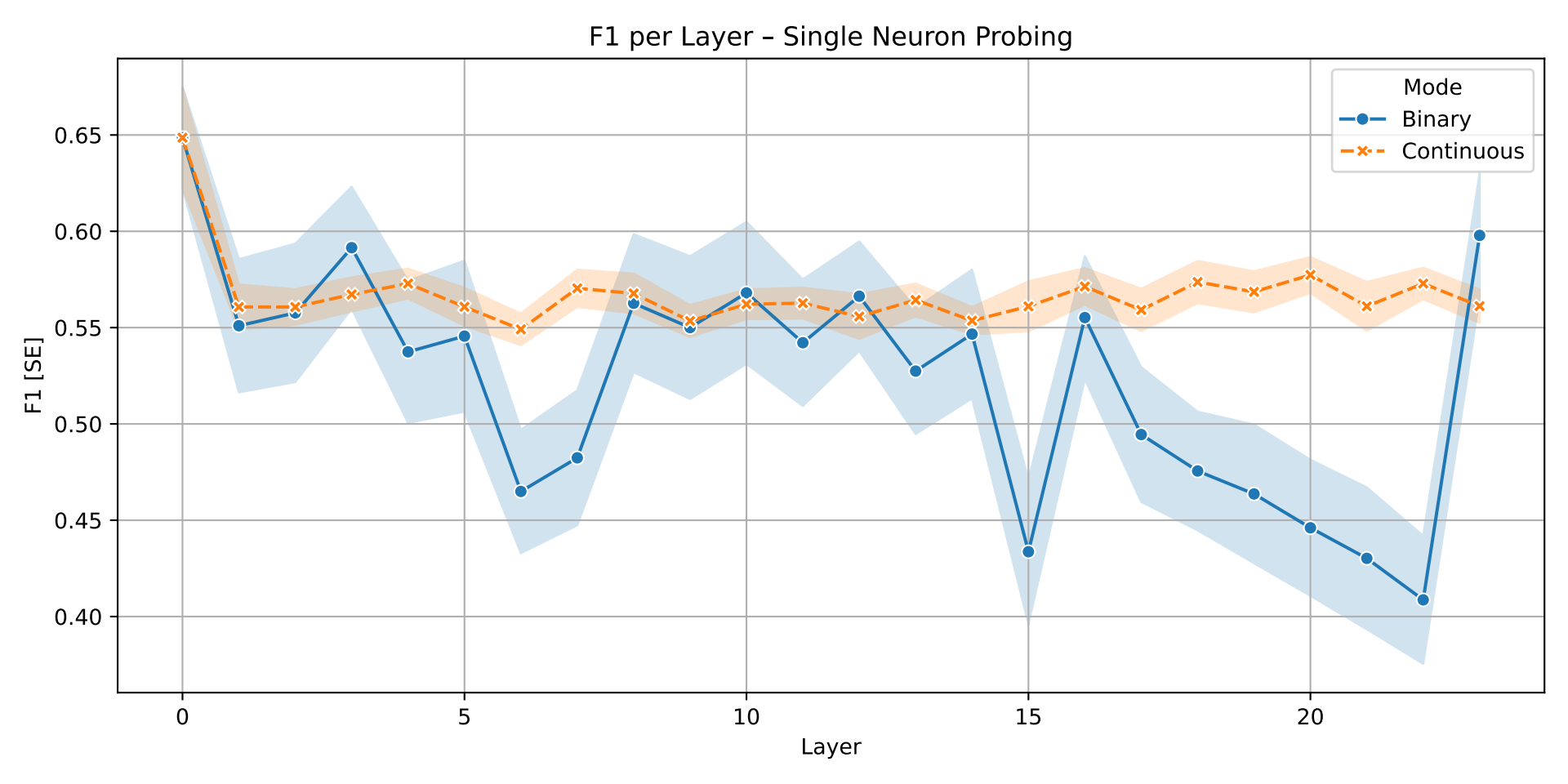
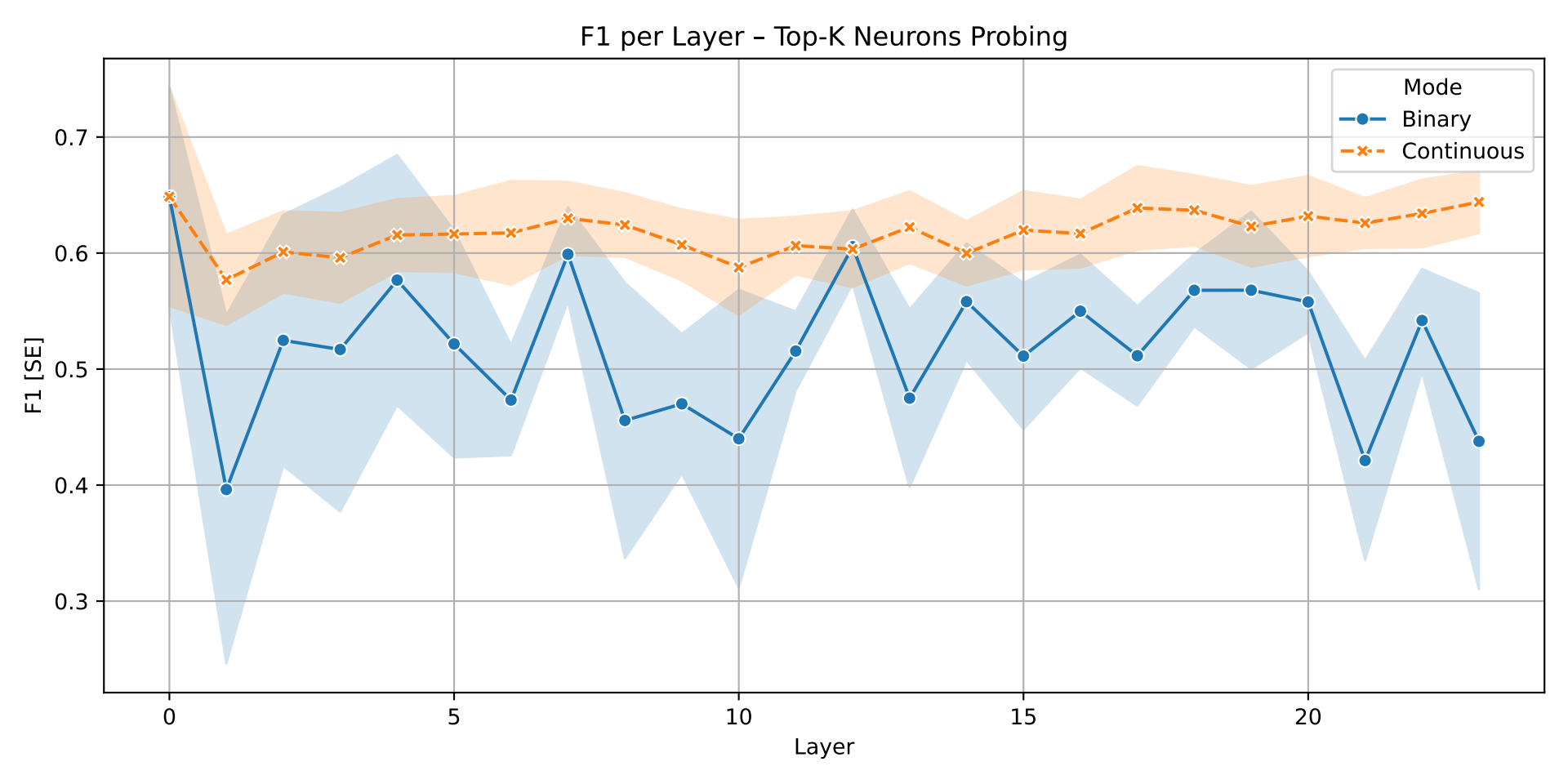
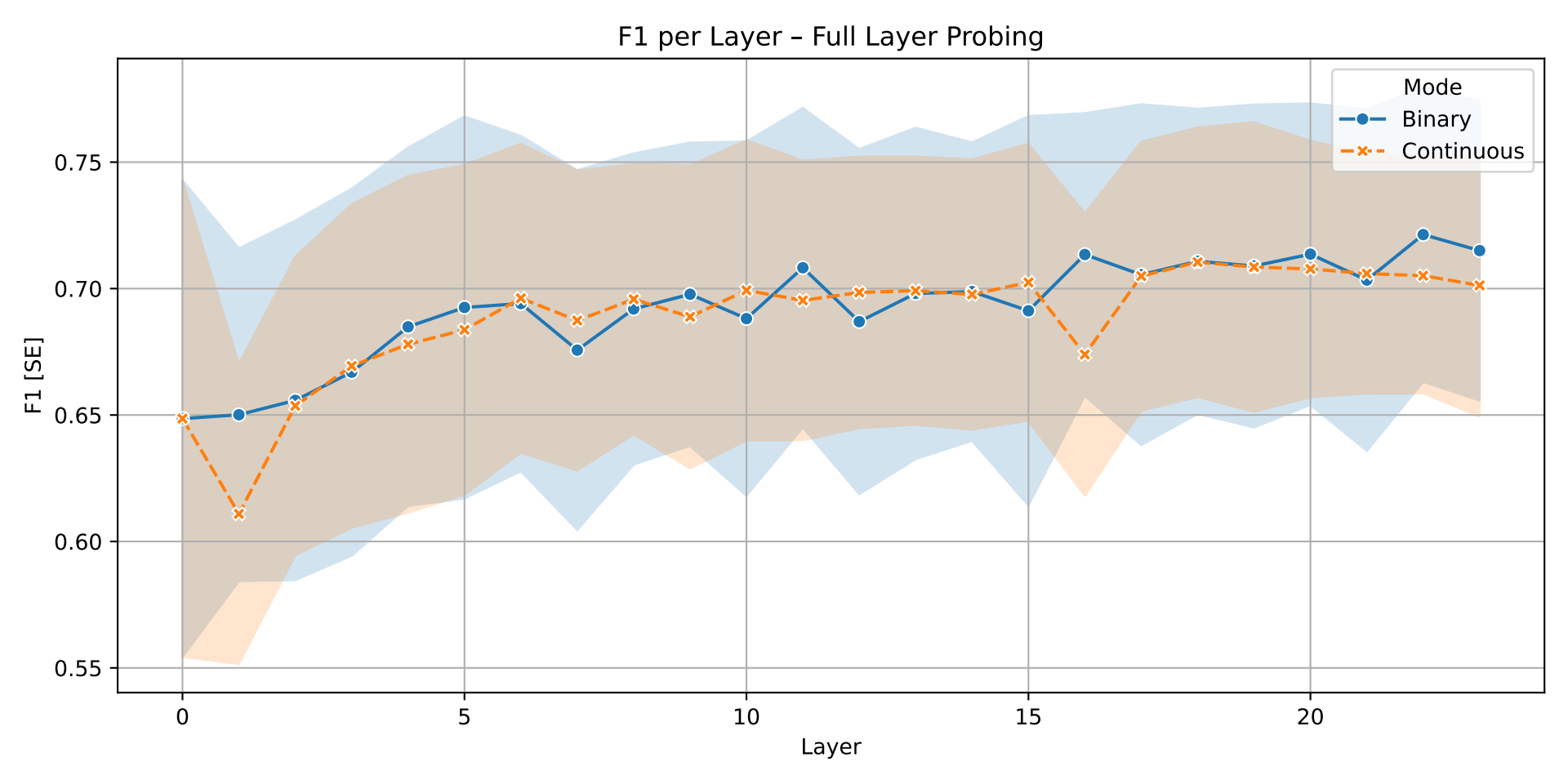
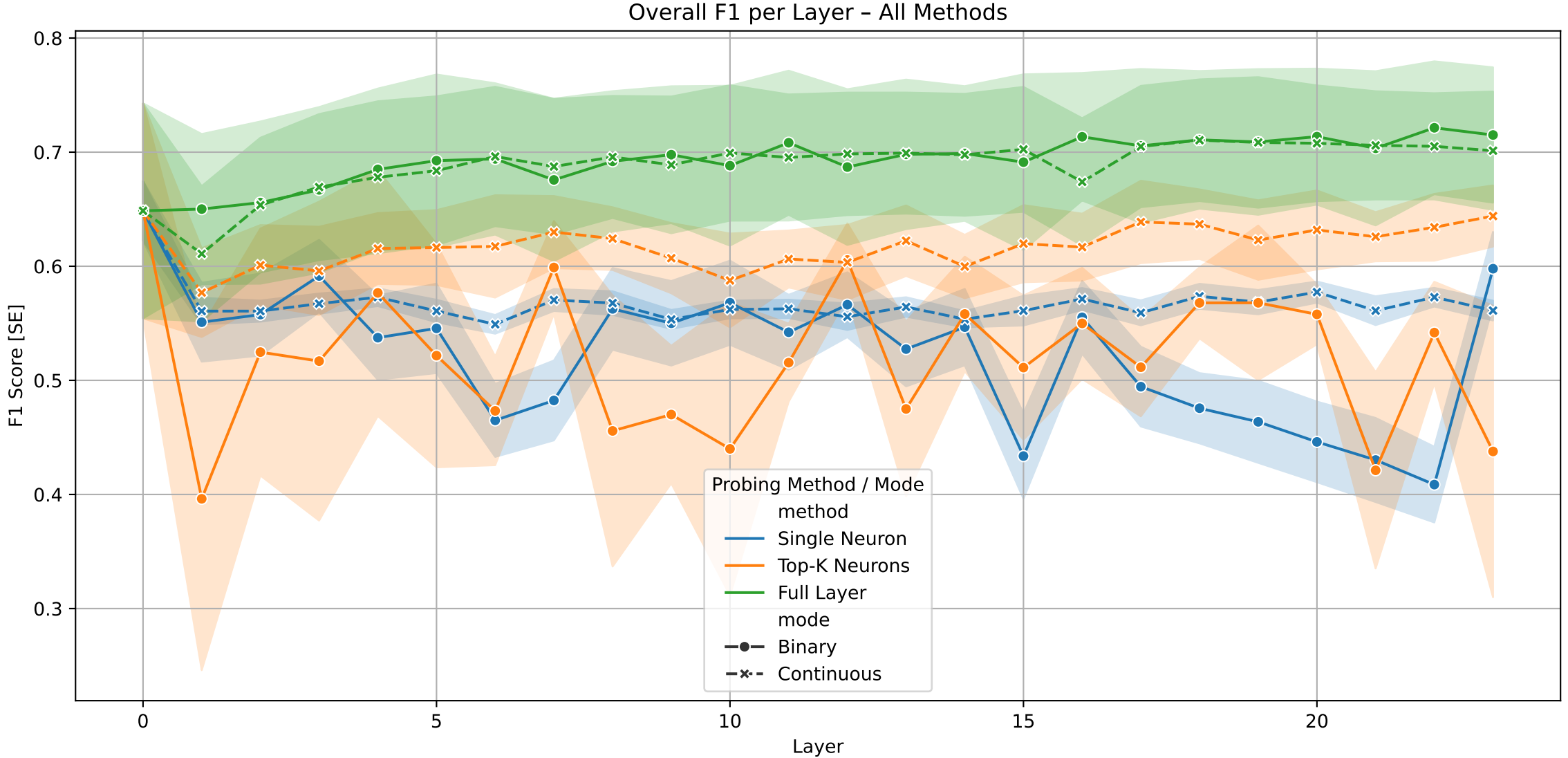

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.65
0.60
0.55
0.50
