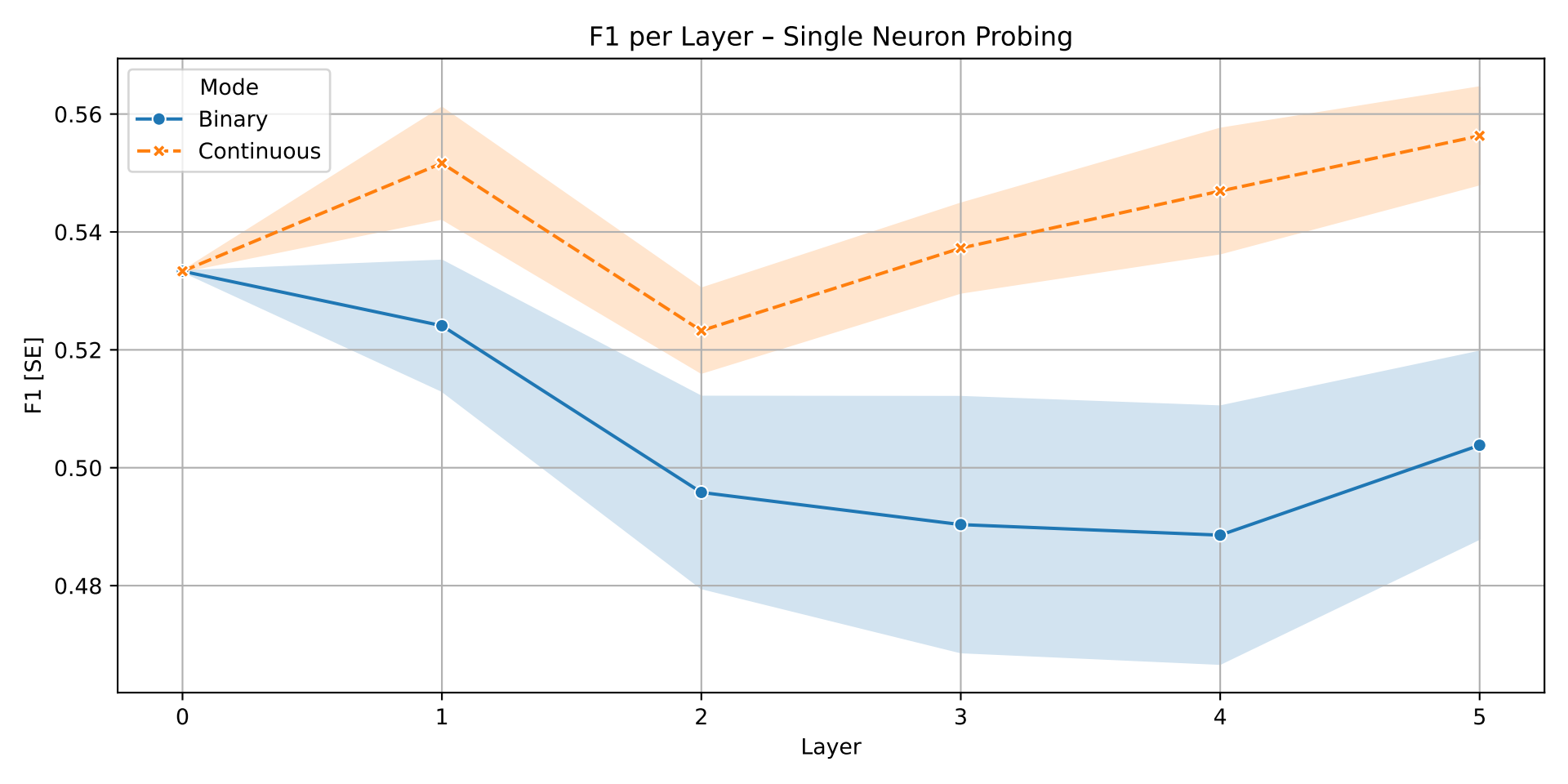
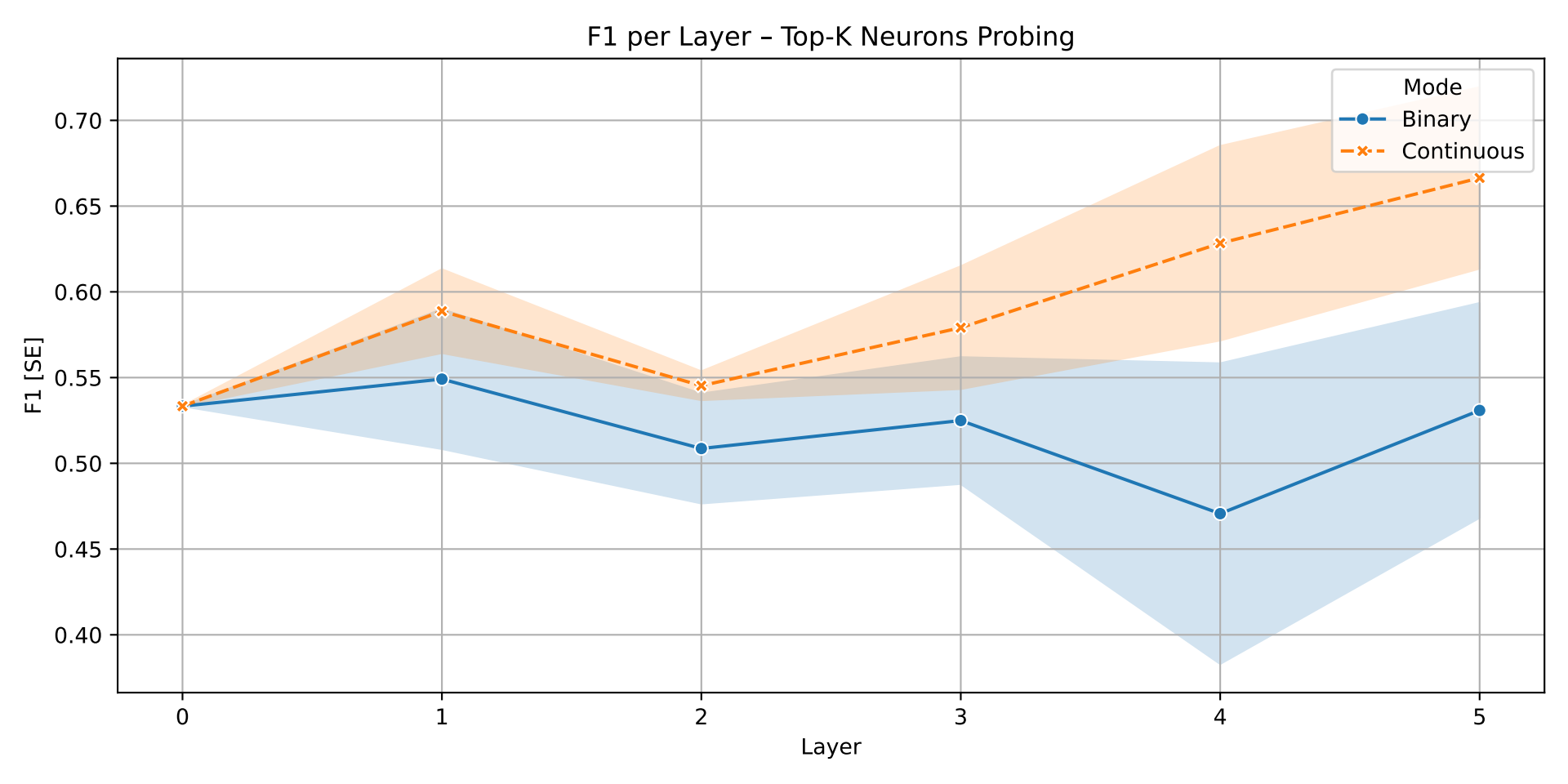
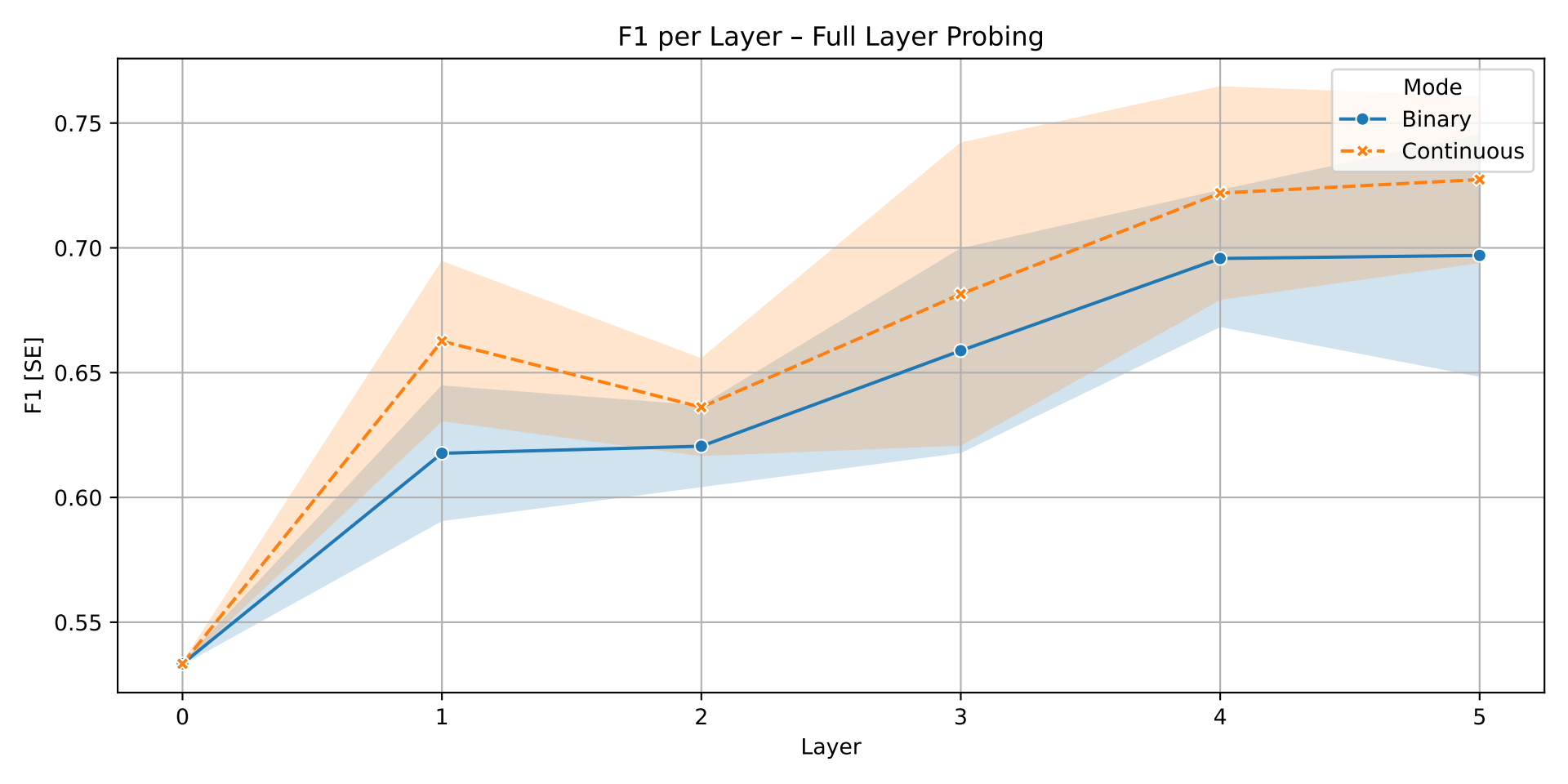
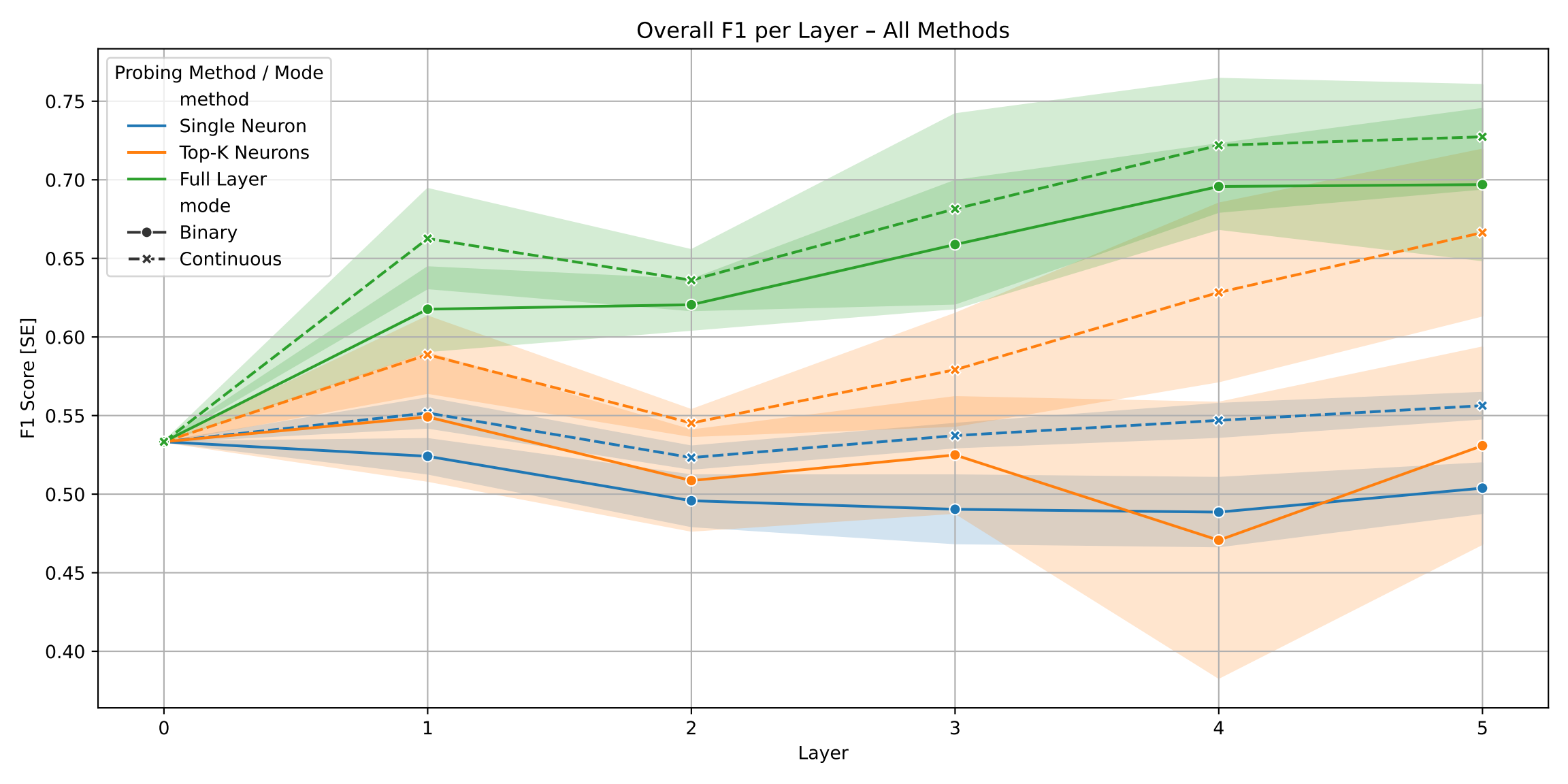

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
0.56
0.54
0.52
0.50
0.48
F1 [SE]
