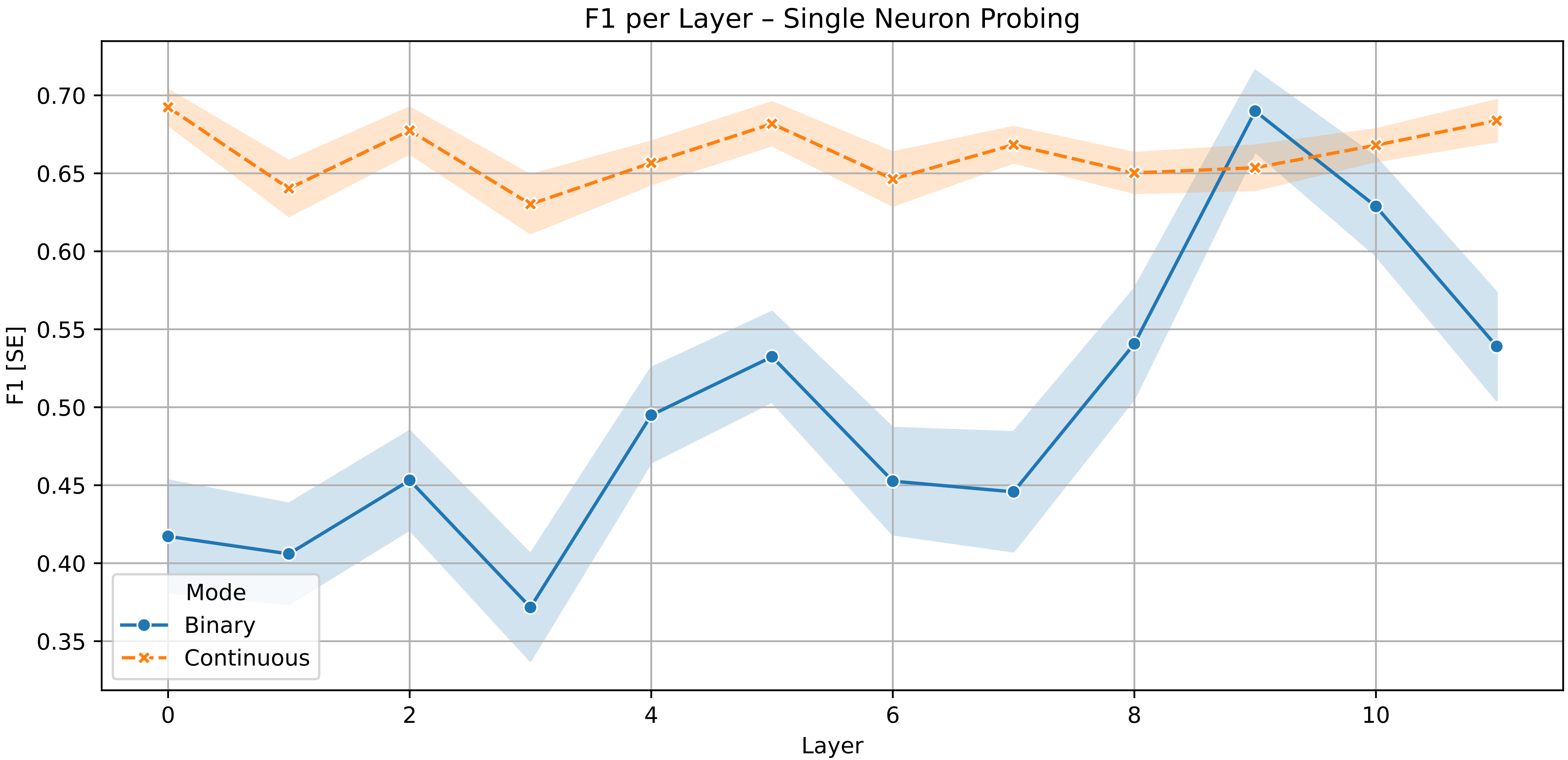
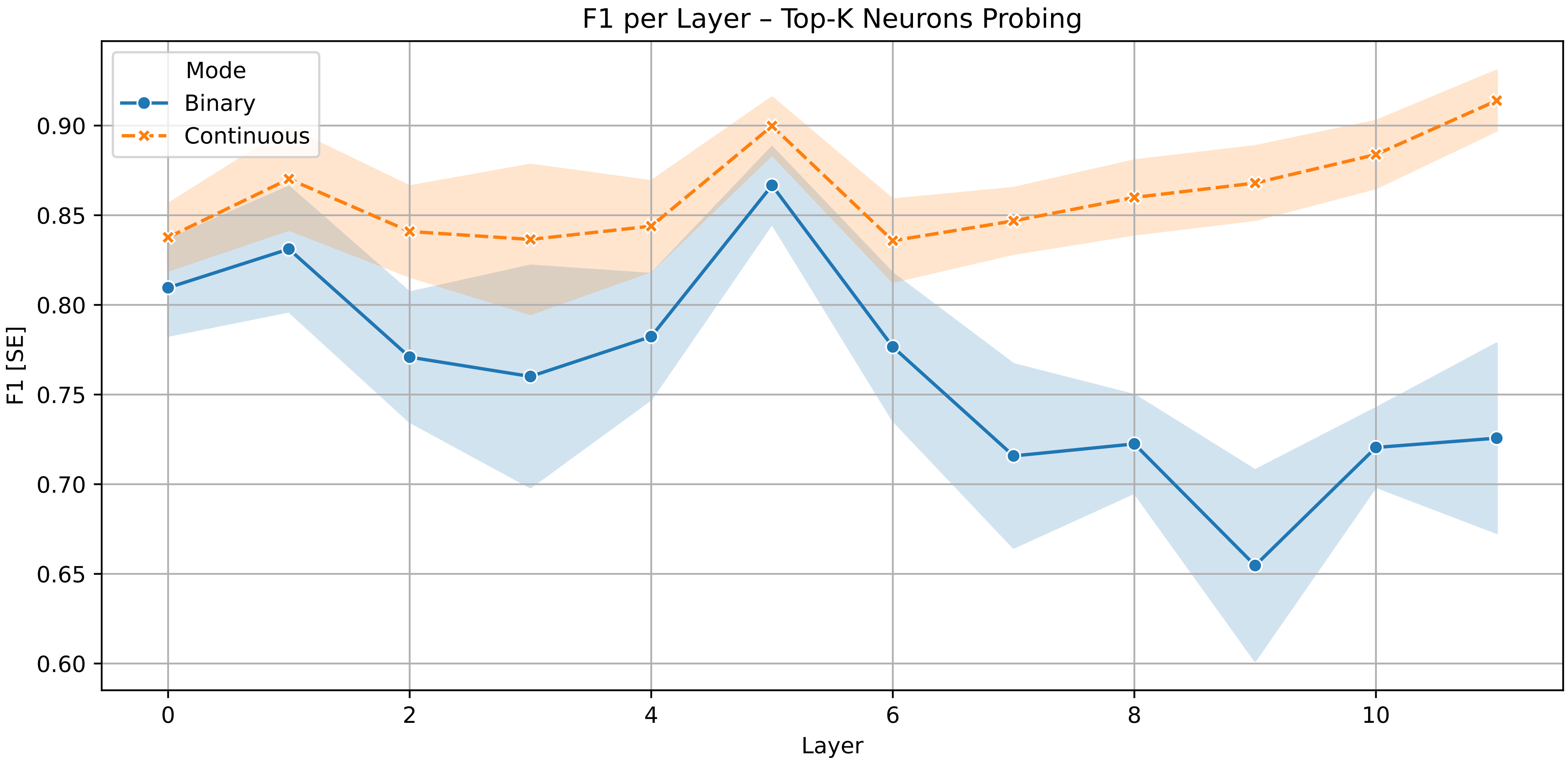
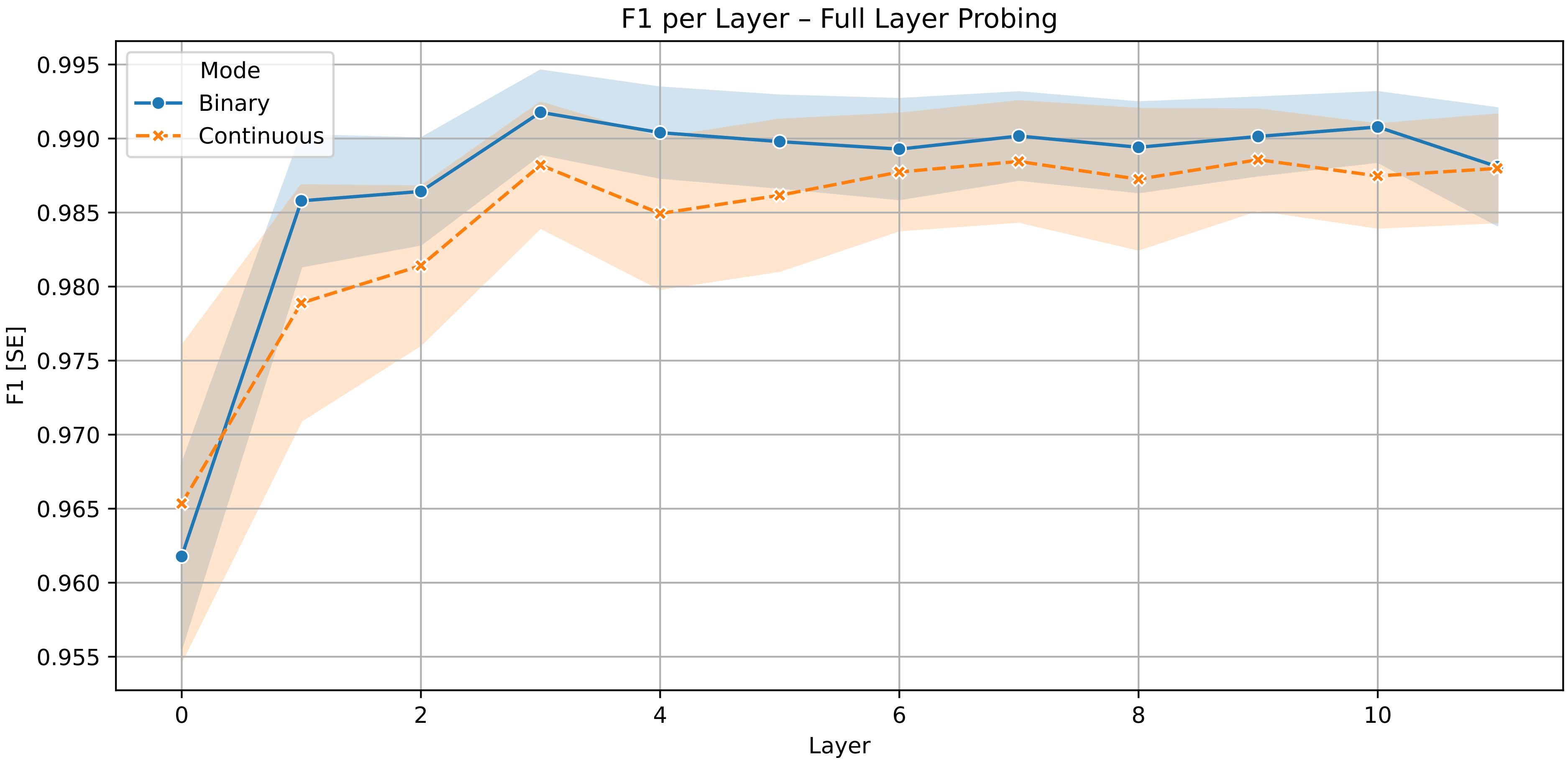
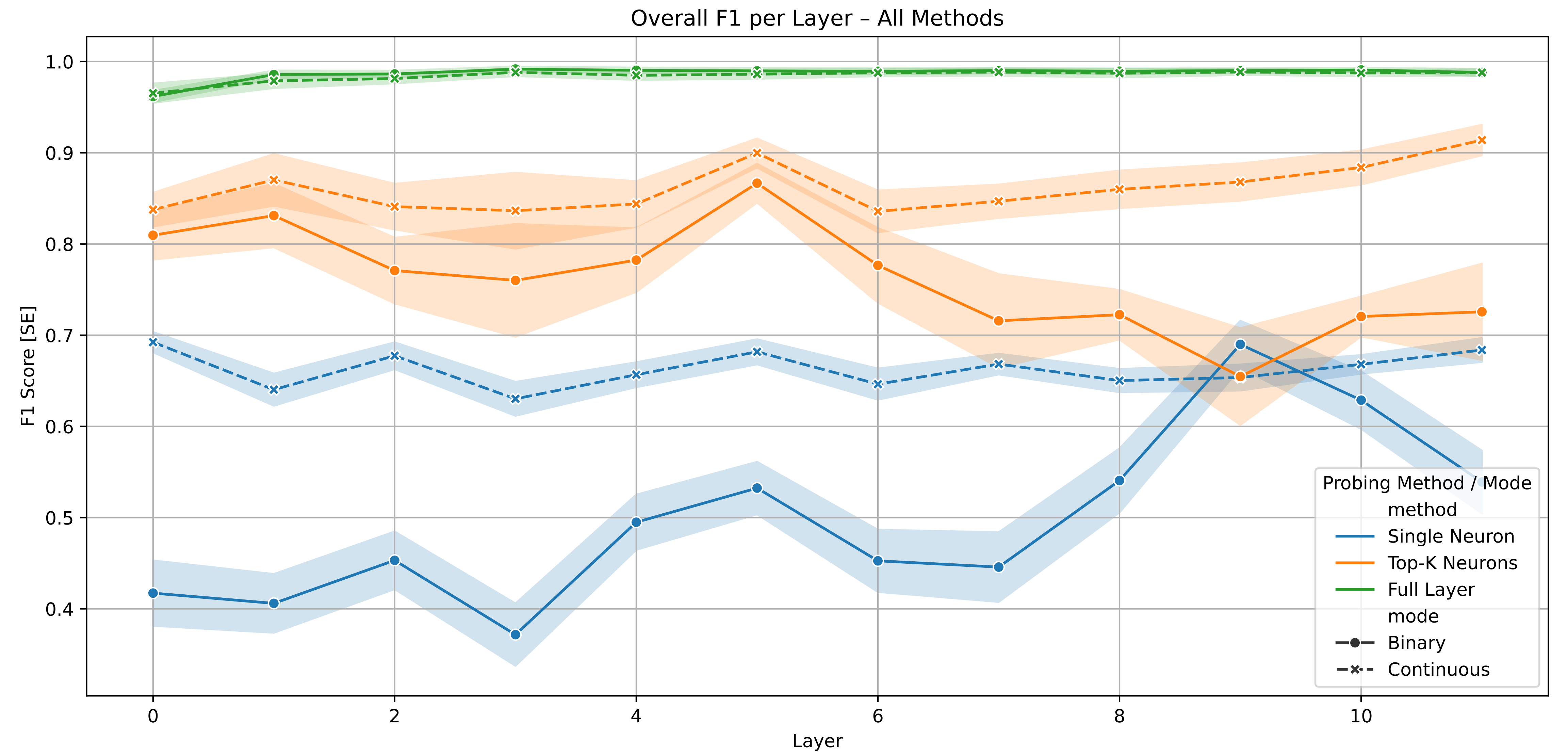
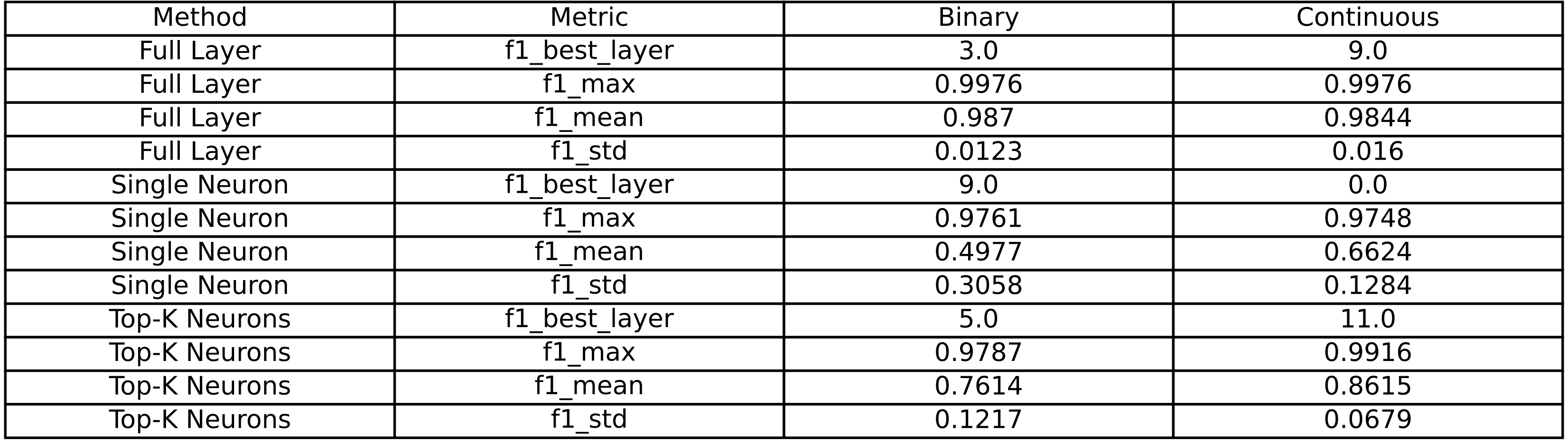

F1 per Layer – Single Neuron Probing
F1 [SE]
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
