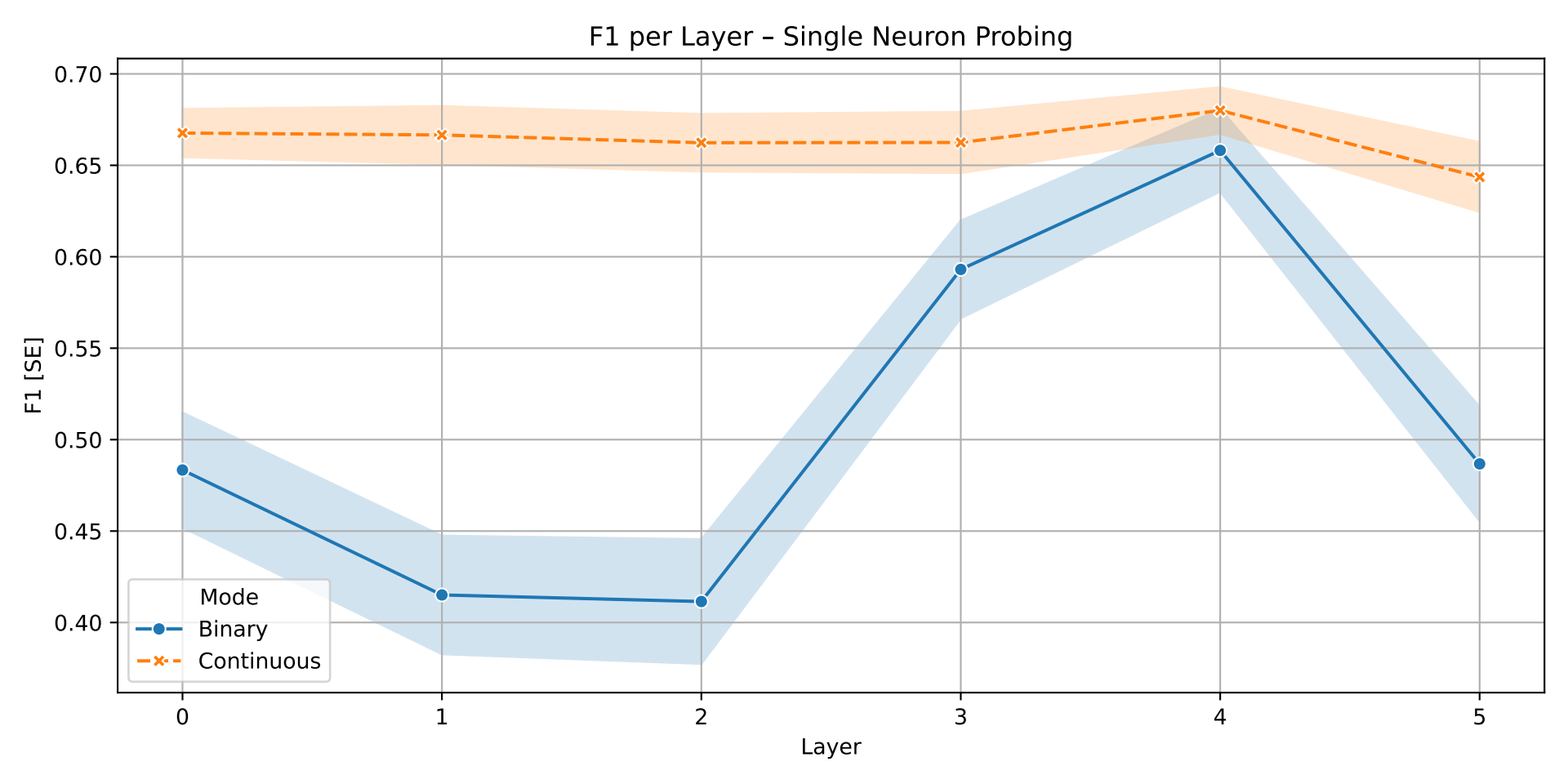
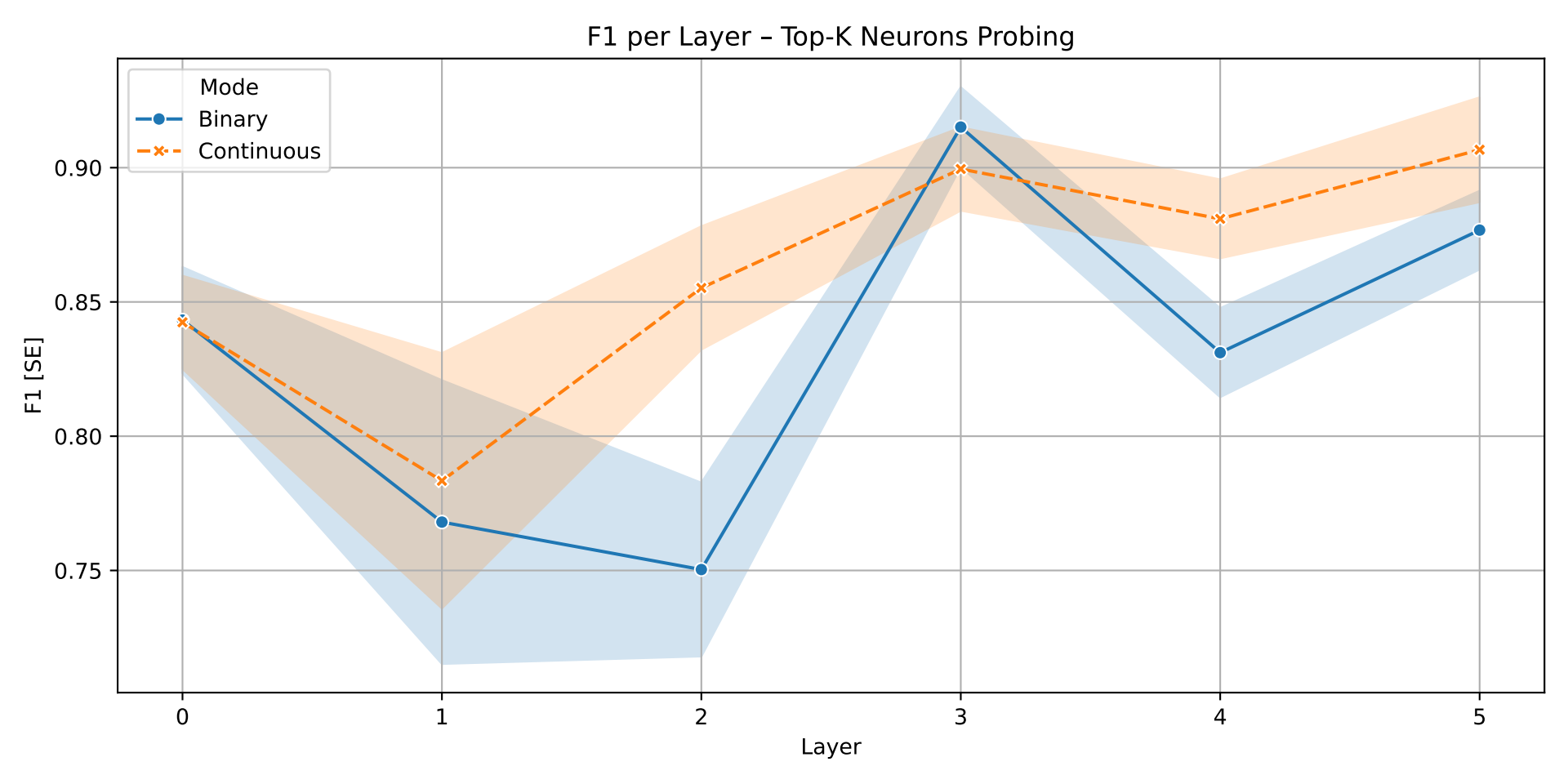
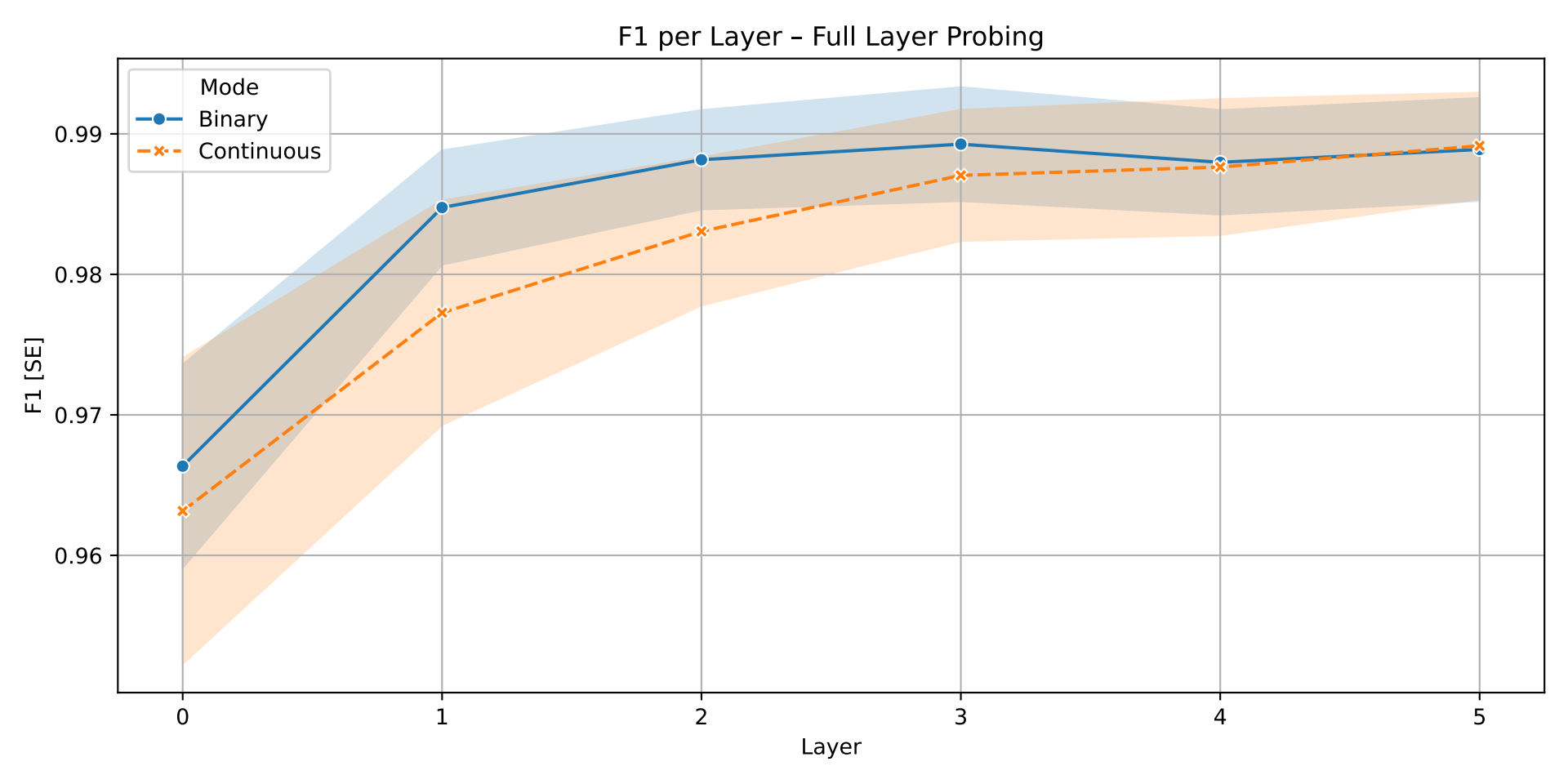
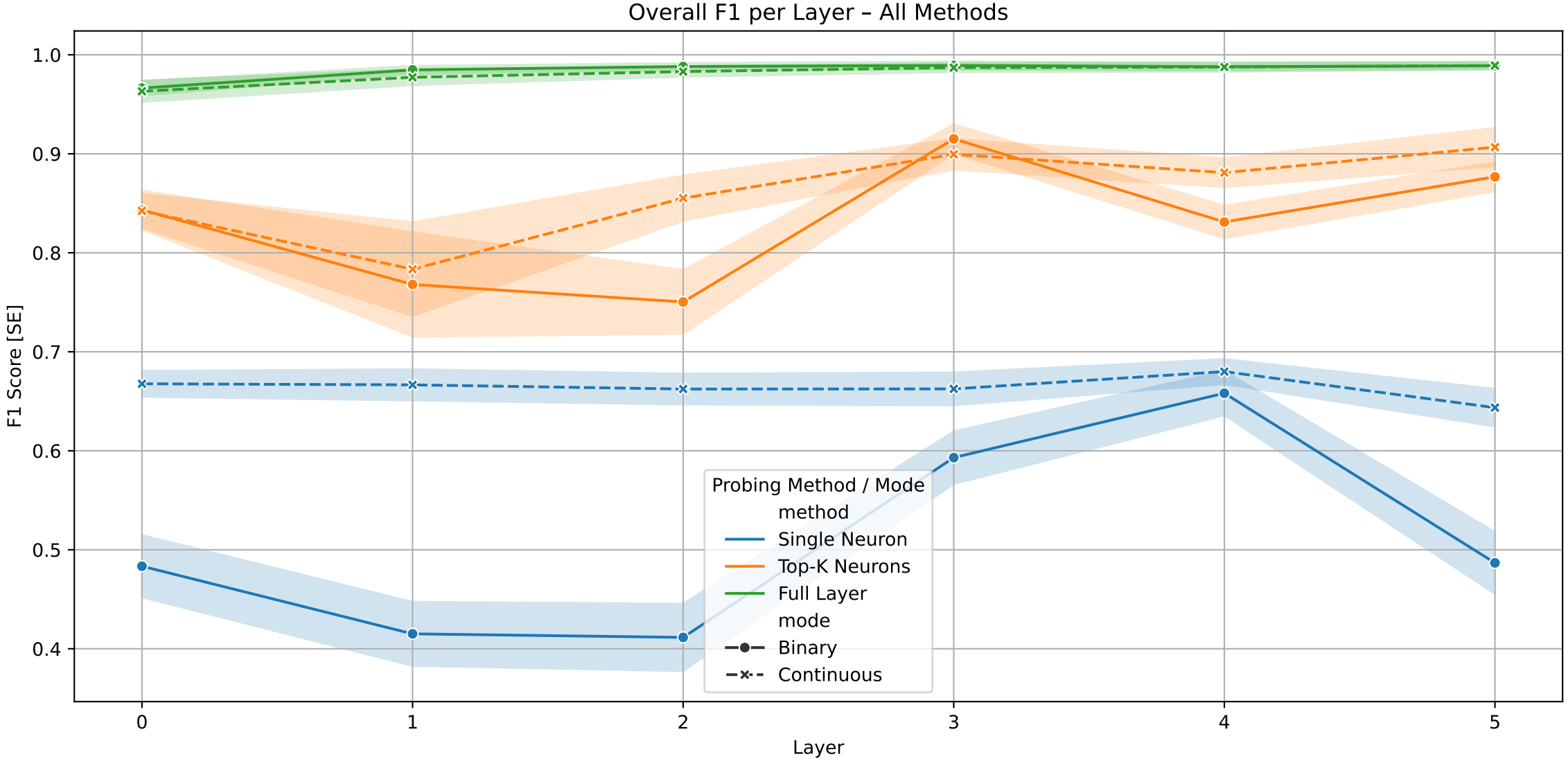

F1 per Layer – Single Neuron Probing
F1 [SE]
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0
