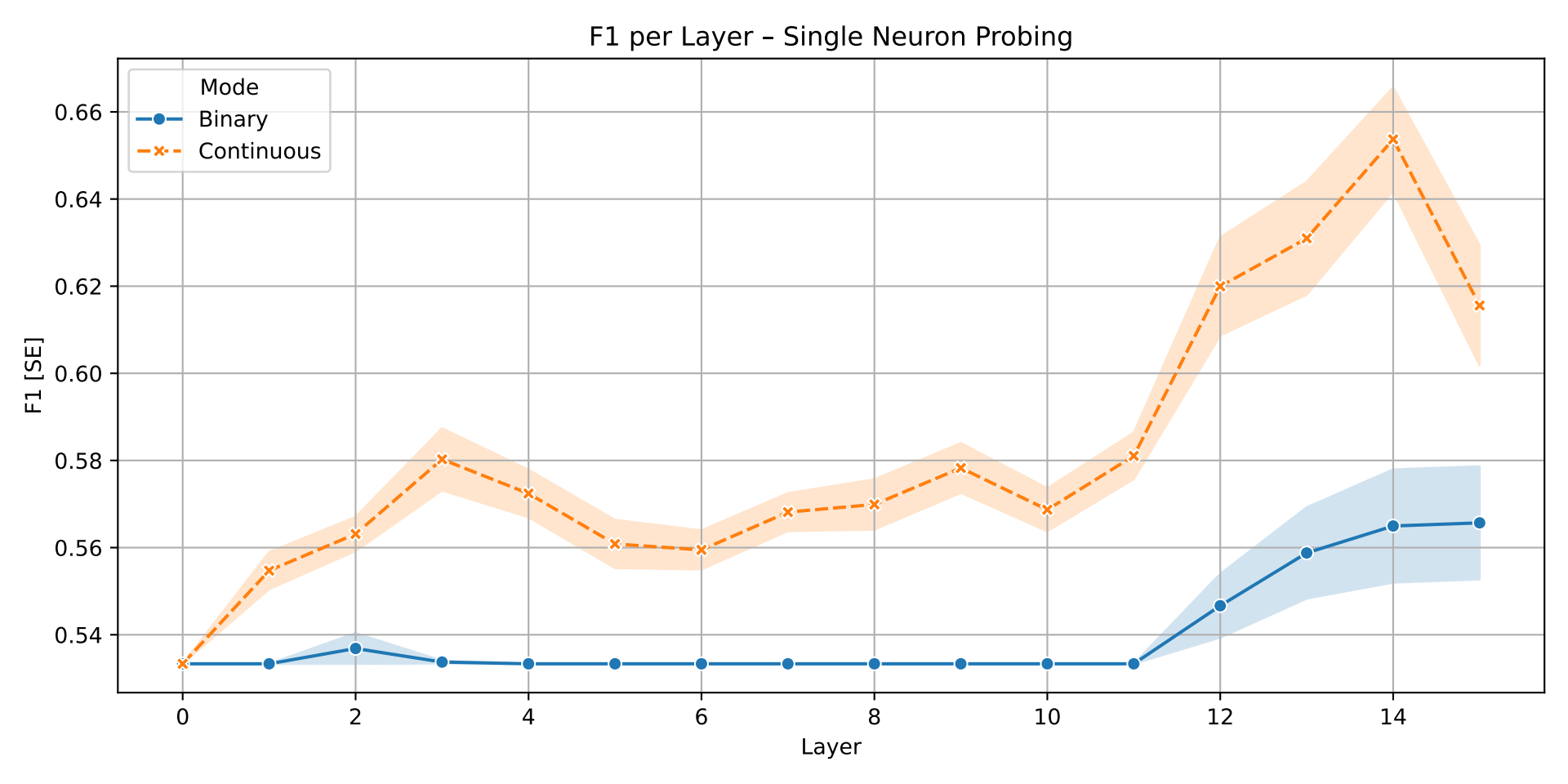
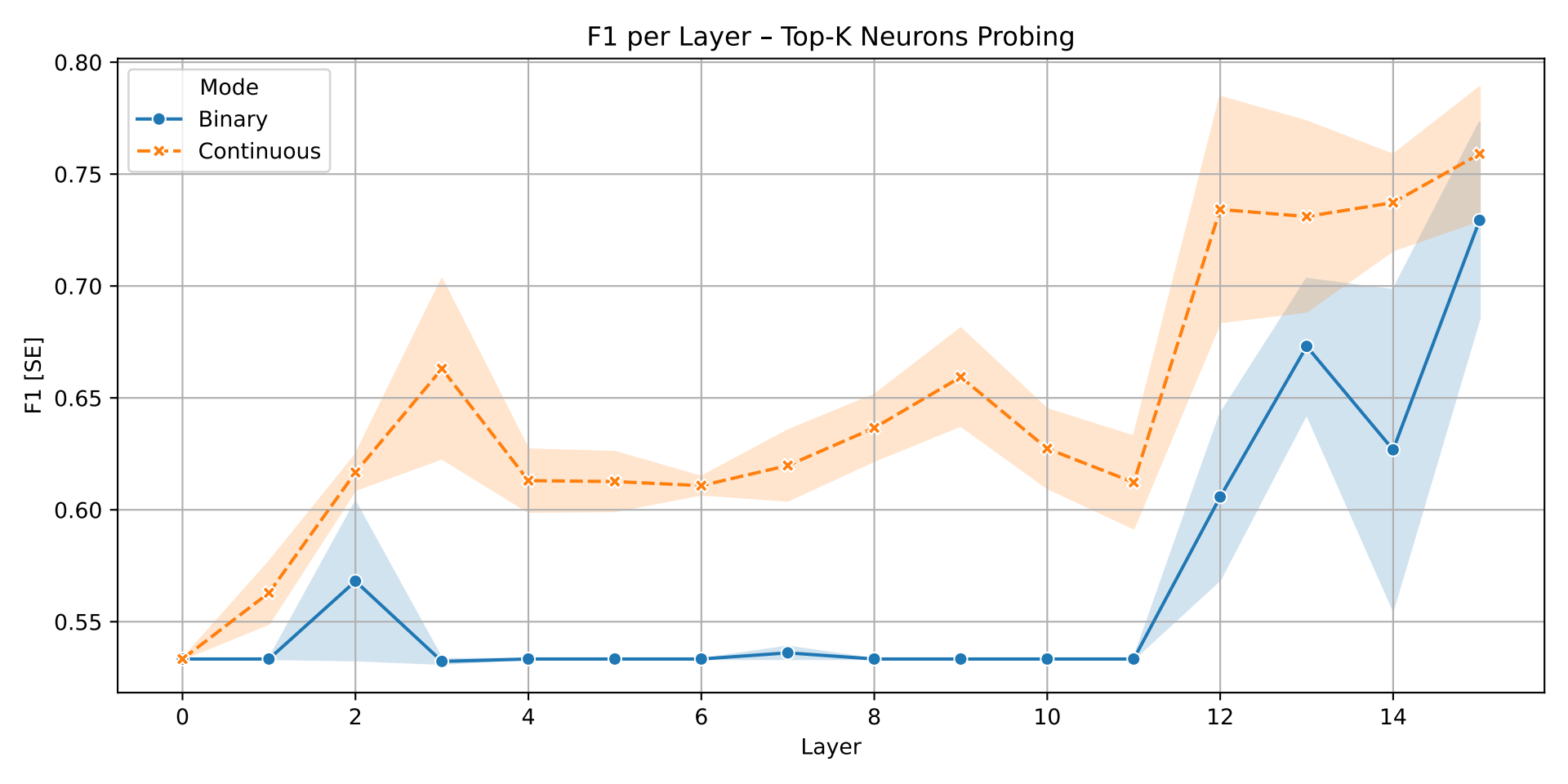
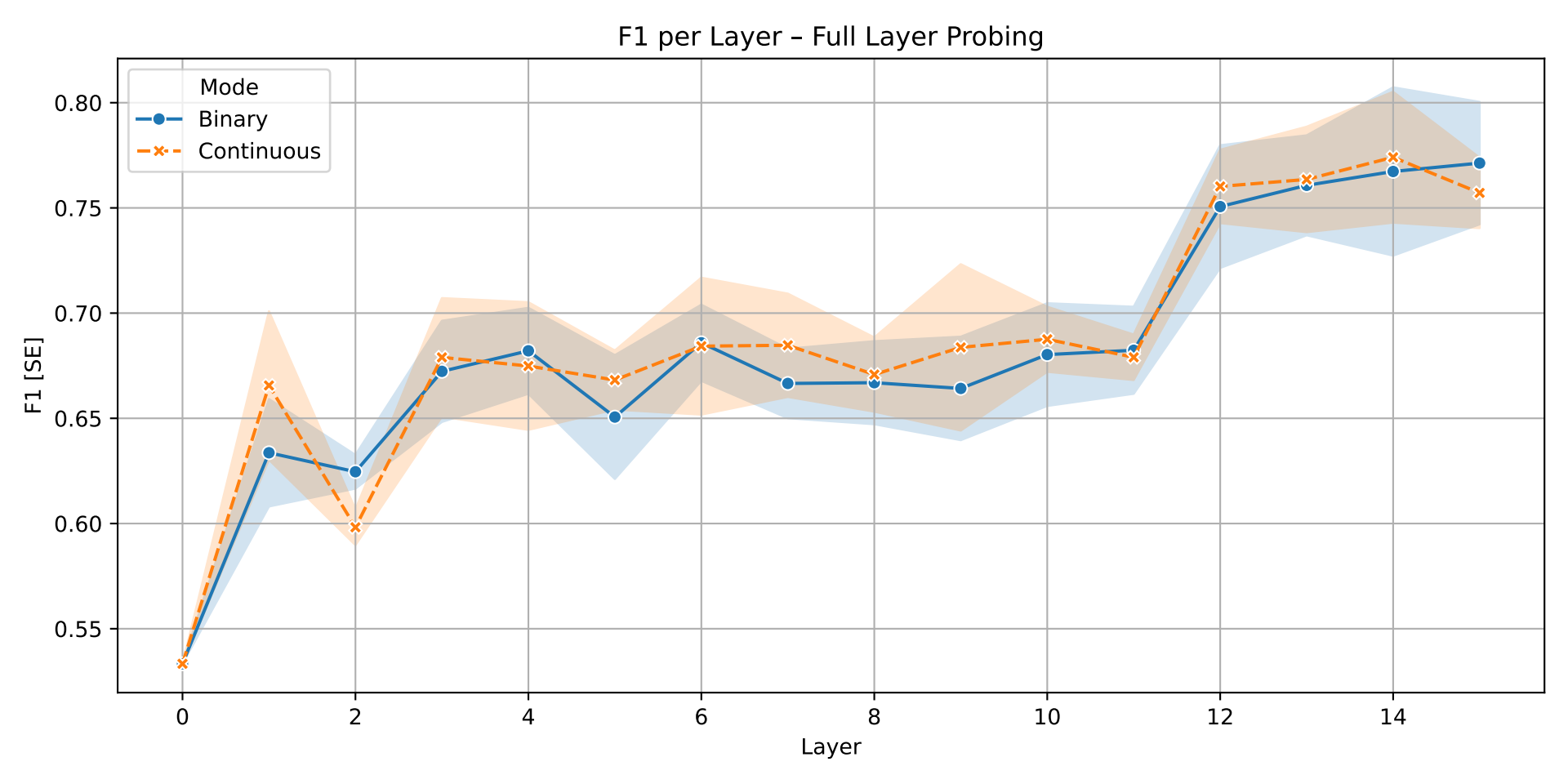
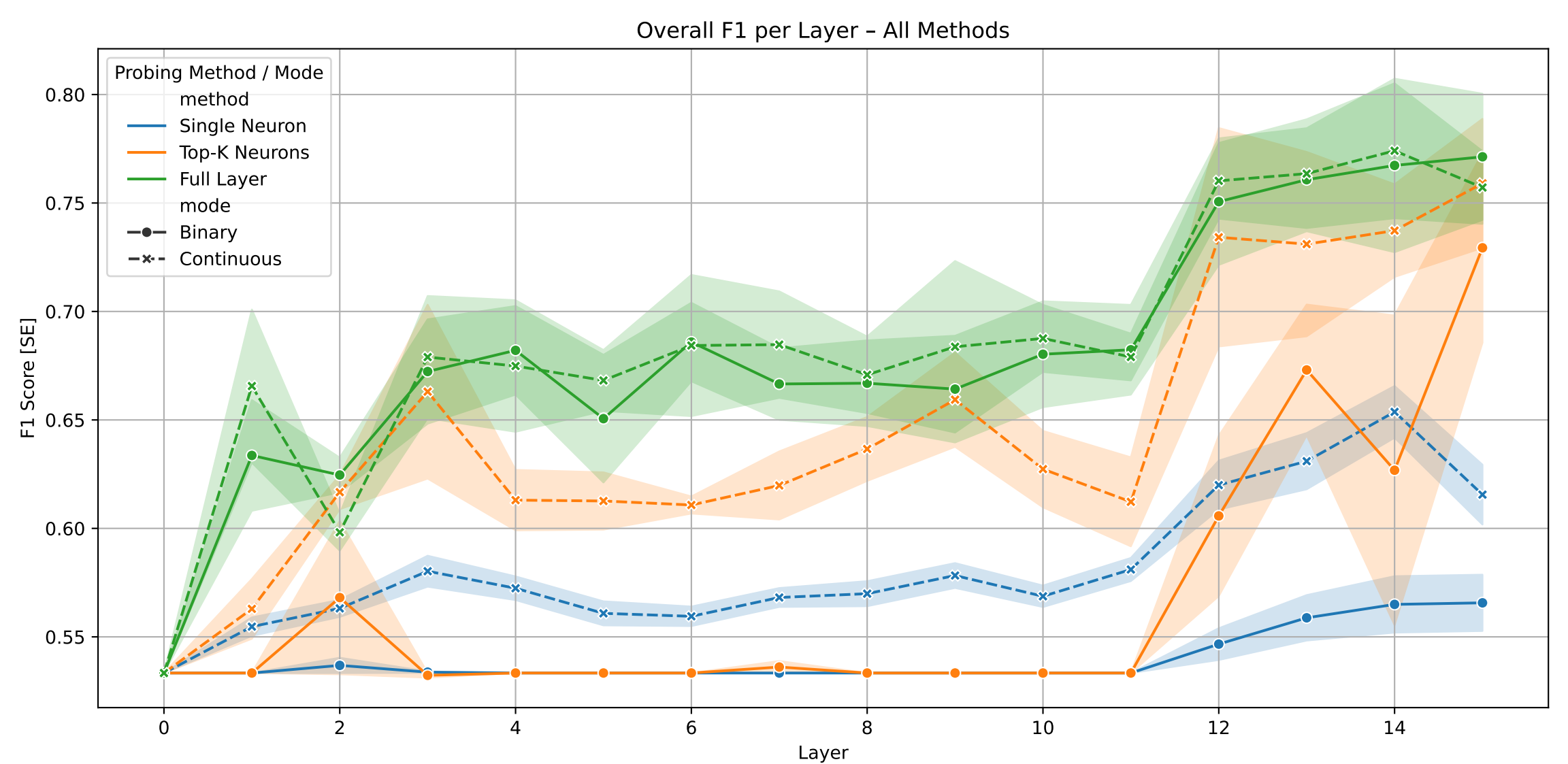

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0
2
4
6
8
10
12
14
Layer

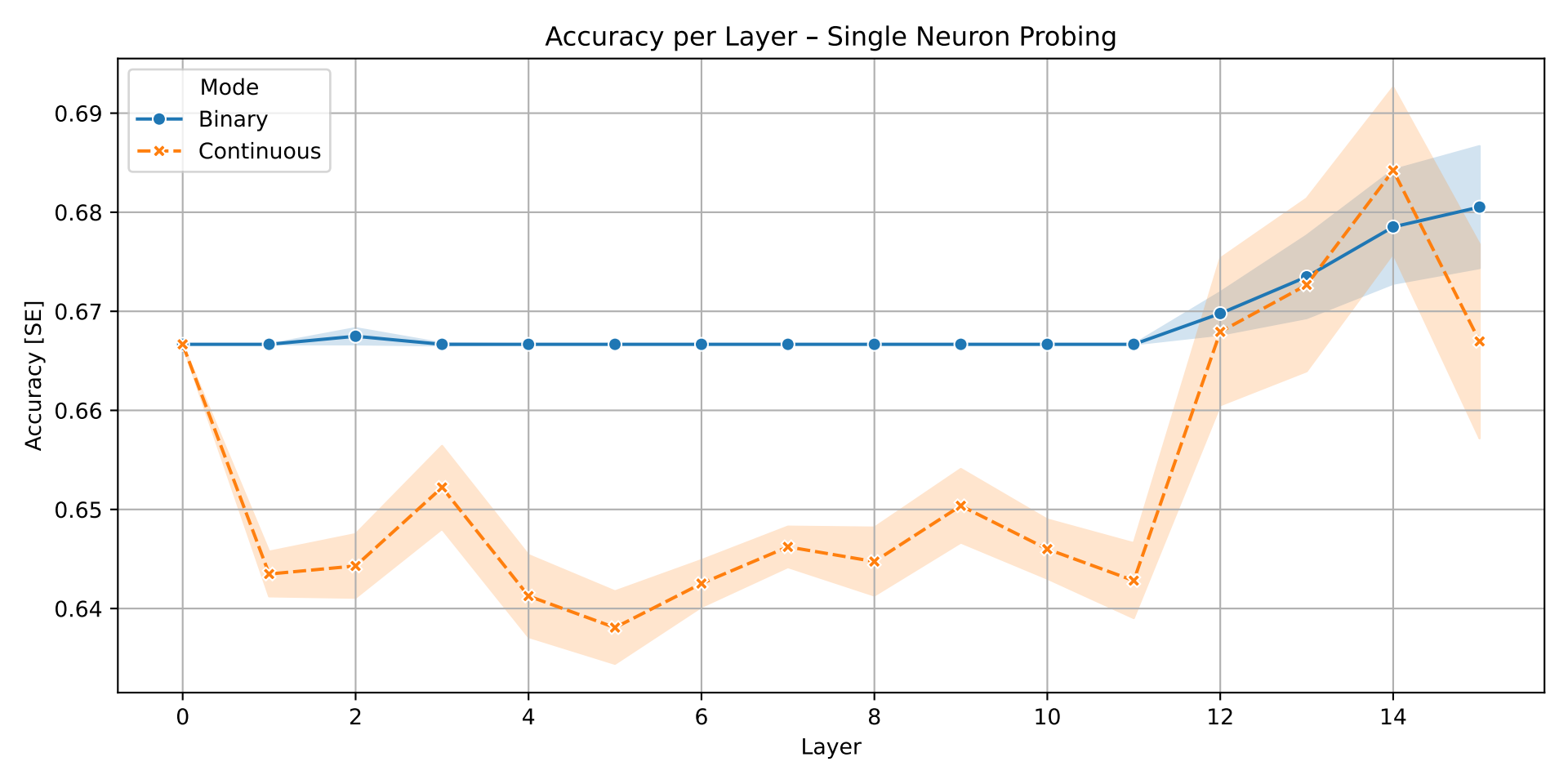

