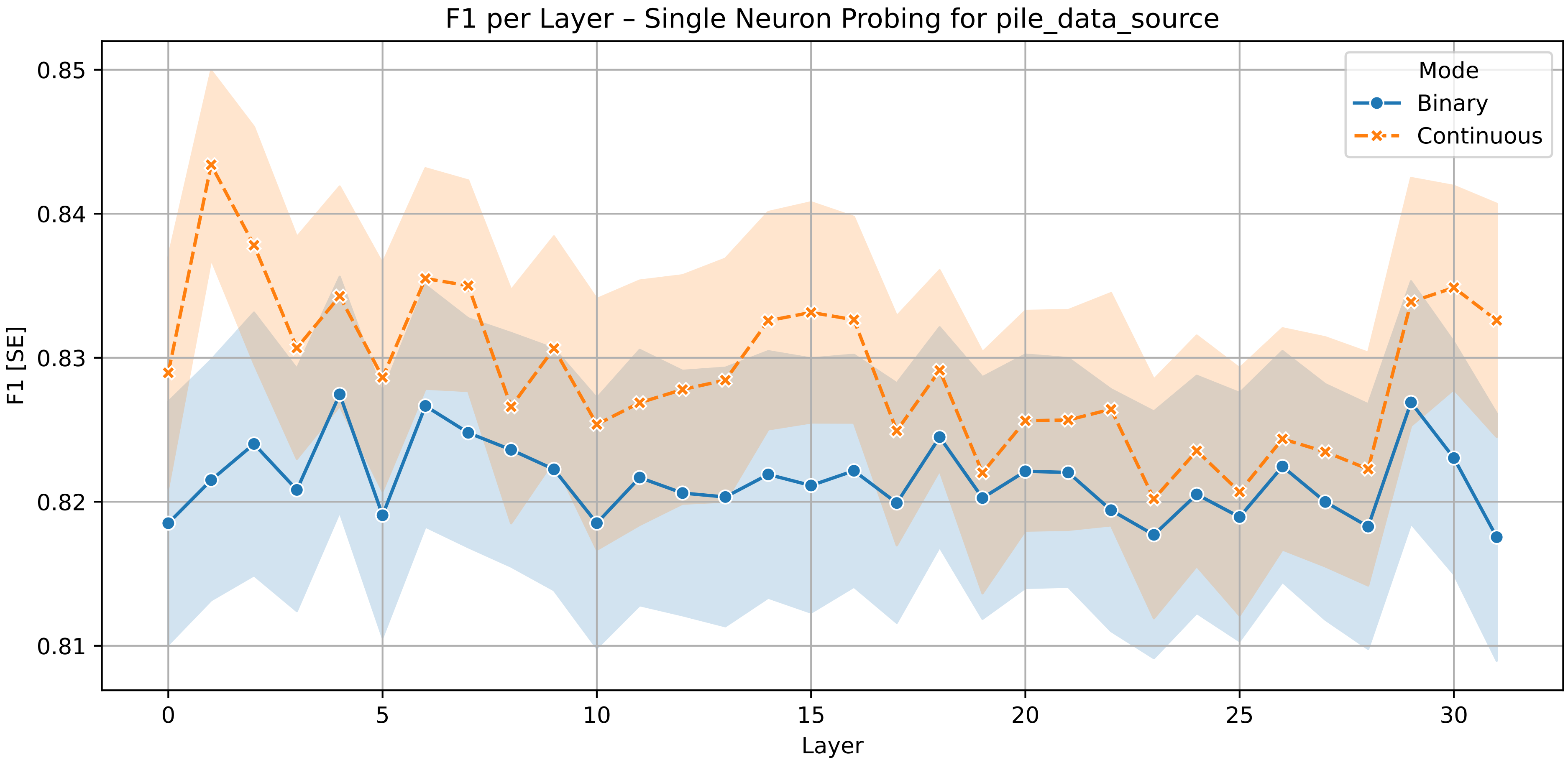
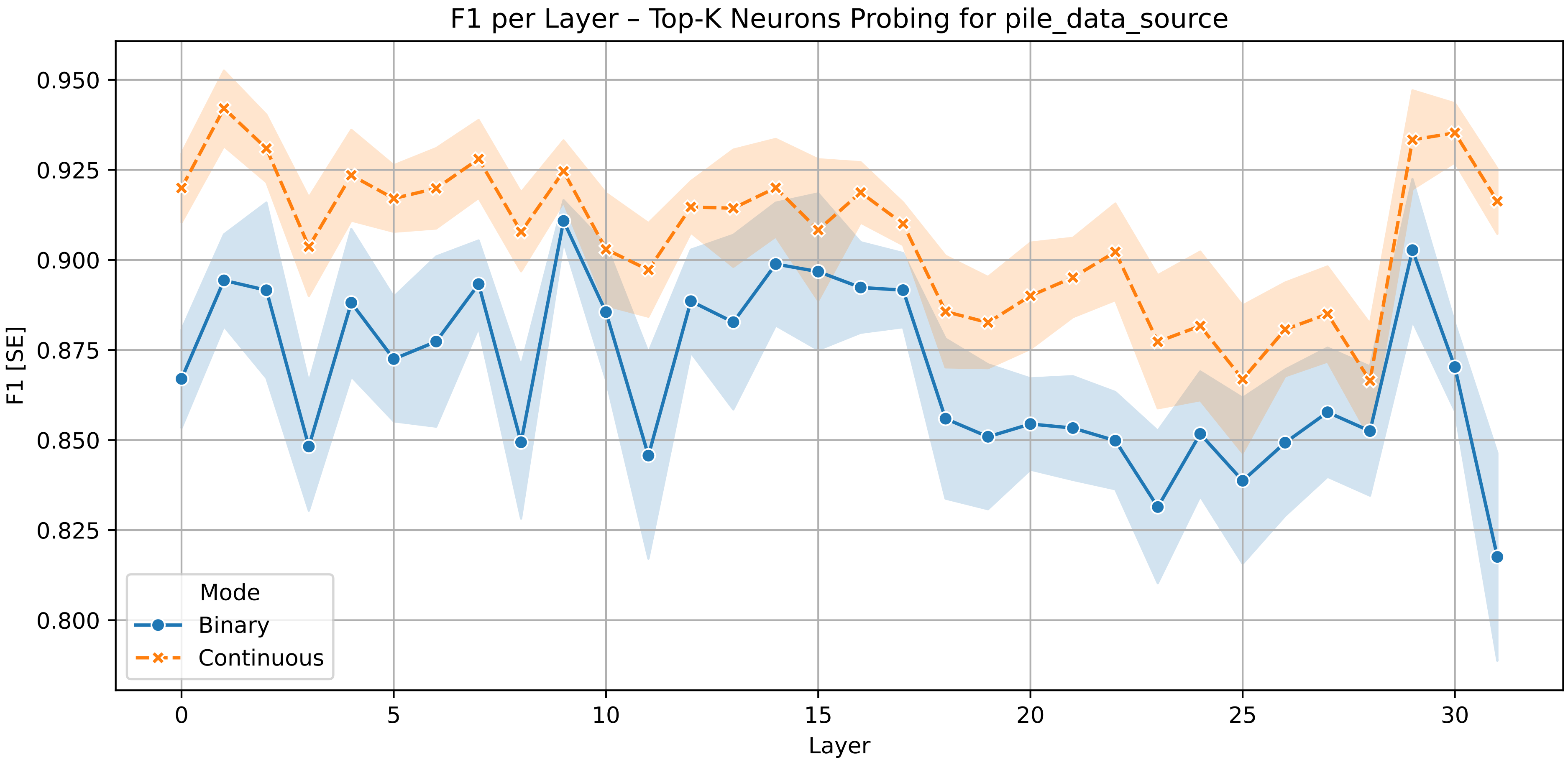
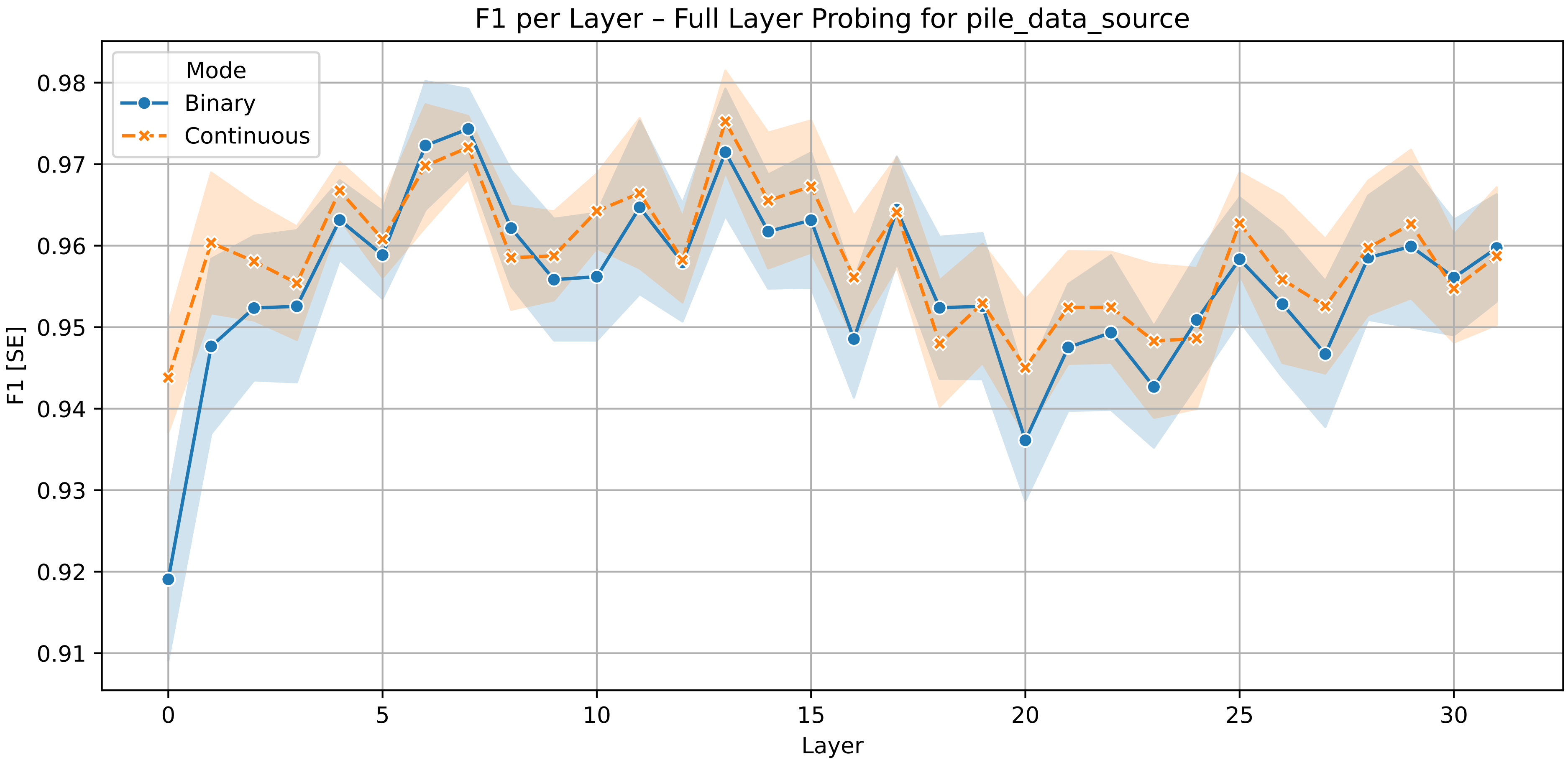
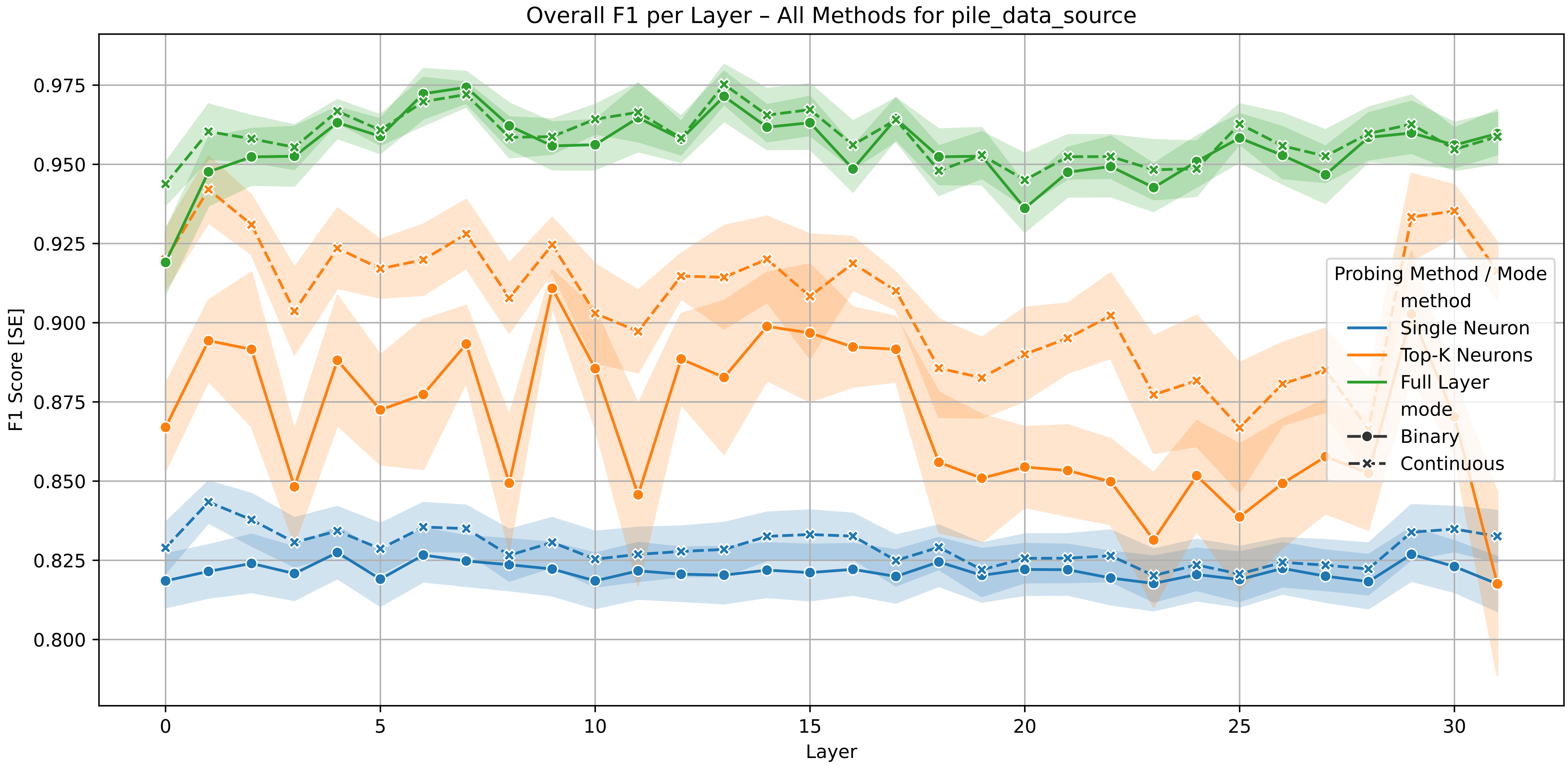
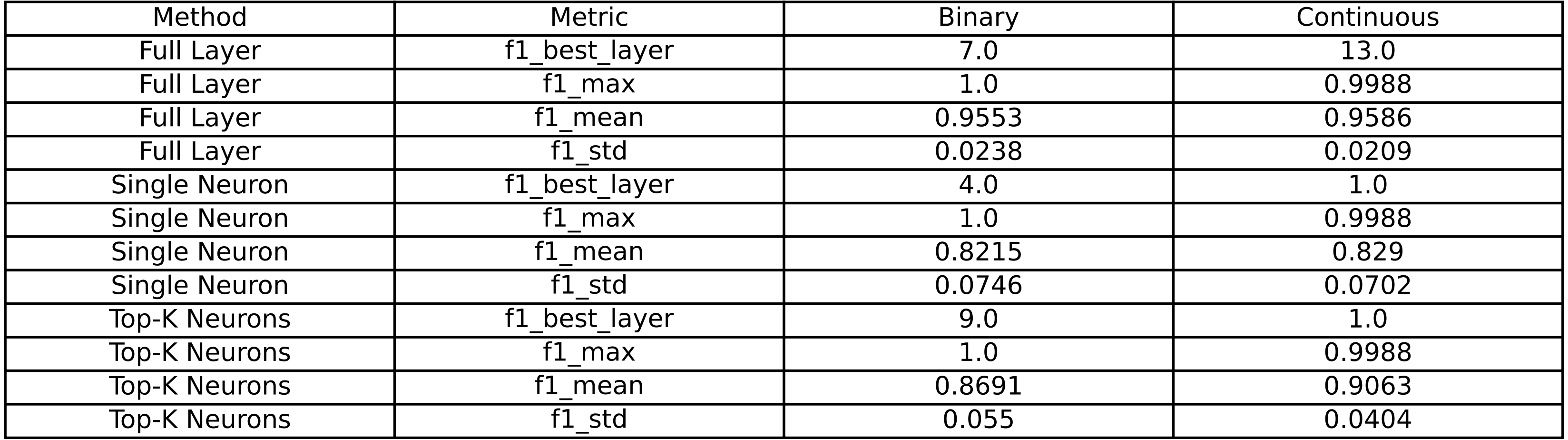

F1 per Layer – Single Neuron Probing for pile_data_source
F1 [SE]
0.85
0.84
0.83
0.82
0.81
Layer
0
5
10
15
20
25
30
Mode
