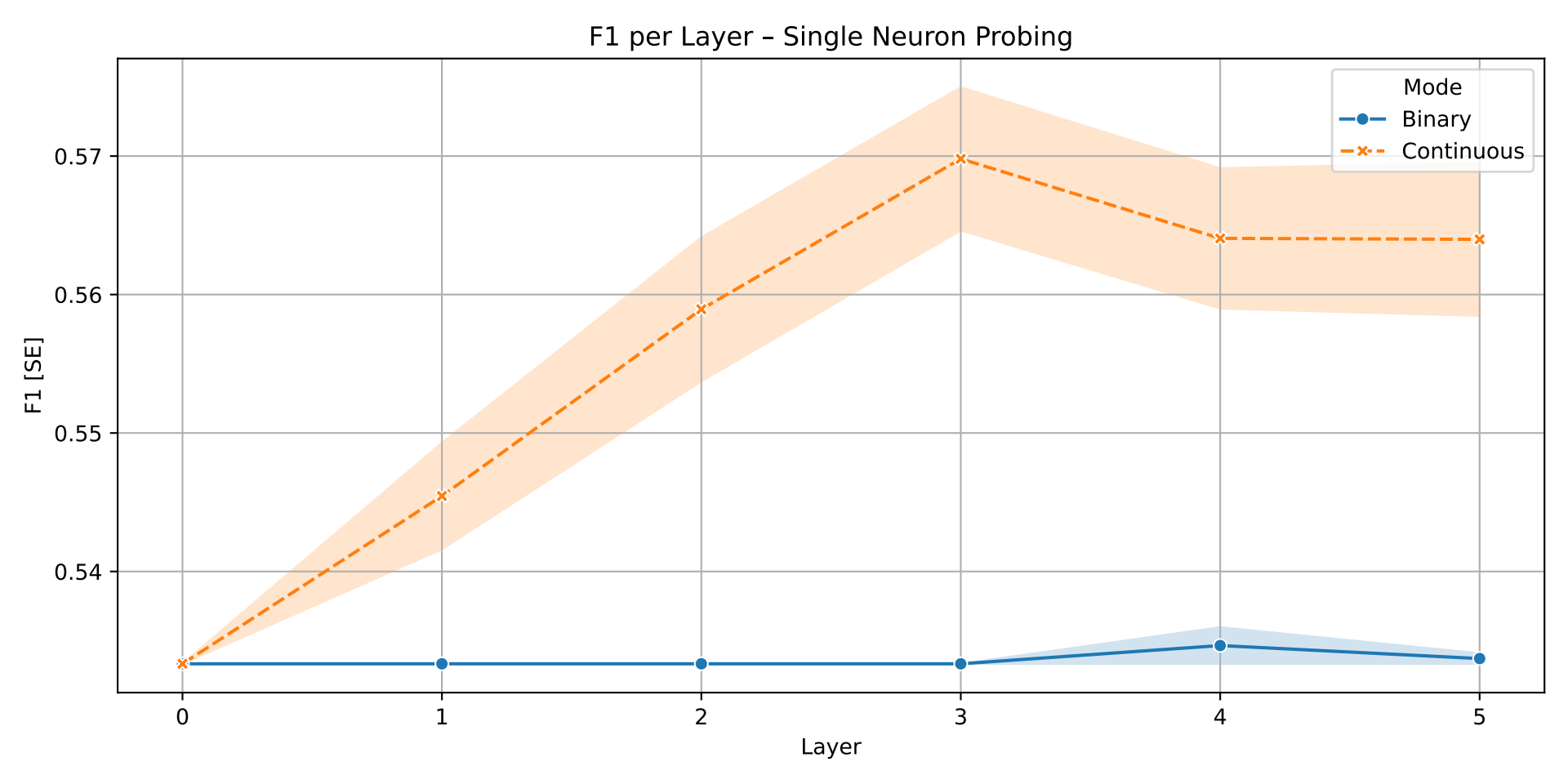
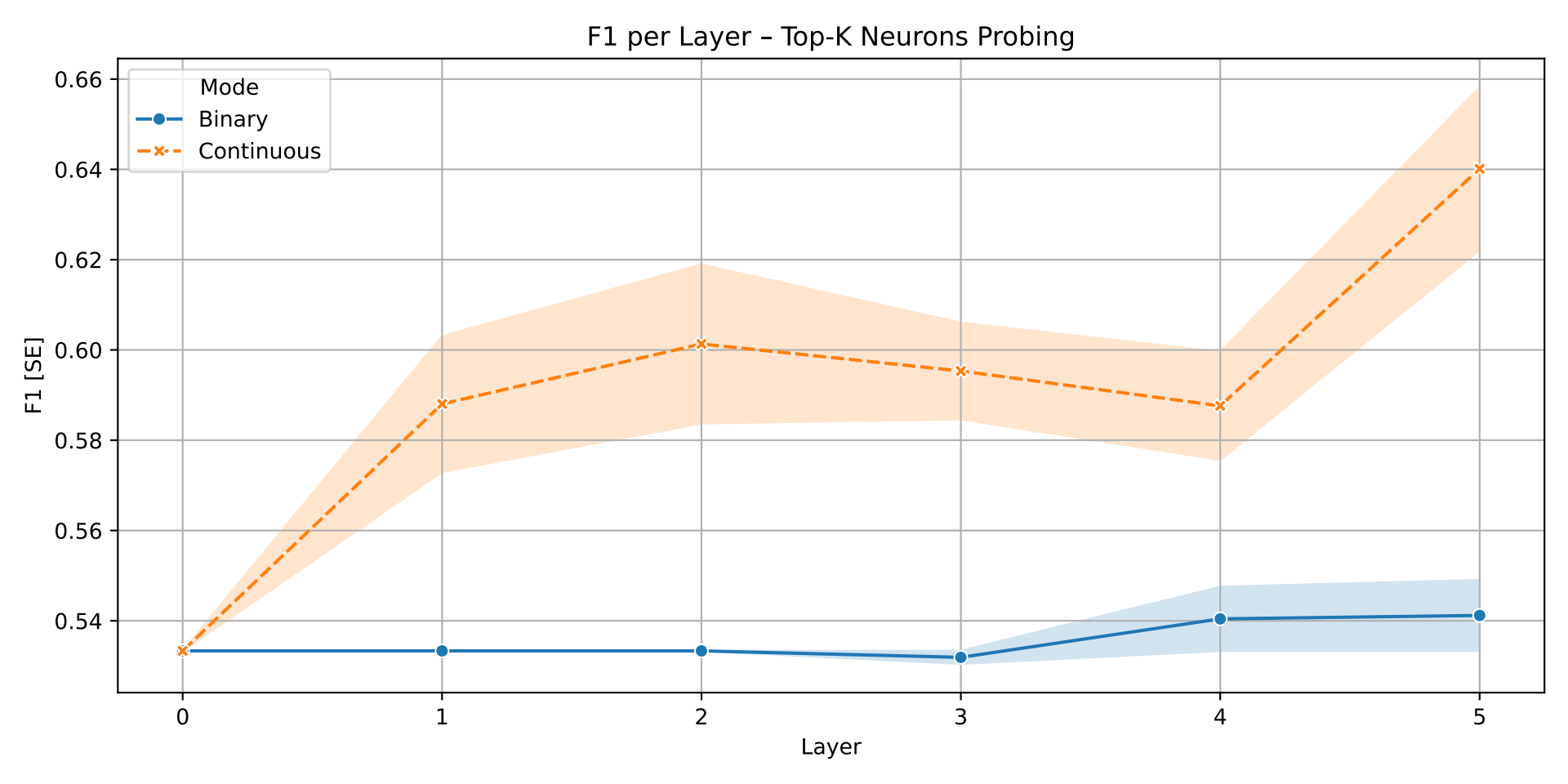
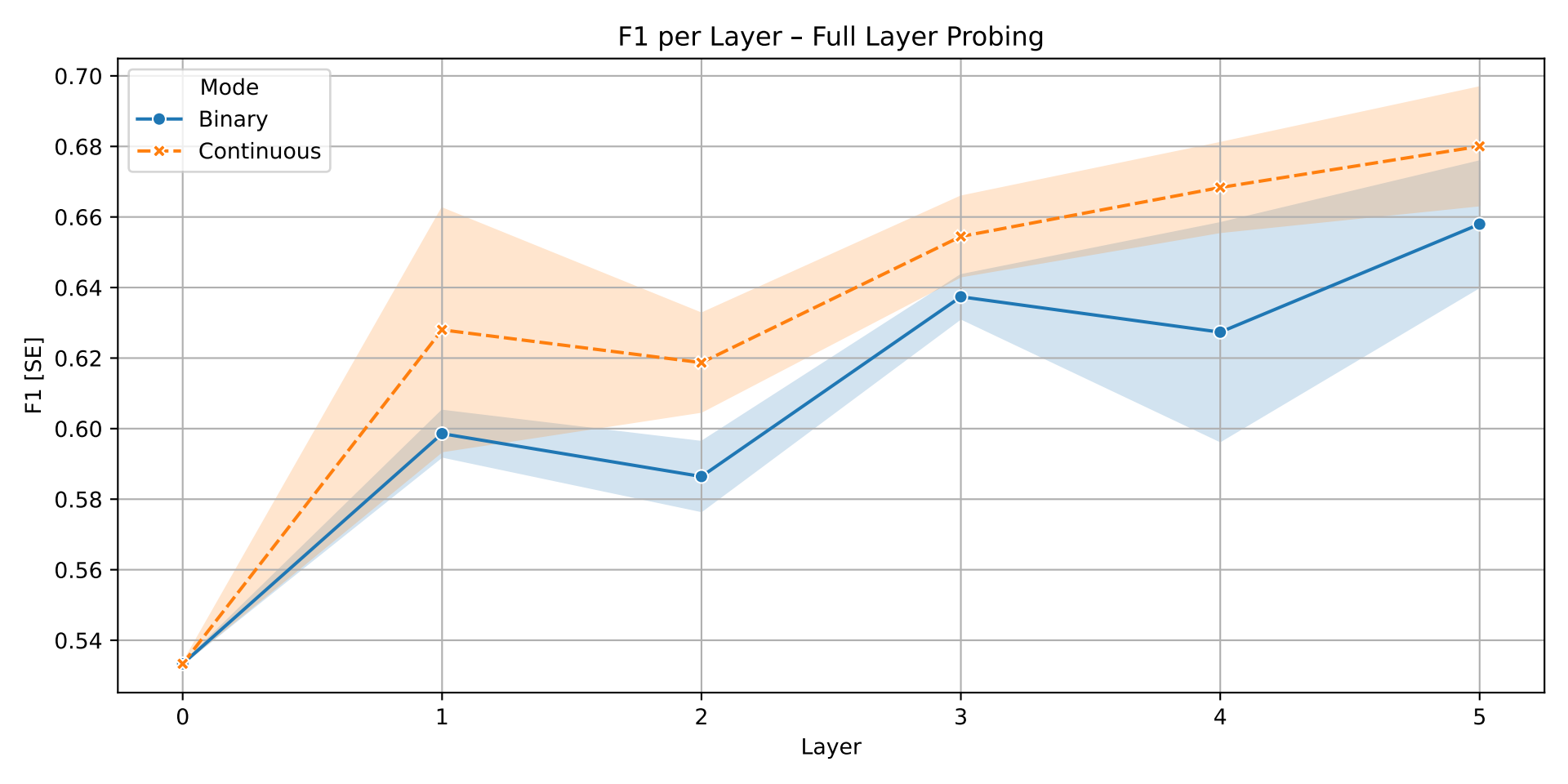
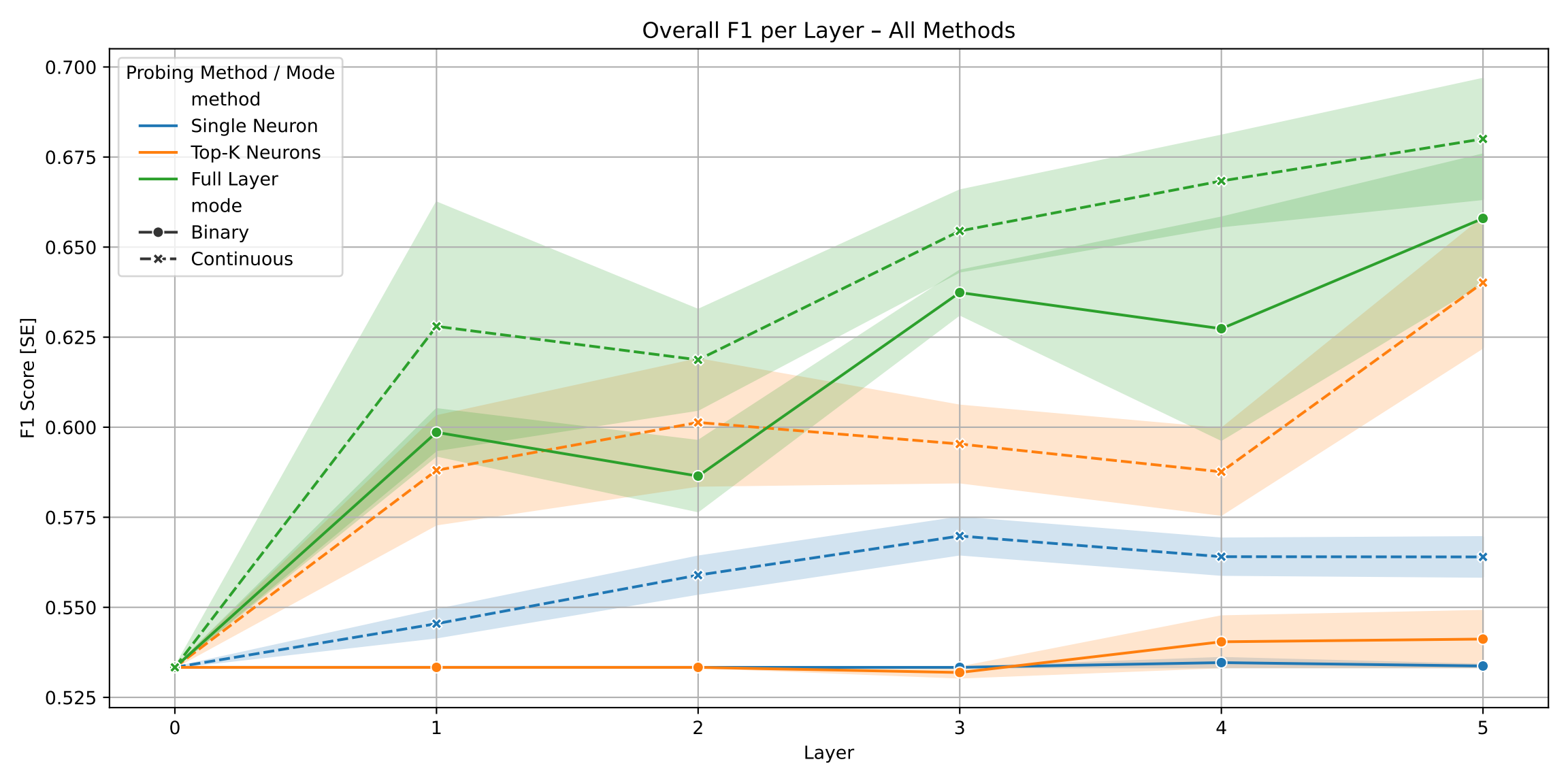

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.57
0.56
0.55
0.54
0
1
2
3
4
5
Layer