0
1
2
3
4
5
Layer

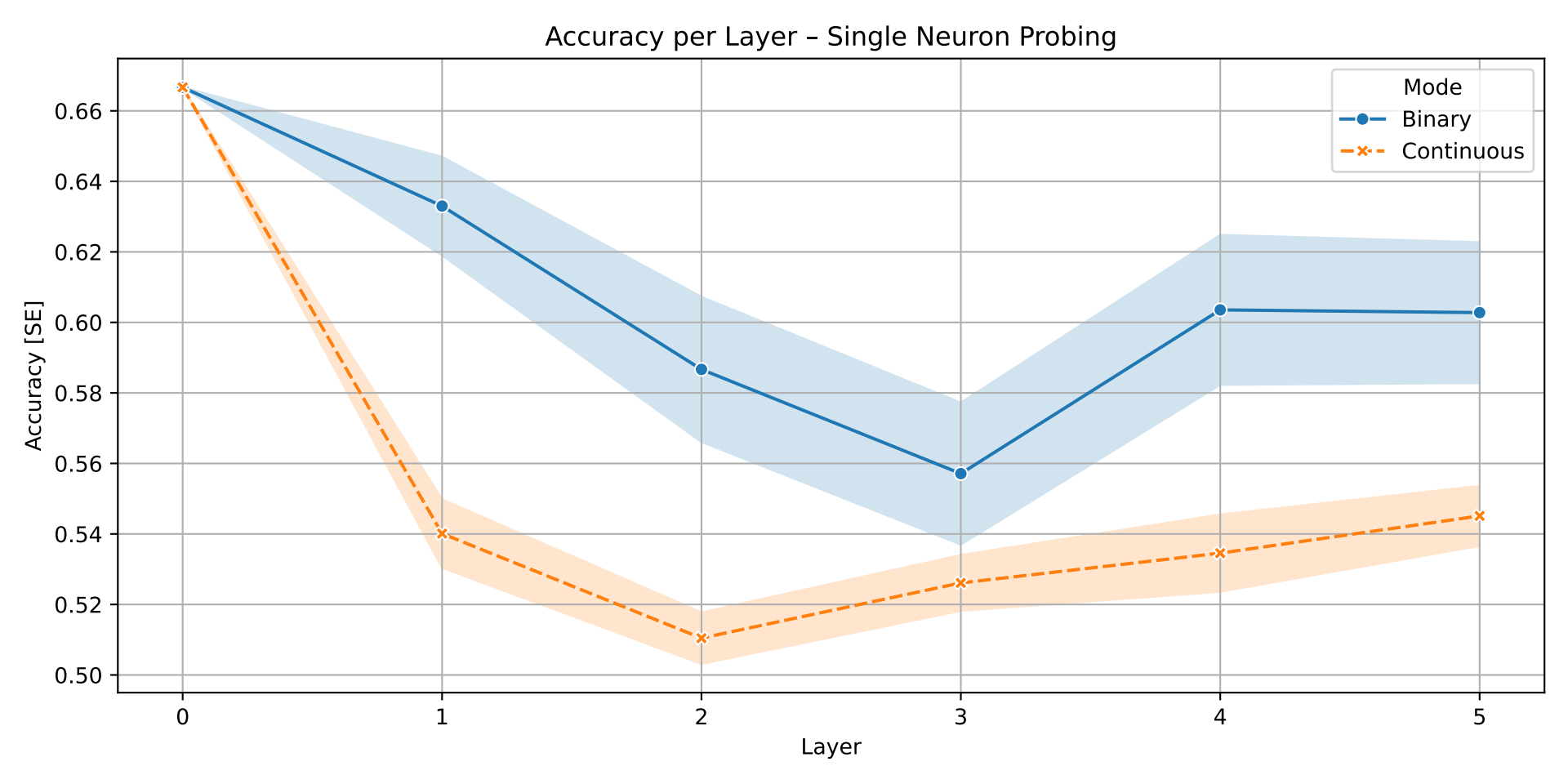

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0
1
2
3
4
5
Layer

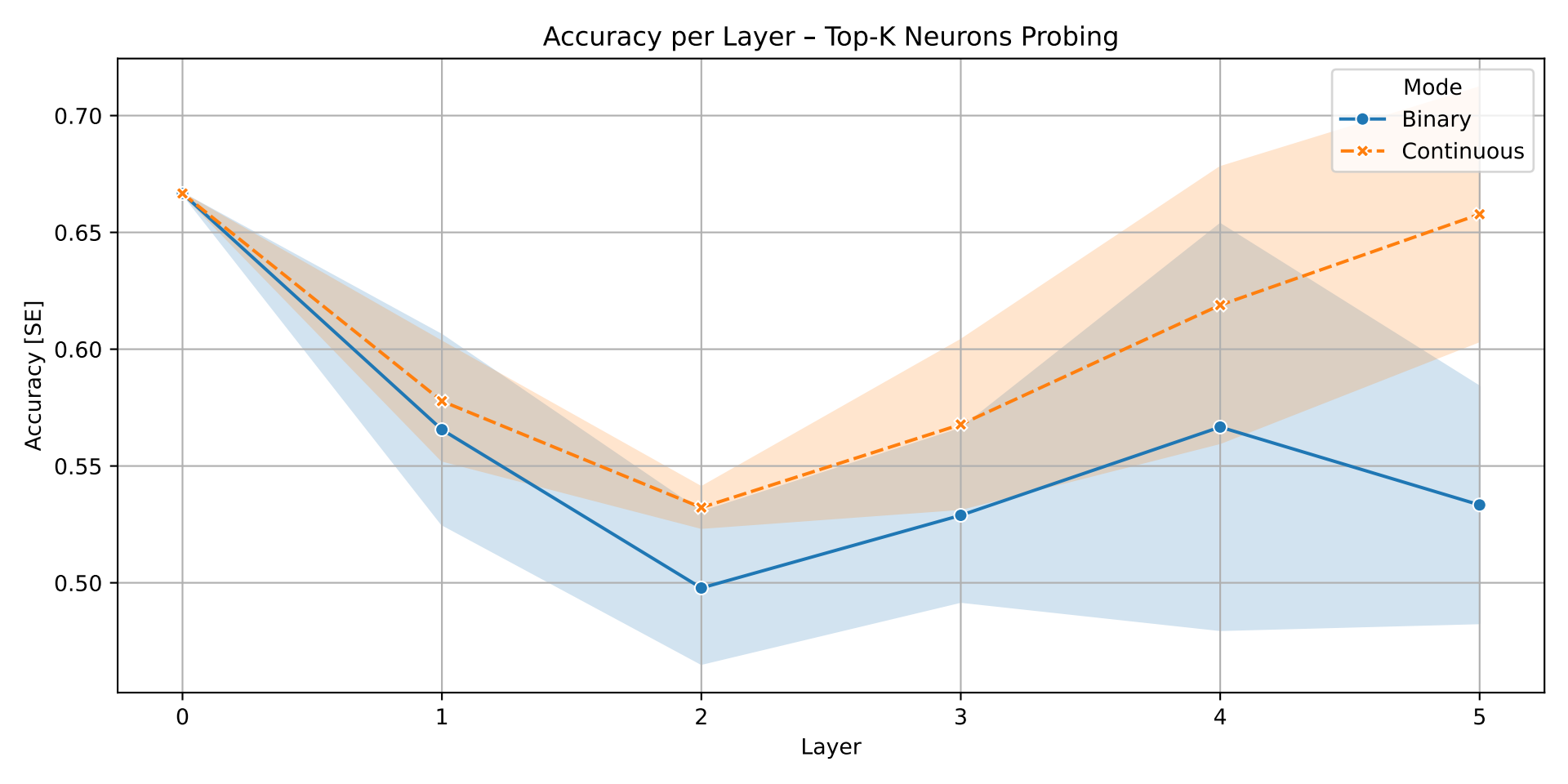

F1 per Layer – Full Layer Probing
F1 [SE]
0.75
0.70
0.65
0.60
0.55
0
1
2
