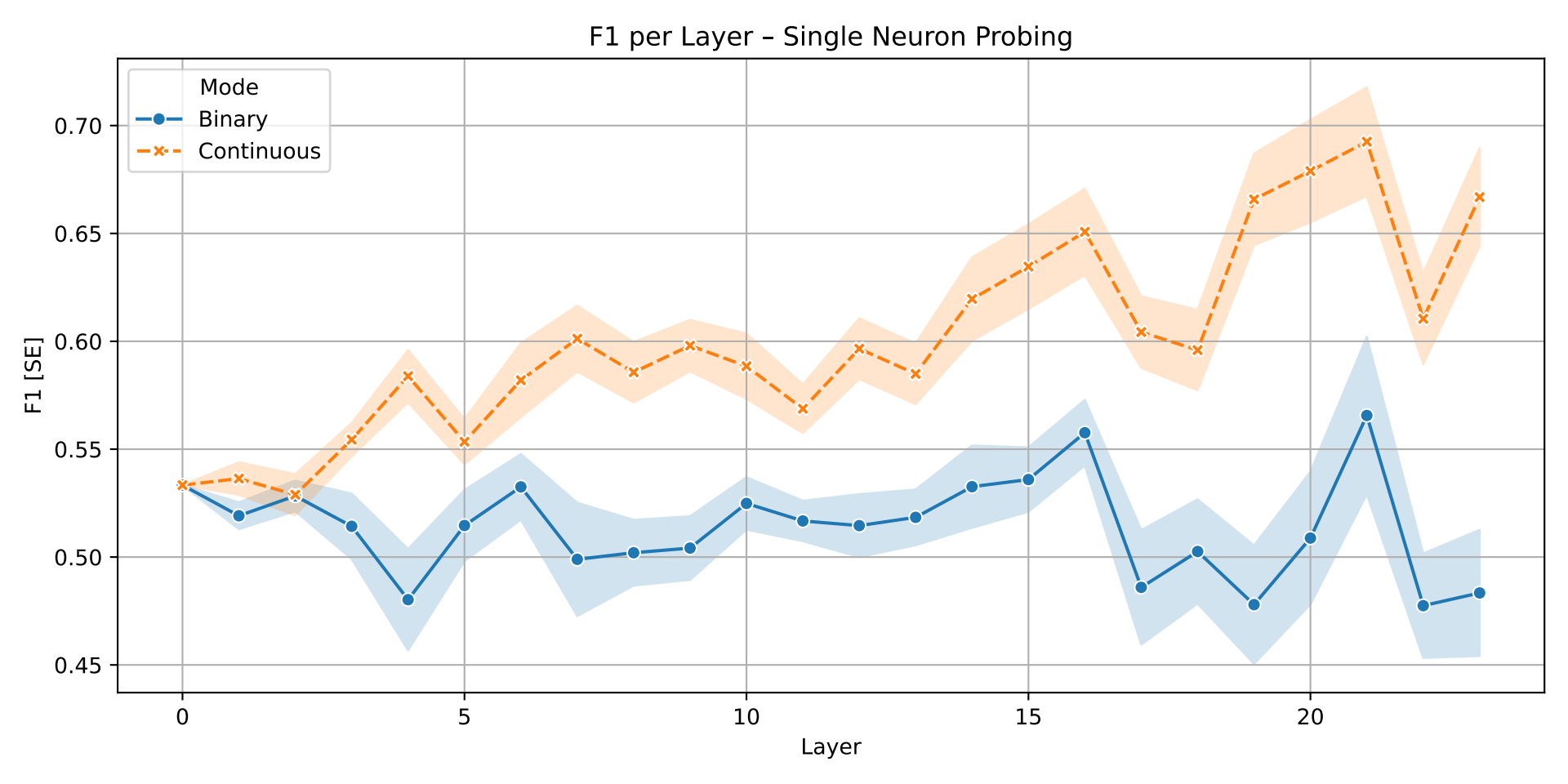
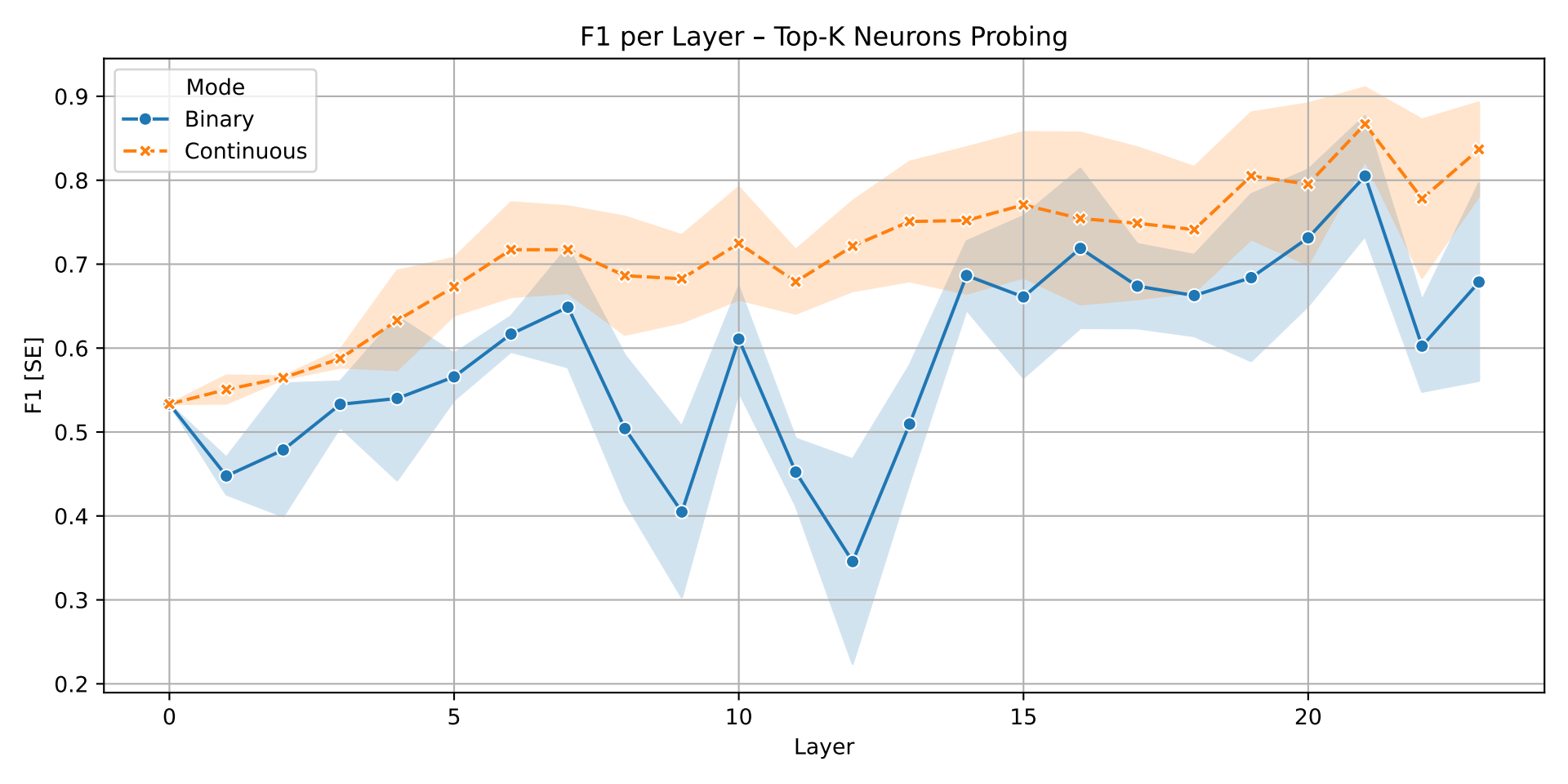
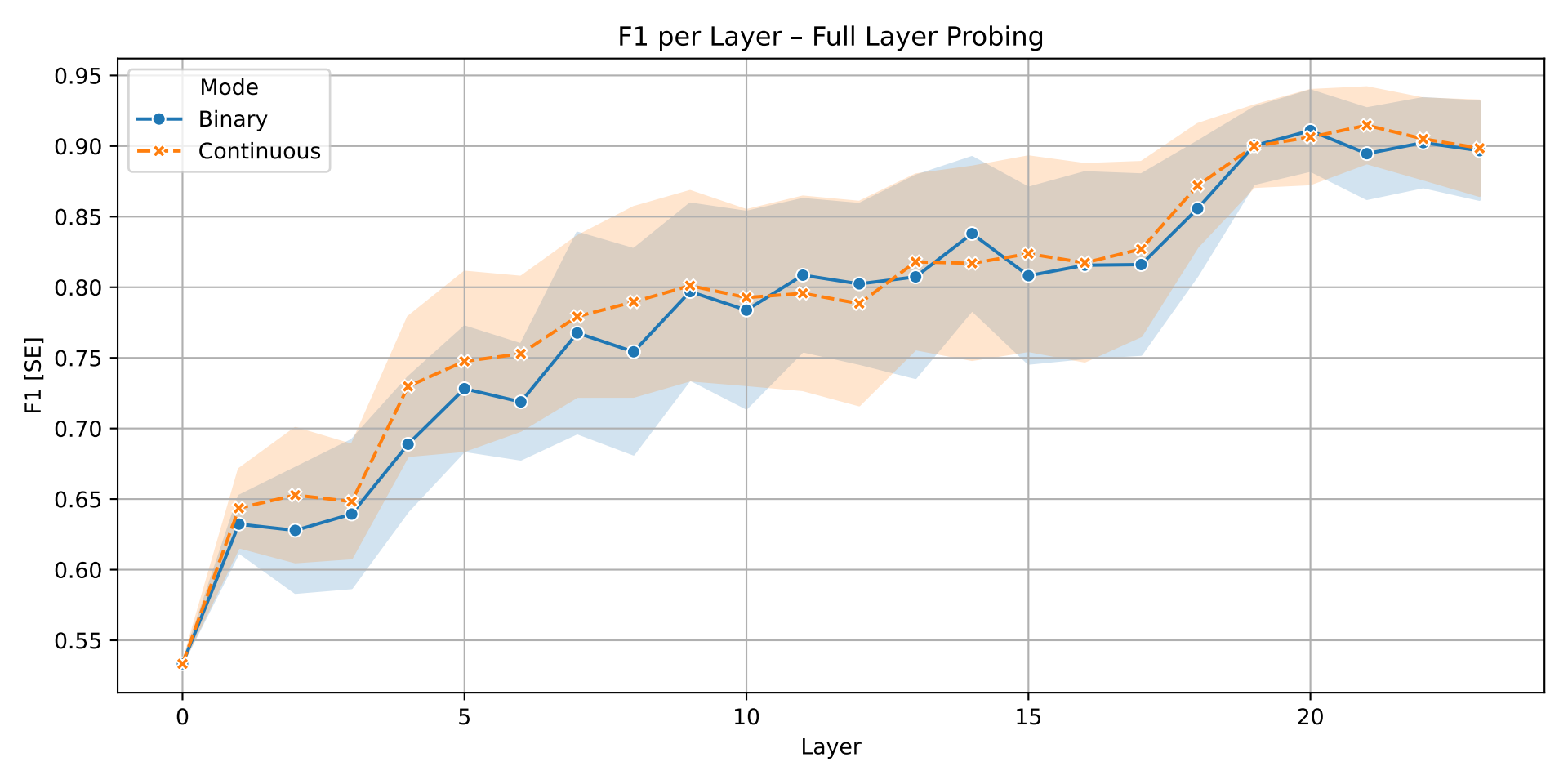
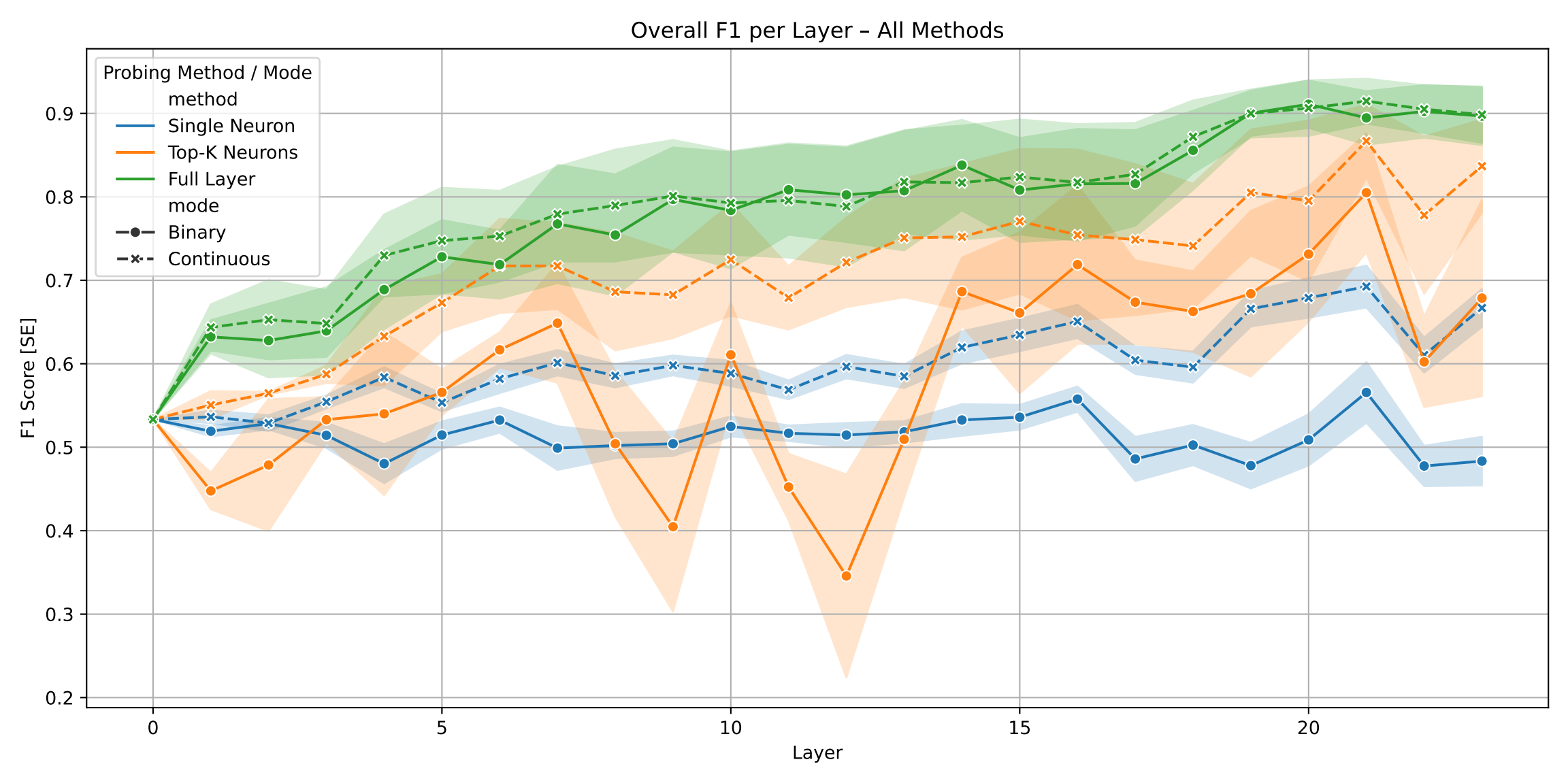

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.70
0.65
0.60
0.55
0.50
0.45
0
5
10
15
20
Layer

