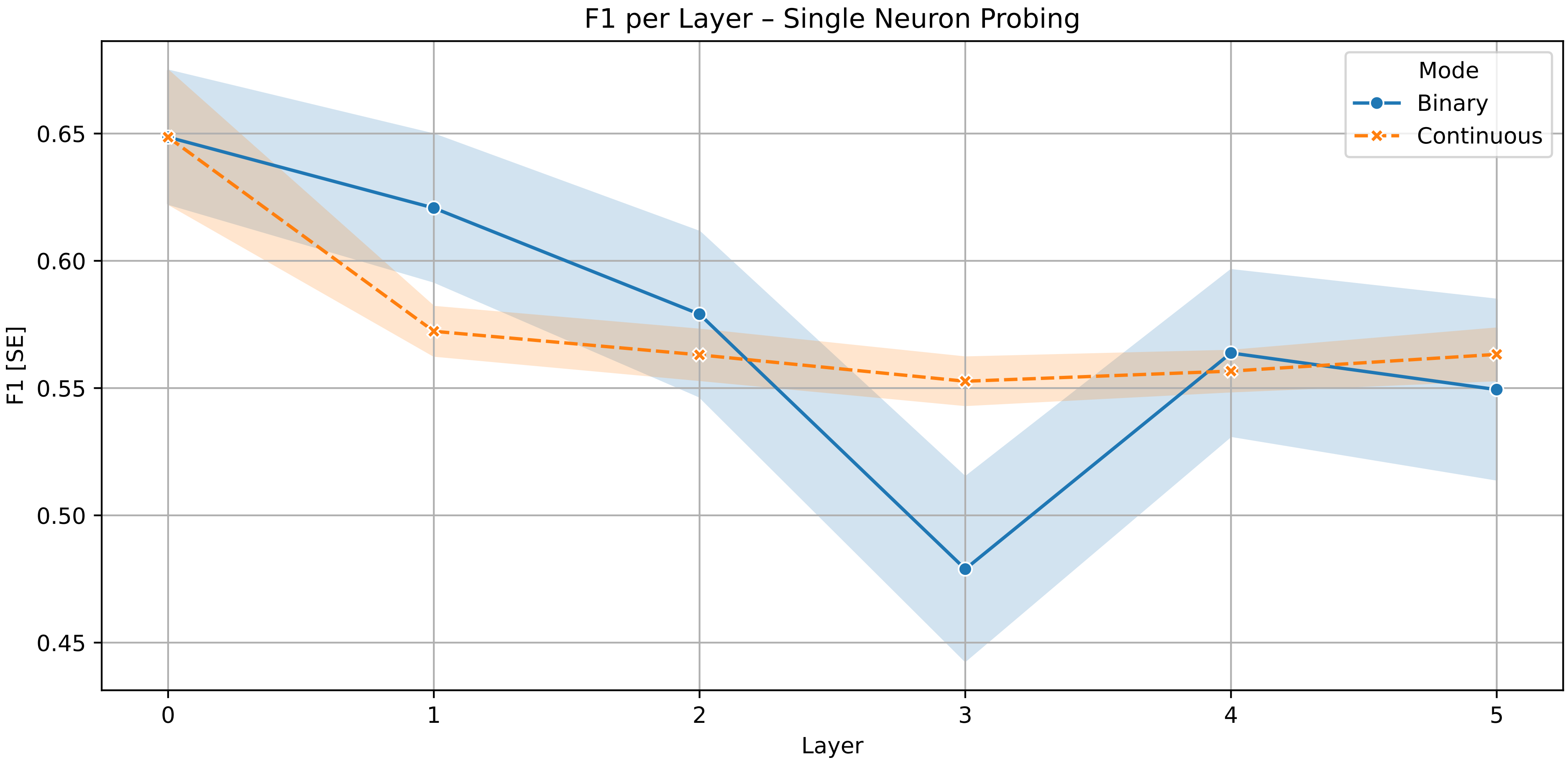
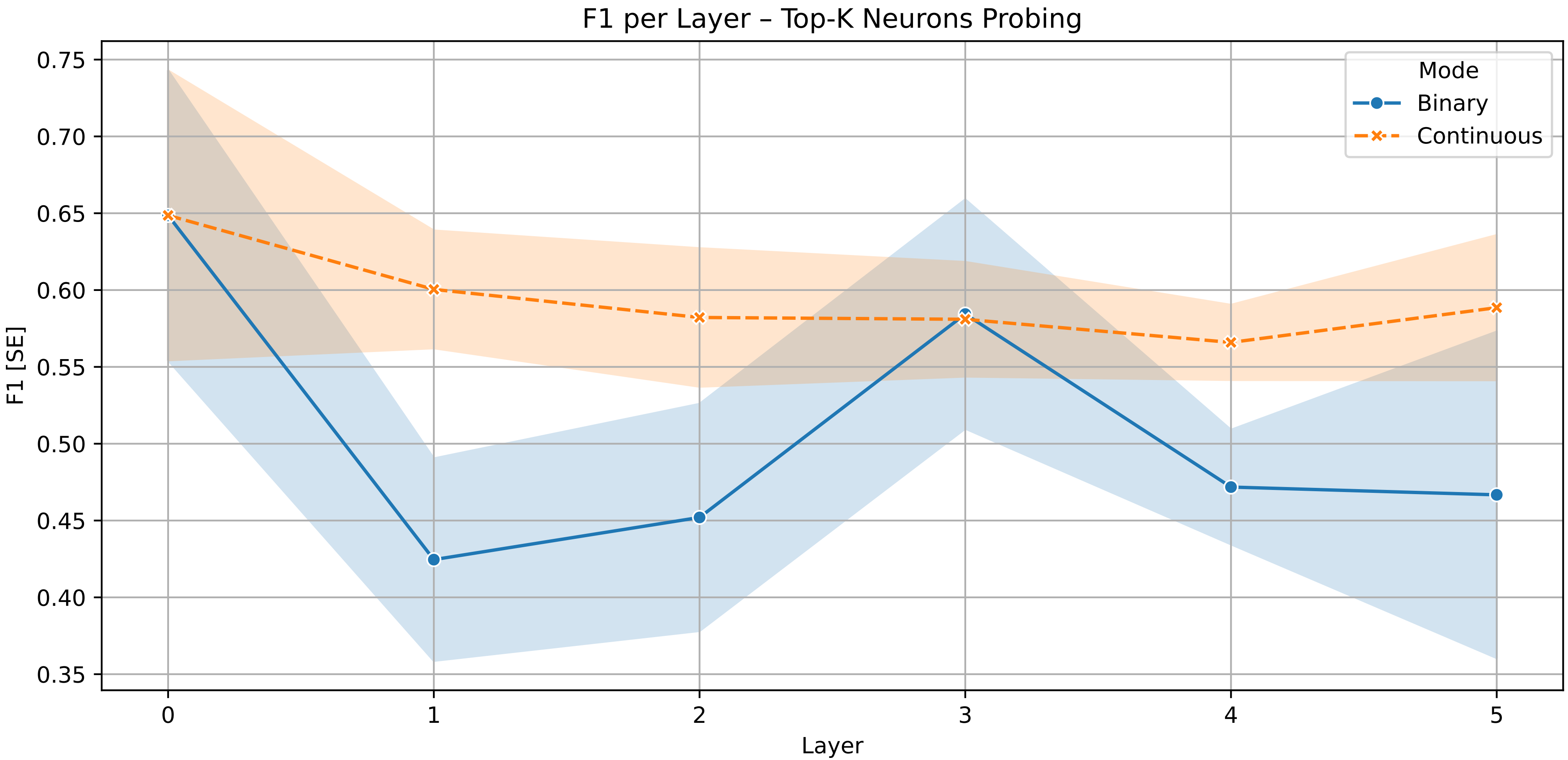
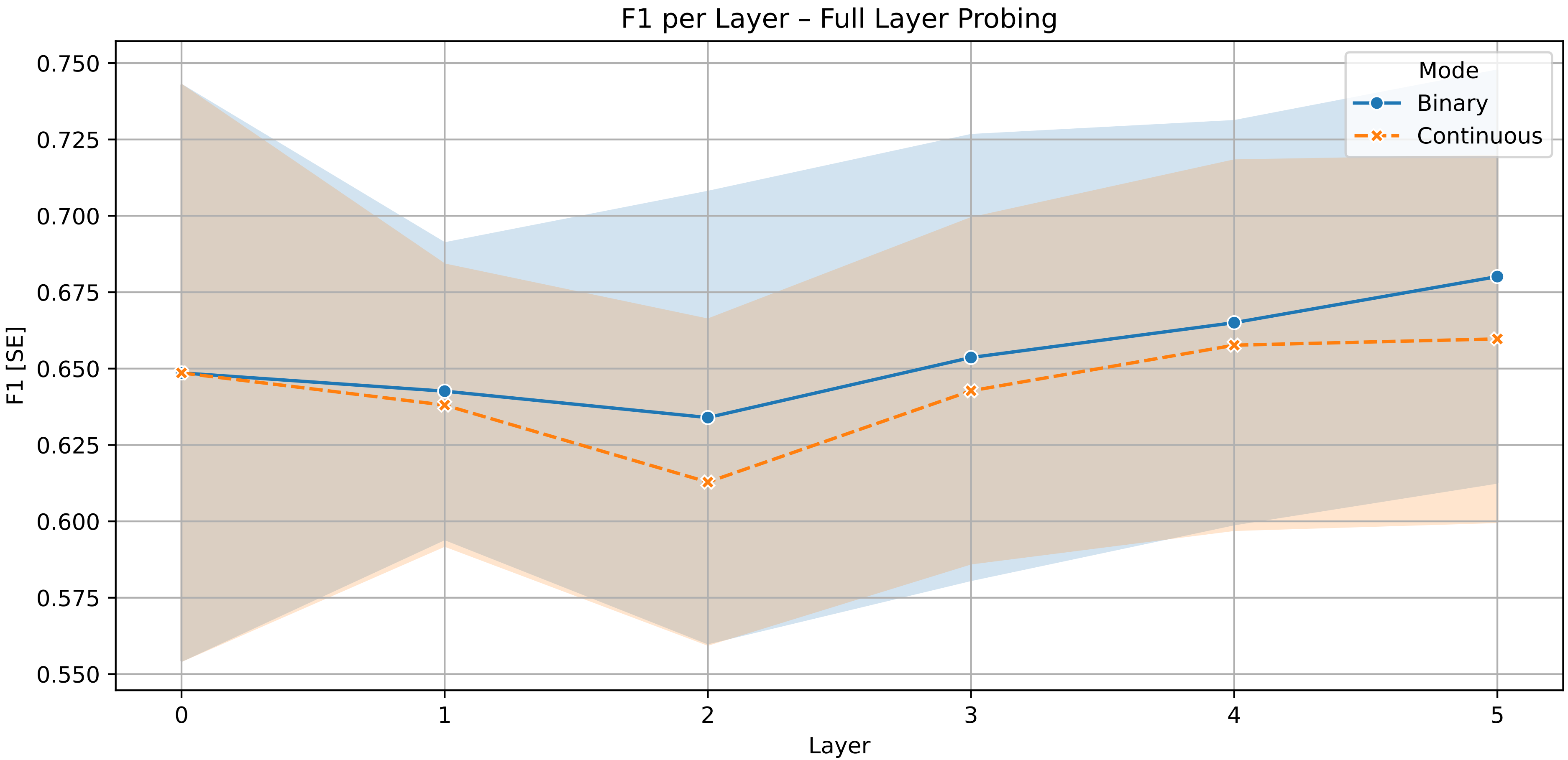
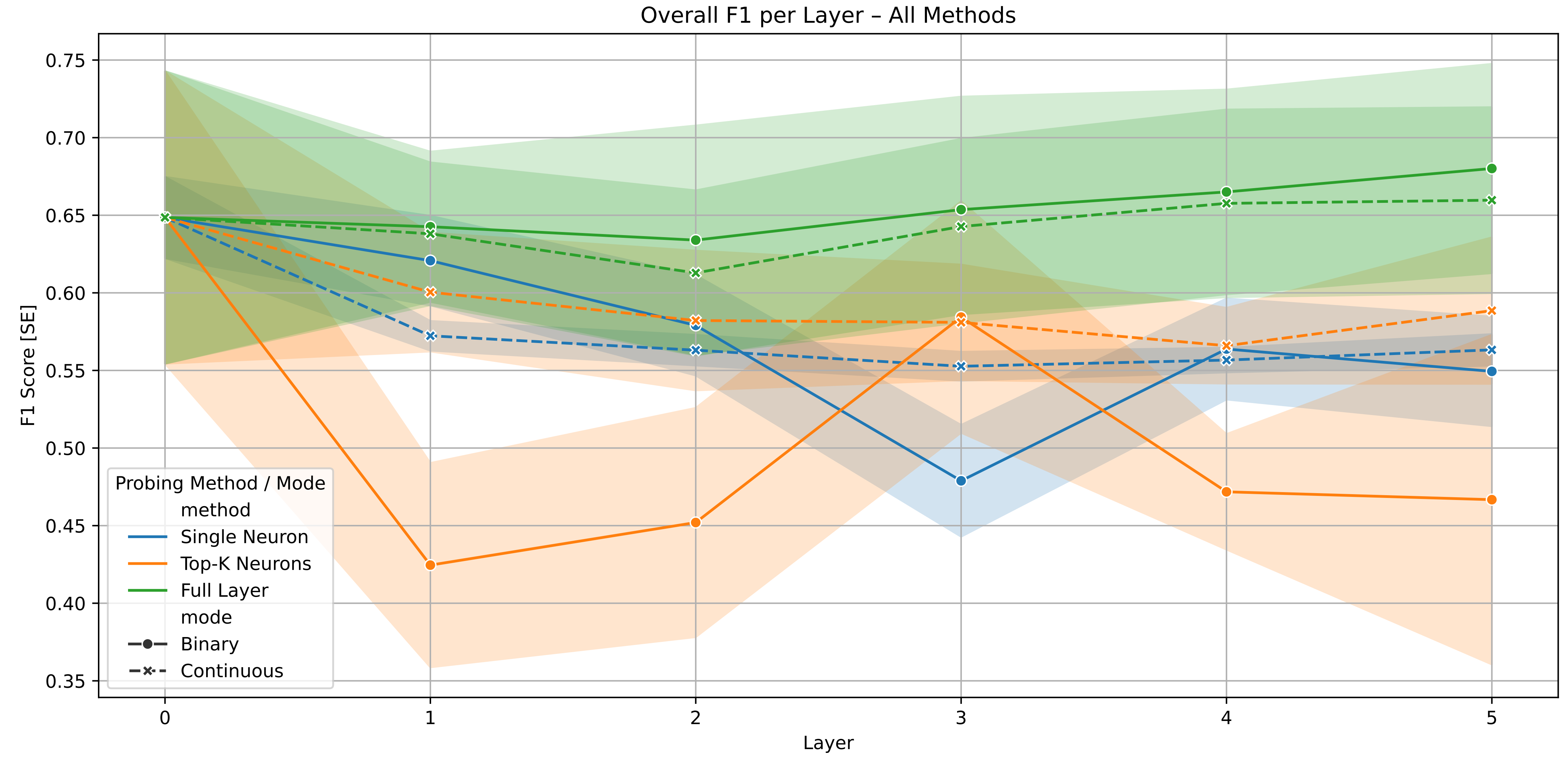
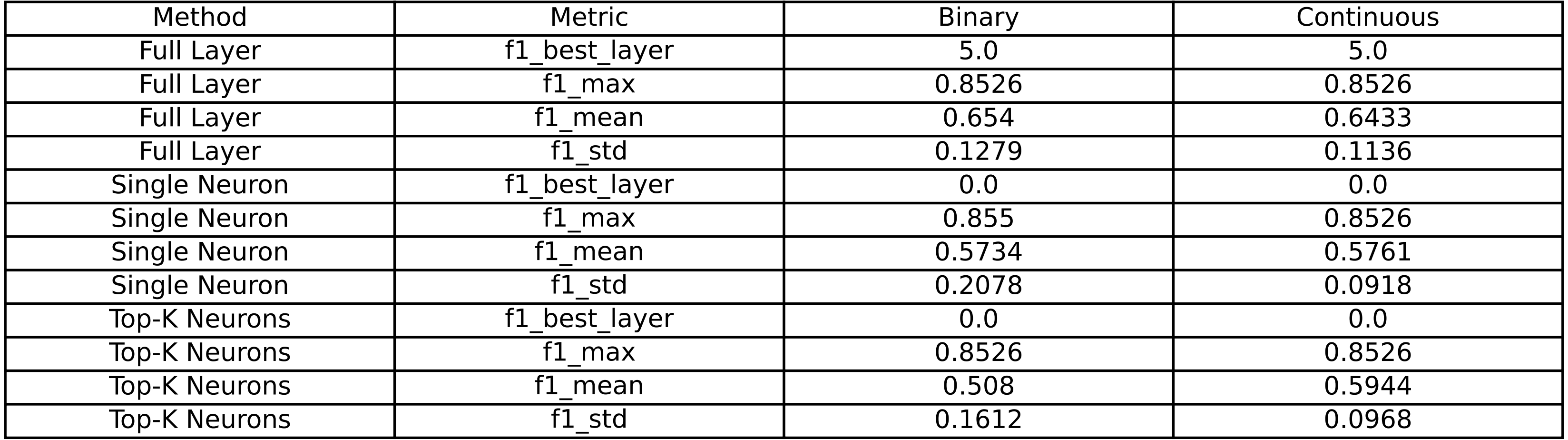

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.65
0.60
0.55
0.50
0.45
Layer
0
1
2
