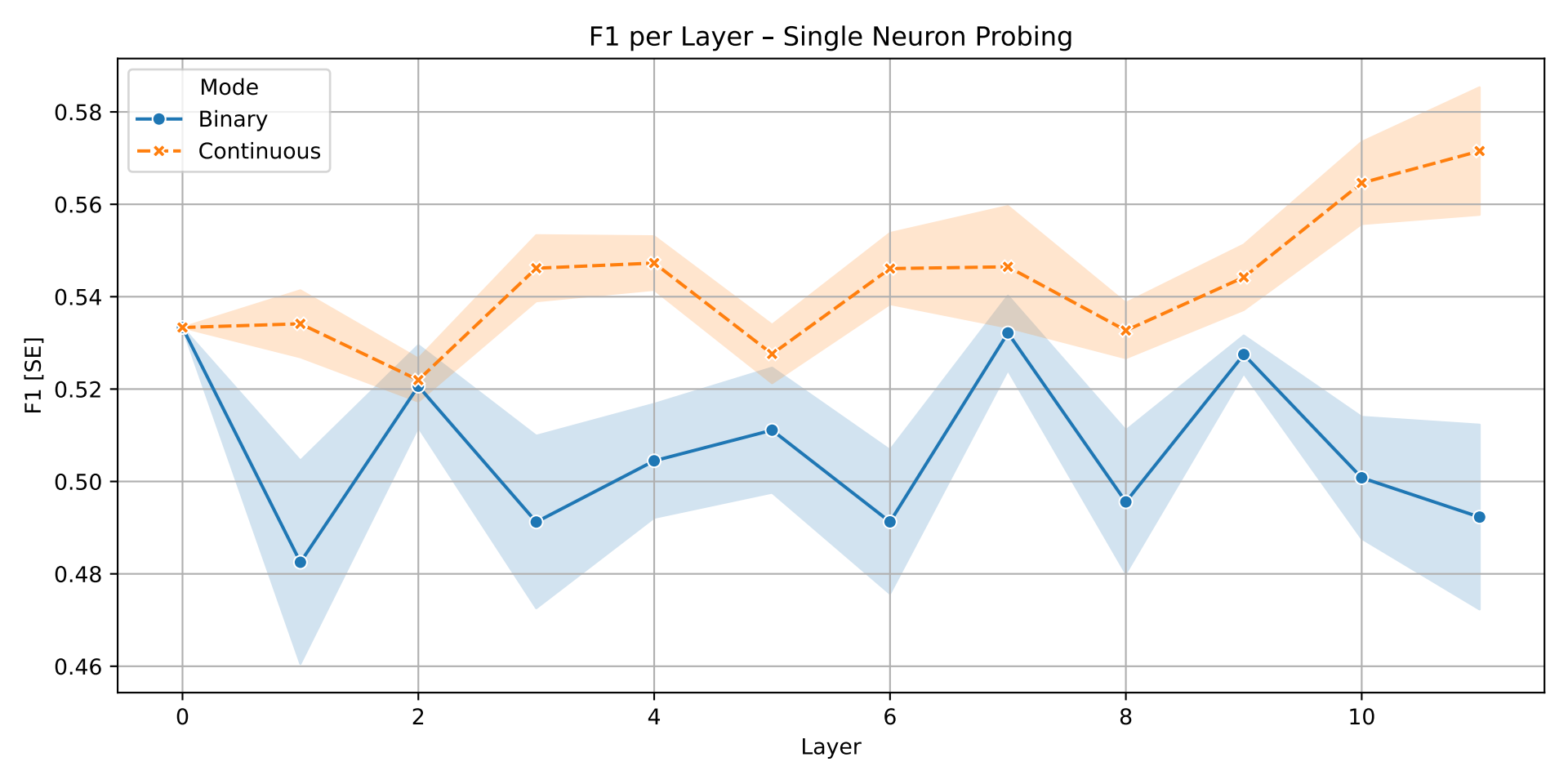
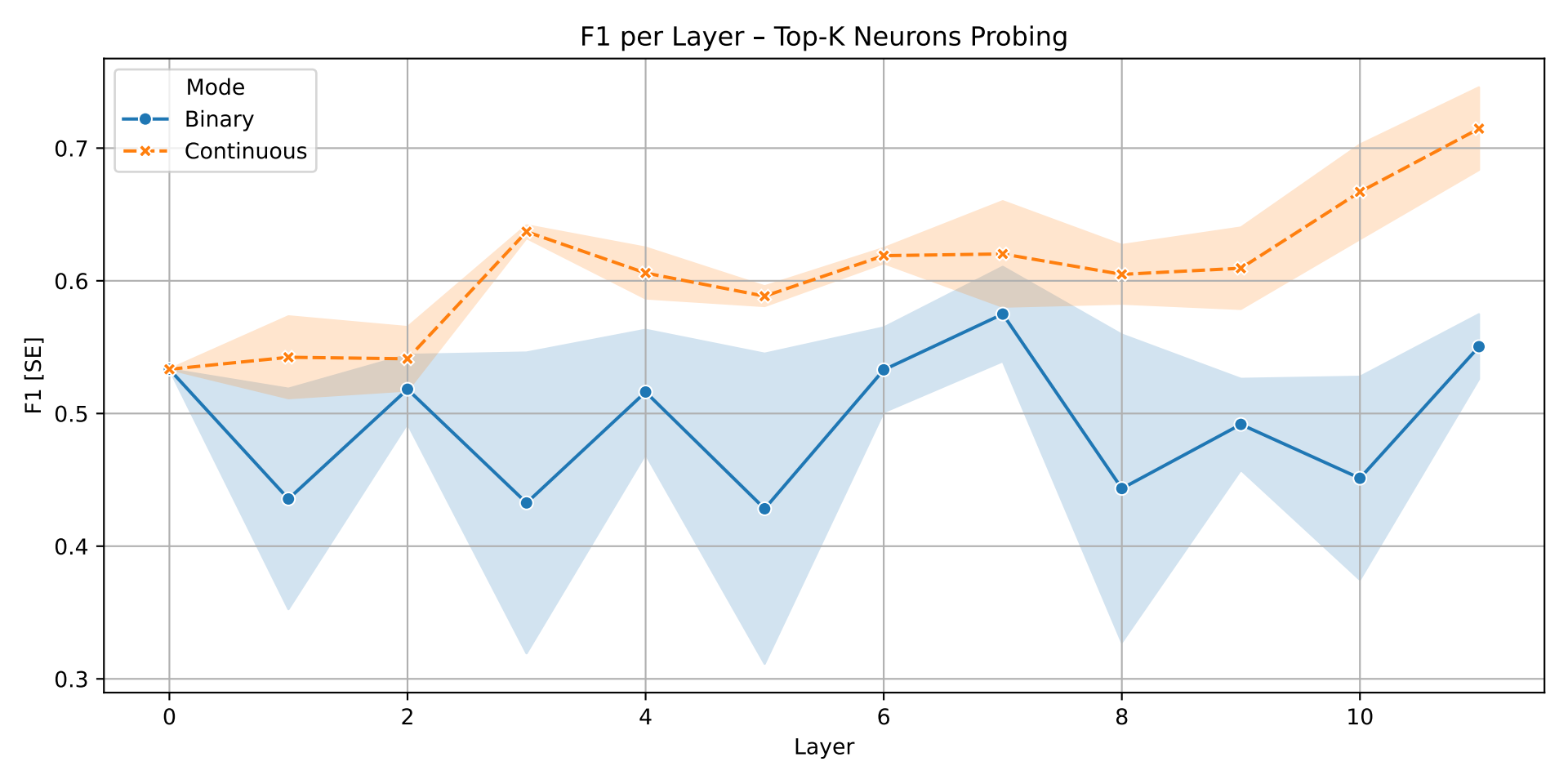
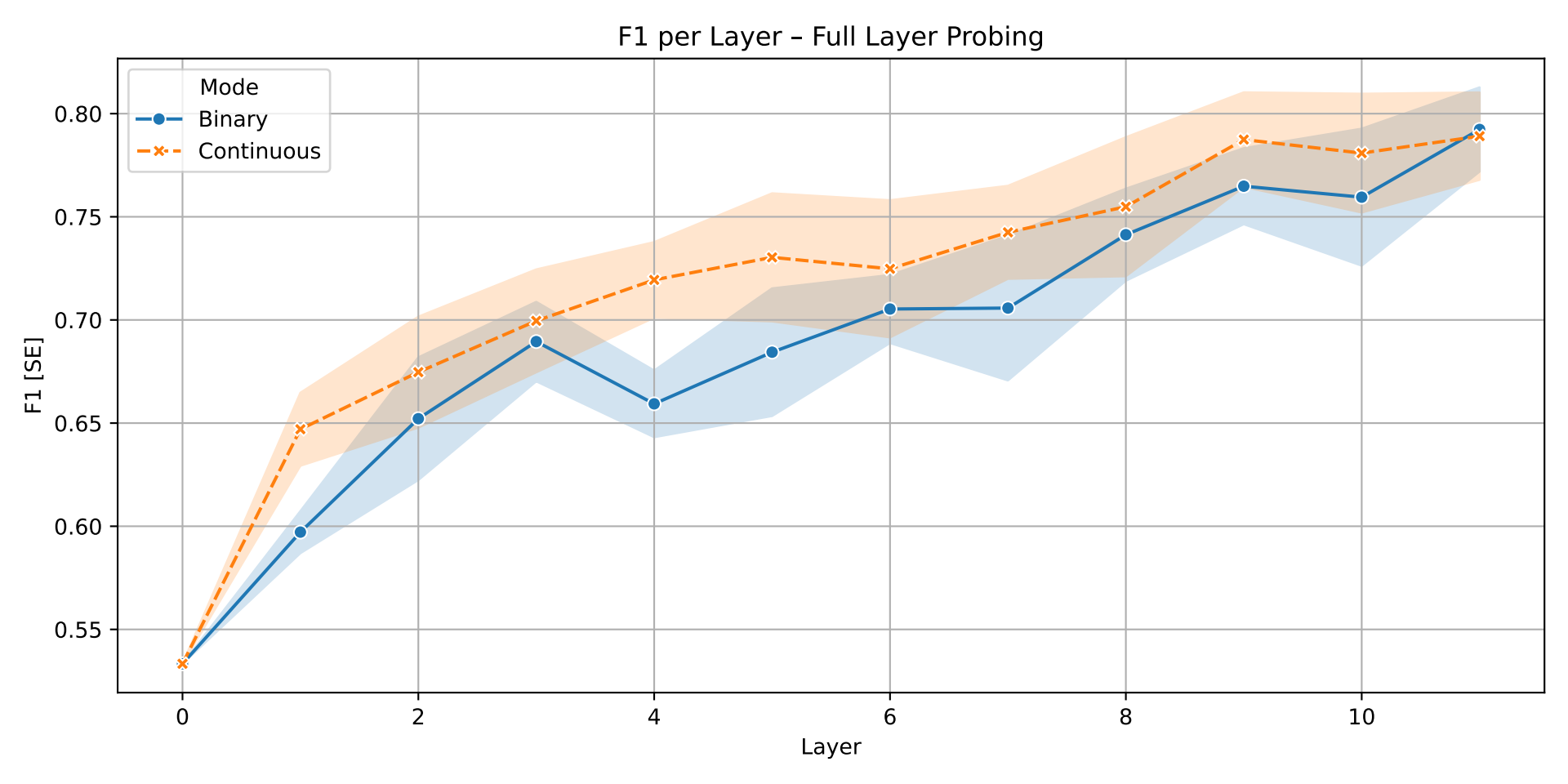
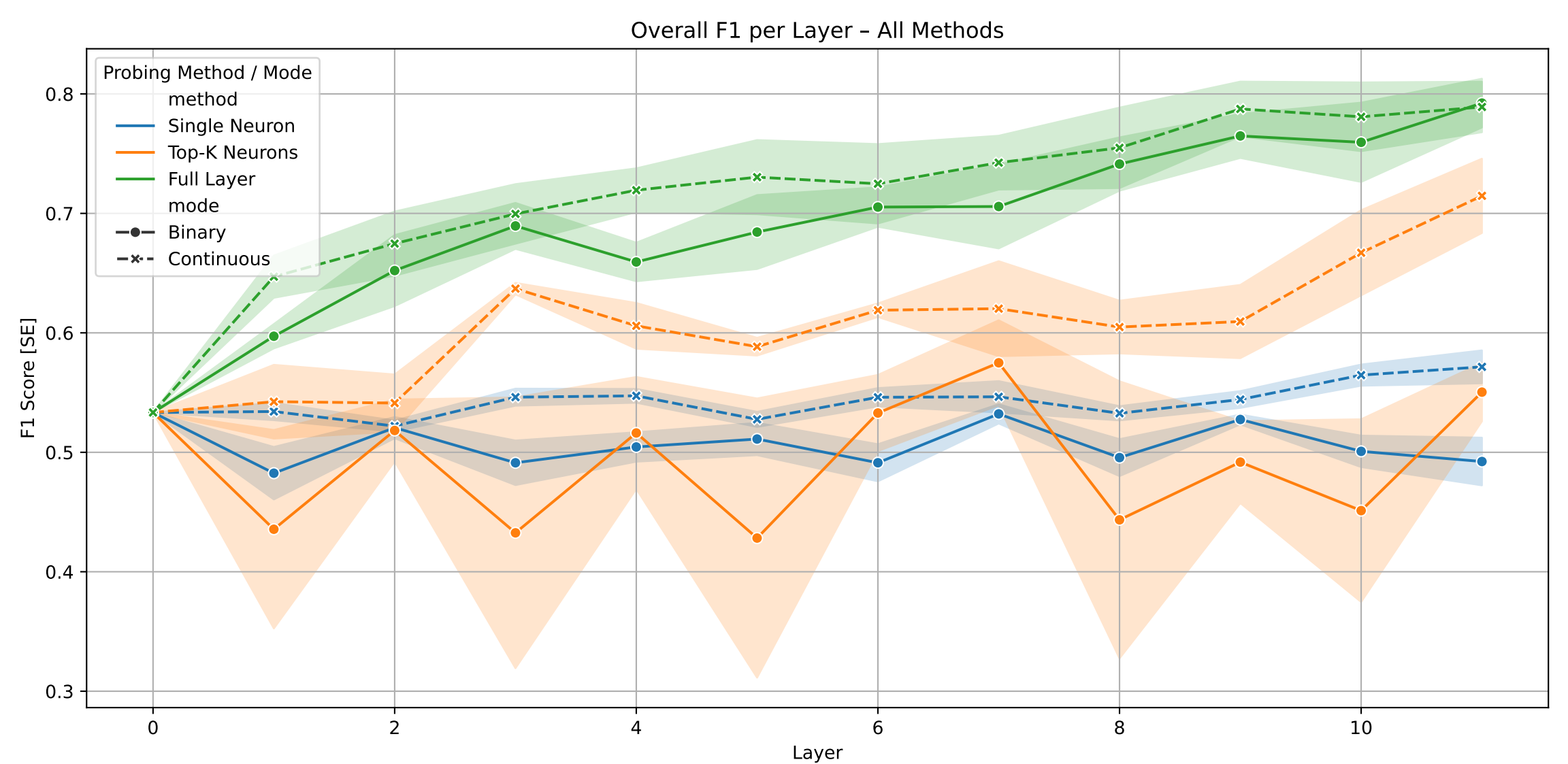

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.58
0.56
0.54
0.52
0.50
0.48
