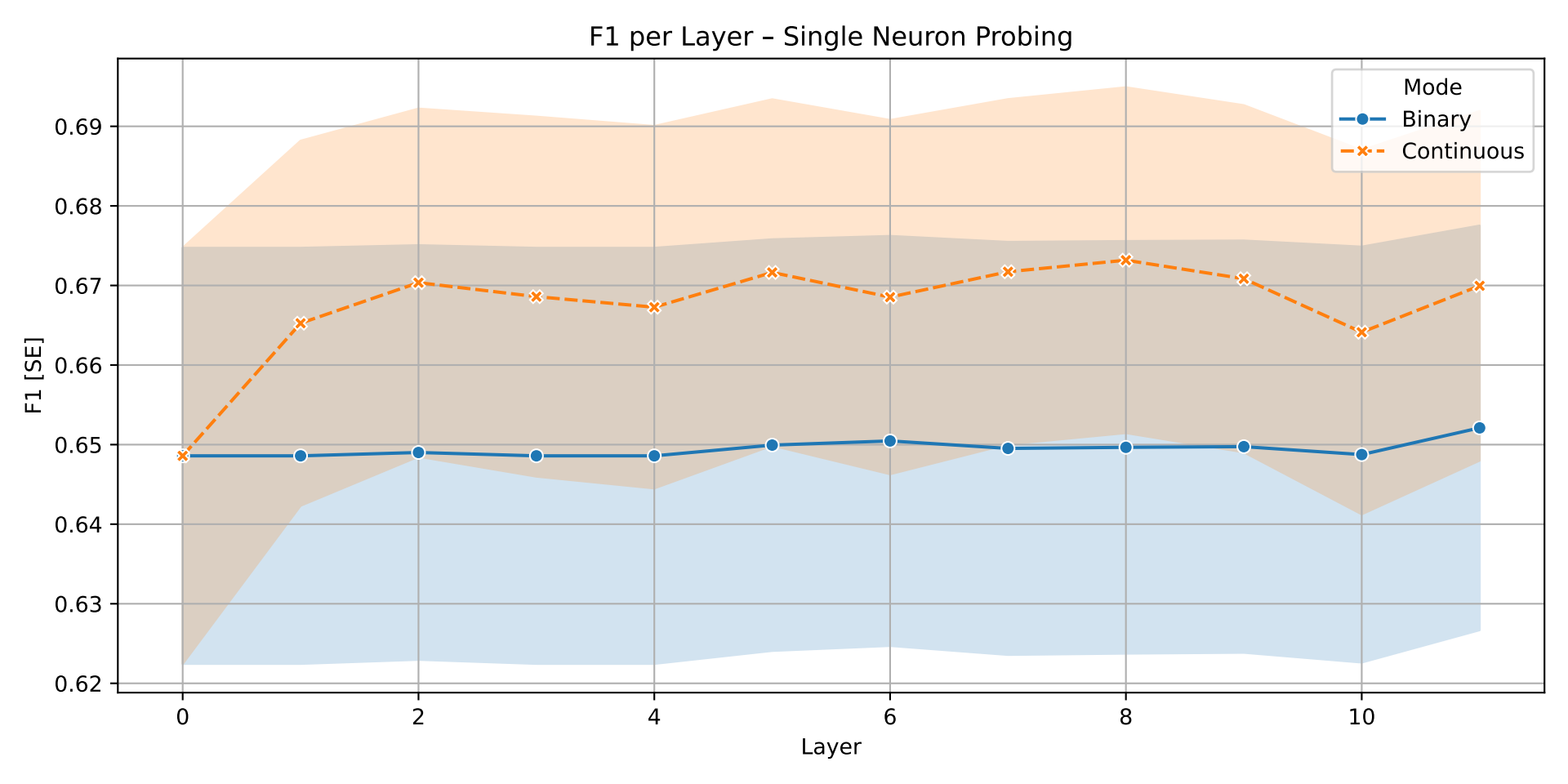
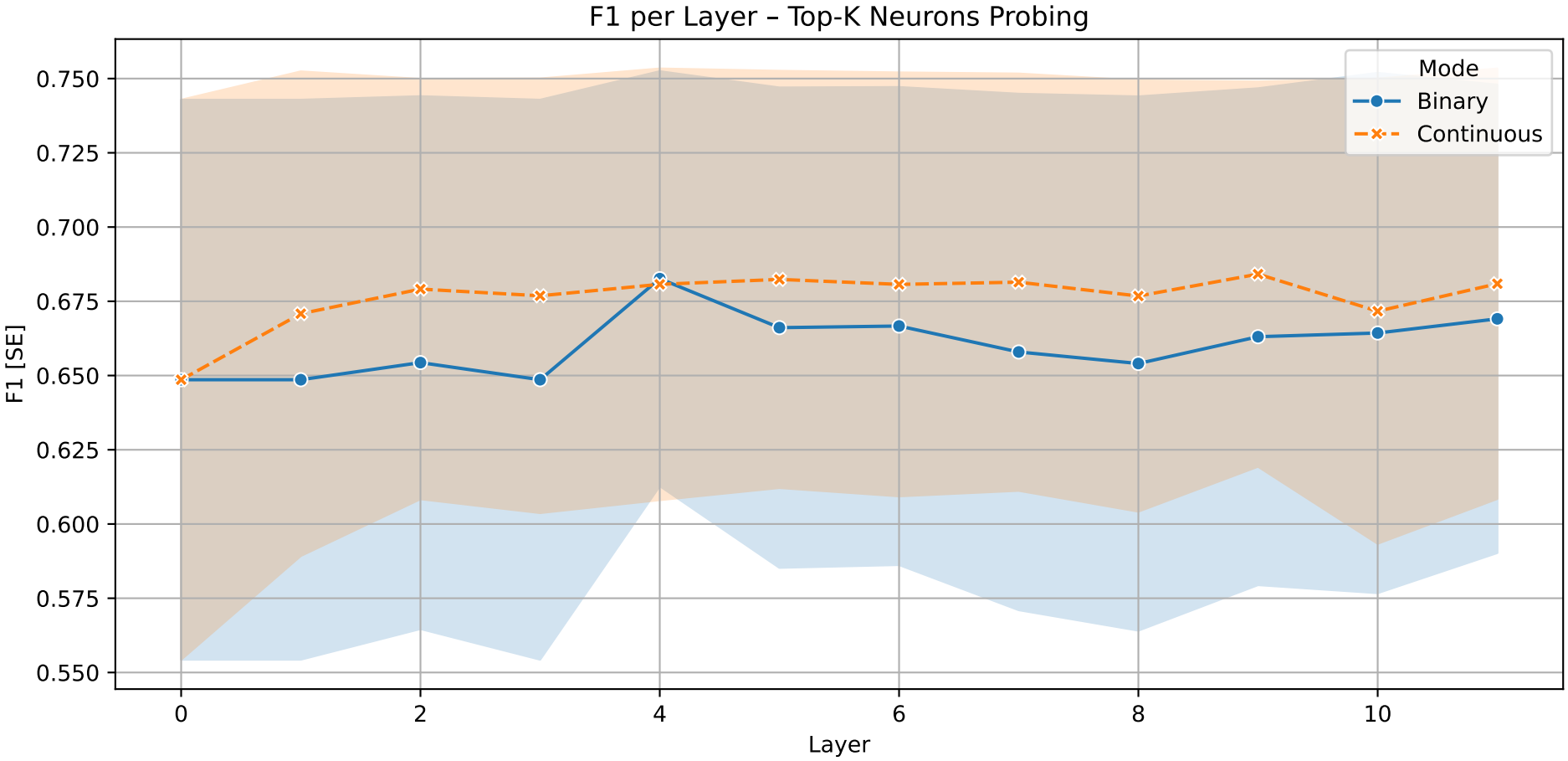
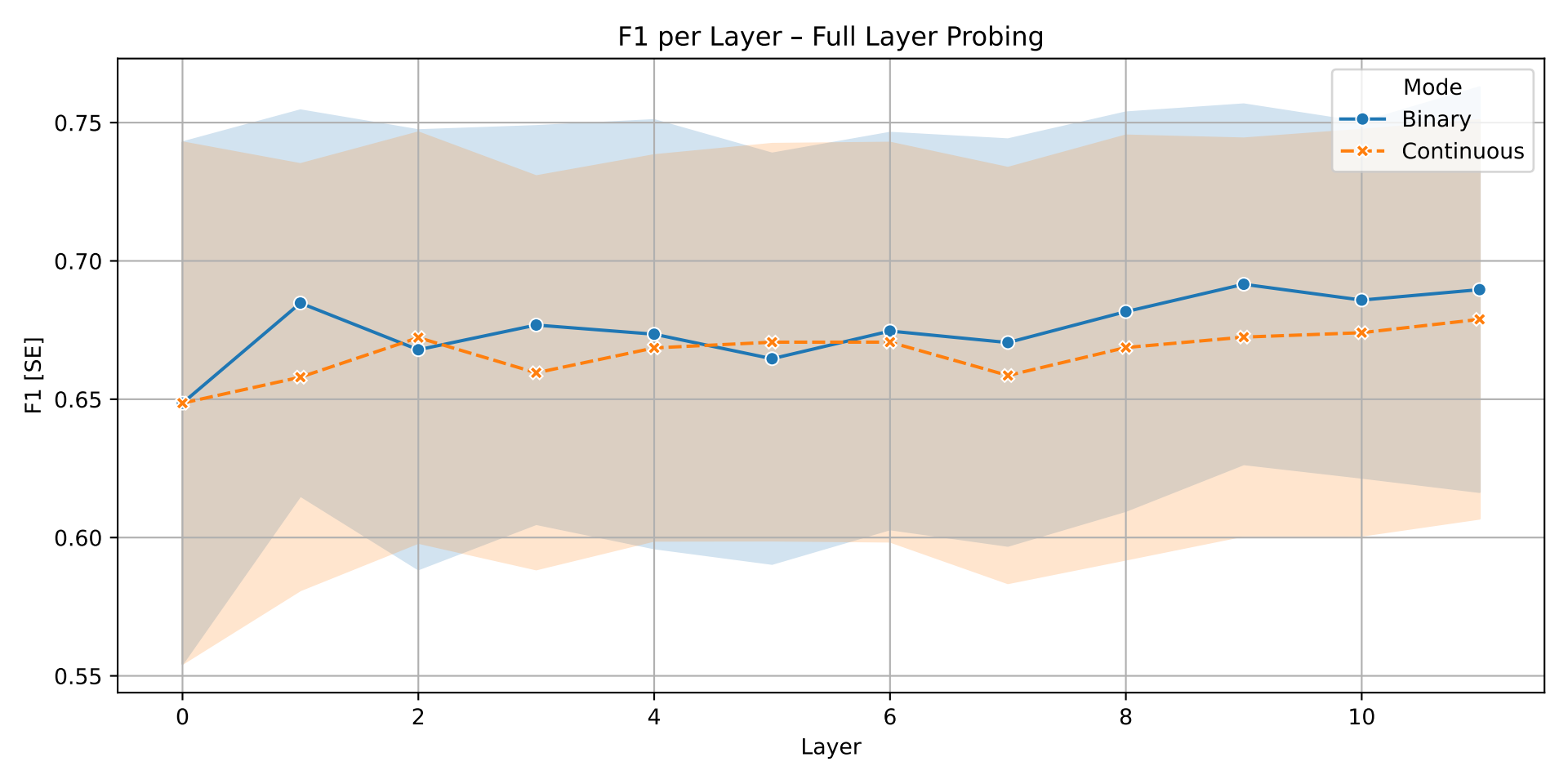
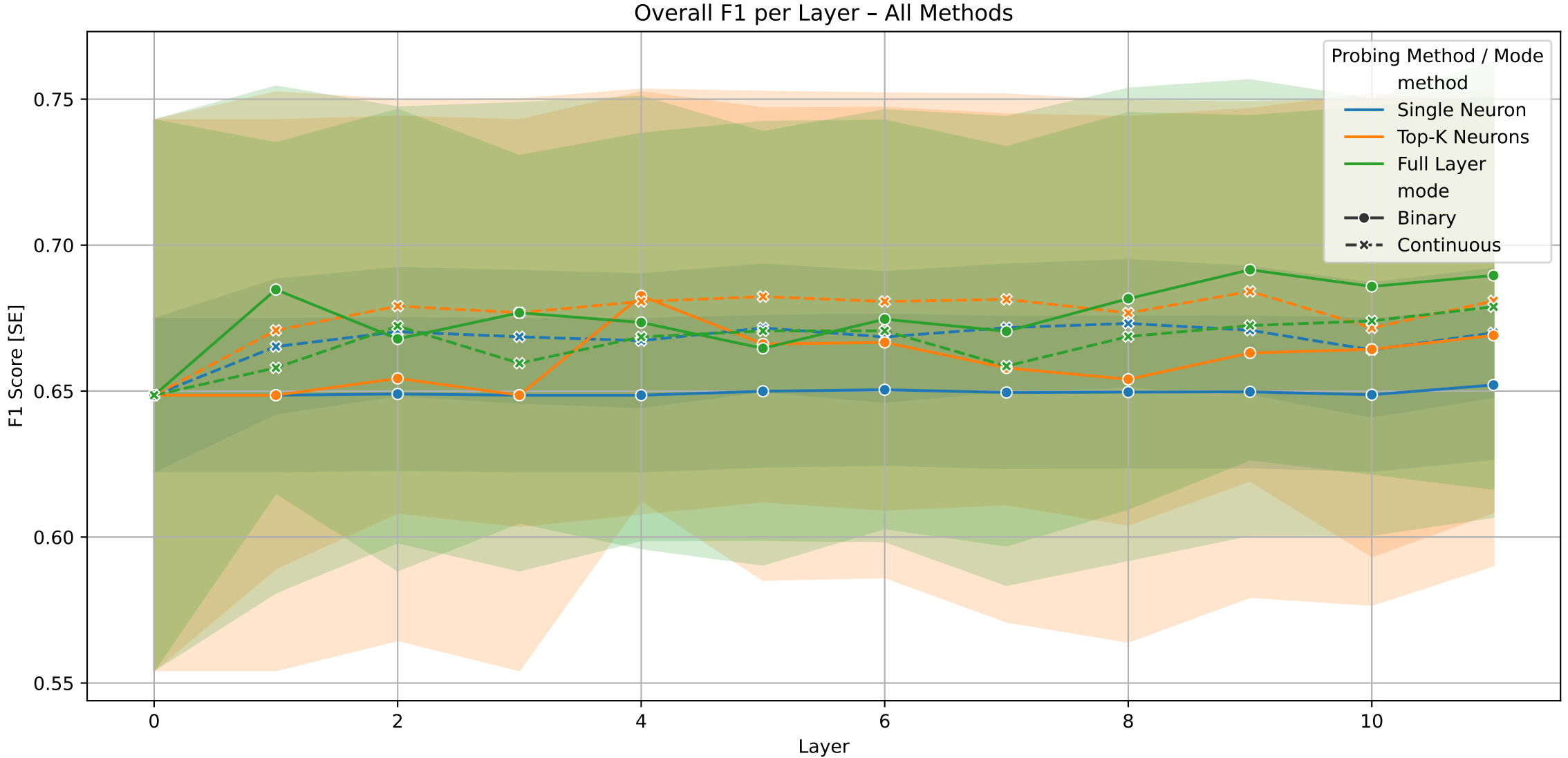

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.69
0.68
0.67
0.66
0.65
0.64
0.63
0.62