3
4
5


F1 per Layer – Top-K Neurons Probing
F1 [SE]
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
Layer
0
1
2
3
4
5
Mode
Binary
Continuous

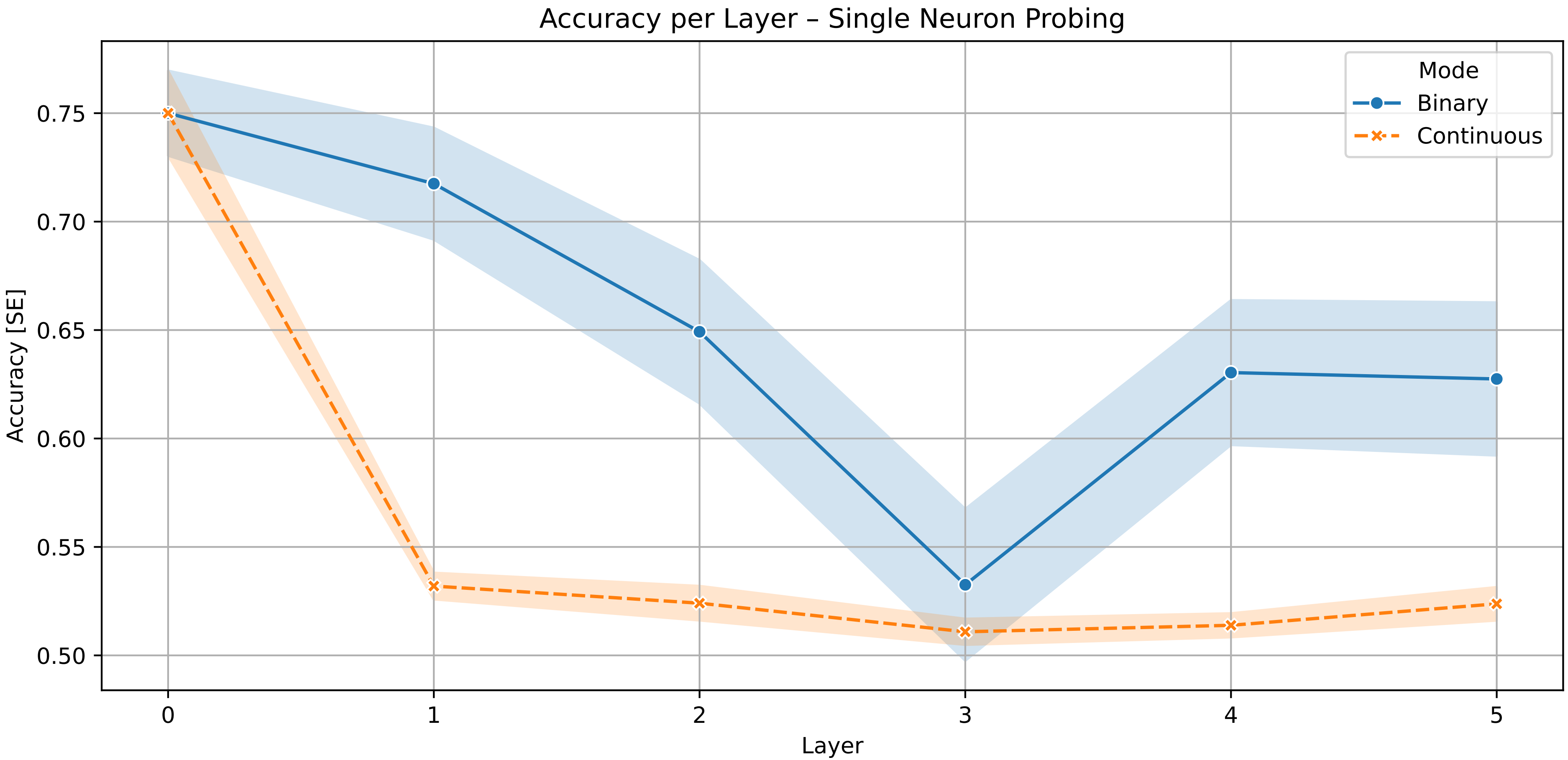

F1 per Layer – Full Layer Probing
F1 [SE]
0.750
0.725
0.700
0.675
0.650
0.625
0.600
0.575
0.550
Layer
0