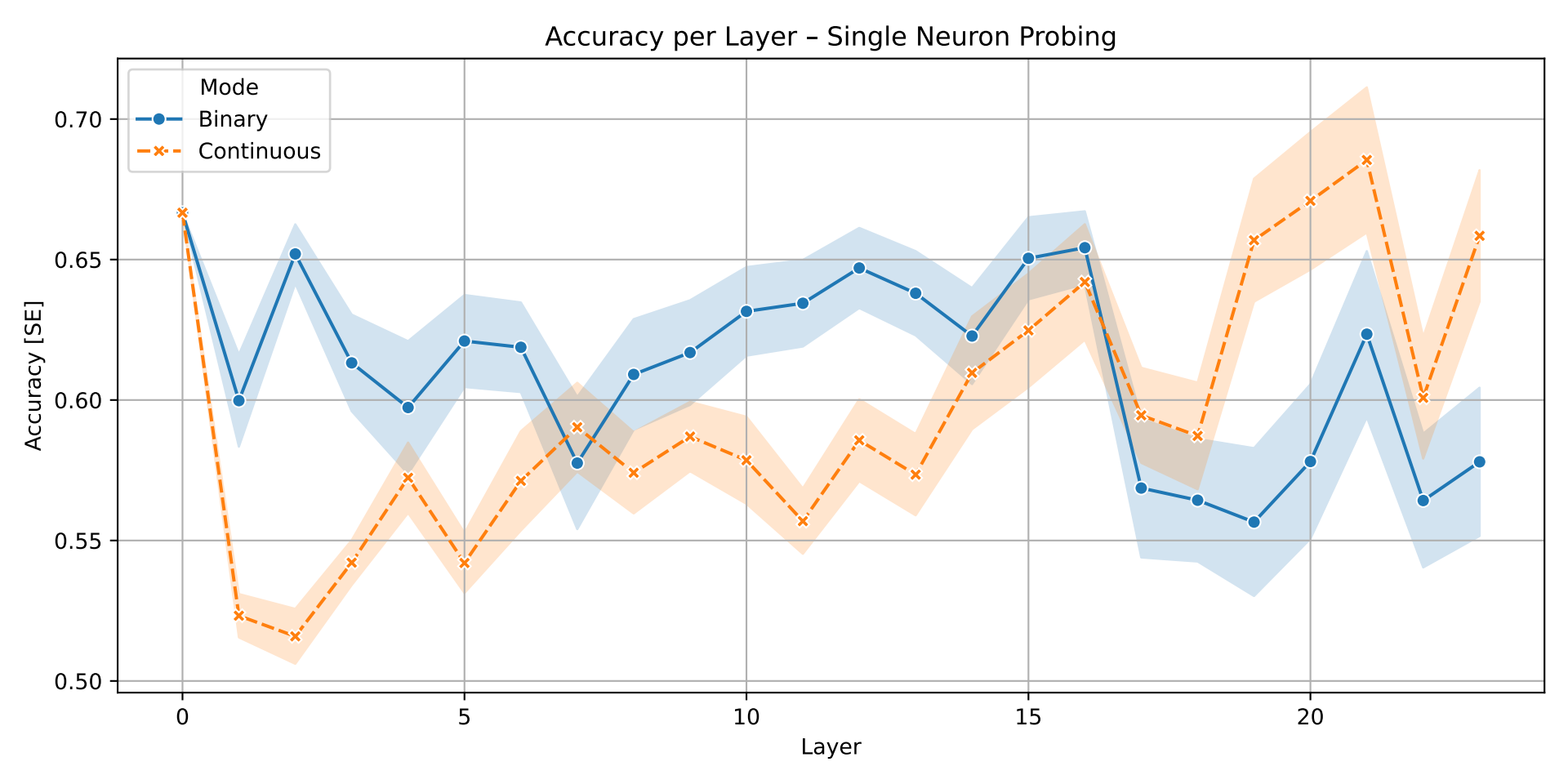

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
Layer
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
5
10
15
20

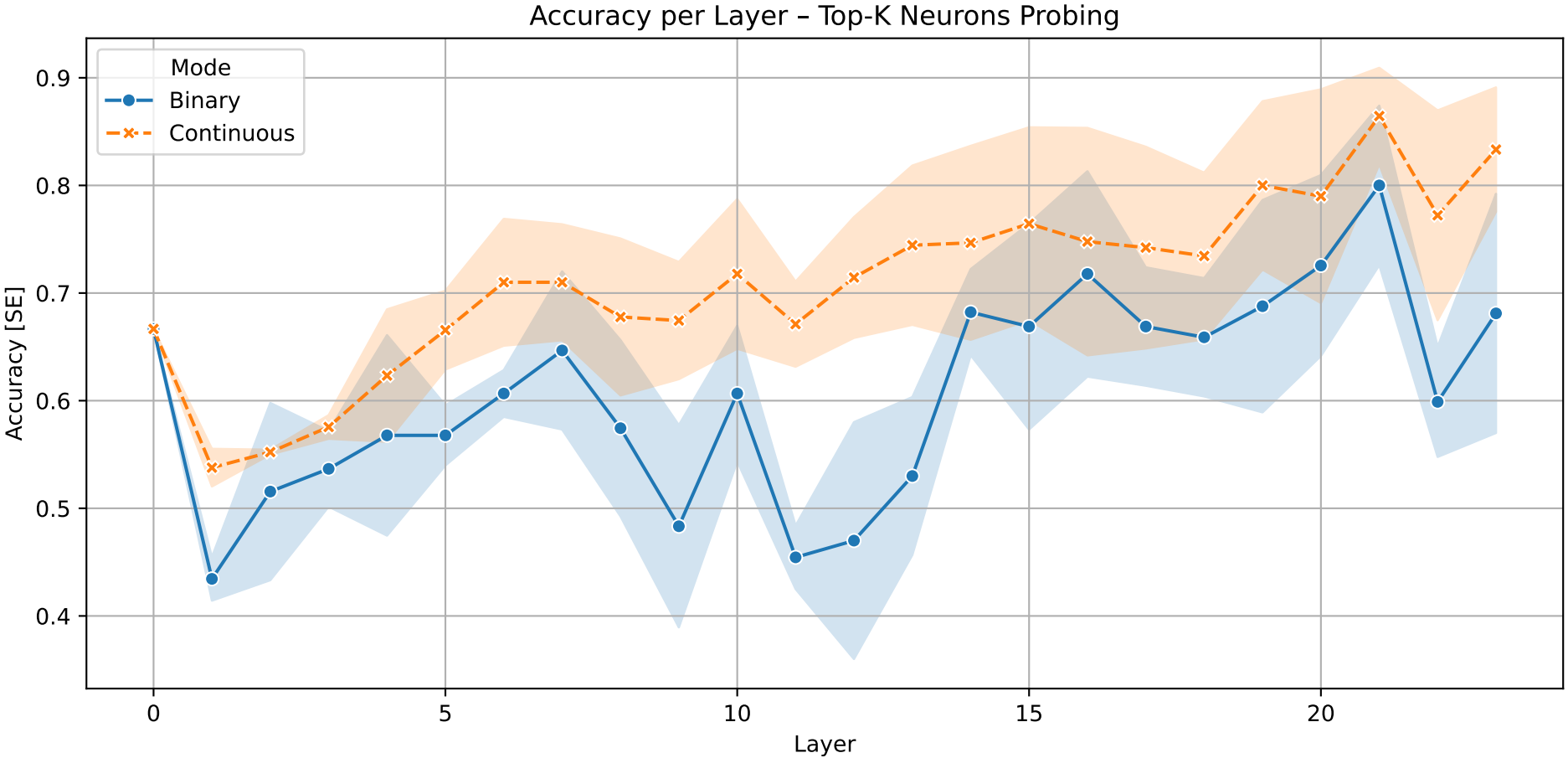

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
Layer
