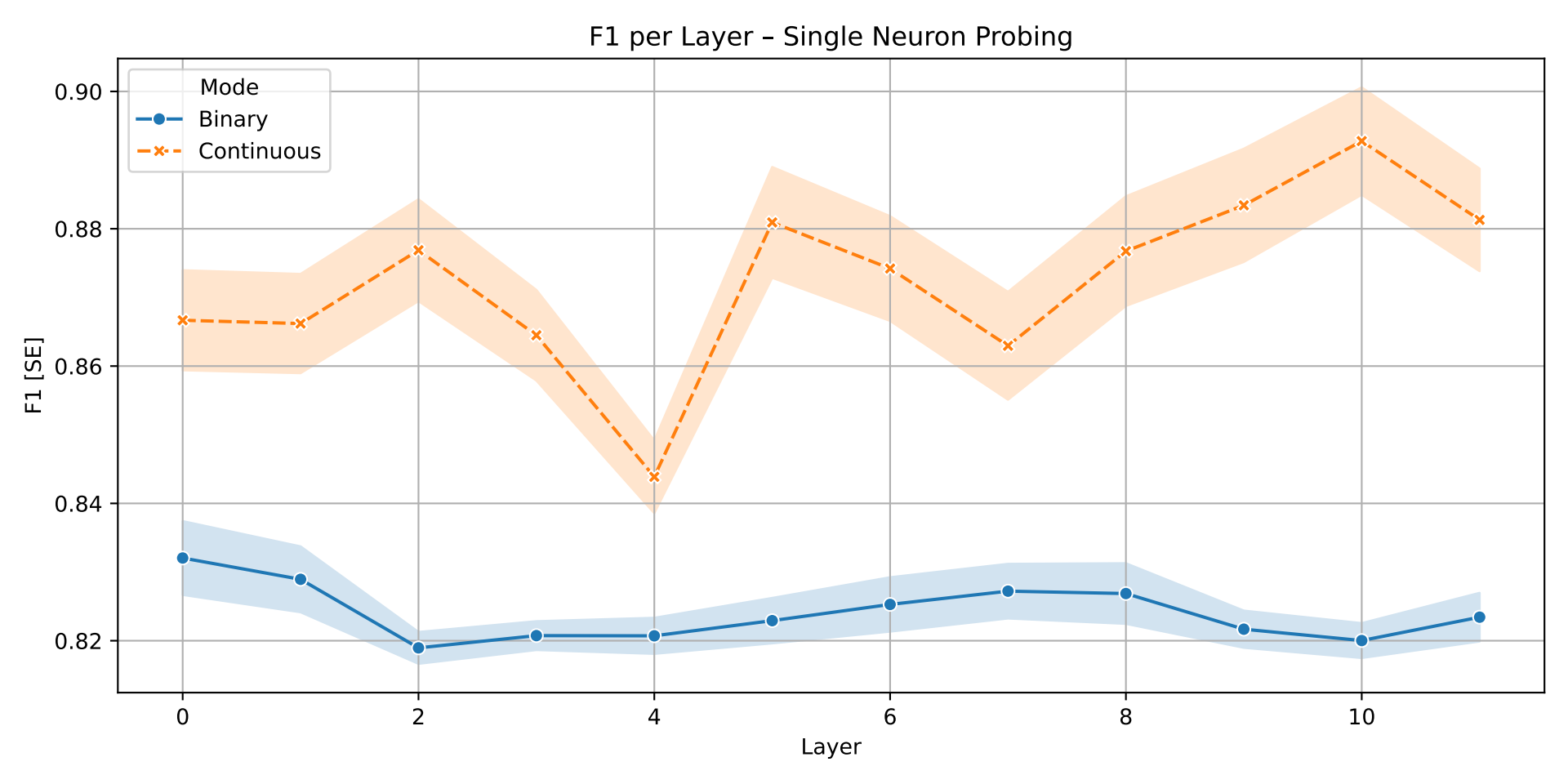
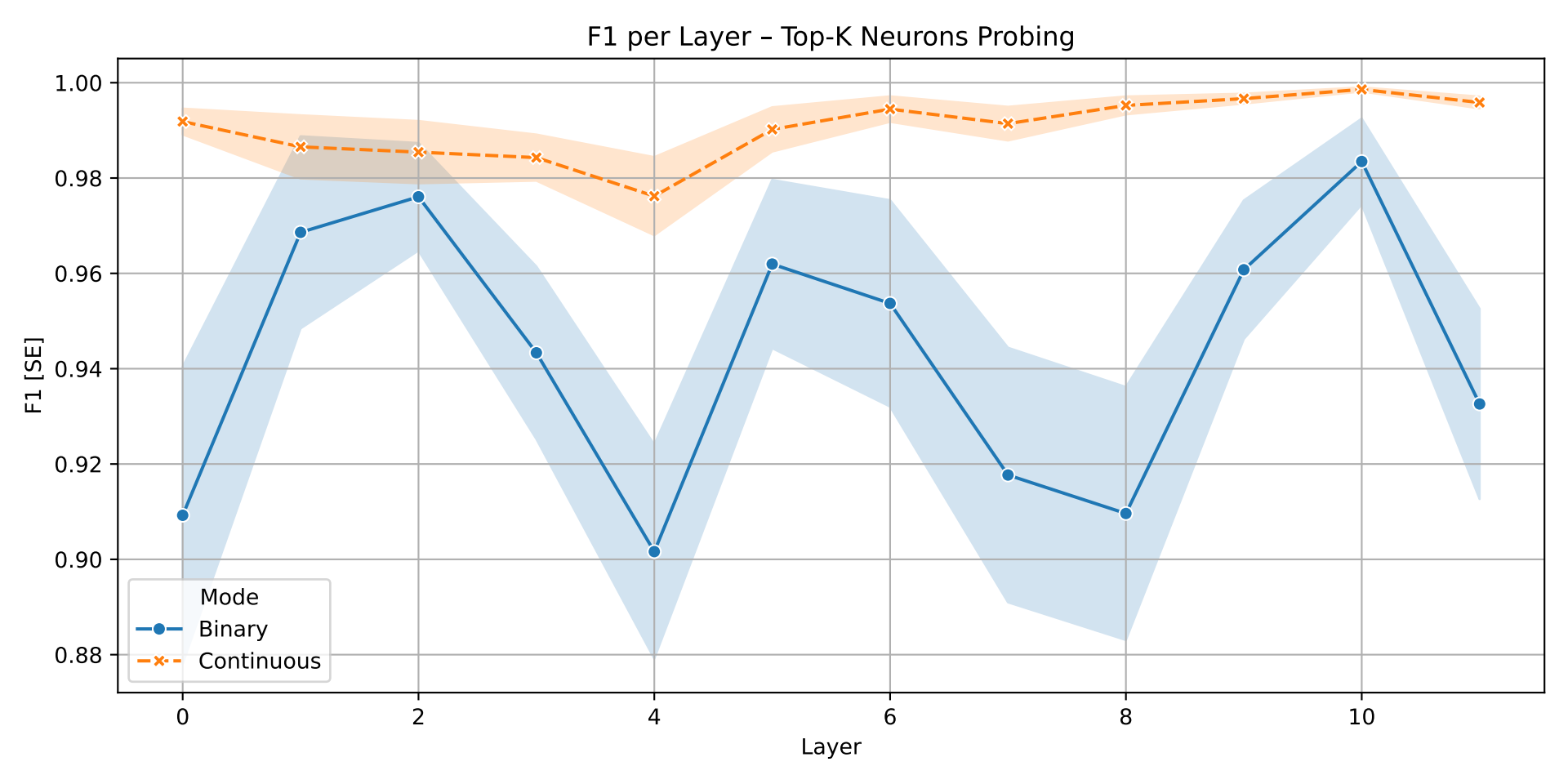
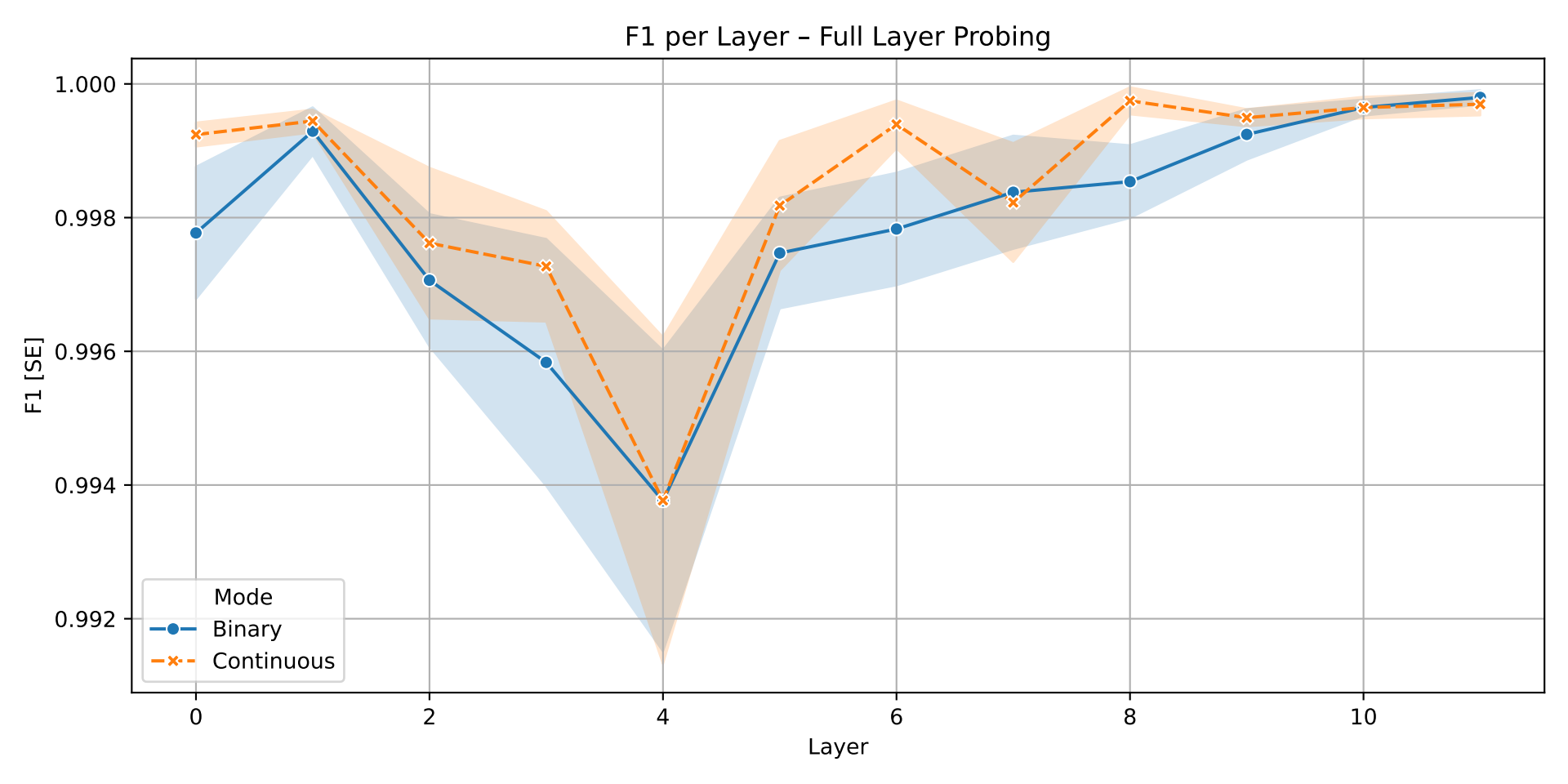
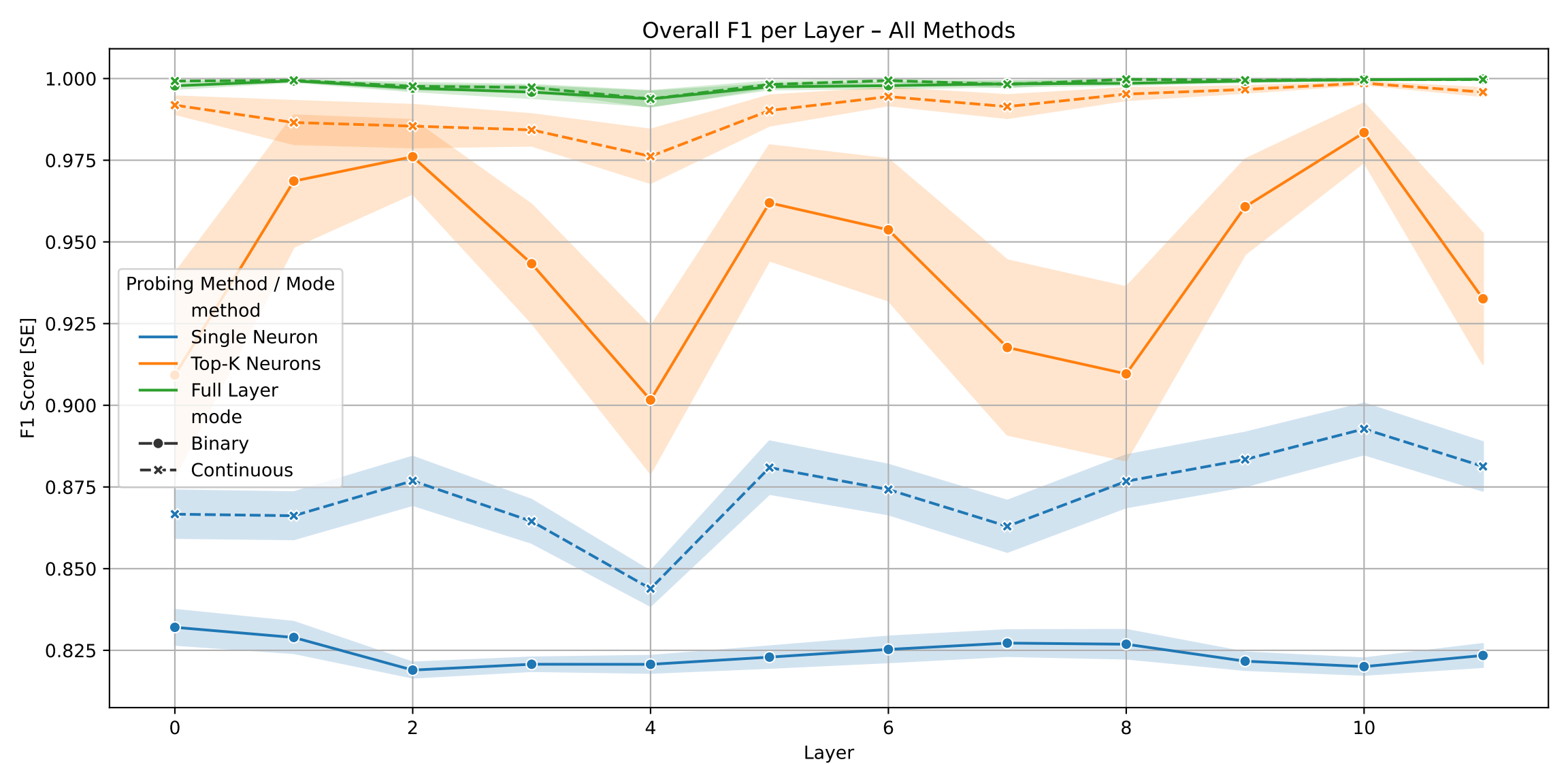

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.90
0.88
0.86
0.84
0.82
0
2
4
6
8
10
Layer

