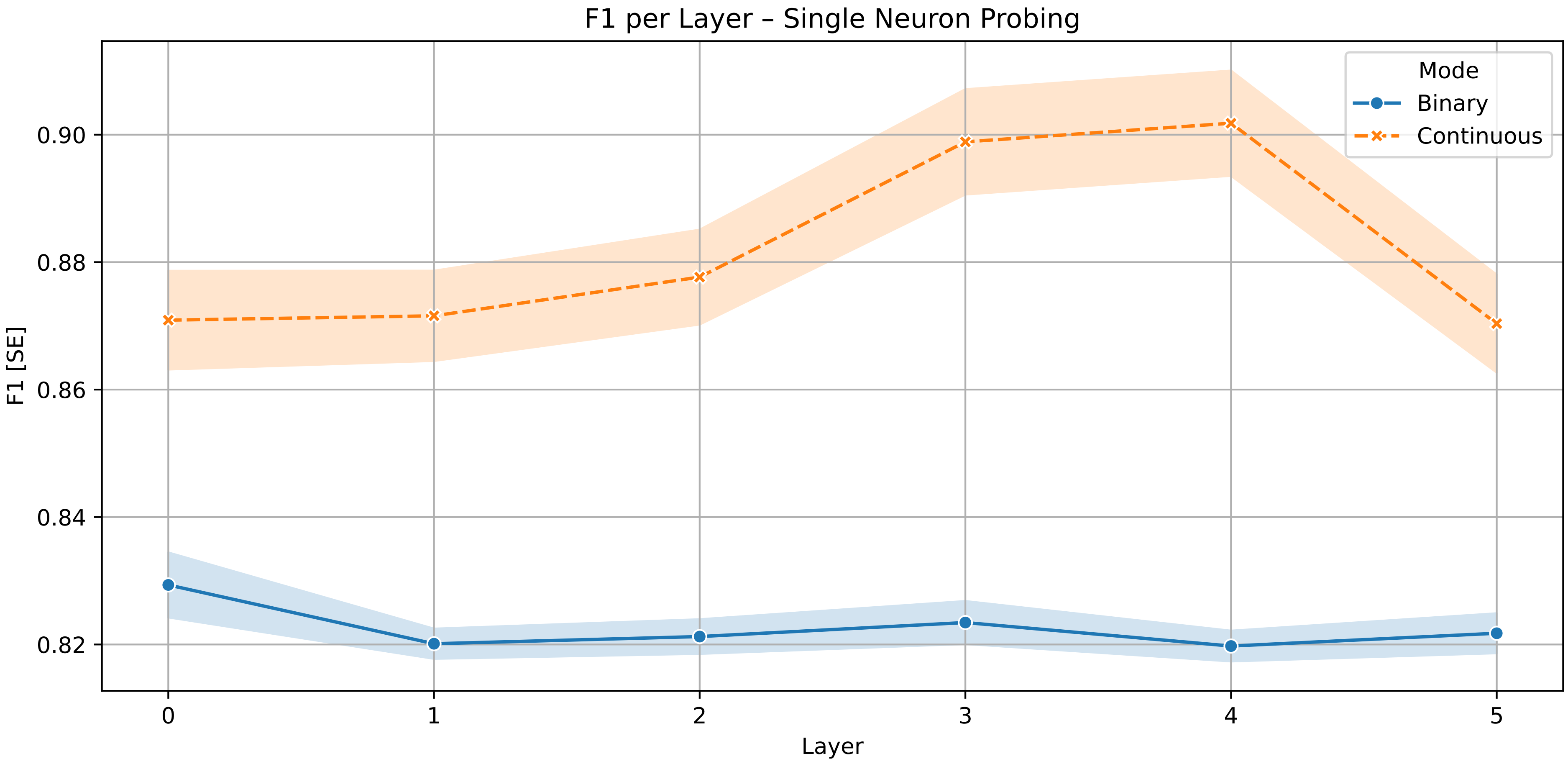
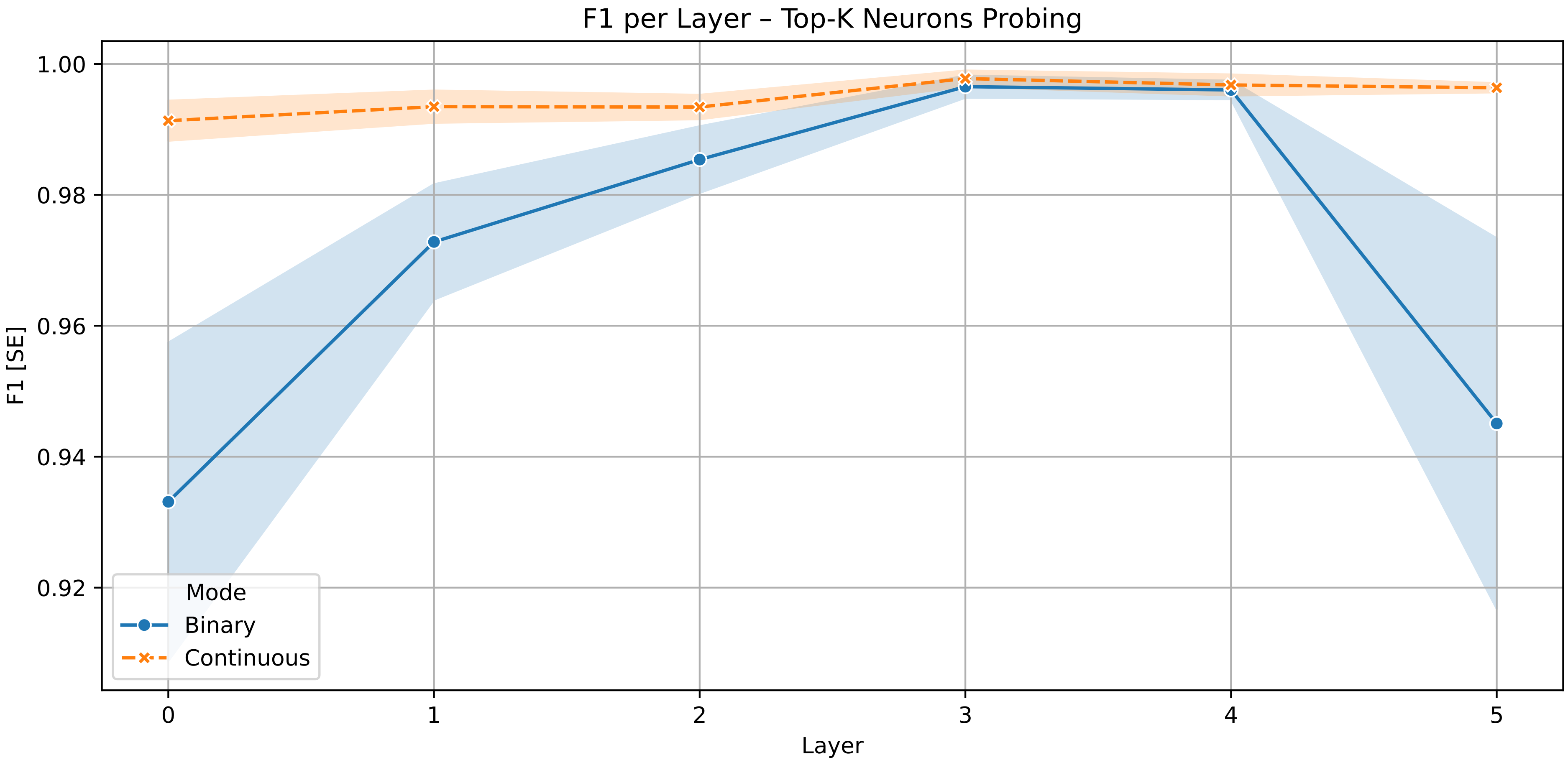
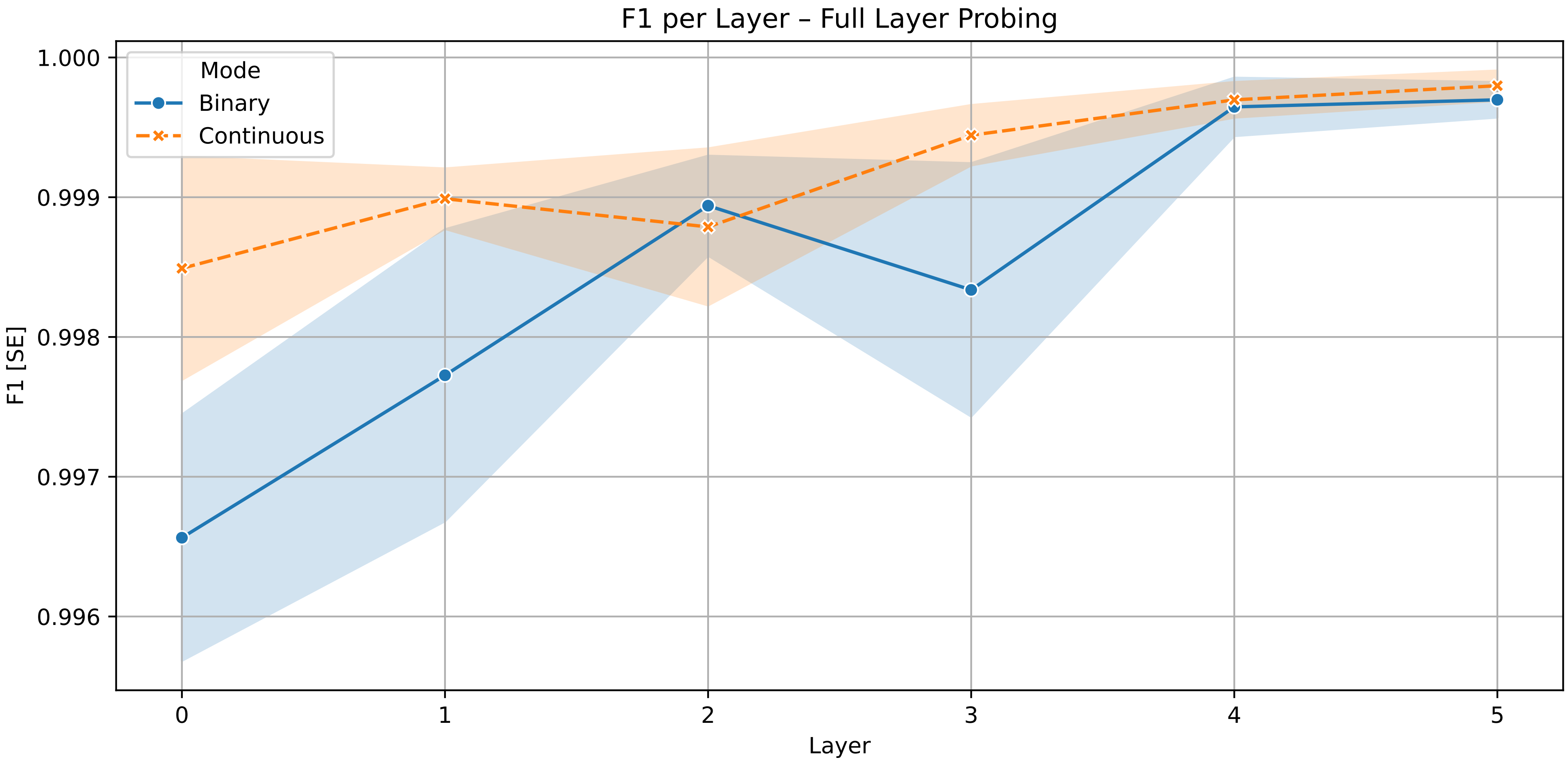
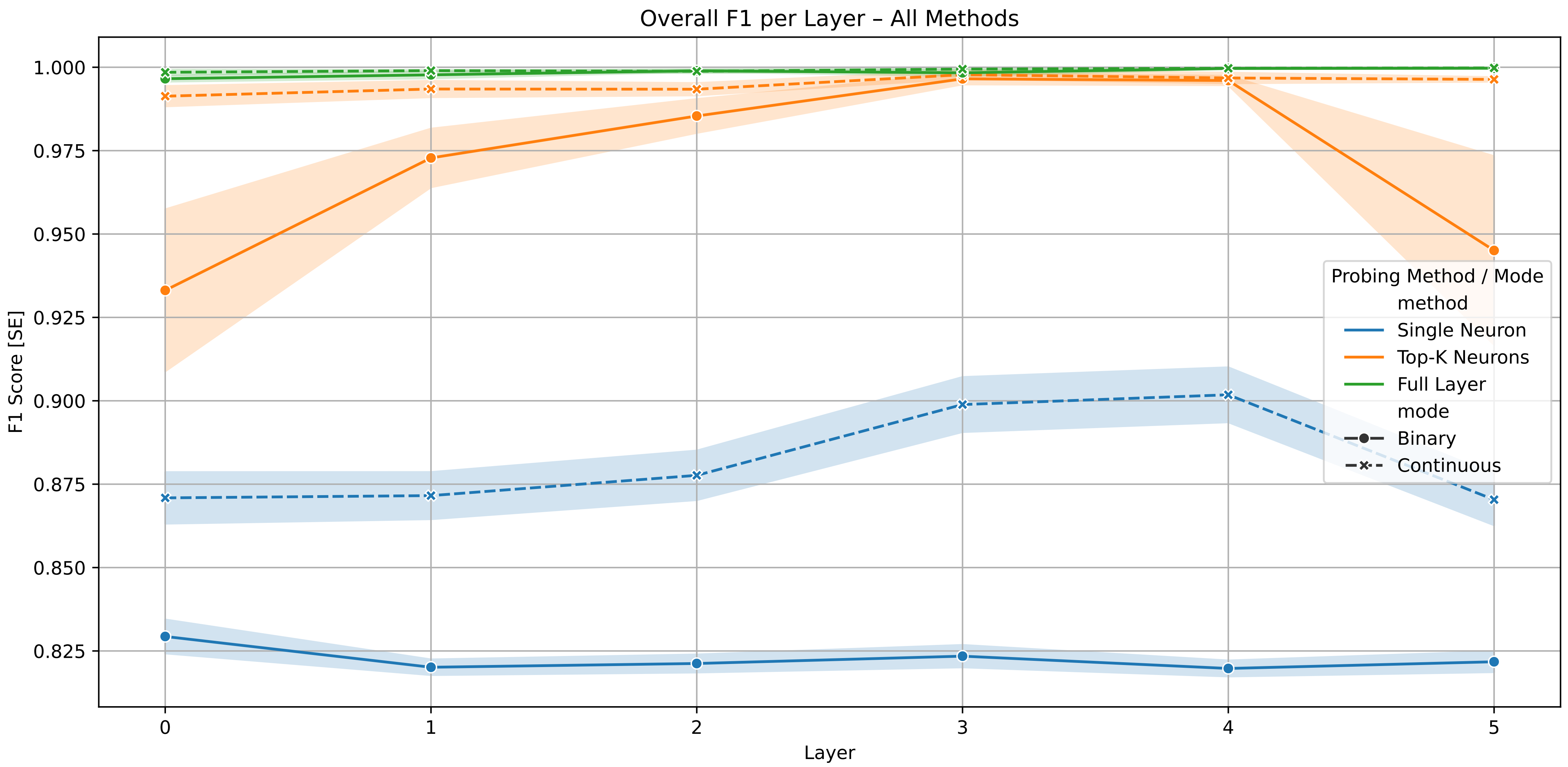
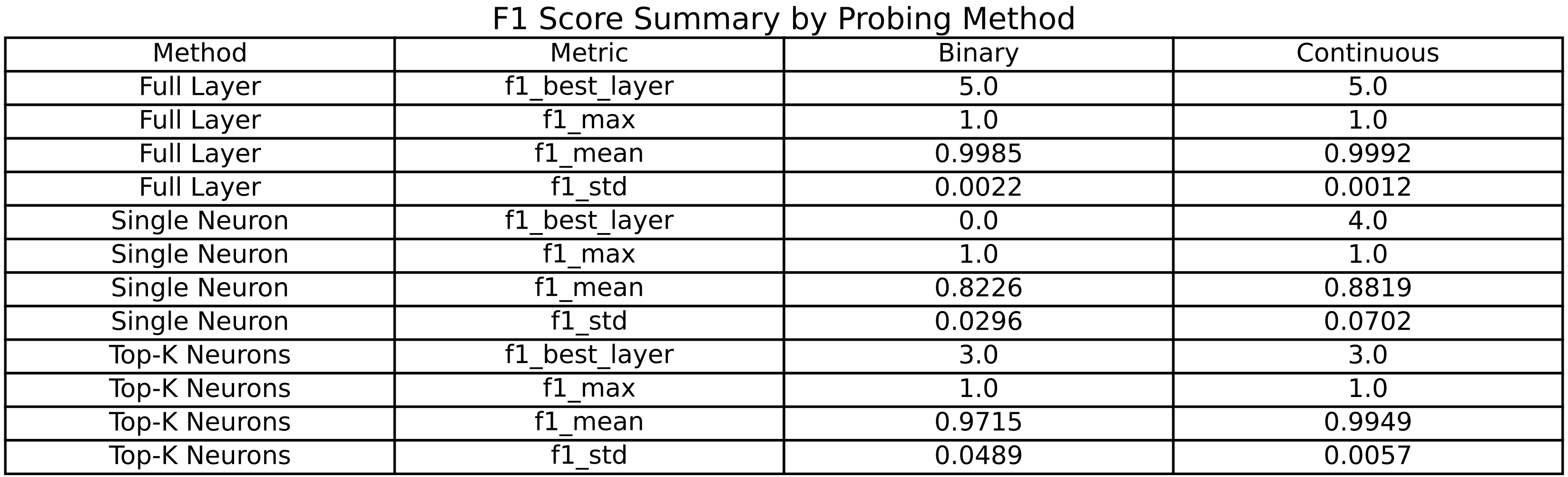

F1 per Layer – Single Neuron Probing
F1 [SE]
0.82
0.84
0.86
0.88
0.90
Layer