Layer
0
2
4
6
8
10
Mode
Binary
Continuous


F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.60
0.65
0.70
0.75
0.80
0.85
0.90
Layer
0
2
4
6
8
10

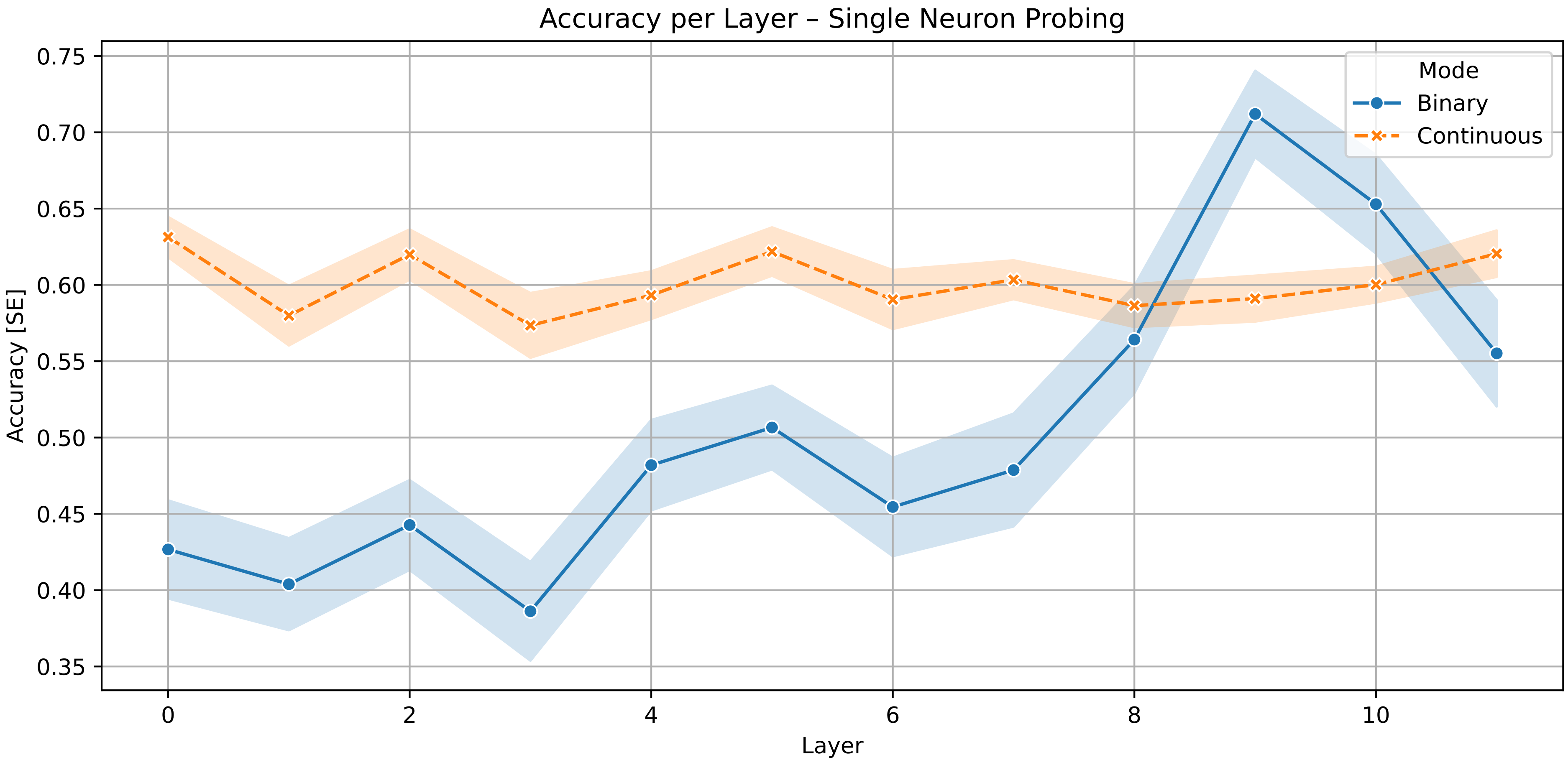

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.995
0.990
