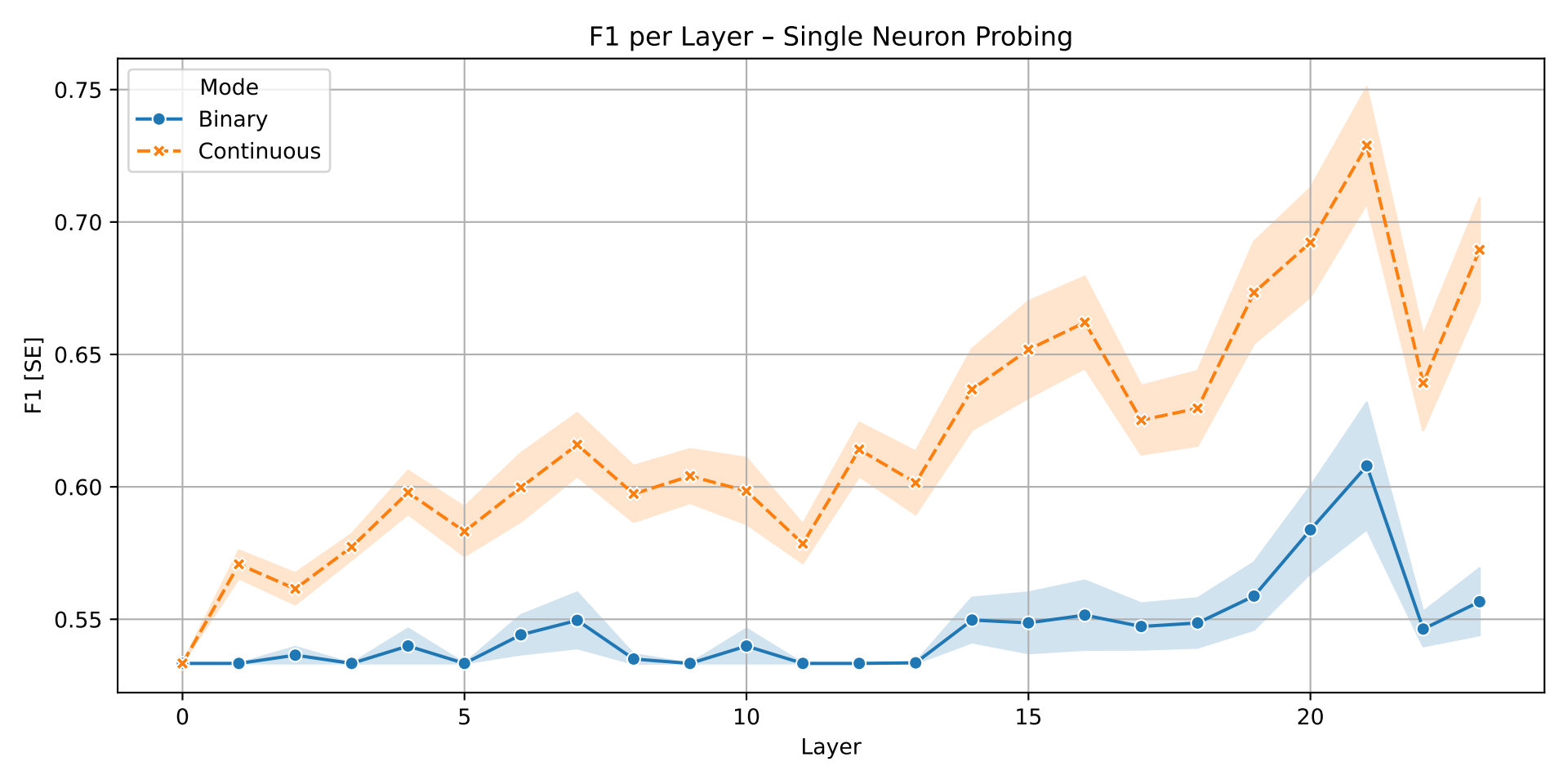
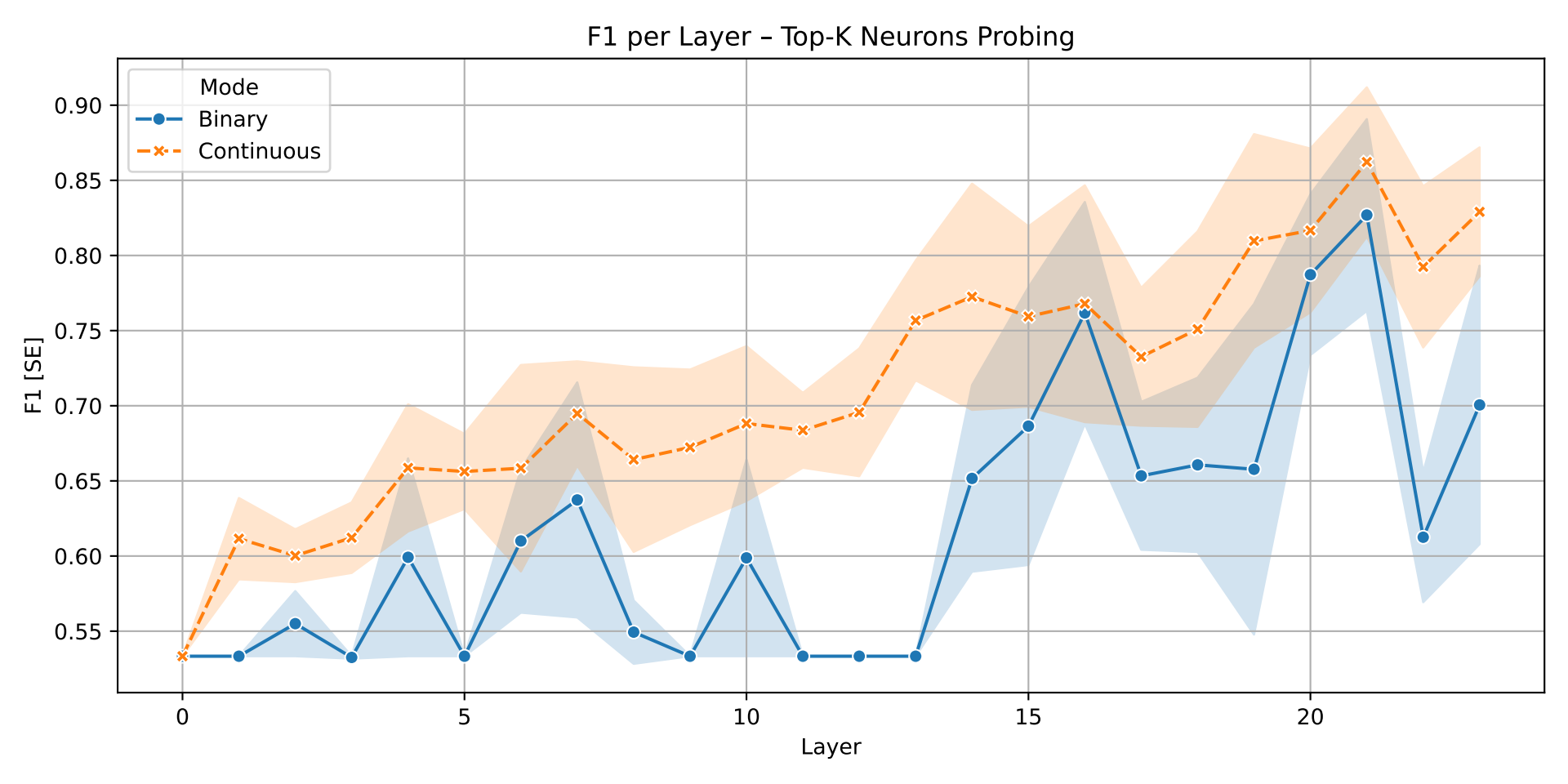
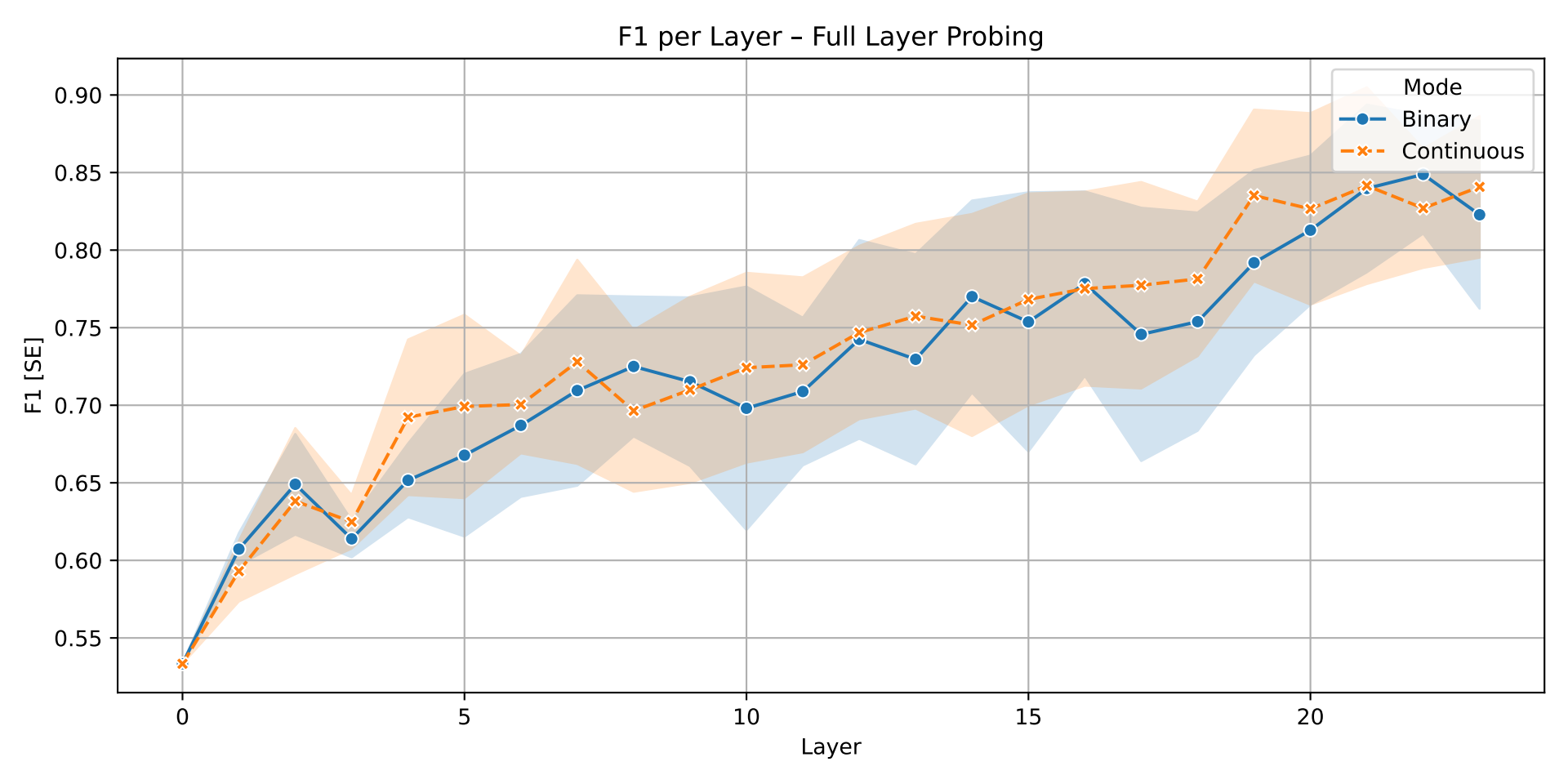
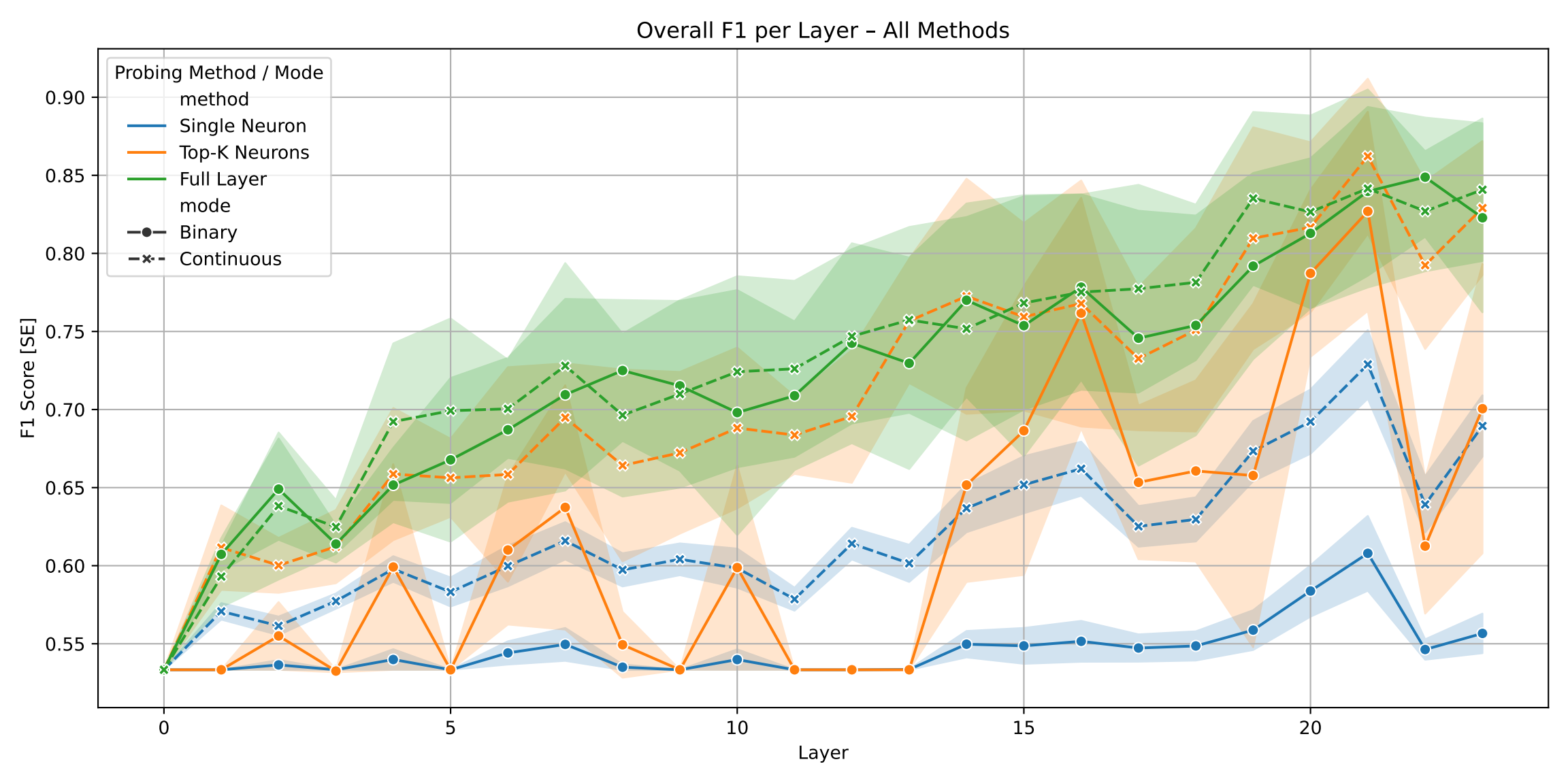

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
Layer

