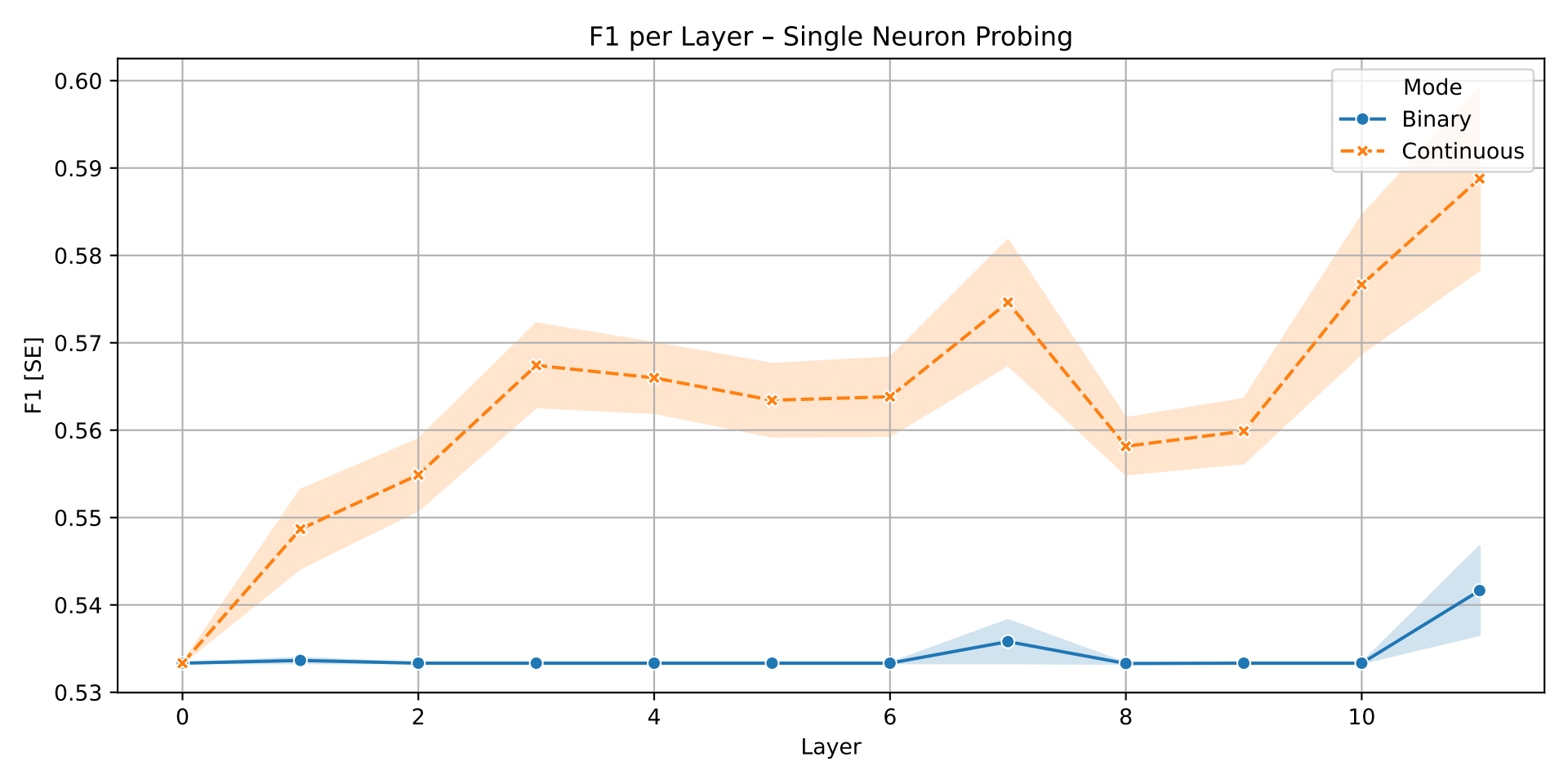
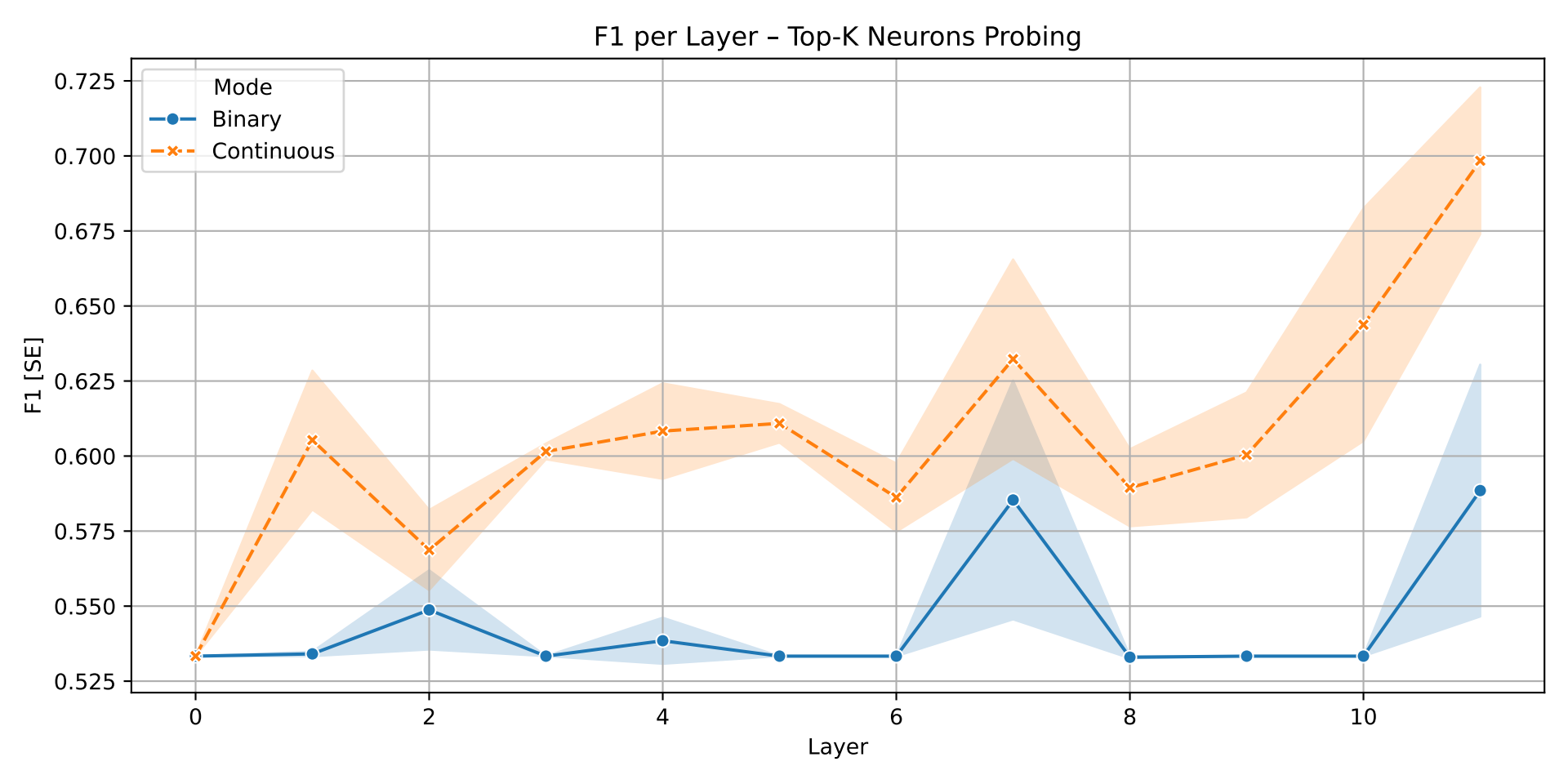
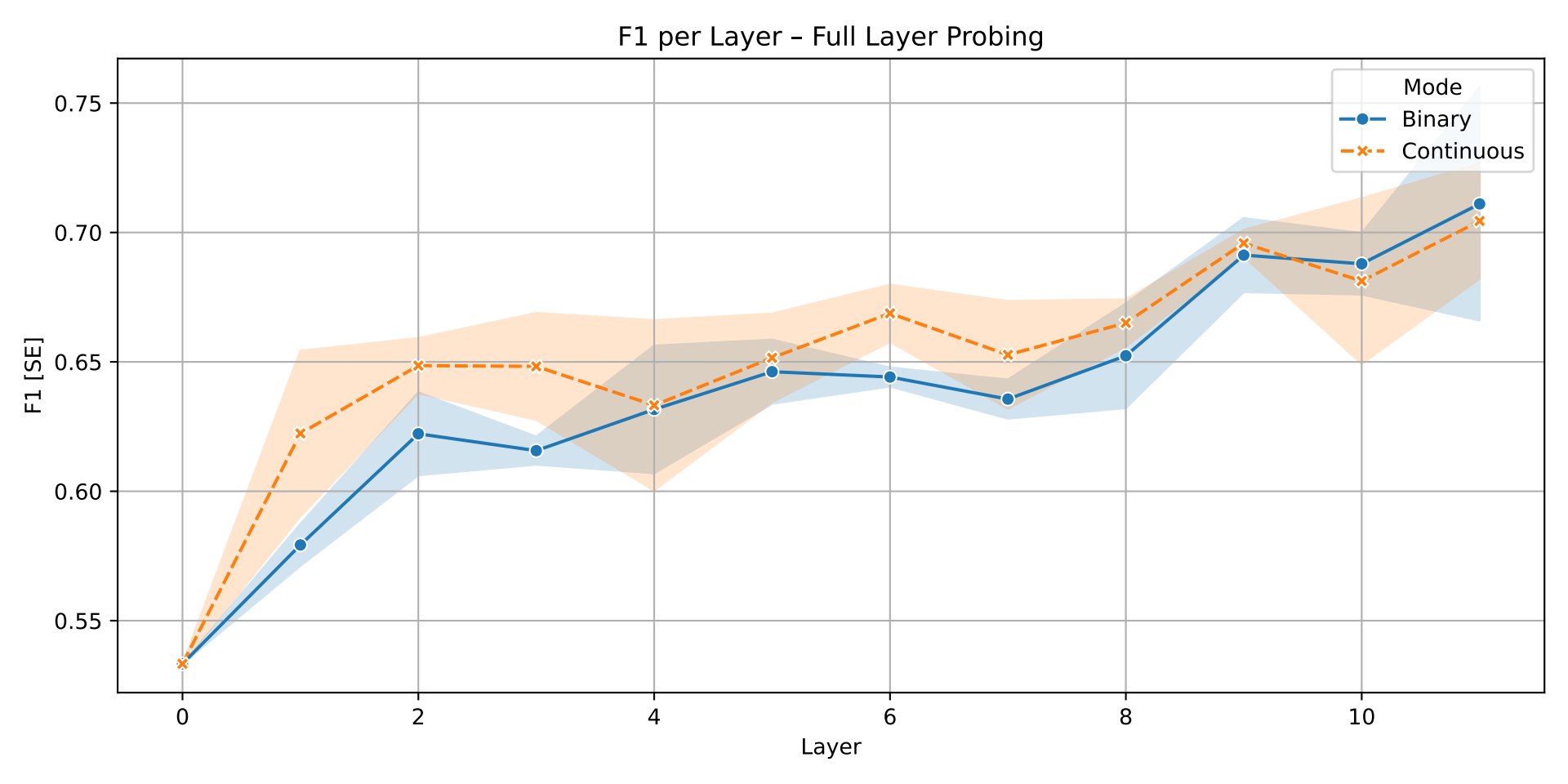
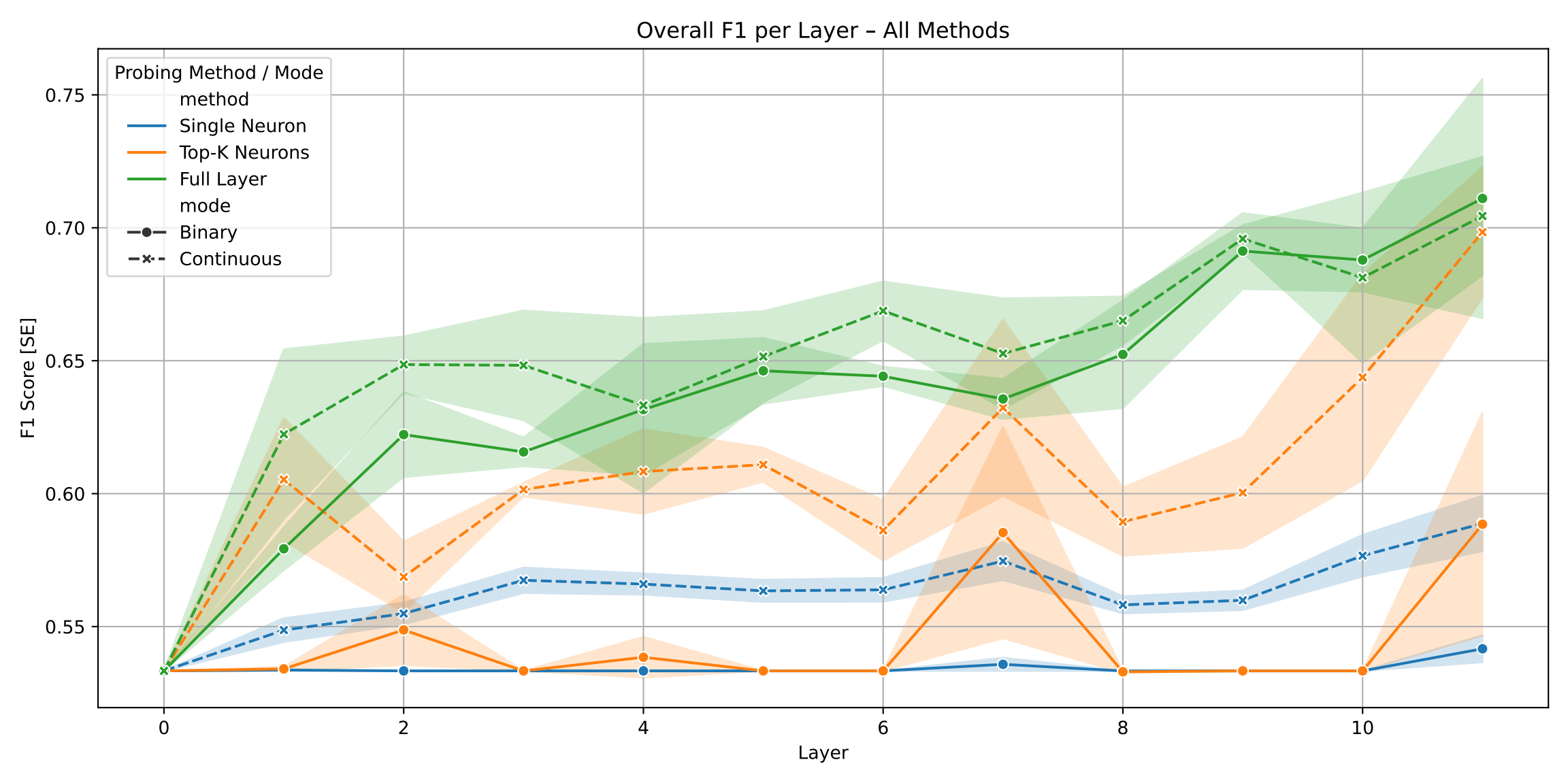

F1 per Layer – Single Neuron Probing
F1 [SE]
Layer
Mode
Binary
Continuous
0.60
0.59
0.58
0.57
0.56
0.55