0.46
0
2
4
6
8
10
Layer

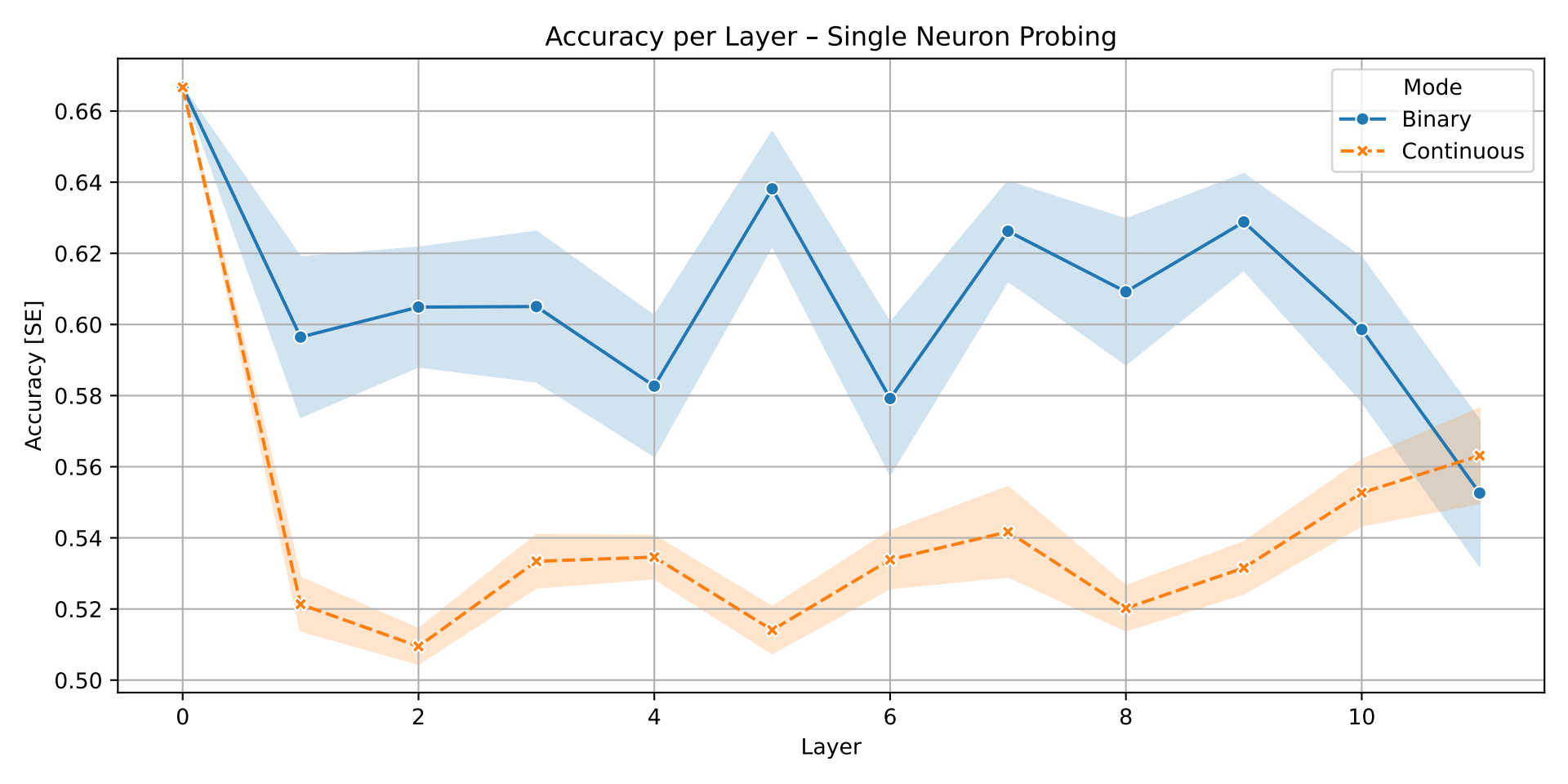

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.7
0.6
0.5
0.4
0.3
0
2
4
6
8
10
Layer

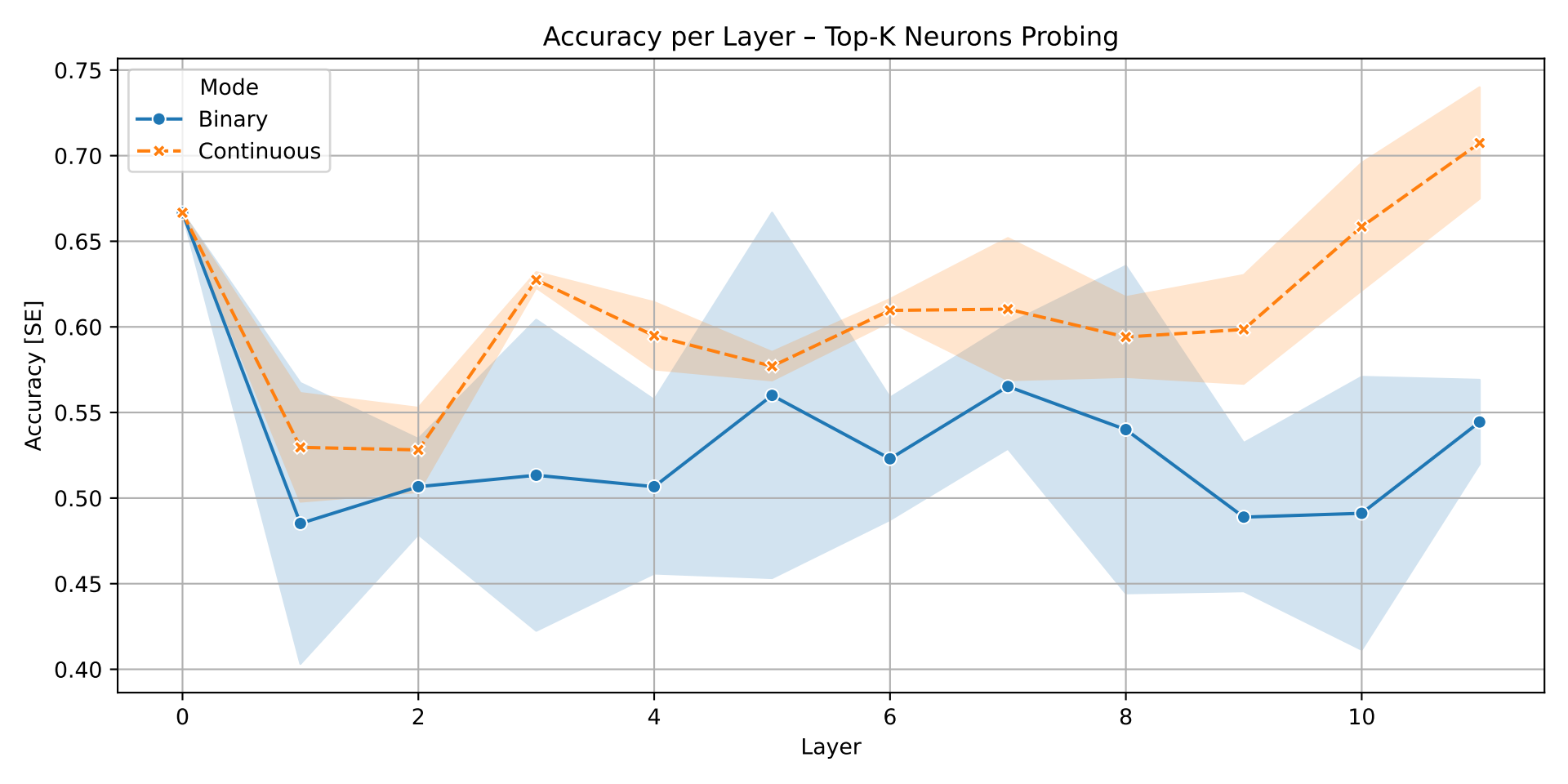

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.80