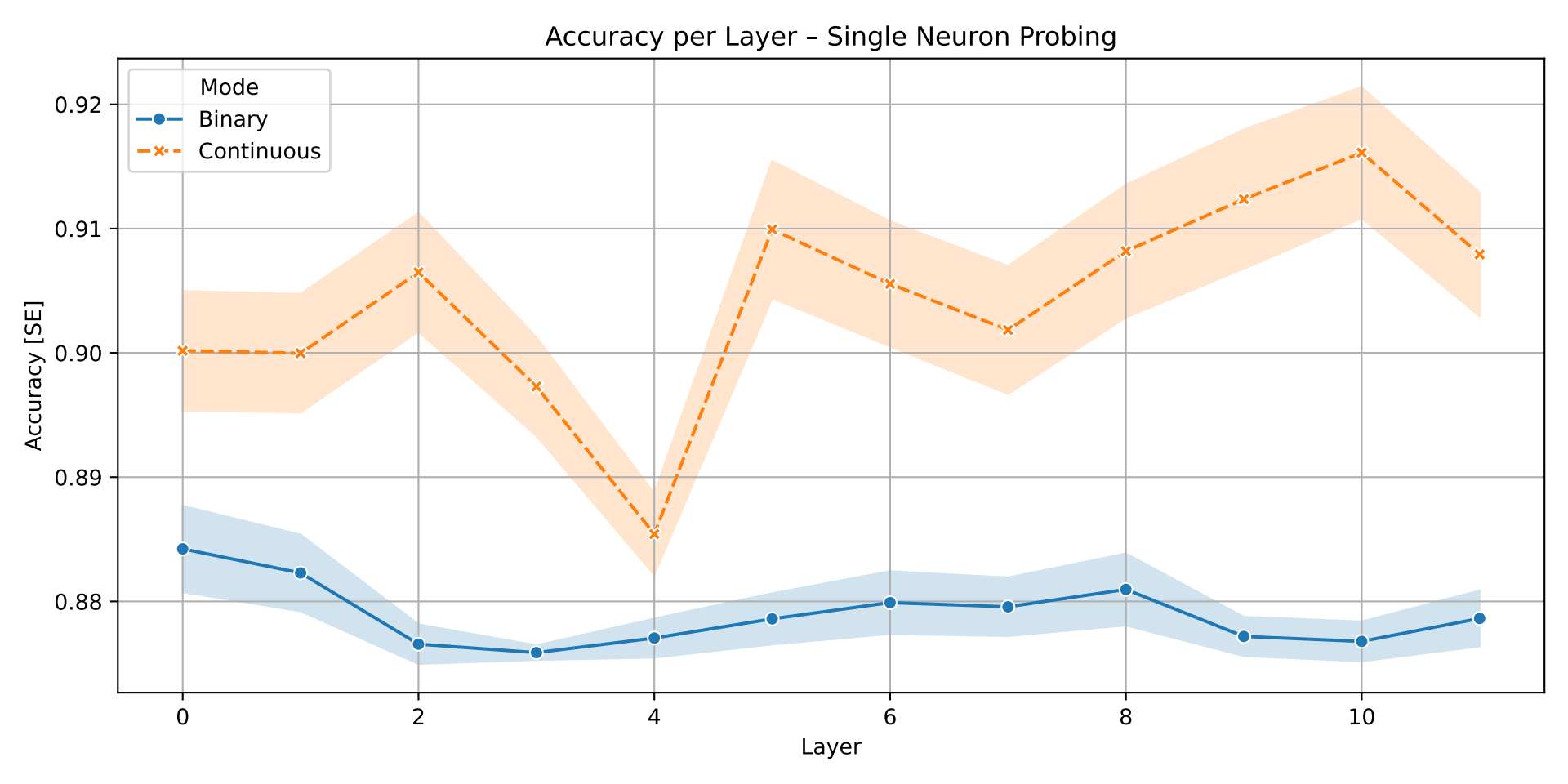

F1 per Layer – Top-K Neurons Probing
F1 [SE]
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0
2
4
6
8
10
Layer
Mode
Binary
Continuous

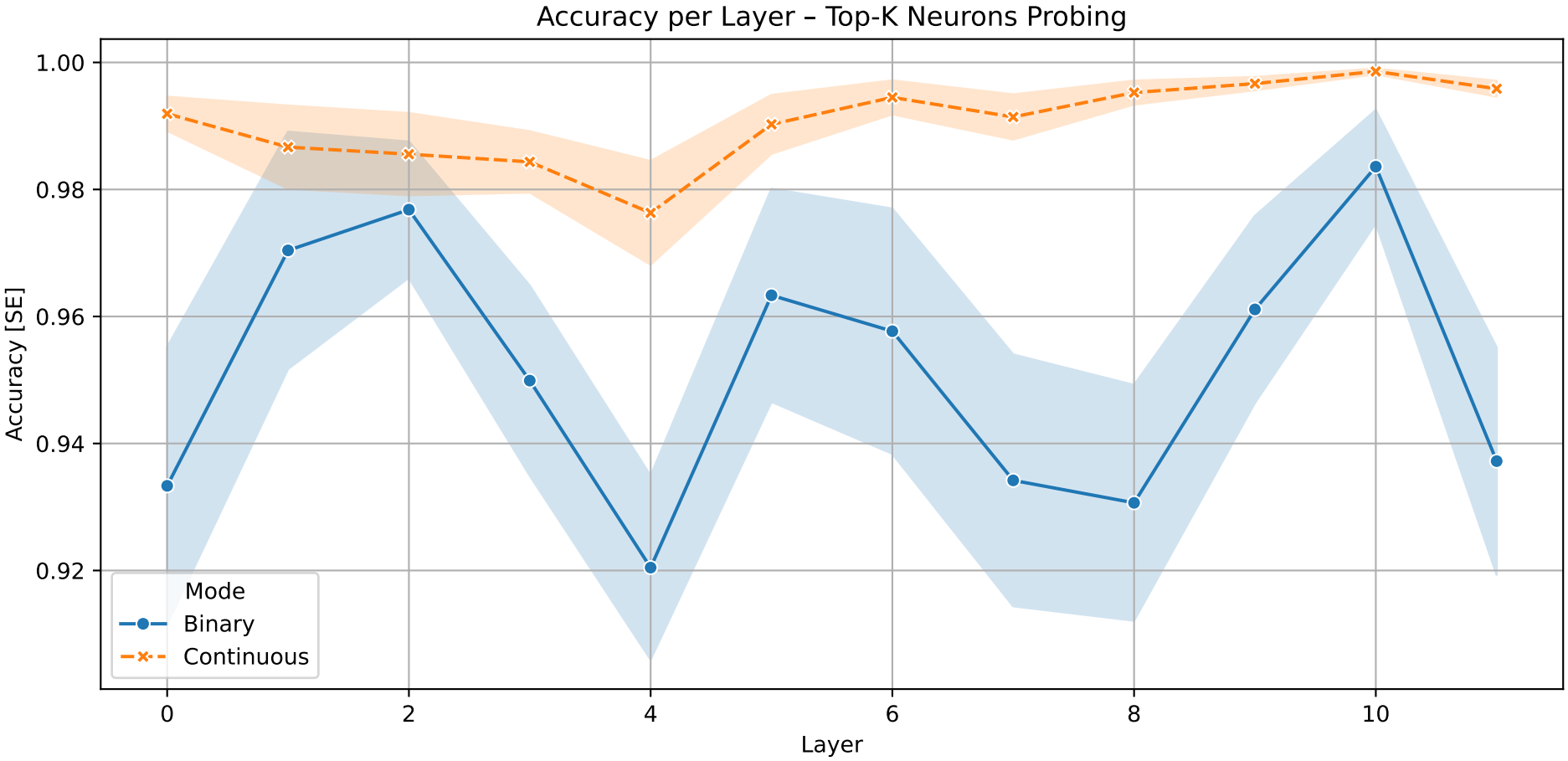

F1 per Layer – Full Layer Probing
1.000
0.998
0.996
0.994
0.992
F1 [SE]
