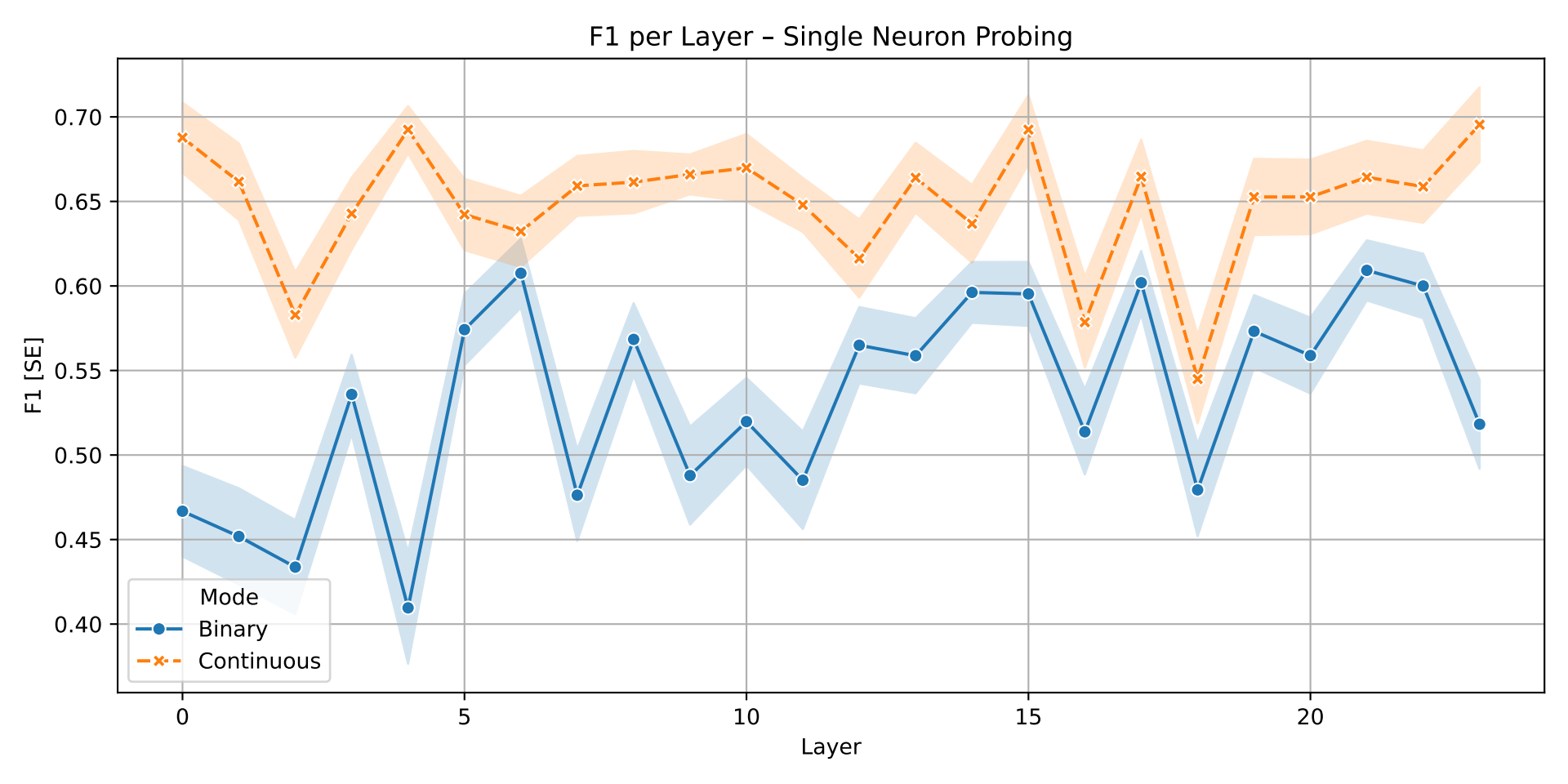
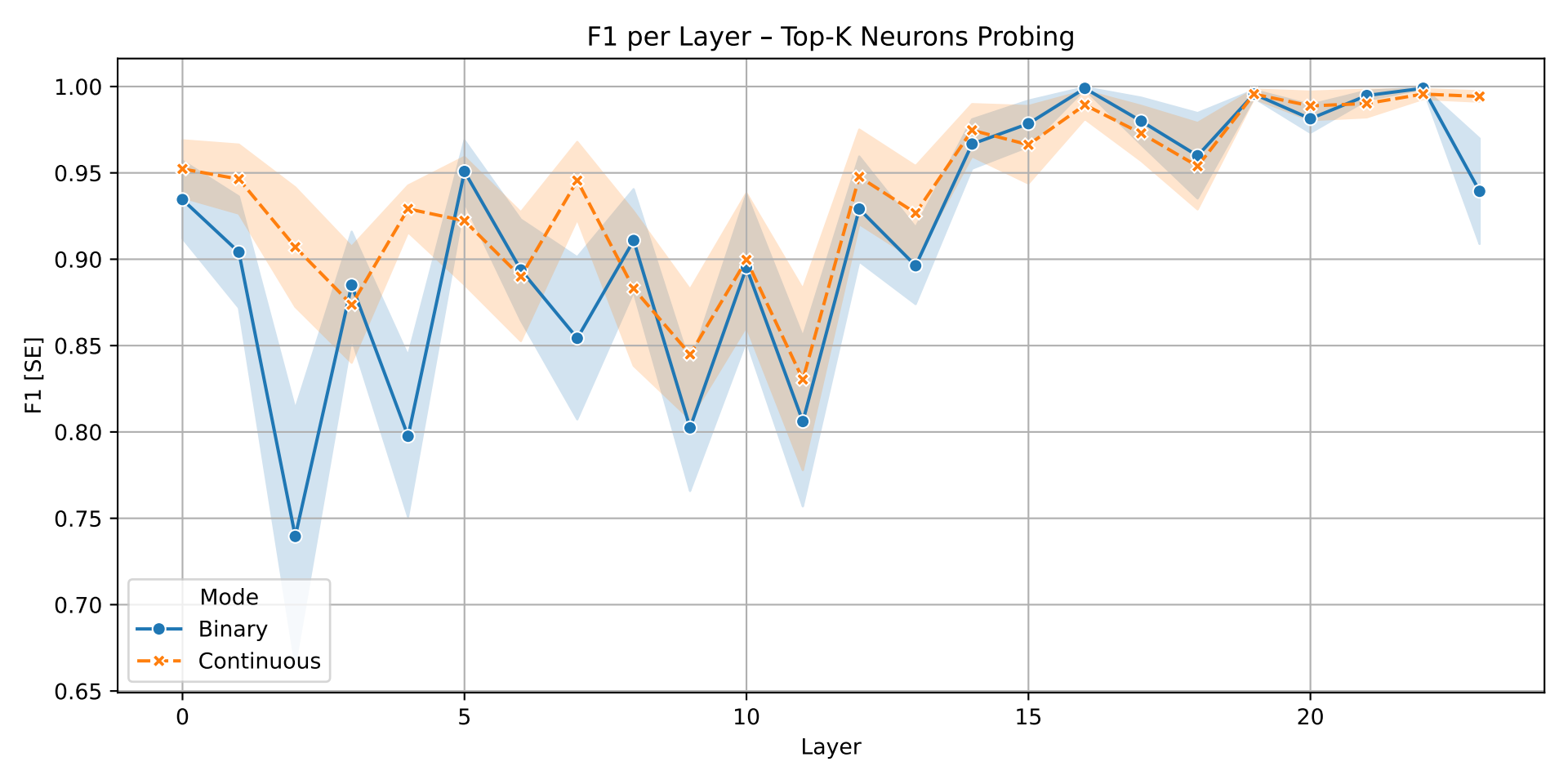
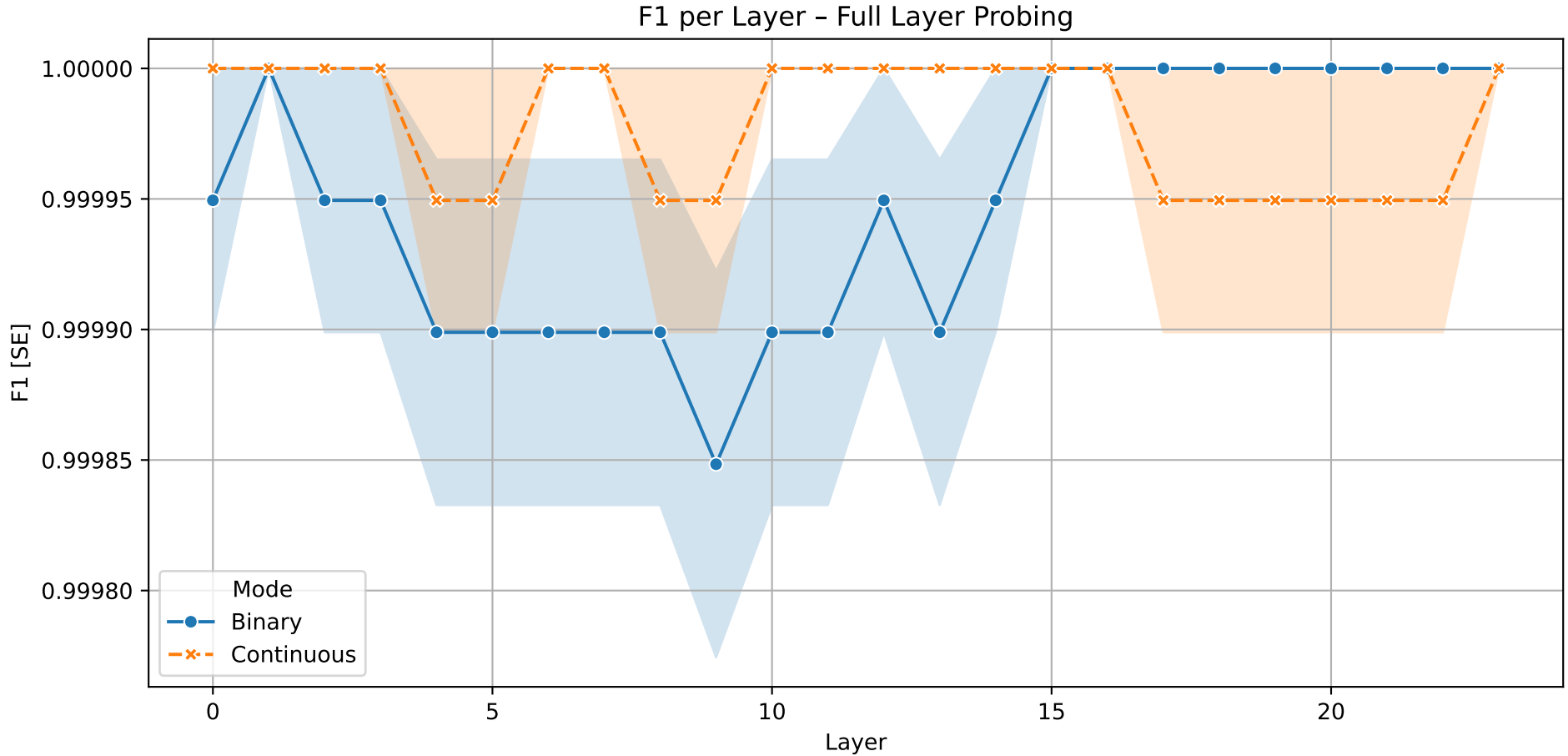
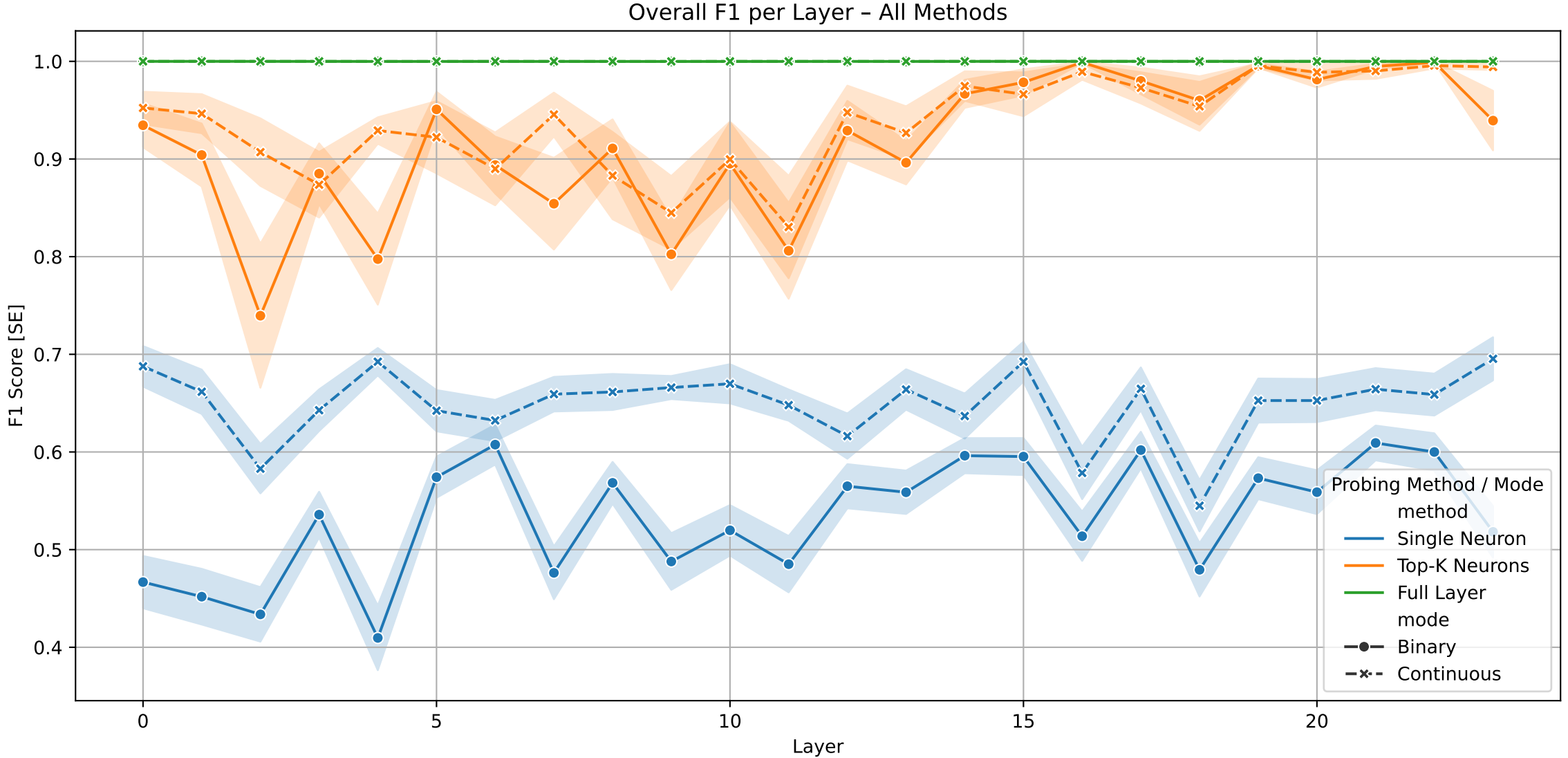

F1 per Layer – Single Neuron Probing
F1 [SE]
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0
