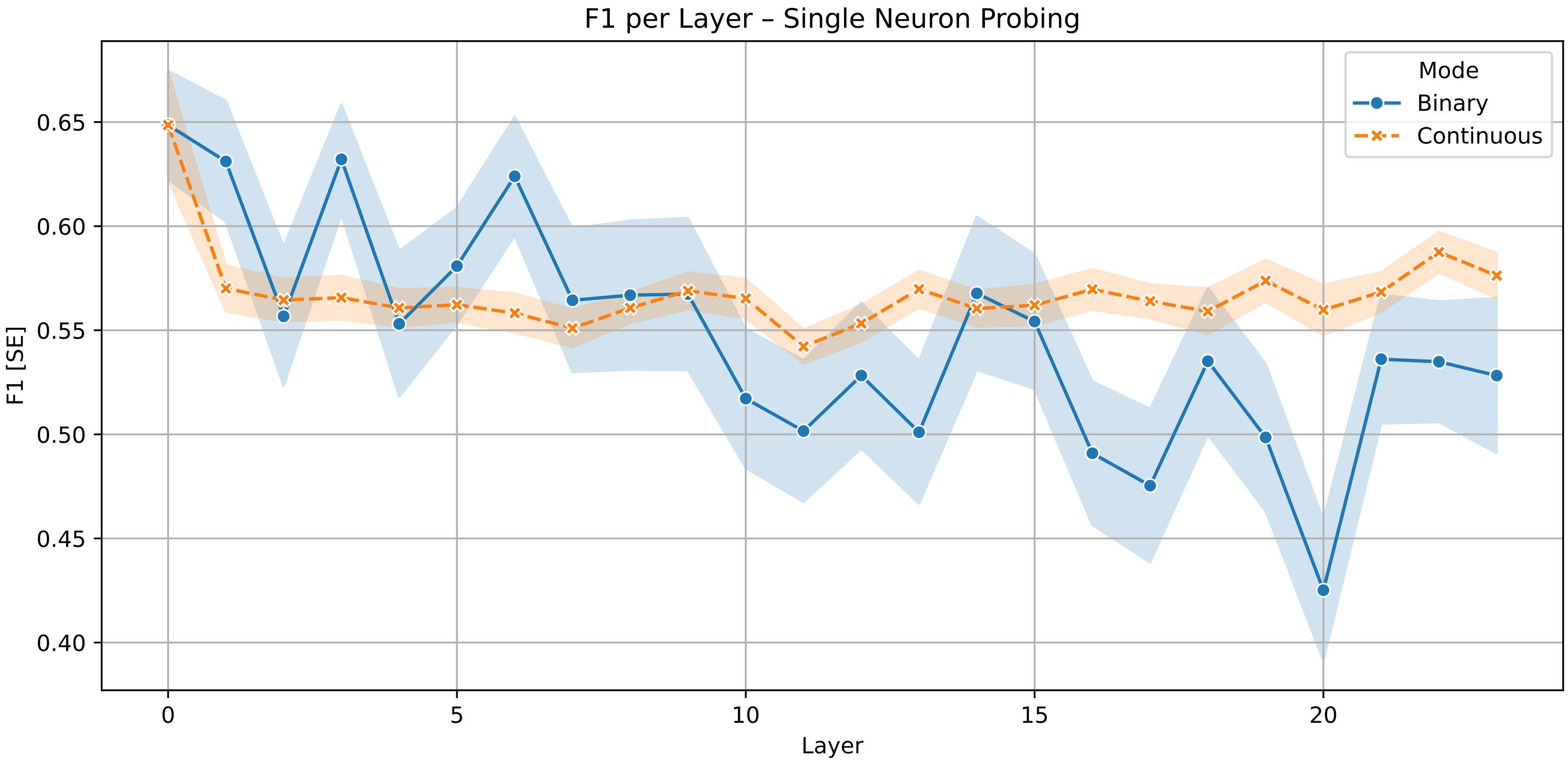
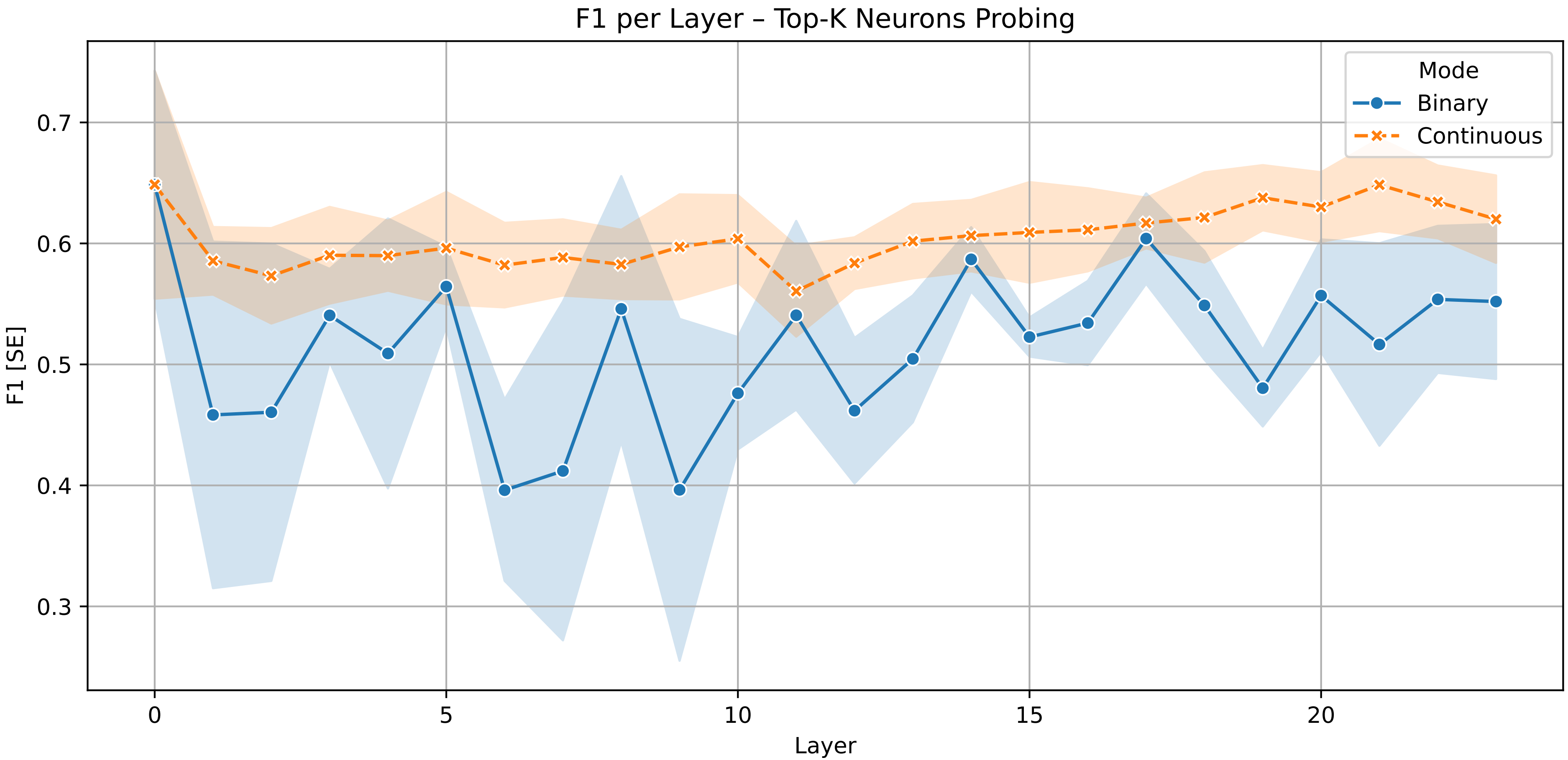
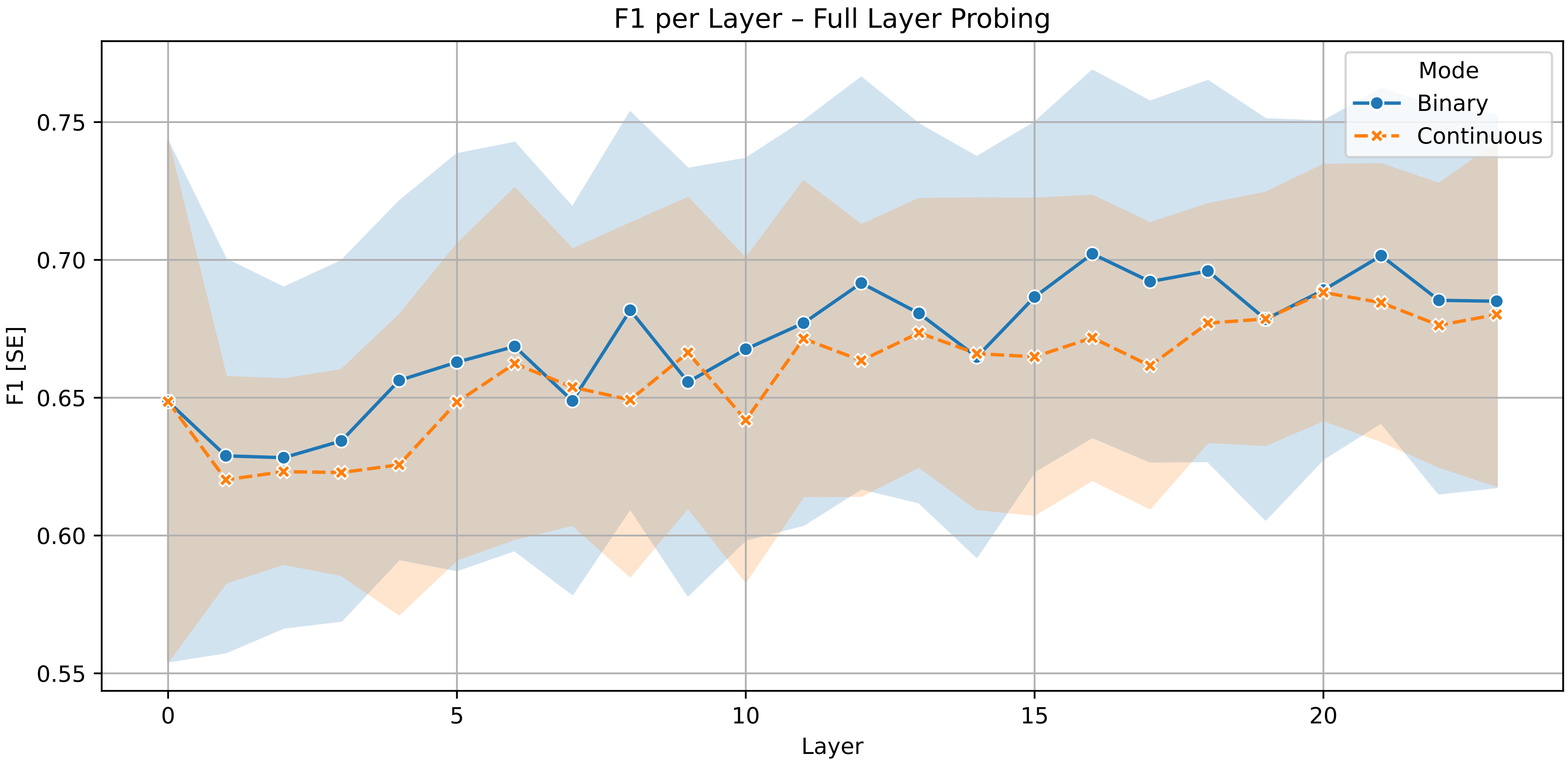
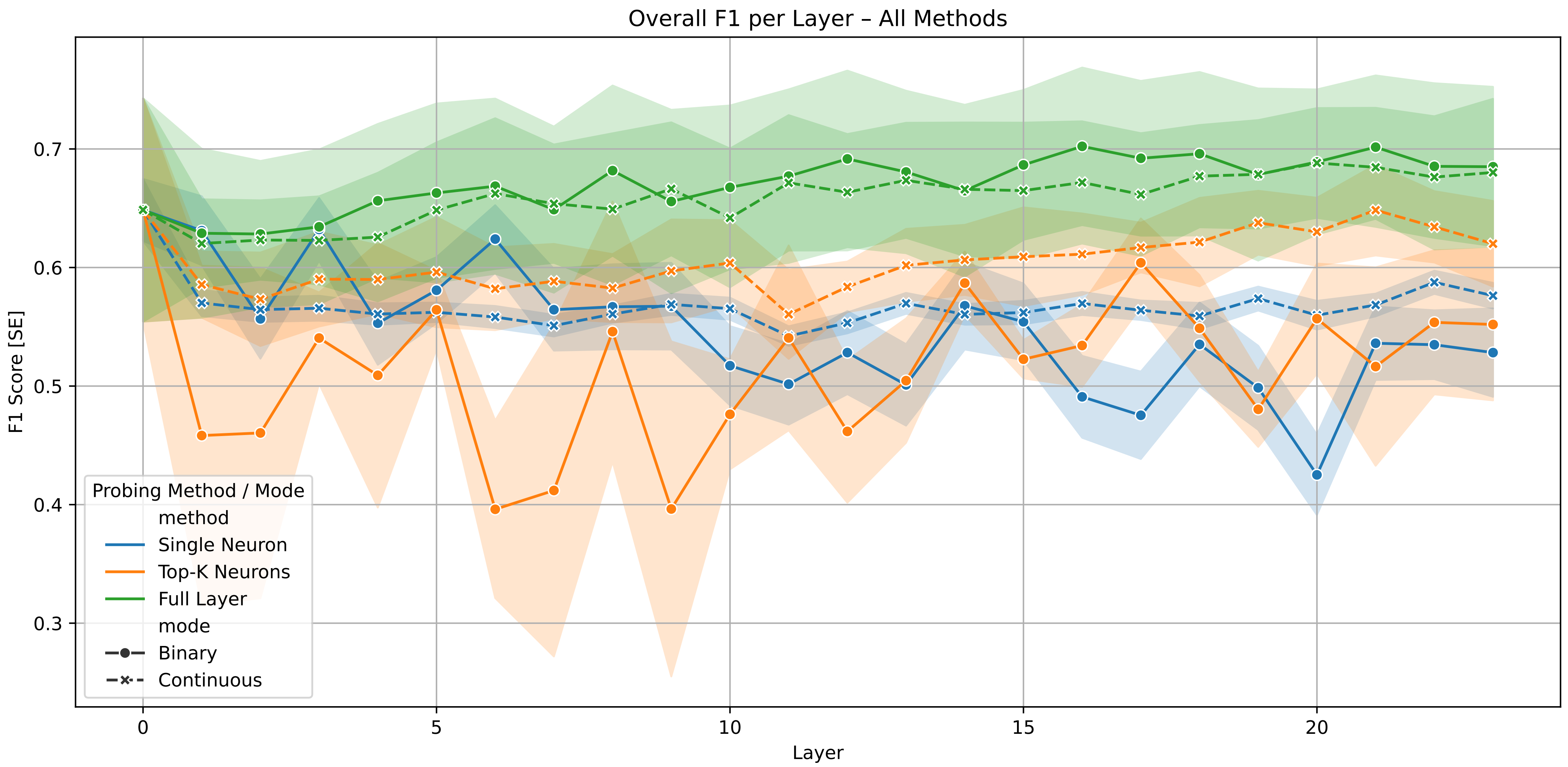
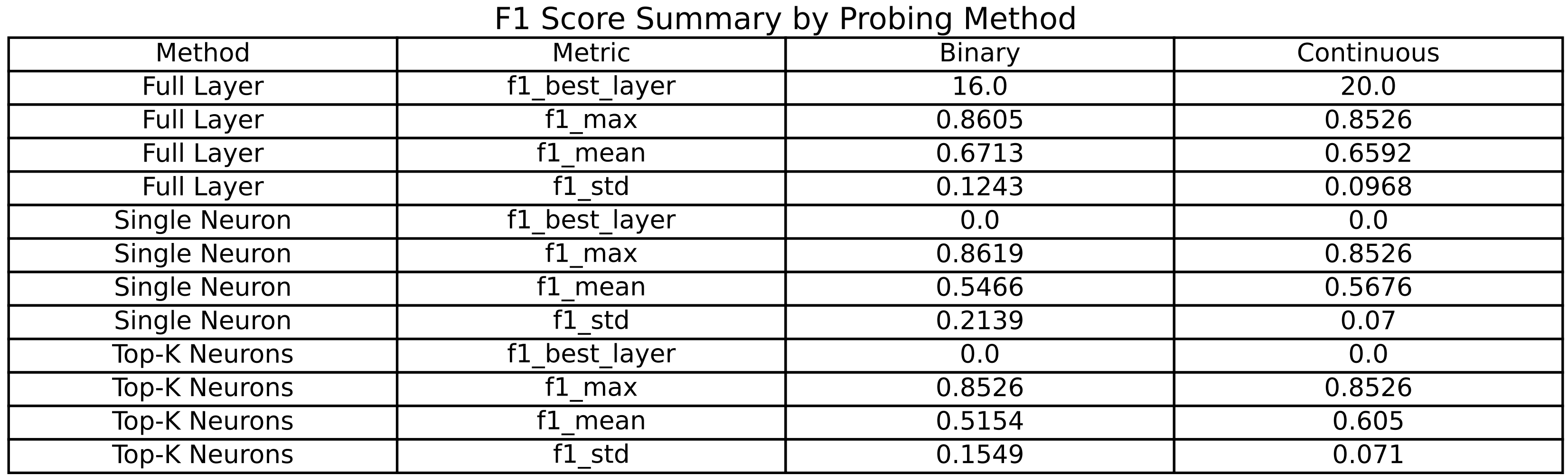

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.65