0
1
2
3
4
5
Mode
Binary
Continuous


F1 per Layer – Top-K Neurons Probing
F1 [SE]
1.00
0.98
0.96
0.94
0.92
Layer
0
1
2
3
4
5
Mode
Binary
Continuous


F1 per Layer – Full Layer Probing
F1 [SE]
1.000
0.999
0.998
0.997
0.996
Layer
0
1
2
3
4
5
Mode
Binary
Continuous

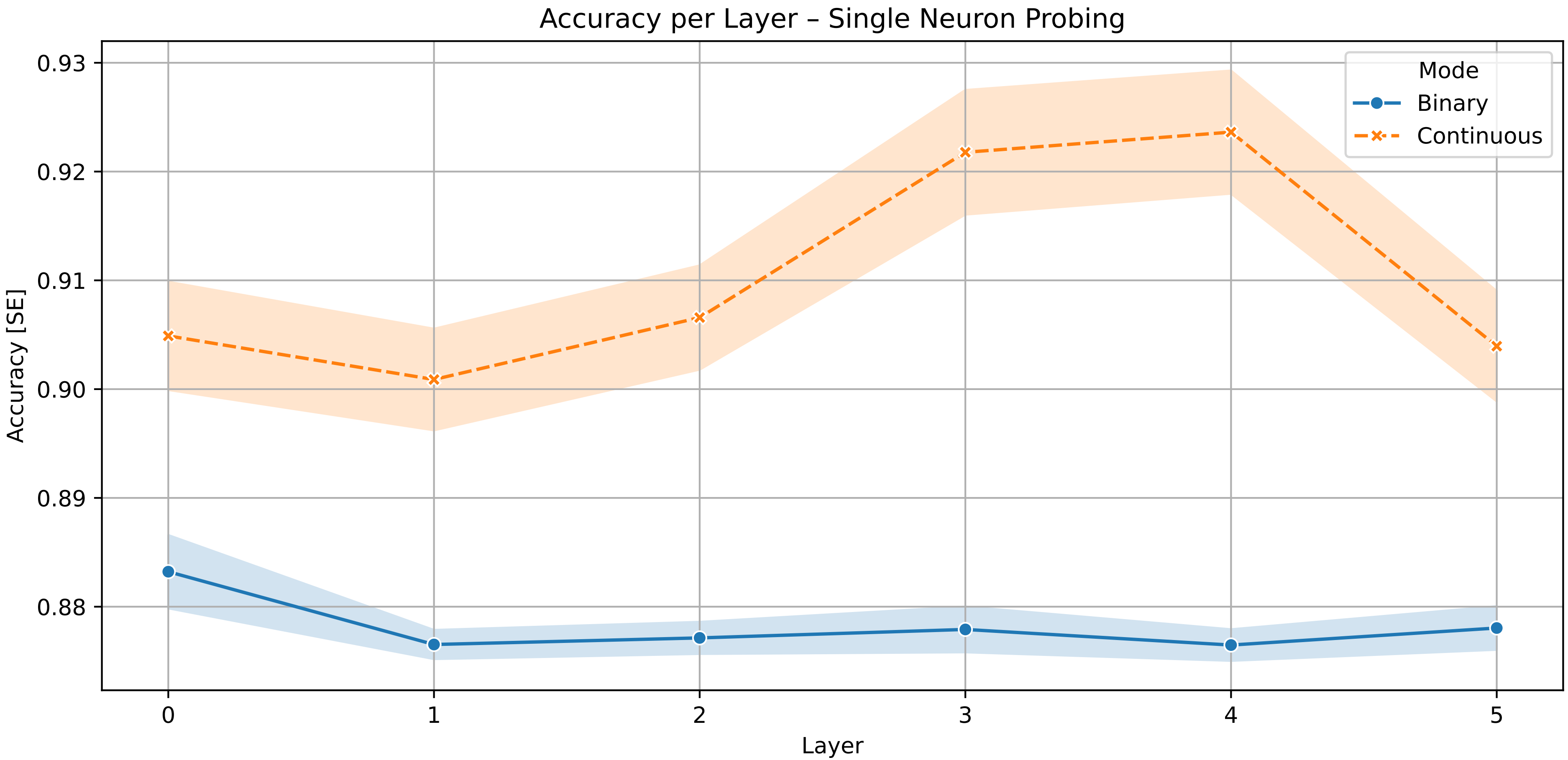


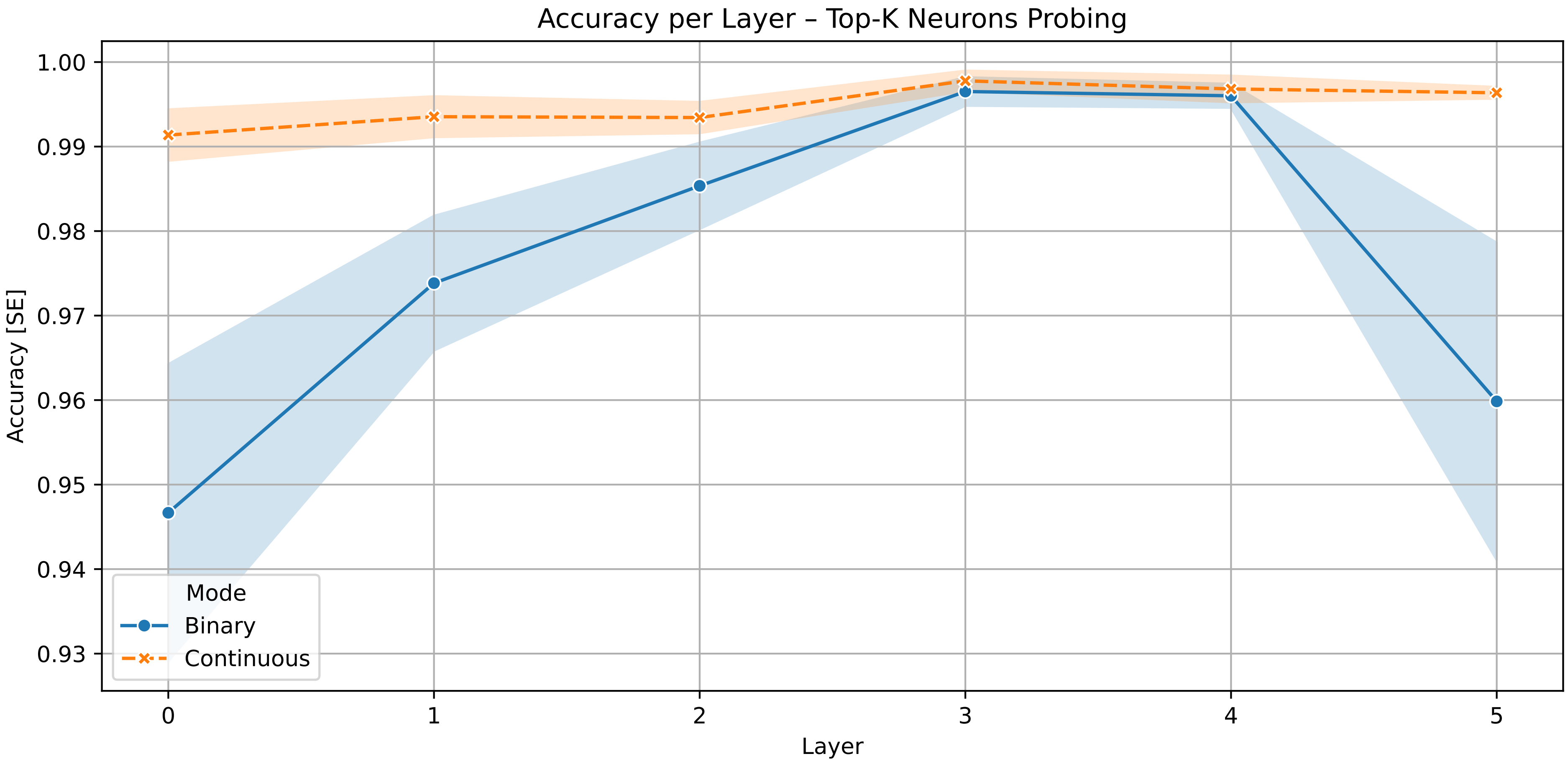
Overall F1 per Layer – All Methods
F1 Score [SE]
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
Layer
0
1
2
3
4
5
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

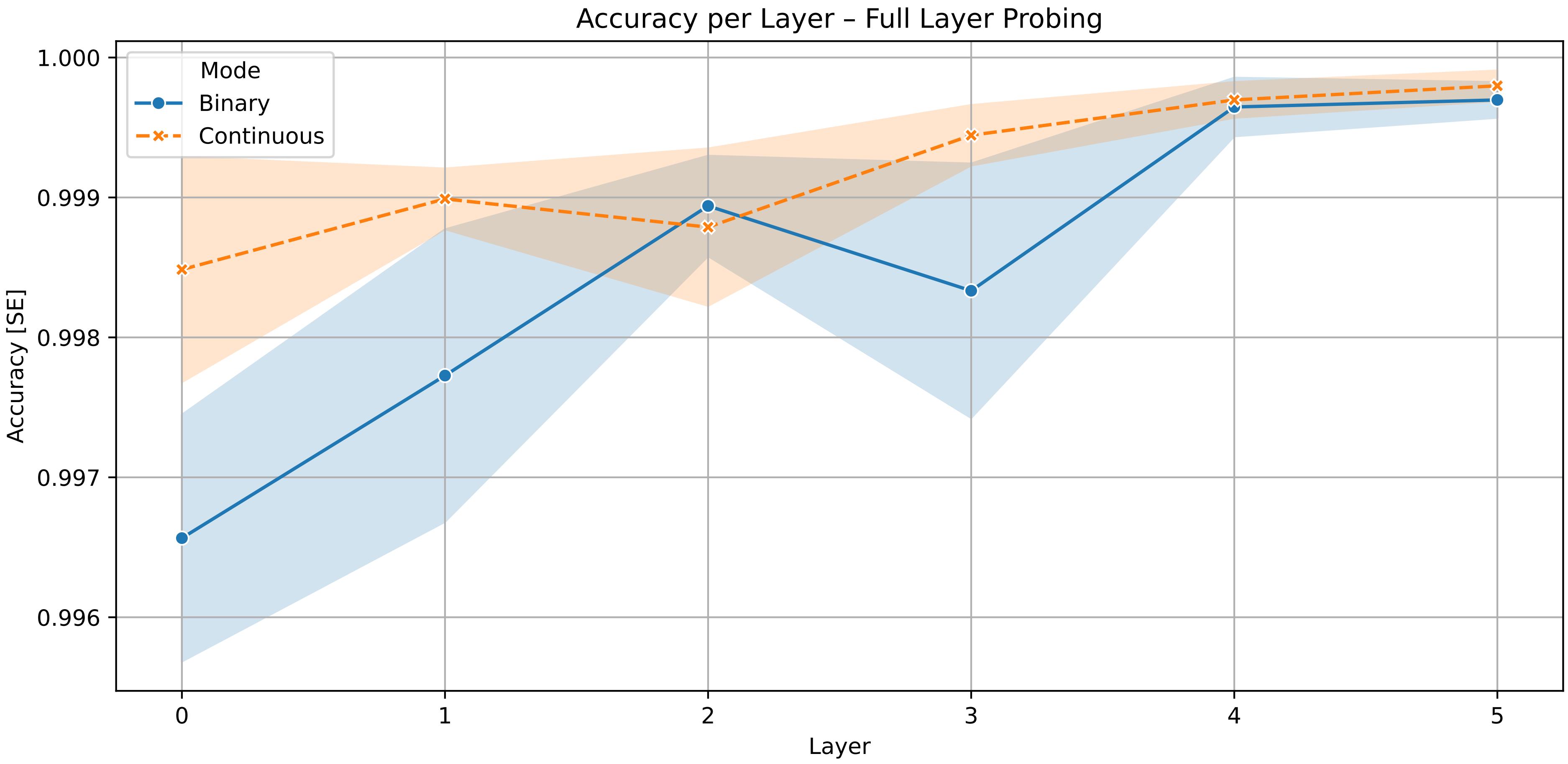