0.985
0.980
0.975
0.970
0.965
0.960
0.955
Layer
0
2
4
6
8
10

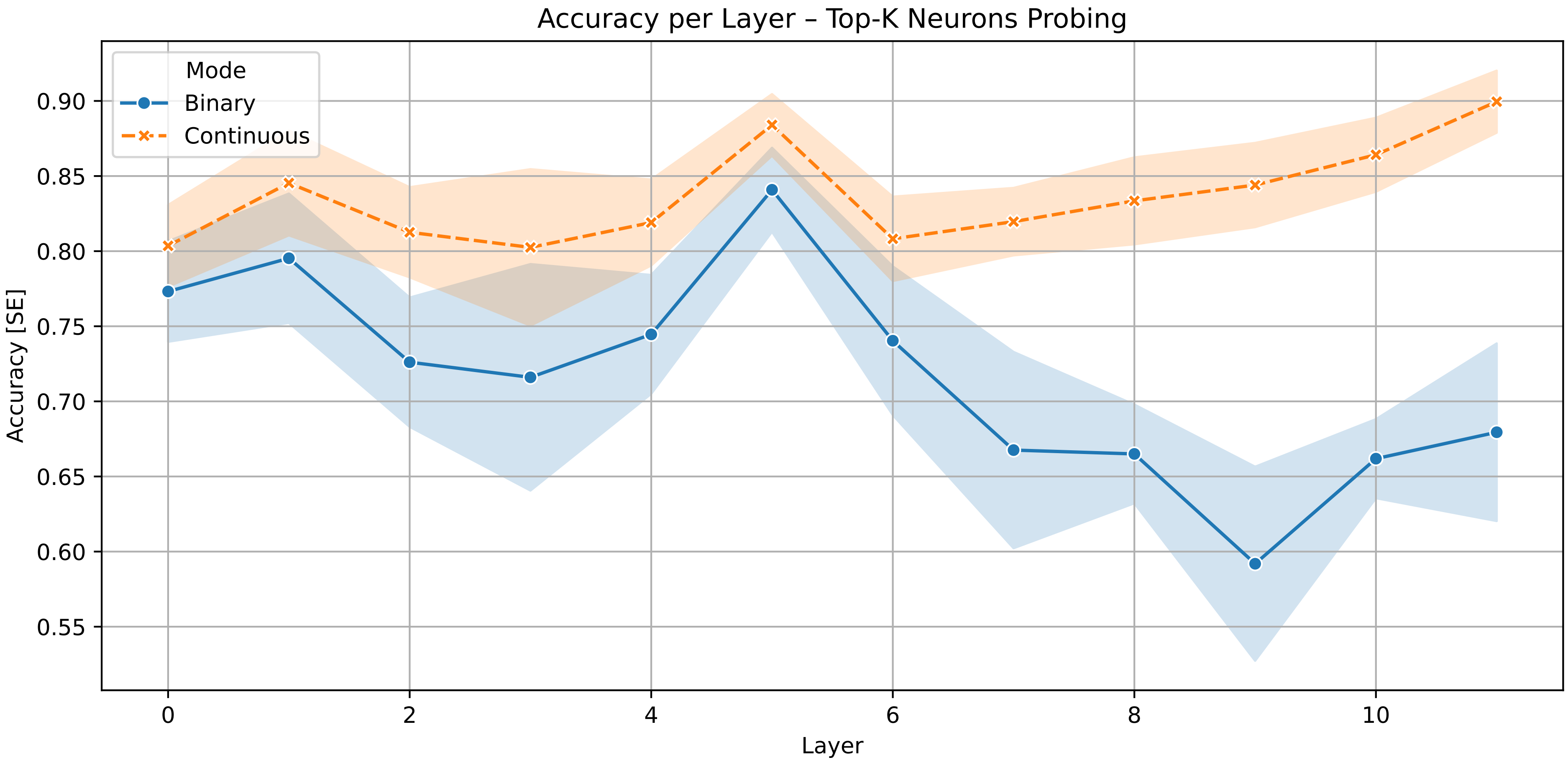

Overall F1 per Layer – All Methods
F1 Score [SE]
1.0
0.9
0.8
0.7
0.6
0.5
0.4
Layer
0
2
4
6
8
10
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

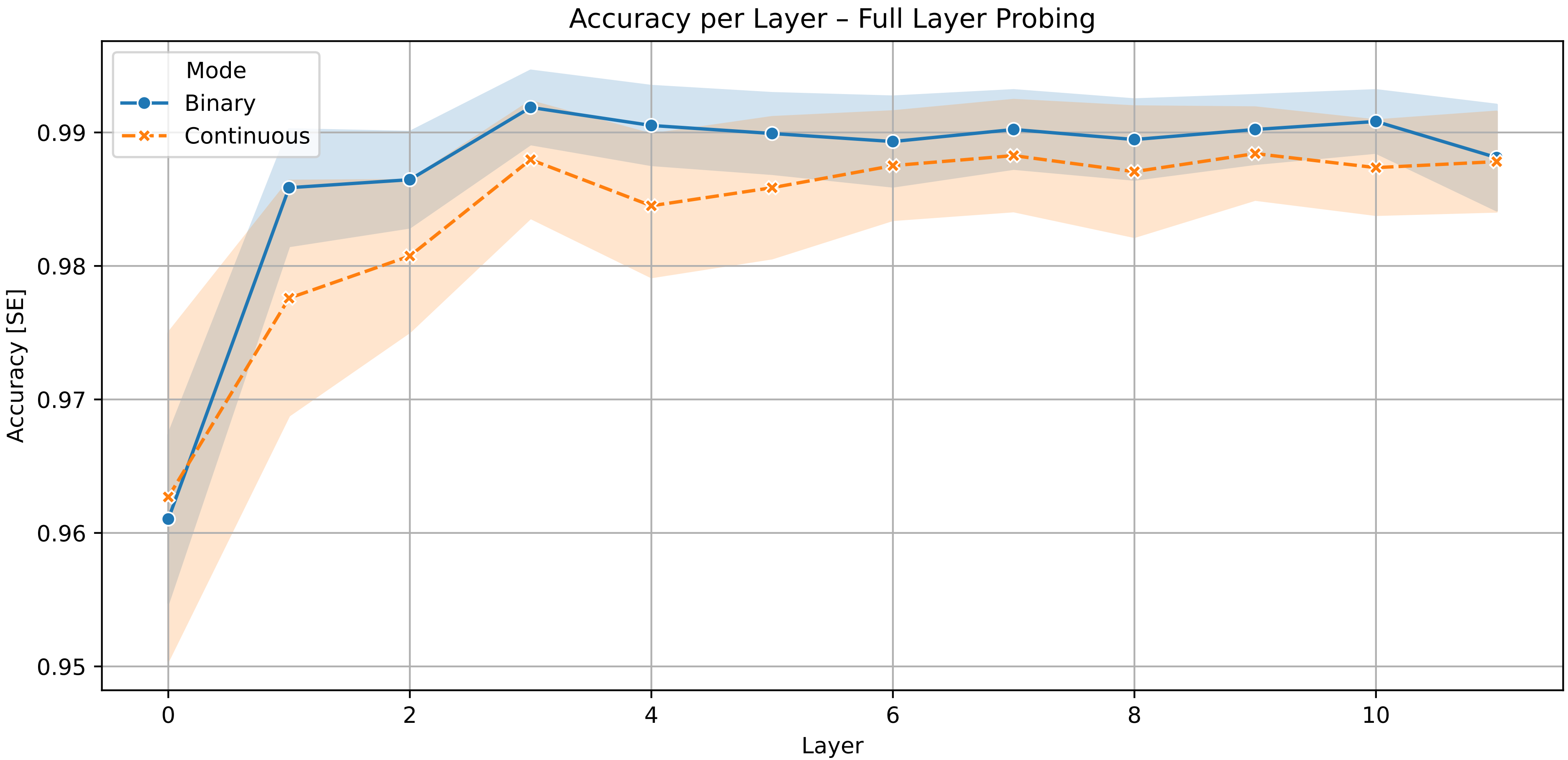