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.80
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
12
14
Layer

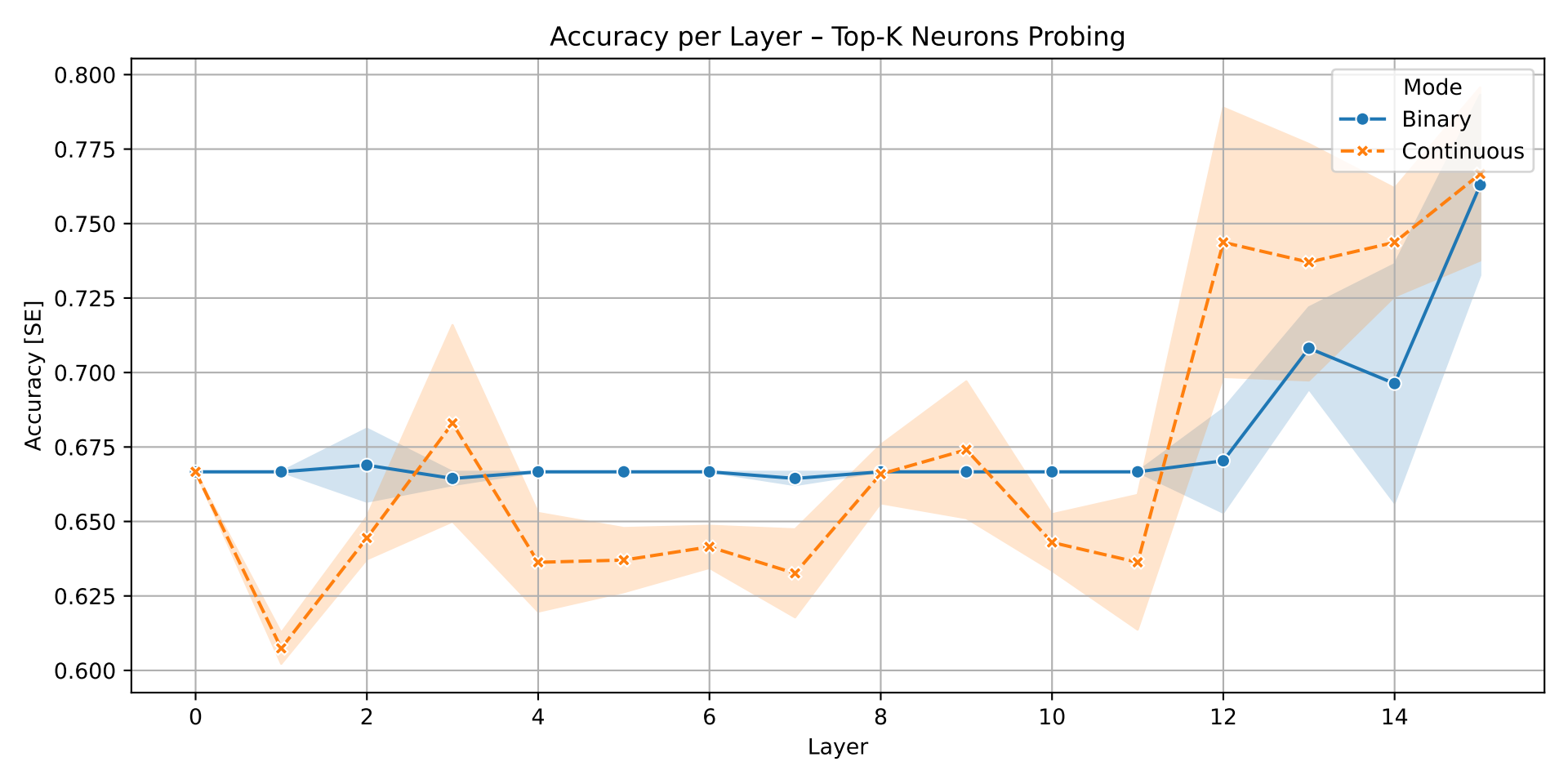

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.80
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
12
