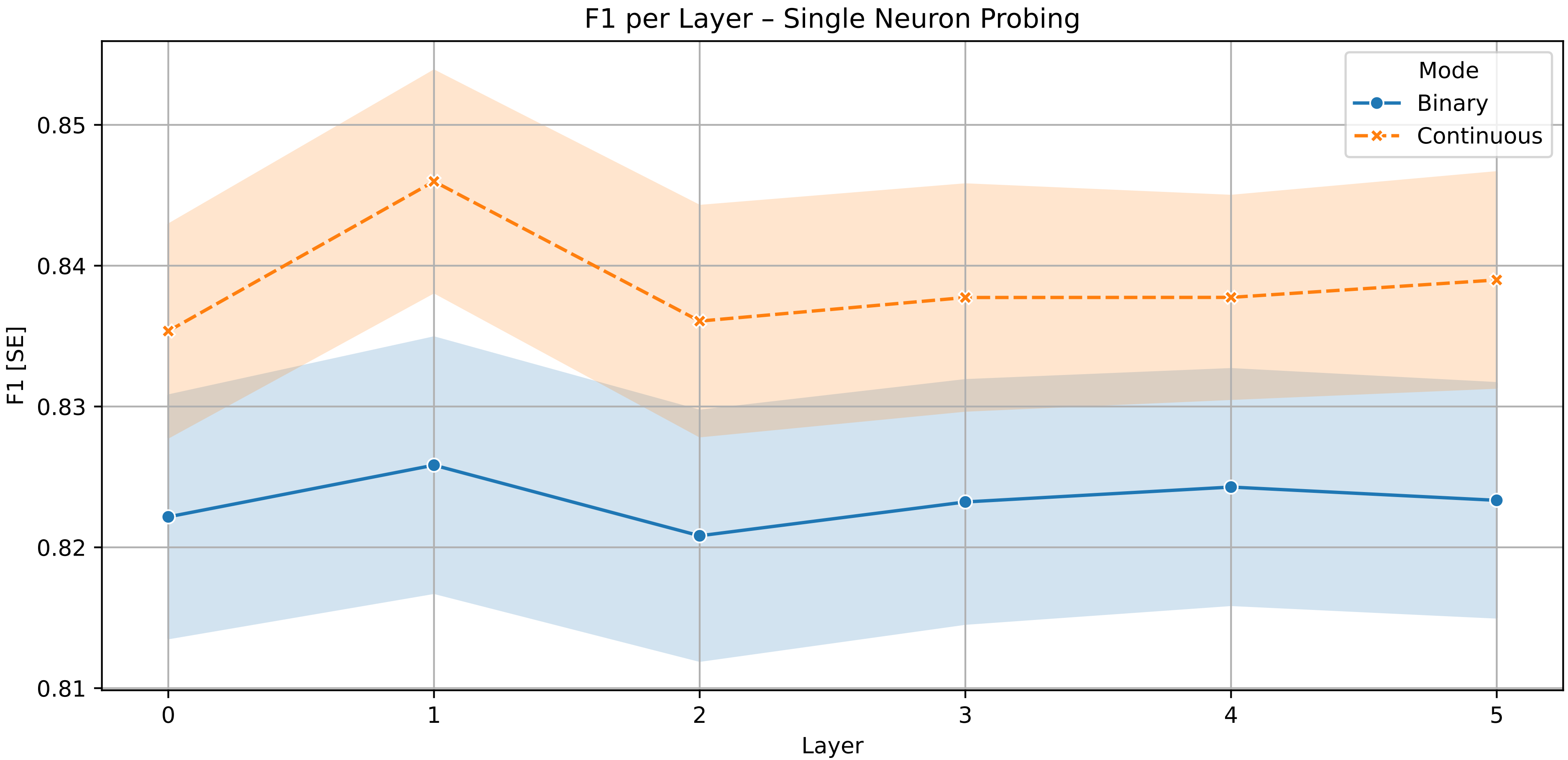
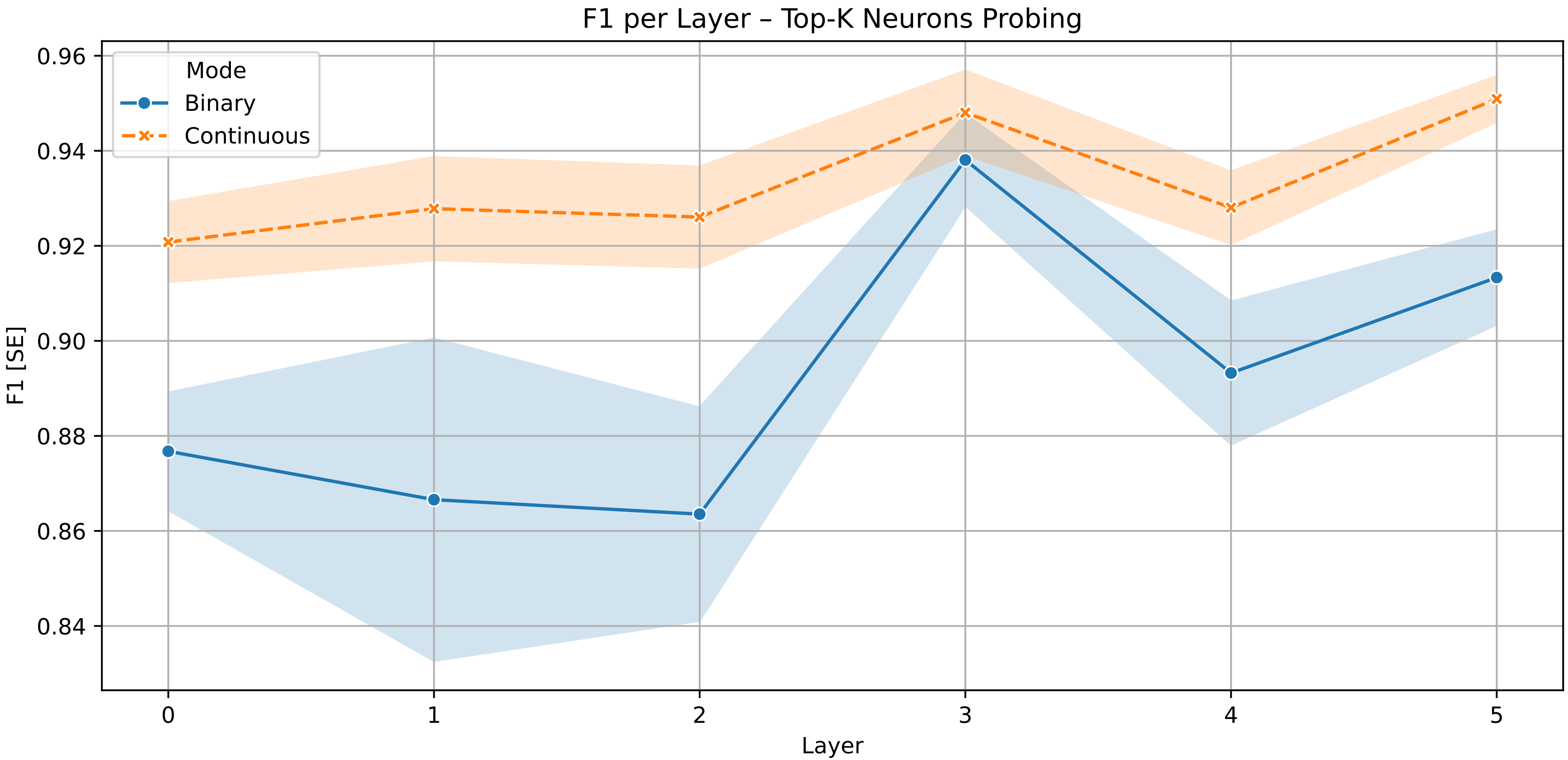
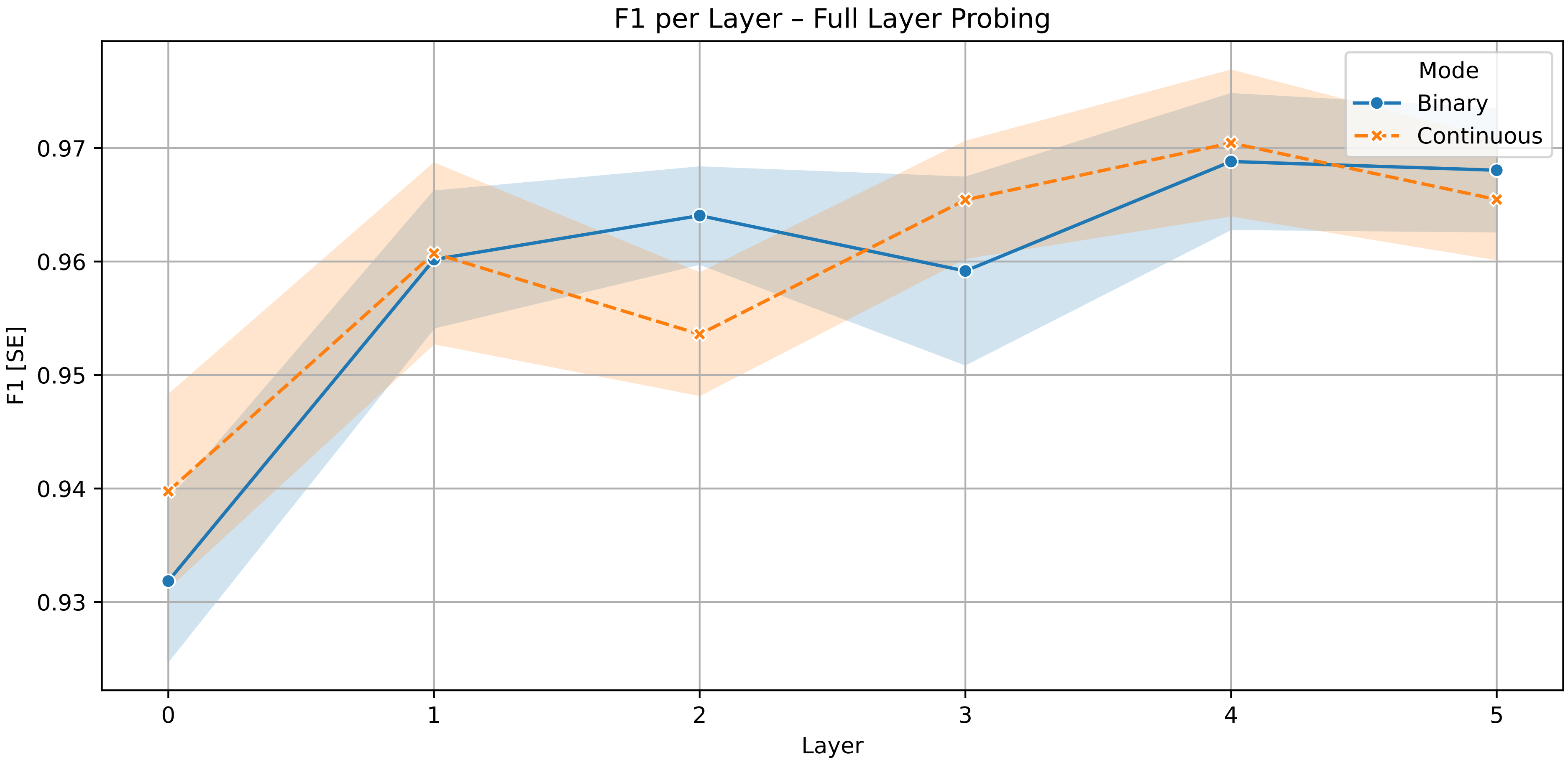
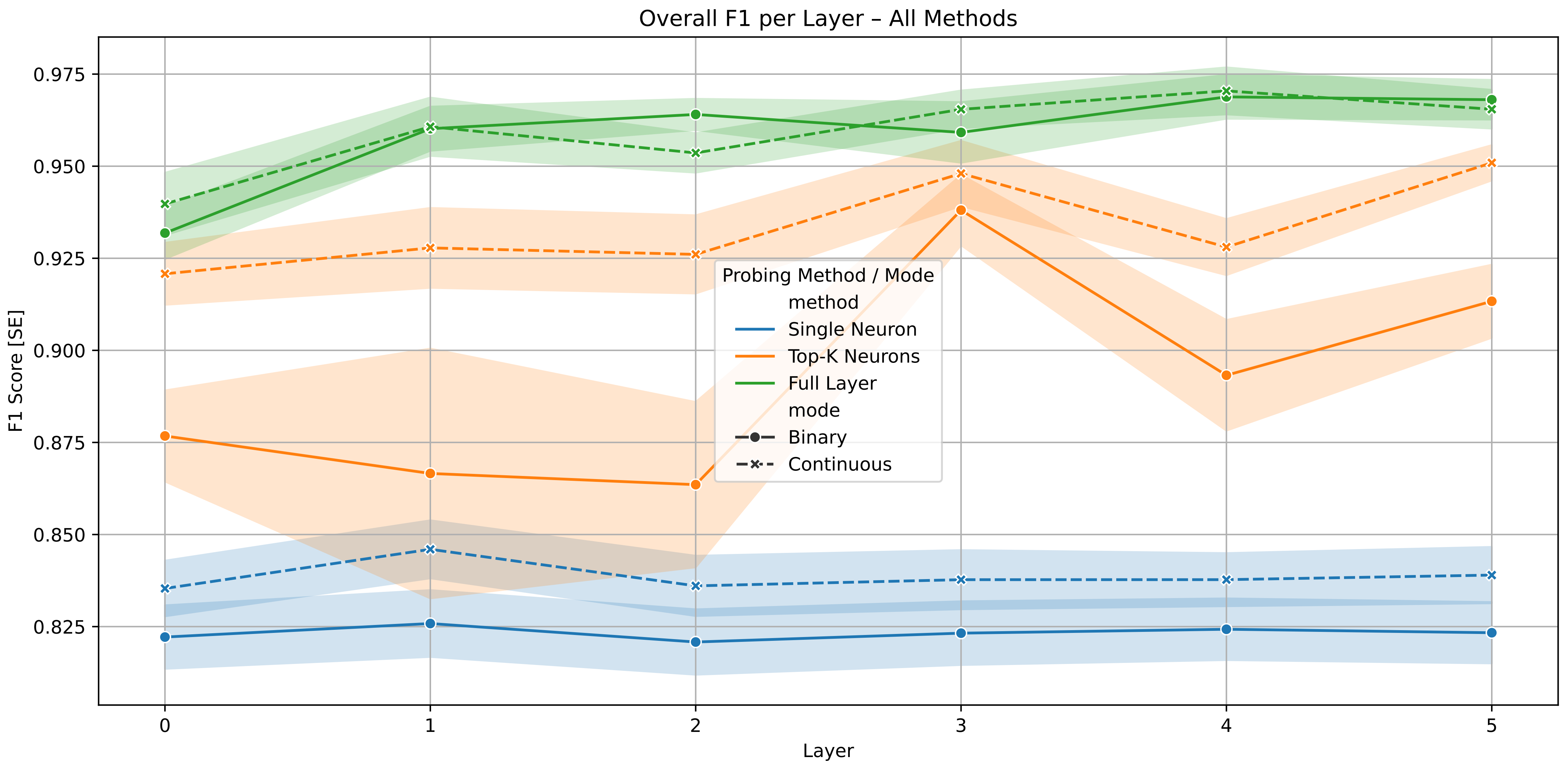
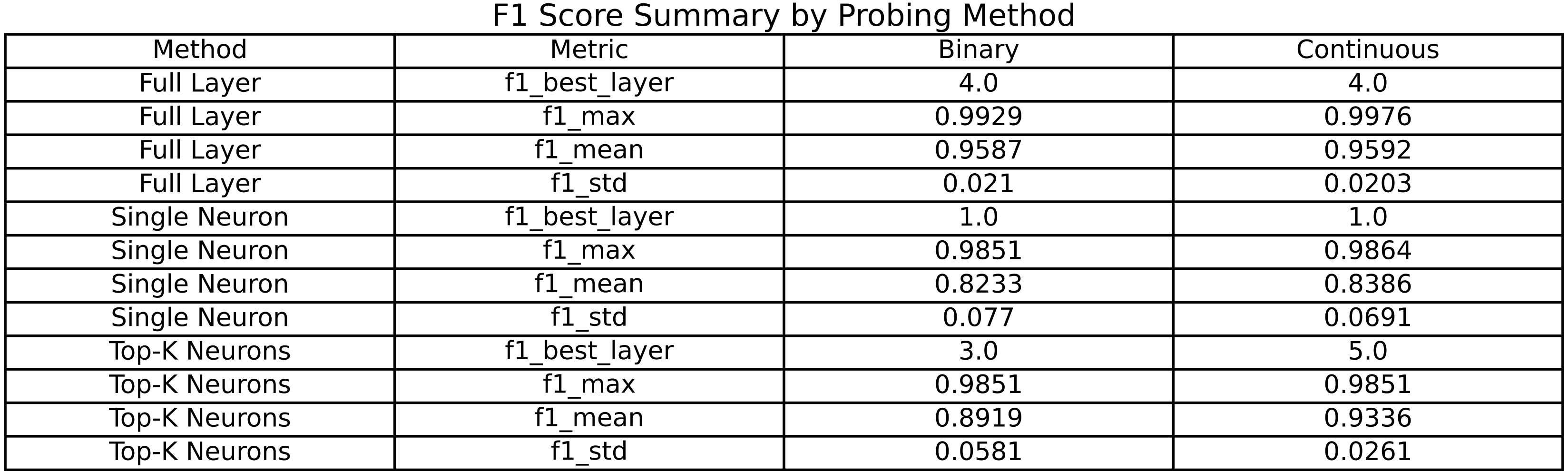

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.85
0.84
0.83
0.82
0.81