Binary
Continuous


F1 per Layer – Top-K Neurons Probing for pile_data_source
F1 [SE]
0.950
0.925
0.900
0.875
0.850
0.825
0.800
Layer
0
5
10
15
20
25
30
Mode
Binary
Continuous

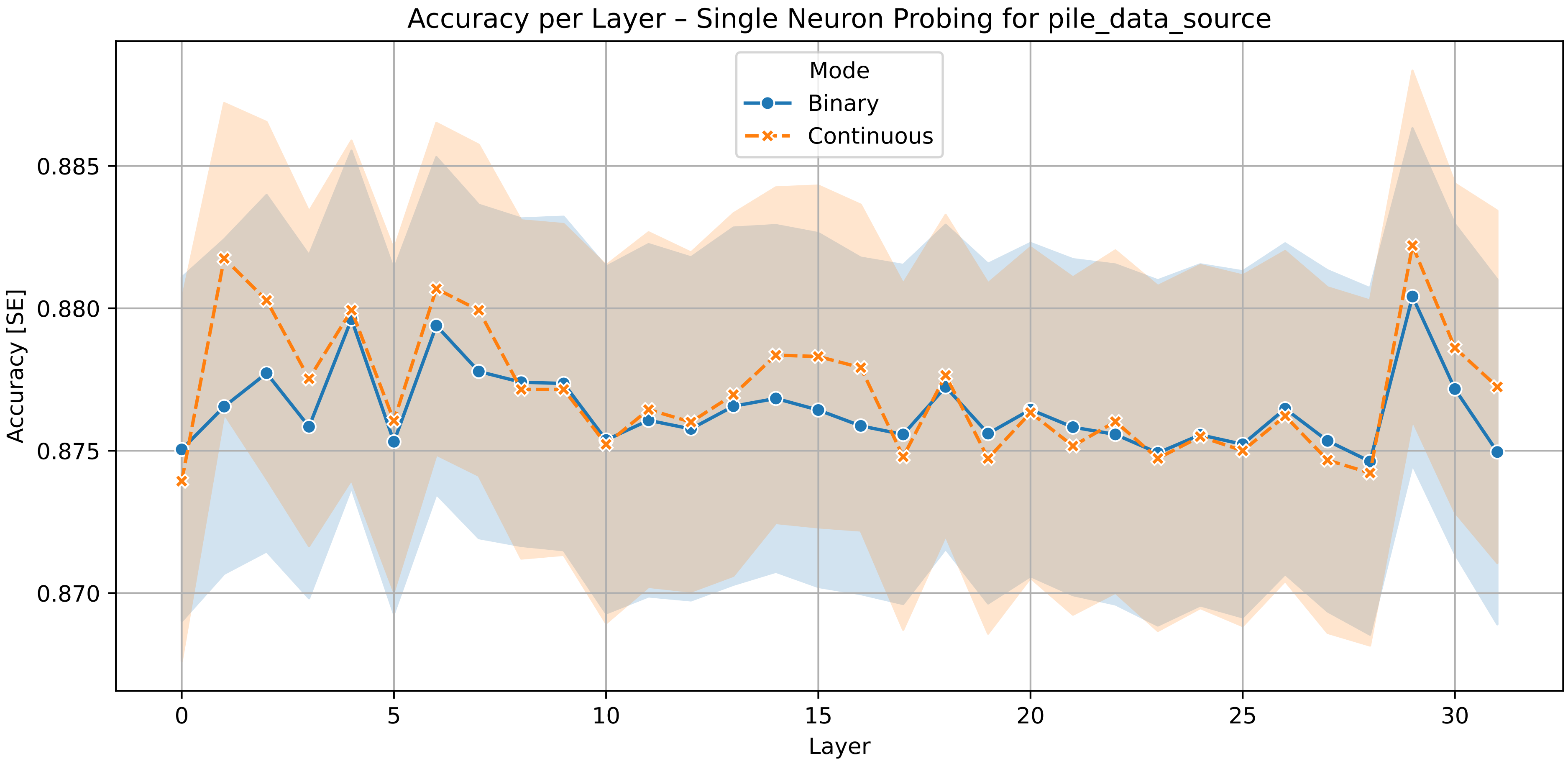

F1 per Layer – Full Layer Probing for pile_data_source
Mode
Binary
Continuous
F1 [SE]
0.91
0.92
0.93
0.94
0.95
0.96
