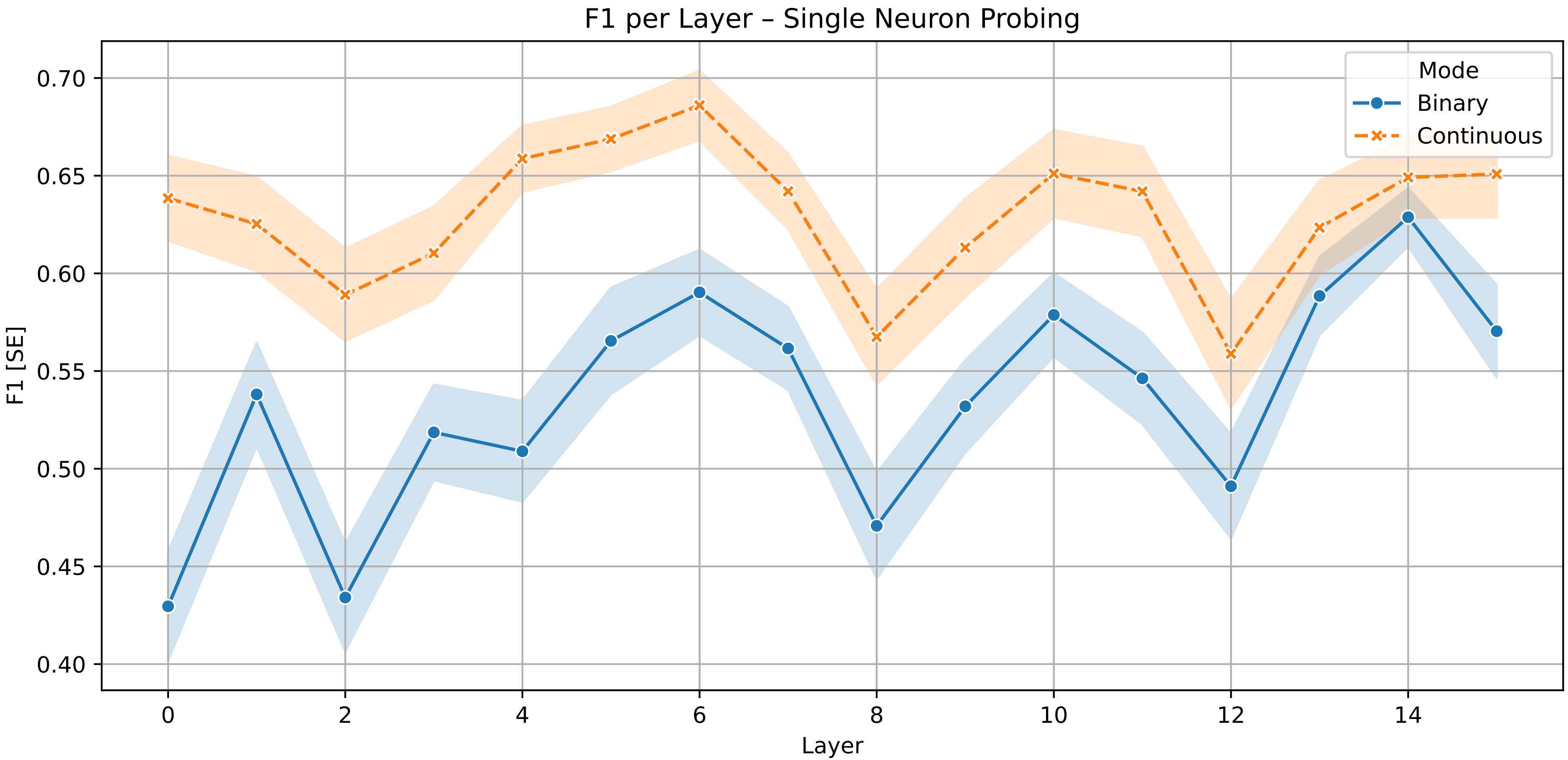
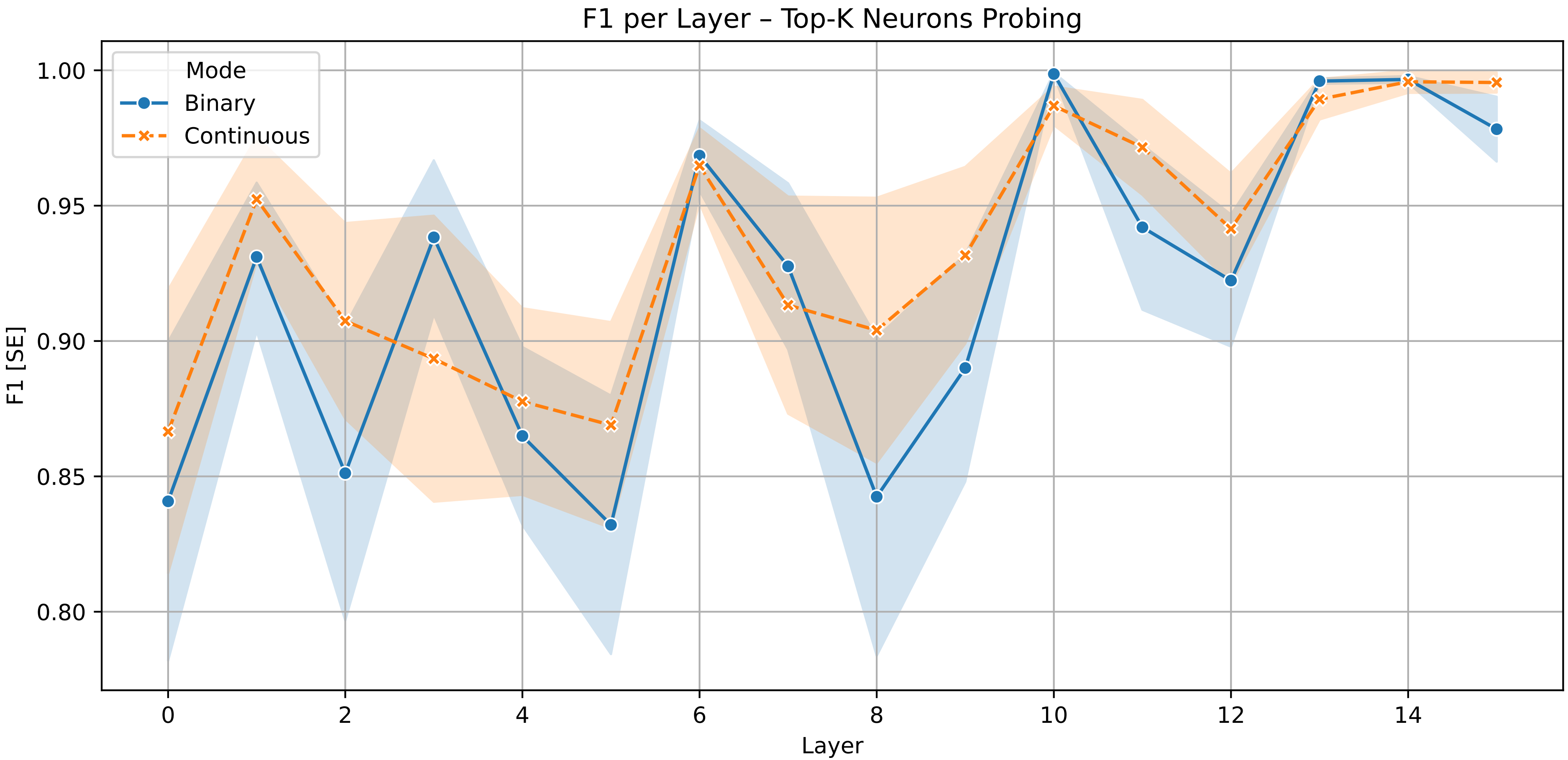
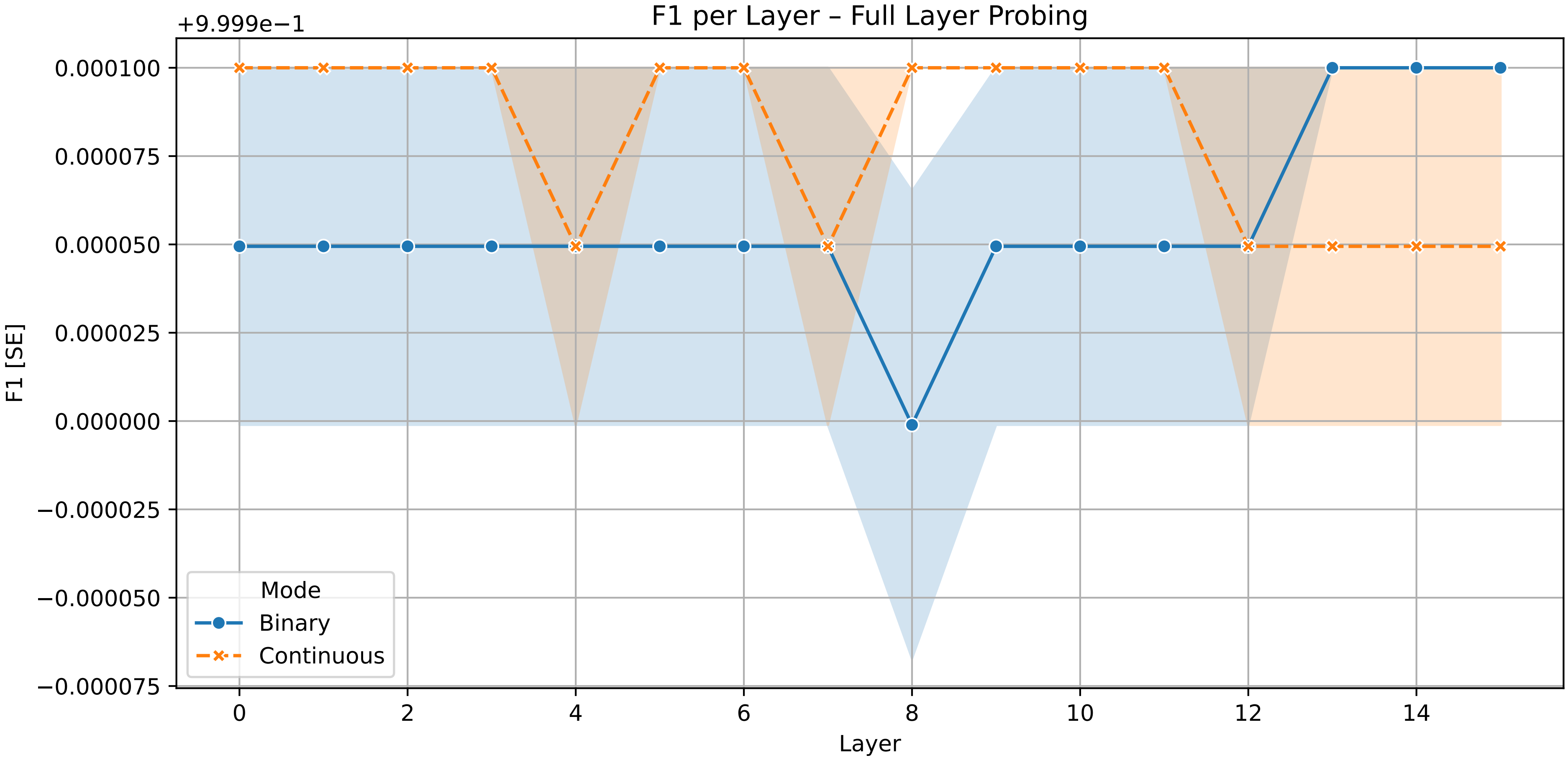
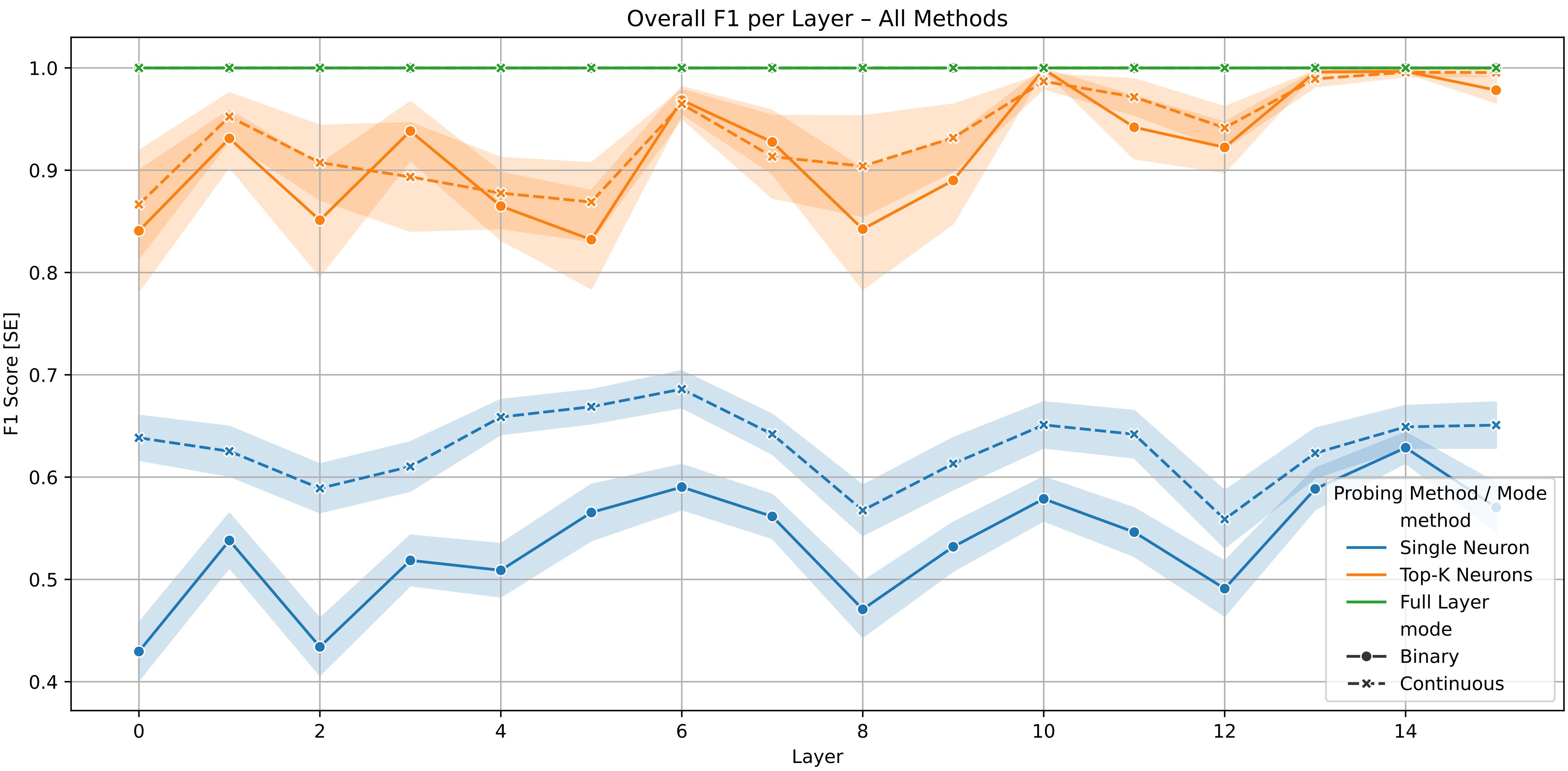
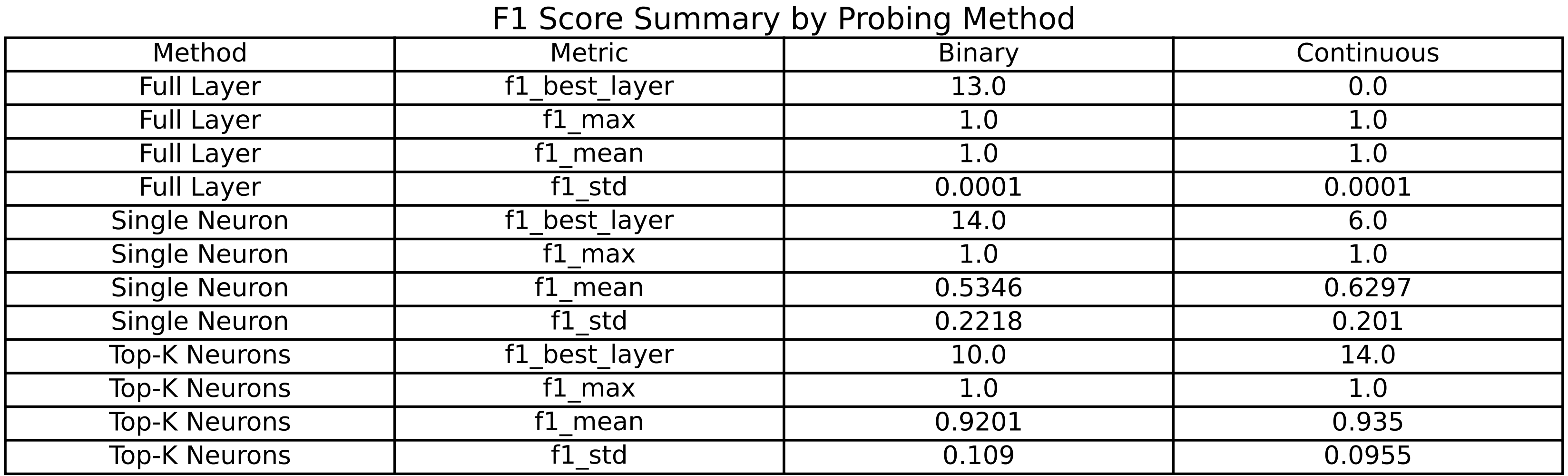

F1 per Layer – Single Neuron Probing
F1 [SE]
0.40
0.45
0.50
0.55