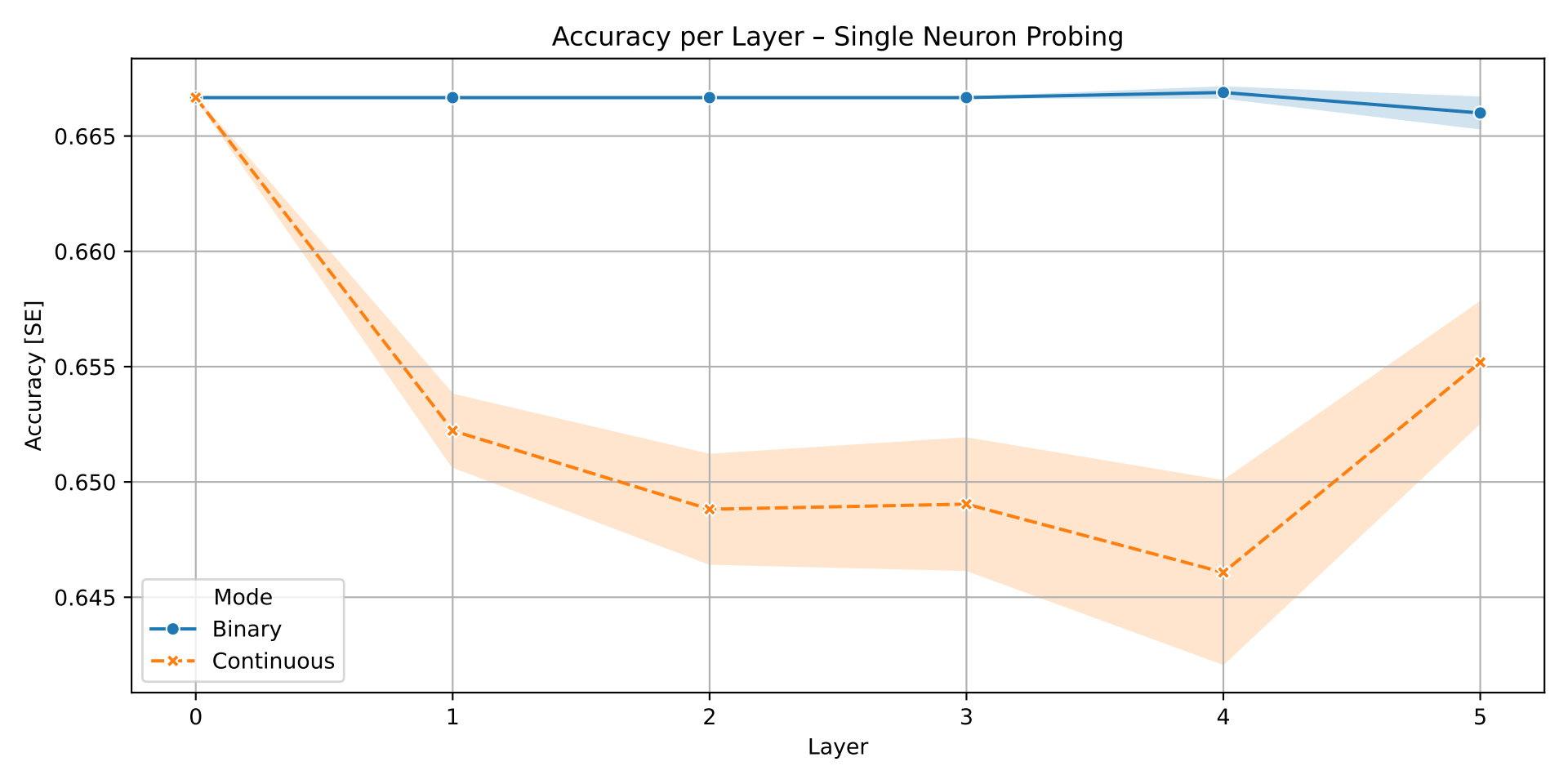

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0
1
2
3
4
5
Layer

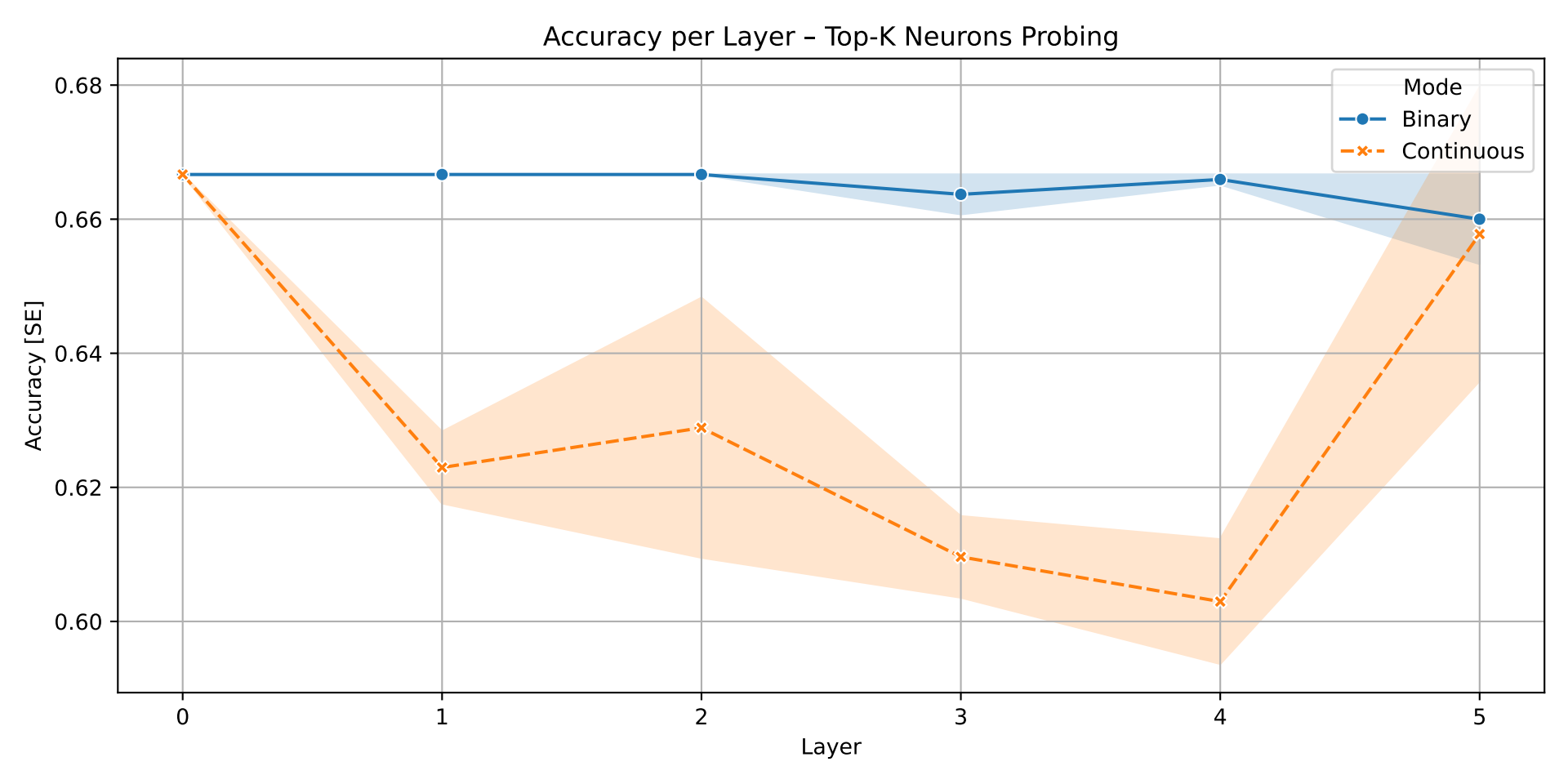

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.70
0.68
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0