0.97
0.98
Layer
0
5
10
15
20
25
30

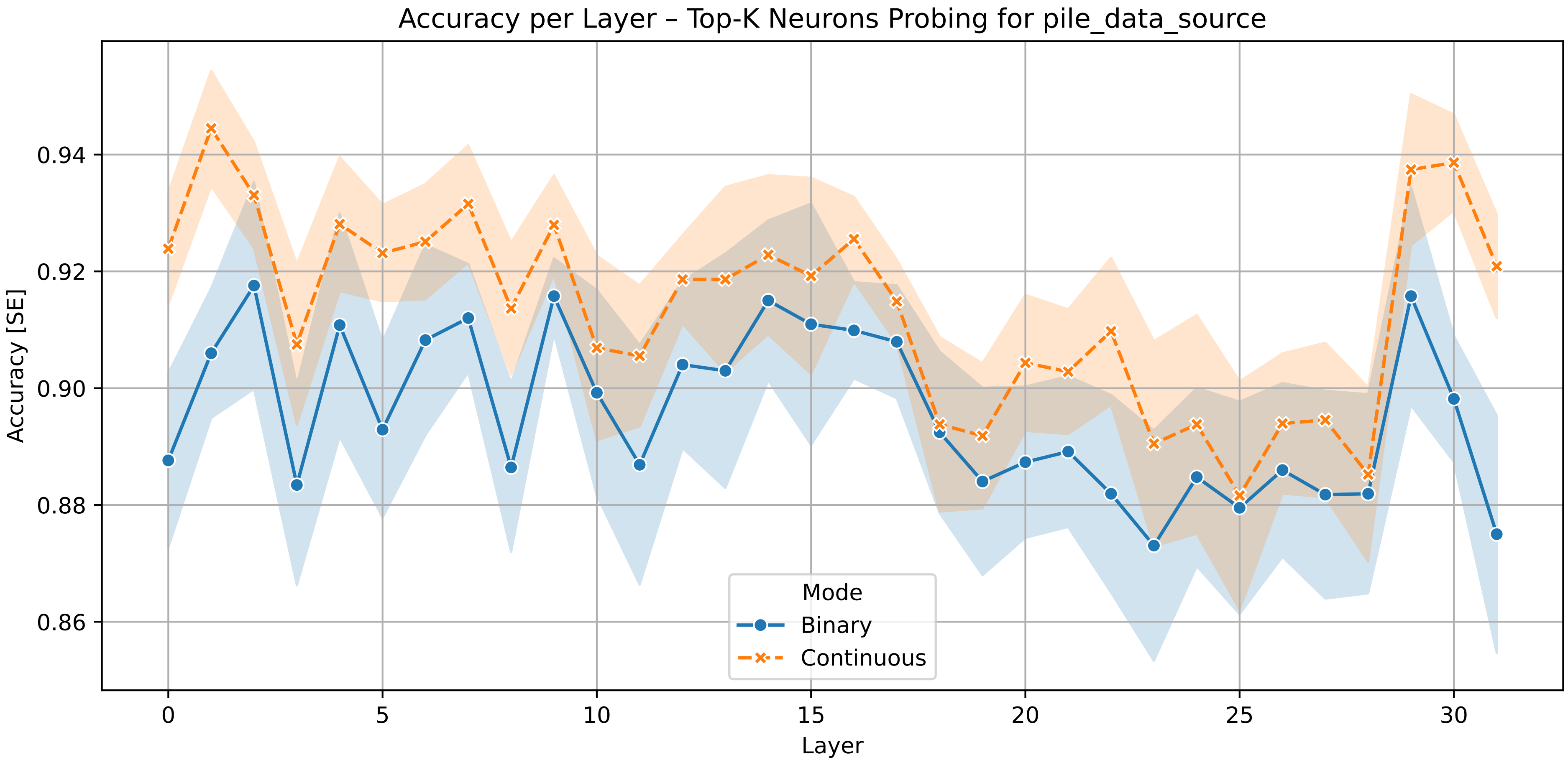

Overall F1 per Layer – All Methods for pile_data_source
F1 Score [SE]
0.800
0.825
0.850
0.875
0.900
0.925
0.950
0.975
Layer
0
5
10
15
20
25
30
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

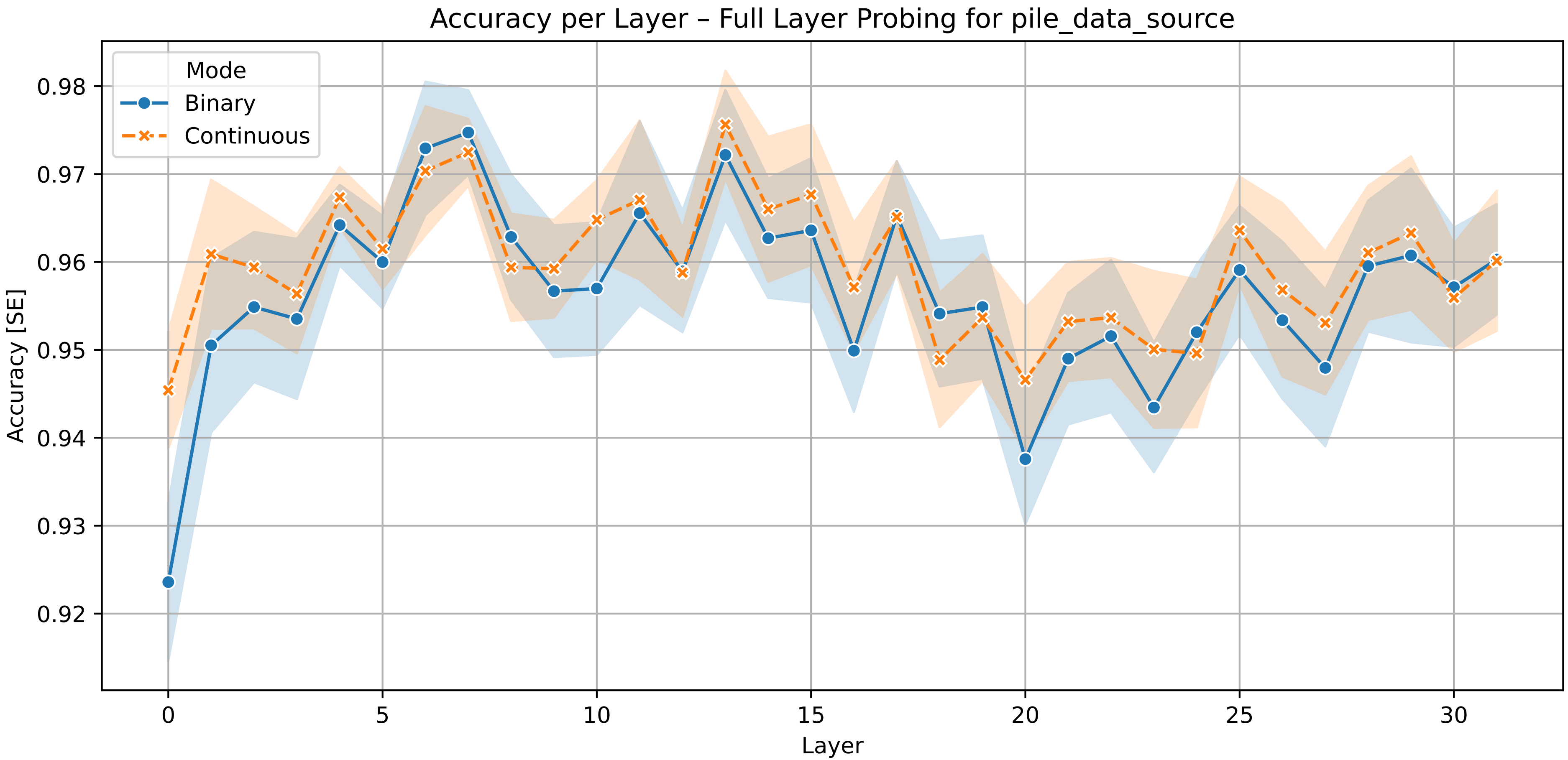