0.60
0.65
0.70
Layer
0
2
4
6
8
10
12
14
Mode
Binary
Continuous


F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
1.00
0.95
0.90
0.85
0.80
Layer
0
2
4
6
8
10
12
14

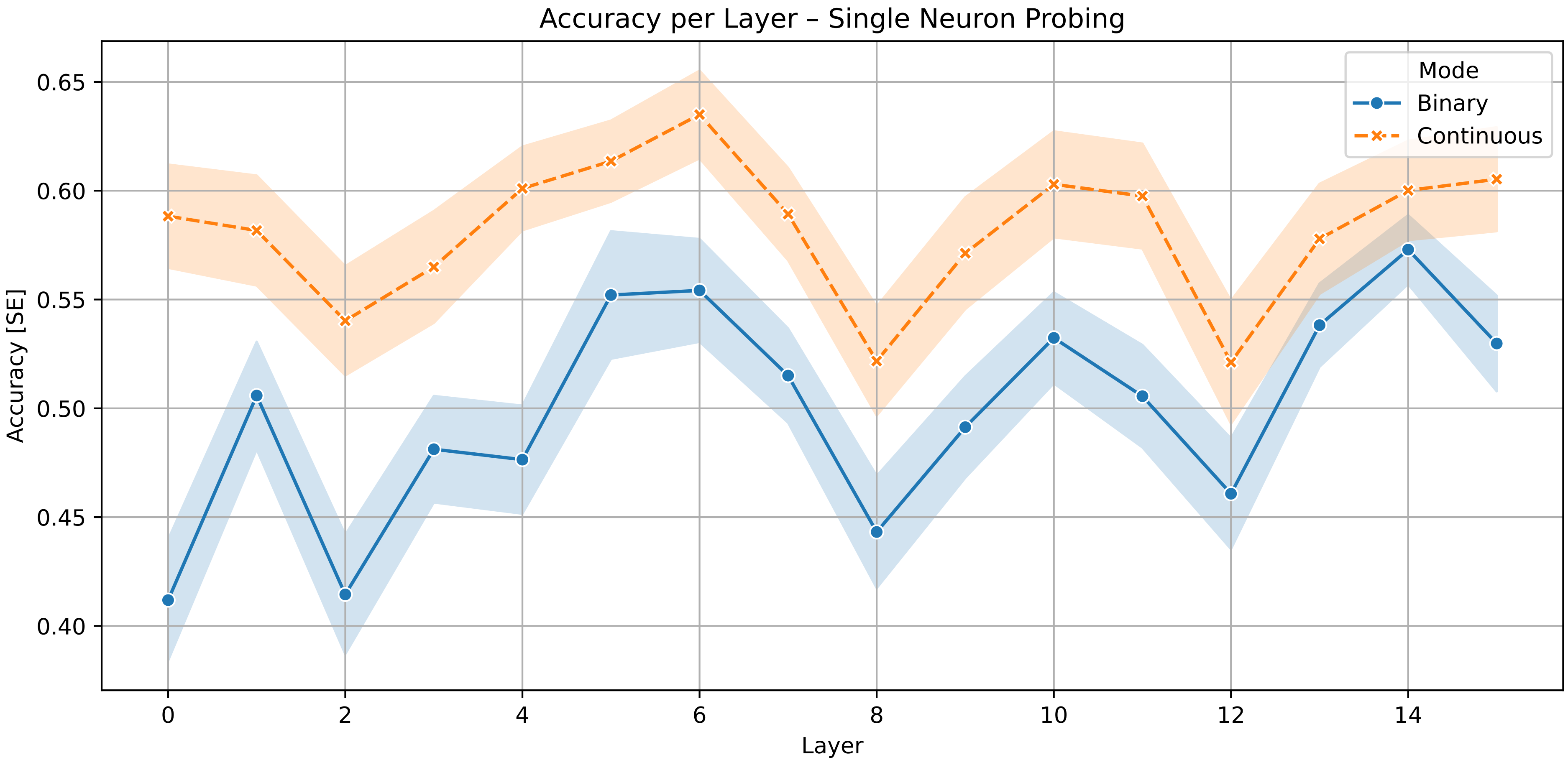

F1 per Layer – Full Layer Probing
+9.999e−1
0.000100
0.000075
