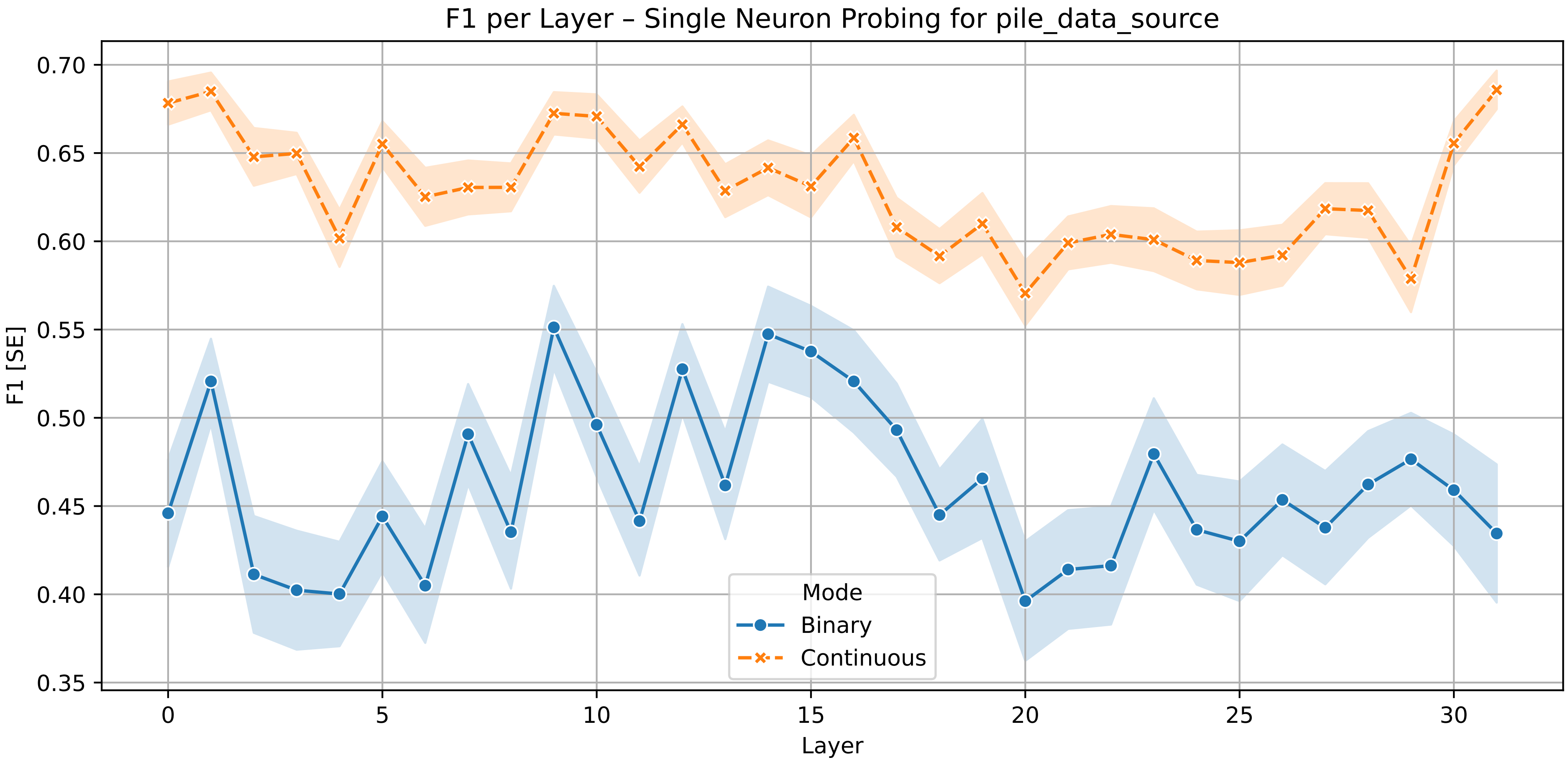
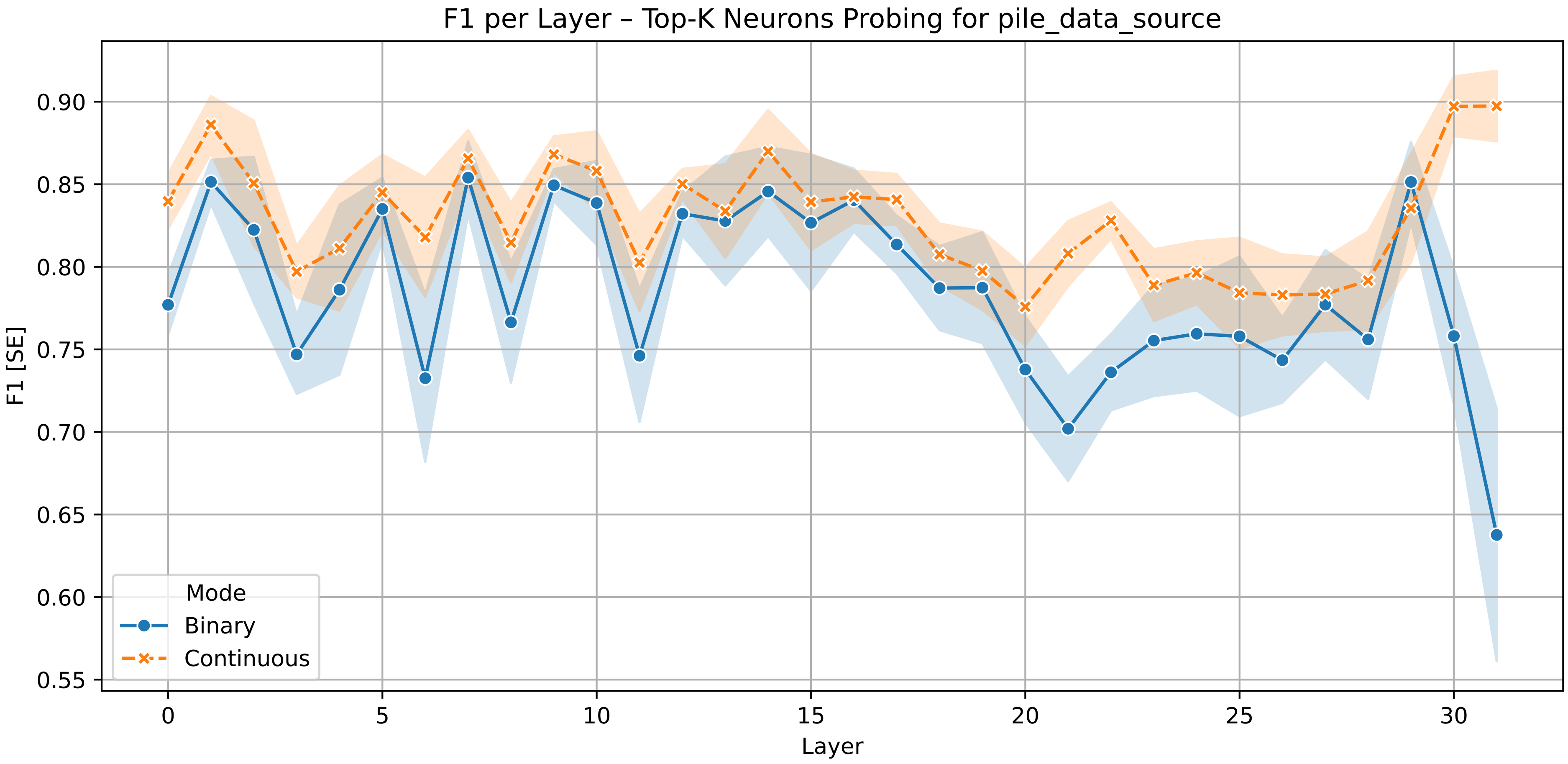
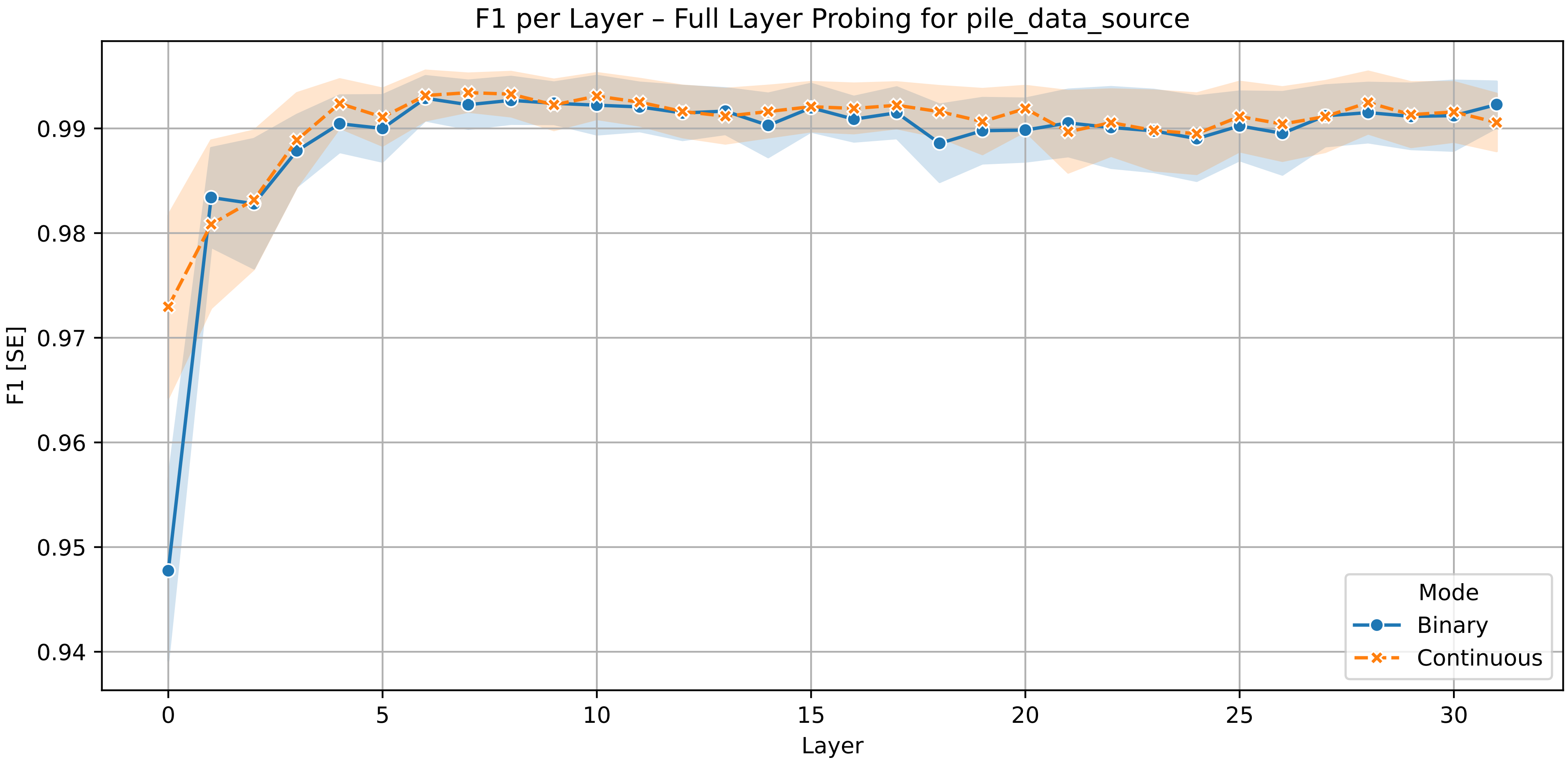
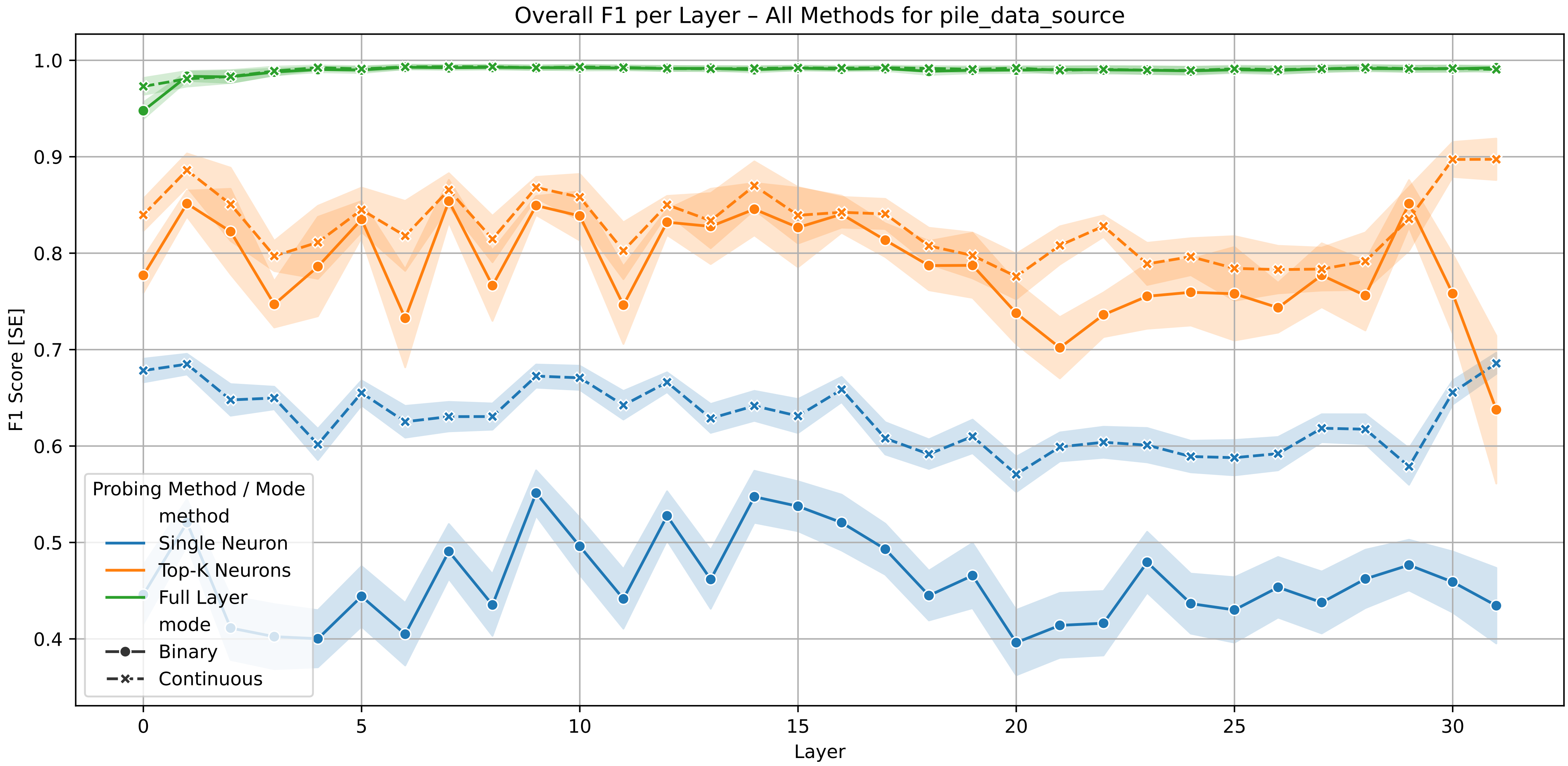
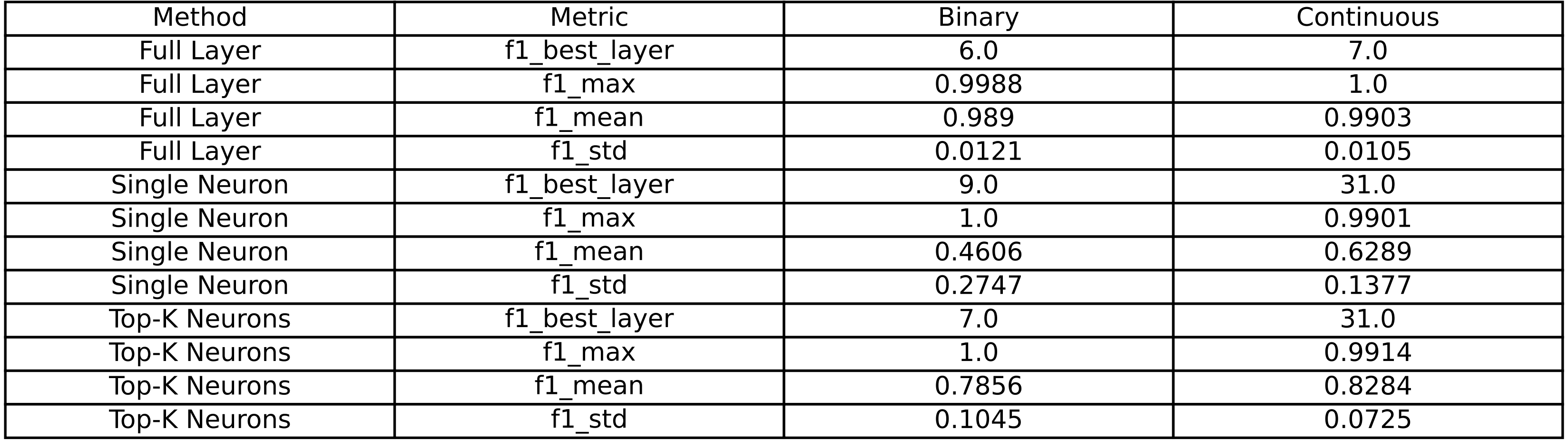

F1 per Layer – Single Neuron Probing for pile_data_source
F1 [SE]
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
Layer