0
2
4
6
8
10
Layer
Mode
Binary
Continuous

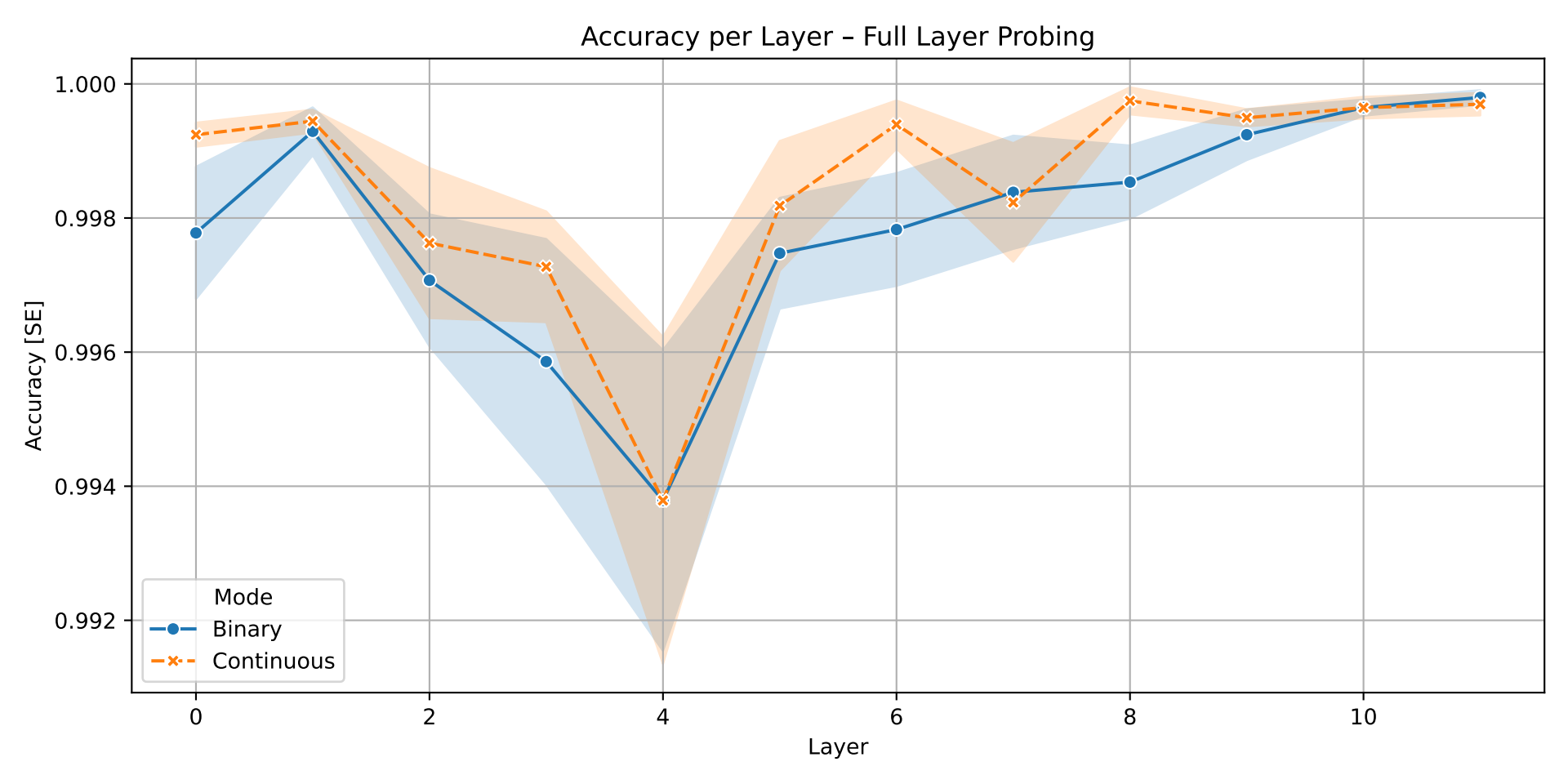

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0
2
4
6
8
10
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

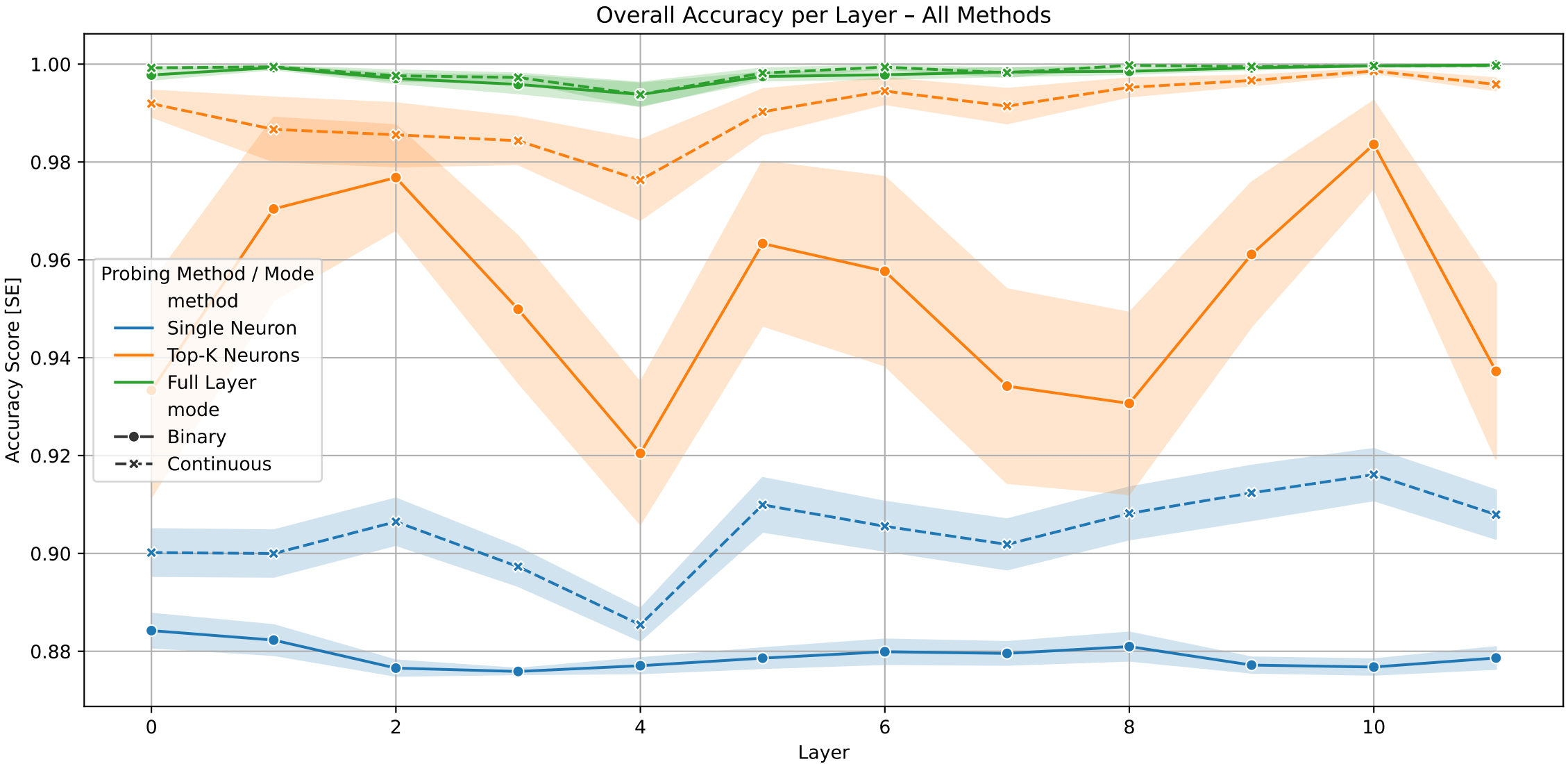
