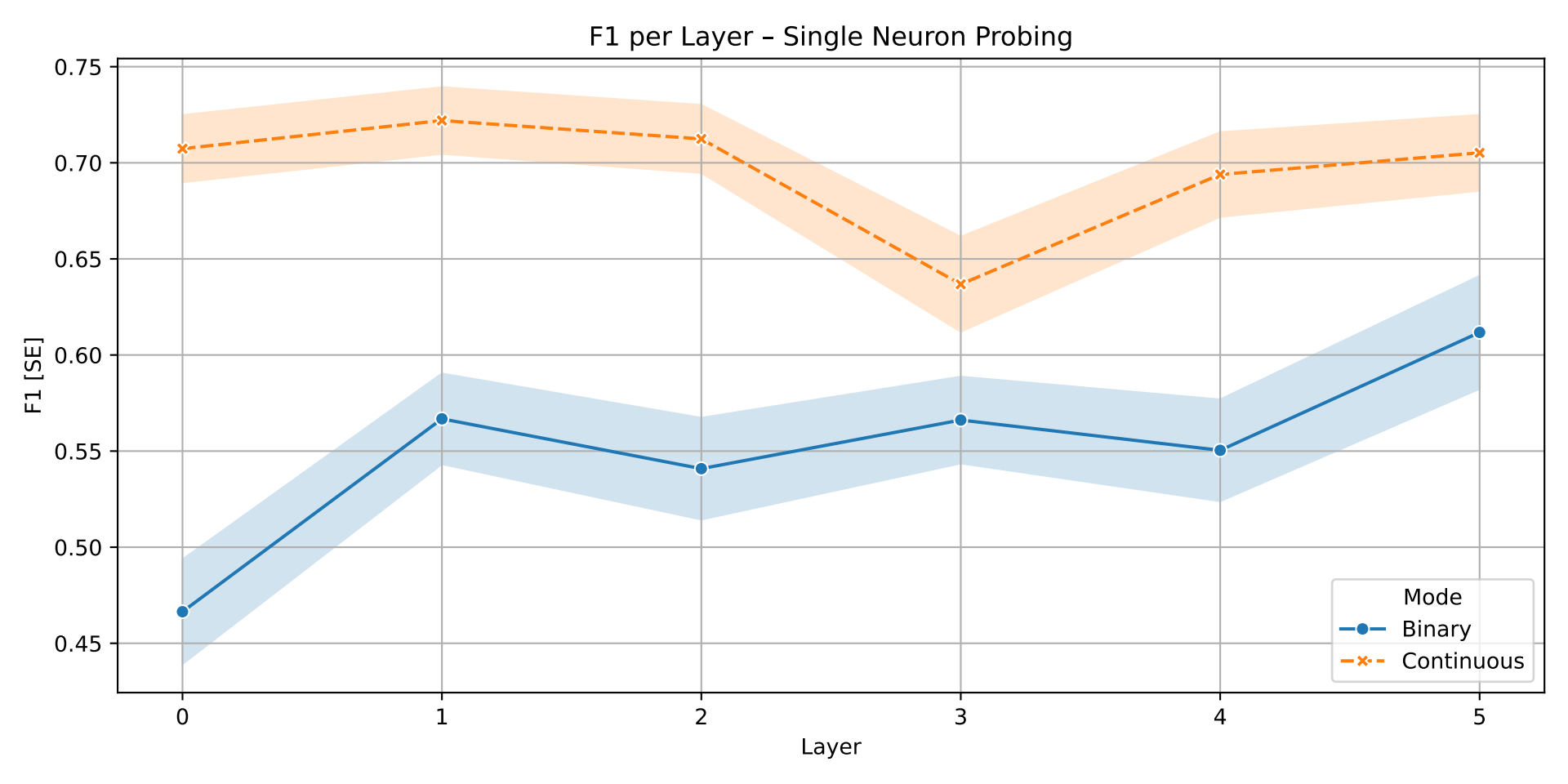
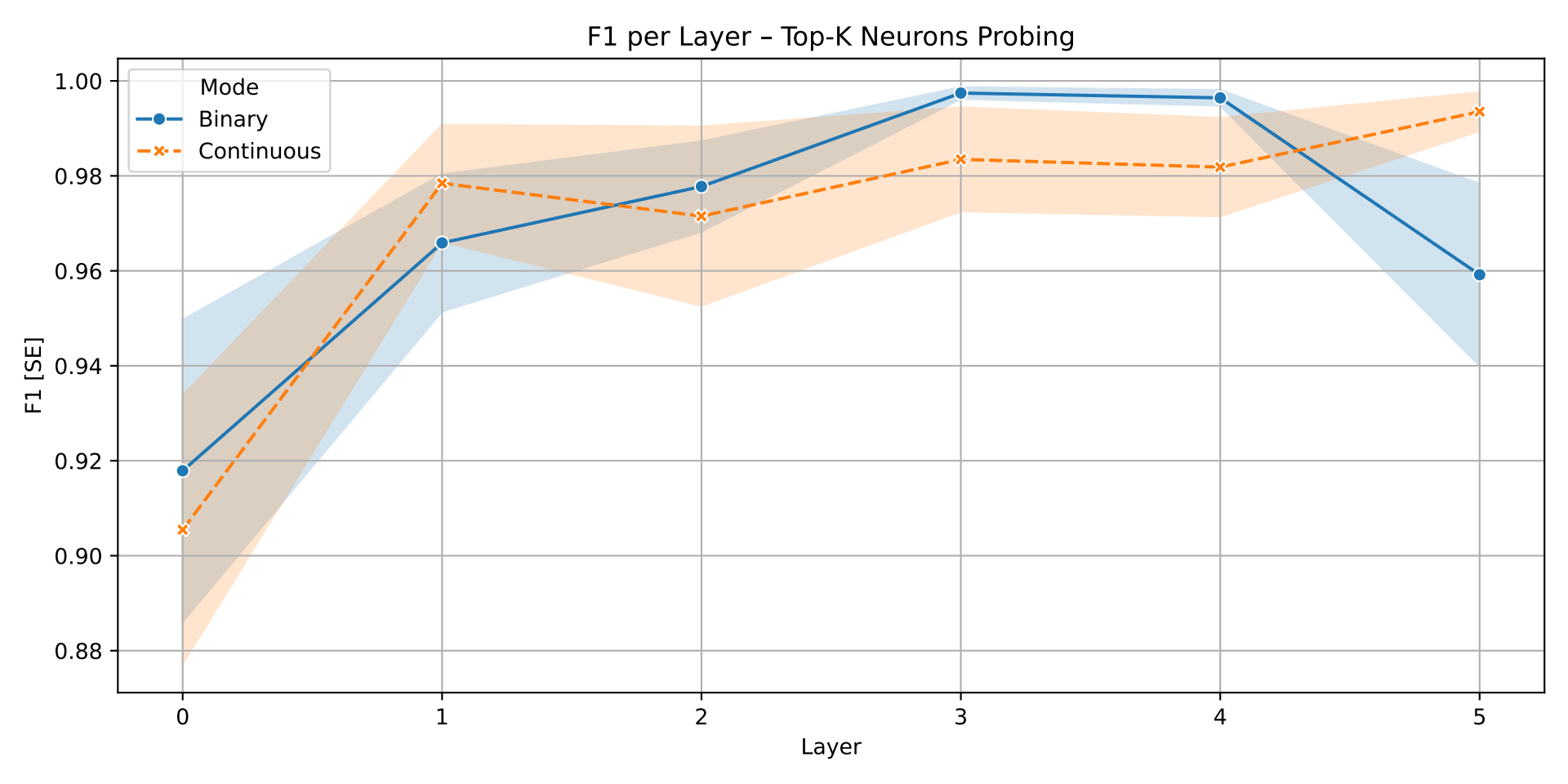
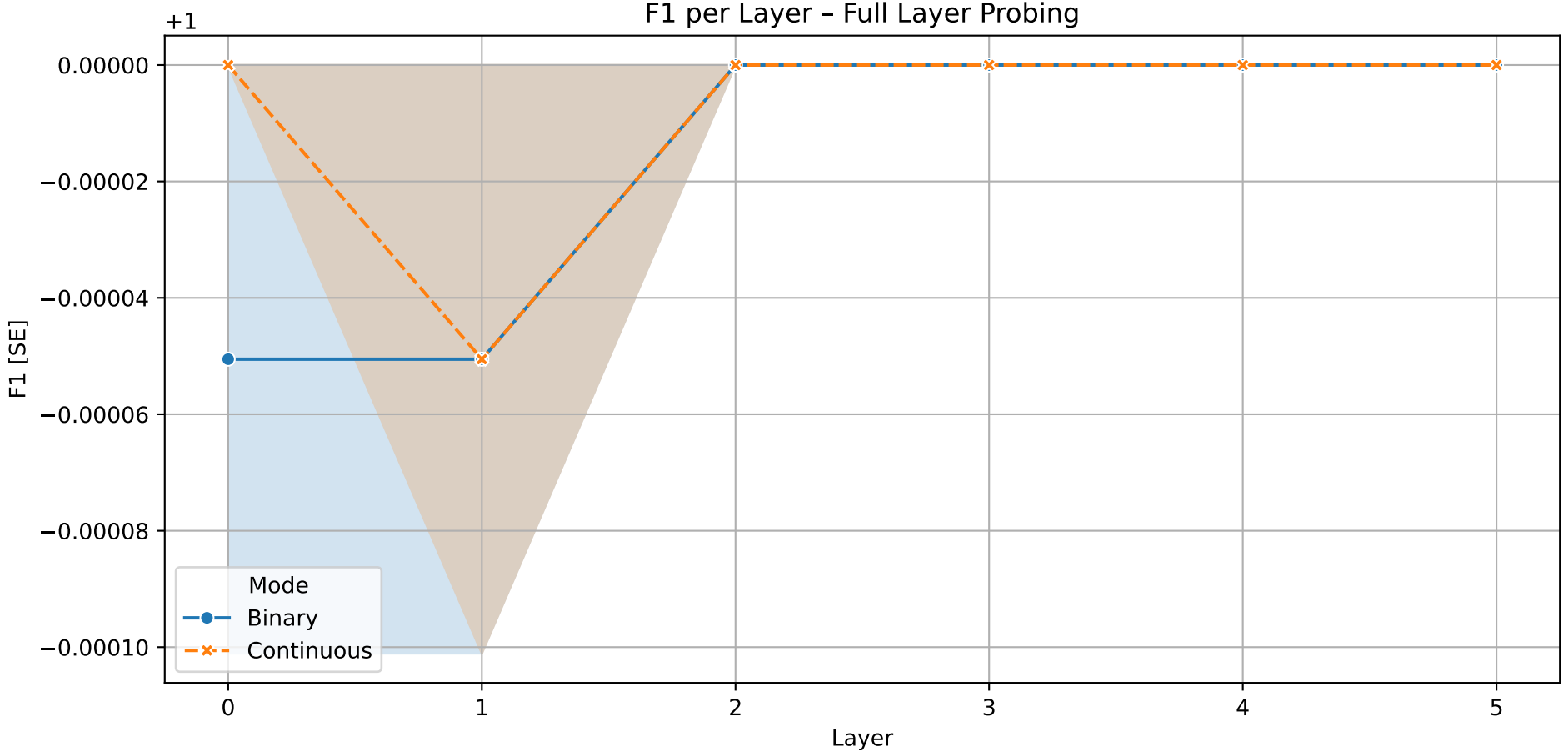
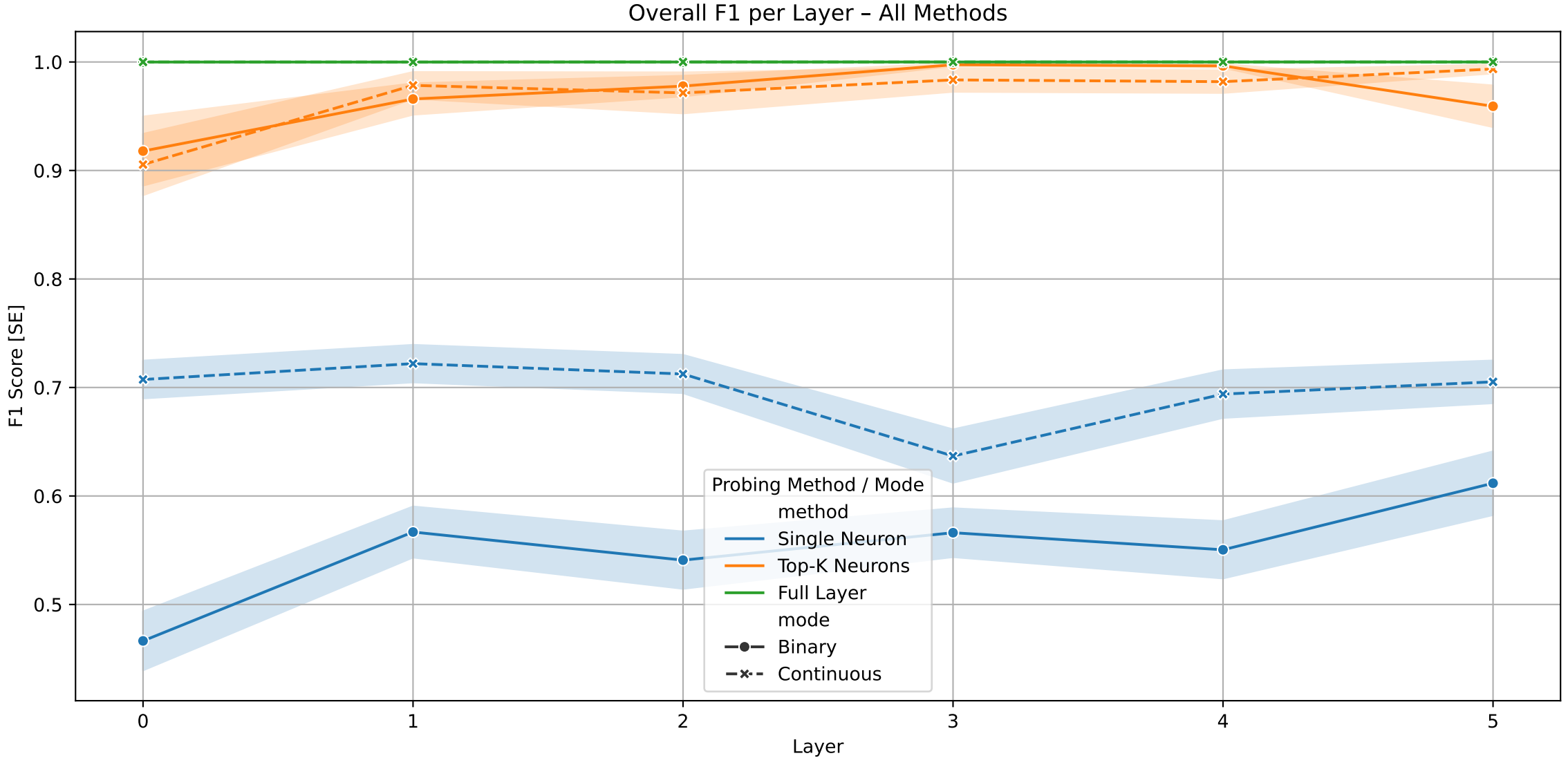

F1 per Layer – Single Neuron Probing
F1 [SE]
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0