0
5
10
15
20
25
30
Mode
Binary
Continuous


F1 per Layer – Top-K Neurons Probing for pile_data_source
F1 [SE]
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
Layer
0
5
10
15
20
25
30
Mode
Binary
Continuous

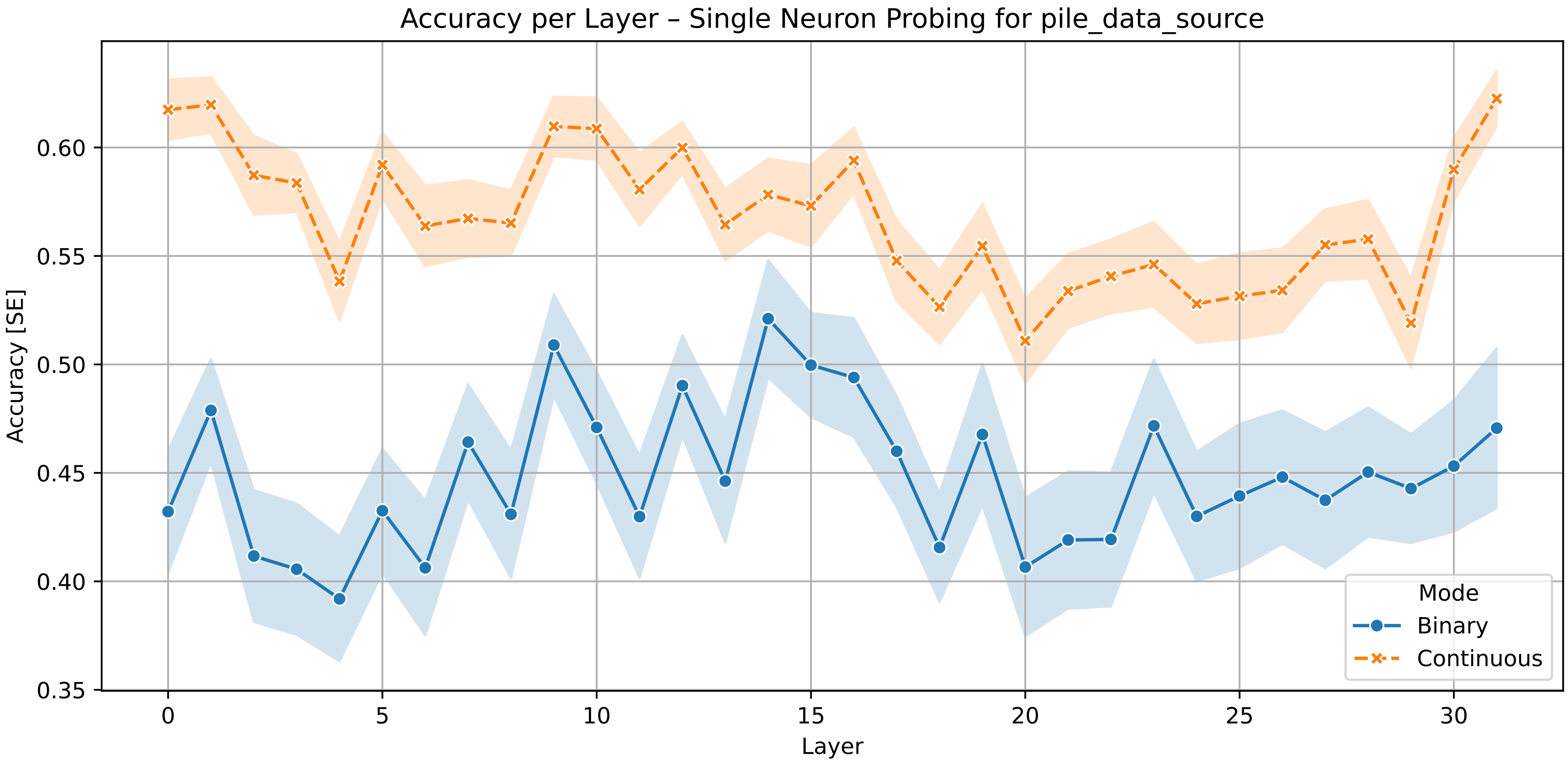

F1 per Layer – Full Layer Probing for pile_data_source
F1 [SE]
0.99
0.98
0.97