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	11.0	8.0
Full Layer	f1_max	1.0	1.0
Full Layer	f1_mean	0.9979	0.9985
Full Layer	f1_std	0.0032	0.0029
Single Neuron	f1_best_layer	0.0	10.0
Single Neuron	f1_max	1.0	1.0
Single Neuron	f1_mean	0.8241	0.8725
Single Neuron	f1_std	0.0324	0.0673
Top-K Neurons	f1_best_layer	10.0	10.0
Top-K Neurons	f1_max	1.0	1.0
Top-K Neurons	f1_mean	0.9432	0.9906
Top-K Neurons	f1_std	0.0619	0.0132


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
0.92
0.91
0.90
0.89
0.88
Accuracy [SE]
0
2
4
6
8
10
Layer


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
1.00
0.98
0.96
0.94
0.92
0
2
4
6
8
10
Layer
Mode
Binary
Continuous


Accuracy per Layer – Full Layer Probing
1.000
0.998
0.996
0.994
0.992
Accuracy [SE]
0
2
4
6
8
10
Layer
Mode
Binary
Continuous


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
Layer
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0
2
4
6
8
10
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	11.0	8.0
Full Layer	accuracy_max	1.0	1.0
Full Layer	accuracy_mean	0.9979	0.9985
Full Layer	accuracy_std	0.0032	0.0029
Single Neuron	accuracy_best_layer	0.0	10.0
Single Neuron	accuracy_max	1.0	1.0
Single Neuron	accuracy_mean	0.879	0.9043
Single Neuron	accuracy_std	0.0202	0.0443
Top-K Neurons	accuracy_best_layer	10.0	10.0
Top-K Neurons	accuracy_max	1.0	1.0
Top-K Neurons	accuracy_mean	0.9516	0.9906
Top-K Neurons	accuracy_std	0.0487	0.0131