1
2
3
4
5
Layer
Mode
Binary
Continuous

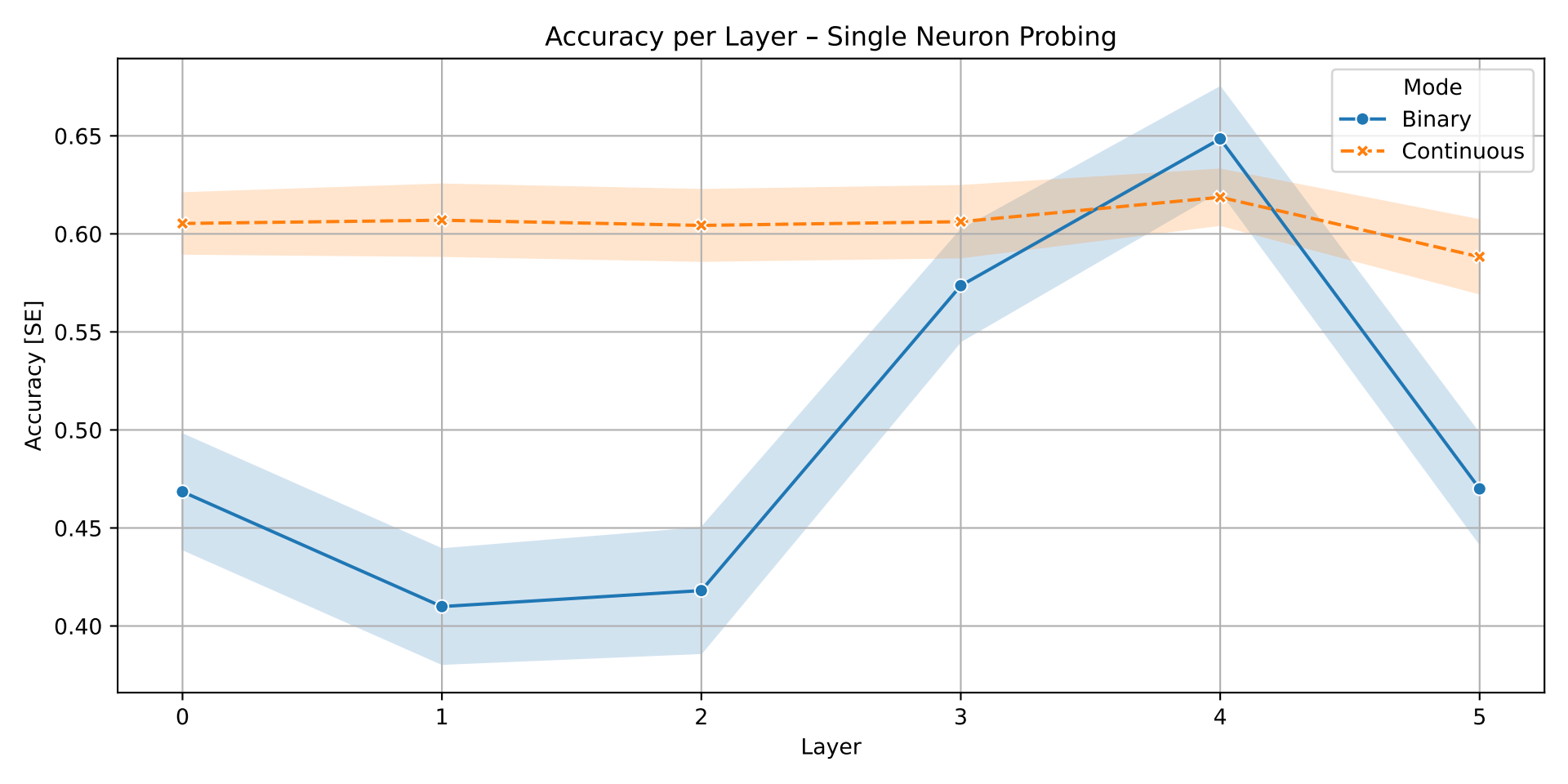

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
0.90
0.85
0.80
0.75
F1 [SE]
0
1
2
3
4
5
Layer

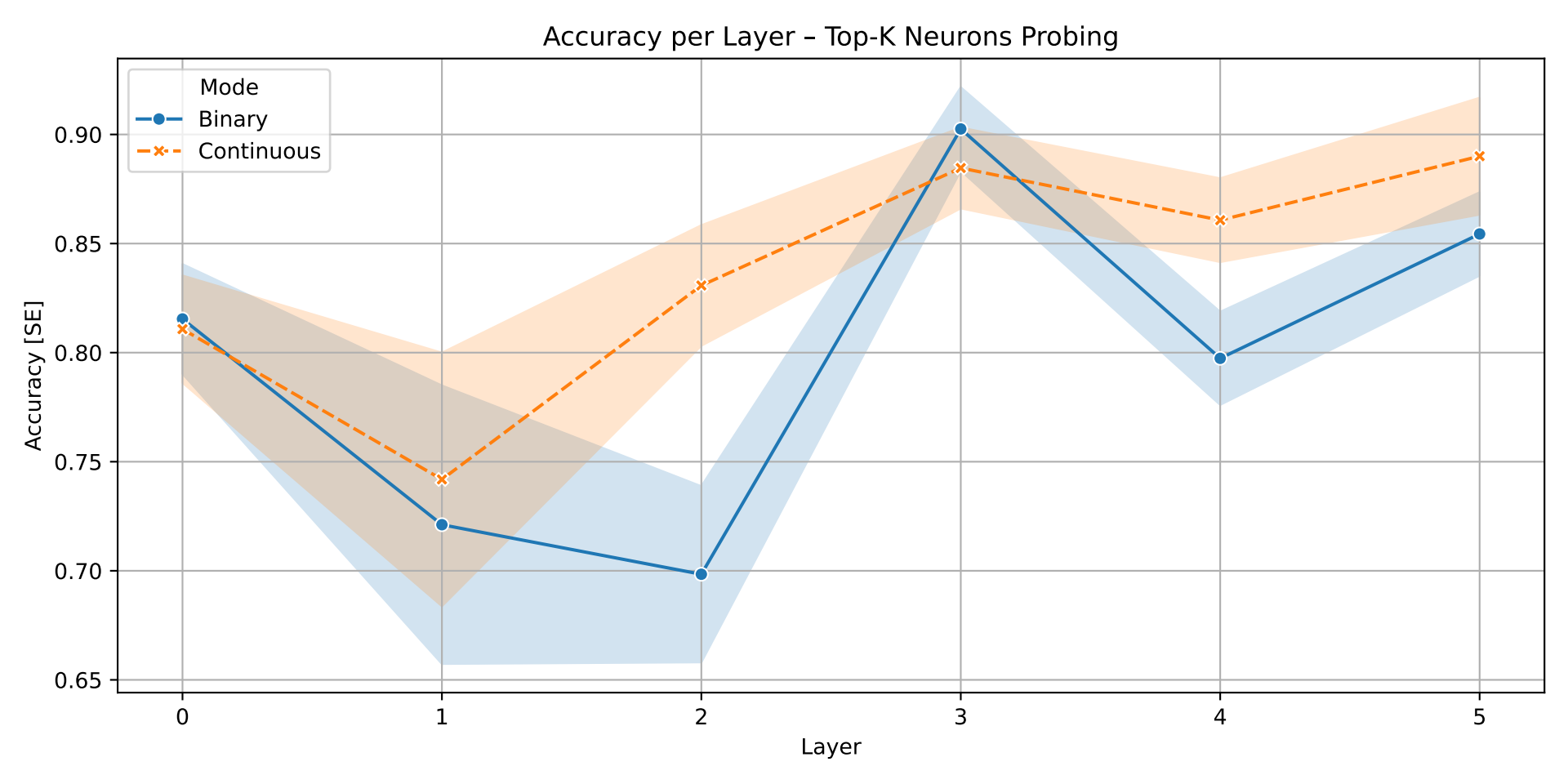

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