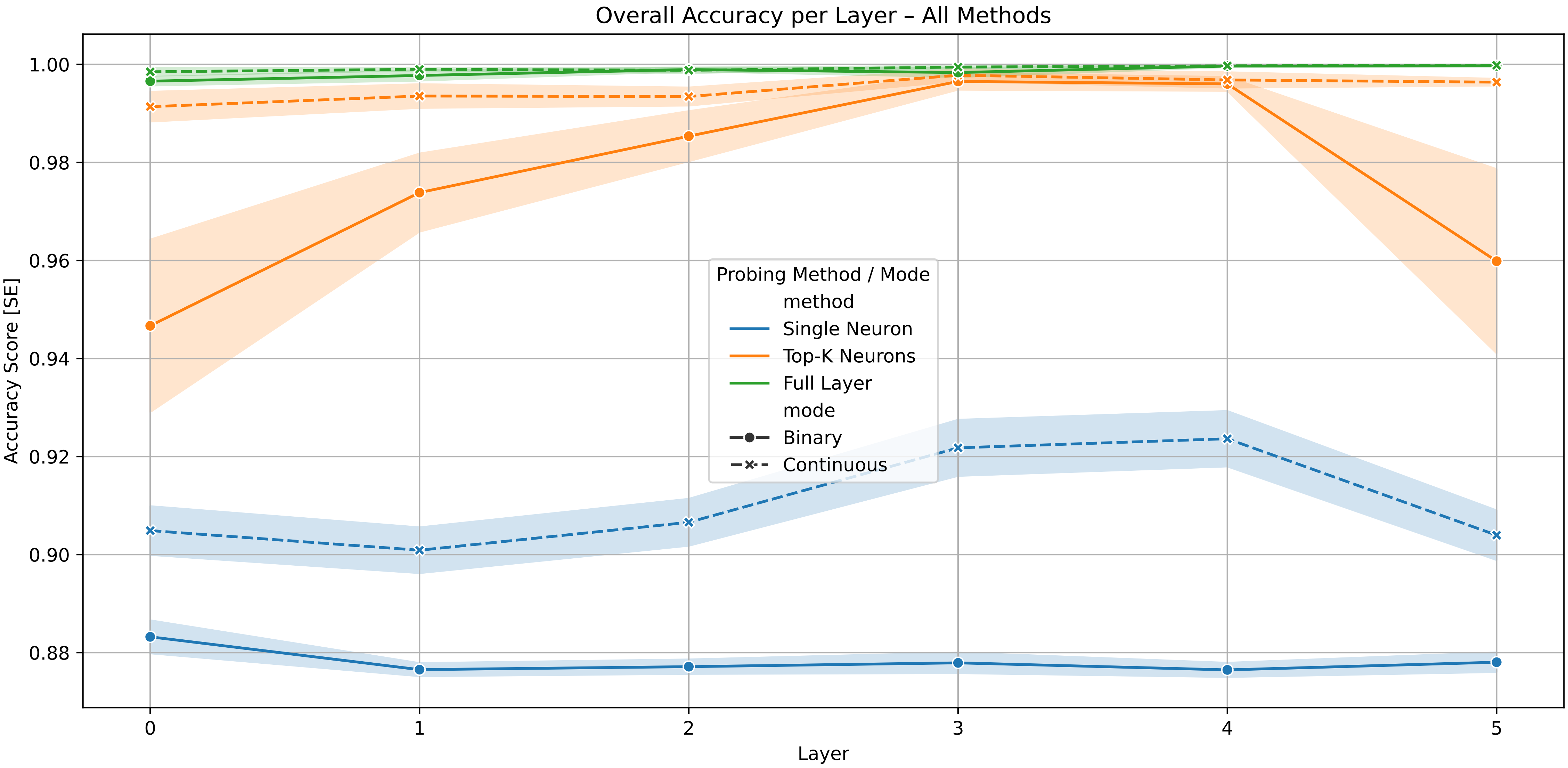
F1 Score Summary by Probing Method

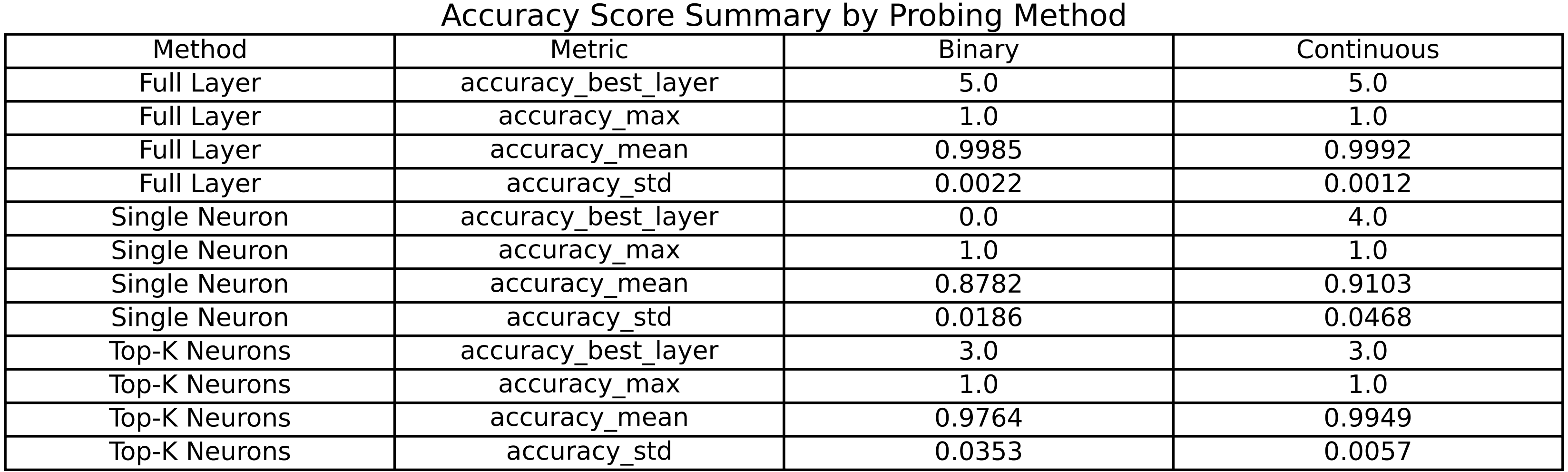
Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	5.0	5.0
Full Layer	f1_max	1.0	1.0
Full Layer	f1_mean	0.9985	0.9992
Full Layer	f1_std	0.0022	0.0012