F1 Score Summary by Probing Method

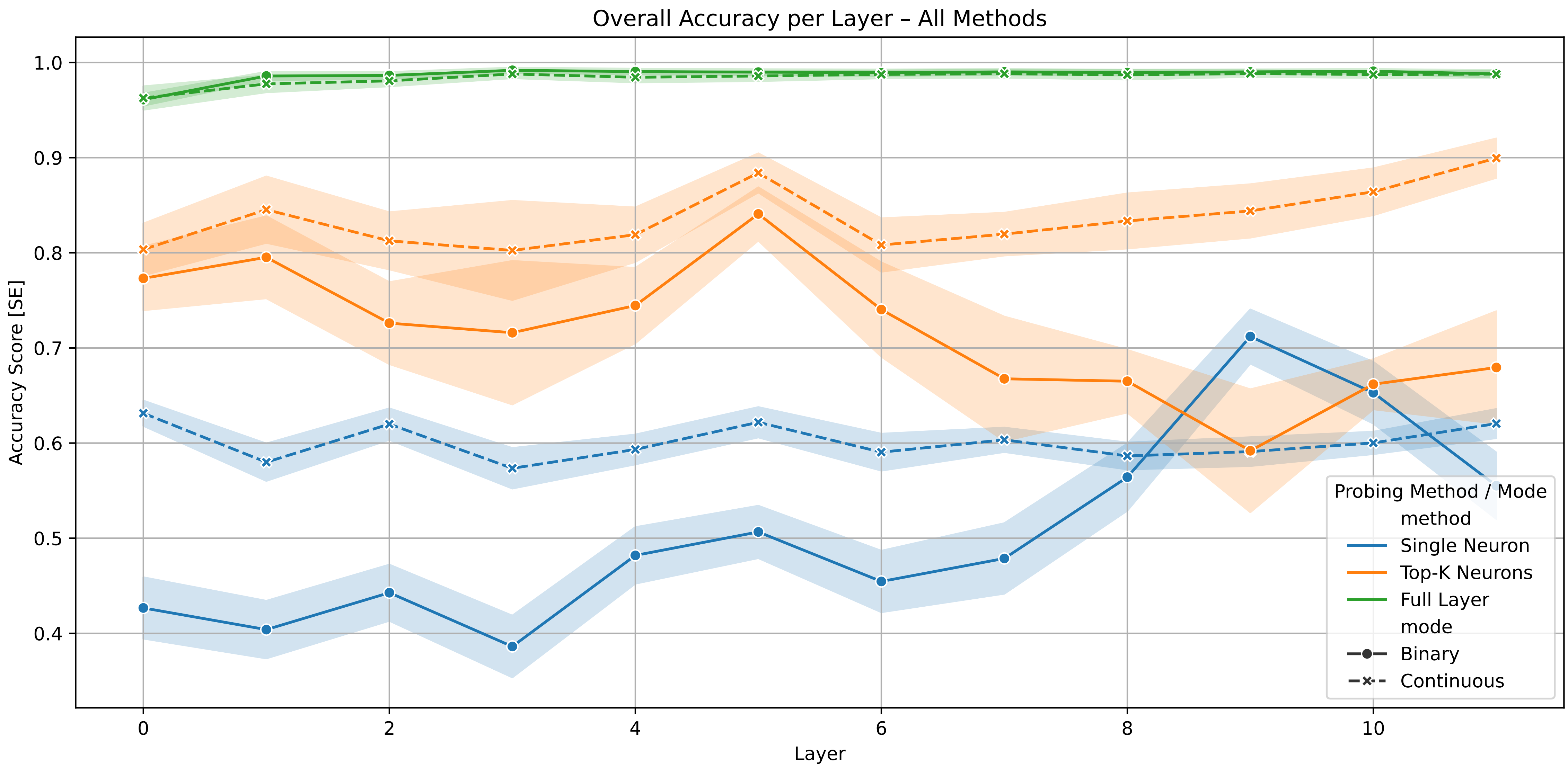


Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	3.0	9.0
Full Layer	f1_max	0.9976	0.9976
Full Layer	f1_mean	0.987	0.9844
Full Layer	f1_std	0.0123	0.016
Single Neuron	f1_best_layer	9.0	0.0