0.99
0.98
0.97
0.96
F1 [SE]
0
1
2
3
4
5
Layer

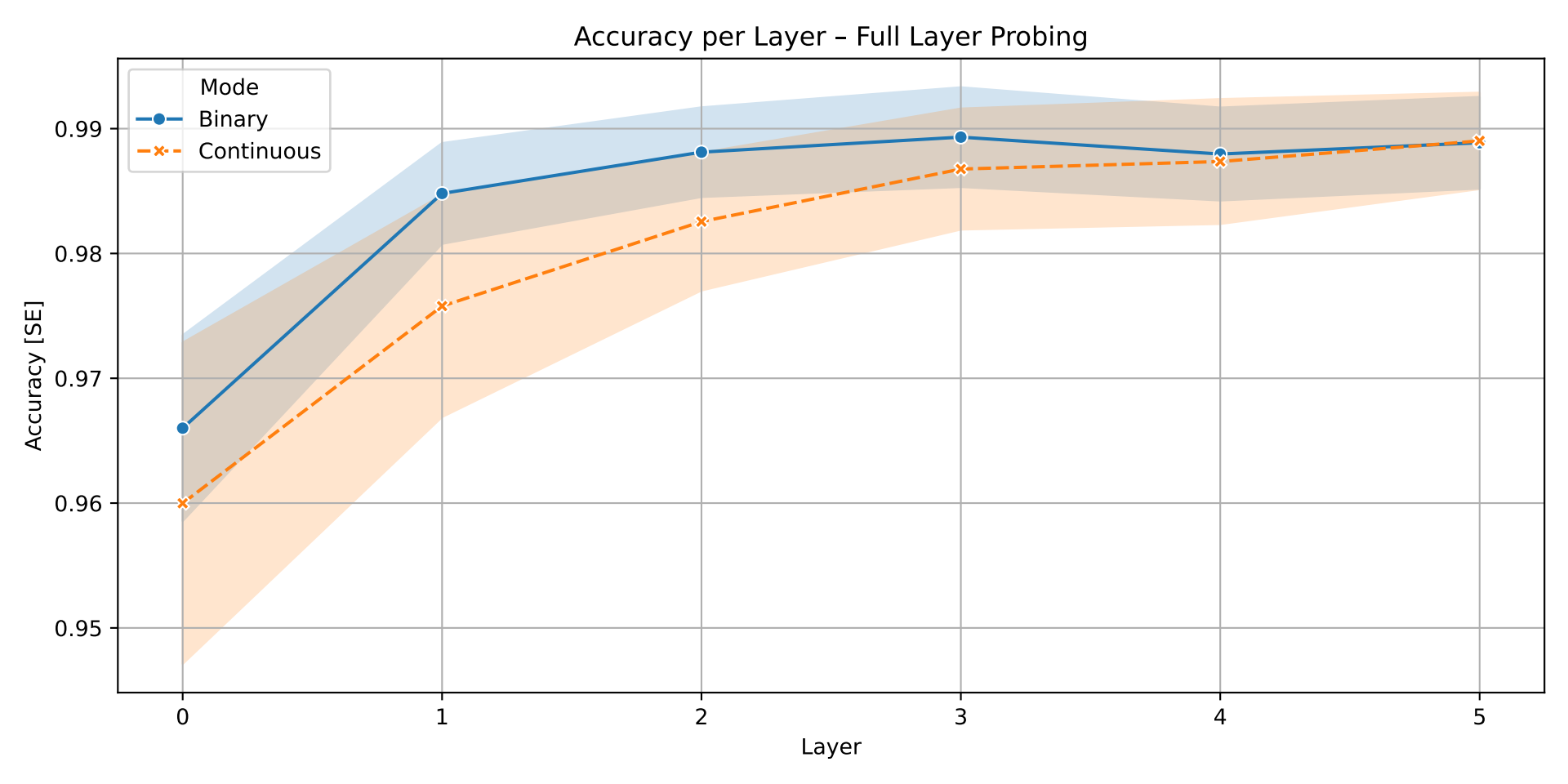

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
1
2
3
4
5
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

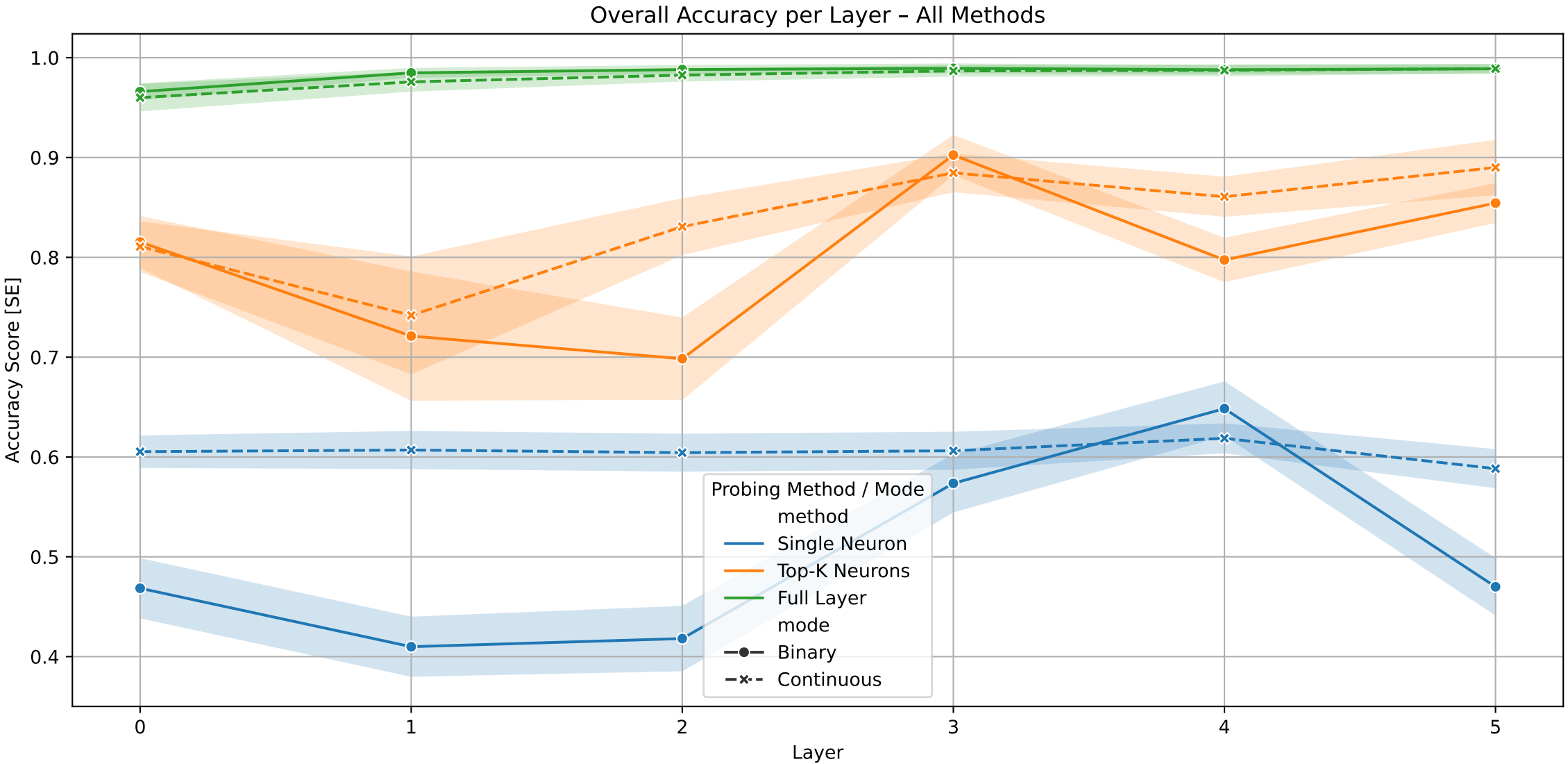