0.96
0.95
0.94
Layer
0
5
10
15
20
25
30
Mode
Binary
Continuous

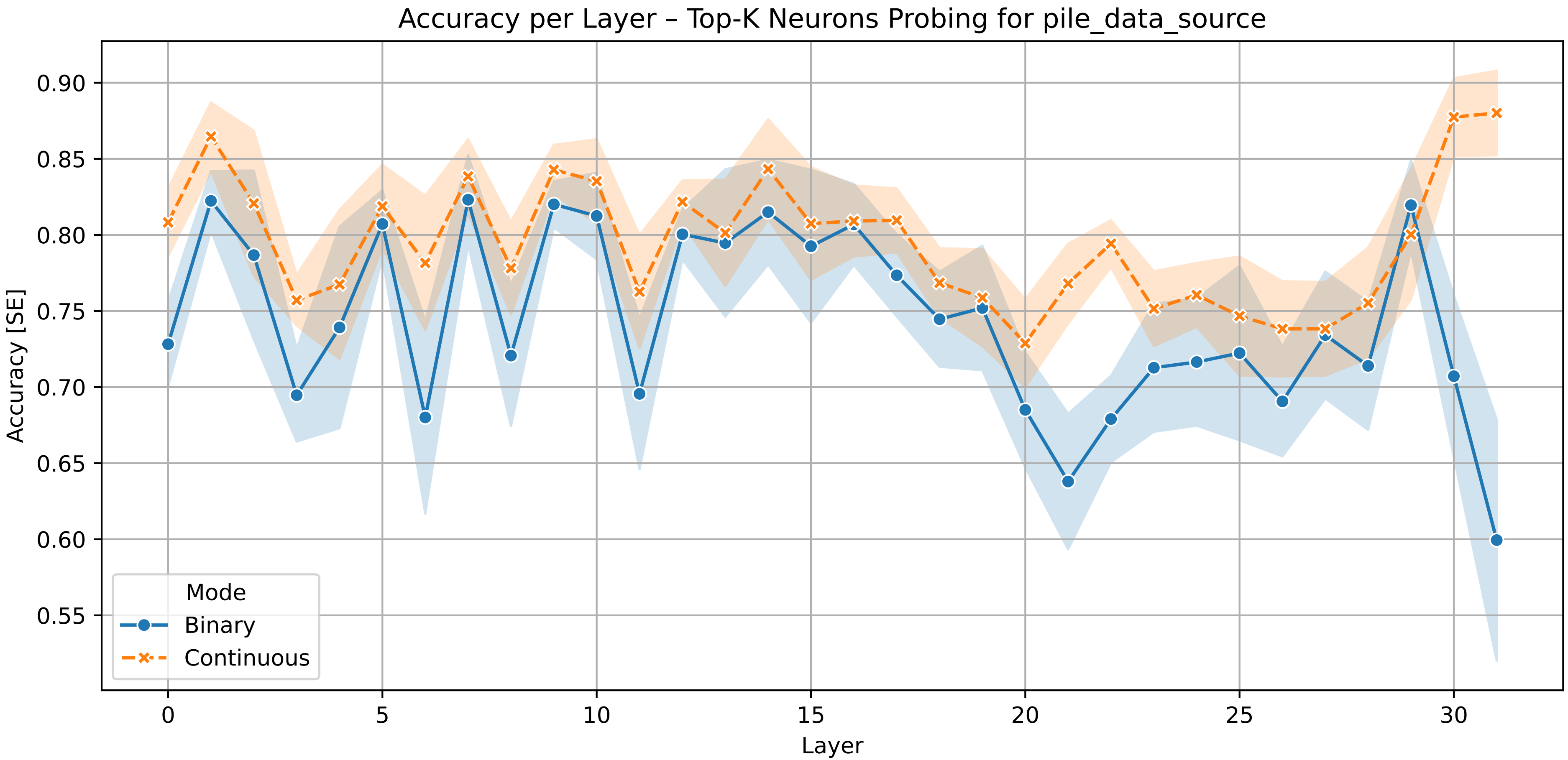

Overall F1 per Layer – All Methods for pile_data_source
F1 Score [SE]
1.0
0.9
0.8
0.7
0.6
0.5
0.4
Layer
0
5
10
15
20
25
30
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

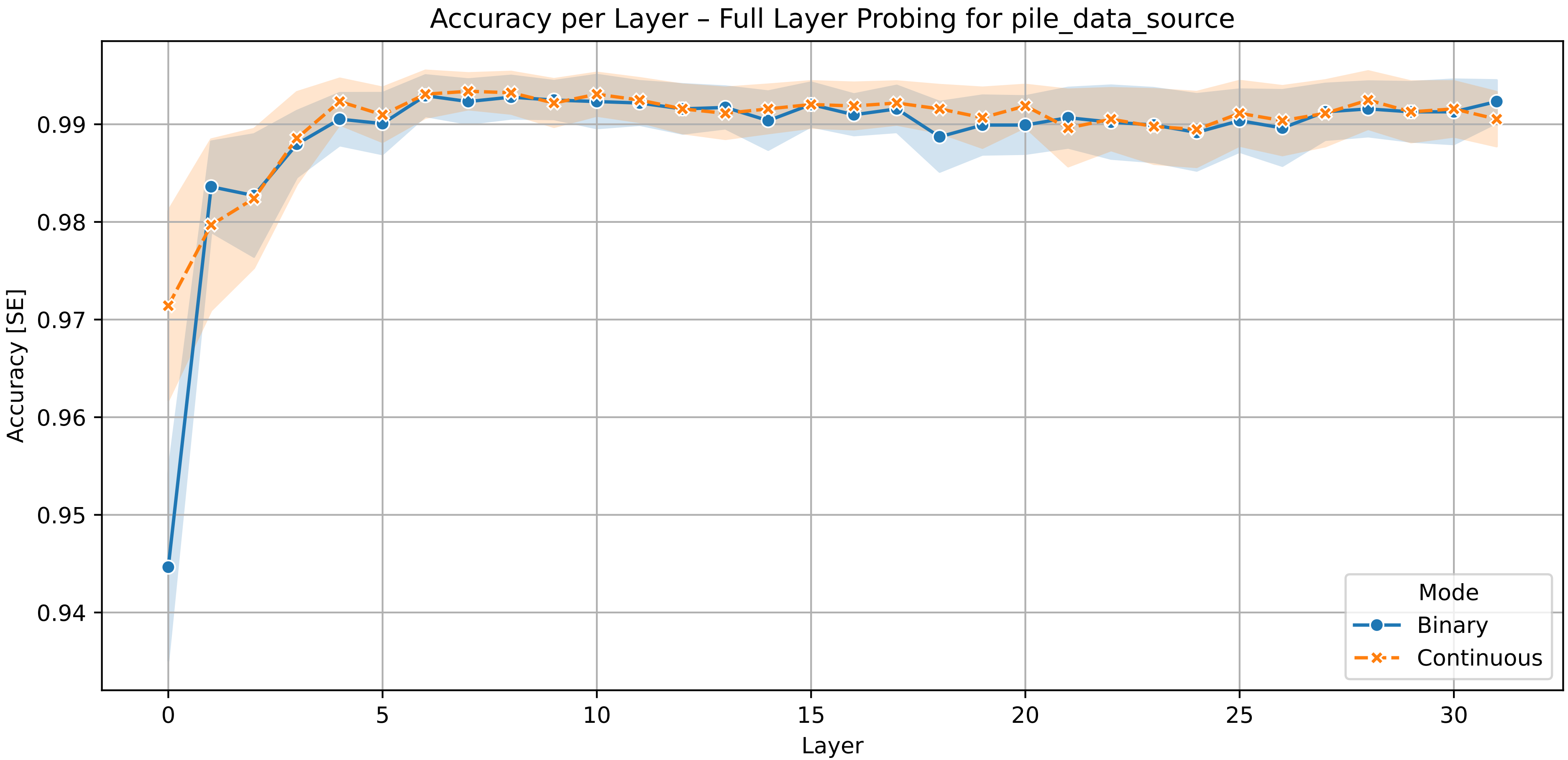
