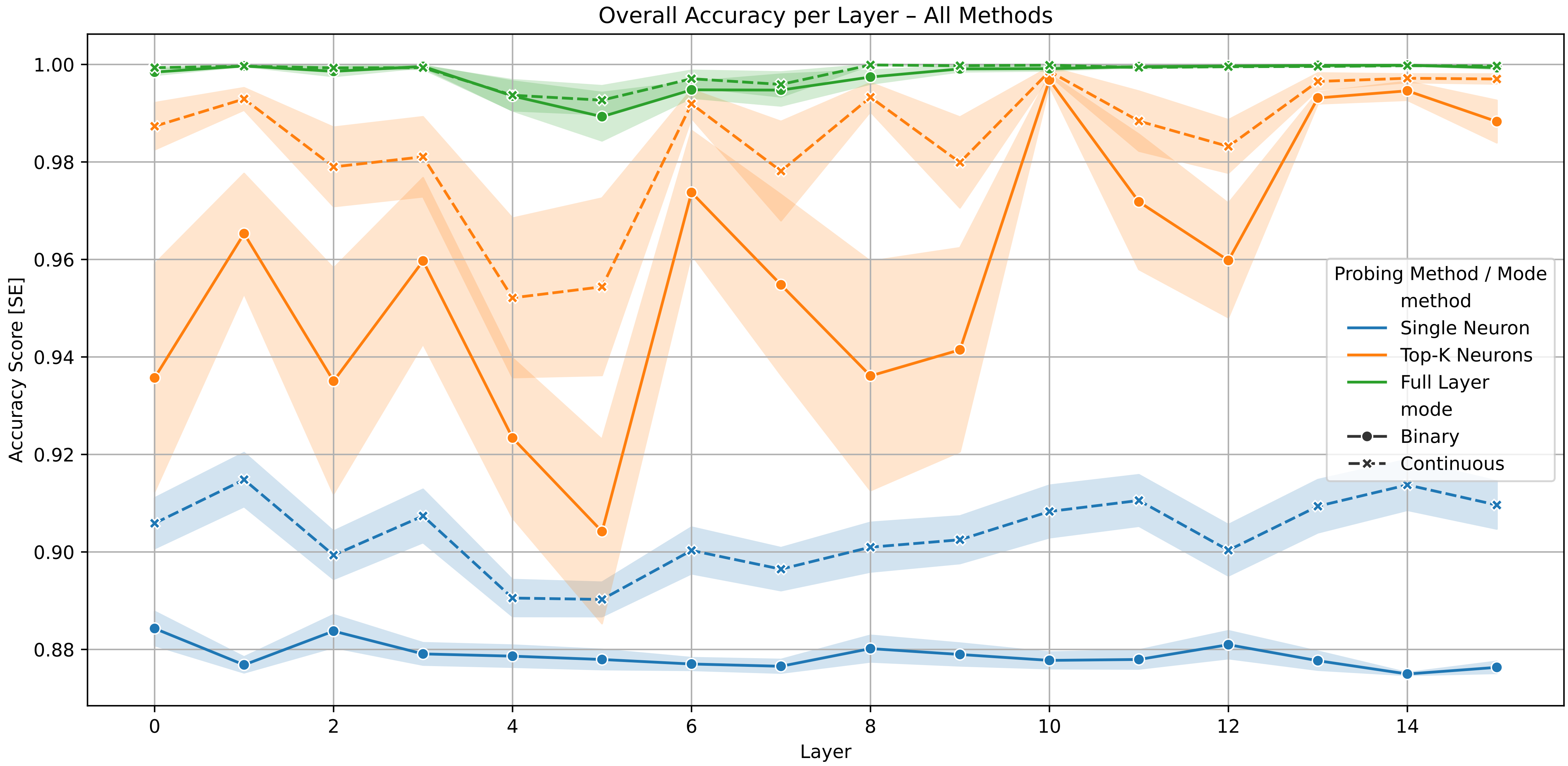
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	14.0	8.0
Full Layer	f1_max	1.0	1.0
Full Layer	f1_mean	0.9976	0.9984
Full Layer	f1_std	0.0056	0.0042
Single Neuron	f1_best_layer	0.0	1.0
Single Neuron	f1_max	1.0	1.0
Single Neuron	f1_mean	0.8241	0.869
Single Neuron	f1_std	0.0319	0.0684
Top-K Neurons	f1_best_layer	10.0	10.0
Top-K Neurons	f1_max	1.0	1.0
Top-K Neurons	f1_mean	0.9488	0.9841
Top-K Neurons	f1_std	0.0667	0.0265