Single Neuron	f1_best_layer	0.0	4.0
Single Neuron	f1_max	1.0	1.0
Single Neuron	f1_mean	0.8226	0.8819
Single Neuron	f1_std	0.0296	0.0702
Top-K Neurons	f1_best_layer	3.0	3.0
Top-K Neurons	f1_max	1.0	1.0
Top-K Neurons	f1_mean	0.9715	0.9949
Top-K Neurons	f1_std	0.0489	0.0057


Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
0.93
0.92
0.91
0.90
0.89
0.88
Layer
0
1
2
3
4
5
Mode
Binary
Continuous


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.93
Layer
0
1
2
3
4
5
Mode
Binary
Continuous


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
1.000
0.999
0.998
0.997
0.996
Layer
0
1
2
3
4
5


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
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	1.0	1.0
Full Layer	accuracy_mean	0.9985	0.9992
Full Layer	accuracy_std	0.0022	0.0012
Single Neuron	accuracy_best_layer	0.0	4.0
Single Neuron	accuracy_max	1.0	1.0
Single Neuron	accuracy_mean	0.8782	0.9103
Single Neuron	accuracy_std	0.0186	0.0468
Top-K Neurons	accuracy_best_layer	3.0	3.0
Top-K Neurons	accuracy_max	1.0	1.0
Top-K Neurons	accuracy_mean	0.9764	0.9949
Top-K Neurons	accuracy_std	0.0353	0.0057