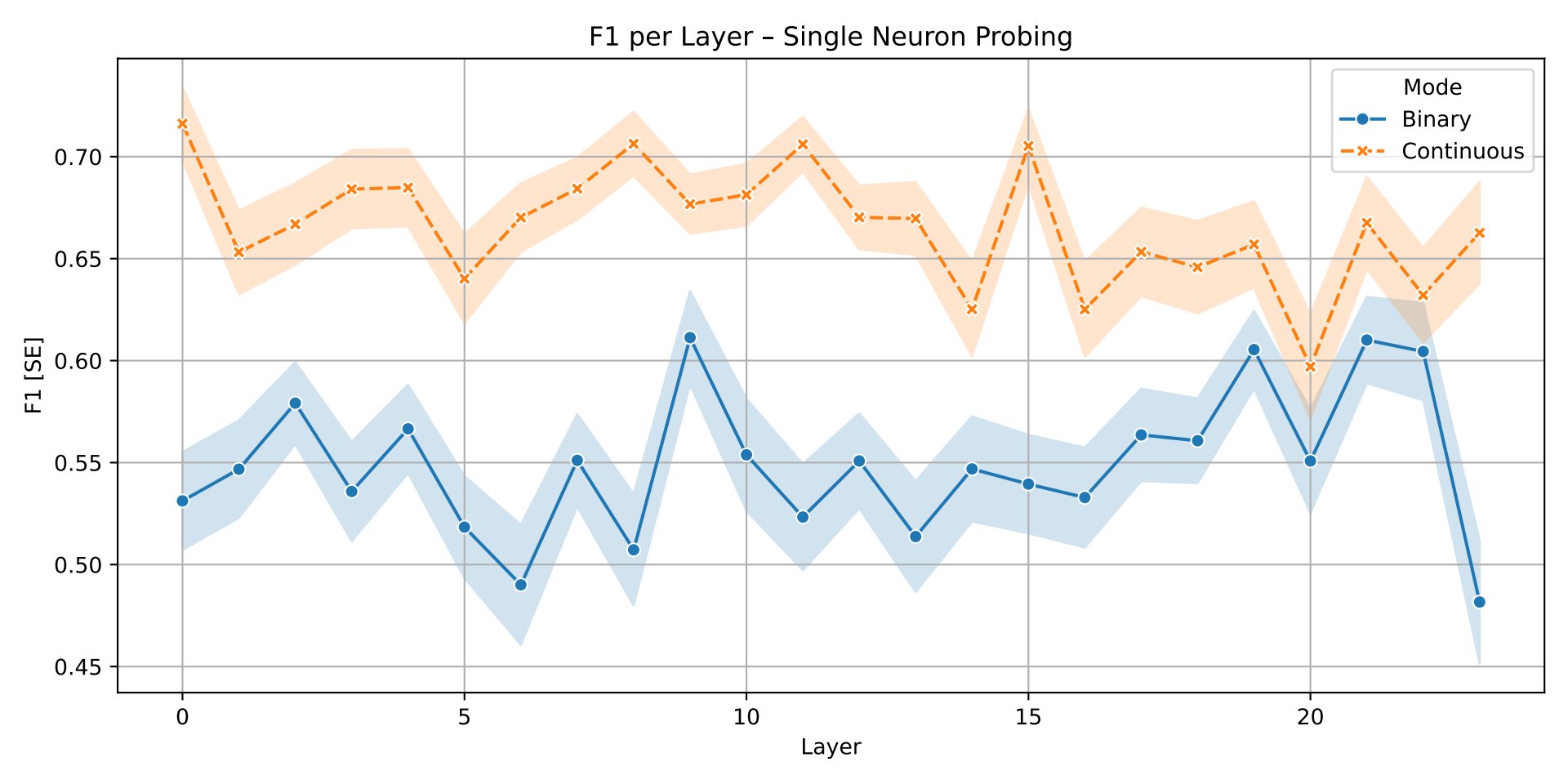
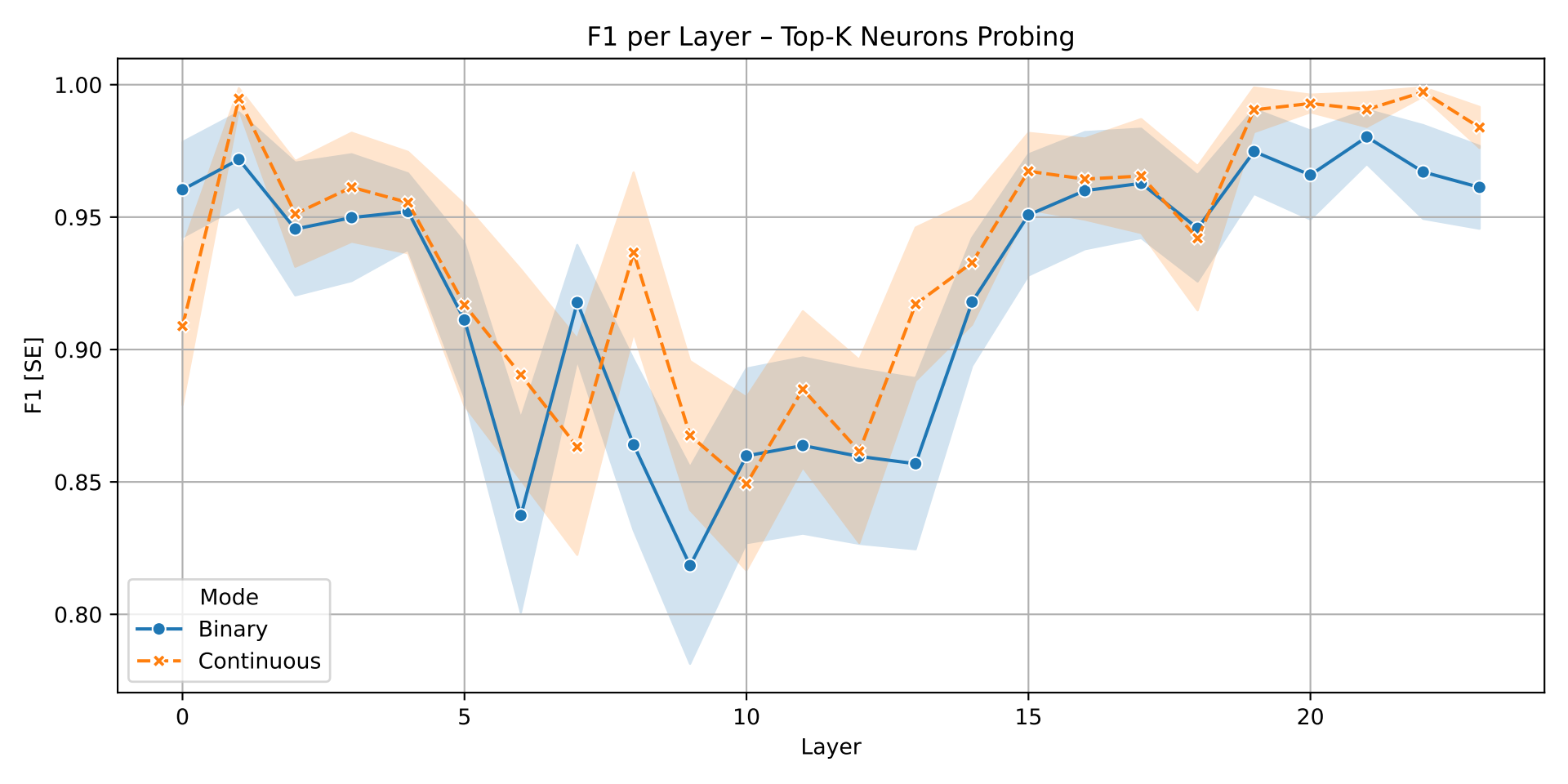
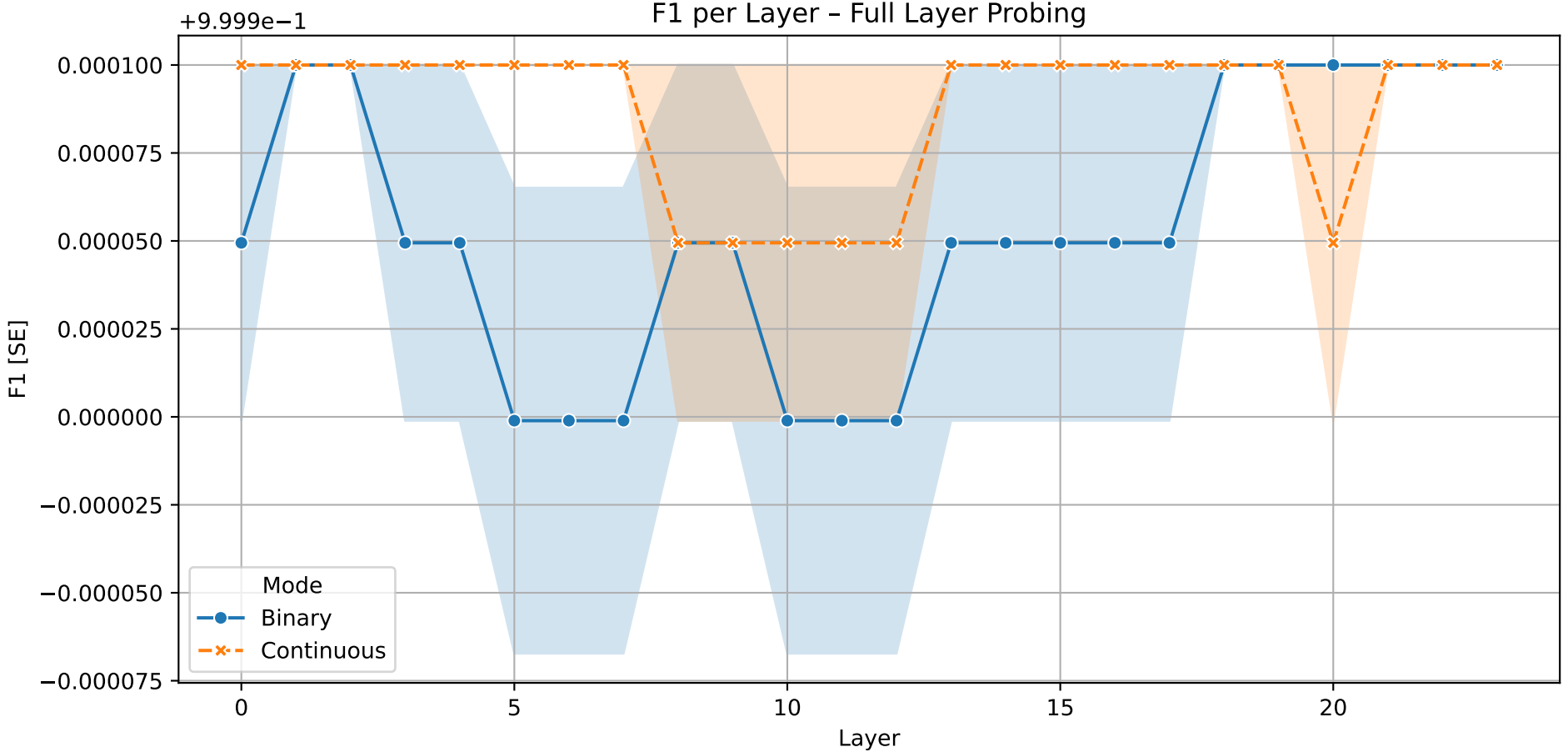
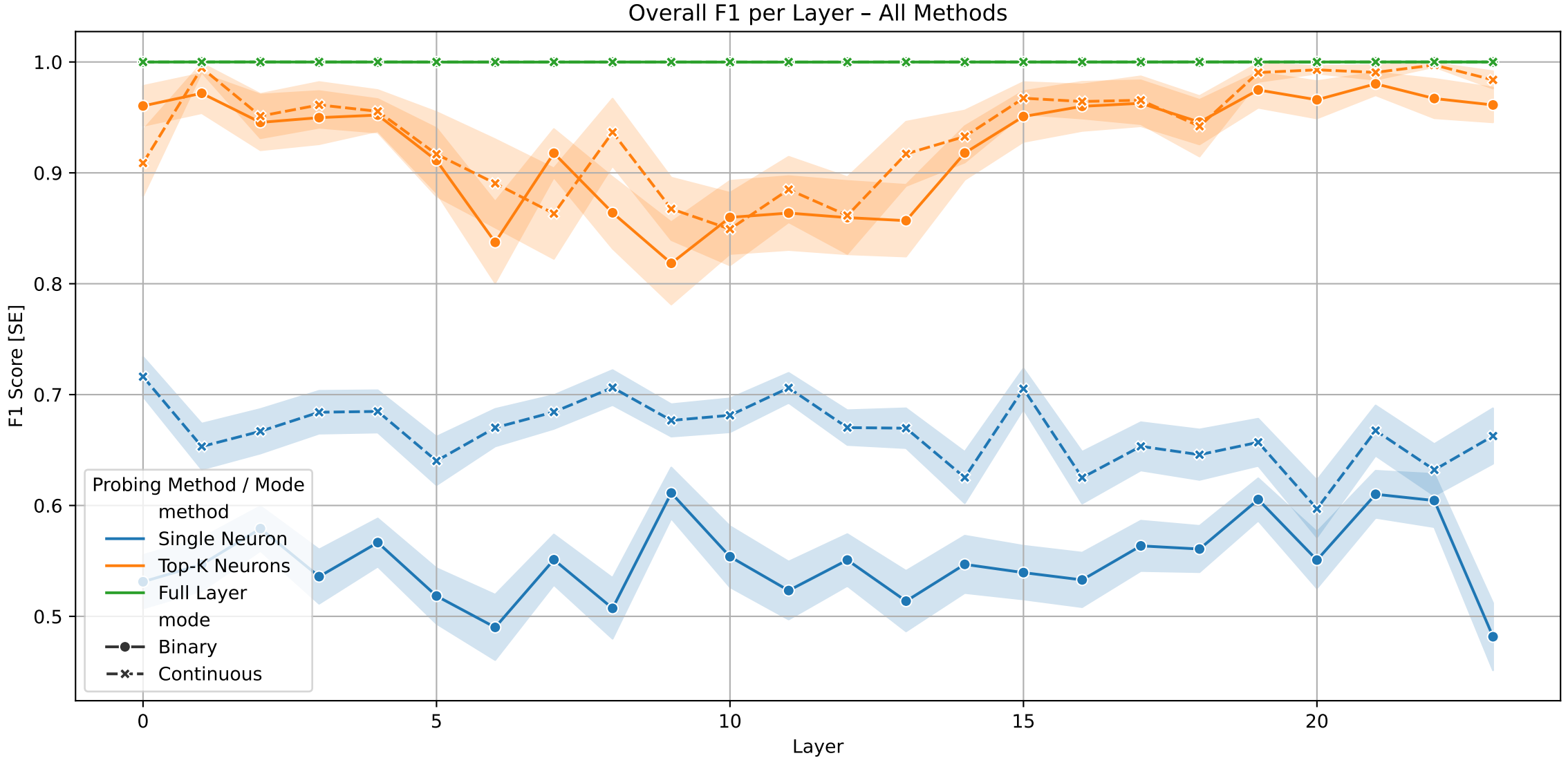

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.70
0.65
0.60
0.55
0.50
0.45
0
5
10
15
20
Layer