0.60
0.55
0.50
0.45
0.40
0
5
10
15
20
Layer


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

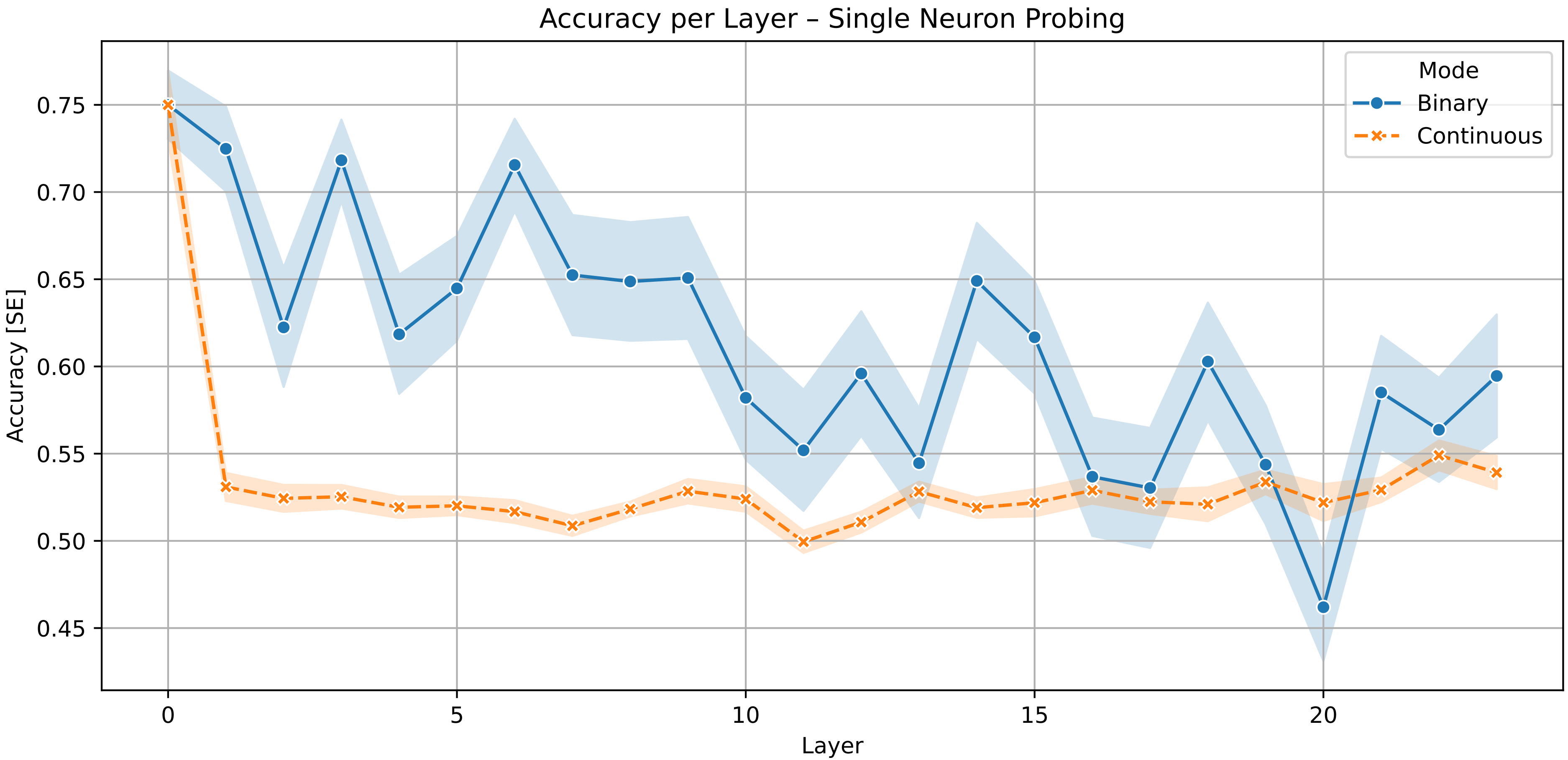

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.75
0.70
0.65
0.60
0.55
Layer
0
5
10
15
20