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20

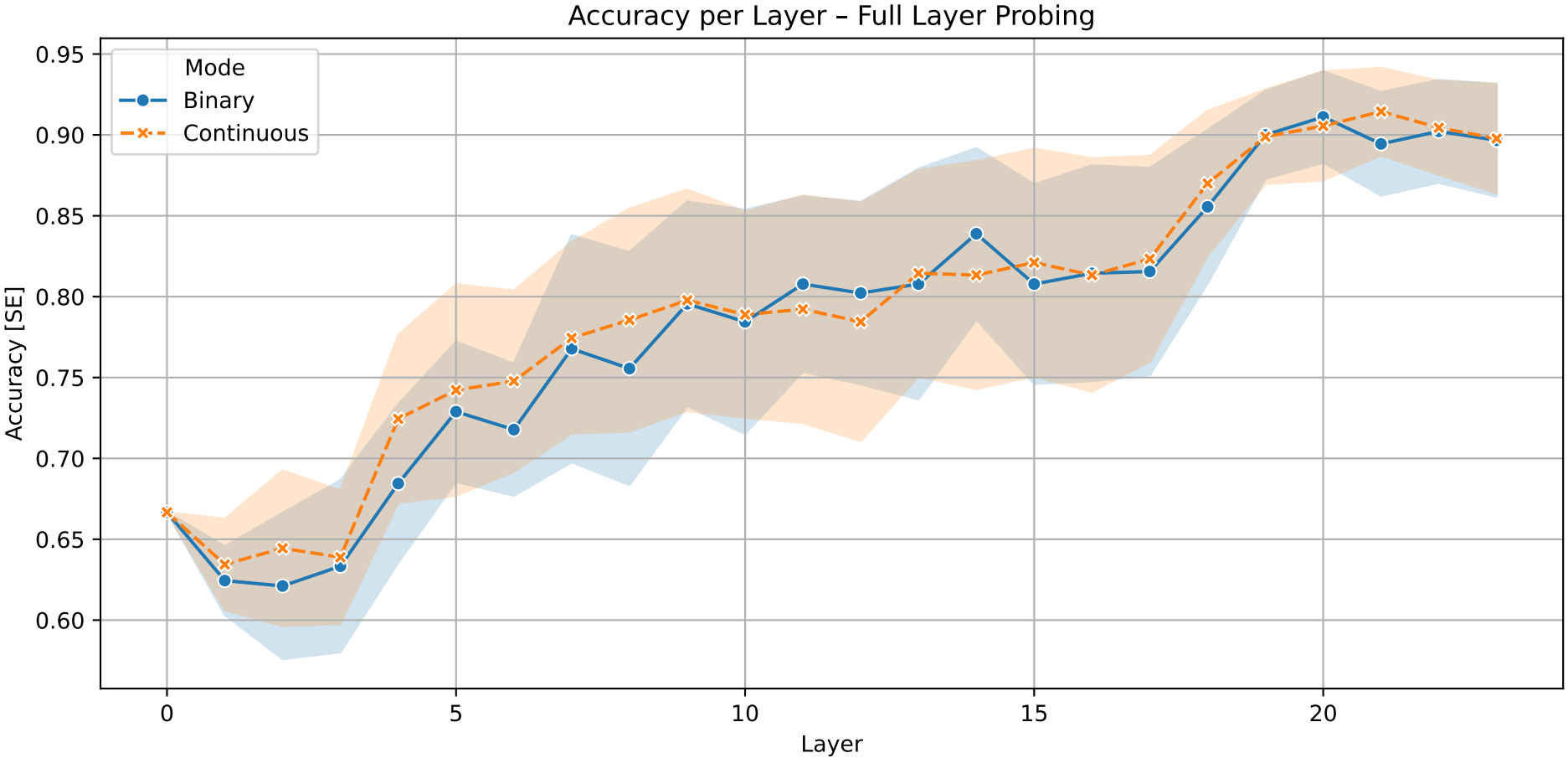

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
Layer
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
5
10
15
20

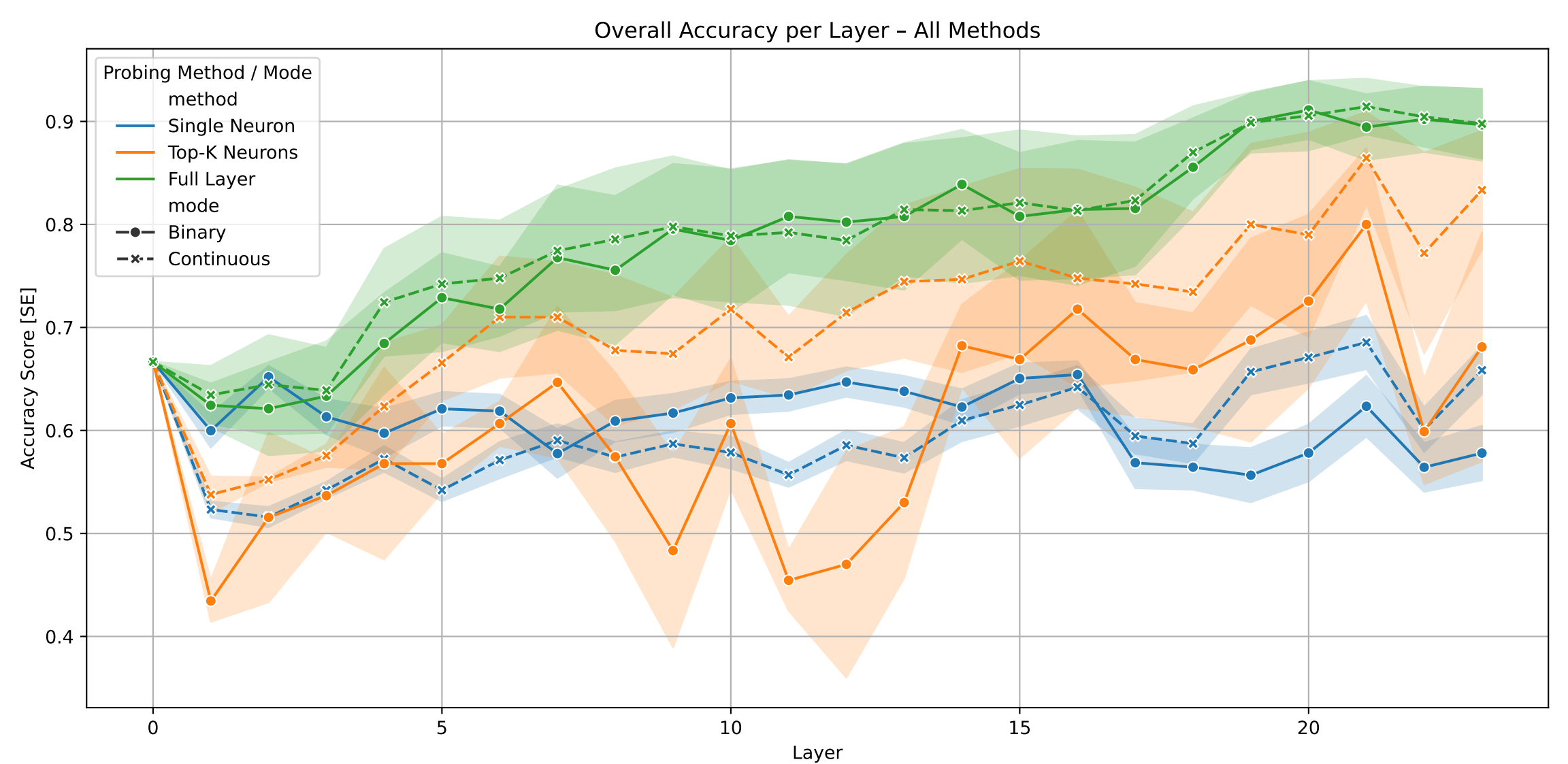