0.45
0.40
0
5
10
15
20
Layer

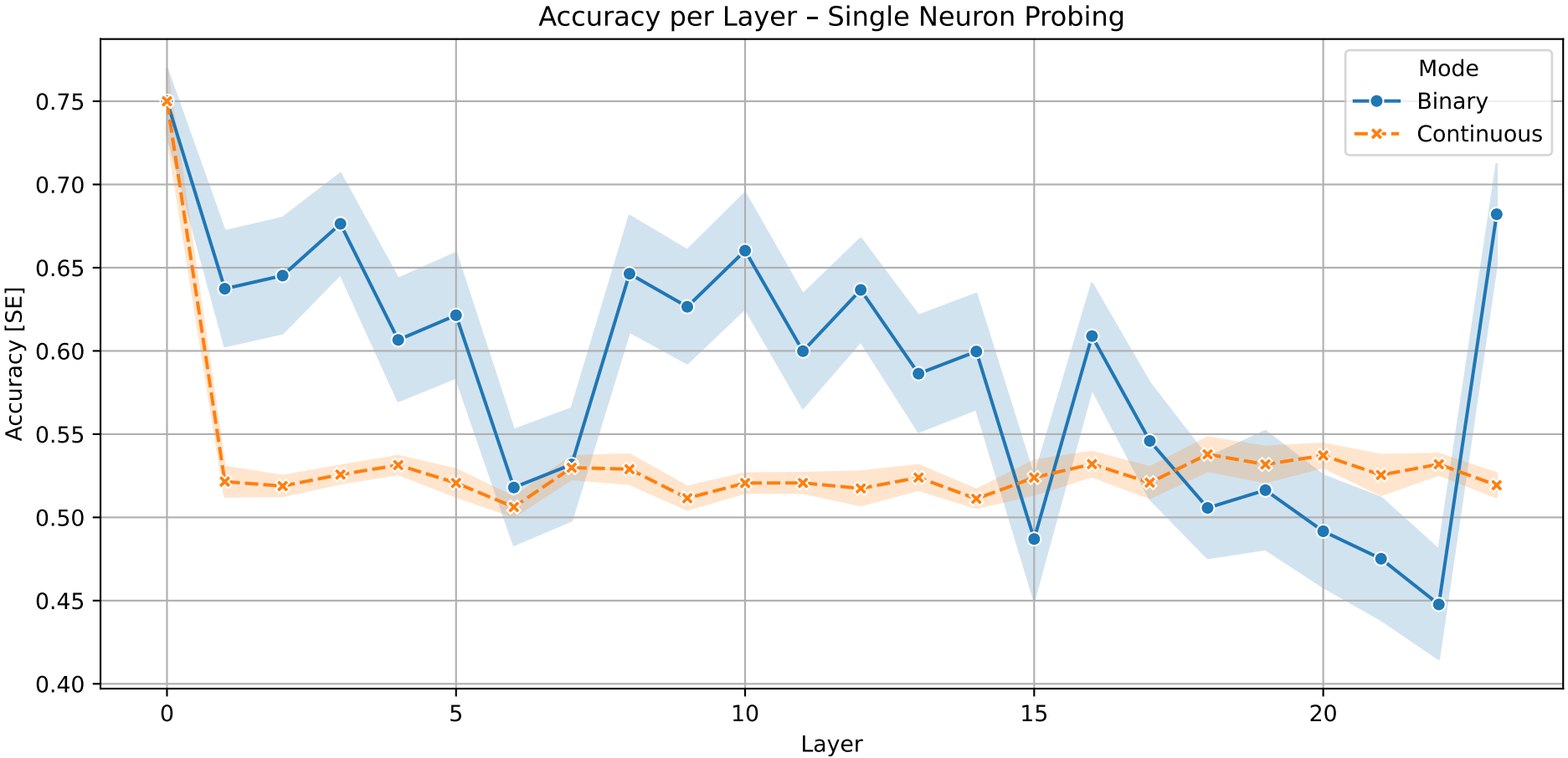

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.7
0.6
0.5
0.4
0.3
0
5
10
15
20
Layer

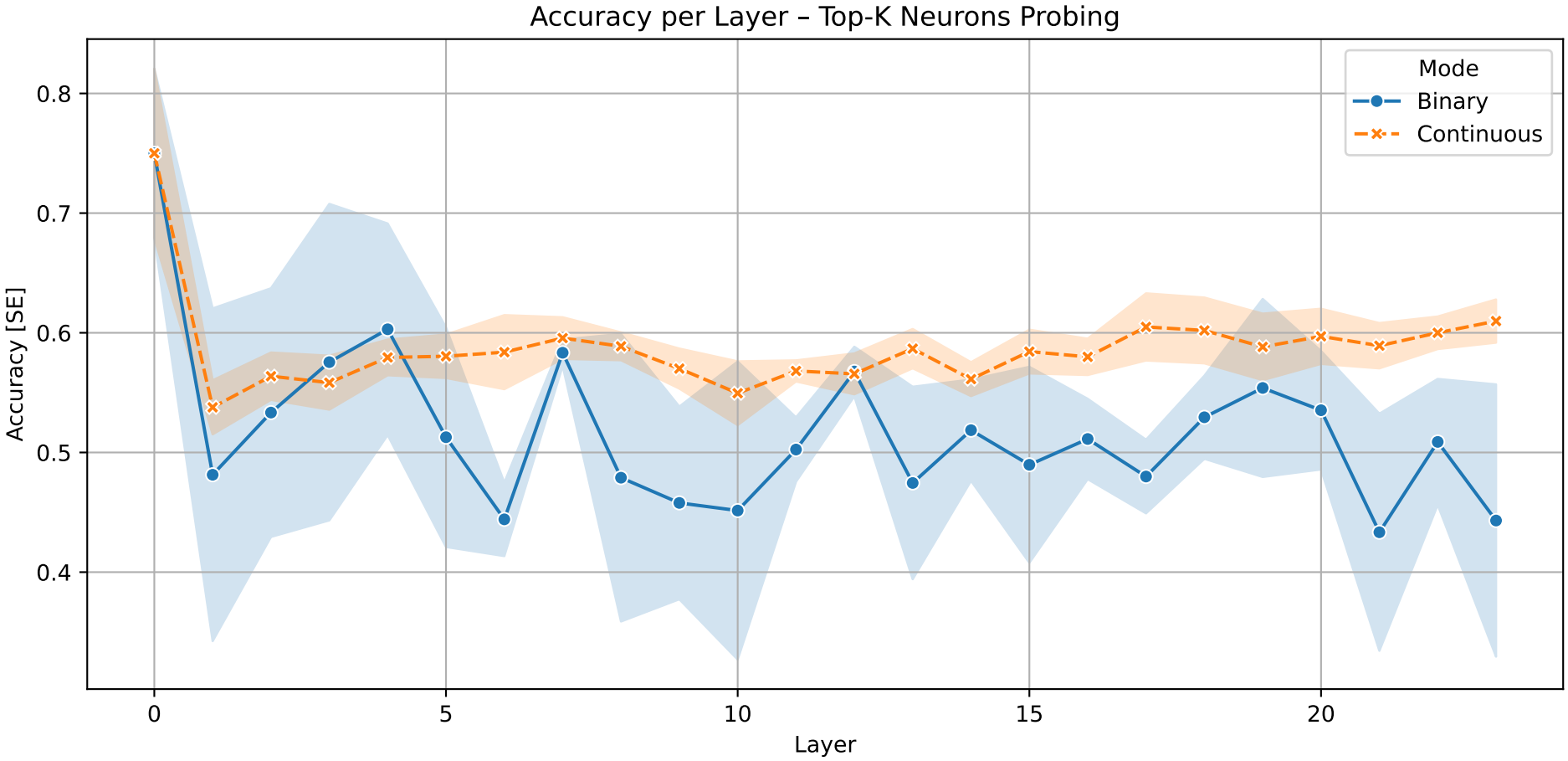

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