Layer
0
1
2
3
4
5


F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.96
0.94
0.92
0.90
0.88
0.86
0.84
Layer
0
1
2
3
4
5

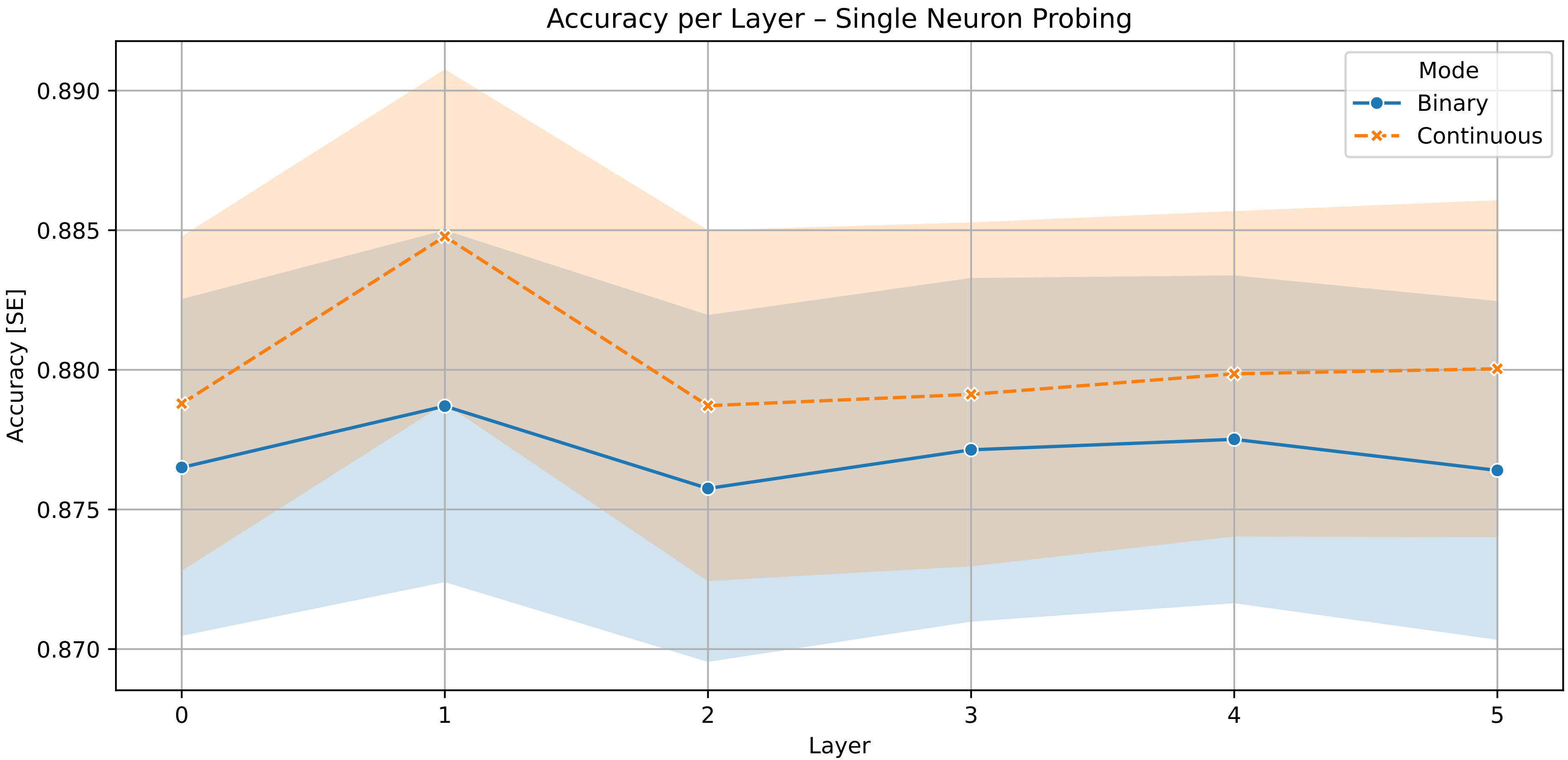

F1 per Layer – Full Layer Probing
F1 [SE]
0.97
0.96
0.95
0.94
0.93
Layer
0
1
2
3
4
5
Mode
Binary
Continuous