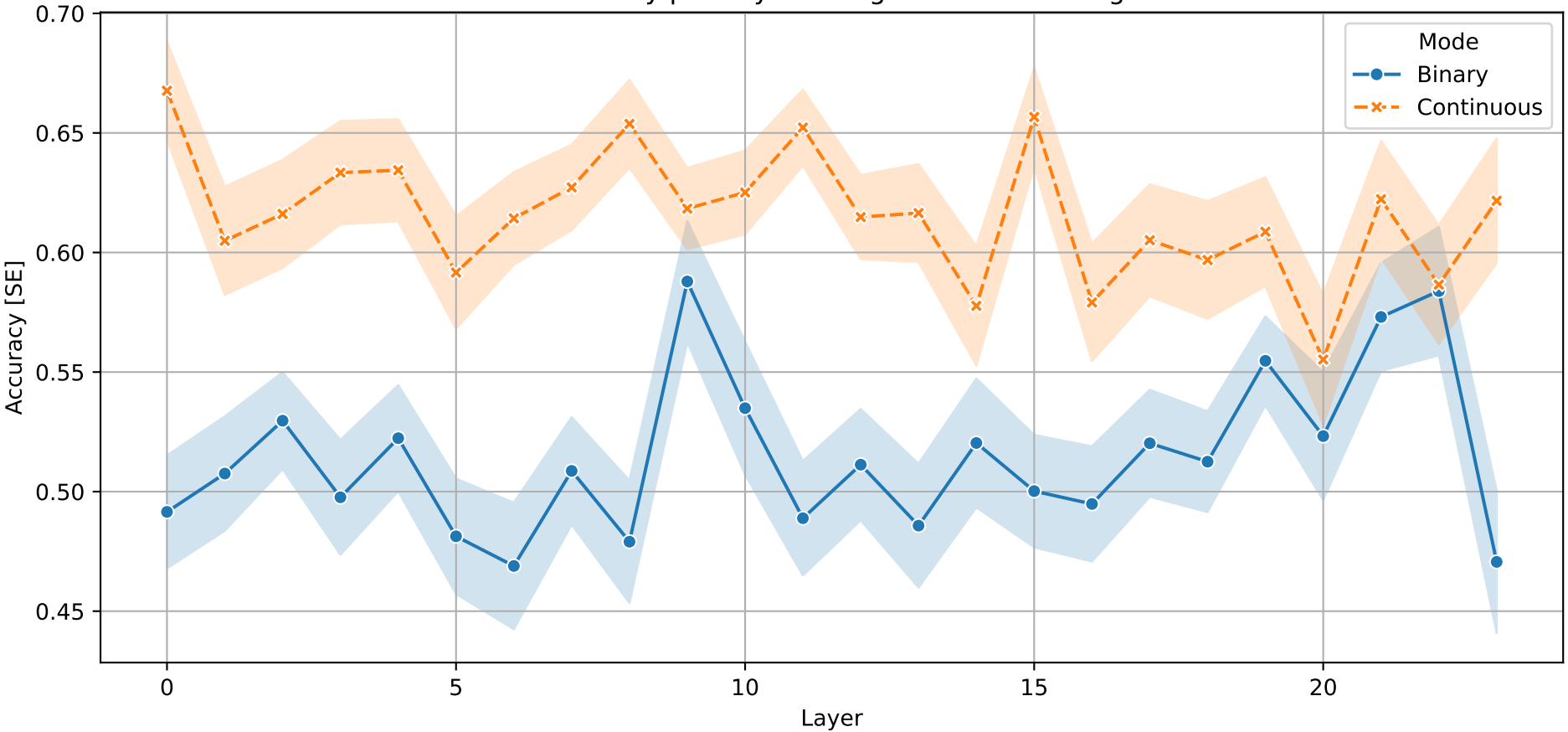

F1 per Layer – Top-K Neurons Probing
F1 [SE]
1.00
0.95
0.90
0.85
0.80
0
5
10
15
20
Layer
Mode
Binary
Continuous

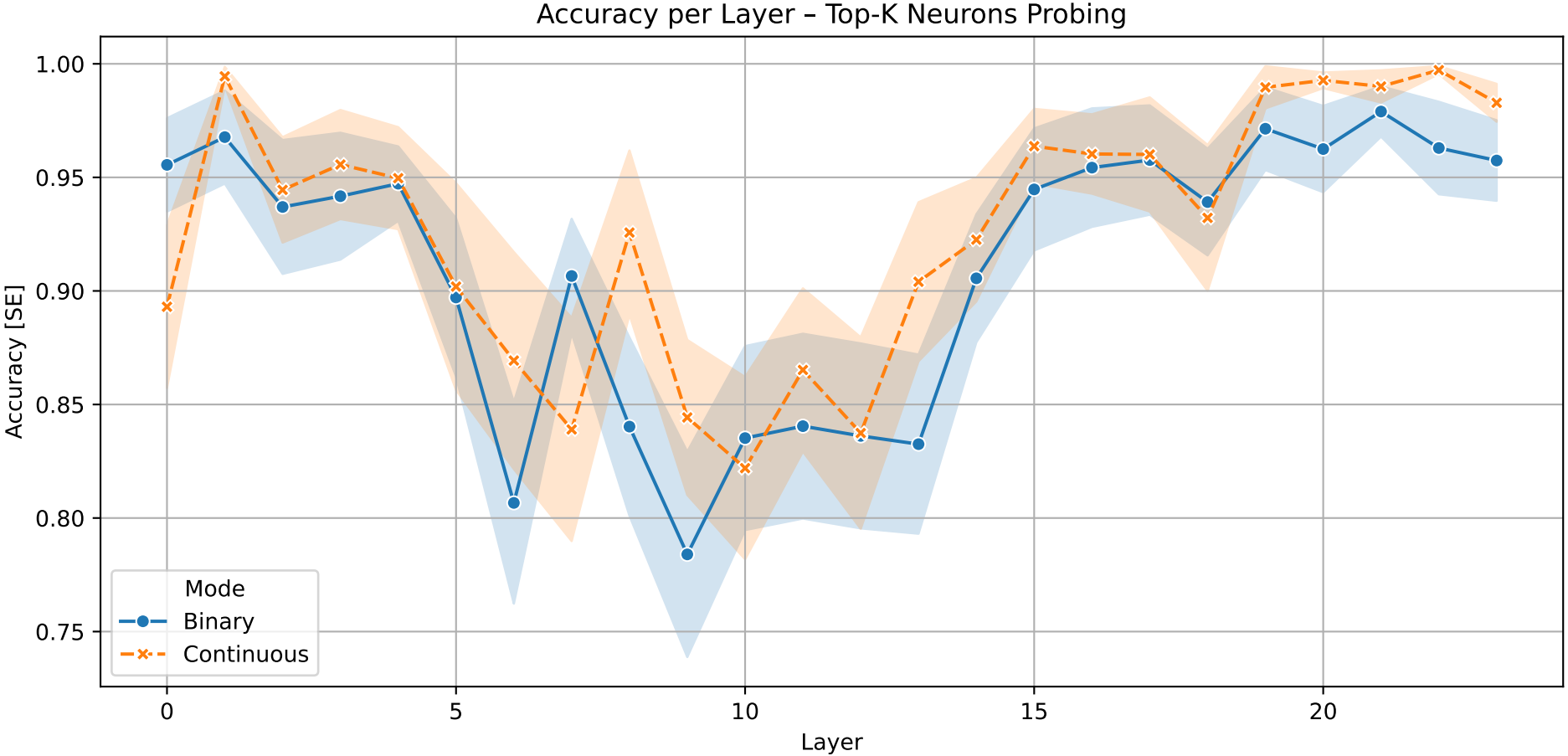

F1 per Layer – Full Layer Probing
+9.999e−1
0.000100
0.000075
0.000050
0.000025
0.000000
