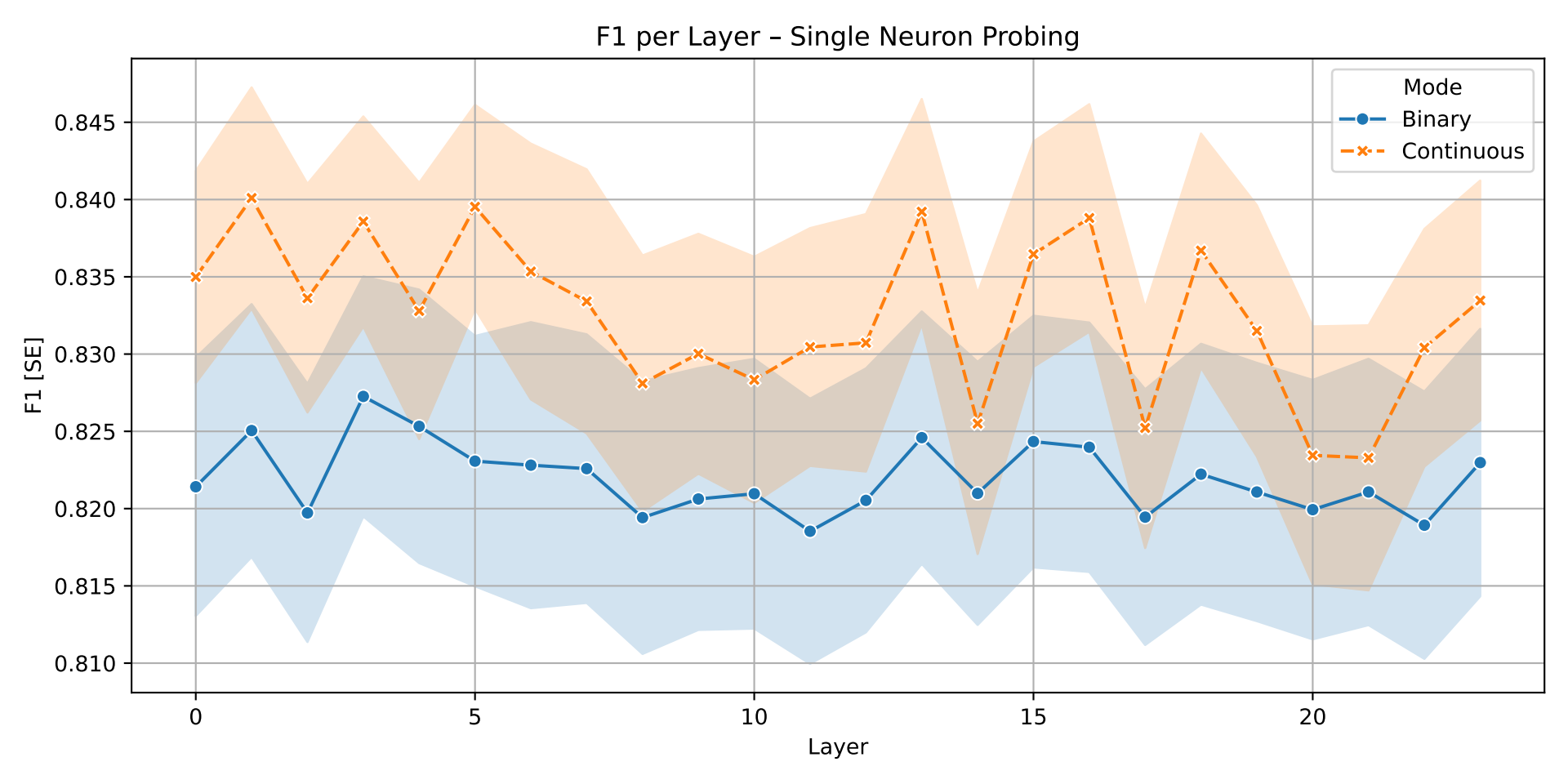
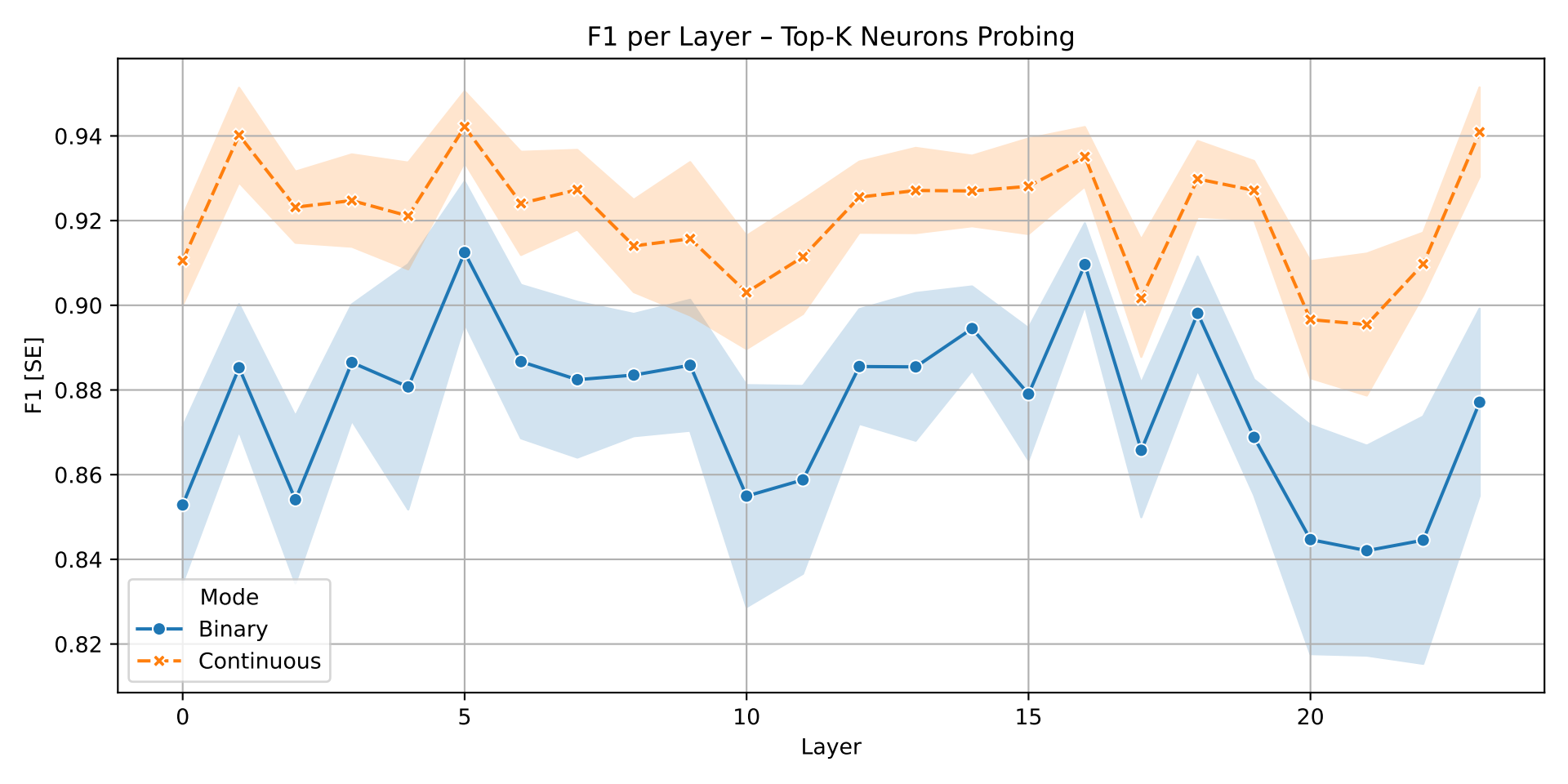
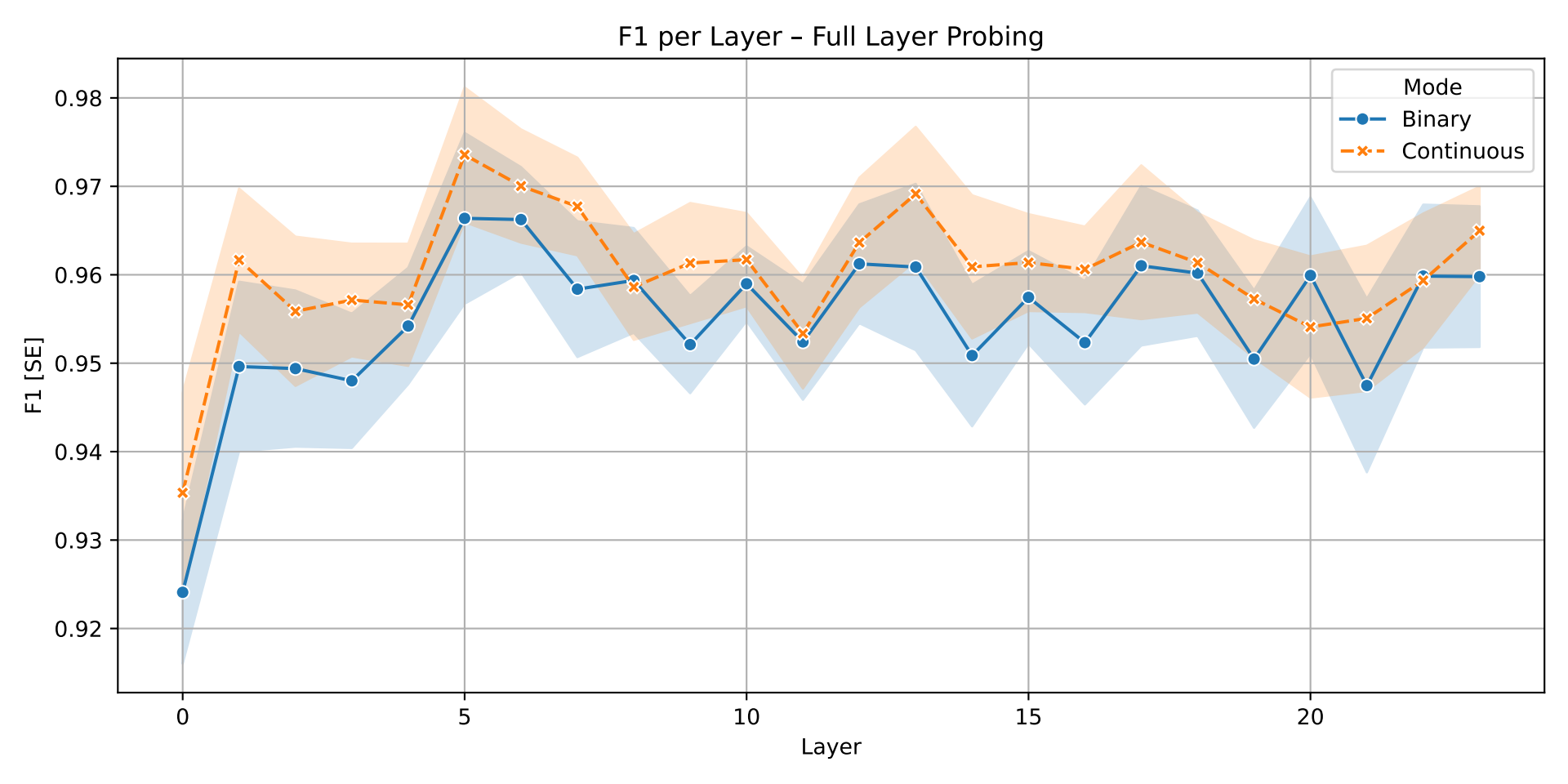
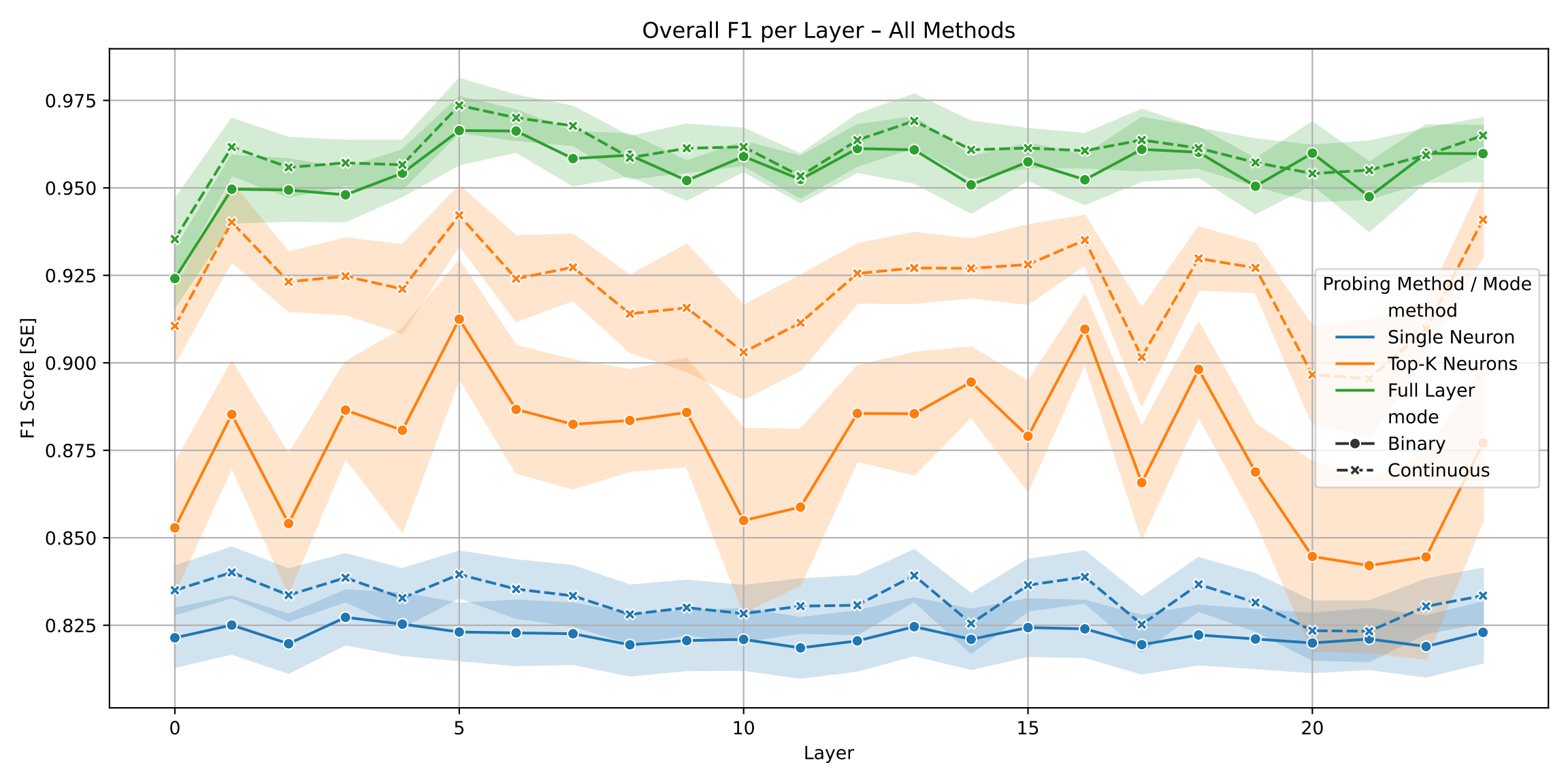

F1 per Layer – Single Neuron Probing
F1 [SE]
Layer
Mode
Binary
Continuous
0.845
0.840
0.835
0.830
0.825
0.820
0.815
0.810
0
5