5
10
15
20
Layer
Mode
Binary
Continuous

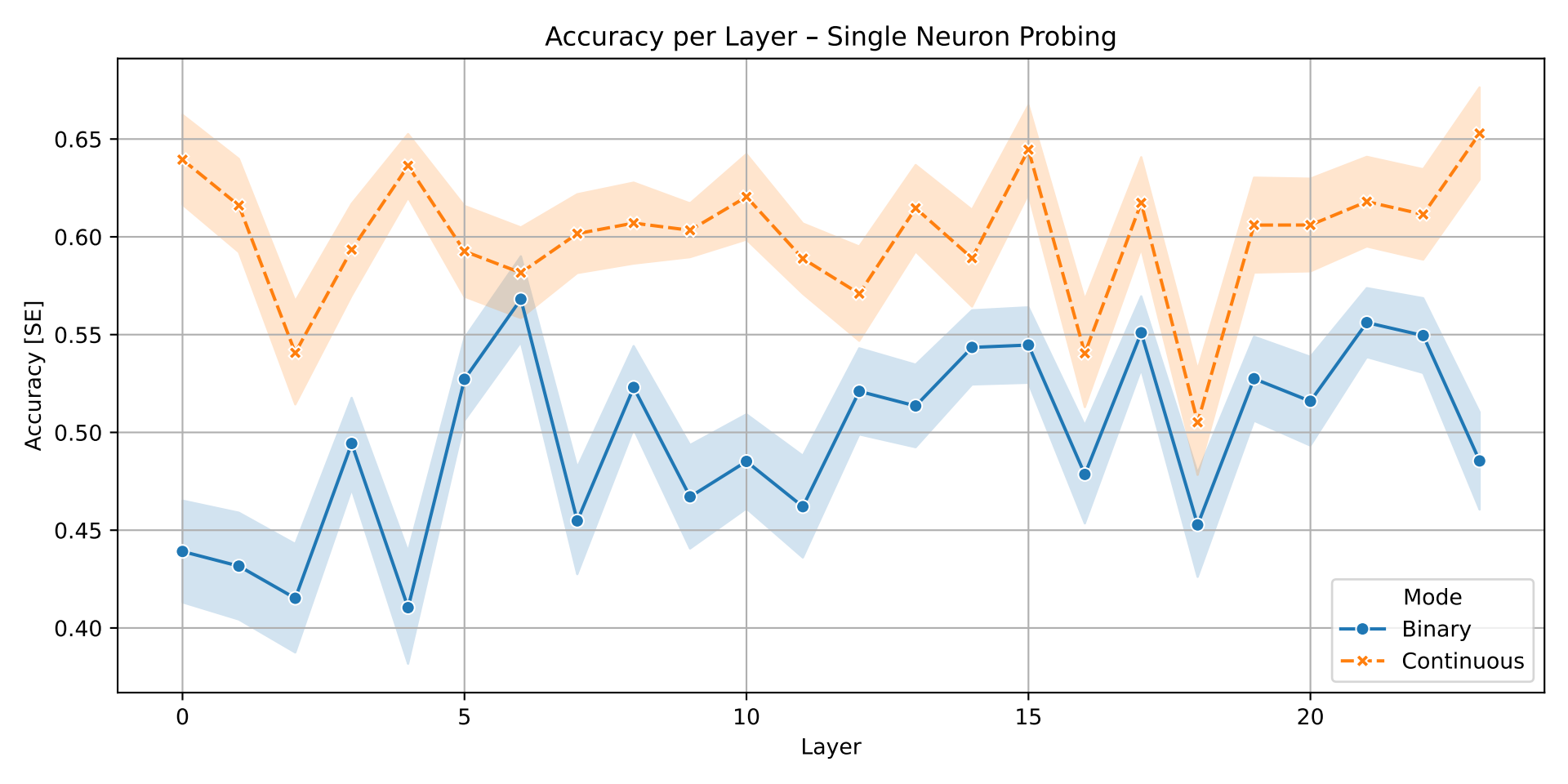

F1 per Layer – Top-K Neurons Probing
F1 [SE]
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0
5
10
15
20
Layer
Mode
Binary
Continuous