−0.000025
−0.000050
−0.000075
F1 [SE]
0
5
10
15
20
Layer
Mode
Binary
Continuous

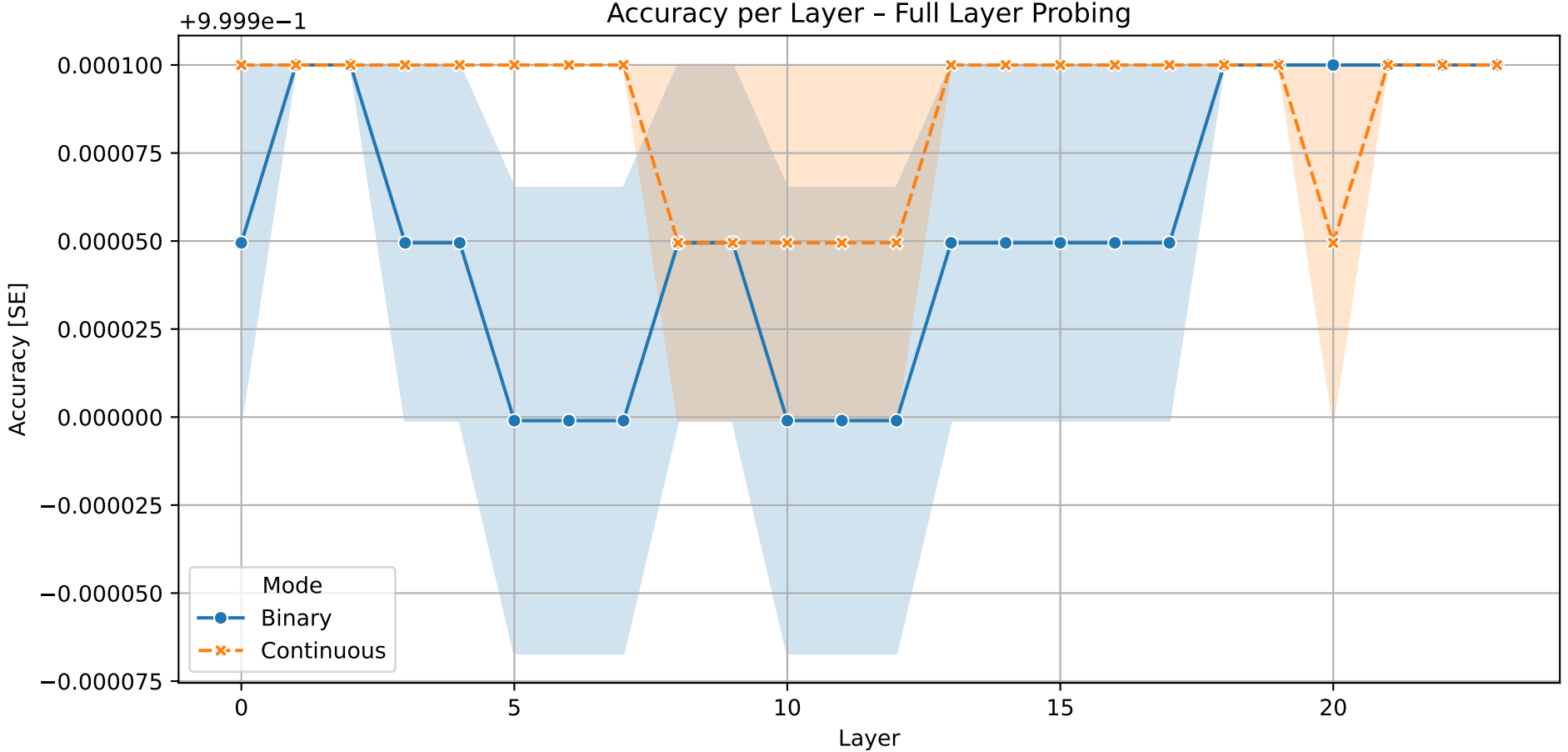

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
1.0
0.9
0.8
0.7
0.6
0.5
0
5
10
15
20
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