10
15
20

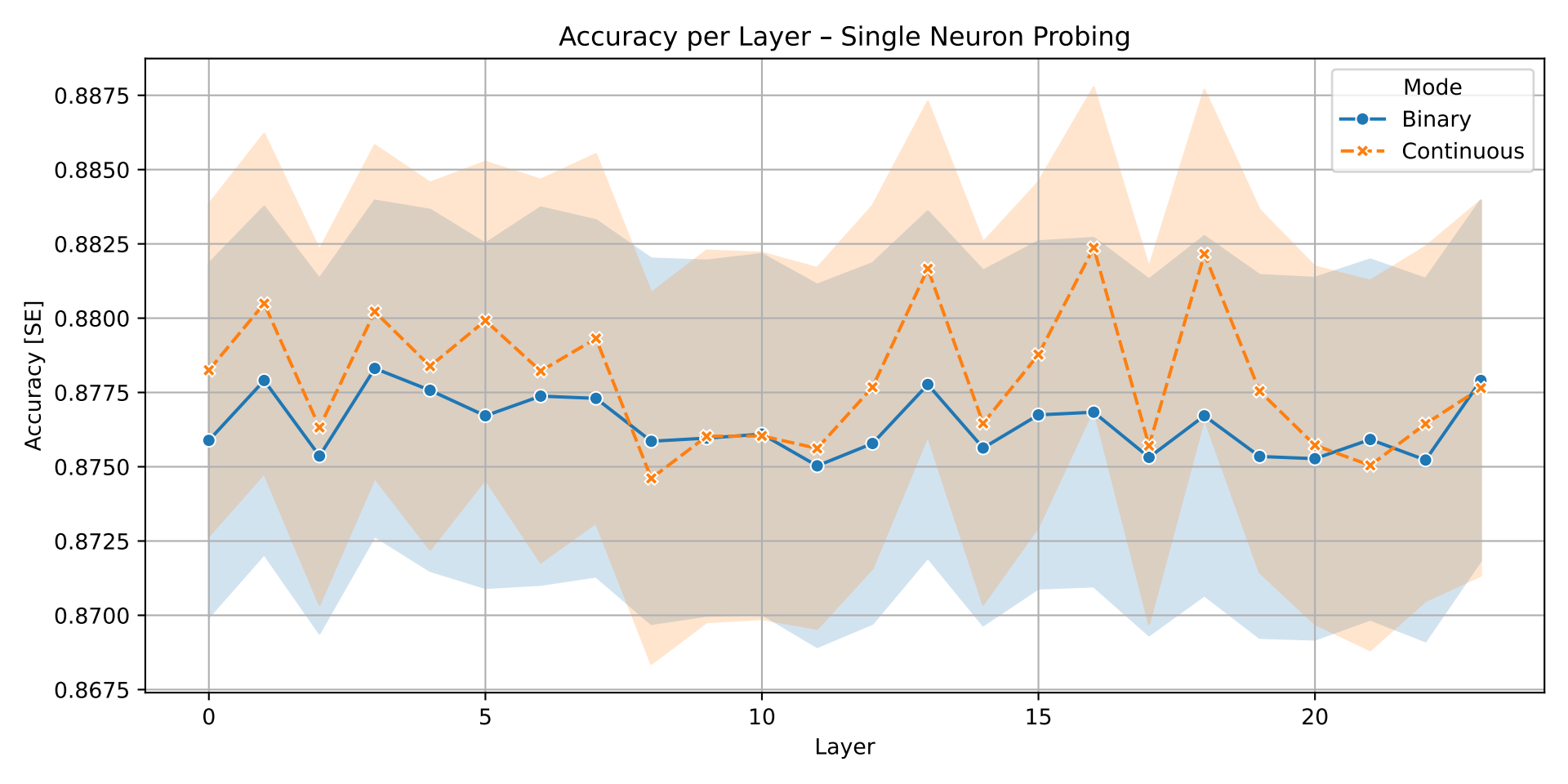

F1 per Layer – Top-K Neurons Probing
F1 [SE]
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0
5
10
15
20
Layer
Mode
Binary
Continuous

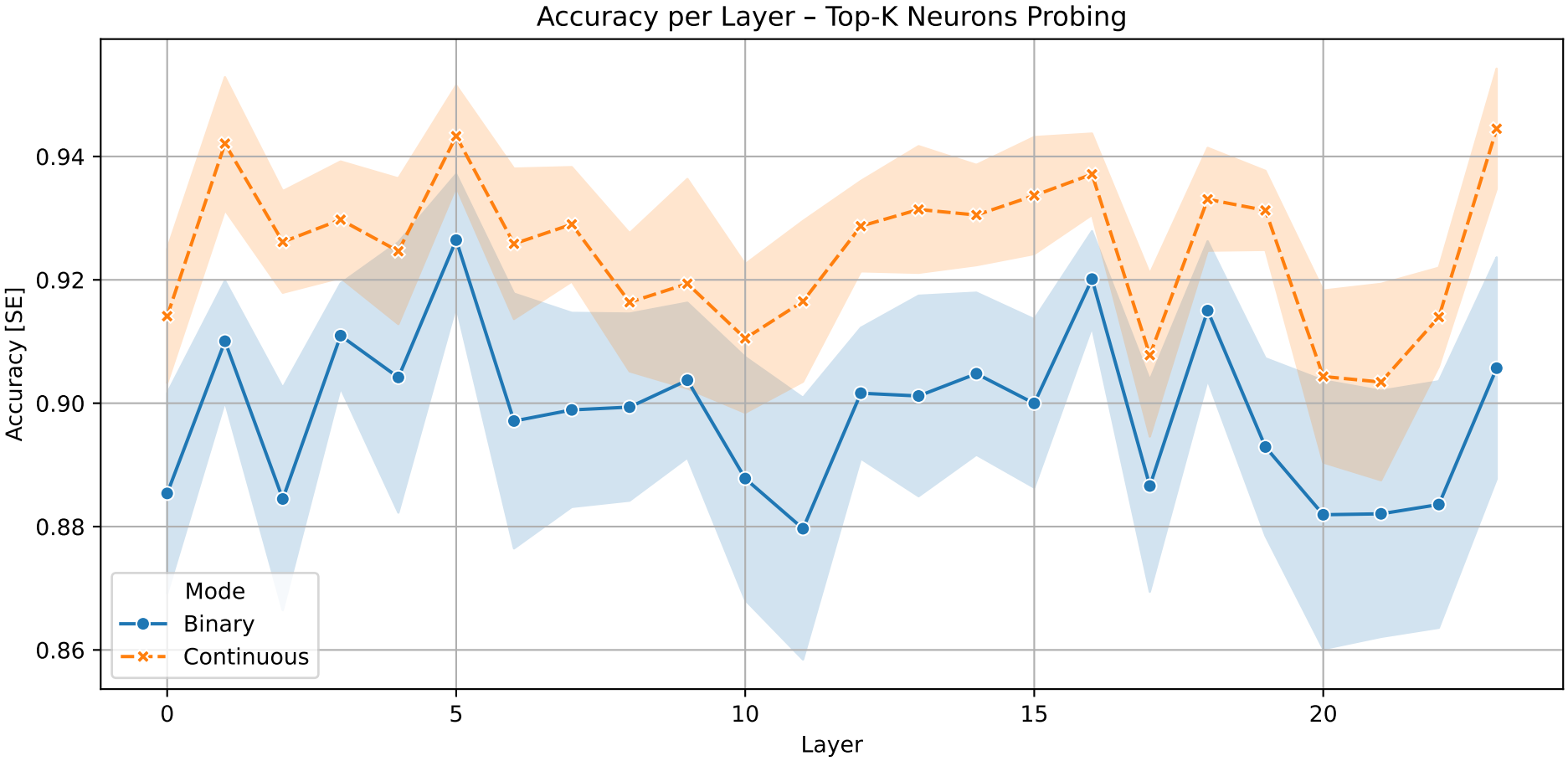

F1 per Layer – Full Layer Probing
F1 [SE]
Layer
Mode