1
2
3
4
5
Mode
Binary
Continuous

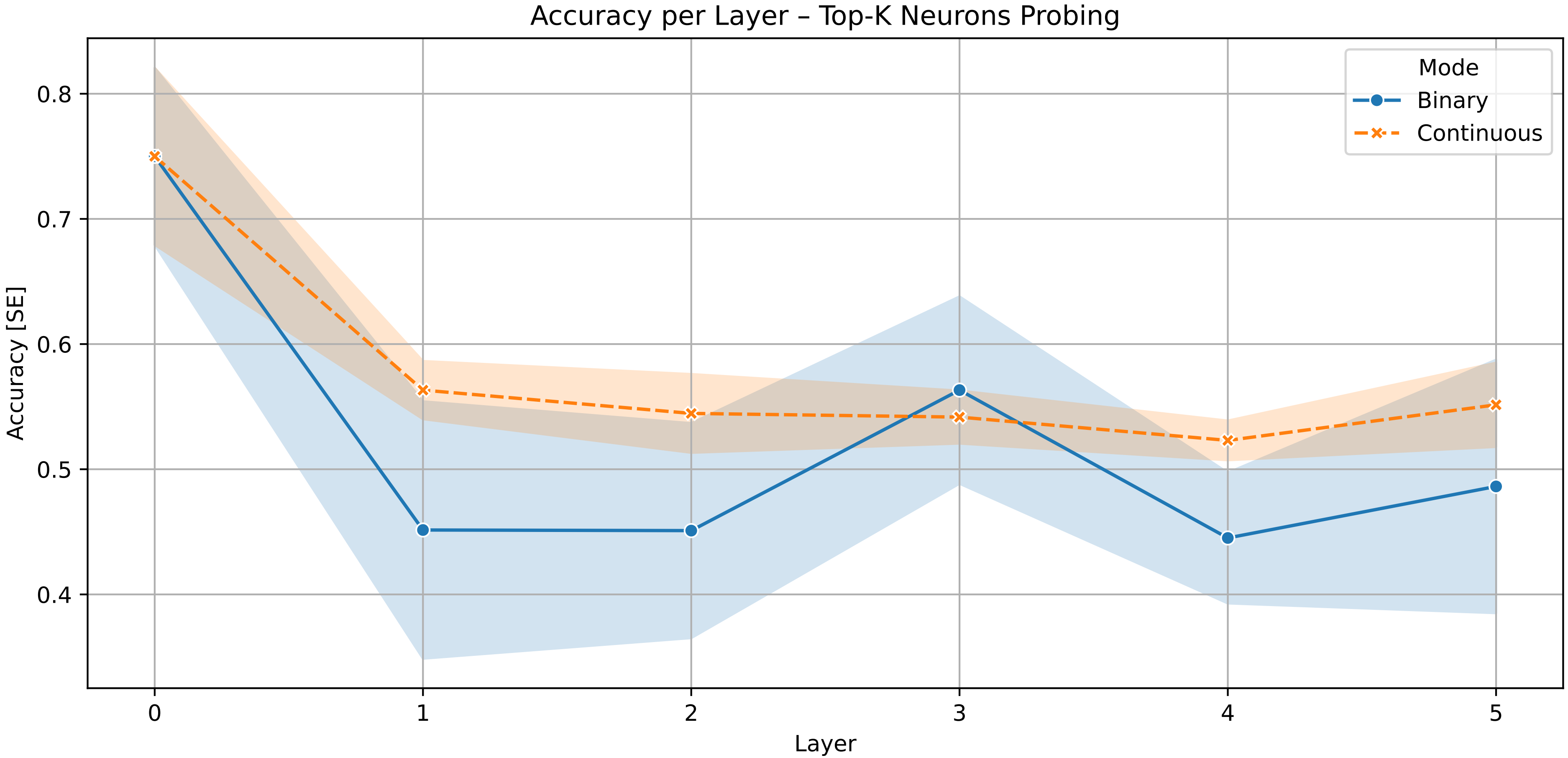

Overall F1 per Layer – All Methods
F1 Score [SE]
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
Layer
0
1
2
3
4
5
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

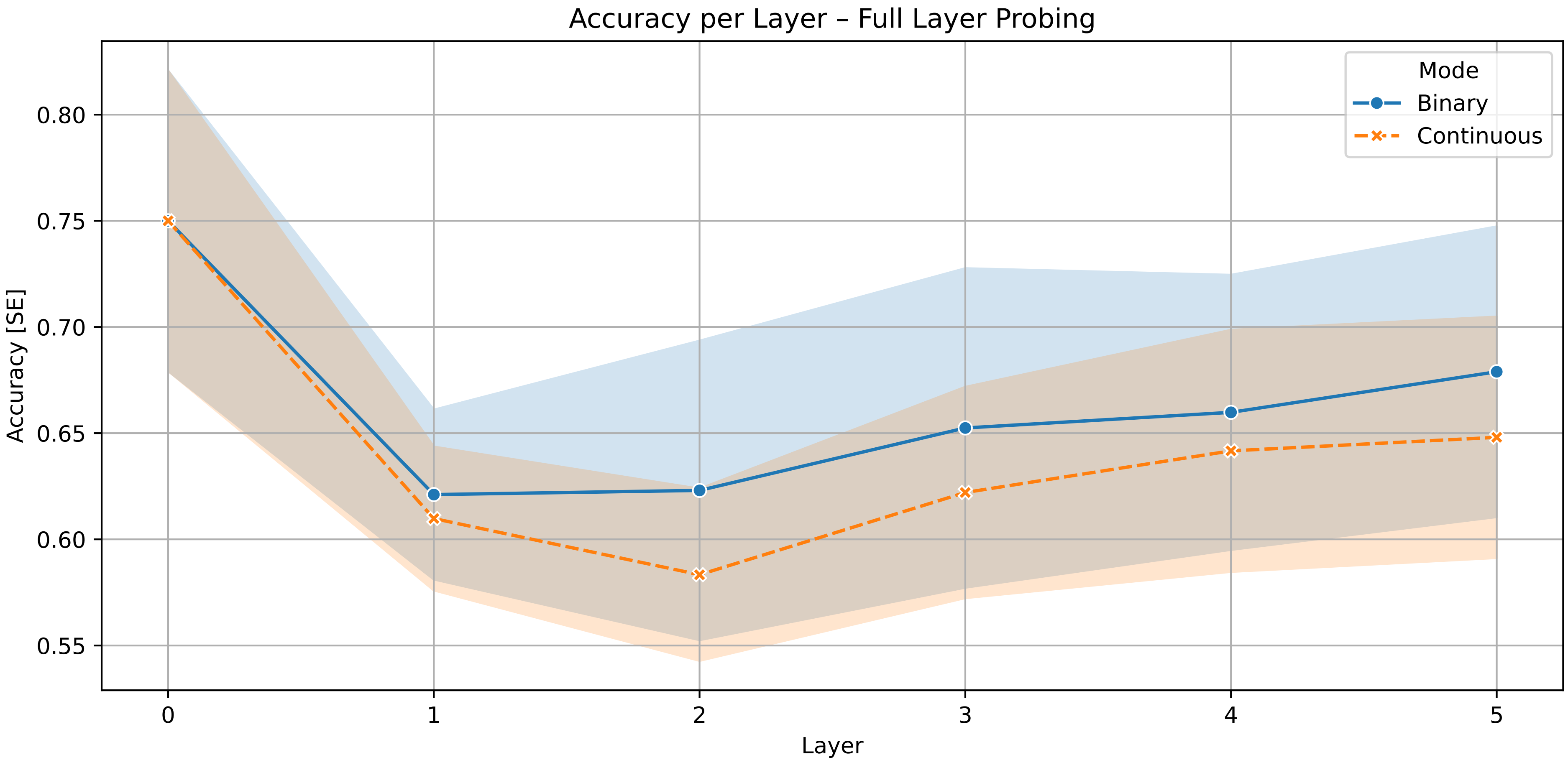