1
2
3
4
5
Layer
Mode
Binary
Continuous

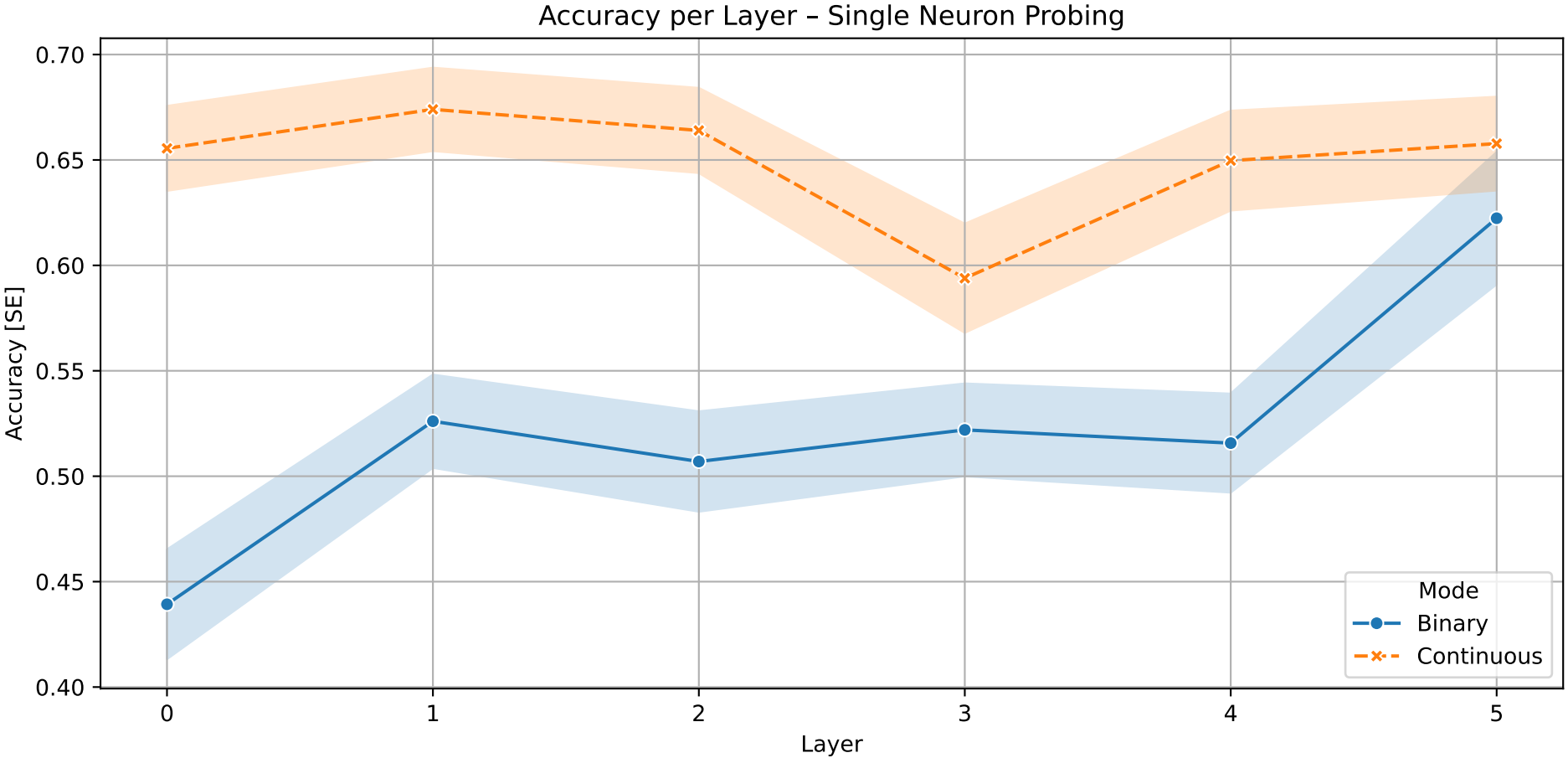

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0
1
2
3
4
5
Layer

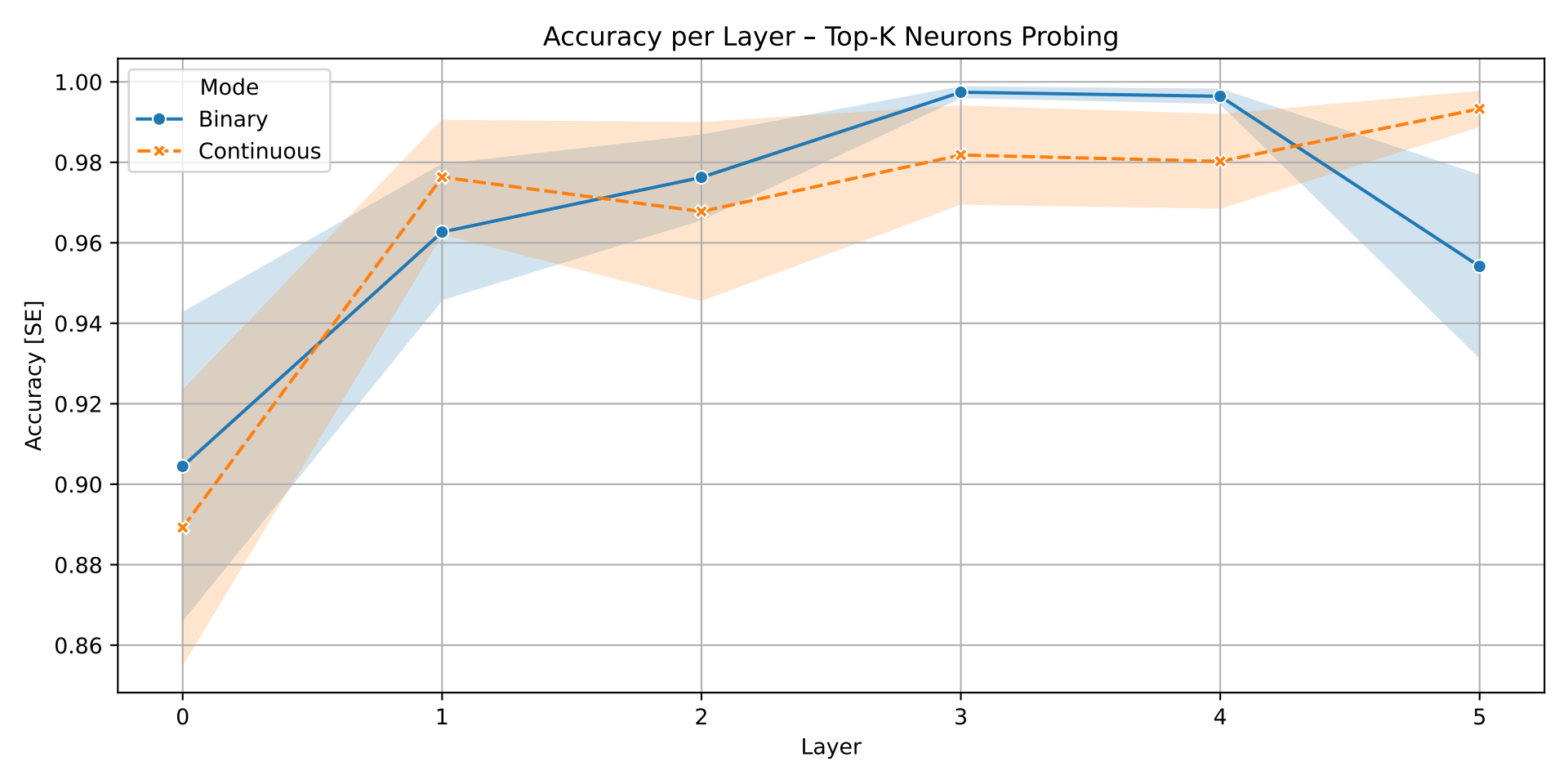

F1 per Layer – Full Layer Probing
+1
0.00000
−0.00002
−0.00004
−0.00006