Binary
Continuous
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0
5
10
15
20

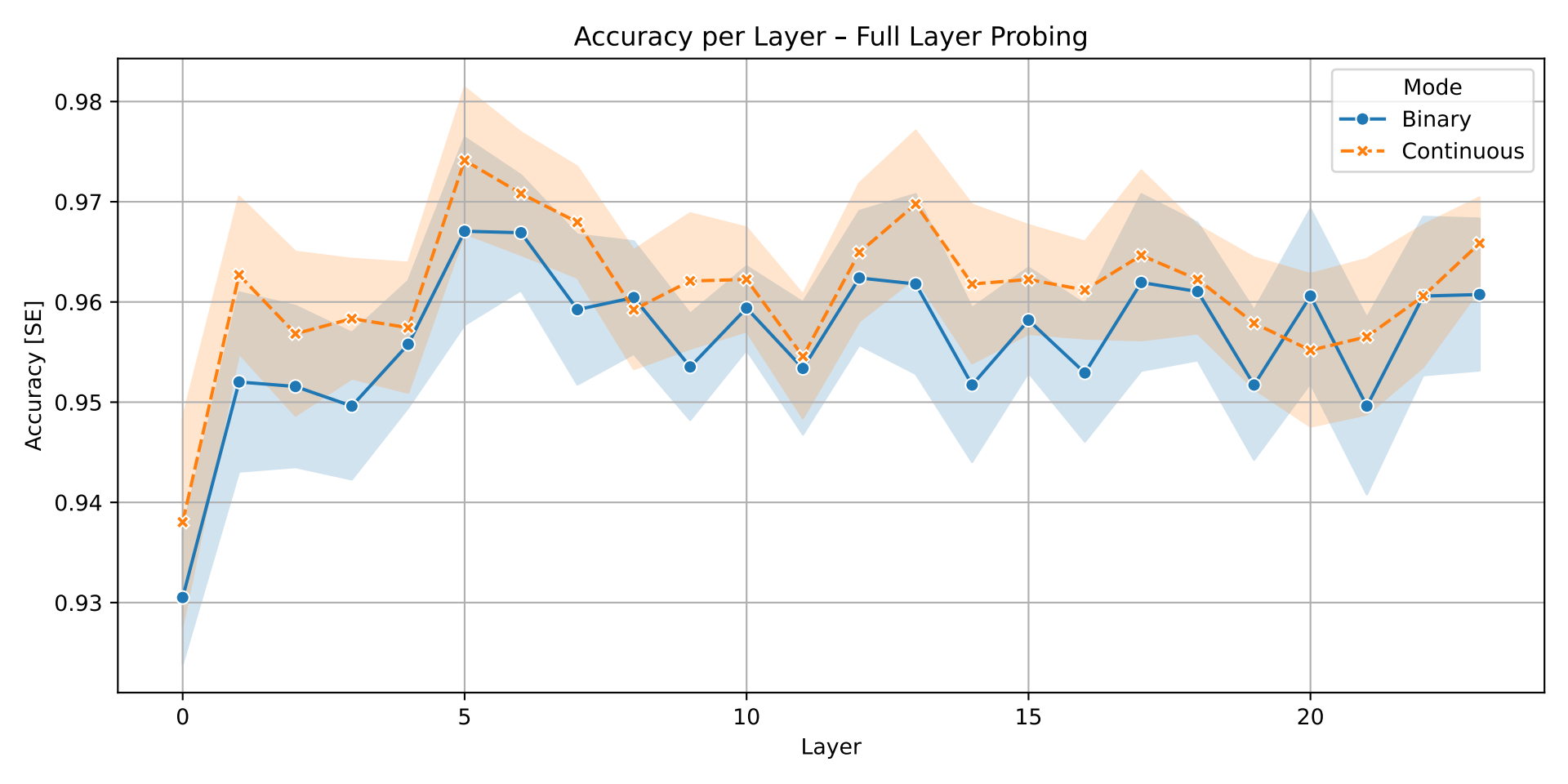

Overall F1 per Layer – All Methods
F1 Score [SE]
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0
5
10
15
20
Layer
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

