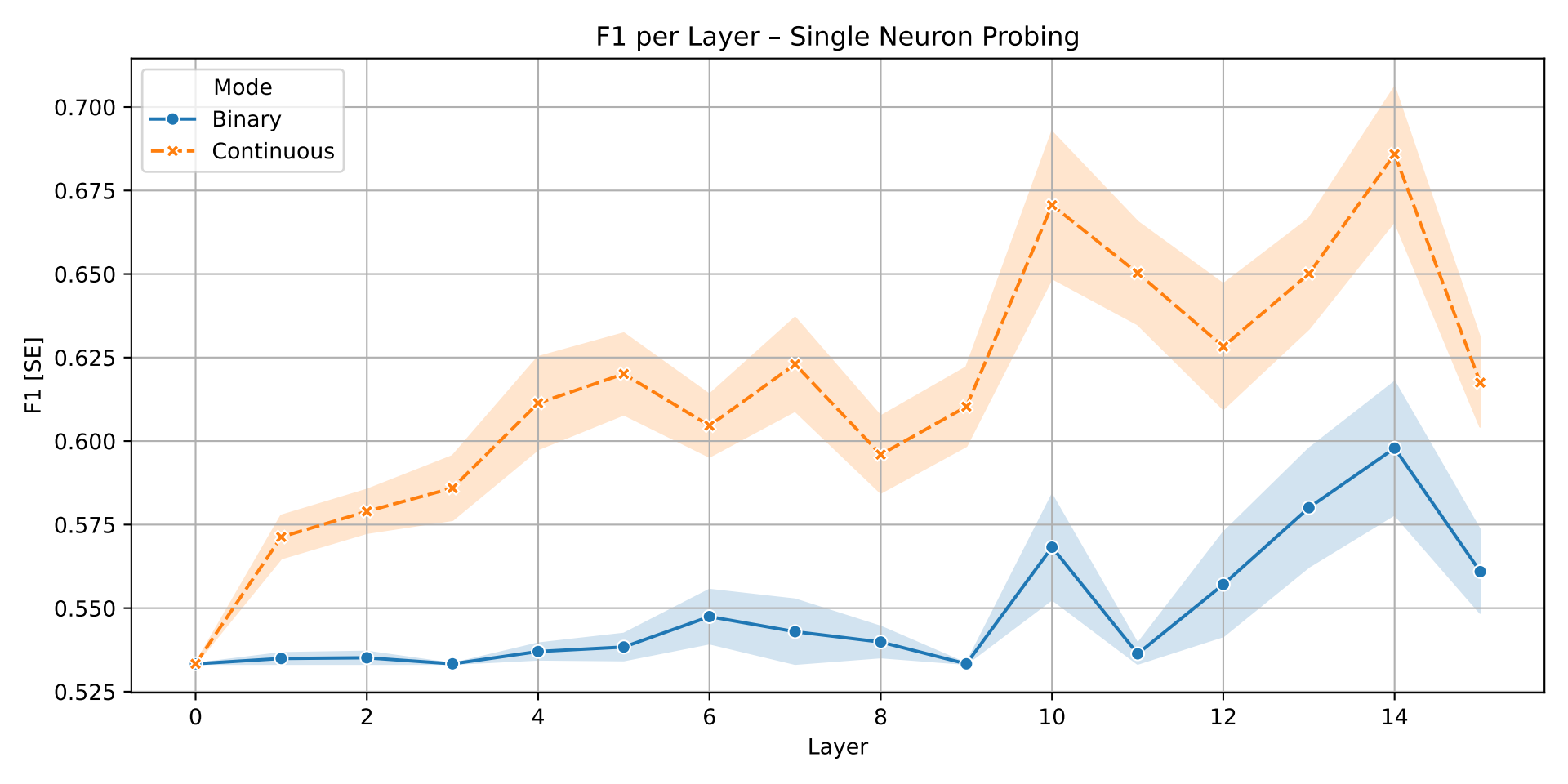
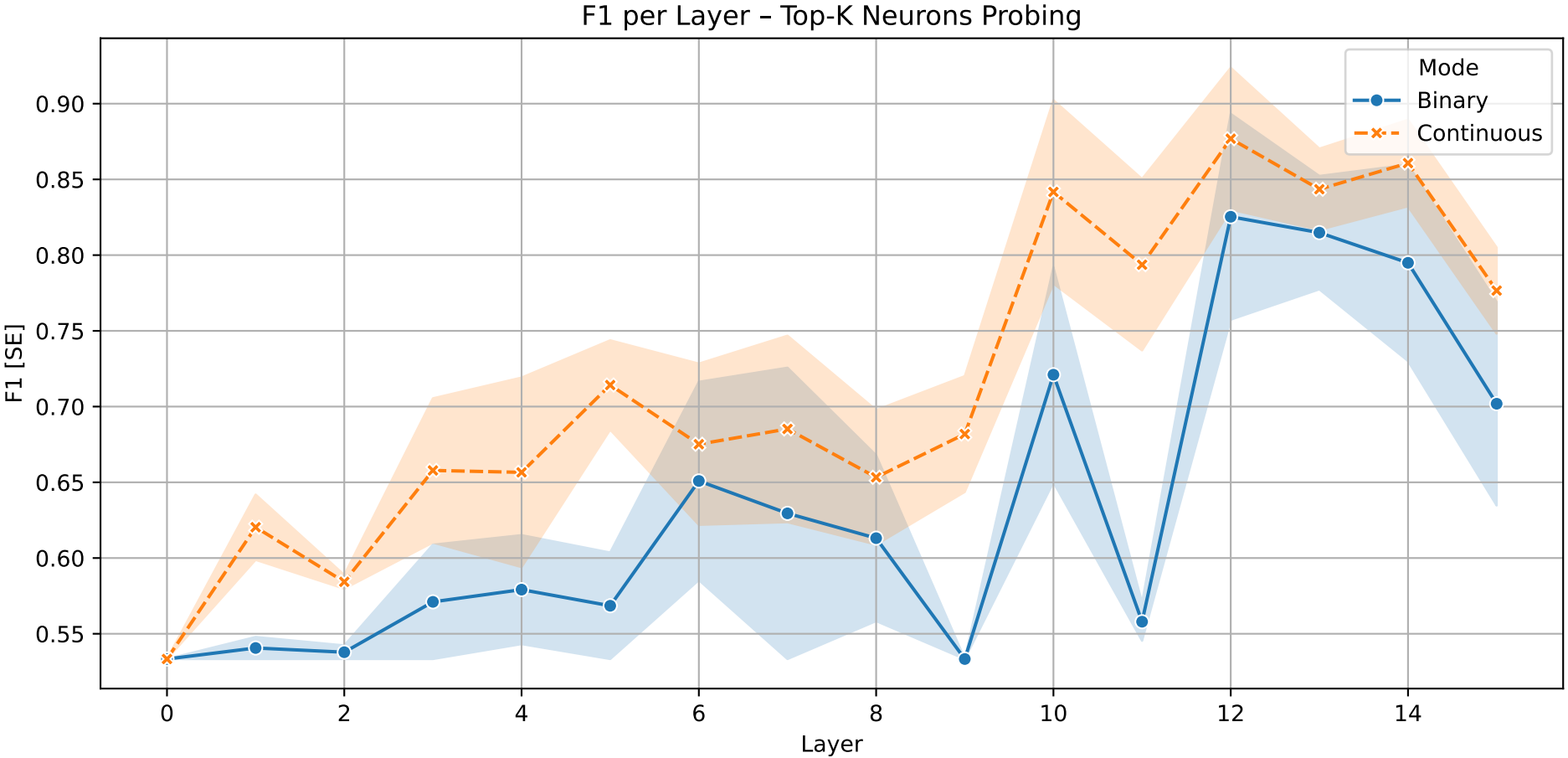
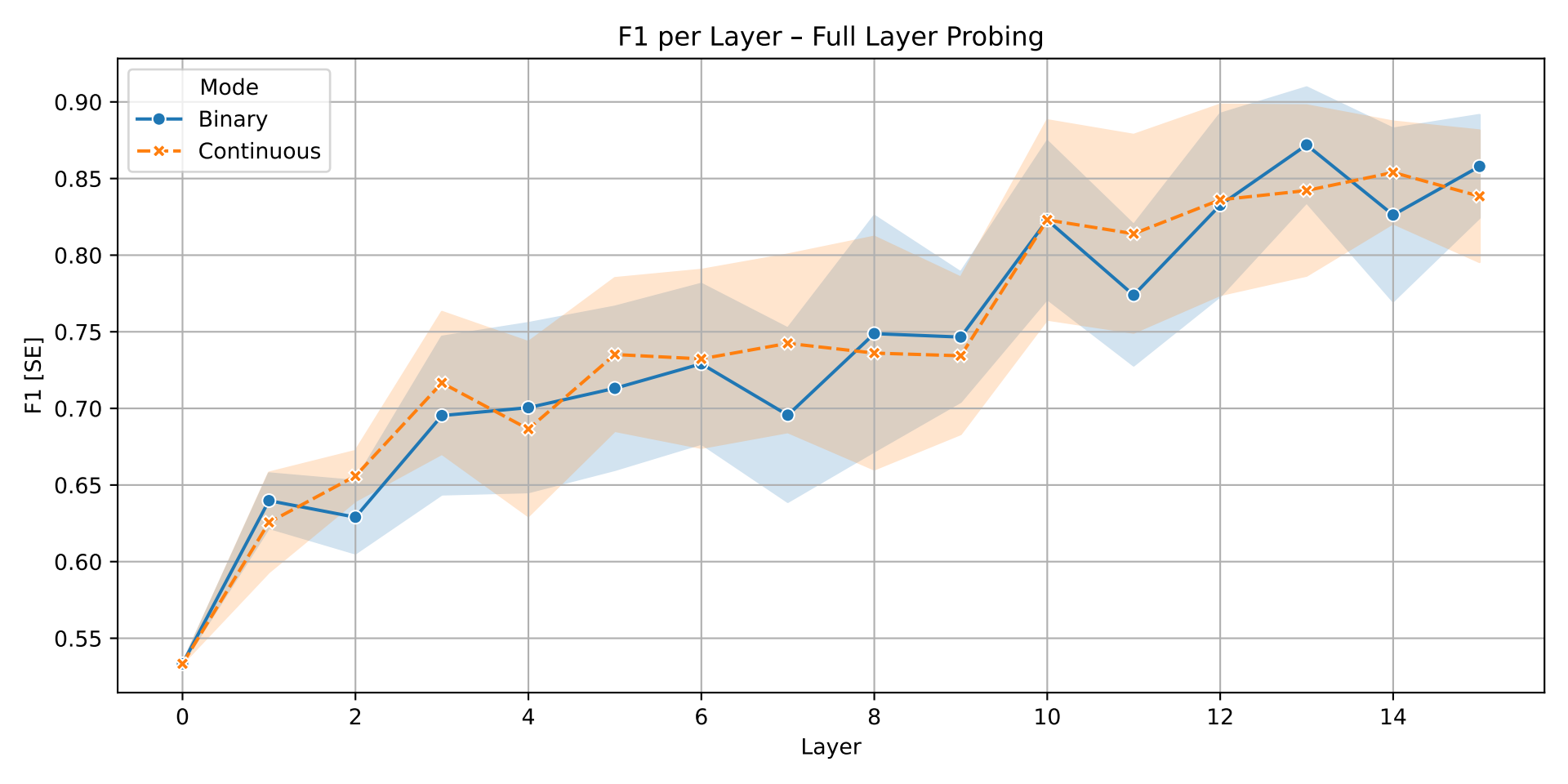
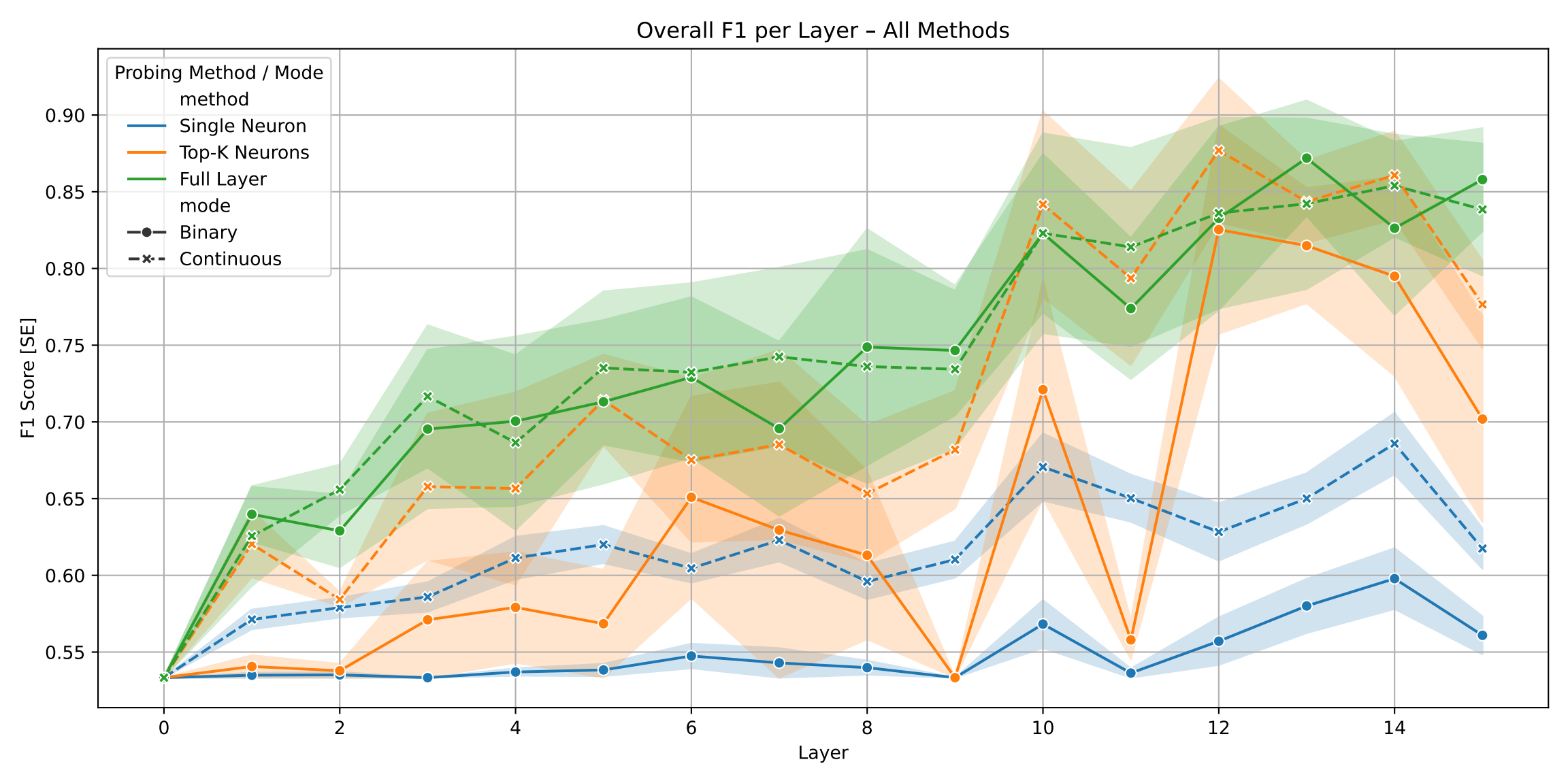

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0.525
0
2
4
6
8
10
12
14
Layer