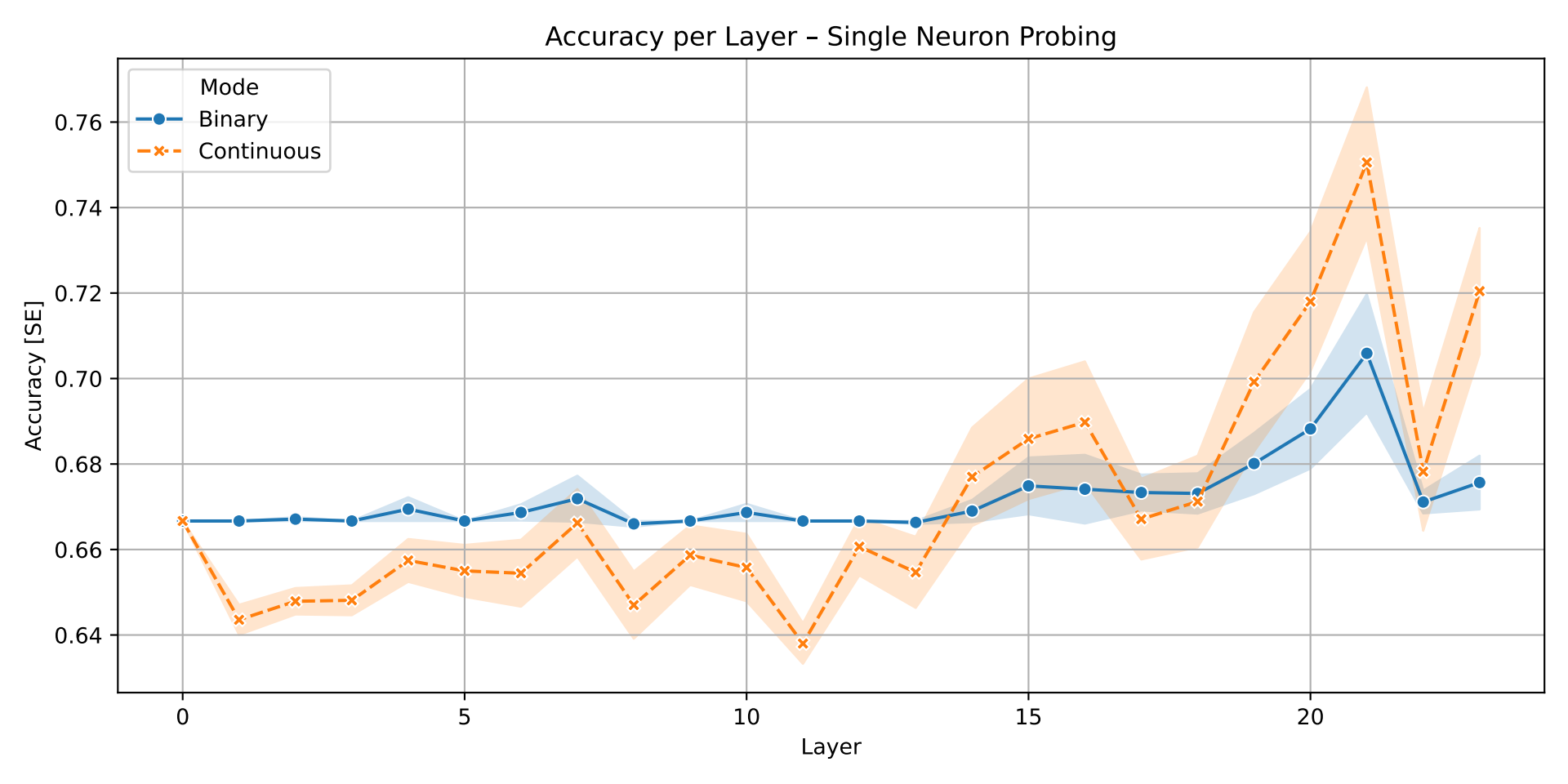

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
Layer

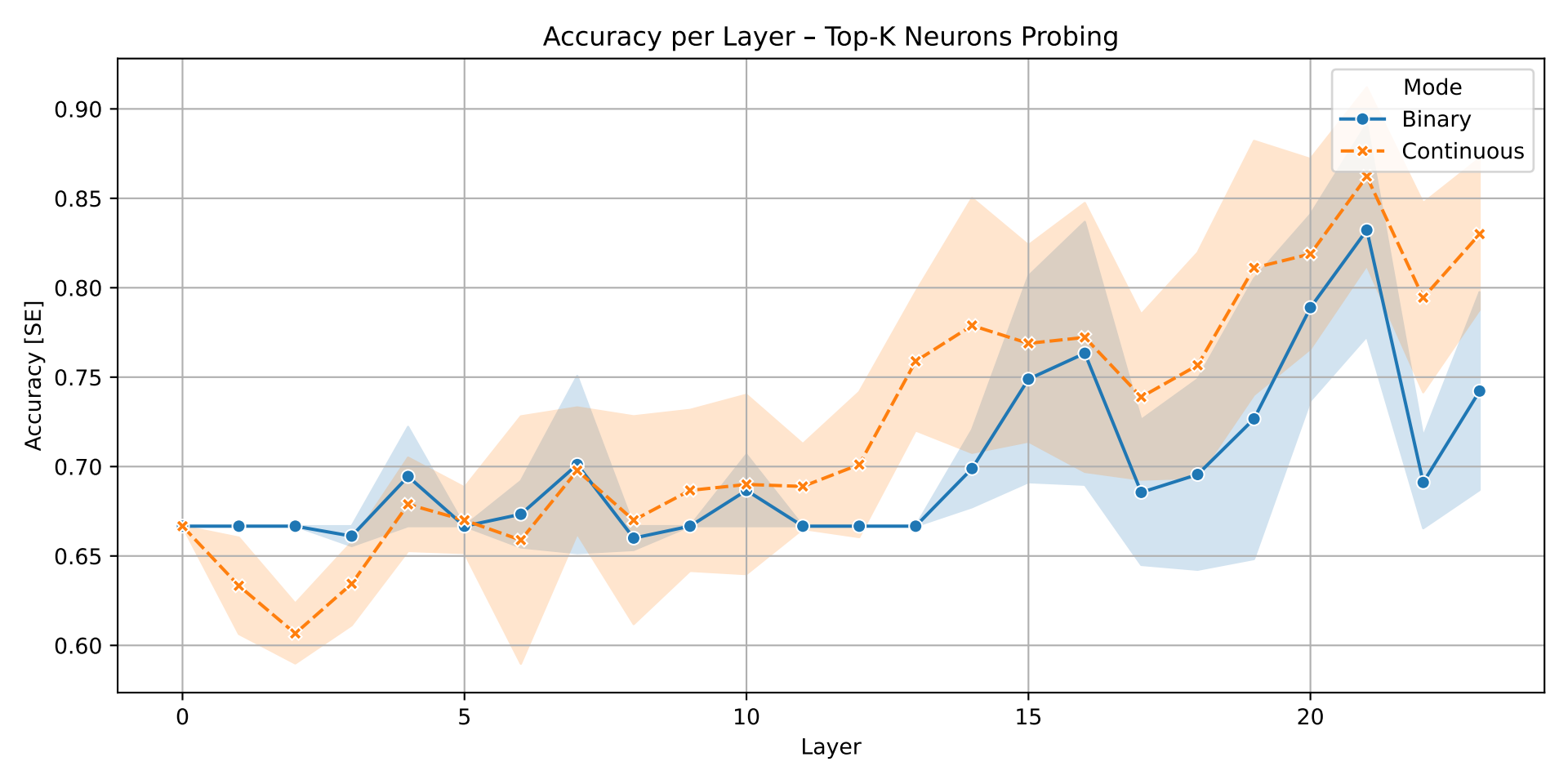

F1 per Layer – Full Layer Probing
F1 [SE]
Layer
Mode
Binary
Continuous