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
Layer

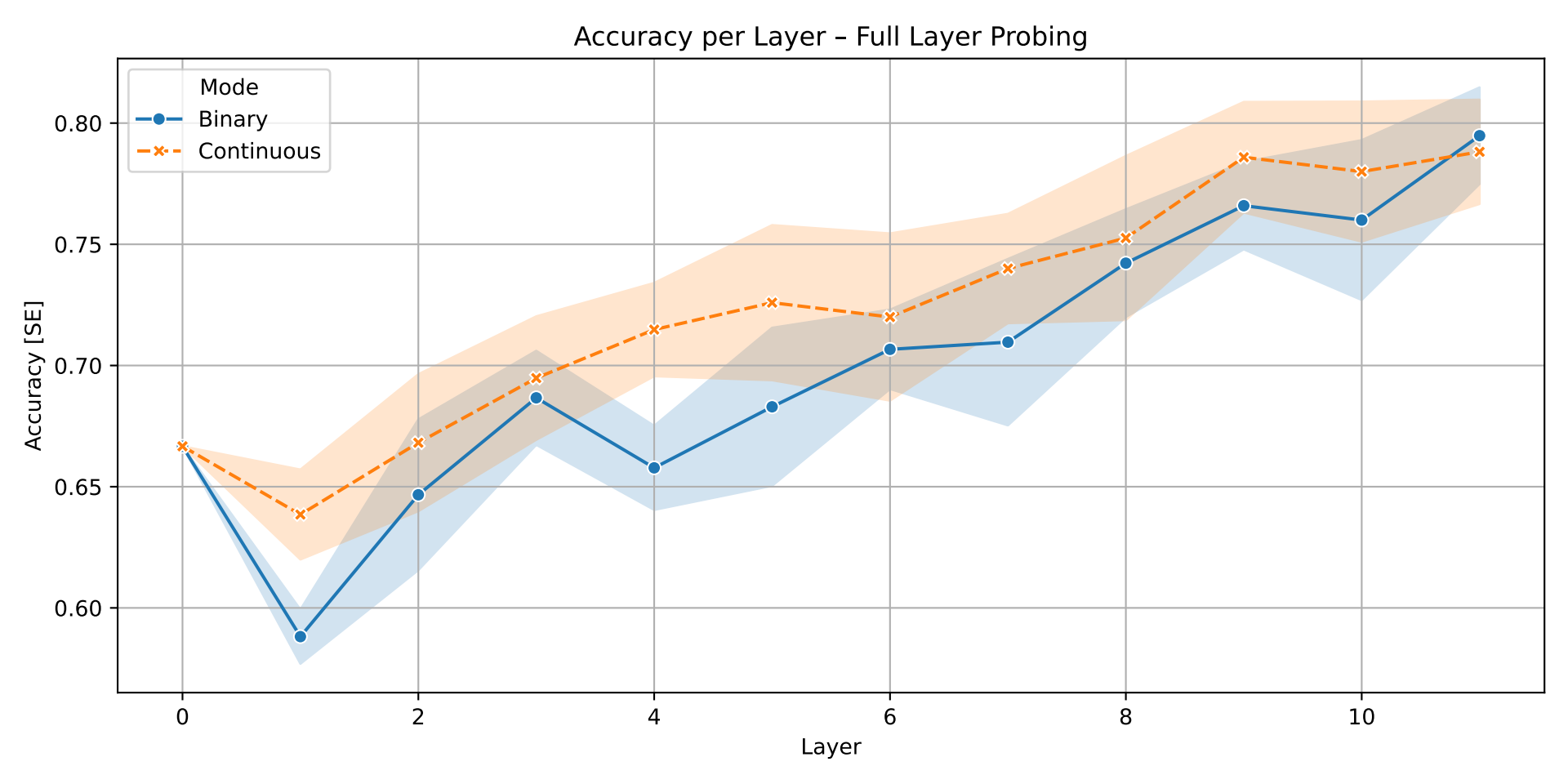

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.8
0.7
0.6
0.5
0.4
0.3
0
2
4
6
8
10
Layer

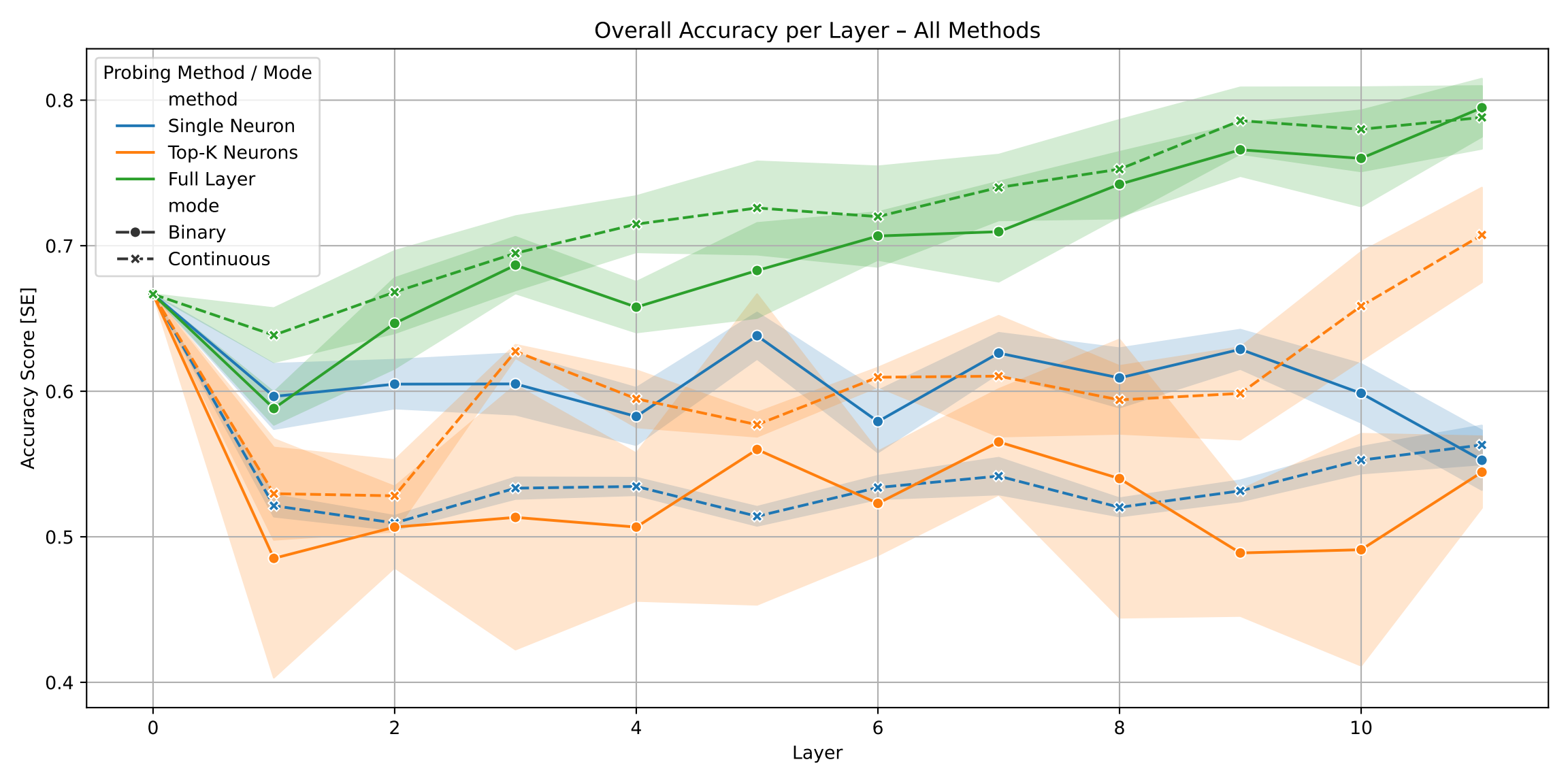