Single Neuron	f1_max	0.9761	0.9748
Single Neuron	f1_mean	0.4977	0.6624
Single Neuron	f1_std	0.3058	0.1284
Top-K Neurons	f1_best_layer	5.0	11.0
Top-K Neurons	f1_max	0.9787	0.9916
Top-K Neurons	f1_mean	0.7614	0.8615
Top-K Neurons	f1_std	0.1217	0.0679

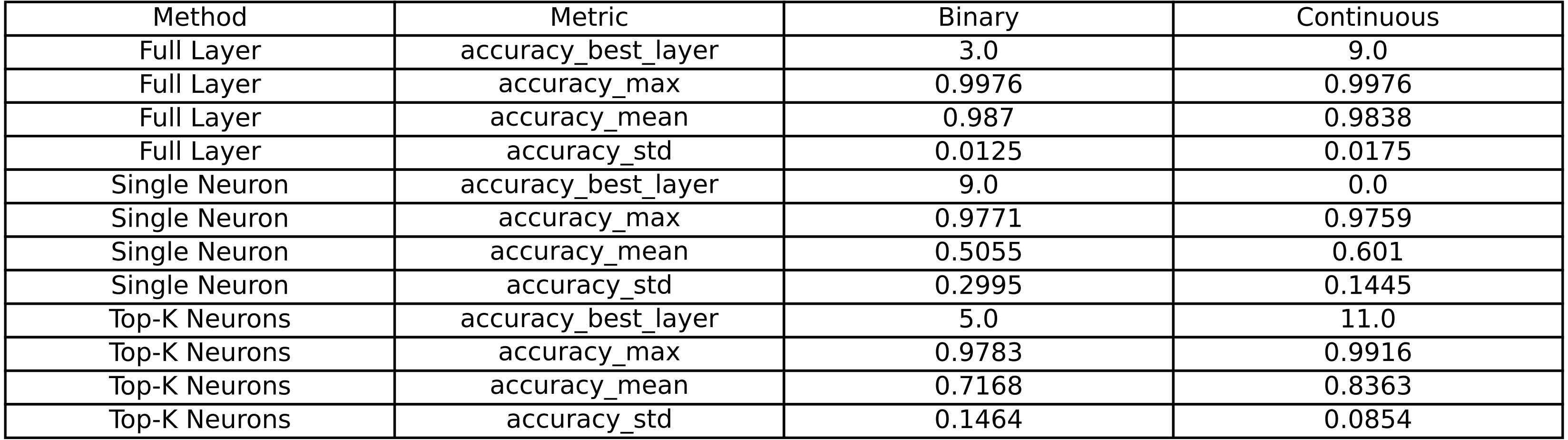

Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
Layer
0
2
4
6
8
10


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
Layer
0
2
4
6
8
10


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.95
0.96
0.97
0.98
0.99
Layer
0
2
4
6
8
10


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
2
4
6
8
10
Layer
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	3.0	9.0
Full Layer	accuracy_max	0.9976	0.9976
Full Layer	accuracy_mean	0.987	0.9838
Full Layer	accuracy_std	0.0125	0.0175
Single Neuron	accuracy_best_layer	9.0	0.0
Single Neuron	accuracy_max	0.9771	0.9759
Single Neuron	accuracy_mean	0.5055	0.601
Single Neuron	accuracy_std	0.2995	0.1445
Top-K Neurons	accuracy_best_layer	5.0	11.0
Top-K Neurons	accuracy_max	0.9783	0.9916
Top-K Neurons	accuracy_mean	0.7168	0.8363
Top-K Neurons	accuracy_std	0.1464	0.0854