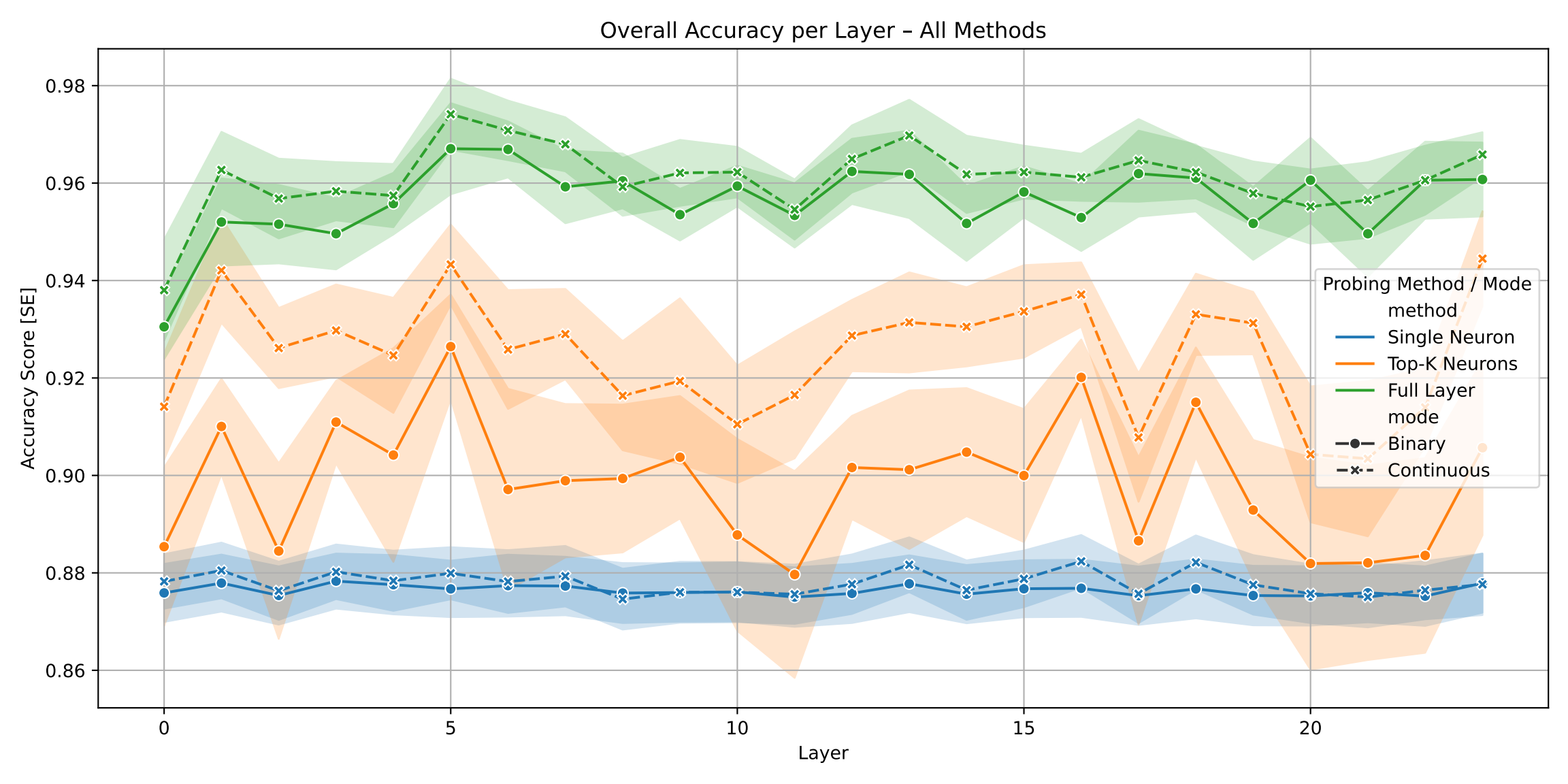
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	5.0	5.0
Full Layer	f1_max	0.9988	0.9988
Full Layer	f1_mean	0.955	0.9602
Full Layer	f1_std	0.0221	0.0203
Single Neuron	f1_best_layer	3.0	1.0
Single Neuron	f1_max	0.9833	0.9876