F1 Score Summary by Probing Method

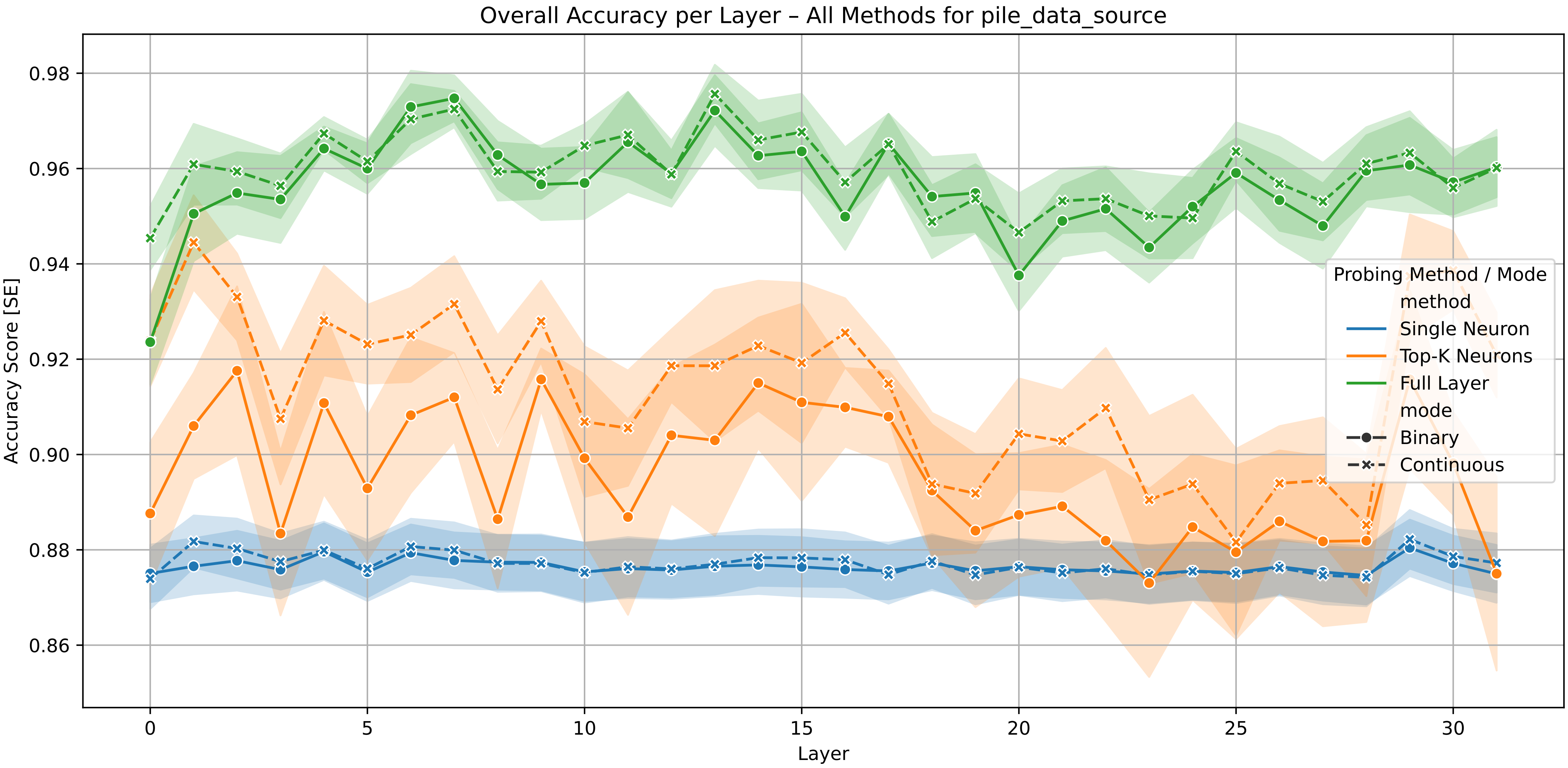


Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	7.0	13.0
Full Layer	f1_max	1.0	0.9988
Full Layer	f1_mean	0.9553	0.9586
Full Layer	f1_std	0.0238	0.0209
Single Neuron	f1_best_layer	4.0	1.0
Single Neuron	f1_max	1.0	0.9988
Single Neuron	f1_mean	0.8215	0.829