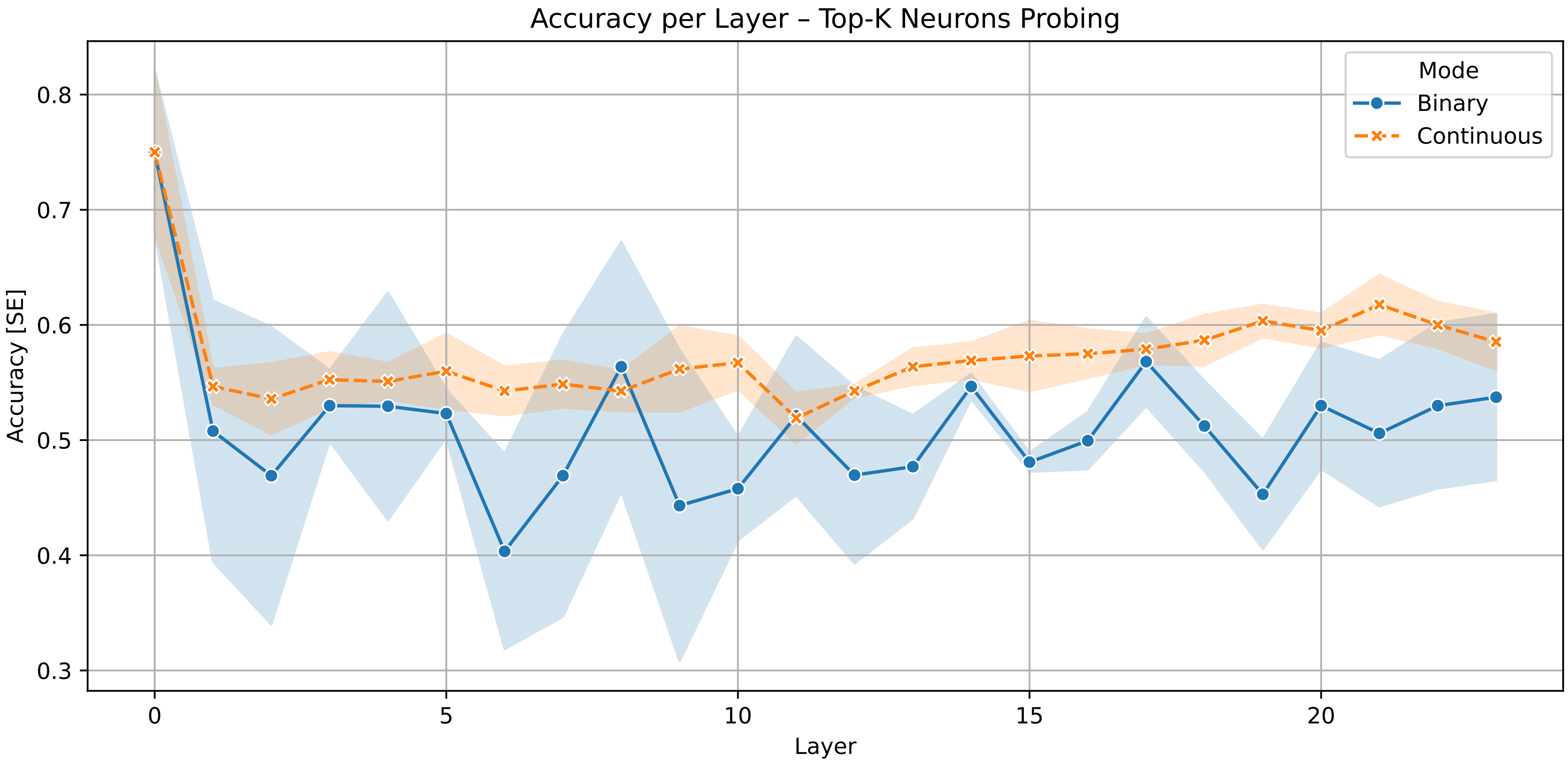

Overall F1 per Layer – All Methods
F1 Score [SE]
0.3
0.4
0.5
0.6
0.7
Layer
0
5
10
15
20
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

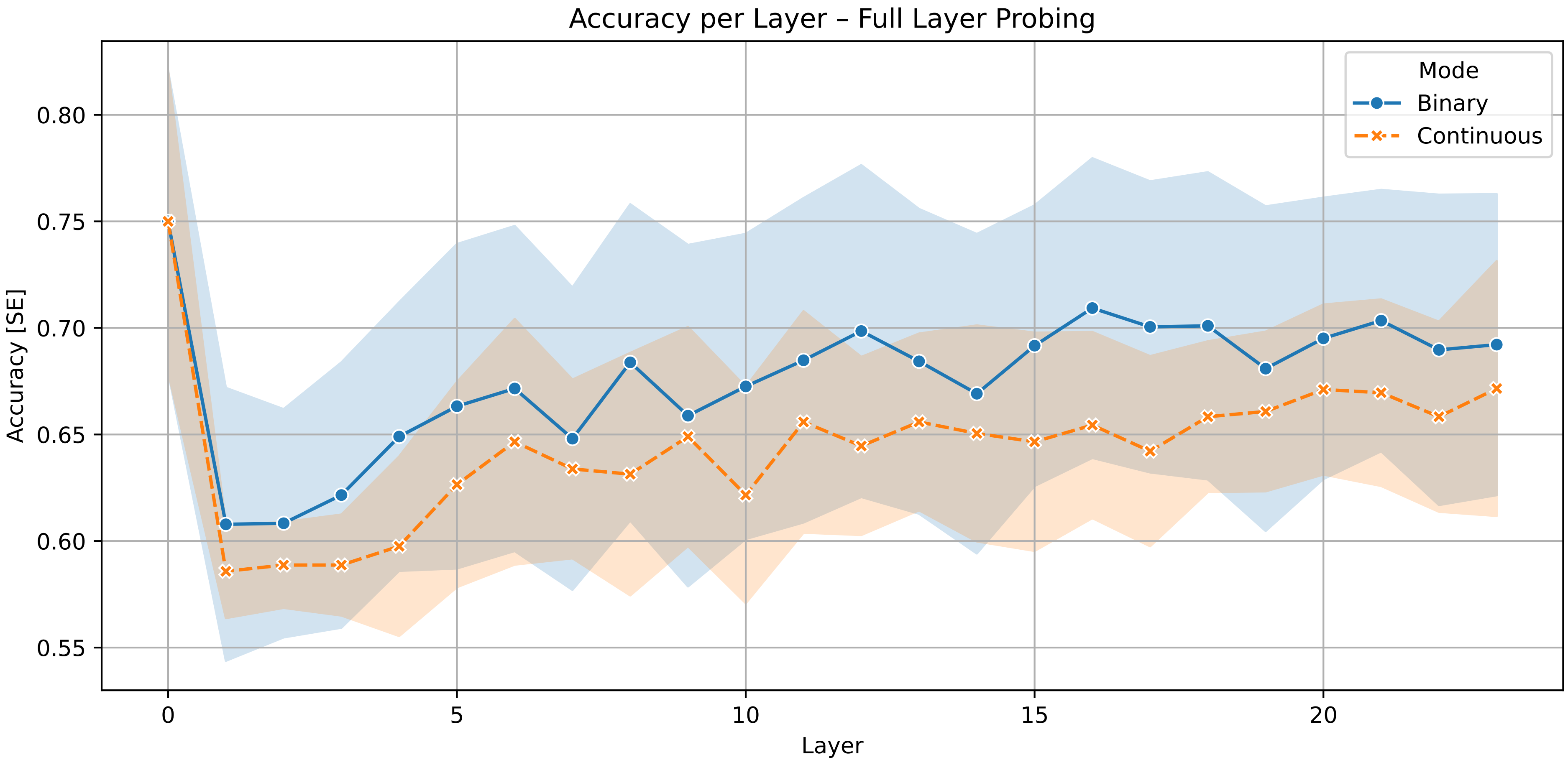


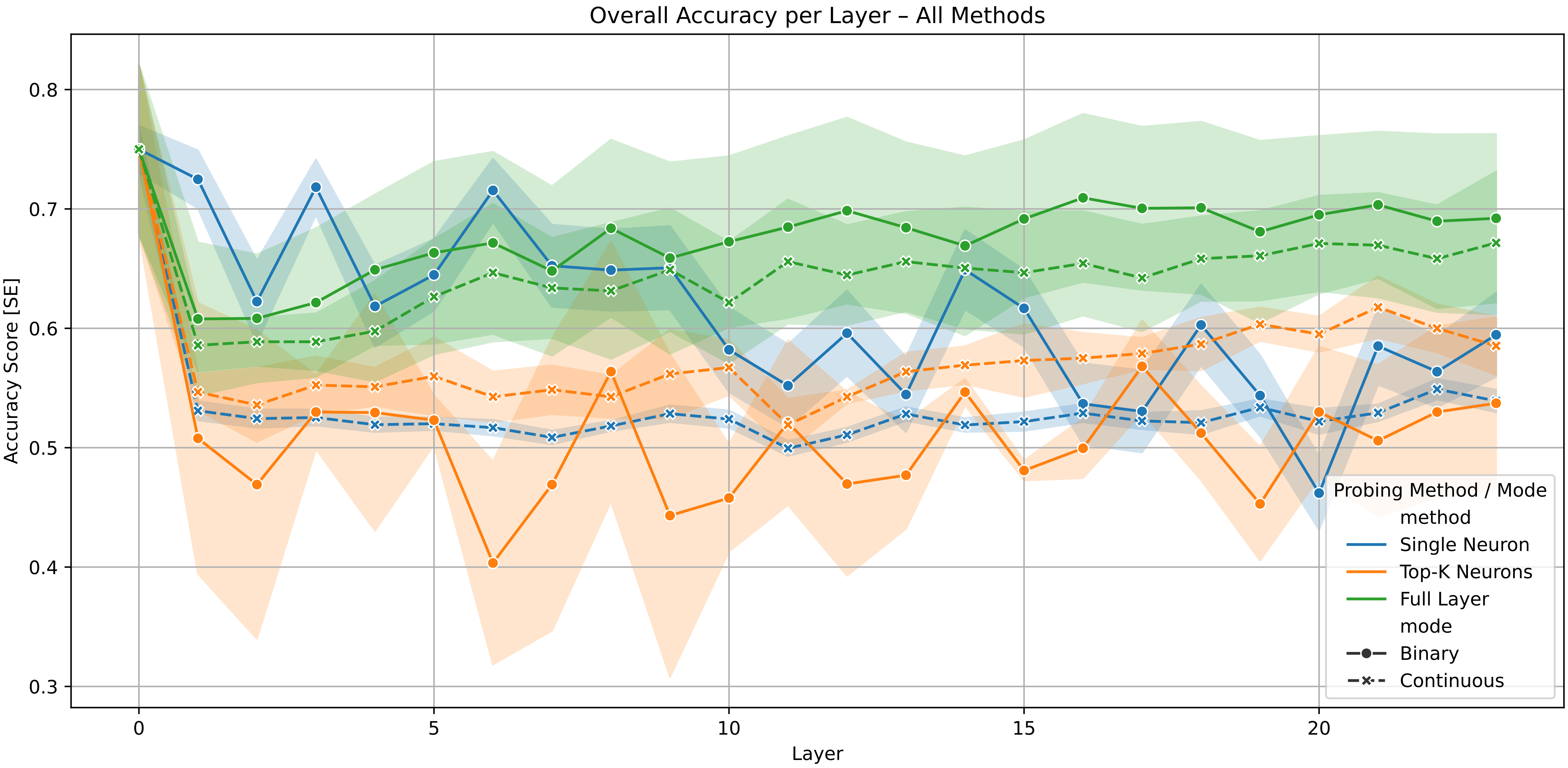