0.54
0.53
0
2
4
6
8
10

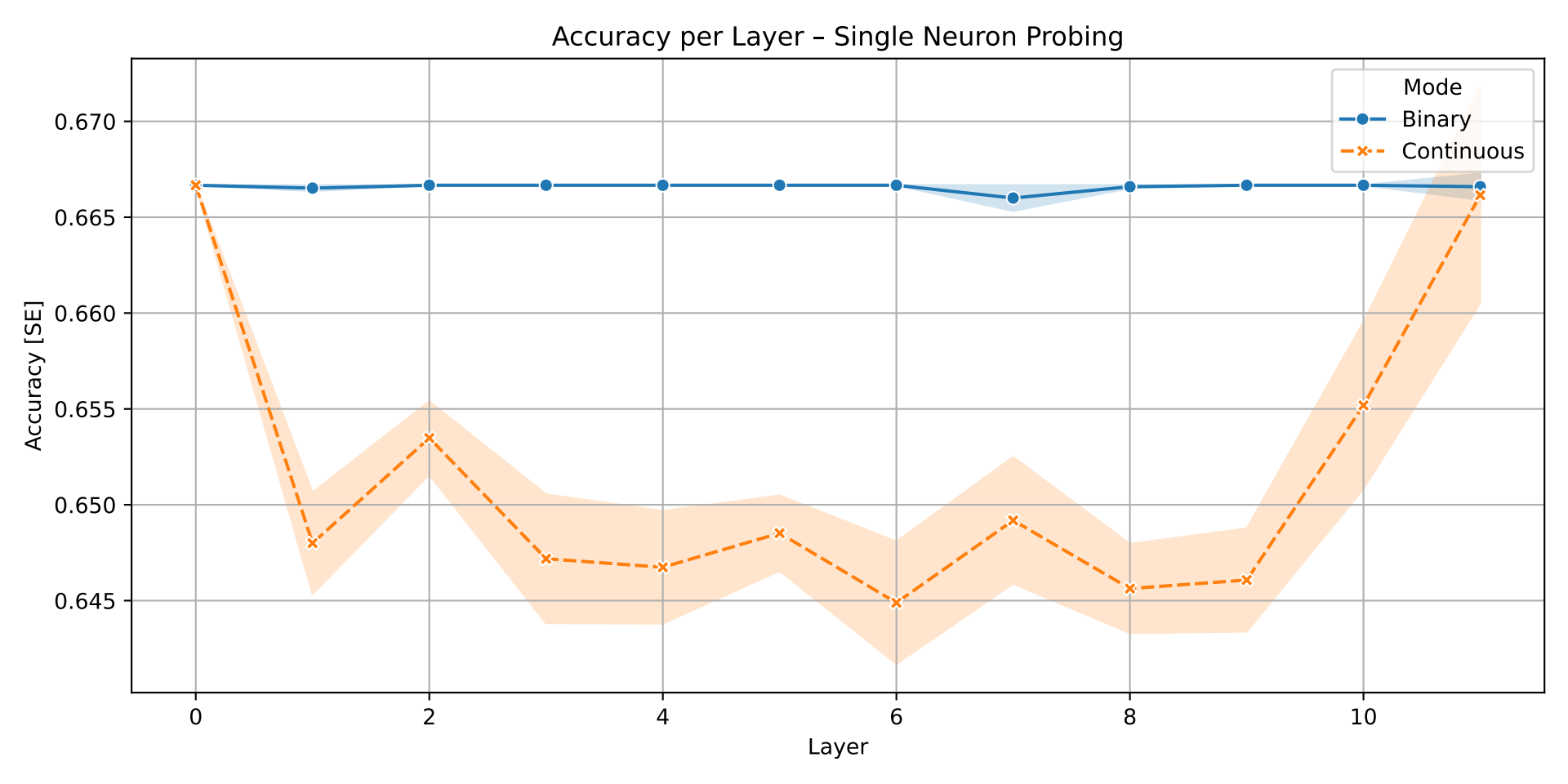

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.725
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0.525
0
2
4
6
8
10
Layer

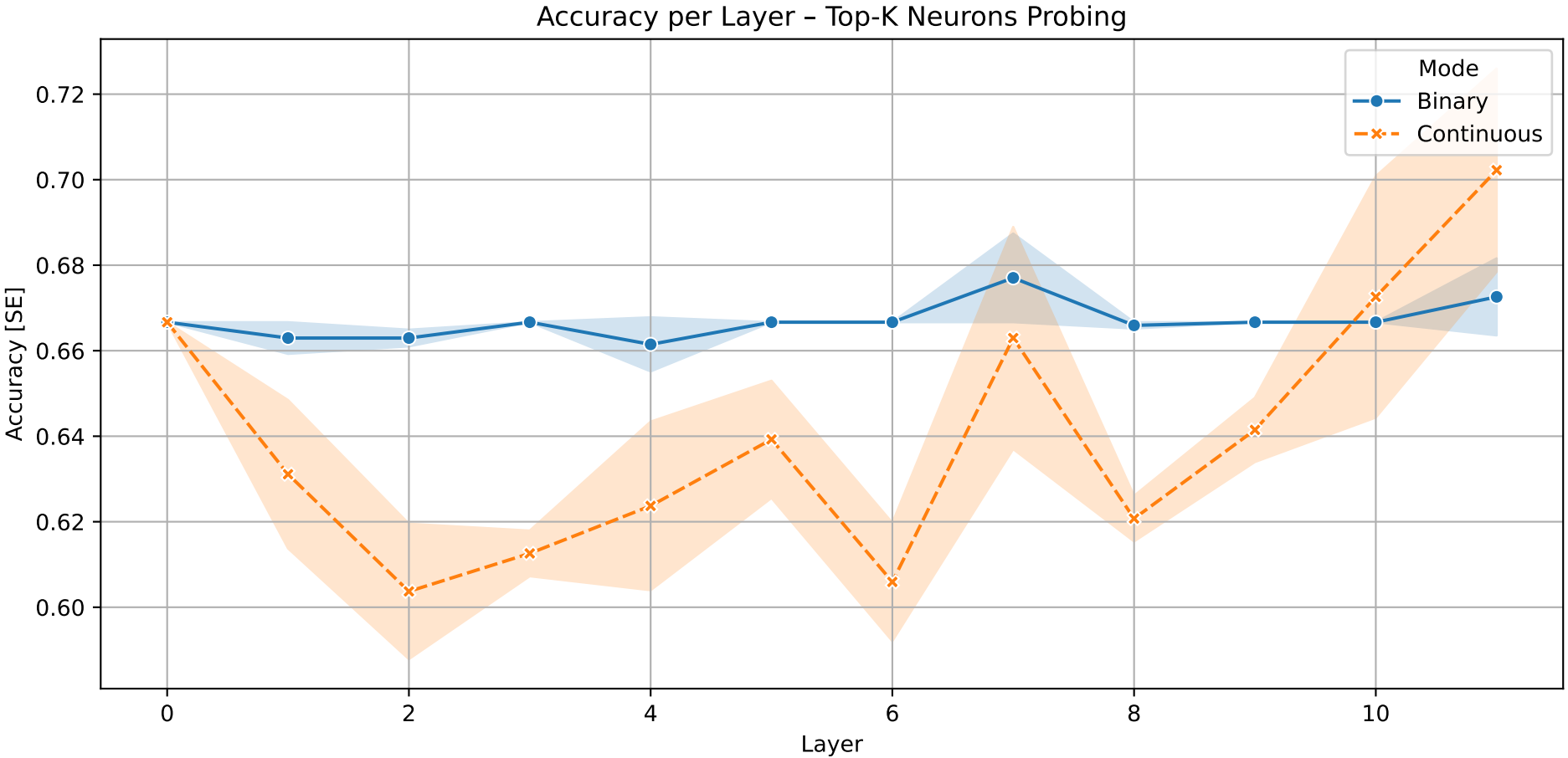

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