F1 Score Summary by Probing Method

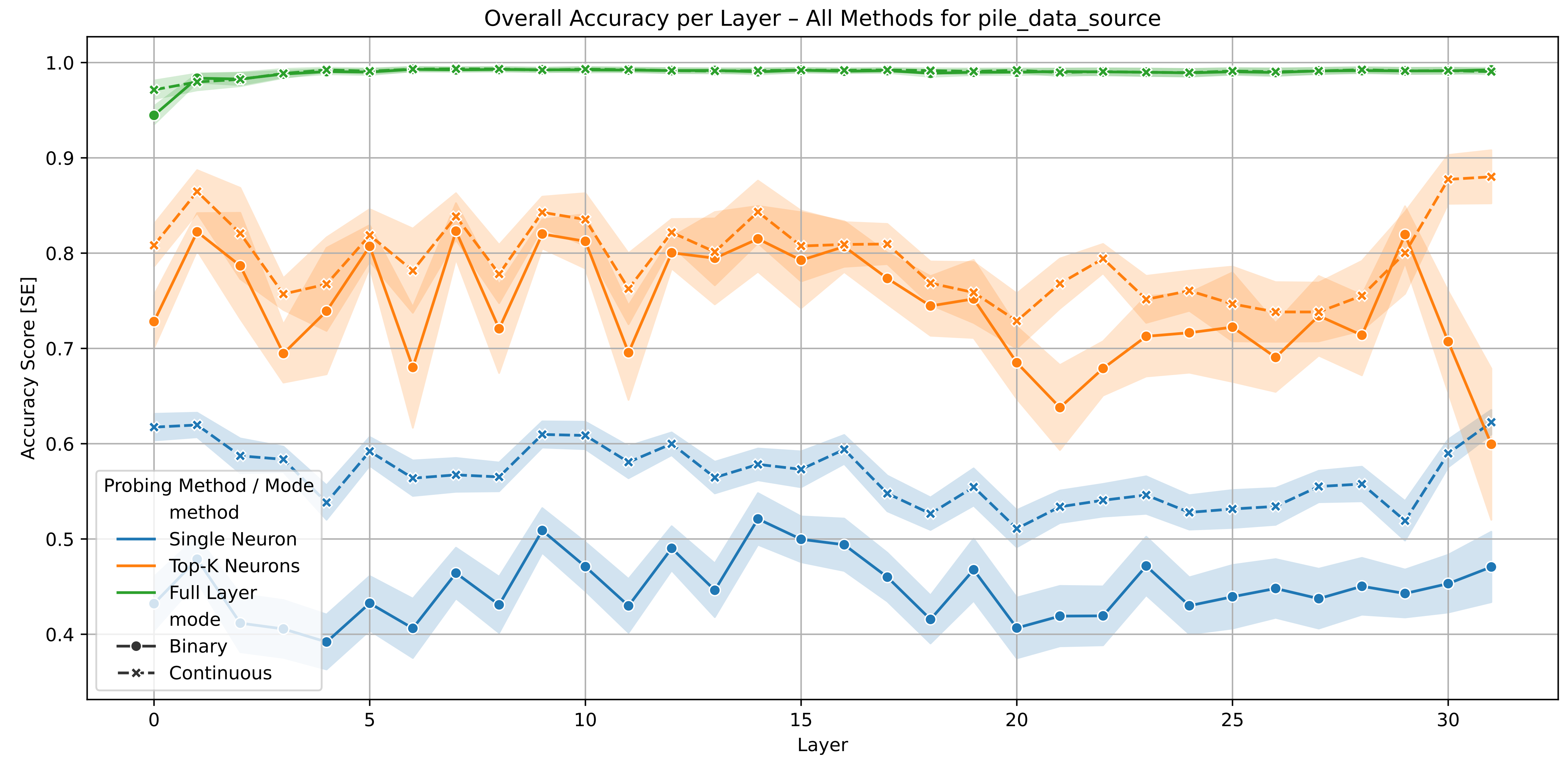


Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	6.0	7.0
Full Layer	f1_max	0.9988	1.0
Full Layer	f1_mean	0.989	0.9903
Full Layer	f1_std	0.0121	0.0105
Single Neuron	f1_best_layer	9.0	31.0
Single Neuron	f1_max	1.0	0.9901
Single Neuron	f1_mean	0.4606	0.6289