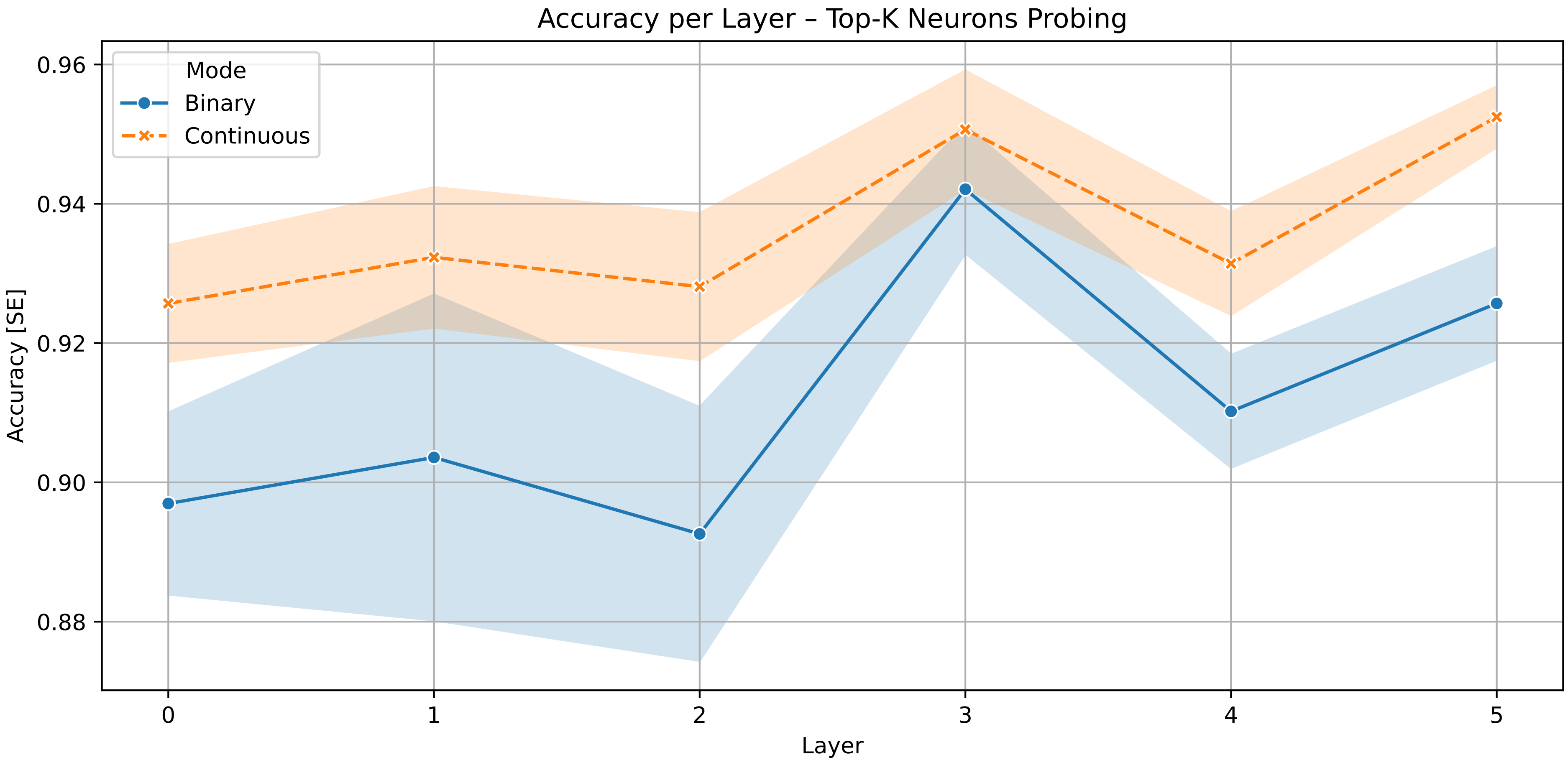

Overall F1 per Layer – All Methods
F1 Score [SE]
0.825
0.850
0.875
0.900
0.925
0.950
0.975
Layer
0
1
2
3
4
5
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

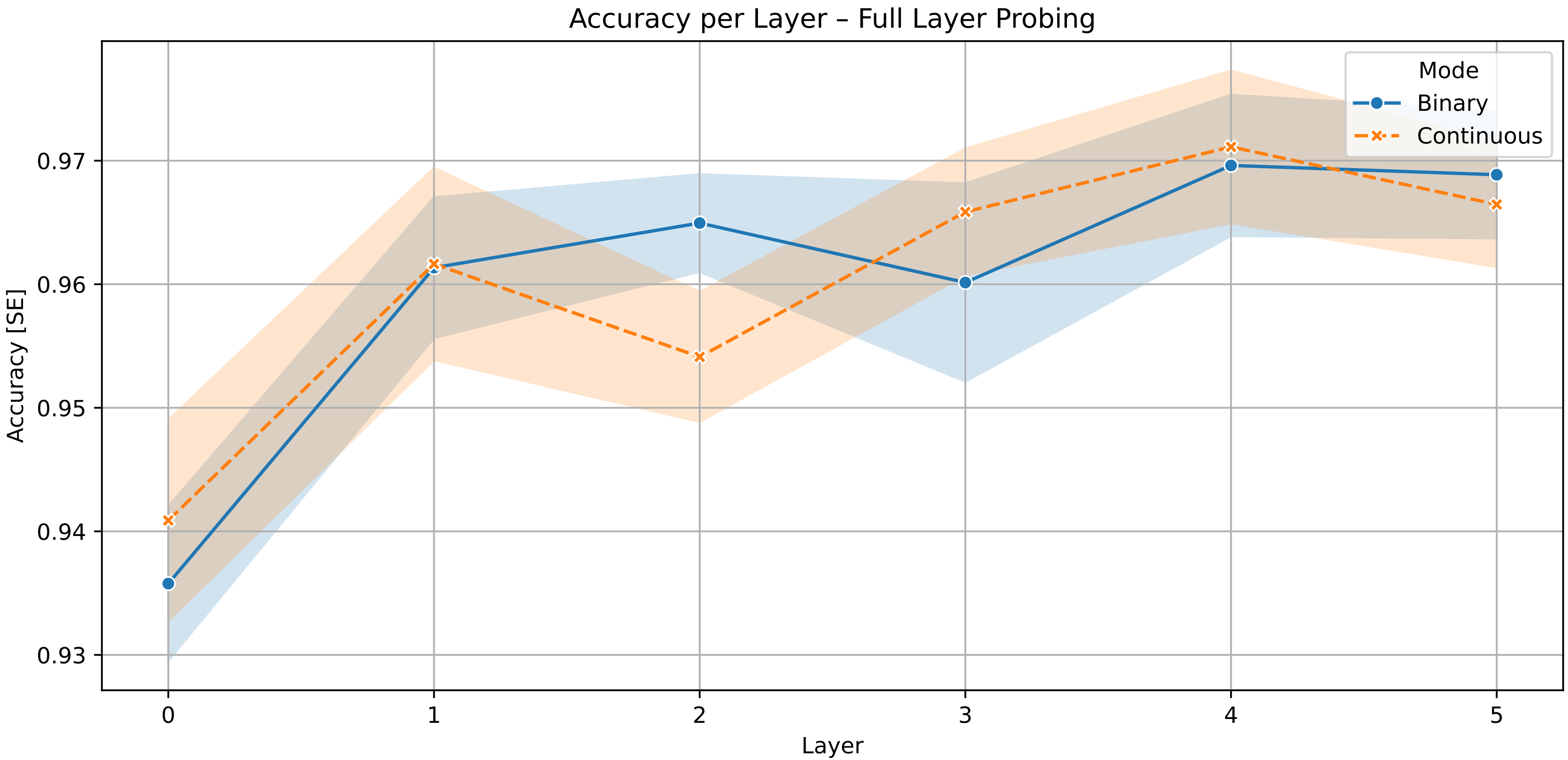


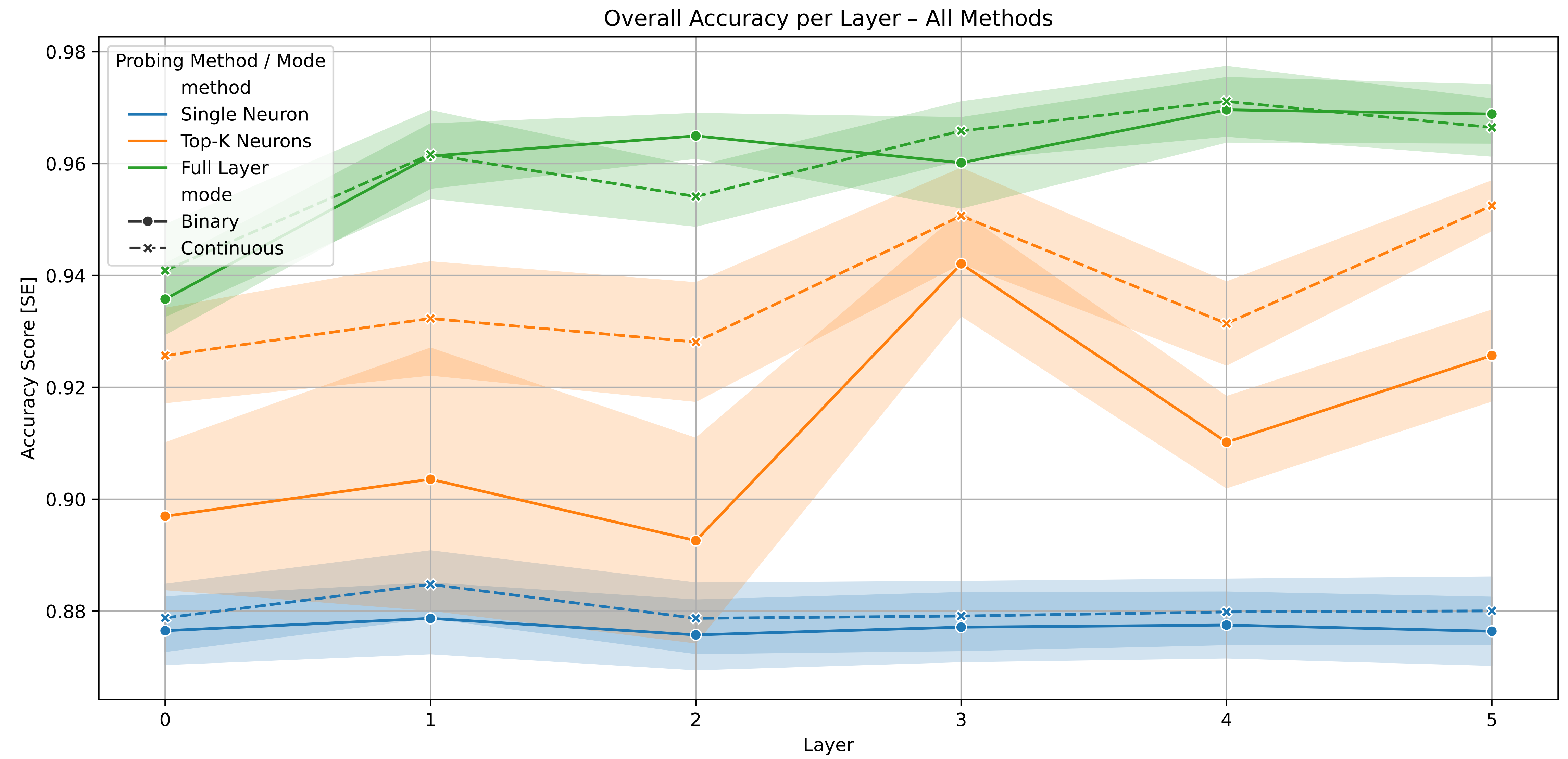