Single Neuron	f1_std	0.0746	0.0702
Top-K Neurons	f1_best_layer	9.0	1.0
Top-K Neurons	f1_max	1.0	0.9988
Top-K Neurons	f1_mean	0.8691	0.9063
Top-K Neurons	f1_std	0.055	0.0404

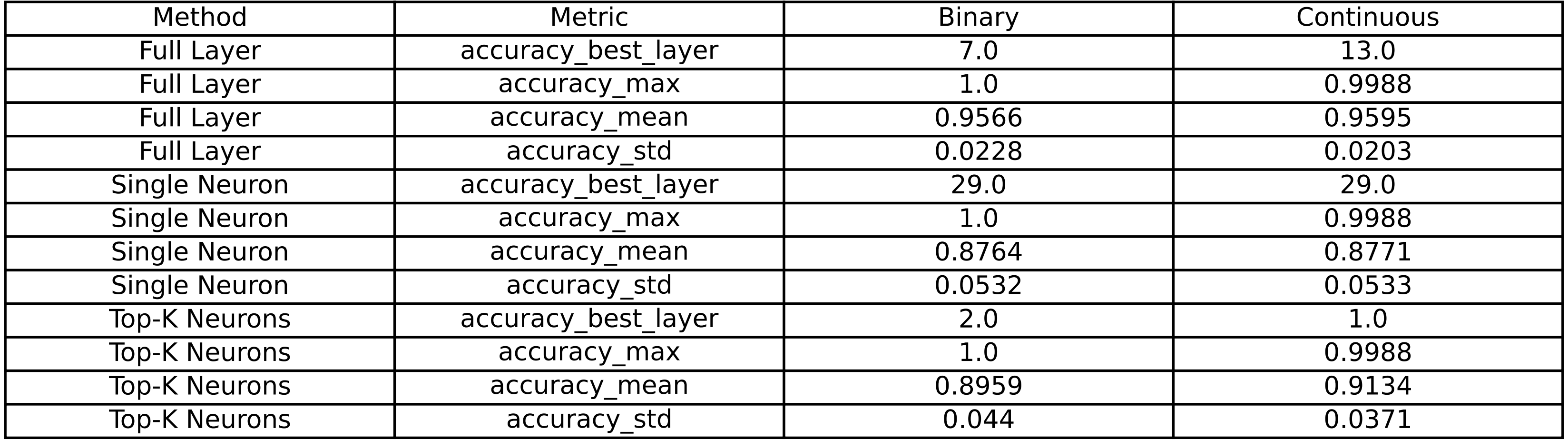

Accuracy per Layer – Single Neuron Probing for pile_data_source
Mode
Binary
Continuous
Accuracy [SE]
0.885
0.880
0.875