1
2
3
4
5
Layer

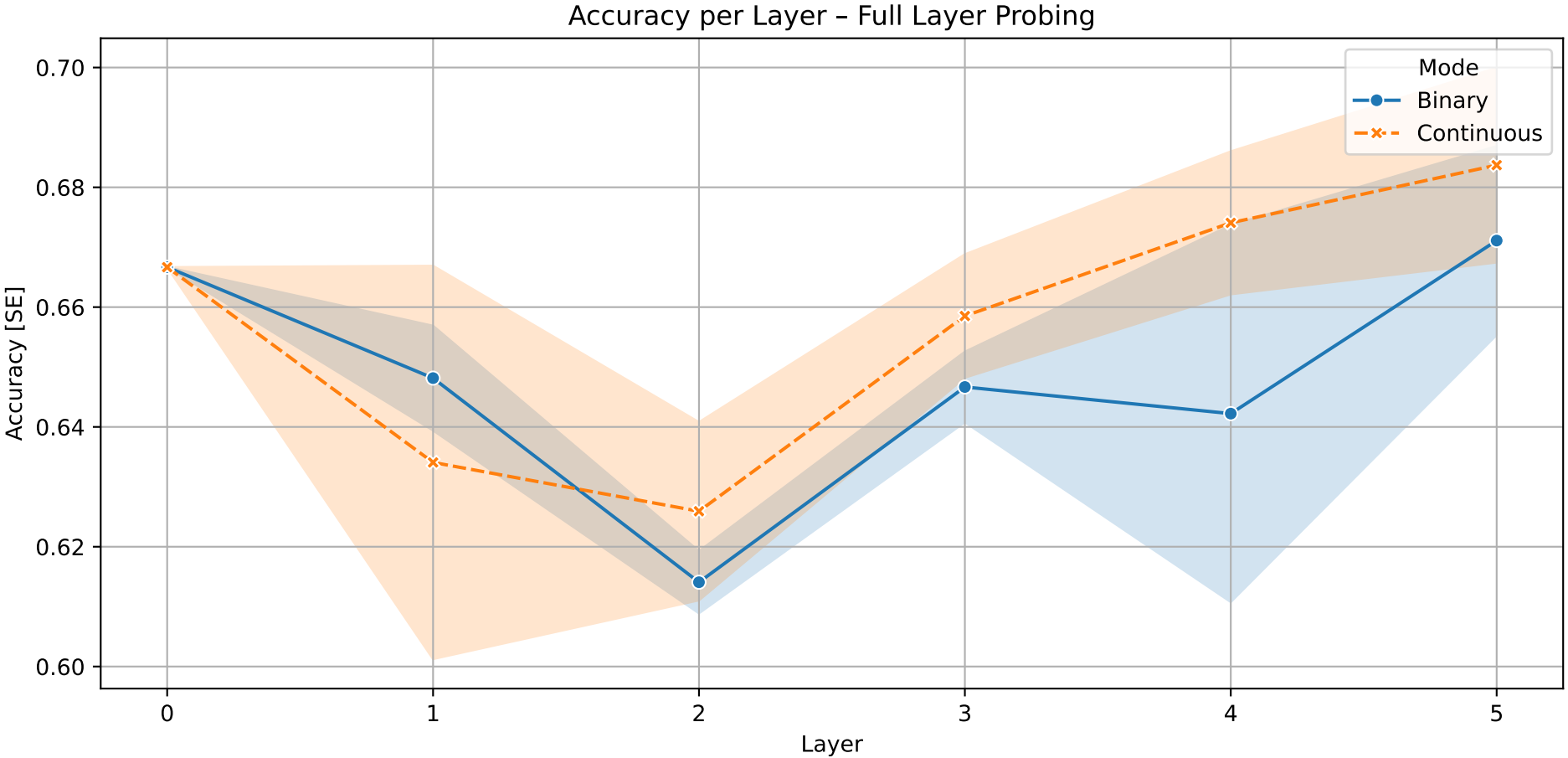

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0.525
0
1
2
3
4
5
Layer

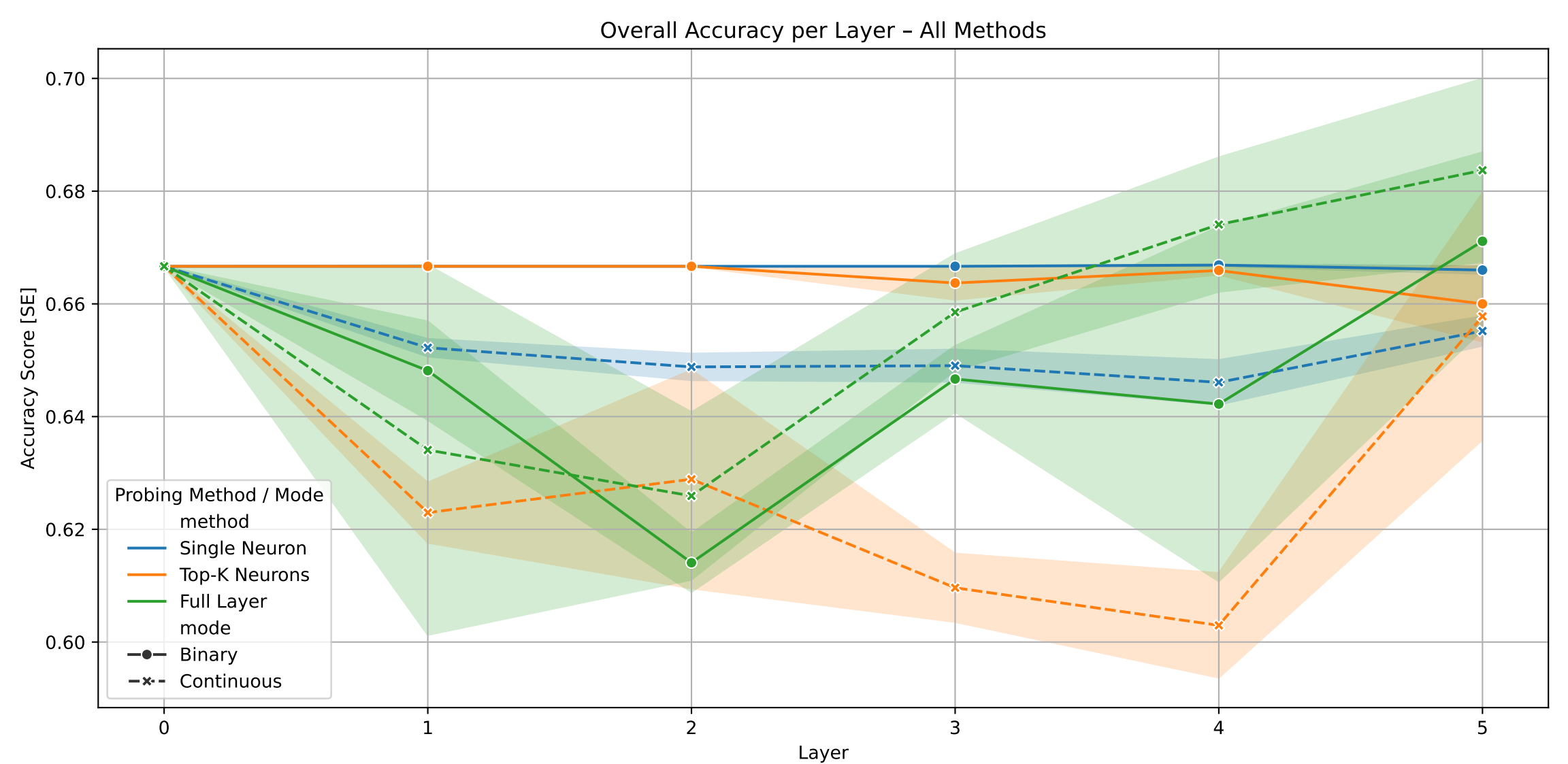
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	5.0	5.0