3
4
5
Layer
Mode
Binary
Continuous

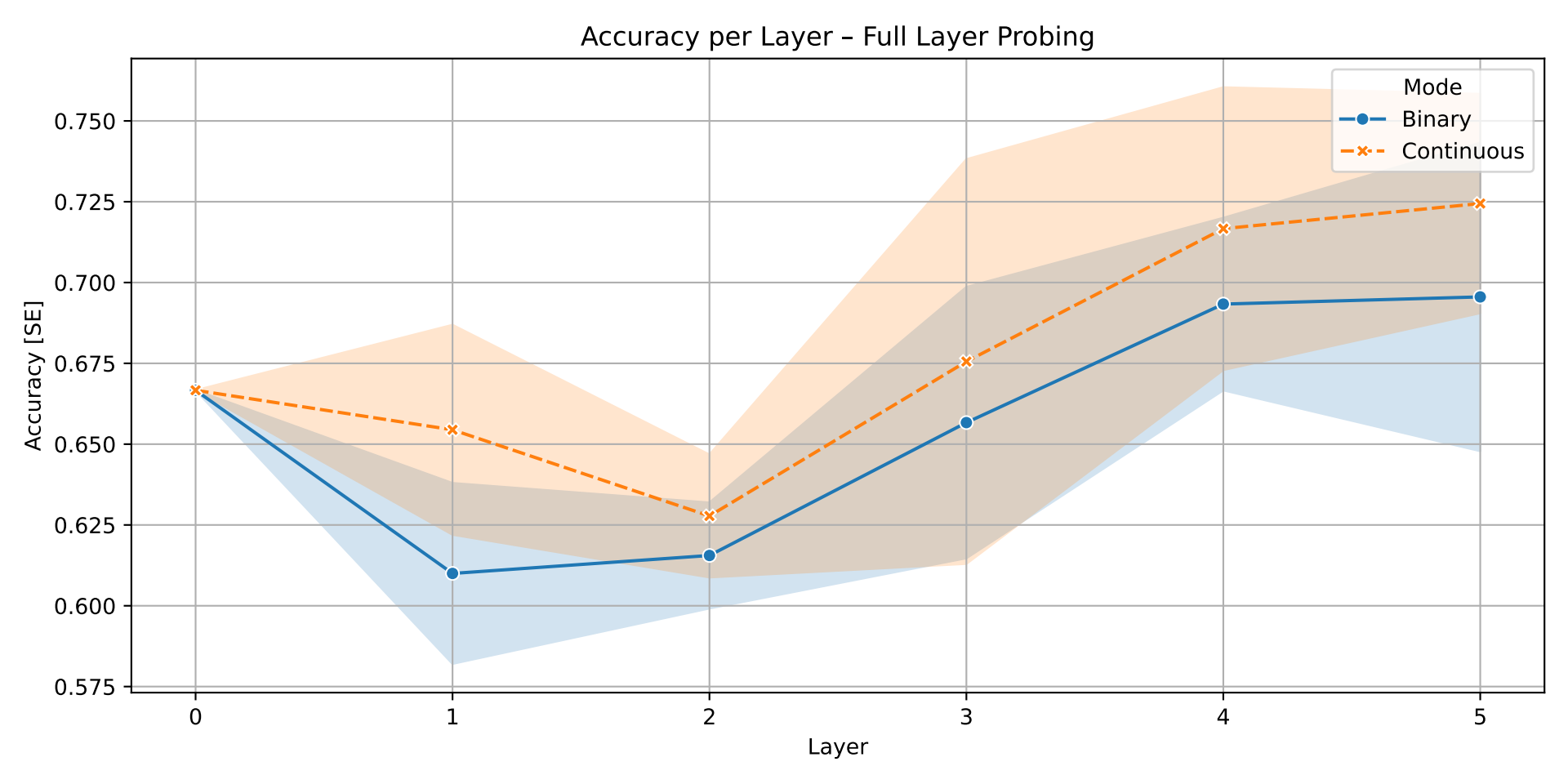

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0
1
2
3
4
5
Layer

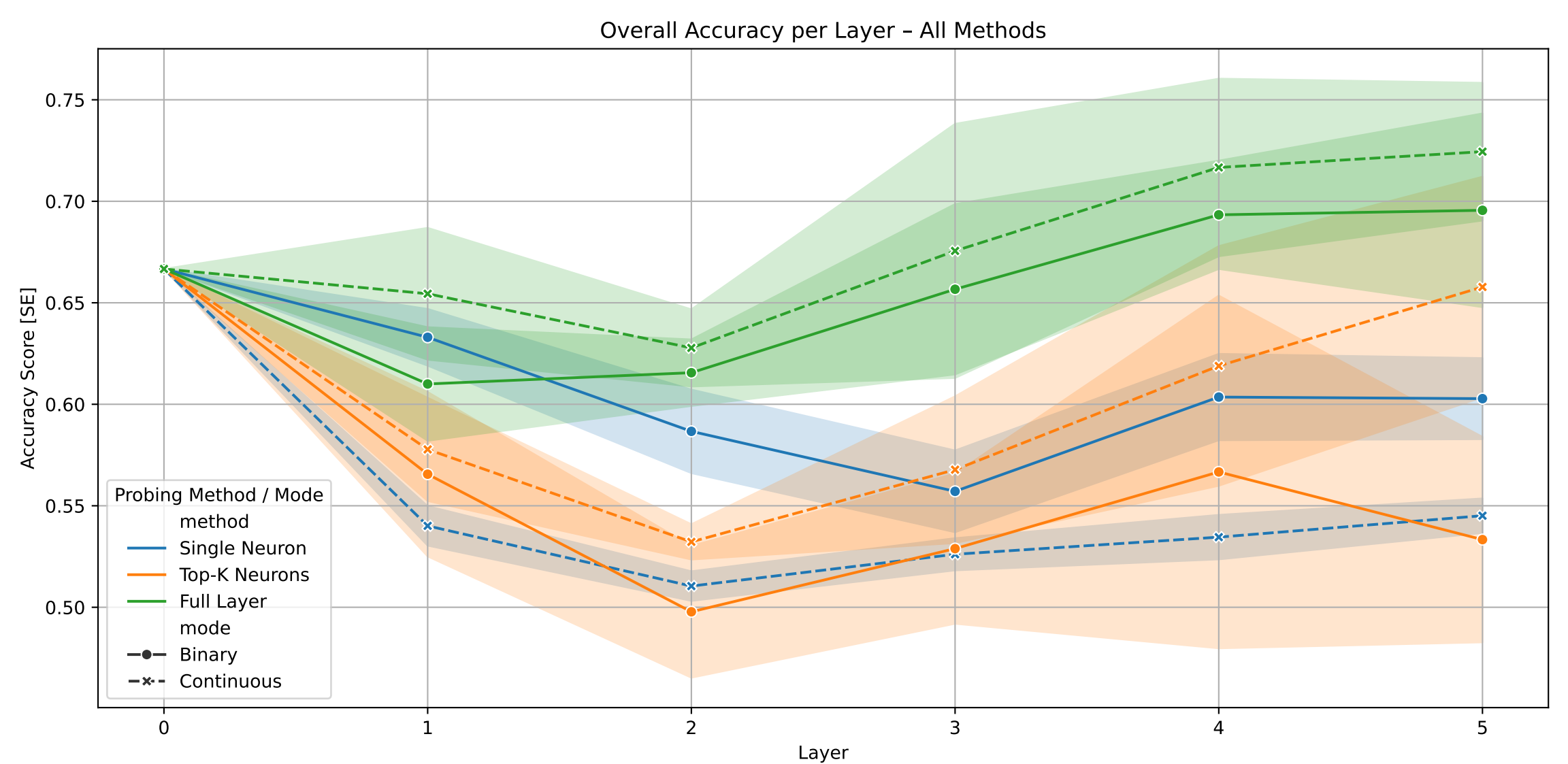
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	5.0	5.0