0
2
4
6
8
10
Layer

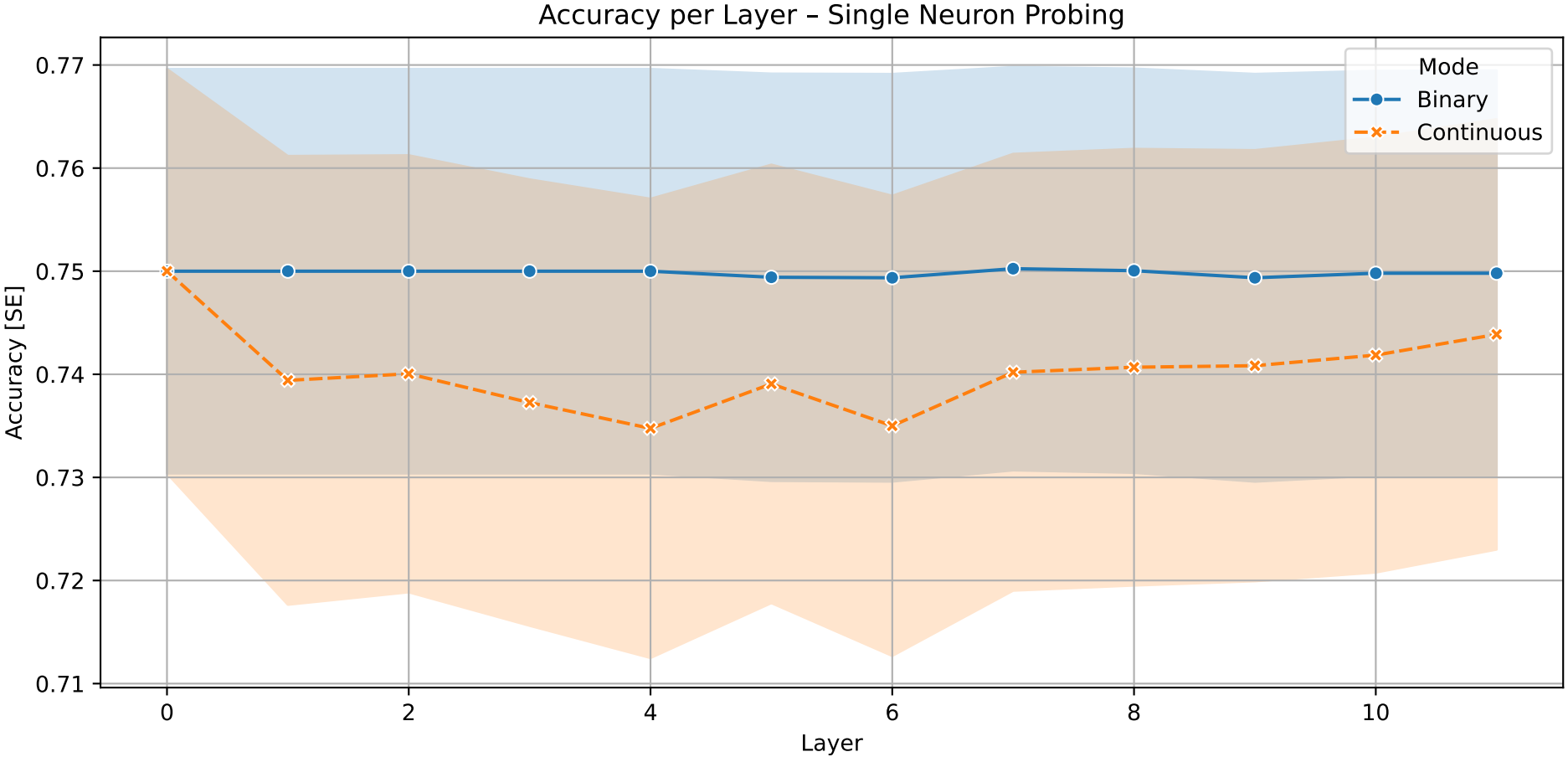

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.750
0.725
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0
2
4
6
8
10
Layer

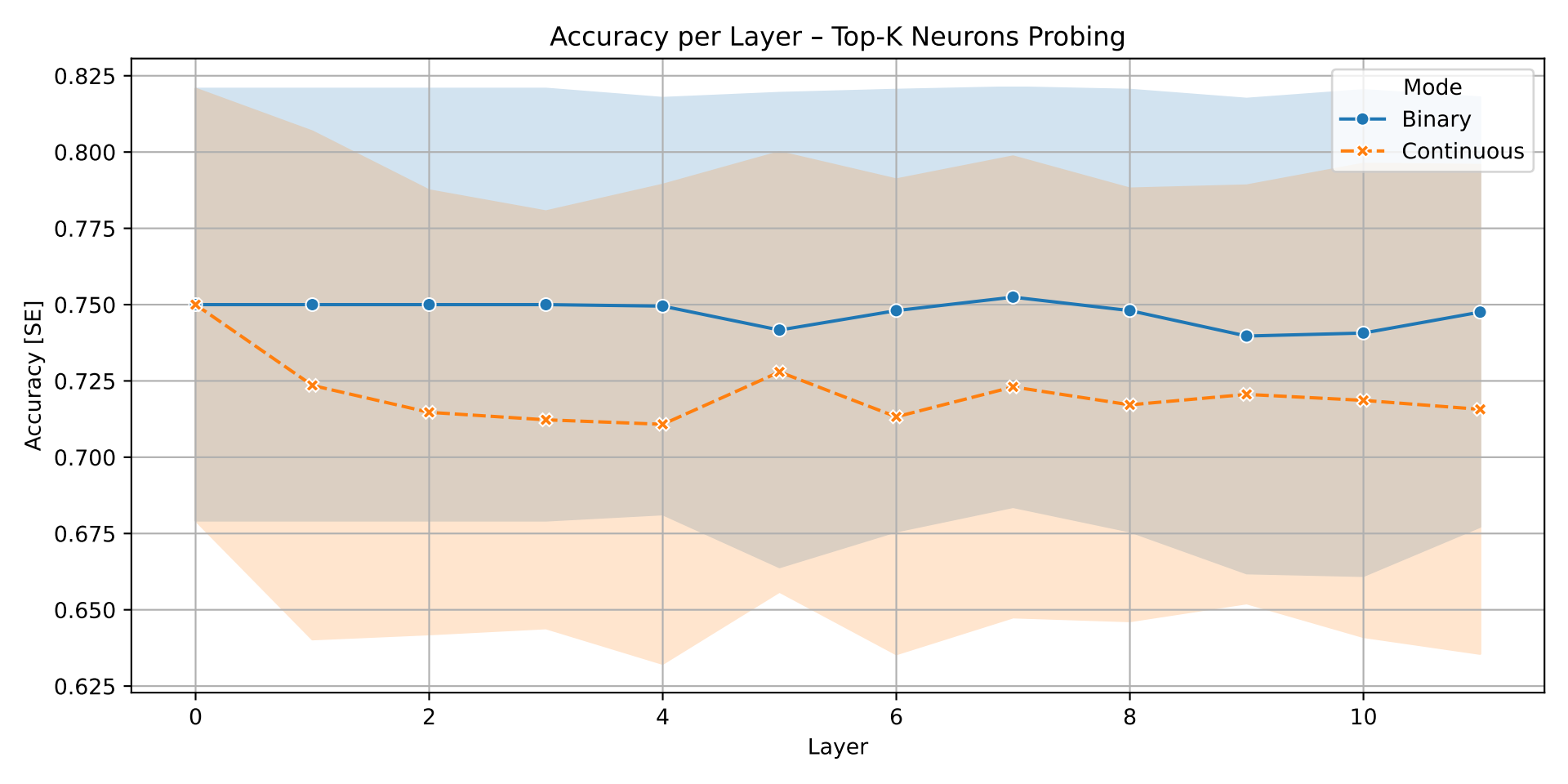