−0.00008
−0.00010
F1 [SE]
0
1
2
3
4
5
Layer
Mode
Binary
Continuous

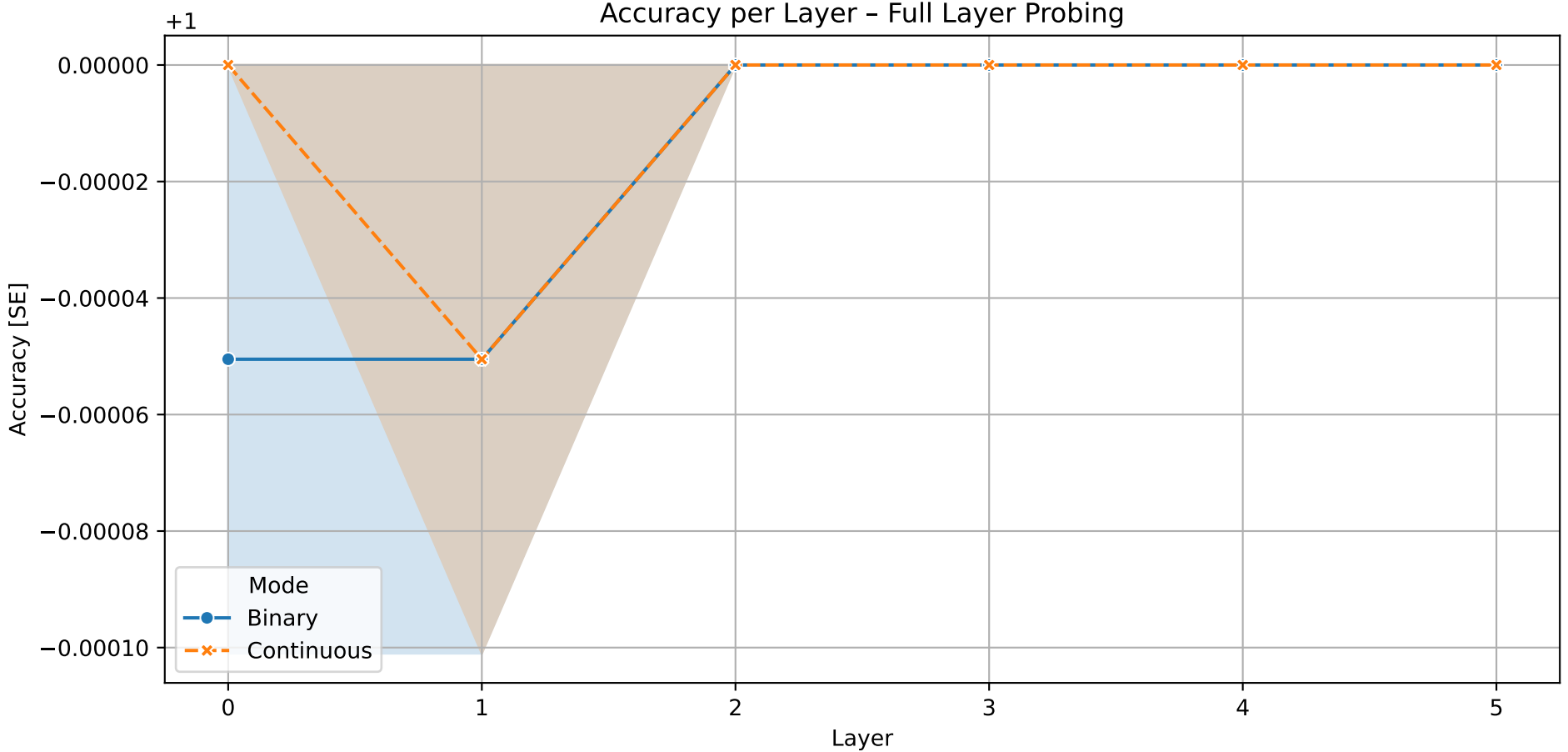

Overall F1 per Layer – All Methods
F1 Score [SE]
1.0
0.9
0.8
0.7
0.6
0.5
0
1
2
3
4
5
Layer
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