Full Layer	f1_max	0.7713	0.7751
Full Layer	f1_mean	0.6372	0.6605
Full Layer	f1_std	0.0731	0.0853
Single Neuron	f1_best_layer	0.0	5.0
Single Neuron	f1_max	0.6521	0.7102
Single Neuron	f1_mean	0.506	0.5415
Single Neuron	f1_std	0.0893	0.0442
Top-K Neurons	f1_best_layer	1.0	5.0
Top-K Neurons	f1_max	0.6171	0.7303
Top-K Neurons	f1_mean	0.5196	0.5902
Top-K Neurons	f1_std	0.0785	0.0709


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
0.50
0
1
2
3
4
5
Layer


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.70
0.65
0.60
0.55
0.50
0
1
2
3
4
5
Layer


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.750
0.725
0.700
0.675
0.650
0.625
0.600
0.575
0
1
2
3
4
5
Layer


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
0.75
0.70
0.65
0.60
0.55
0.50
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	5.0	5.0
Full Layer	accuracy_max	0.77	0.77
Full Layer	accuracy_mean	0.6563	0.6776
Full Layer	accuracy_std	0.057	0.0646
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.71	0.7033
Single Neuron	accuracy_mean	0.6083	0.5538
Single Neuron	accuracy_std	0.1015	0.0681
Top-K Neurons	accuracy_best_layer	0.0	0.0
Top-K Neurons	accuracy_max	0.6667	0.7233
Top-K Neurons	accuracy_mean	0.5598	0.6035
Top-K Neurons	accuracy_std	0.0894	0.0741