Single Neuron	f1_mean	0.822	0.8325
Single Neuron	f1_std	0.0753	0.0693
Top-K Neurons	f1_best_layer	5.0	5.0
Top-K Neurons	f1_max	0.9857	0.9903
Top-K Neurons	f1_mean	0.8758	0.9209
Top-K Neurons	f1_std	0.0543	0.0327


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.8875
0.8850
0.8825
0.8800
0.8775
0.8750
0.8725
0.8700
0.8675
0
5
10
15
20
Layer


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
0.94
0.92
0.90
0.88
0.86
0
5
10
15
20
Layer
Mode
Binary
Continuous


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.98
0.97
0.96
0.95
0.94
0.93
0
5
10
15
20
Layer


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
Layer
0.98
0.96
0.94
0.92
0.90
0.88
0.86
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	5.0	5.0
Full Layer	accuracy_max	0.9988	0.9988
Full Layer	accuracy_mean	0.9564	0.9611
Full Layer	accuracy_std	0.0208	0.0195
Single Neuron	accuracy_best_layer	3.0	16.0
Single Neuron	accuracy_max	0.9832	0.988
Single Neuron	accuracy_mean	0.8764	0.8779
Single Neuron	accuracy_std	0.0533	0.053
Top-K Neurons	accuracy_best_layer	5.0	23.0
Top-K Neurons	accuracy_max	0.9856	0.9904
Top-K Neurons	accuracy_mean	0.8985	0.9249
Top-K Neurons	accuracy_std	0.0447	0.031