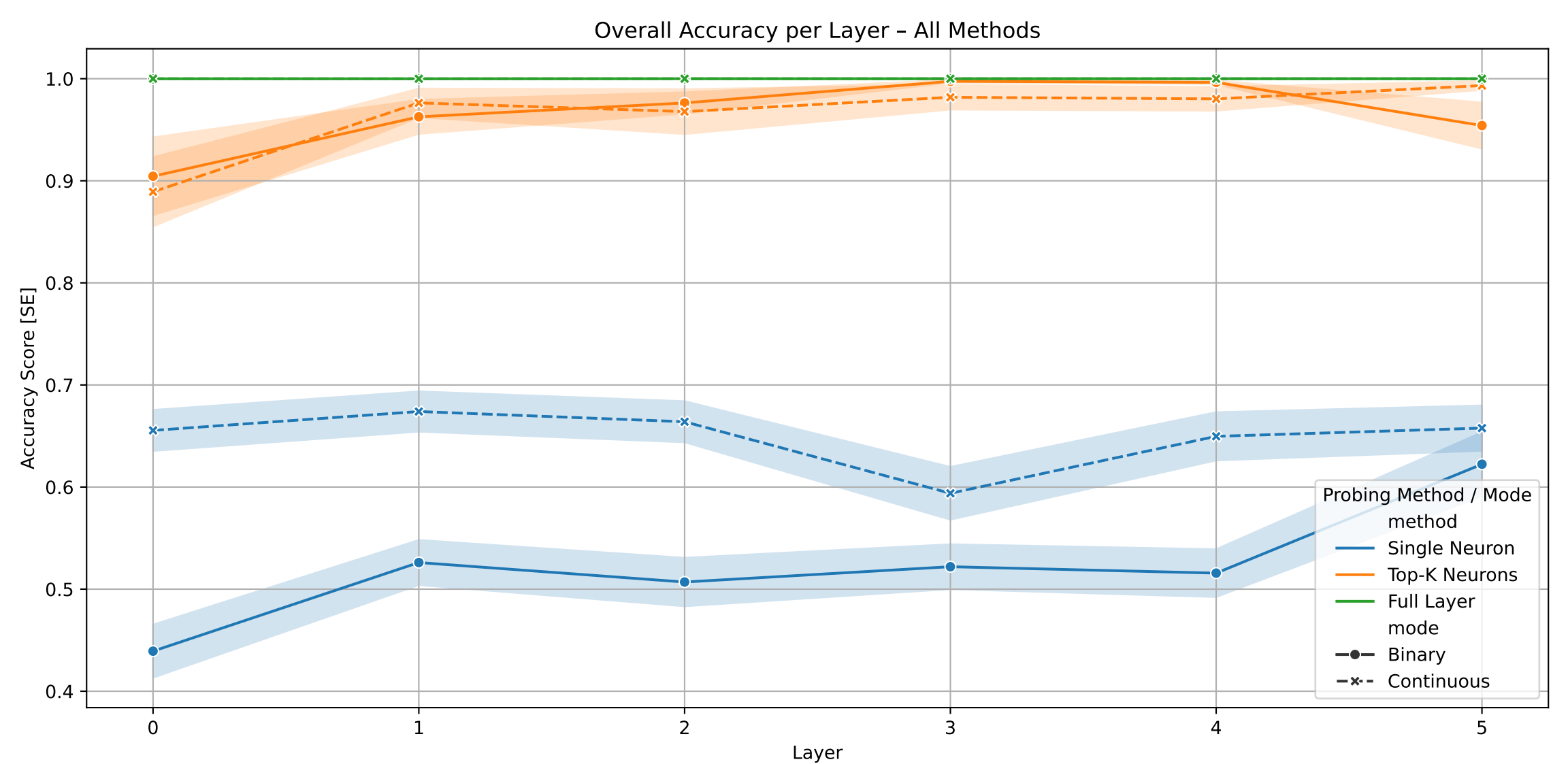
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	2.0	0.0
Full Layer	f1_max	1.0	1.0
Full Layer	f1_mean	1.0	1.0
Full Layer	f1_std	0.0001	0.0001
Single Neuron	f1_best_layer	5.0	1.0
Single Neuron	f1_max	1.0	1.0
Single Neuron	f1_mean	0.5504	0.6963