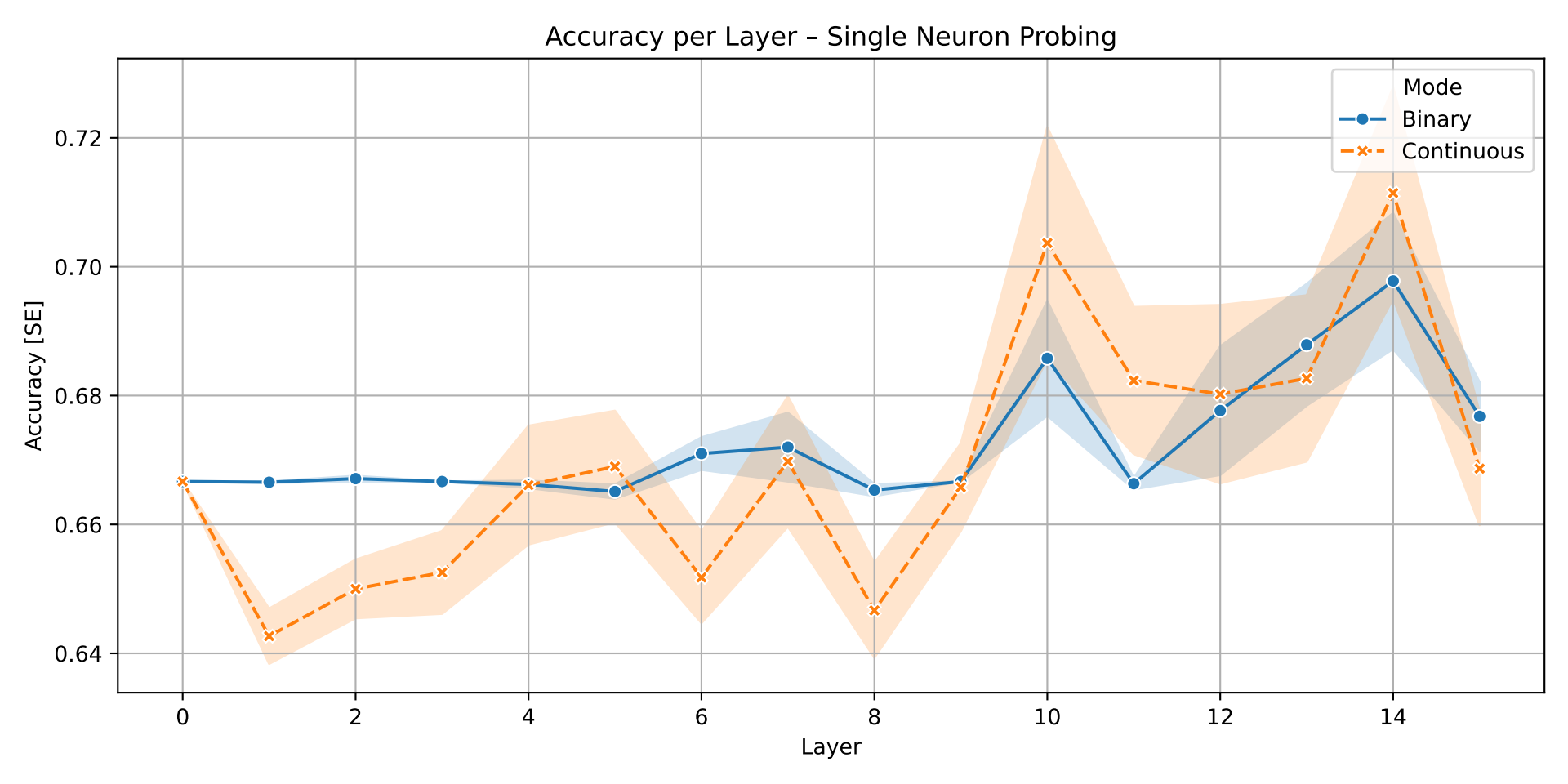

F1 per Layer – Top-K Neurons Probing
F1 [SE]
Layer
Mode
Binary
Continuous
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
12
14

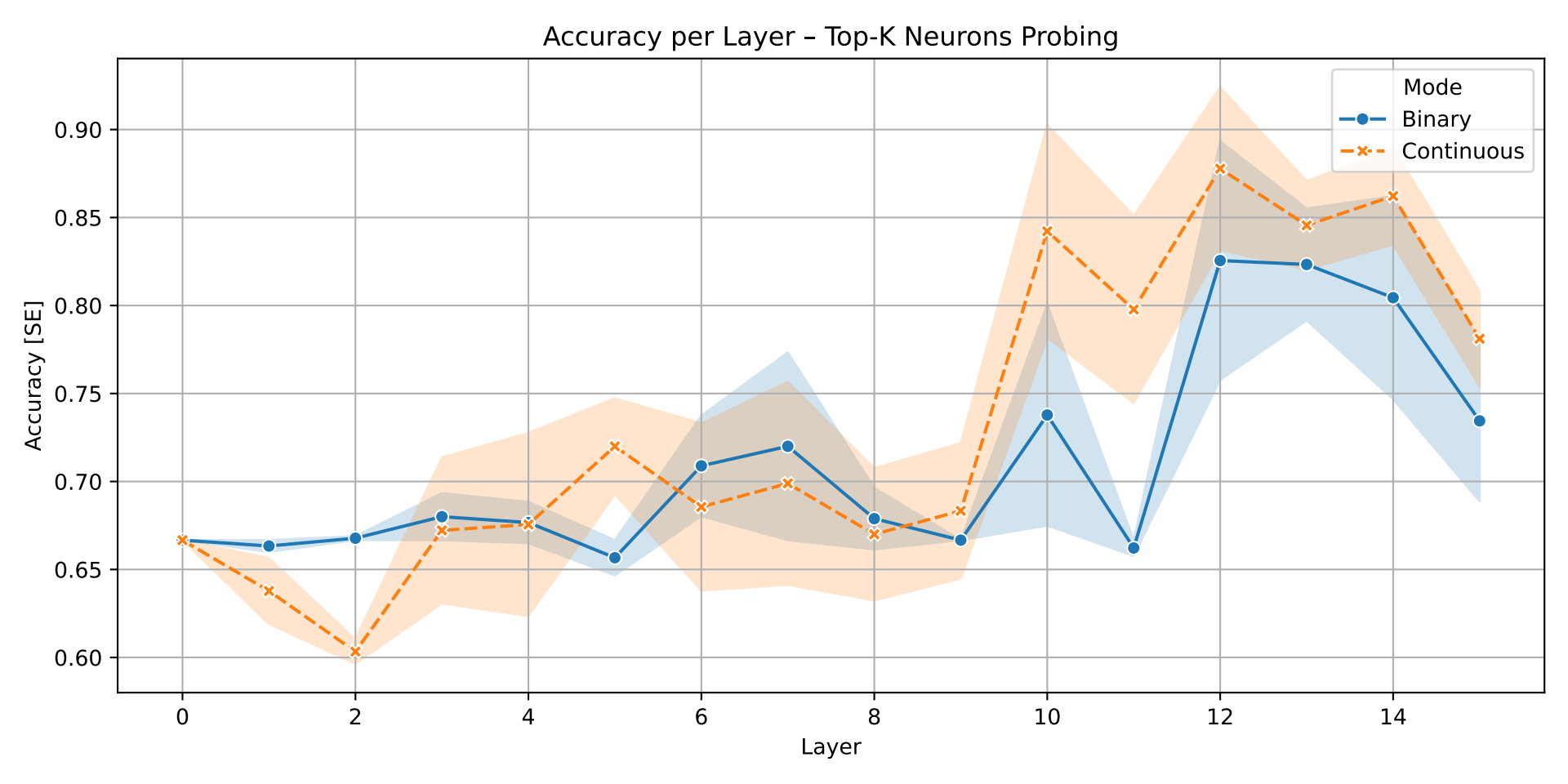

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