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	20.0	21.0
Full Layer	f1_max	0.9634	0.9635
Full Layer	f1_mean	0.7803	0.7898
Full Layer	f1_std	0.1227	0.1212
Single Neuron	f1_best_layer	21.0	21.0
Single Neuron	f1_max	0.9454	0.9342
Single Neuron	f1_mean	0.5138	0.6006
Single Neuron	f1_std	0.1107	0.099
Top-K Neurons	f1_best_layer	21.0	21.0
Top-K Neurons	f1_max	0.9268	0.957
Top-K Neurons	f1_mean	0.5873	0.7113
Top-K Neurons	f1_std	0.1537	0.1263


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


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.9
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]
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0
5
10
15
20
Layer


Overall Accuracy per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
Accuracy Score [SE]
0.9
0.8
0.7
0.6
0.5
0.4
0
5
10
15
20
Layer

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	20.0	21.0
Full Layer	accuracy_max	0.9633	0.9633
Full Layer	accuracy_mean	0.7848	0.7916
Full Layer	accuracy_std	0.1153	0.1148
Single Neuron	accuracy_best_layer	0.0	21.0
Single Neuron	accuracy_max	0.9467	0.9333
Single Neuron	accuracy_mean	0.6118	0.5963
Single Neuron	accuracy_std	0.1108	0.1021
Top-K Neurons	accuracy_best_layer	21.0	21.0
Top-K Neurons	accuracy_max	0.9267	0.9567
Top-K Neurons	accuracy_mean	0.6063	0.7099
Top-K Neurons	accuracy_std	0.1389	0.125