F1 Score Summary by Probing Method

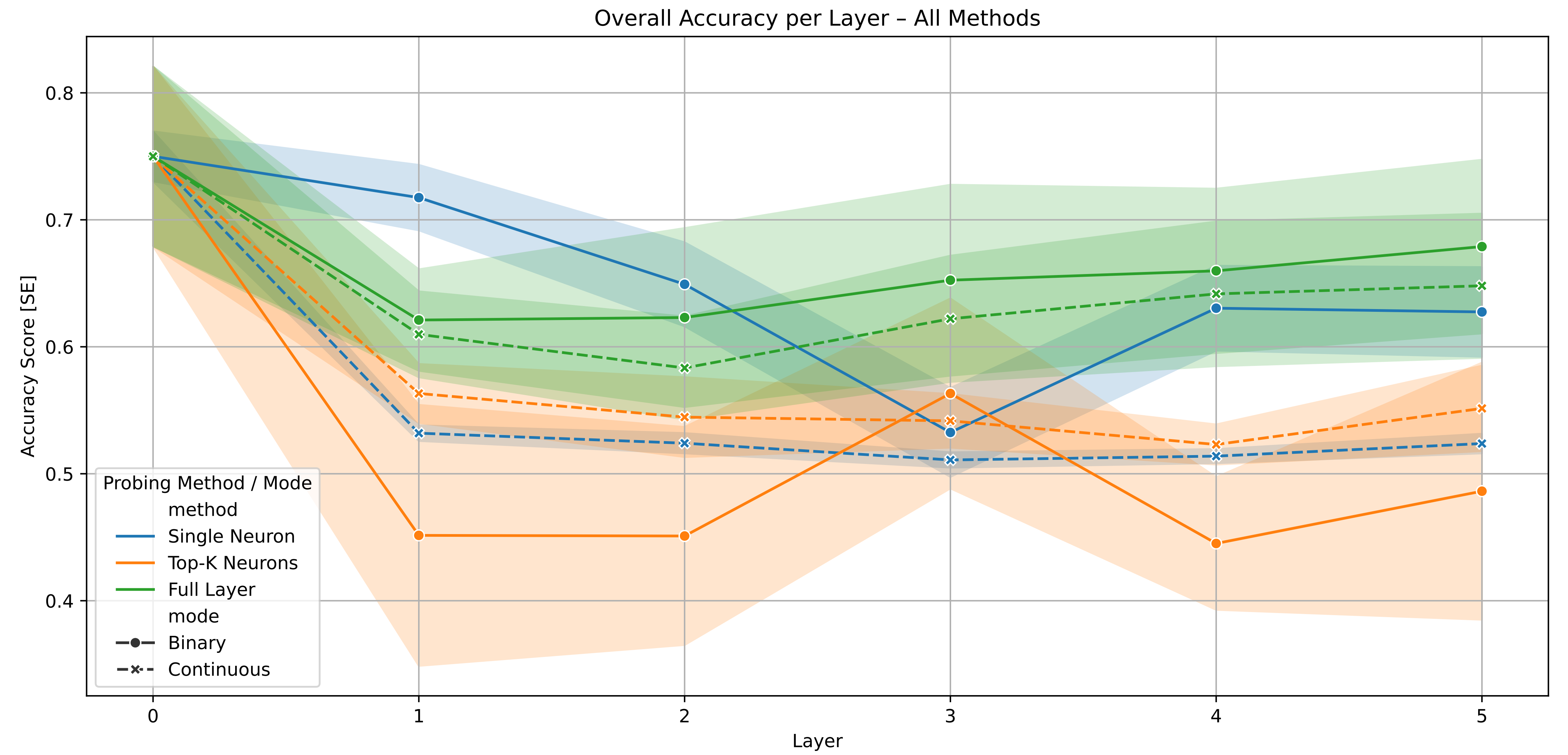


Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	5.0	5.0
Full Layer	f1_max	0.8526	0.8526
Full Layer	f1_mean	0.654	0.6433
Full Layer	f1_std	0.1279	0.1136
Single Neuron	f1_best_layer	0.0	0.0
Single Neuron	f1_max	0.855	0.8526