Layer

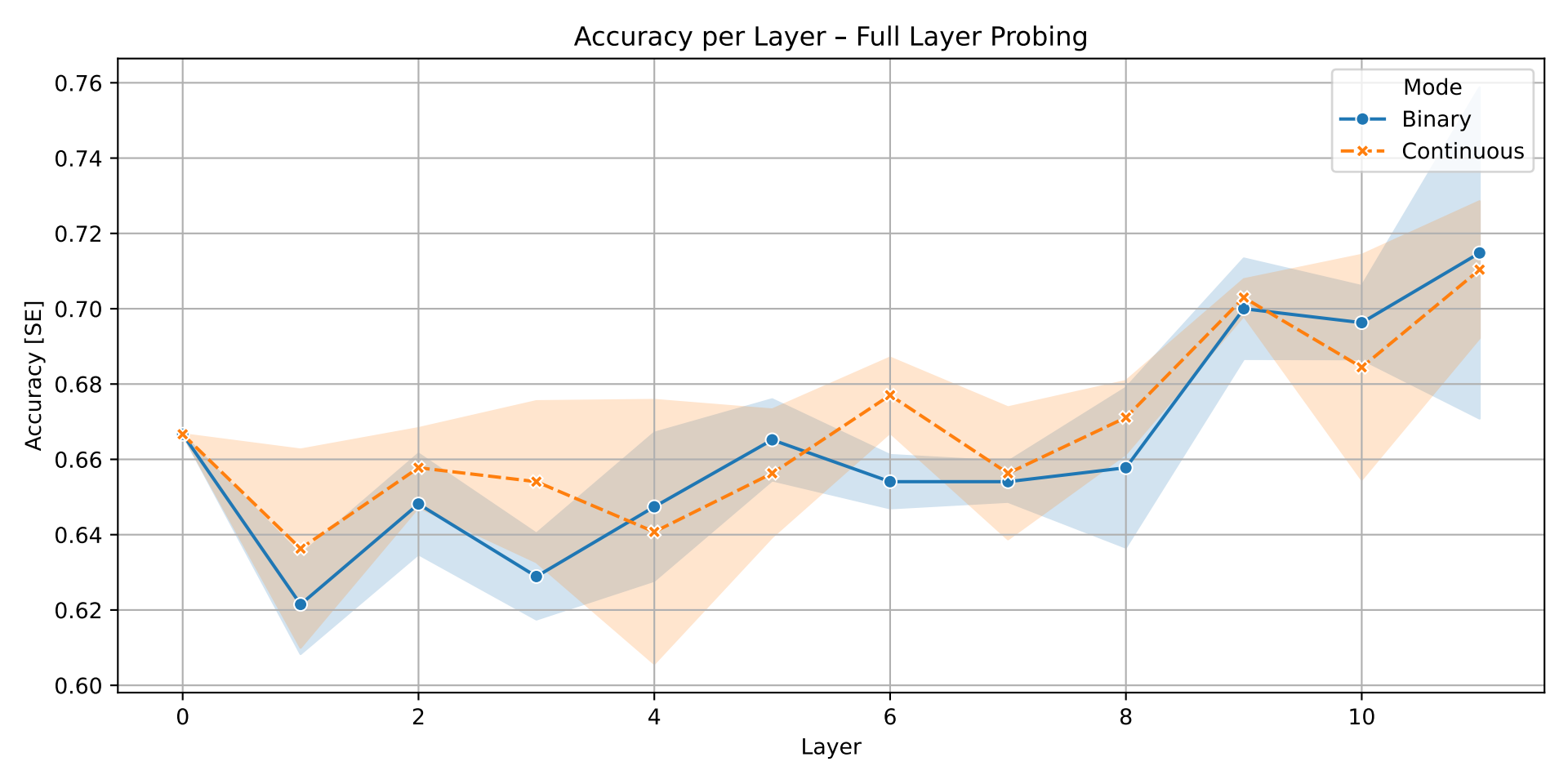

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
Layer

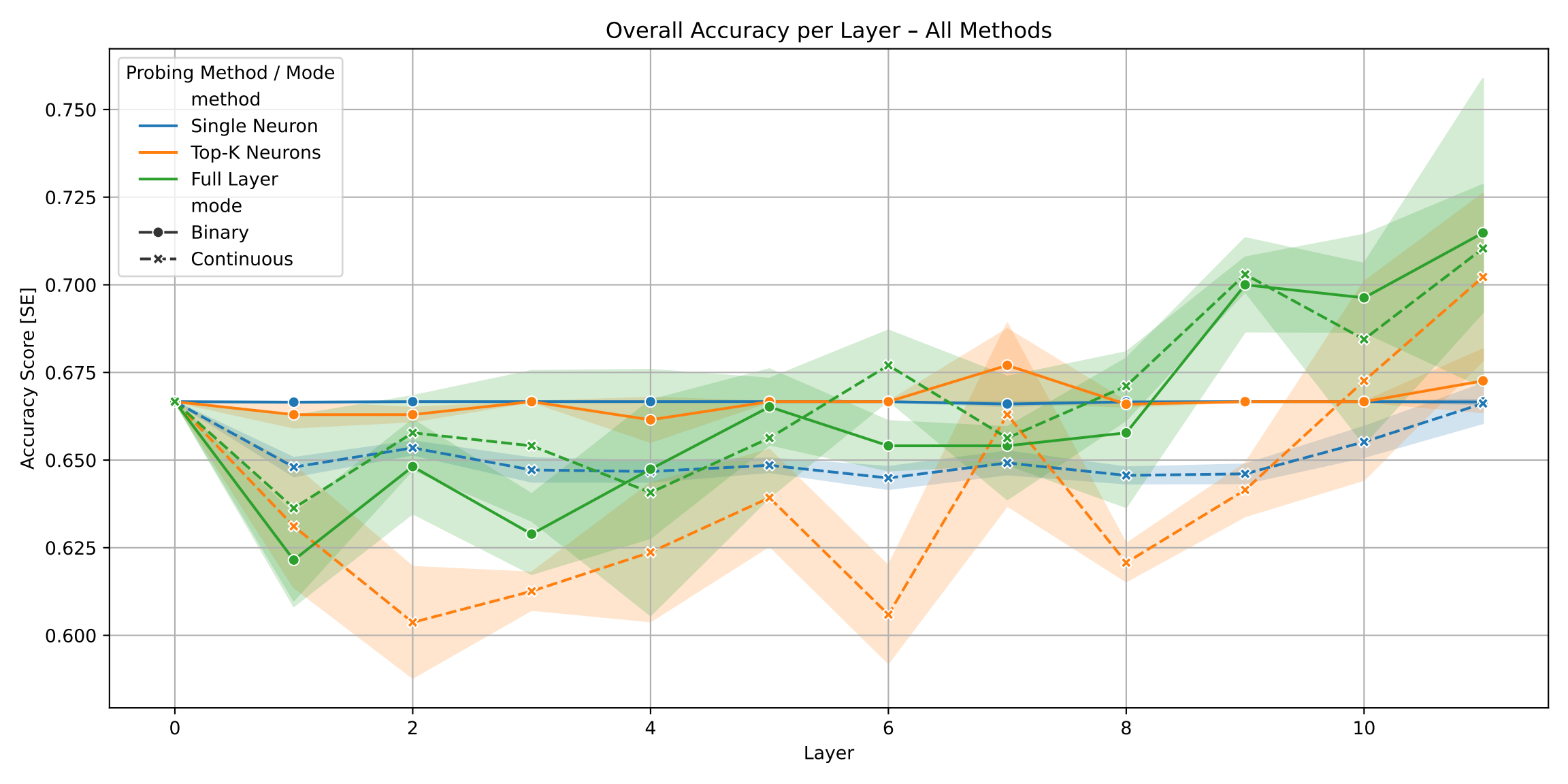
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	11.0	11.0
Full Layer	f1_max	0.7812	0.749