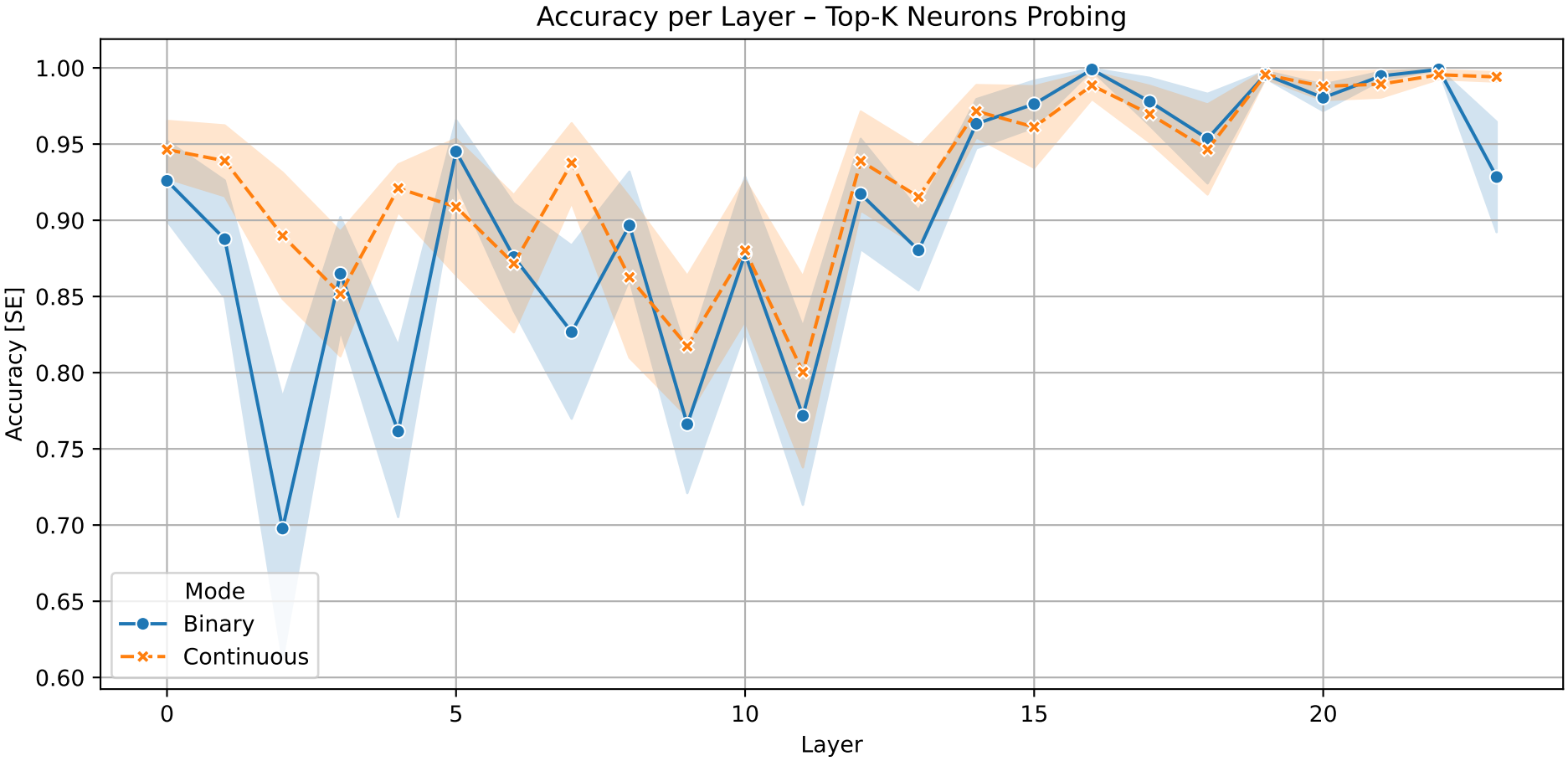

F1 per Layer – Full Layer Probing
1.00000
0.99995
0.99990
0.99985
0.99980
F1 [SE]
0
5
10
15
20
Layer
Mode
Binary
Continuous

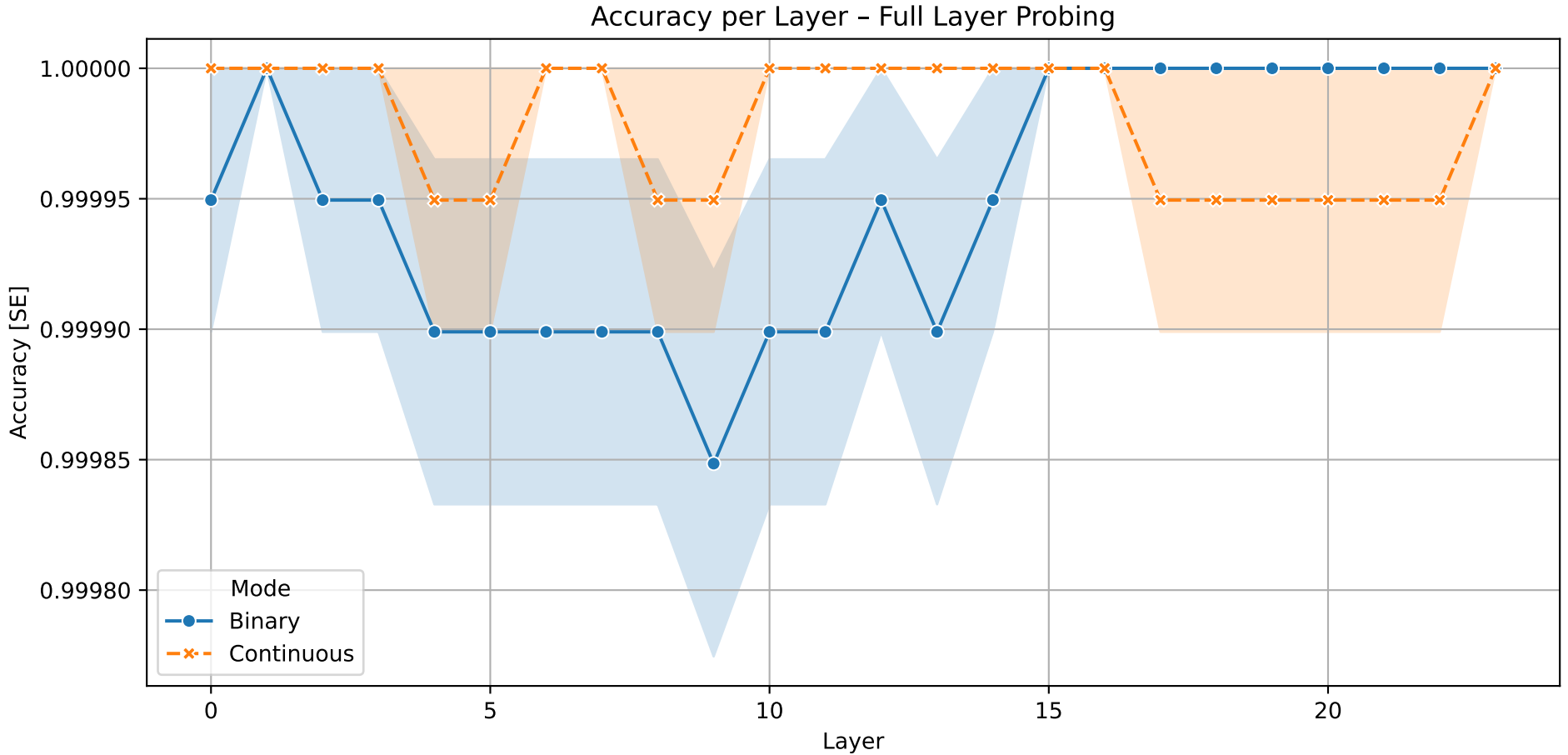

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
5
10
15
20