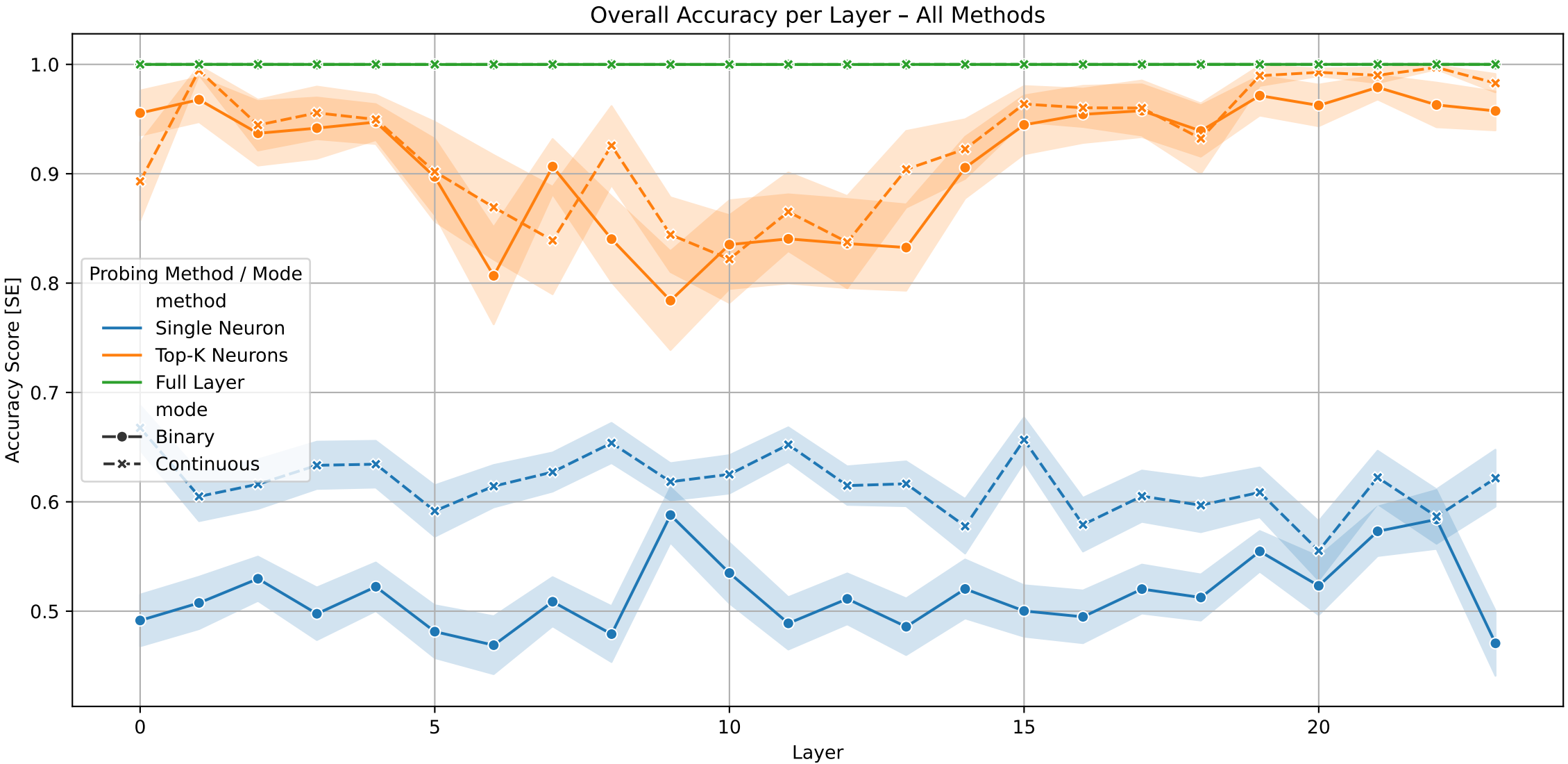
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	1.0	0.0
Full Layer	f1_max	1.0	1.0
Full Layer	f1_mean	1.0	1.0
Full Layer	f1_std	0.0001	0.0001
Single Neuron	f1_best_layer	9.0	0.0
Single Neuron	f1_max	1.0	1.0
Single Neuron	f1_mean	0.549	0.6659