Single Neuron	f1_mean	0.5734	0.5761
Single Neuron	f1_std	0.2078	0.0918
Top-K Neurons	f1_best_layer	0.0	0.0
Top-K Neurons	f1_max	0.8526	0.8526
Top-K Neurons	f1_mean	0.508	0.5944
Top-K Neurons	f1_std	0.1612	0.0968

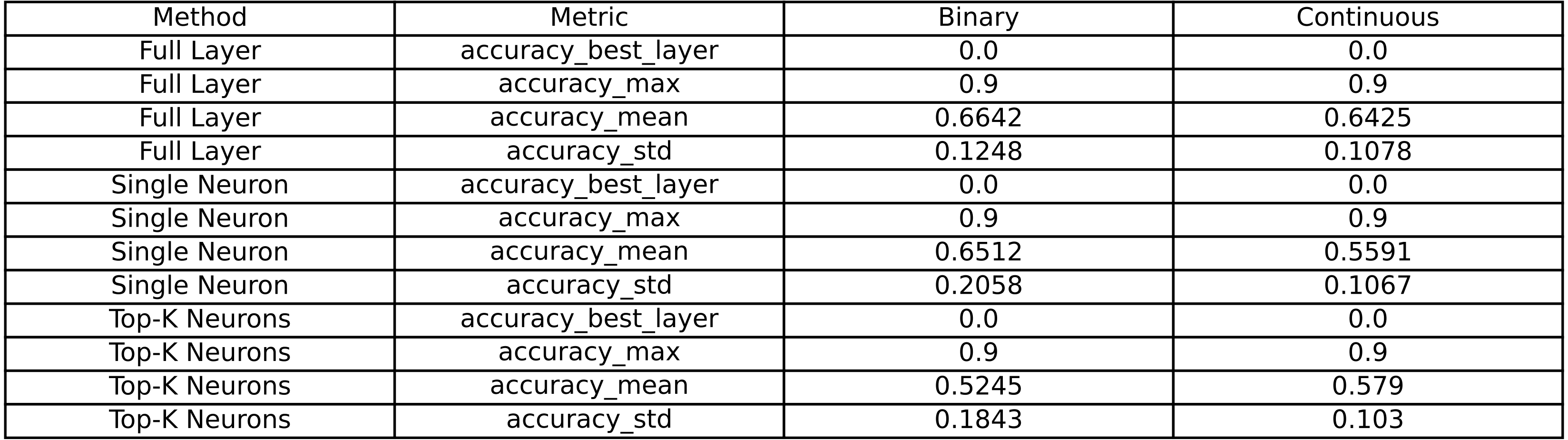

Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
0.75
0.70
0.65
0.60
0.55
0.50
Layer
0
1
2
3
4
5
Mode
Binary
Continuous


Accuracy per Layer – Top-K Neurons Probing
Accuracy [SE]
0.4
0.5
0.6
0.7
0.8
Layer
0
1
2
3
4
5
Mode
Binary
Continuous


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.80
0.75
0.70
0.65
0.60
0.55
Layer
0
1
2
3
4
5


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
0.4
0.5
0.6
0.7
0.8
Layer
0
1
2
3
4
5
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	0.0	0.0
Full Layer	accuracy_max	0.9	0.9
Full Layer	accuracy_mean	0.6642	0.6425
Full Layer	accuracy_std	0.1248	0.1078
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.9	0.9
Single Neuron	accuracy_mean	0.6512	0.5591
Single Neuron	accuracy_std	0.2058	0.1067
Top-K Neurons	accuracy_best_layer	0.0	0.0
Top-K Neurons	accuracy_max	0.9	0.9
Top-K Neurons	accuracy_mean	0.5245	0.579
Top-K Neurons	accuracy_std	0.1843	0.103