14
Layer

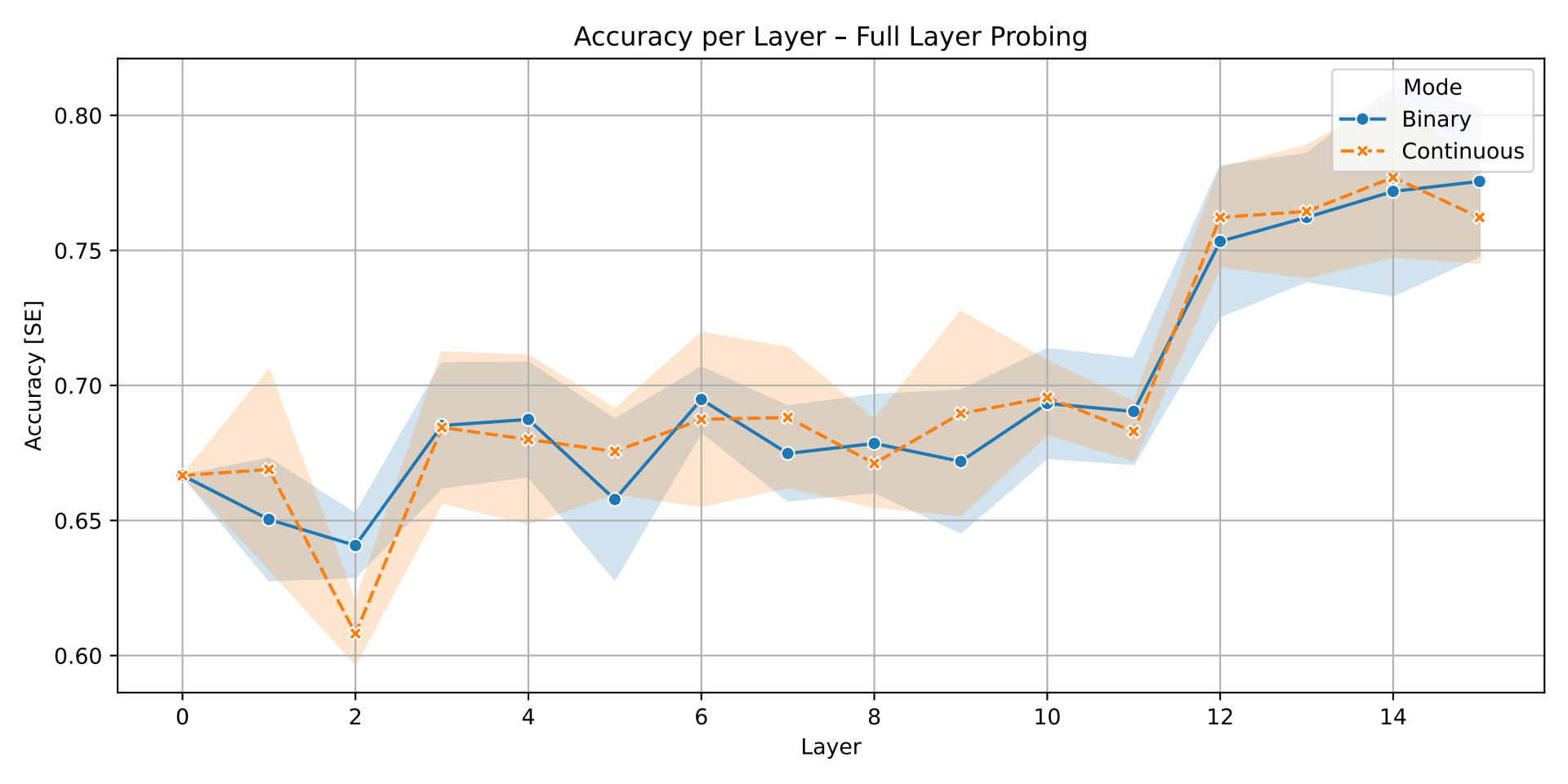

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.80
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
12
14
Layer

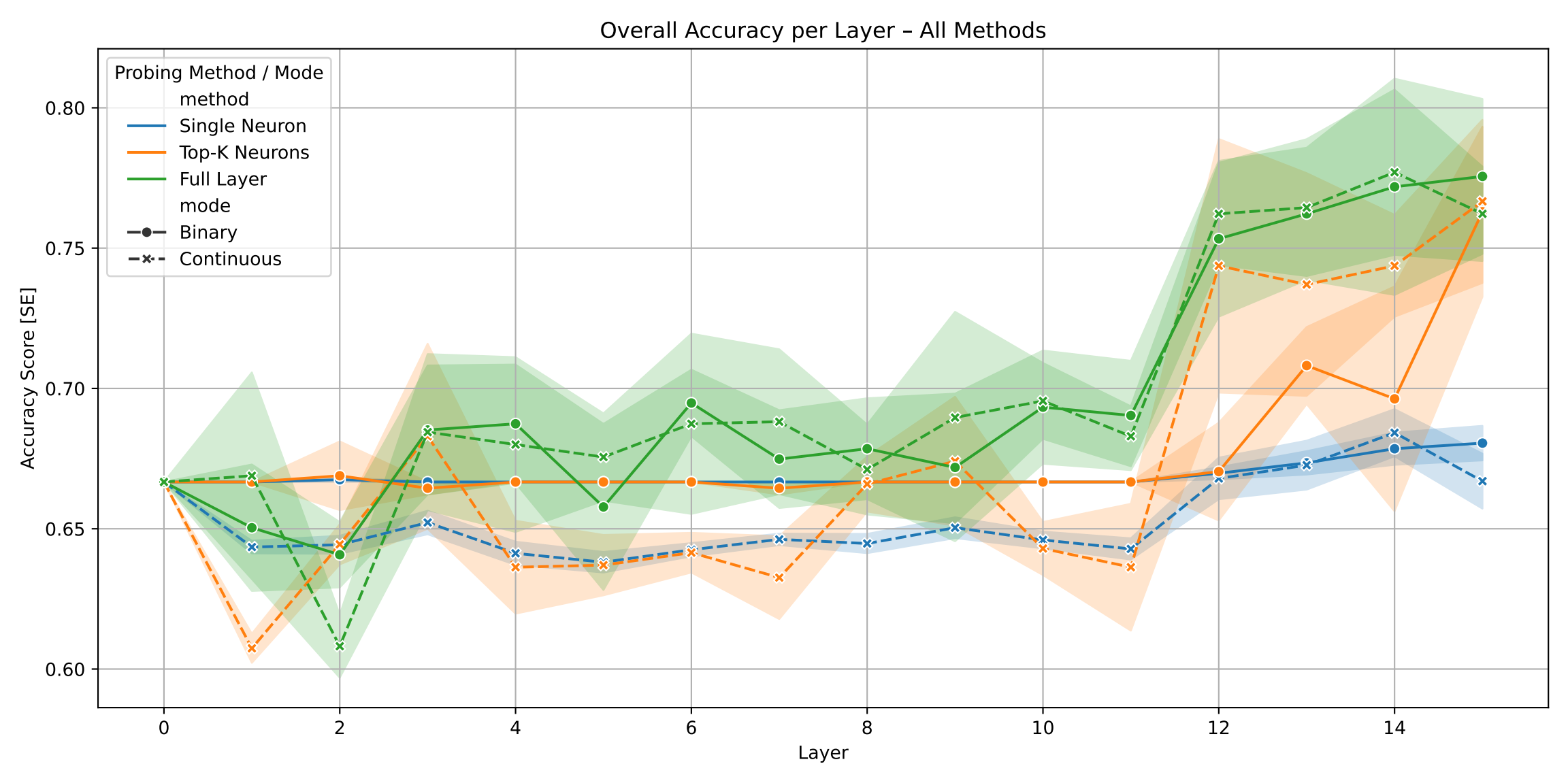
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	15.0	14.0