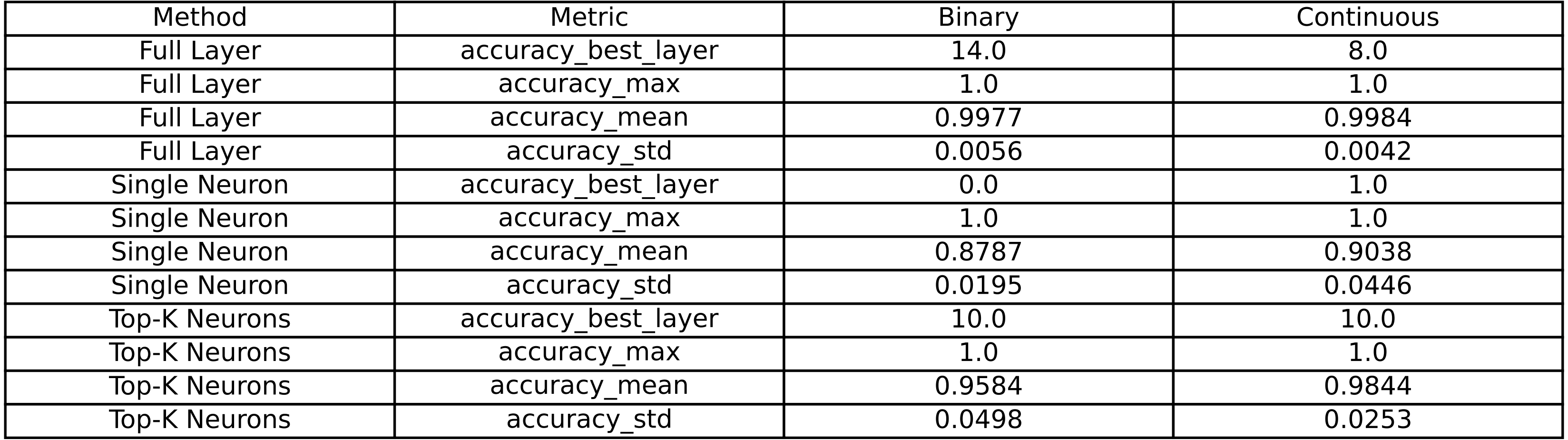

Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
0.92
0.91
0.90
0.89
0.88
Layer
0
2
4
6
8
10
12
14
Mode
Binary
Continuous


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
1.00
0.98
0.96
0.94
0.92
0.90
0.88
Layer
0
2
4
6
8
10
12
14
Mode
Binary
Continuous


Accuracy per Layer – Full Layer Probing
Accuracy [SE]
1.000
0.998
0.996
0.994
0.992
0.990
0.988
0.986
0.984
Layer
0
2
4
6
8
10
12
14
Mode
Binary
Continuous


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
1.00
0.98
0.96
0.94
0.92
0.90
0.88
Layer
0
2
4
6
8
10
12
14
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	14.0	8.0
Full Layer	accuracy_max	1.0	1.0
Full Layer	accuracy_mean	0.9977	0.9984
Full Layer	accuracy_std	0.0056	0.0042
Single Neuron	accuracy_best_layer	0.0	1.0
Single Neuron	accuracy_max	1.0	1.0
Single Neuron	accuracy_mean	0.8787	0.9038
Single Neuron	accuracy_std	0.0195	0.0446
Top-K Neurons	accuracy_best_layer	10.0	10.0
Top-K Neurons	accuracy_max	1.0	1.0
Top-K Neurons	accuracy_mean	0.9584	0.9844
Top-K Neurons	accuracy_std	0.0498	0.0253