Single Neuron	f1_std	0.235	0.1794
Top-K Neurons	f1_best_layer	3.0	5.0
Top-K Neurons	f1_max	1.0	1.0
Top-K Neurons	f1_mean	0.9691	0.969
Top-K Neurons	f1_std	0.0524	0.0521


Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0
1
2
3
4
5
Layer
Mode
Binary
Continuous


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0
1
2
3
4
5
Layer


Accuracy per Layer – Full Layer Probing
+1
0.00000
−0.00002
−0.00004
−0.00006
−0.00008
−0.00010
Accuracy [SE]
0
1
2
3
4
5
Layer
Mode
Binary
Continuous


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
1
2
3
4
5
Layer
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	2.0	0.0
Full Layer	accuracy_max	1.0	1.0
Full Layer	accuracy_mean	1.0	1.0
Full Layer	accuracy_std	0.0001	0.0001
Single Neuron	accuracy_best_layer	5.0	1.0
Single Neuron	accuracy_max	1.0	1.0
Single Neuron	accuracy_mean	0.5221	0.6491
Single Neuron	accuracy_std	0.2283	0.1977
Top-K Neurons	accuracy_best_layer	3.0	5.0
Top-K Neurons	accuracy_max	1.0	1.0
Top-K Neurons	accuracy_mean	0.9652	0.9648
Top-K Neurons	accuracy_std	0.0617	0.0613