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20

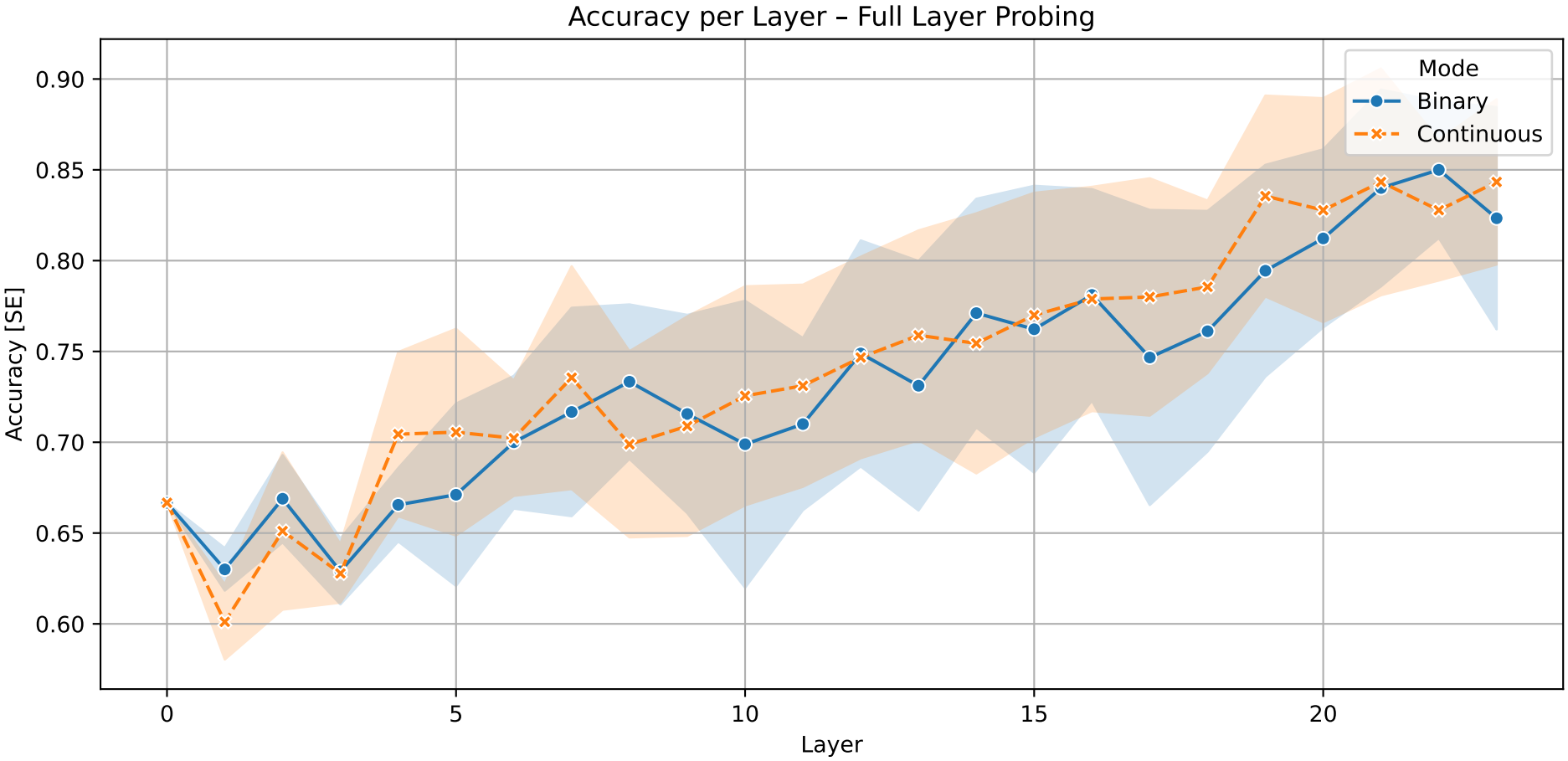

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
Layer

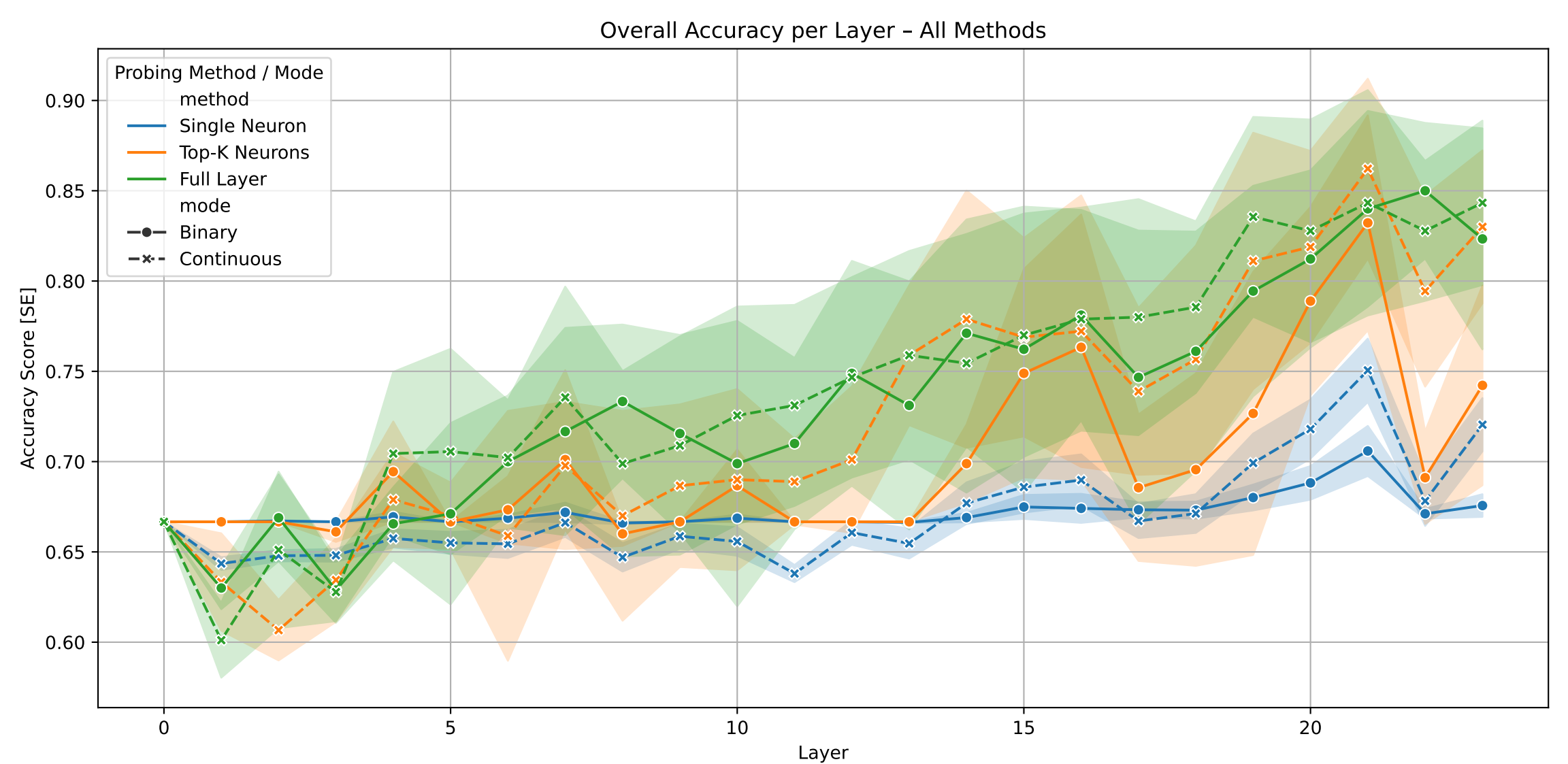
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	22.0	21.0