12
14
Layer

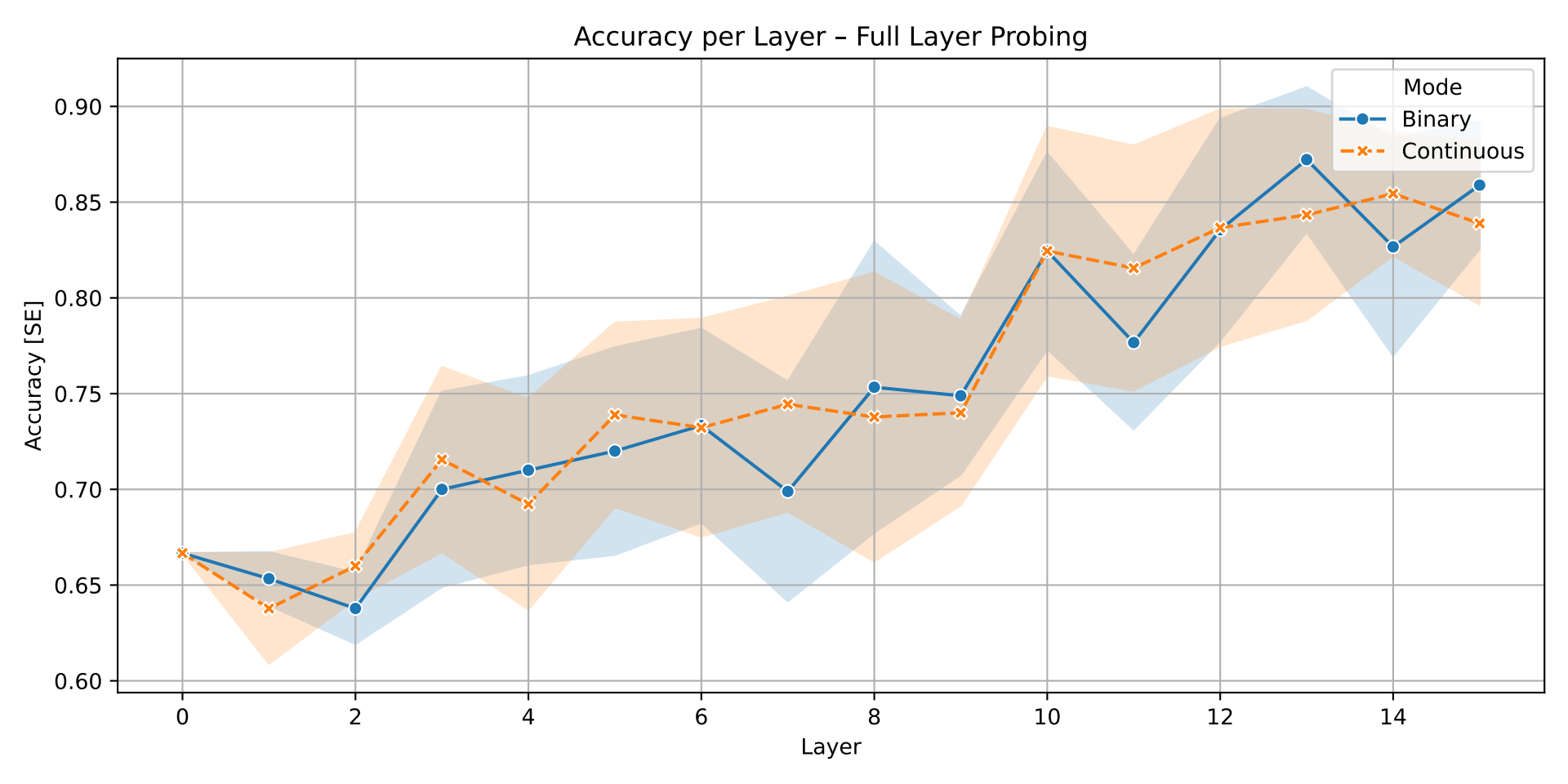

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
12
14
Layer

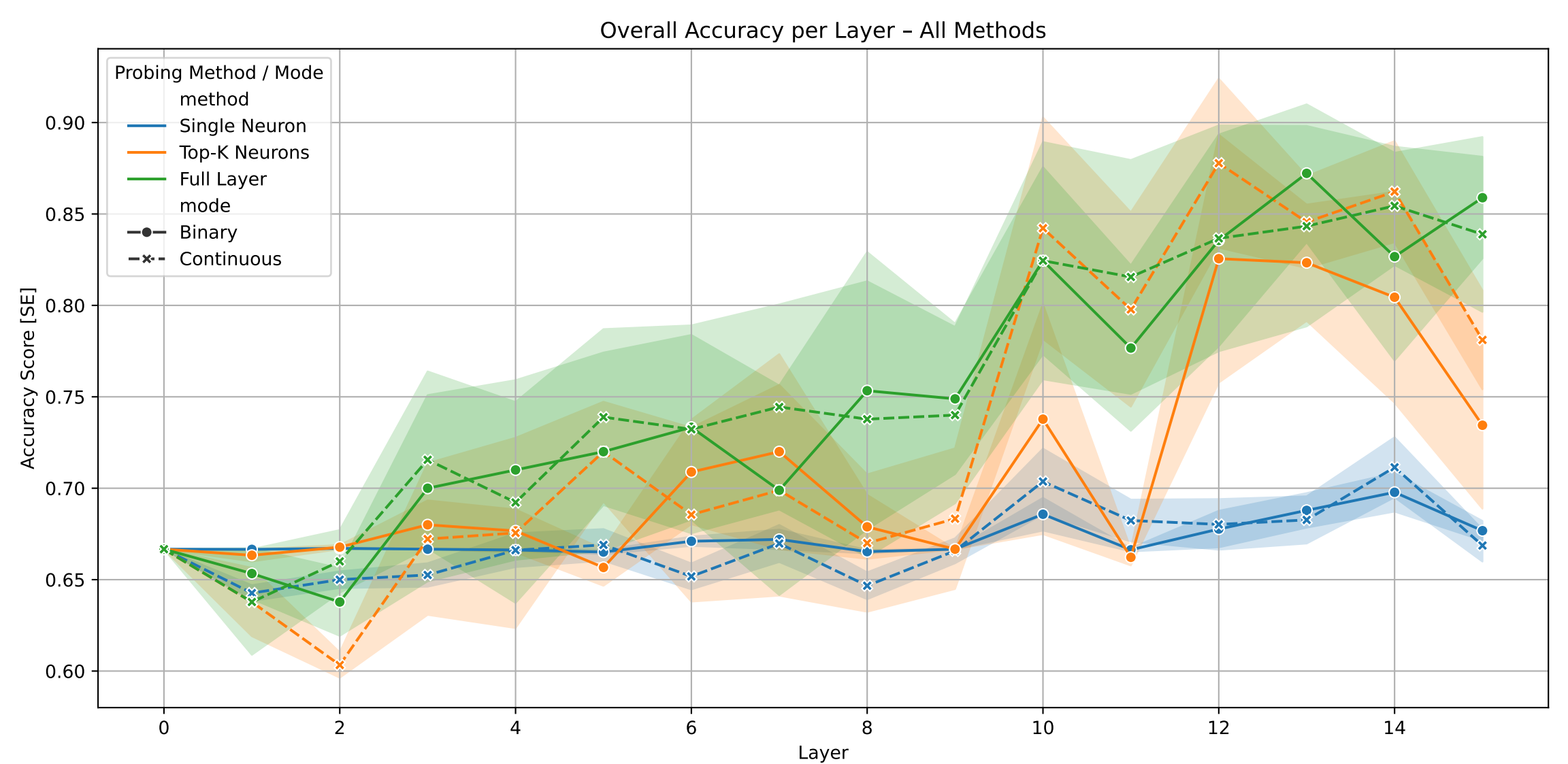
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	13.0	14.0
Full Layer	f1_max	0.9498	0.9528