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	4.0	4.0
Full Layer	f1_max	0.9929	0.9976
Full Layer	f1_mean	0.9587	0.9592
Full Layer	f1_std	0.021	0.0203
Single Neuron	f1_best_layer	1.0	1.0
Single Neuron	f1_max	0.9851	0.9864
Single Neuron	f1_mean	0.8233	0.8386
Single Neuron	f1_std	0.077	0.0691
Top-K Neurons	f1_best_layer	3.0	5.0
Top-K Neurons	f1_max	0.9851	0.9851
Top-K Neurons	f1_mean	0.8919	0.9336
Top-K Neurons	f1_std	0.0581	0.0261


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.890
0.885
0.880
0.875
0.870
Layer
0
1
2
3
4
5


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.96
0.94
0.92
0.90
0.88
Layer
0
1
2
3
4
5


Accuracy per Layer – Full Layer Probing
Accuracy [SE]
0.97
0.96
0.95
0.94
0.93
Layer
0
1
2
3
4
5
Mode
Binary
Continuous


Overall Accuracy per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
Accuracy Score [SE]
0.98
0.96
0.94
0.92
0.90
0.88
Layer
0
1
2
3
4
5

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	4.0	4.0
Full Layer	accuracy_max	0.9928	0.9976
Full Layer	accuracy_mean	0.9601	0.96
Full Layer	accuracy_std	0.0196	0.0198
Single Neuron	accuracy_best_layer	1.0	1.0
Single Neuron	accuracy_max	0.9856	0.9868
Single Neuron	accuracy_mean	0.877	0.8802
Single Neuron	accuracy_std	0.054	0.0535
Top-K Neurons	accuracy_best_layer	3.0	5.0
Top-K Neurons	accuracy_max	0.9856	0.9856
Top-K Neurons	accuracy_mean	0.9119	0.9368
Top-K Neurons	accuracy_std	0.0425	0.025