0.870
Layer
0
5
10
15
20
25
30


Accuracy per Layer – Top-K Neurons Probing for pile_data_source
Accuracy [SE]
0.86
0.88
0.90
0.92
0.94
Layer
0
5
10
15
20
25
30
Mode
Binary
Continuous


Accuracy per Layer – Full Layer Probing for pile_data_source
Mode
Binary
Continuous
Accuracy [SE]
0.92
0.93
0.94
0.95
0.96
0.97
0.98
Layer
0
5
10
15
20
25
30


Overall Accuracy per Layer – All Methods for pile_data_source
Accuracy Score [SE]
0.98
0.96
0.94
0.92
0.90
0.88
0.86
Layer
0
5
10
15
20
25
30
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	7.0	13.0
Full Layer	accuracy_max	1.0	0.9988
Full Layer	accuracy_mean	0.9566	0.9595
Full Layer	accuracy_std	0.0228	0.0203
Single Neuron	accuracy_best_layer	29.0	29.0
Single Neuron	accuracy_max	1.0	0.9988
Single Neuron	accuracy_mean	0.8764	0.8771
Single Neuron	accuracy_std	0.0532	0.0533
Top-K Neurons	accuracy_best_layer	2.0	1.0
Top-K Neurons	accuracy_max	1.0	0.9988
Top-K Neurons	accuracy_mean	0.8959	0.9134
Top-K Neurons	accuracy_std	0.044	0.0371