Full Layer	f1_mean	0.7385	0.7441
Full Layer	f1_std	0.1133	0.1147
Single Neuron	f1_best_layer	14.0	14.0
Single Neuron	f1_max	0.9602	0.9629
Single Neuron	f1_mean	0.5485	0.6148
Single Neuron	f1_std	0.0563	0.0822
Top-K Neurons	f1_best_layer	12.0	12.0
Top-K Neurons	f1_max	0.9602	0.9629
Top-K Neurons	f1_mean	0.6358	0.716
Top-K Neurons	f1_std	0.1249	0.1188


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.72
0.70
0.68
0.66
0.64
0
2
4
6
8
10
12
14
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
0.60
0
2
4
6
8
10
12
14
Layer


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
Layer
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0
2
4
6
8
10
12
14


Overall Accuracy per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
Accuracy Score [SE]
0.90
0.85
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	13.0	14.0
Full Layer	accuracy_max	0.95	0.9533
Full Layer	accuracy_mean	0.751	0.7549
Full Layer	accuracy_std	0.0992	0.1009
Single Neuron	accuracy_best_layer	14.0	14.0
Single Neuron	accuracy_max	0.96	0.9633
Single Neuron	accuracy_mean	0.6728	0.6694
Single Neuron	accuracy_std	0.03	0.0579
Top-K Neurons	accuracy_best_layer	12.0	12.0
Top-K Neurons	accuracy_max	0.96	0.9633
Top-K Neurons	accuracy_mean	0.7108	0.7325
Top-K Neurons	accuracy_std	0.0763	0.1028