Single Neuron	f1_std	0.2747	0.1377
Top-K Neurons	f1_best_layer	7.0	31.0
Top-K Neurons	f1_max	1.0	0.9914
Top-K Neurons	f1_mean	0.7856	0.8284
Top-K Neurons	f1_std	0.1045	0.0725

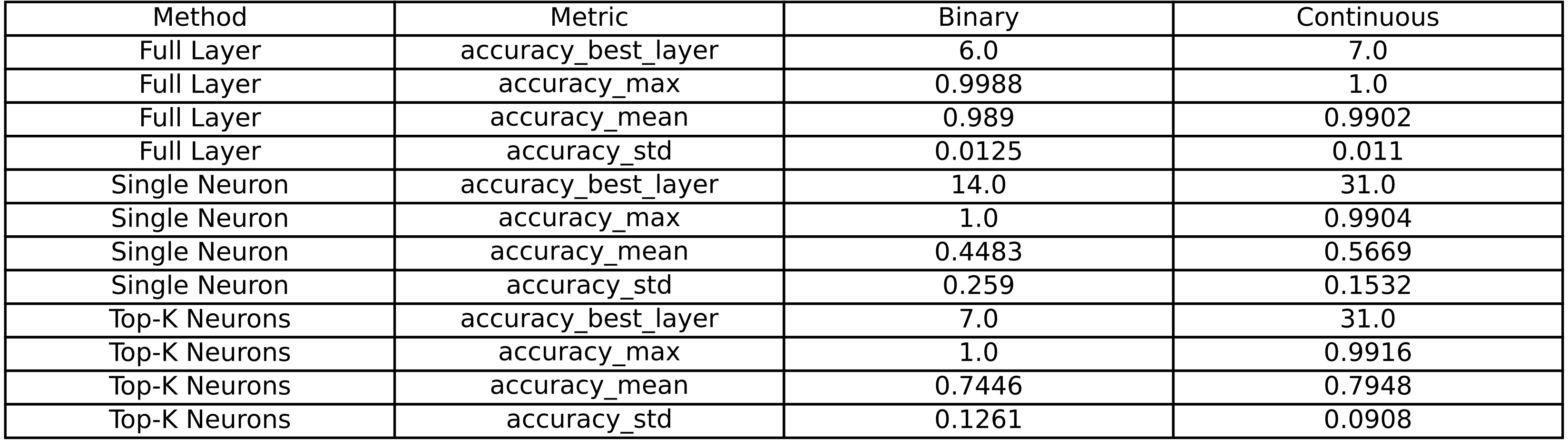

Accuracy per Layer – Single Neuron Probing for pile_data_source
Accuracy [SE]
0.35
0.40
0.45
0.50
0.55
0.60
Layer