Layer

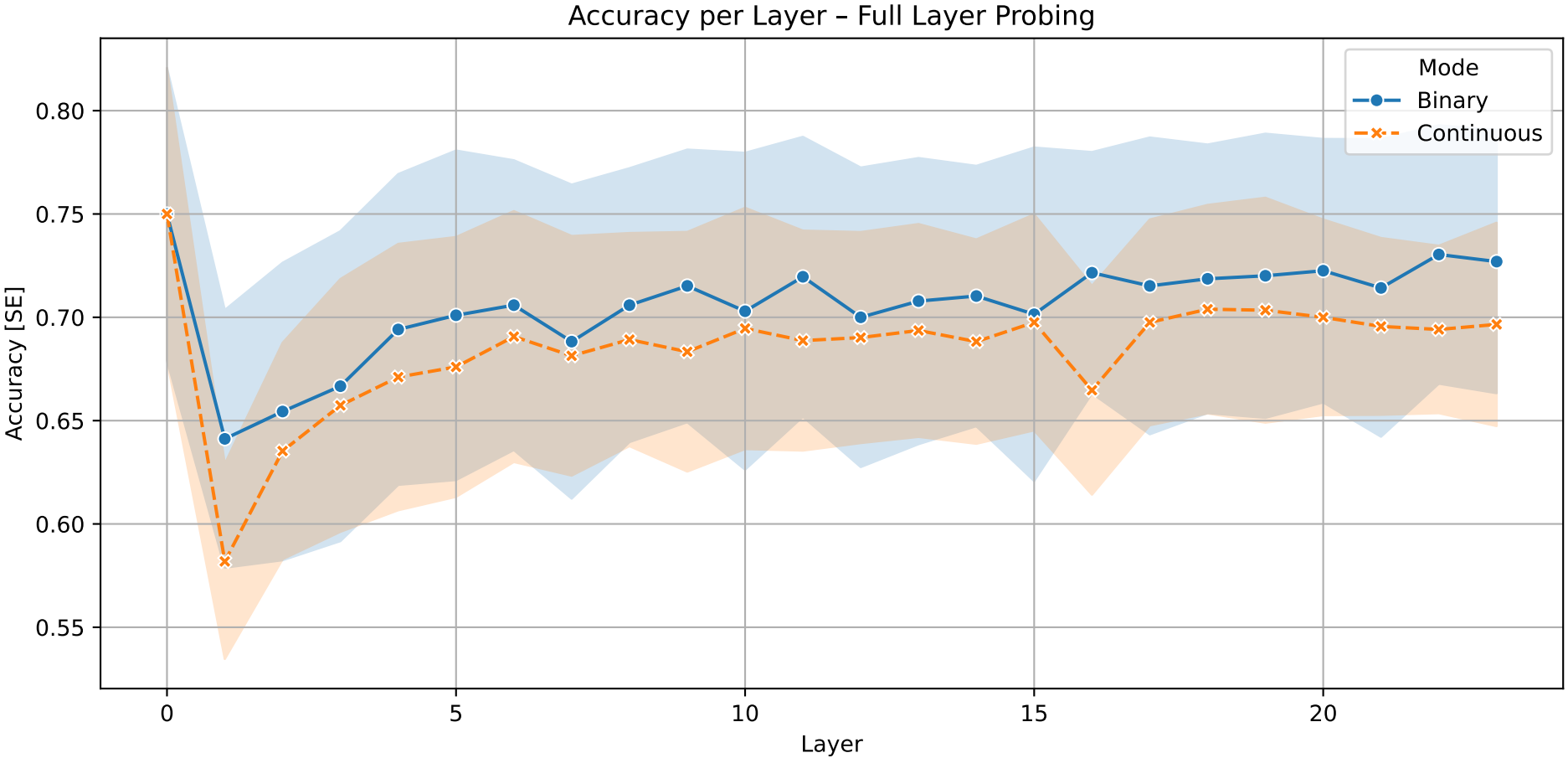

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
0.8
0.7
0.6
0.5
0.4
0.3
0
5
10
15
20
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

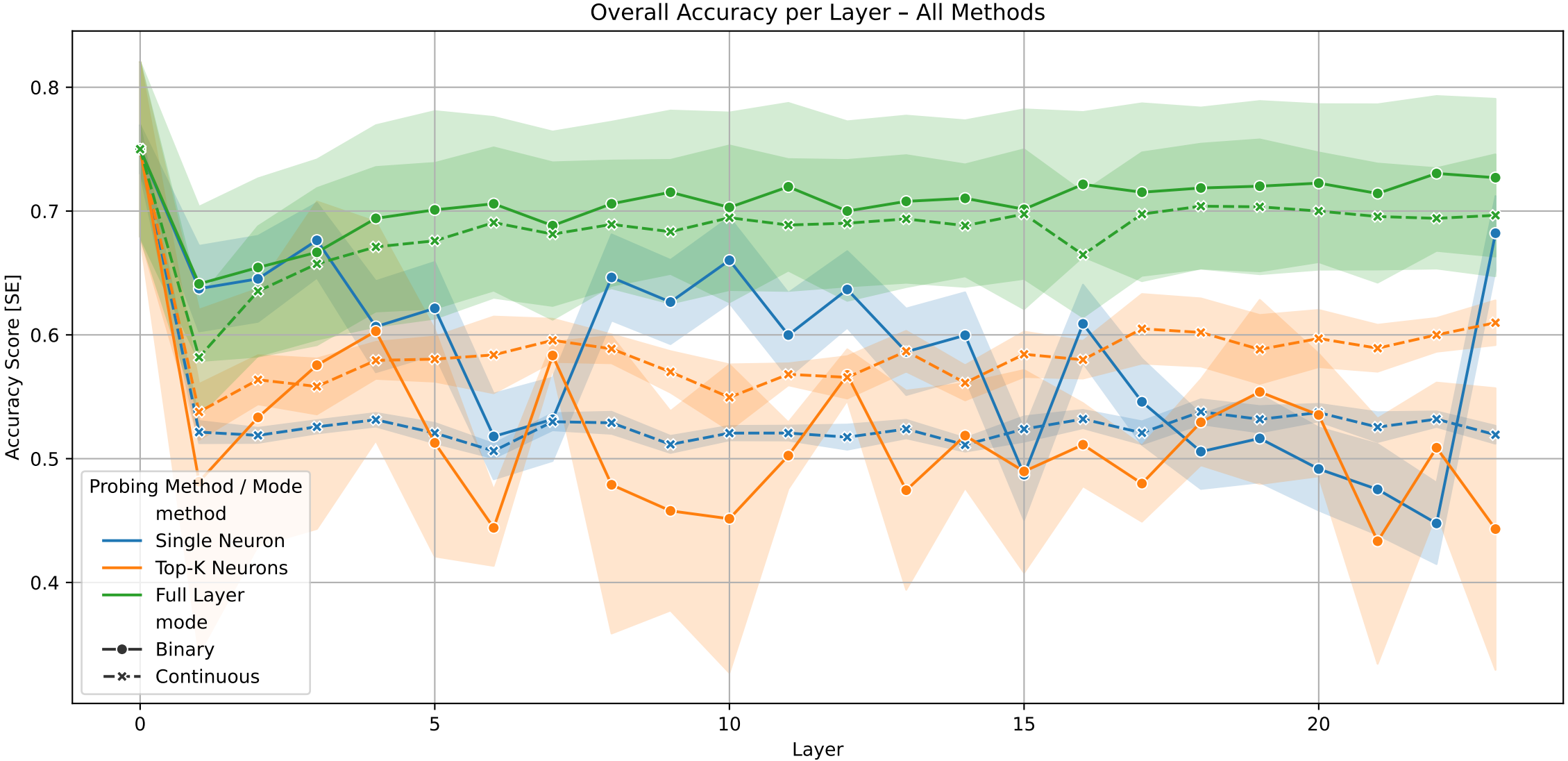