Single Neuron	f1_std	0.2202	0.1789
Top-K Neurons	f1_best_layer	21.0	22.0
Top-K Neurons	f1_max	1.0	1.0
Top-K Neurons	f1_mean	0.9231	0.9369
Top-K Neurons	f1_std	0.083	0.0807


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


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
1.00
0.95
0.90
0.85
0.80
0.75
0
5
10
15
20
Layer
Mode
Binary
Continuous


Accuracy per Layer – Full Layer Probing
+9.999e−1
0.000100
0.000075
0.000050
0.000025
0.000000
−0.000025
−0.000050
−0.000075
Accuracy [SE]
0
5
10
15
20
Layer
Mode
Binary
Continuous


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
Layer
1.0
0.9
0.8
0.7
0.6
0.5
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	1.0	0.0
Full Layer	accuracy_max	1.0	1.0
Full Layer	accuracy_mean	1.0	1.0
Full Layer	accuracy_std	0.0001	0.0001
Single Neuron	accuracy_best_layer	9.0	0.0
Single Neuron	accuracy_max	1.0	1.0
Single Neuron	accuracy_mean	0.5145	0.6159
Single Neuron	accuracy_std	0.2187	0.1959
Top-K Neurons	accuracy_best_layer	21.0	22.0
Top-K Neurons	accuracy_max	1.0	1.0
Top-K Neurons	accuracy_mean	0.9109	0.9266
Top-K Neurons	accuracy_std	0.0993	0.0964