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	3.0	5.0
Full Layer	f1_max	0.9976	0.9976
Full Layer	f1_mean	0.9842	0.9812
Full Layer	f1_std	0.0147	0.0201
Single Neuron	f1_best_layer	4.0	4.0
Single Neuron	f1_max	0.9851	0.9837
Single Neuron	f1_mean	0.508	0.6638
Single Neuron	f1_std	0.2816	0.1396
Top-K Neurons	f1_best_layer	3.0	5.0
Top-K Neurons	f1_max	0.9792	0.9662
Top-K Neurons	f1_mean	0.8308	0.8614
Top-K Neurons	f1_std	0.0961	0.0801


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


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.90
0.85
0.80
0.75
0.70
0.65
0
1
2
3
4
5
Layer


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
0.99
0.98
0.97
0.96
0.95
Accuracy [SE]
0
1
2
3
4
5
Layer


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
Layer
1.0
0.9
0.8
0.7
0.6
0.5
0.4
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	3.0	5.0
Full Layer	accuracy_max	0.9976	0.9976
Full Layer	accuracy_mean	0.9842	0.9802
Full Layer	accuracy_std	0.0149	0.0225
Single Neuron	accuracy_best_layer	4.0	4.0
Single Neuron	accuracy_max	0.9856	0.9844
Single Neuron	accuracy_mean	0.4981	0.605
Single Neuron	accuracy_std	0.2702	0.153
Top-K Neurons	accuracy_best_layer	3.0	5.0
Top-K Neurons	accuracy_max	0.9795	0.9651
Top-K Neurons	accuracy_mean	0.7982	0.8365
Top-K Neurons	accuracy_std	0.1187	0.0998