0
5
10
15
20
25
30
Mode
Binary
Continuous


Accuracy per Layer – Top-K Neurons Probing for pile_data_source
Accuracy [SE]
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
Layer
0
5
10
15
20
25
30
Mode
Binary
Continuous


Accuracy per Layer – Full Layer Probing for pile_data_source
Accuracy [SE]
0.94
0.95
0.96
0.97
0.98
0.99
Layer
0
5
10
15
20
25
30
Mode
Binary
Continuous


Overall Accuracy per Layer – All Methods for pile_data_source
Accuracy Score [SE]
Layer
1.0
0.9
0.8
0.7
0.6
0.5
0.4
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	6.0	7.0
Full Layer	accuracy_max	0.9988	1.0
Full Layer	accuracy_mean	0.989	0.9902
Full Layer	accuracy_std	0.0125	0.011
Single Neuron	accuracy_best_layer	14.0	31.0
Single Neuron	accuracy_max	1.0	0.9904
Single Neuron	accuracy_mean	0.4483	0.5669
Single Neuron	accuracy_std	0.259	0.1532
Top-K Neurons	accuracy_best_layer	7.0	31.0
Top-K Neurons	accuracy_max	1.0	0.9916
Top-K Neurons	accuracy_mean	0.7446	0.7948
Top-K Neurons	accuracy_std	0.1261	0.0908