0.000050
0.000025
0.000000
−0.000025
−0.000050
−0.000075
F1 [SE]
0
2
4
6
8
10
12
14
Layer
Mode
Binary
Continuous

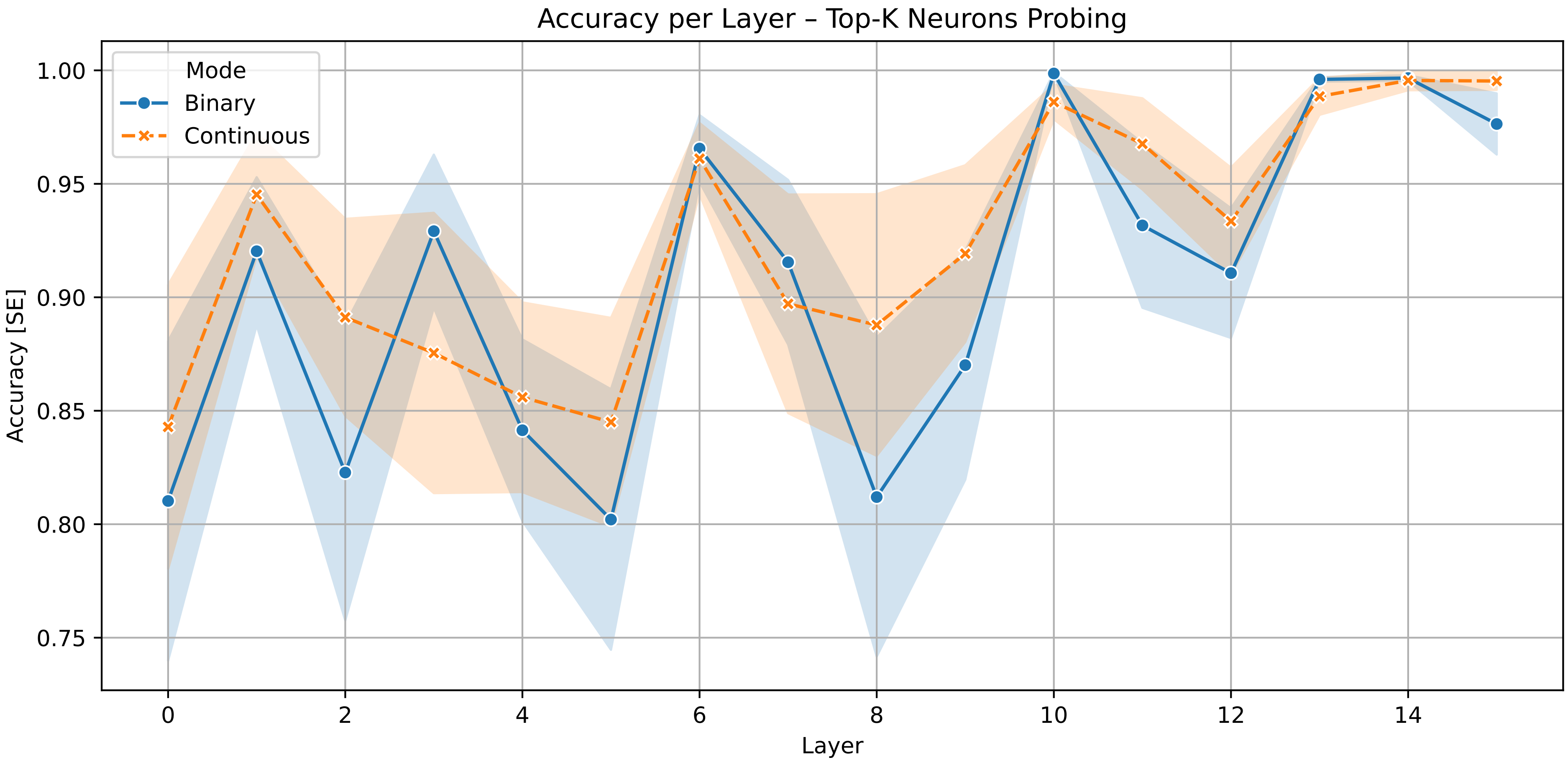

Overall F1 per Layer – All Methods
F1 Score [SE]
1.0
0.9
0.8
0.7
0.6
0.5
0.4
Layer
0
2
4
6
8
10
12
14