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
Layer

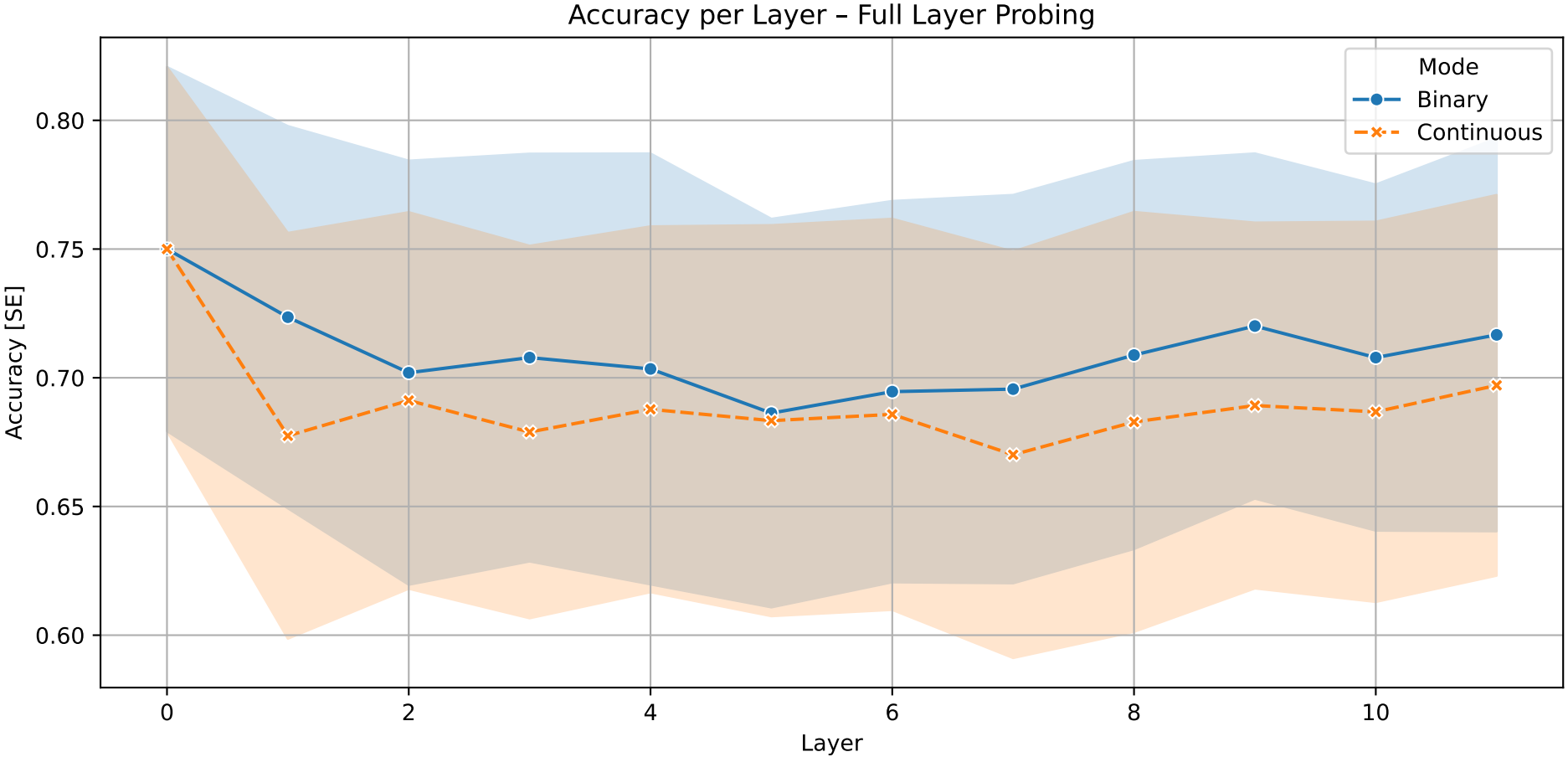

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

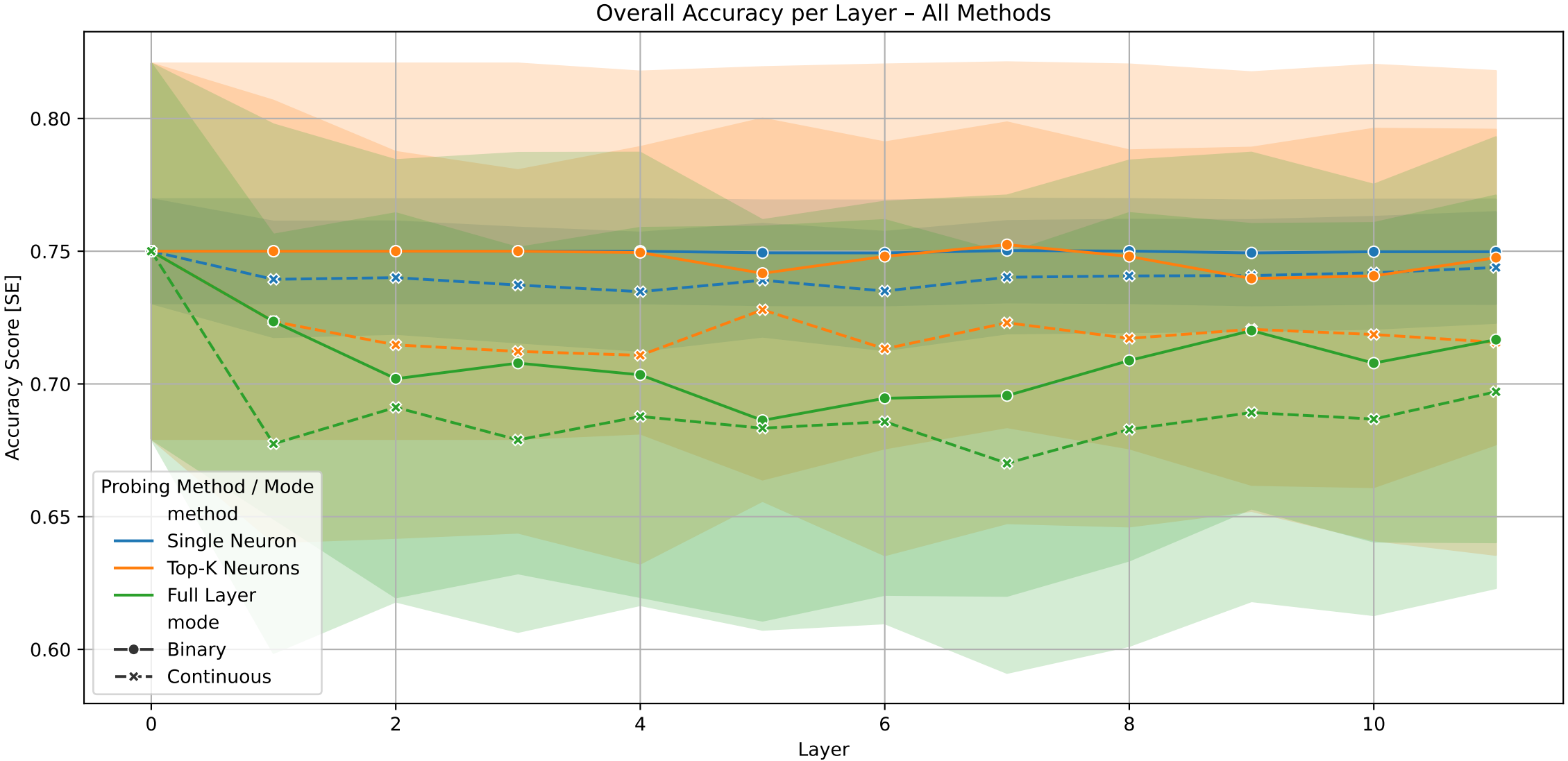
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	9.0	11.0