Full Layer	f1_mean	0.6376	0.6504
Full Layer	f1_std	0.0541	0.052
Single Neuron	f1_best_layer	11.0	11.0
Single Neuron	f1_max	0.6725	0.749
Single Neuron	f1_mean	0.5343	0.563
Single Neuron	f1_std	0.0091	0.0325
Top-K Neurons	f1_best_layer	11.0	11.0
Top-K Neurons	f1_max	0.6708	0.7429
Top-K Neurons	f1_mean	0.544	0.6065
Top-K Neurons	f1_std	0.0318	0.049


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.670
0.665
0.660
0.655
0.650
0.645
0
2
4
6
8
10
Layer


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.72
0.70
0.68
0.66
0.64
0.62
0.60
0
2
4
6
8
10
Layer


Accuracy per Layer – Full Layer Probing
Accuracy [SE]
Layer
Mode
Binary
Continuous
0.76
0.74
0.72
0.70
0.68
0.66
0.64
0.62
0.60
0
2
4
6
8
10


Overall Accuracy per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
Accuracy Score [SE]
0.750
0.725
0.700
0.675
0.650
0.625
0.600
0
2
4
6
8
10
Layer

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	11.0	11.0
Full Layer	accuracy_max	0.78	0.7467
Full Layer	accuracy_mean	0.6629	0.6678
Full Layer	accuracy_std	0.0372	0.0355
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.6822	0.7556
Single Neuron	accuracy_mean	0.6666	0.6515
Single Neuron	accuracy_std	0.0015	0.0184
Top-K Neurons	accuracy_best_layer	7.0	11.0
Top-K Neurons	accuracy_max	0.6978	0.7444
Top-K Neurons	accuracy_mean	0.6669	0.6402
Top-K Neurons	accuracy_std	0.0077	0.0382