Full Layer	f1_max	0.9465	0.9599
Full Layer	f1_mean	0.7232	0.7331
Full Layer	f1_std	0.1093	0.1096
Single Neuron	f1_best_layer	21.0	21.0
Single Neuron	f1_max	0.9454	0.9599
Single Neuron	f1_mean	0.5463	0.6193
Single Neuron	f1_std	0.0523	0.0862
Top-K Neurons	f1_best_layer	21.0	21.0
Top-K Neurons	f1_max	0.9454	0.9599
Top-K Neurons	f1_mean	0.6173	0.7116
Top-K Neurons	f1_std	0.1162	0.1079


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.76
0.74
0.72
0.70
0.68
0.66
0.64
0
5
10
15
20
Layer


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


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

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	22.0	21.0
Full Layer	accuracy_max	0.9467	0.96
Full Layer	accuracy_mean	0.7345	0.7421
Full Layer	accuracy_std	0.0983	0.1001
Single Neuron	accuracy_best_layer	21.0	21.0
Single Neuron	accuracy_max	0.9467	0.96
Single Neuron	accuracy_mean	0.6721	0.6713
Single Neuron	accuracy_std	0.0278	0.0614
Top-K Neurons	accuracy_best_layer	21.0	21.0
Top-K Neurons	accuracy_max	0.9467	0.96
Top-K Neurons	accuracy_mean	0.6993	0.7239
Top-K Neurons	accuracy_std	0.0698	0.0969