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	16.0	20.0
Full Layer	f1_max	0.8605	0.8526
Full Layer	f1_mean	0.6713	0.6592
Full Layer	f1_std	0.1243	0.0968
Single Neuron	f1_best_layer	0.0	0.0
Single Neuron	f1_max	0.8619	0.8526
Single Neuron	f1_mean	0.5466	0.5676
Single Neuron	f1_std	0.2139	0.07
Top-K Neurons	f1_best_layer	0.0	0.0
Top-K Neurons	f1_max	0.8526	0.8526
Top-K Neurons	f1_mean	0.5154	0.605
Top-K Neurons	f1_std	0.1549	0.071

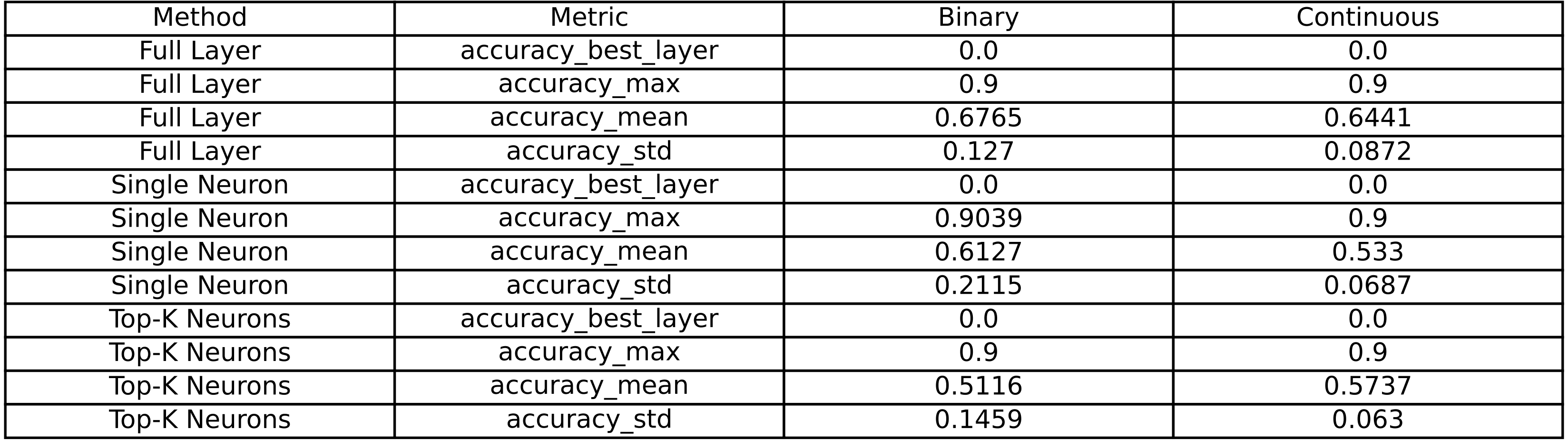

Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
0.75
0.70
0.65
0.60
0.55
0.50
0.45
Layer
0
5
10
15
20
Mode
Binary
Continuous


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
0.3
0.4
0.5
0.6
0.7
0.8
Layer
0
5
10
15
20
Mode
Binary
Continuous


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


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
0.3
0.4
0.5
0.6
0.7
0.8
Layer
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	0.6765	0.6441
Full Layer	accuracy_std	0.127	0.0872
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.9039	0.9
Single Neuron	accuracy_mean	0.6127	0.533
Single Neuron	accuracy_std	0.2115	0.0687
Top-K Neurons	accuracy_best_layer	0.0	0.0
Top-K Neurons	accuracy_max	0.9	0.9
Top-K Neurons	accuracy_mean	0.5116	0.5737
Top-K Neurons	accuracy_std	0.1459	0.063