Full Layer	f1_max	0.8559	0.8574
Full Layer	f1_mean	0.6758	0.6667
Full Layer	f1_std	0.1304	0.1318
Single Neuron	f1_best_layer	11.0	8.0
Single Neuron	f1_max	0.8526	0.8526
Single Neuron	f1_mean	0.6495	0.6675
Single Neuron	f1_std	0.1623	0.1412
Top-K Neurons	f1_best_layer	4.0	9.0
Top-K Neurons	f1_max	0.8526	0.8563
Top-K Neurons	f1_mean	0.6603	0.6762
Top-K Neurons	f1_std	0.1511	0.1311


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.77
0.76
0.75
0.74
0.73
0.72
0.71
0
2
4
6
8
10
Layer


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
Layer
Mode
Binary
Continuous
0.825
0.800
0.775
0.750
0.725
0.700
0.675
0.650
0.625
0
2
4
6
8
10


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
Layer


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
Layer
0.80
0.75
0.70
0.65
0.60
0
2
4
6
8
10
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	0.0	0.0
Full Layer	accuracy_max	0.9	0.9
Full Layer	accuracy_mean	0.7097	0.69
Full Layer	accuracy_std	0.1327	0.1325
Single Neuron	accuracy_best_layer	7.0	0.0
Single Neuron	accuracy_max	0.9	0.9
Single Neuron	accuracy_mean	0.7498	0.7402
Single Neuron	accuracy_std	0.1229	0.1331
Top-K Neurons	accuracy_best_layer	7.0	0.0
Top-K Neurons	accuracy_max	0.9	0.9
Top-K Neurons	accuracy_mean	0.7473	0.7206
Top-K Neurons	accuracy_std	0.1271	0.1313