Full Layer	f1_max	0.8472	0.8356
Full Layer	f1_mean	0.6808	0.6853
Full Layer	f1_std	0.0687	0.0692
Single Neuron	f1_best_layer	15.0	14.0
Single Neuron	f1_max	0.7776	0.8178
Single Neuron	f1_mean	0.54	0.5819
Single Neuron	f1_std	0.0327	0.0513
Top-K Neurons	f1_best_layer	15.0	15.0
Top-K Neurons	f1_max	0.8151	0.8332
Top-K Neurons	f1_mean	0.567	0.6456
Top-K Neurons	f1_std	0.0697	0.0723


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.69
0.68
0.67
0.66
0.65
0.64
0
2
4
6
8
10
12
14
Layer


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
Layer
Mode
Binary
Continuous
0.800
0.775
0.750
0.725
0.700
0.675
0.650
0.625
0.600
0
2
4
6
8
10
12
14


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.80
0.75
0.70
0.65
0.60
0
2
4
6
8
10
12
14
Layer


Overall Accuracy per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
Accuracy Score [SE]
0.80
0.75
0.70
0.65
0.60
0
2
4
6
8
10
12
14
Layer

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	15.0	14.0
Full Layer	accuracy_max	0.8489	0.8356
Full Layer	accuracy_mean	0.6972	0.6978
Full Layer	accuracy_std	0.0538	0.0562
Single Neuron	accuracy_best_layer	15.0	14.0
Single Neuron	accuracy_max	0.7911	0.8178
Single Neuron	accuracy_mean	0.6689	0.6532
Single Neuron	accuracy_std	0.0137	0.0307
Top-K Neurons	accuracy_best_layer	15.0	15.0
Top-K Neurons	accuracy_max	0.8222	0.8333
Top-K Neurons	accuracy_mean	0.6772	0.6725
Top-K Neurons	accuracy_std	0.0324	0.0573