Full Layer	f1_max	0.6935	0.7071
Full Layer	f1_mean	0.6068	0.6305
Full Layer	f1_std	0.0474	0.0563
Single Neuron	f1_best_layer	4.0	3.0
Single Neuron	f1_max	0.5729	0.6662
Single Neuron	f1_mean	0.5336	0.5559
Single Neuron	f1_std	0.0031	0.0277
Top-K Neurons	f1_best_layer	5.0	5.0
Top-K Neurons	f1_max	0.5569	0.6718
Top-K Neurons	f1_mean	0.5356	0.591
Top-K Neurons	f1_std	0.0074	0.0379


Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
0.665
0.660
0.655
0.650
0.645
Mode
Binary
Continuous
0
1
2
3
4
5
Layer


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.68
0.66
0.64
0.62
0.60
0
1
2
3
4
5
Layer


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.70
0.68
0.66
0.64
0.62
0.60
0
1
2
3
4
5
Layer


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
0.70
0.68
0.66
0.64
0.62
0.60
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	0.7022	0.7089
Full Layer	accuracy_mean	0.6481	0.6572
Full Layer	accuracy_std	0.0292	0.0331
Single Neuron	accuracy_best_layer	4.0	0.0
Single Neuron	accuracy_max	0.6733	0.6978
Single Neuron	accuracy_mean	0.6666	0.653
Single Neuron	accuracy_std	0.0016	0.0153
Top-K Neurons	accuracy_best_layer	0.0	0.0
Top-K Neurons	accuracy_max	0.6667	0.6889
Top-K Neurons	accuracy_mean	0.6649	0.6315
Top-K Neurons	accuracy_std	0.005	0.0306