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

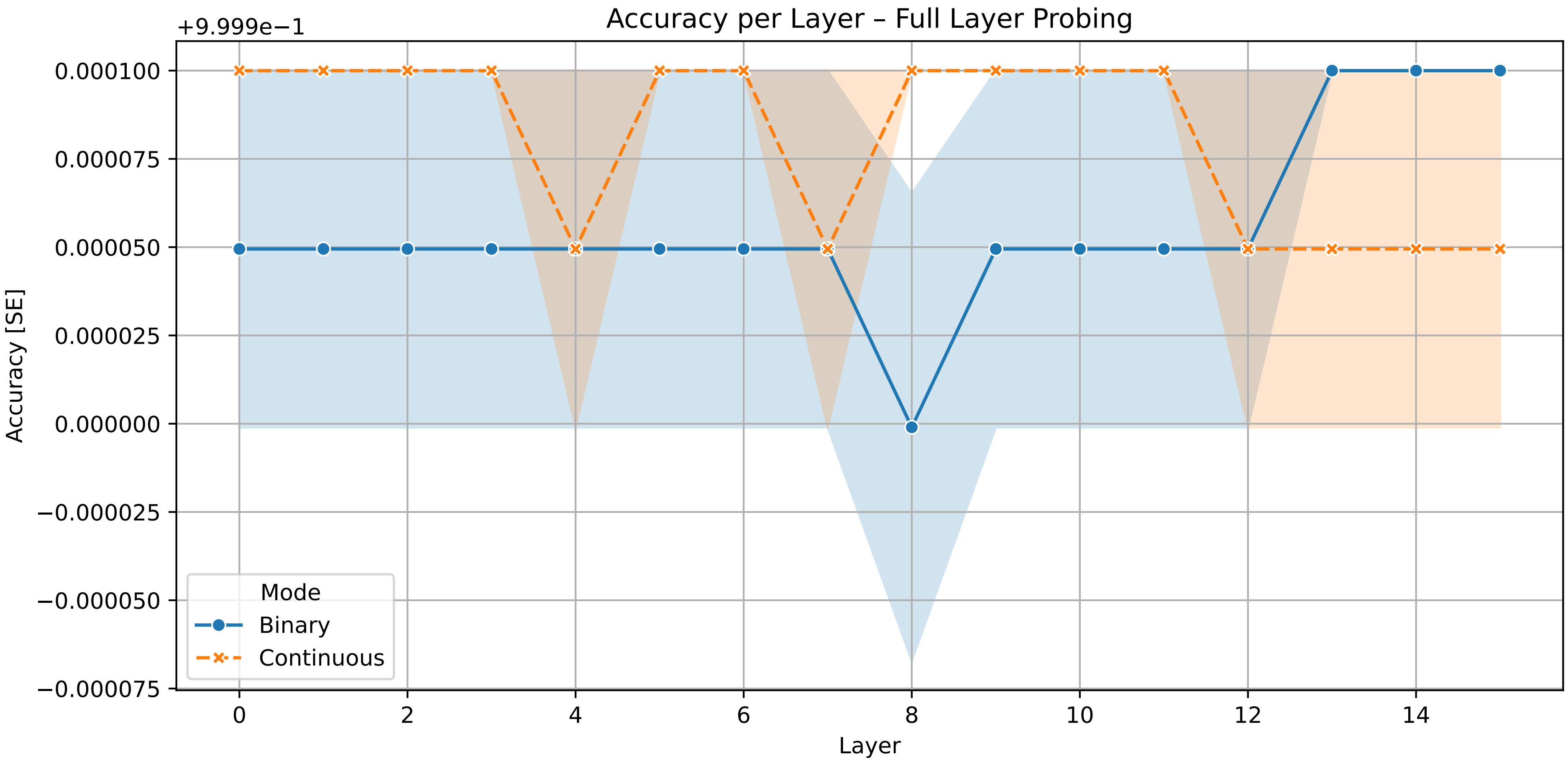


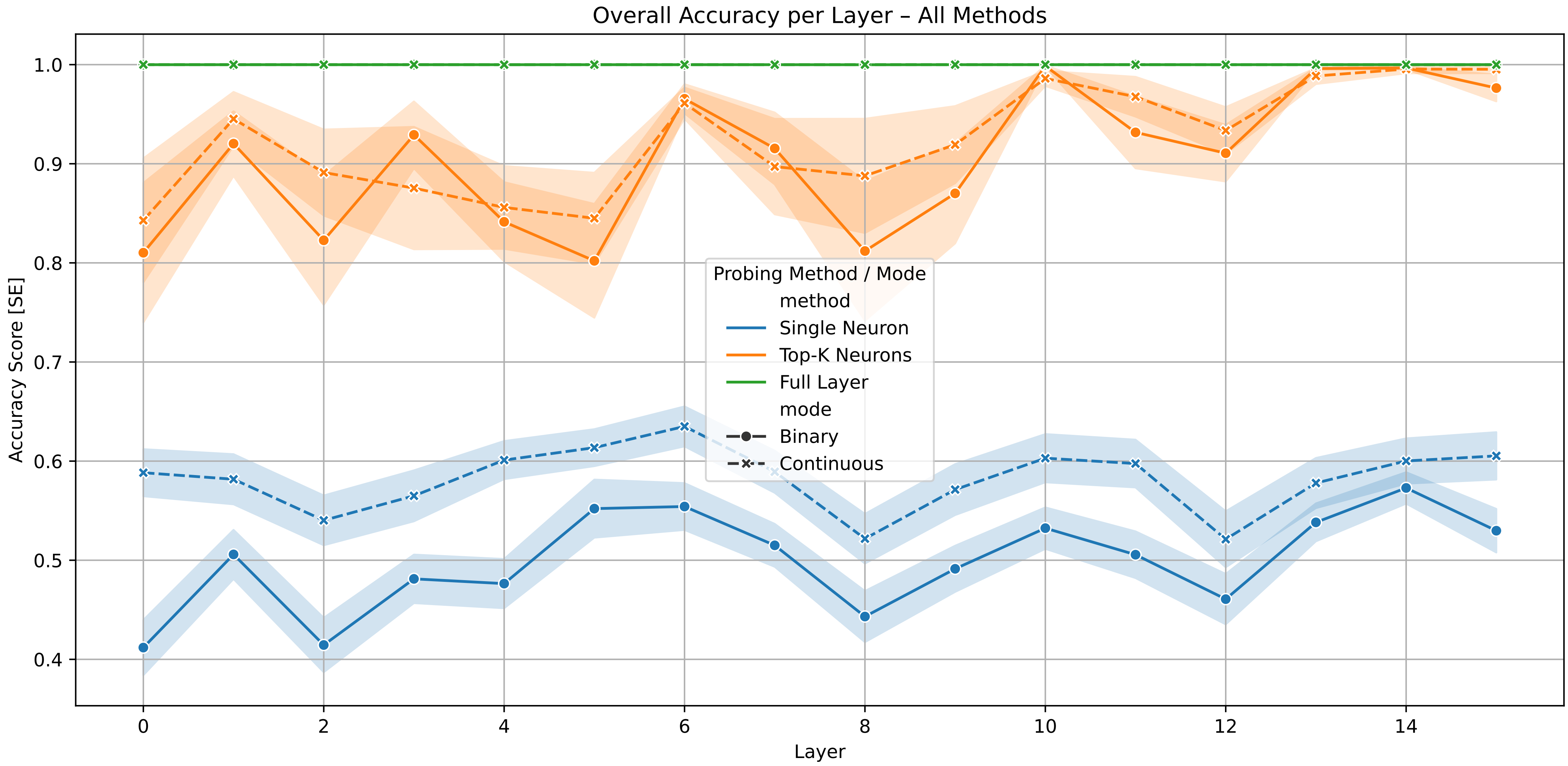
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	13.0	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	14.0	6.0
Single Neuron	f1_max	1.0	1.0