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

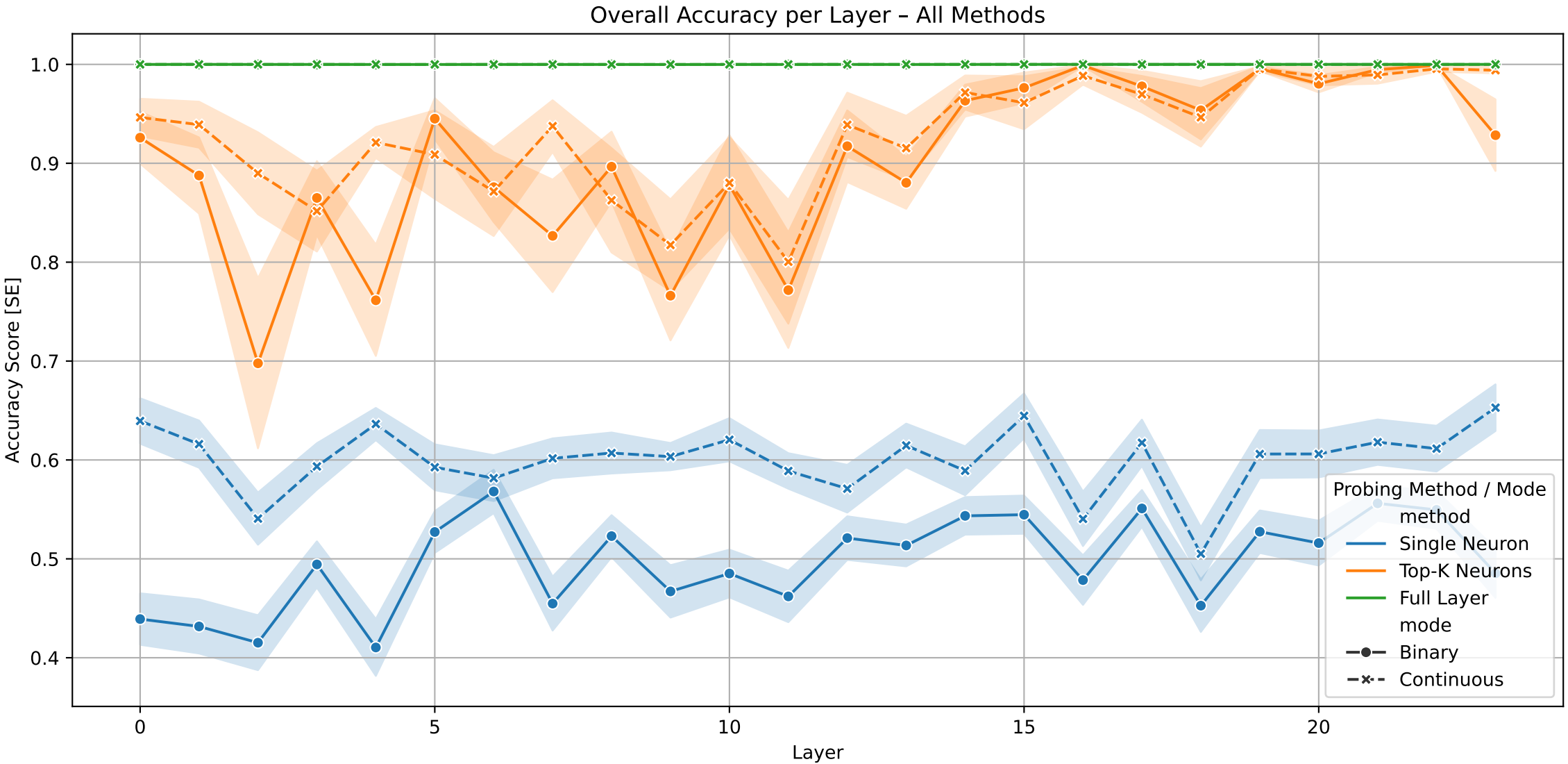
F1 Score Summary by Probing Method

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	0.9999	1.0
Full Layer	f1_std	0.0001	0.0001
Single Neuron	f1_best_layer	21.0	23.0
Single Neuron	f1_max	1.0	1.0
Single Neuron	f1_mean	0.5328	0.6487
Single Neuron	f1_std	0.2233	0.1918
Top-K Neurons	f1_best_layer	22.0	22.0
Top-K Neurons	f1_max	1.0	1.0
Top-K Neurons	f1_mean	0.9164	0.9383
Top-K Neurons	f1_std	0.1095	0.0859


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


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
Layer
Mode
Binary
Continuous
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0
5
10
15
20


Accuracy per Layer – Full Layer Probing
1.00000
0.99995
0.99990
0.99985
0.99980
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
0.4
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	0.9999	1.0
Full Layer	accuracy_std	0.0001	0.0001
Single Neuron	accuracy_best_layer	6.0	23.0
Single Neuron	accuracy_max	1.0	1.0
Single Neuron	accuracy_mean	0.4966	0.5999
Single Neuron	accuracy_std	0.2138	0.2055
Top-K Neurons	accuracy_best_layer	22.0	19.0
Top-K Neurons	accuracy_max	1.0	1.0
Top-K Neurons	accuracy_mean	0.9026	0.9284
Top-K Neurons	accuracy_std	0.1298	0.1023