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	11.0	11.0
Full Layer	f1_max	0.8323	0.8384
Full Layer	f1_mean	0.6904	0.7153
Full Layer	f1_std	0.0788	0.0787
Single Neuron	f1_best_layer	0.0	11.0
Single Neuron	f1_max	0.6725	0.7557
Single Neuron	f1_mean	0.5069	0.543
Single Neuron	f1_std	0.0778	0.0457
Top-K Neurons	f1_best_layer	7.0	11.0
Top-K Neurons	f1_max	0.646	0.7756
Top-K Neurons	f1_mean	0.4924	0.607
Top-K Neurons	f1_std	0.113	0.0622


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
0.50
0
2
4
6
8
10
Layer


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
Layer
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
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
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
Accuracy Score [SE]
0.8
0.7
0.6
0.5
0.4
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	0.8333	0.8378
Full Layer	accuracy_mean	0.7007	0.723
Full Layer	accuracy_std	0.0659	0.0604
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.6822	0.7533
Single Neuron	accuracy_mean	0.6074	0.5436
Single Neuron	accuracy_std	0.1017	0.0592
Top-K Neurons	accuracy_best_layer	0.0	11.0
Top-K Neurons	accuracy_max	0.6689	0.7711
Top-K Neurons	accuracy_mean	0.5326	0.6085
Top-K Neurons	accuracy_std	0.1048	0.0627