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	22.0	18.0
Full Layer	f1_max	0.8651	0.8526
Full Layer	f1_mean	0.6926	0.6884
Full Layer	f1_std	0.1189	0.1045
Single Neuron	f1_best_layer	0.0	0.0
Single Neuron	f1_max	0.8658	0.8526
Single Neuron	f1_mean	0.5228	0.5673
Single Neuron	f1_std	0.2216	0.0705
Top-K Neurons	f1_best_layer	0.0	0.0
Top-K Neurons	f1_max	0.8598	0.8526
Top-K Neurons	f1_mean	0.5185	0.6176
Top-K Neurons	f1_std	0.1603	0.0675


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0
5
10
15
20
Layer


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.8
0.7
0.6
0.5
0.4
0
5
10
15
20
Layer


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
Layer


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
Layer
0.8
0.7
0.6
0.5
0.4
0
5
10
15
20
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	0.7056	0.6844
Full Layer	accuracy_std	0.1238	0.0987
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.9	0.9
Single Neuron	accuracy_mean	0.5876	0.5333
Single Neuron	accuracy_std	0.2229	0.0713
Top-K Neurons	accuracy_best_layer	0.0	0.0
Top-K Neurons	accuracy_max	0.9	0.9
Top-K Neurons	accuracy_mean	0.5175	0.5873
Top-K Neurons	accuracy_std	0.1568	0.0575