Single Neuron	f1_mean	0.5346	0.6297
Single Neuron	f1_std	0.2218	0.201
Top-K Neurons	f1_best_layer	10.0	14.0
Top-K Neurons	f1_max	1.0	1.0
Top-K Neurons	f1_mean	0.9201	0.935
Top-K Neurons	f1_std	0.109	0.0955

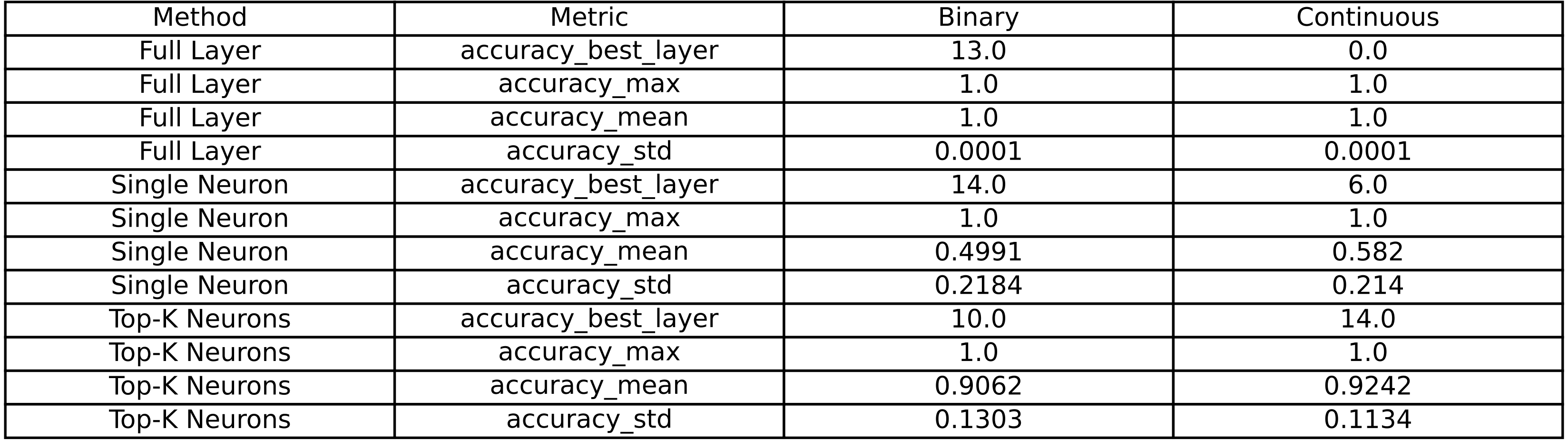

Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.40
0.45
0.50
0.55
0.60
0.65
Layer
0
2
4
6
8
10
12
14


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
1.00
0.95
0.90
0.85
0.80
0.75
Layer
0
2
4
6
8
10
12
14


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
2
4
6
8
10
12
14
Layer
Mode
Binary
Continuous


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
1.0
0.9
0.8
0.7
0.6
0.5
0.4
Layer
0
2
4
6
8
10
12
14
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	13.0	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	14.0	6.0
Single Neuron	accuracy_max	1.0	1.0
Single Neuron	accuracy_mean	0.4991	0.582
Single Neuron	accuracy_std	0.2184	0.214
Top-K Neurons	accuracy_best_layer	10.0	14.0
Top-K Neurons	accuracy_max	1.0	1.0
Top-K Neurons	accuracy_mean	0.9062	0.9242
Top-K Neurons	accuracy_std	0.1303	0.1134